7450 Ethernet Service Switch
7750 Service Router
7950 Extensible Routing System

Advanced Configuration Guide - Part I
Releases Up To 16.0.R4

3HE 14990 AAAA TQZZA 01

Issue: 01

December 2018

The information presented is subject to change without notice. No responsibility is assumed for inaccuracies contained herein.

Table of Contents

List of tables

List of figures

IEEE 1588 for Frequency, Phase, and Time Distribution 31

Synchronous Ethernet..65

Distributed CPU Protection 83

Event Handling System 107

SR OS NETCONF Server Basics 121

Multi-Chassis APS and Pseudowire Redundancy Interworking....................127

BFD for RSVP-TE LSPs with Failure-Action 933

Class-Based Forwarding 949

DiffServ Traffic Engineering 985

Entropy Label 1023

IGP Shortcuts 1037

MPLS Transport Profile 1247

Multicast Label Distribution Protocol 1273

Path MTU Discovery 1307

PCEP Support for RSVP-TE LSPs 1327

RSVP Point-to-Point LSPs 1353

Preface

About This Guide

The Advanced Configuration Guide is divided into three volumes, the Part I Guide, the Part II Guide, and the Part III Guide.

- Part I provides advanced configurations for basic systems, system management, interface configuration, router configuration, unicast routing protocols, and MPLS.
- Part II provides advanced configurations for services overview, Layer 2 and EVPN services, Layer 3 services, and Quality of Service.
- Part III provides advanced configurations for Multi-Service Integrated Service Adapter and Triple Play Service Delivery Architecture.

The guide is organized alphabetically within each category and provides feature and configuration explanations, CLI descriptions and overall solutions. The chapters in the Advanced Configuration Guide are written for and based on several releases, up to 16.0.R4. The Applicability section in each chapter specifies on which release the configuration is based.

The Advanced Configuration Guide supplements the user configuration guides listed in the 7450 ESS, 7750 SR, and 7950 XRS Documentation Suite Overview Card.

Audience

This manual is intended for network administrators who are responsible for configuring the routers. It is assumed that the network administrators have a detailed understanding of networking principles and configurations.

Basic System

In this section

This section provides configuration information for the following topics:

- IEEE 1588 for Frequency, Phase, and Time Distribution
- Synchronous Ethernet

IEEE 1588 for Frequency, Phase, and Time Distribution

This chapter provides information about IEEE 1588 for frequency, phase, and time distribution.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This section is applicable to all of the 7750 SR and 7450 ESS series, except for the SR-1, ESS-1, and ESS-6/6v. It is not applicable to t.he 7710 SR nor the 7950 XRS series. Description and examples are based on release 12.0.R2. The only software pre-requisites are IP reachability between the node and neighboring 1588 clocks.

IEEE 1588 has several hardware dependencies both for the basic functionality as well as the 1588 port based timestamping necessary for high accuracy time distribution. Please consult the related Nokia documentation for the details of all the hardware requirements.

Overview

Defined in IEEE Std 1588™-2008 (1588v2), Precision Time Protocol (PTP) is a protocol that distributes frequency, phase and time over packet based networks.

Note: Many applications do not need time alignment but only phase alignment. However, phase is derived from time and so, for the remainder of the document, discussion refers to time but those references imply both time and phase.

The IEEE 1588 protocol has become the standard for distribution of high accuracy time. Following guidelines for specific network architectures allows the delivery of time to accuracies of one microsecond. This level of accuracy is required for mobile base stations using either Time Division Duplex technology and/or advanced LTE functions, as well as in the power industry for intelligent electronic device alignment.

More lenient architectures can still achieve 100 microseconds or better accuracies which can greatly enhance the usefulness of event logging and network one way delay measurements.

In addition, 1588 has been used to deliver a frequency reference for T1/E1 ports or for mobile base station frequency alignment. This is useful in environments where the transport network does not provide physical layer synchronization services.

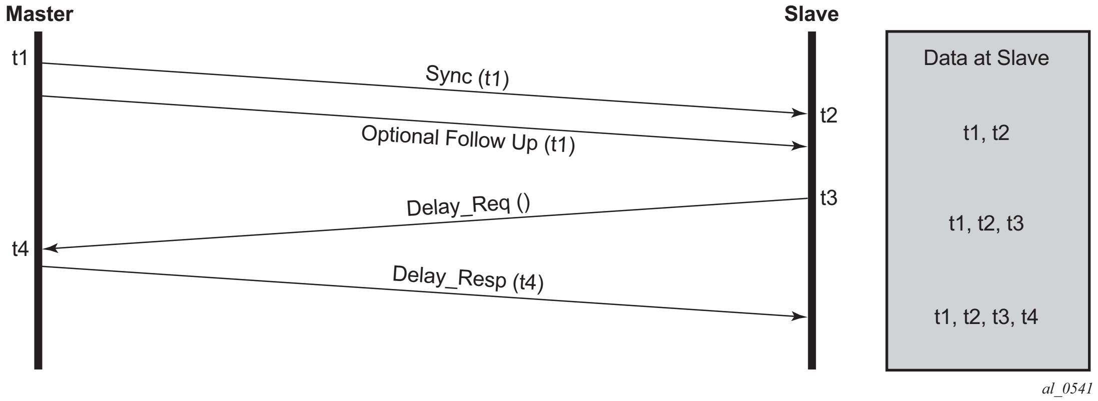
The following 1588 capabilities are provided within the 7750 SR and 7450 ESS nodes:

- CPM/CFM based 1588 master, boundary, and slave clock functionality
- Transport over Unicast UDP/IPv4 packets
- Access to 1588 process through base routing, IES, and VPRNs
- Port based timestamping of 1588 packets
- IEEE 1588 Profiles: 2008 standard default and ITU-T G.8265.1
- Utilization of 1588 derived time for NTP and System time.

PTP Basics

PTP uses an exchange of four timestamps between a reference clock (master port) and the clock to be synchronized (slave port). A simplified illustration of this mechanism is shown in Figure 1.

Figure 1 **PTP Messages and Timestamp Exchange**



The master sends a PTP Sync message containing a timestamp of when the Sync message is transmitted (t1) to the slave. In a two-step master clock, the t1 timestamp is sent in a Follow_Up message. The slave records the time it receives the Sync message (t2). At some point after receiving the Sync message, the slave sends a Delay_Req message back to the master. The slave records the time of transmission of the Delay_Req message (t3) locally. The master records the time it receives the Delay_Req message (t4) and sends this timestamp back to the slave in a Delay_Resp message.

Note: The Follow_Up message was defined to allow for implementations to generate a timestamp for the transmission of the Sync message but not have to try to insert that timestamp into the Sync message and update any frame checksums on the fly as it is in the process of transmission. While many recent implementations can perform the timestamping, update and checksum calculation on the fly, not all devices could perform this three step process with the desired accuracy. By using the Follow_Up message to transmit the timestamp of the Sync message, the master port can still provide extremely accurate timestamps for the transmission of the Sync message to the slave port. Apart from the extra message required, there is no detriment to a master port using one-step clock versus a two-step clock procedures. All PTP clocks that have slave port capability must accept timing information from both types of master port. There is no requirement to force a clock that is a one-step clock to use two-step clock procedures on its master ports. The nodes covered by this example all support one-step clock master port procedures.

After the four timestamp exchange the slave can calculate the mean path delay and the clock offset from master using the following two equations:

- mean_path_delay = [(t4-t1) – (t3-t2)] /2
- offset_from_master = [(t2 –t1) – mean_path_delay]

These calculations can occur on every message exchange or some initial packet selection can be performed so that only optimal message exchanges are used. The latter is useful if there is variable delay between the master and slave ports.

If only frequency is necessary, then the slave may use one or both pairs of timestamps (t1, t2) and (t3, t4). The slave can monitor the change in the perceived delay master-to-slave (t2 - t1) or slave-to-master (t4 - t3) over time. If the delay (t2 – t1) decreases over time, it means the t2 timestamps are not progressing quickly enough and the slave clock frequency needs to be increased.

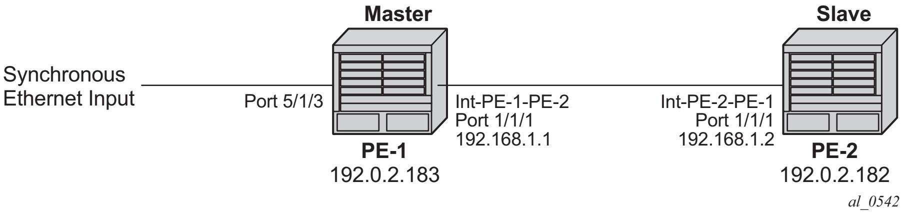
If time is necessary, then all four timestamps must be used. It is also important to note how the equation for offset uses the mean_path_delay. If the delays in the two directions are actually different, then the equation will introduce an error in the offset_from_master that is half of the difference of the two delays. The IEEE 1588 standard includes procedures to compensate for this asymmetry, if it is known, but if it is uncompensated it does introduce time error.

PTP Deployment Architectures

It is important to understand that there are very different topologies recommended for using 1588 for frequency distribution and using 1588 for time distribution.

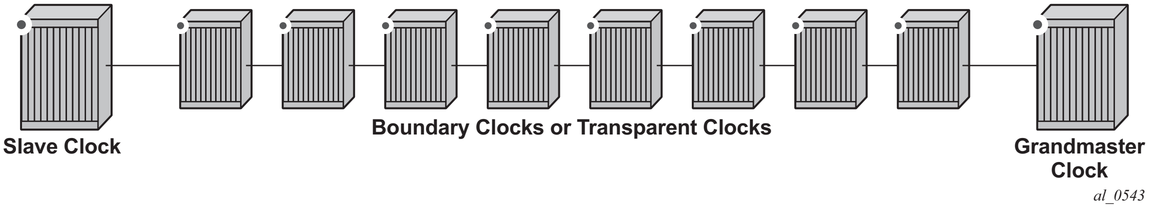
Frequency distribution was developed for an architecture where there are mobile providers who have points of presence at the mobile telephony switching offices (MTSOs) and the cell site locations which depend on other parties for the connectivity between the MTSOs and the cell site locations. The mobile providers wanted a solution that could span the transport networks with minimal dependence on that network. This can be achieved by placing a 1588 grandmaster at the MTSO and a slave in a cell site router or directly in the basestation and distributing the timestamped packets between the two, as shown in Figure 2. The transport network does introduce packet delay variation (PDV) to the 1588 messages which makes it more difficult to track the frequency of the grandmaster's clock. However, the slaves have been designed to perform packet selection and noise filtering to allow for the recovery of a frequency within the required accuracies of the mobile basestations. This architecture and the performance requirements are covered by the ITU-T G.826x series of recommendations.

Figure 2 **1588 Topology for Frequency Distribution**



For time distribution, it has been recognized that the architecture used above is extremely unlikely to be successful. The fundamental reason is that the performance requirement is much tighter and the network introduces not only PDV but also potentially asymmetric delay which causes time error in the slave. The topology recommended for time distribution is what is sometimes referred to as “Full On-Path Support (OPS)”. Full OPS means that every network element between the grandmaster clock and the slave clock is either a 1588 boundary clock or a 1588 transparent clock, as shown in Figure 3. Boundary clocks and transparent clocks process the 1588 messages and remove the PDV noise that would be present in a non 1588 network element. By using network elements that have very tight constraints on the time error they introduced, the network can be built to guarantee time accuracy under all network traffic conditions. This architecture and the performance requirements are covered by the ITU-T G.827x series of recommendations.

Figure 3 **1588 Topology for Time Distribution**



PTP Profiles

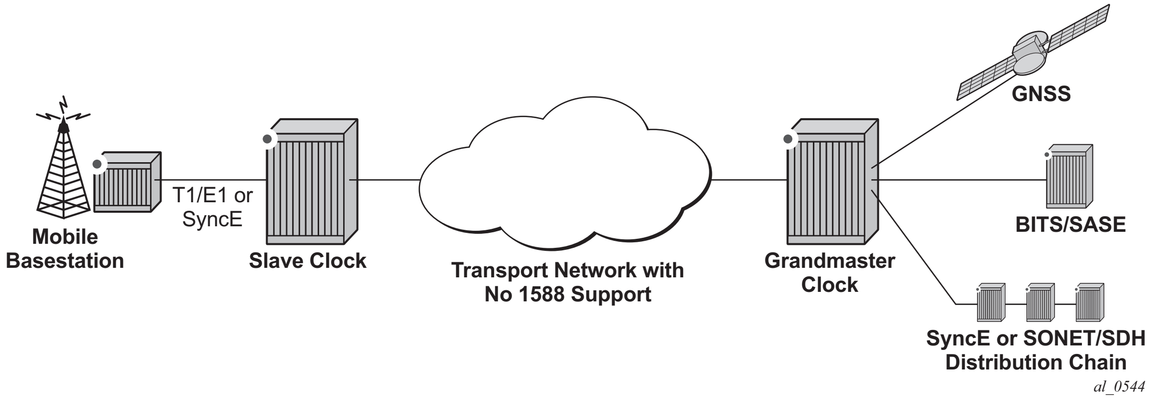
The 1588v2 standard includes the concept of a PTP profile. A PTP profile allows standardization bodies or industry groups to adapt the 1588v2 standard to a particular application. A profile defines which aspects of the 1588v2 standard are included or excluded, along with configurable ranges and defaults necessary for the application.

The 1588 standard itself includes a **default** profile that can be used for either time or frequency distribution. The default profile was defined principally for multicast operation. However, it can be used with the unicast sessions as described below. The default profile supports all 1588 clock types and includes the Best Master Clock Algorithm (BMCA) that automatically builds the synchronization distribution hierarchy amongst the PTP clocks. The SR OS only supports the unicast session version of the default profile.

In the telecommunications industry, the ITU-T is the body that develops these profiles. They have generated a profile for frequency distribution (G.8265.1) and a profile for time distribution (G.8275.1). The frequency profile permits only grandmaster and slave clocks and can be used to extended a traditional physical layer synchronization distribution (SONET/SDH, PDH, or SyncE) with a final leg of 1588 messages. The frequency source of the 1588 grandmaster could be a GPS receiver, a central office BITS or SASE device or it could use the frequency recovered from a Synchronous Ethernet or SONET/SDH interface. This is shown in Figure 4

Because a 1588 distribution system is significantly noisier than a physical layer distribution system, it should only be used as the final segment to connect the end application into the synchronization network. It should not be used to connect two Synchronous Ethernet or SONET/SDH islands.

Figure 4 **Frequency Distribution with 1588 as Last Mile**



The important features defined in the G.8265.1 profile are:

- Only master clocks and slave clocks are allowed.
- Unicast Message Negotiation using Signaling messages from the slave clocks toward the master clocks is used to establish communications.
- PTP messages are encapsulated over UDP over IPv4.
- PTP clock class values are based on a mapping of traditional quality levels from SSM/ESMC.

Note: SSM stands for Synchronization Status Messages and ESMC stands for Ethernet Synchronization Messaging Channel. These are two capabilities in SDH/SONET and Synchronous Ethernet respectively for the relaying of source clock quality information.

The slave clock uses an alternate BMCA to select the grandmaster clock from the available master clocks based on:

- Quality Level.
- Relative Priority.

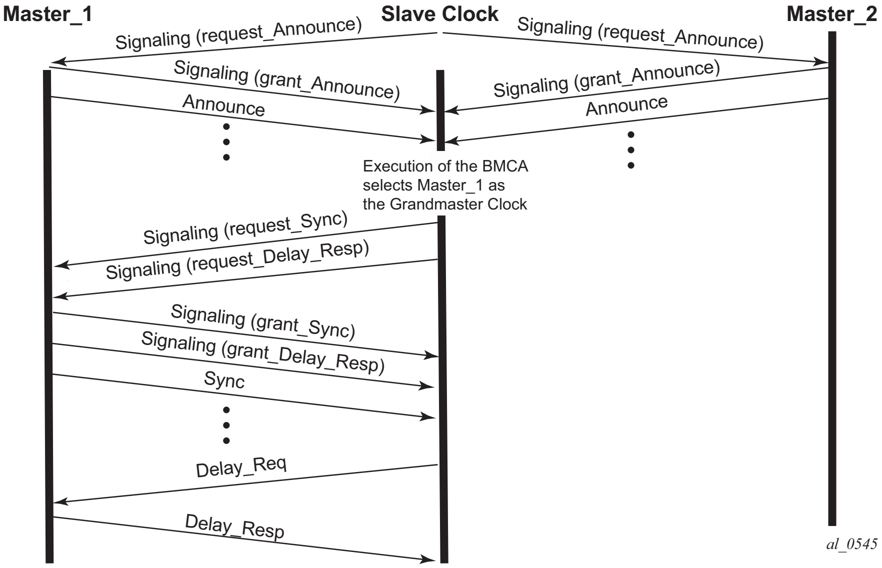
The ITU-T has defined the first time distribution profile in G.8275.1. It uses an architecture of a Global Navigation Satellite System (GNSS) based grandmaster clock distributing time through a chain of boundary clocks to a final slave device and end application. It includes the use of Synchronous Ethernet and 1588 at the same time for optimal performance. Physical layer Synchronous Ethernet is an excellent tool for the distribution of an accurate and stable frequency. This frequency can be used to advance time between offset adjustments made using the 1588 information.

Unicast Message Negotiation

The initial IEEE 1588-2002 standard defined a multicast messaging model. IEEE 1588-2008 introduced the option of using unicast messaging with unicast discovery to establish a message exchange between a master and slave.

The typical unicast message flow between a master and slave is illustrated in Figure 5.

Figure 5 **Unicast Message Negotiation**



A slave clock initiates unicast discovery by sending a Signaling message to one of its configured master clocks requesting the master send unicast Announce messages to the slave. The request includes the desired rate for the Announce messages and the duration over which the messages should be sent. If the master can support the request it replies with a Signaling message indicating that the session for unicast Announce messages has been granted.

From this point on, the master sends unicast Announce messages to the slave at the rate requested. A slave will generally establish an Announce message session with at least two master clocks.

The slave then uses the Announce messages it receives from all masters as input to the BMCA that determines which master clock is the best source for information. The selected master becomes the grandmaster clock to the slave. The slave then sends additional Signaling messages to the grandmaster to request unicast delivery of Sync and Delay_Resp messages. Assuming the grandmaster clock has sufficient resources, the request is granted and unicast Sync and Delay_Resp messages are sent from the grandmaster to the slave.

As with the Announce messages, the rate at which the Sync and Delay_Resp messages are sent and the duration of the unicast sessions is requested by the slave in the initial Signaling messages.

The unicast sessions for Announce, Sync and Delay Response messages have an expiry time. The slave renews all three sessions before this time is reached.

Network Limits

A common concern around 1588 is whether it will work on or over a specific customer network. For time distribution using full OPS as shown in Figure 3, there are well defined limits on the number of network elements allowed in the distribution chain (see below). However, for the frequency distribution using the architecture shown in Figure 2, it is a more difficult question to answer. There are so many different types of network elements and inter-node links that a simple limit on the number of network elements is not adequate. What has been specified is a limit to the noise that the network can introduce to the 1588 message flow between the grandmaster and slave clocks. This noise occurs as packet delay variation (PDV). The following sections provide some description of this PDV and a new metric that has been defined for PDV as well as the recommended limit to PDV for 1588 deployments.

Packet Delay Variation

If the packet delay through the packet network is constant, then it is relatively easy to use a series of timestamp exchanges to remove the delay as an unknown and track the master clock frequency. However, in most network technologies, the packet delay will be different for each individual packet. This PDV makes it more difficult to track the master clock since observations have both the master information and PDV noise included.

PDV is introduced when packets get placed in queues before they are forwarded. The time each packet sits in any one queue is influenced by multiple factors:

- the speed of the interface toward which the queue drains, for example 100Mbps versus 100Gbps,
- the traffic load on the interface, for example 20% versus 100% of line capacity,
- the distribution of packet sizes and priorities in the traffic load toward the interface, and
- the underlying physical technology used, xPON, xDSL, Ethernet, or microwave.

In addition, the load and packet distribution within the load will vary over time so the distribution of the PDV can shift rapidly such as when a network event triggers congestion or slowly, for example as end customers gradually come online over a period of several hours.

Also there are pipeline effects that can occur in a chain of queuing devices, where the small timing packets can catch up to a large packets moving across the network. Once behind such a packet, the timing packet can remain stuck behind that packet on all subsequent transmit queues.

QoS prioritization of packets helps reduce PDV significantly during congestion periods, but does not remove the PDV effects during lighter loading. This is due to the fact that a timing packet may be delivered to the egress queue for an interface while the interface is busy transmitting a packet. Pre-emption of packet transmissions is not used in today's networks.

Having stated all of the above, most of the time, the network will still present a percentage of packets that get across the network with minimal queuing delays. These are often referred to as 'lucky' or 'fastest' packets. Since these lucky packets are never waiting in queues or have minimal wait times, their transit across the network is relatively consistent. By running a selection filter on all 1588 packets to find these lucky packets, a level of variation of network delay can be removed or reduced significantly. Then the slave clocks have a much easier time determining the frequency of the grandmaster.

However, there will always be a limit to the amount of PDV that can be tolerated. The ITU has defined a metric to quantify the PDV, the limit of the PDV for a compliant network, and the required tolerance of a slave clock.

PDV Metrics

In order to know whether a particular timing-over-packet implementation will meet the performance targets in a given network deployment, it is desirable to both characterize the limits on the PDV that the implementation can tolerate and to measure the network against these limits. In 2012, the ITU-T published three documents that address these requirements:

- G.8260 defines the Floor Packet Percentage (FPP) metric.
- G.8261.1 defines a network limit for PDV in terms of FPP.
- G.8263 defines the input tolerance expected of a 1588 frequency slave in terms of FPP.

The Floor Packet Percentage (FPP) metric provides an indication of the guarantee that there are packets experiencing minimal delay across the network. The rationale behind this focus on 'fastest' packets is that many networks do provide good consistency of these packets in most operating conditions and because most slave clocks are capable of operating using only the information from these fastest packets.

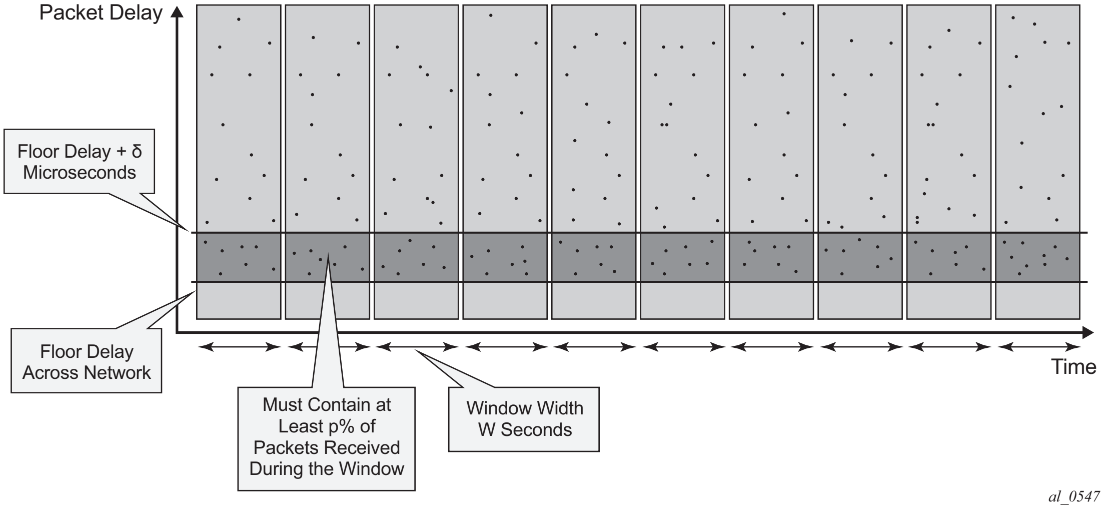
There are four parameters associated with the metric:

- **W** is the width of the windows used to monitor for the presence of fastest packets.

- **Floor Delay** is a value that is as close as possible to the absolute minimum transit delay across the network. Every actual delay measurement must be equal to or larger than this value.
- δ is the range above the floor to be analyzed for the presence of fastest packets.
- ρ is the percentage of all the packets received in a window whose delay must be within the range floor delay to floor delay + δ.

Figure 6 illustrates how these parameters and the metric work. First the delays of all individual 1588 packets are plotted over the period of observation. Next the observation period is broken down into a series of consecutive windows of width **W** seconds. Then for each window a count is made of all the 1588 packets whose delays are within the range **floor delay** to **floor delay + δ** and this count is compared with all the 1588 packets received during the window to turn the count into a percentage. Finally the percentage of each window is checked against the threshold percentage ρ. For the FPP metric to be met, every window must have a percentage greater or equal to the threshold. If even one single window does not meet this threshold then the metric condition is not met.

Figure 6 **Floor Packet Counting for FPP (n, W, δ)**


Packet Delay
Floor Delay + δ Microseconds
Floor Delay Across Network
Must Contain at Least p% of Packets Received During the Window
Window Width W Seconds
Time
al_0547

Note: This metric is not perfect as it does not take into account slaves that use other aspects of the packet delay distribution (such as average delay), nor does it discuss the impact of reroutes, nor do the limits discuss how to apply these limits to the forward and backward message exchanges at the same time. However, it was agreed that this metric was a good start for the definitions of the network and slave limits. Expect to see timing test equipment vendors providing the tools to generate 1588 PDV profiles providing FPP based distributions.

ITU-T Budget for Frequency

The network limit on PDV for frequency distribution is defined in G.8271.1 using the FPP metrics defined above.

In general most carrier grade networks with spans of up to 10 nodes and which do not exceed 80% load on their internode links should meet the requirement. However, very low (sub 50 Mbps) shaping or very long networks or last mile technologies such as xDSL or xPON may need to be studied to determine their acceptability.

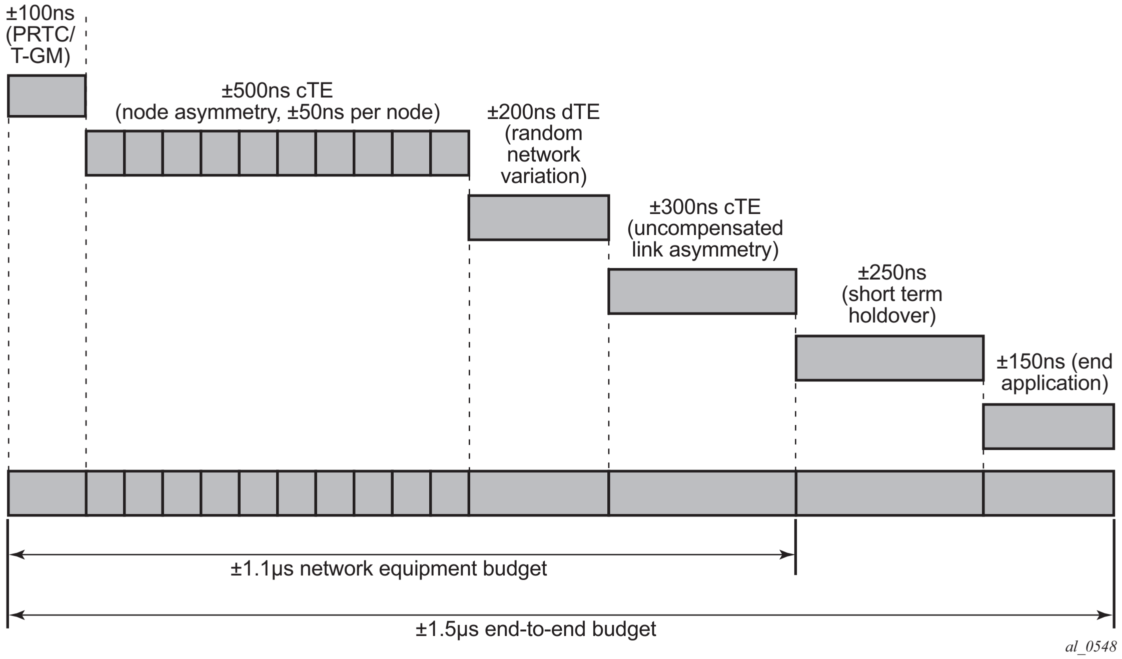
A general strategy for rolling out 1588 frequency distribution is to evaluate the specific grandmaster and slave pairing in a lab environment using a network emulator to introduce controlled PDV. Once the grandmaster and slave have passed the lab tests, then field trial locations should be identified. Ideally, the sites should include locations where the PDV of the network will likely be at its worst. This would be sites with the most intervening network elements between the grandmasters and the slaves and include segments of the network that have a high load. The slaves' clocks should be deployed and monitored over several days to ensure that their frequency recovery engines can maintain lock with the grandmasters. During the initial field trails, it is beneficial to use external frequency test equipment at the slave locations to accurately monitor the frequency generated out of the slaves and ensure it stays within limits. As more sites are evaluated and confidence in the PDV environment increases, more deployments can be rolled out. In the deployed network, PTP frequency recovery slave states can be monitored to ensure the solution continues to work.

There may be some locations in the network where the PDV will be too large preventing the slaves to achieve or maintain lock. If it is possible to utilize an alternate network interface to obtain a frequency such as a leased T1 or E1 interface then that could be used. A last resort would be the deployment of a GNSS receiver at the location to provide the frequency reference.

ITU-T Budget for Time

The ITU-T has defined a topology for time distribution based on a full OPS environment. This means that every network element in the time distribution chain is a 1588 clock of some type. Currently the work has defined an environment using Boundary Clocks, but this might be modified in the future to include transparent clocks. The ITU-T tackled the time distribution problem in a more traditional way when compared with the frequency distribution. The ITU-T first defined specific network element clock performance constraints and then defined a longest chain network permitted to ensure that the solution meets the end to end budget. The breakdown of the chain and the budget is shown in Figure 7.

Figure 7 **G.8271.1 Time Error Budget**



The overall end to end budget is defined as ±1.5 microseconds. From this the following allocations are made:

- ±100ns Time error due to the GNSS receiver and the 1588 grandmaster.
- ±500ns Constant Time error due to ten Telecom Boundary Clocks (50 ns limit per

  boundary clock).
- ±200ns Dynamic Time error presented at the end of the boundary clock chain into the

  end slave.
- ±300ns Time error due to errors in cable latency asymmetry compensation (see below).
- ±150ns Time error due to the end slave and any internals of the basestation between

  the recovery and the presentation on the air interface.
- ±200ns Time error in the end application during short term holdovers such as network topology re-arrangements.

Note there is discussion that some of these elements could be traded-off against each other. For example, if the link asymmetry needs a higher budget then the holdover budget would have to be less – implying a better end device or a shorter duration of holdover.

The link asymmetries are an important aspect of this budget. The network topology not only has to have the network elements that meet the clock specifications but it also needs to have links that meet certain requirements. As explained above, the time offset calculation makes the assumptions that the master–to-slave latency is the same from the slave–to-master latency. When the latency is not equal, an error is introduced. Some analysis of network intersite connections may need to be performed to determine the budget for the link asymmetries.

Configuration

IP Addressing for PTP Communication

The system supports communication to the PTP process on the CPM using any of the IPv4 local interface addresses or an IPv4 local loopback addresses. The system will record both the source and destination address information from the received Signaling message which establishes the unicast session. The system will then swap these addresses for use for the Sync and/or Delay_Req messages generated toward the external clock.

The IP address becomes more significant when 1588 port based timestamping is enabled. The port level functionality will filter received PTP packets for a known IP address. This ensures that only PTP messages intended for the node are modified and not PTP messages merely transiting the node.

If the 1588 nodes are directly connected or it is ensured that the PTP messages for a peer shall always enter/exit the system through a single interface, then the IP address of that interface can be used for the PTP message communication. If the PTP messages from a peer could enter through more than one interface, then it may be easier to utilize a loopback address for the PTP message communication.

If using a loopback address and 1588 port based timestamping is also to be used, then the specific loopback address must be assigned to PTP for use using the source-address command. An example is provided in the “Port Based Timestamping” section below. Note: When a source address is defined for the PTP process within a given routing context, then the source address for all Signaling messages originating out of the node within that routing context shall use that address.

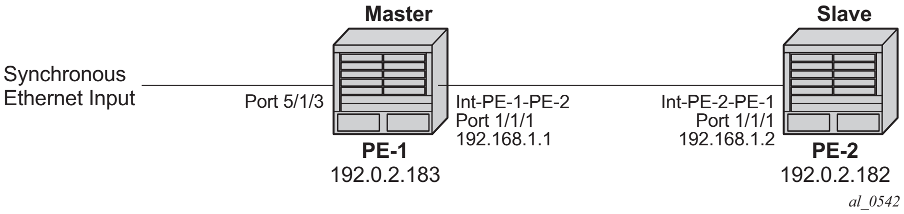
Note: The procedures to establish IP connectivity for the specific addresses used in these examples are not included.

Master and Slave Clocks for Frequency

A typical deployment scenario for a system configured as an ordinary master to distribute frequency to an external slave clock, often a cell site router or a base station, is shown in Figure 8. The central clock of the system is locked via its BITS ports or a Synchronous Ethernet port to an external source that is traceable to a primary reference. The frequency of the central clock is used to generate the timestamps contained in PTP event messages. The timestamps generated do not correlate to any standard epoch and therefore indicate an arbitrary timescale. As such it is only the rate of progression of the timestamps that has meaning.

The 7750 SR and the 7450 ESS can be configured as a 1588 slave clock for frequency recovery. In real deployments, it is more likely for the slave devices to be smaller cell site routers or basestations instead of another 7750 SR or 7450 ESS.

Figure 8 **Master and Slave Clocks for Frequency**



In the topology in Figure 8, the systems will most likely be configured with the ITU-T G.8265.1 Profile.

For this example, a loopback address is used for PTP communication between the nodes.

Ordinary Master Configuration

The steps to configure PE-1 as a PTP ordinary-clock master for frequency distribution using the G.8265.1 Telecom profile are outlined below:

Configure a /32 IPv4 system address on PE-1 and an interface to reach PE-2.

```
*A:PE-1#
        configure
            router
                interface "system"
                    address 192.0.2.183/32
                    no shutdown
                exit
                interface "int-PE-1-PE-2"
                    address 192.168.1.1/30
                    port 1/1/1
                    no shutdown
                exit
            exit
```

Configure an input reference for the central clock on PE-1. In this example, Synchronous Ethernet port 5/1/3 is used as the source for **ref2**.

```
*A:PE-1#
        configure
            port 5/1/3
                description "Sync-E reference for node"
                ethernet
                    ssm
                        code-type sonet
                        no shutdown
                    exit
                exit
                no shutdown
            exit
            system
                sync-if-timing
                    begin
                    ql-selection
                    ref2
                        source-port 5/1/3
                        no shutdown
                    exit
                    commit
                exit
            exit
```

The default clock type is set to ordinary slave so that must be changed to ordinary master. The only other relevant configuration parameter for the master clock running the G.8265.1 profile is the network-type. The coding of the SSM/ESMC Quality Level into PTP clock Class must match the environment. The system supports both SONET and SDH networks. The default network-type is sdh but for this example, the system is configured for the North American market so the network-type is set to sonet.

```
*A:PE-1#
        configure
            system
                ptp
                    clock-type ordinary master
                    network-type sonet
                    no shutdown
                exit
            exit
```

Ordinary Slave Configuration

To configure PE-2 as a PTP ordinary slave for frequency distribution using the G.8265.1 Telecom profile, firstly configure a /32 IPv4 system address on PE-2 and an interface to reach PE-1.

```
*A:PE-2#
        configure
            router
                interface "system"
                    address 192.0.2.182/32
                    no shutdown
                exit
                interface "int-PE-2-PE-1"
                    address 192.168.1.2/30
                    port 1/1/1
                    no shutdown
                exit
            exit
```

As the default clock type is ordinary slave, PE-1 is configured as a peer clock, and the PTP process is enabled. In this example, the Quality Level encoding is also changed to sonet in order to match the North American market

```
*A:PE-1#
        configure
            system
                ptp
                    network-type sonet
                    peer 192.0.2.183 create
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

Usually a 1588 slave has at least two peers configured in order to provide redundant sources.

Configure PTP as the reference for the central clock on PE-2.

```
*A:PE-2#
        configure
            system
                sync-if-timing
                    begin
                    ql-selection
                    ptp
                        no shutdown
                    exit
                    commit
                exit
            exit
```

Verification of Session Establishment

When PTP is set to no shutdown on PE-2, it initiates a PTP unicast session with PE-1. Correct session establishment can be verified by checking PTP related information as follows:

```
*A:PE-1# show system ptp unicast
===============================================================================
IEEE 1588/PTP Unicast Negotiation Information
===============================================================================
Router
  IP Address      Dir Type     Rate       Duration State    Time
-------------------------------------------------------------------------------
Base
  192.0.2.182     Tx  Announce 1 pkt/2 s  300      Granted  05/30/2014 06:08:38
  192.0.2.182     Tx  Sync     64 pkt/s   300      Granted  05/30/2014 06:08:43
  192.0.2.182     Rx  DelayReq 64 pkt/s   300      Granted  05/30/2014 06:08:43
  192.0.2.182     Tx  DelayRsp 64 pkt/s   300      Granted  05/30/2014 06:08:43
-------------------------------------------------------------------------------
PTP Peers               : 1
Total Packet Rate       : 192 packets/second
===============================================================================


*A:PE-2# show system ptp unicast
===============================================================================
IEEE 1588/PTP Unicast Negotiation Information
===============================================================================
Router
  IP Address      Dir Type     Rate       Duration State    Time
-------------------------------------------------------------------------------
Base
  192.0.2.183     Rx  Announce 1 pkt/2 s  300      Granted  05/30/2014 09:08:38
  192.0.2.183     Rx  Sync     64 pkt/s   300      Granted  05/30/2014 09:08:43
  192.0.2.183     Tx  DelayReq 64 pkt/s   300      Granted  05/30/2014 09:08:43
  192.0.2.183     Rx  DelayRsp 64 pkt/s   300      Granted  05/30/2014 09:08:43
-------------------------------------------------------------------------------
PTP Peers               : 1
Total Packet Rate       : 192 packets/second
===============================================================================
```

A **Pending** state indicates the system has sent a Unicast Request toward the peer but has not received a response. If the state remains **Pending**, then the IP connectivity between the systems should be verified.

To verify the slave frequency is operating properly, first check the high level information for PTP on PE-2. Note that the PTP Recovery State initially shows phase-tracking and then changes to locked. The time to achieve locked state varies based on the PDV.

```
*A:PE-2# show system ptp
===============================================================================
IEEE 1588/PTP Clock Information
===============================================================================
-------------------------------------------------------------------------------
Local Clock
-------------------------------------------------------------------------------
Clock Type        : ordinary,slave     PTP Profile        : ITU-T G.8265.1
Domain            : 4                  Network Type       : sonet
Admin State       : up                 Oper State         : up
Announce Interval : 1 pkt/2 s          Announce Rx Timeout: 3 intervals
Peer Limit        : none (Base Router)
Clock Id          : 00233efffe808250   Clock Class        : 255 (slave-only)
Clock Accuracy    : unknown            Clock Variance     : ffff (not computed)
Clock Priority1   : 128                Clock Priority2    : 128
PTP Port State    : slave              Last Changed       : 05/30/2014 09:08:42
PTP Recovery State: phase-tracking     Last Changed       : 05/30/2014 09:08:42
Frequency Offset  : -2.704 ppb
-------------------------------------------------------------------------------
Parent Clock
-------------------------------------------------------------------------------
IP Address        : 192.0.2.183        Router             : Base
Parent Clock Id   : 00233efffe69f250   Remote PTP Port    : 1
GM Clock Id       : 00233efffe69f250   GM Clock Class     : 80 (prs)
GM Clock Accuracy : unknown            GM Clock Variance  : ffff (not computed)
GM Clock Priority1: 128                GM Clock Priority2 : 128
-------------------------------------------------------------------------------
Time Information
-------------------------------------------------------------------------------
Timescale           : Arbitrary
Current Time        : 2014/05/30 14:12:52.9 (ARB)
Frequency Traceable : yes
Time Traceable      : no
Time Source         : other
```

```
===============================================================================
```

In addition PTP packet statistics can be checked to verify reception of the PTP messages and the execution of the frequency slave:

```
*A:PE-2# show system ptp statistics
===============================================================================
IEEE 1588/PTP Packet Statistics
===============================================================================
                                                               Input     Output
-------------------------------------------------------------------------------
PTP Packets                                                     5506       2742
  Announce                                                        23          0
  Sync                                                          2740          0
  Follow Up                                                        0          0
  Delay Request                                                    0       2740
  Delay Response                                                2740          0
  Signaling                                                        3          3
    Request Unicast Transmission TLVs                              0          3
      Announce                                                     0          1
      Sync                                                         0          1
      Delay Response                                               0          1
    Grant Unicast Transmission (Accepted) TLVs                     3          0
      Announce                                                     1          0
      Sync                                                         1          0
      Delay Response                                               1          0
    Grant Unicast Transmission (Denied) TLVs                       0          0
      Announce                                                     0          0
      Sync                                                         0          0
      Delay Response                                               0          0
    Cancel Unicast Transmission TLVs                               0          0
      Announce                                                     0          0
      Sync                                                         0          0
      Delay Response                                               0          0
    Ack Cancel Unicast Transmission TLVs                           0          0
      Announce                                                     0          0
      Sync                                                         0          0
      Delay Response                                               0          0
    Other TLVs                                                     0          0
  Other                                                            0          0
  Event Packets timestamped at port                                0          0
  Event Packets timestamped at cpm                              2740       2740
Discards                                                           0          0
  Bad PTP domain                                                   0          0
  Alternate Master                                                 0          0
  Out Of Sequence                                                  0          0
  Peer Disabled                                                    0          0
  Other                                                            0          0
===============================================================================
===============================================================================
IEEE 1588/PTP Frequency Recovery State Statistics
===============================================================================
State                                                                   Seconds
-------------------------------------------------------------------------------
Initial                                                                       0
Acquiring                                                                     0
Phase-Tracking                                                               43
Locked                                                                        0
```

```
Hold-over                                                                     0
===============================================================================

===============================================================================
IEEE 1588/PTP Event Statistics
===============================================================================
Event                                                      Sync Flow Delay Flow
-------------------------------------------------------------------------------
Packet Loss                                                        0          0
Excessive Packet Loss                                              0          0
Excessive Phase Shift Detected                                     0          0
Too Much Packet Delay Variation                                    0          0
===============================================================================
*
```

Secondly, the central clock status on the system can be checked:

```
*A:PE-2# show system sync-if-timing
===============================================================================
System Interface Timing Operational Info
===============================================================================
System Status CPM B                : Master Locked
    Reference Input Mode           : Non-revertive
    Quality Level Selection        : Disabled
    Reference Selected             : ptp
    System Quality Level           : prs
    Current Frequency Offset (ppm) : +0

Reference Order                    : bits ref1 ref2 ptp

Reference Mate CPM
    Qualified For Use              : No
        Not Qualified Due To       :     LOS
    Selected For Use               : No
        Not Selected Due To        :     not qualified

Reference Input 1
    Admin Status                   : down
    Rx Quality Level               : unknown
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
    Source Port                    : None

Reference Input 2
    Admin Status                   : down
    Rx Quality Level               : unknown
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
    Source Port                    : None

Reference BITS B
    Input Admin Status             : down
    Rx Quality Level               : failed
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
    Interface Type                 : DS1
    Framing                        : ESF
    Line Coding                    : B8ZS
    Line Length                    : 0-110ft
    Output Admin Status            : down
    Output Source                  : line reference
    Output Reference Selected      : none
    Tx Quality Level               : N/A

Reference PTP
    Admin Status                   : up
    Rx Quality Level               : prs
    Quality Level Override         : none
    Qualified For Use              : Yes
    Selected For Use               : Yes
```

Optional Configuration Items for Ordinary Master or Slave Configuration

The G.8265.1 profile is the default PTP profile on the system and it uses domain number value of 4. The domain number must match at both ends of the communication path or the PTP messages will be dropped. Some very old 1588 devices, may have the domain number set to zero which is the value used by the IEEE1588 default profile. In this case, the system would need to have its domain number changed to match that of the external slave.

```
configure
    system
        ptp
            shutdown
            domain 0
            no shutdown
        exit
    exit
```

Note that the domain number can only be adjusted if PTP is shutdown and only one common domain number is allowed for all 1588 messages to and from the system.

When using the system as a 1588 slave for frequency distribution, it is strongly recommended to use the default message rate of 64 pps for Sync and Delay_Resp messages. If for some reason the parent 1588 peer cannot offer this rate, then the rate that the system requests must be adjusted. For example, if the maximum rate supported by the external 1588 grandmaster device (with an IP address of 192.0.2.166) only is 32 pps, then the system can be adjusted to request that rate as follows:

```
configure
    system
        ptp
            peer 192.0.2.166 create
                log-sync-interval -5
                no shutdown
            exit
        exit
    exit
```

Note that the Sync message rate can only be adjusted if the peer is shutdown.

The message rates are entered as the base 2 logarithm of the inter-message interval. So 32 pps has an inter message interval of 1/32 seconds and a log-sync-interval of -5.

The Announce message rate impact the speed at which PTP can detect communication failures and the speed at which the PTP topology is re-arranged. The default Announce rate is one message every two seconds and this should be adequate for networks with short chains of PTP clocks, for example G.8265.1 architectures. However, in network with longer chains of PTP clocks (for example, more than 5 boundary clocks), it may be desired to use a faster Announce message rate. In the following example, the slave is configured to request two Announce messages per second:

```
configure
    system
        ptp
            shutdown
            log-anno-interval -1
            no shutdown
        exit
    exit
```

Note that the Announce rate can only be adjusted if PTP is shutdown. In addition, there is one common Announce rate for all unicast sessions; it cannot be configured on an individual peer basis.

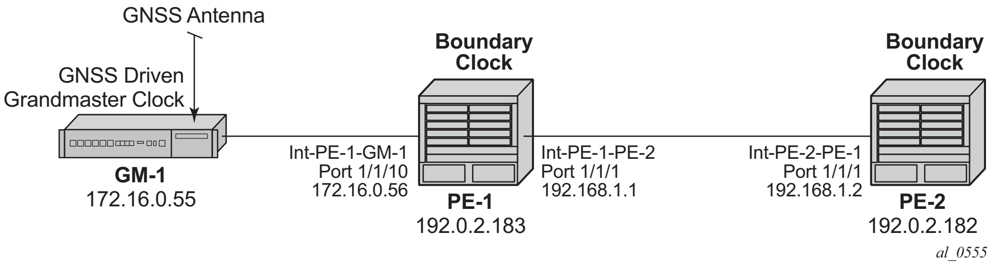
Boundary Clock

With the increase interest in high accuracy time distribution across networks, the system most likely takes on the role of a 1588 boundary clock. In this role, the system requests time from a GNSS driven grandmaster clock or from a neighboring boundary clock. The system only supports boundary clock configuration when the ptp profile is configured as the default profile.

In this mode of operation, it is strongly recommended to have Synchronous Ethernet physical layer frequency distribution configured at the same time.

The example in Figure 9 shows a boundary clock (PE-1) communicating directly with the GNSS driven grandmaster (GM-1) and a second boundary clock (PE-2) communicating with the first boundary clock.

Figure 9 **Boundary Clock**



The steps to configure the systems as boundary clocks running the IEEE default profile are:

On PE-1, configure a /32 IPv4 system address, an interface to reach PE-2, and an interface to reach GM-1.

```
*A:PE-1#
        configure
            router
                interface "system"
                    address 192.0.2.183/32
                    no shutdown
                exit
                interface "int-PE-1-PE-2"
                    address 192.168.1.1/30
                    port 1/1/1
                    no shutdown
                exit
                interface "int-PE-1-GM-1"
                    address 172.16.0.56/30
                    port 1/1/2
                    no shutdown
                exit
            exit
```

On PE-2, configure a /32 IPv4 system address and an interface to reach PE-1.

```
*A:PE-2#
        configure
            router
                interface "system"
                    address 192.0.2.182/32
                    no shutdown
                exit
                interface "int-PE-2-PE-1"
                    address 192.168.1.2/30
                    port 1/1/1
                    no shutdown
                exit
            exit
```

Configure both PE-1 and PE-2 to have physical layer frequency sources into their central clocks. PE-2 is configured to receive Synchronous Ethernet from PE-1 on the same port as is used for PTP. This commonality is not a requirement but might be common in the network topology.

On PE-1, configure the port toward PE-2 as a Synchronous Ethernet port. This will cause the port transmit timing to be sourced from the node timing. Also configure the port to transmit ssm codes using the sonet codes.

```
*A:PE-1#
        configure card 1 mda 1 sync-e
        configure port 1/1/1 ethernet
            code-type sonet
            no shutdown
        exit
```

On PE-2, configure the port on towards PE-1 as a Synchronous Ethernet port and to use sonet codes and to be the reference into the central clock of PE-2.

```
*A:PE-2#
        configure card 1 mda 1 sync-e
        configure port 1/1/1
            ethernet
                ssm
                    code-type sonet
                    no shutdown
                exit
            exit
        exit
        configure system sync-it-timing
            begin
            ql-selection
            ref1
                source-port 1/1/1
                no shutdown
            exit
            commit
        exit
```

Next configure PE-1 as a boundary clock requesting service from GM-1 using the default profile. In this example, the interface address of GM-1 is used for the PTP communication.

```
*A:PE-1#
        configure system ptp
            shutdown
            profile ieee1588-2008
            clock-type boundary
            peer 172.16.0.55 create
                no shutdown
            exit
            no shutdown
        exit
```

If it is desired to operate the network at the default for the G.8275.1 profile, then the Announce messages should be set to 8 pps and the Sync and Delay_Resp messages should be set to 16 pps.

```
*A:PE-1#
        configure system ptp
            shutdown
            log-anno-interval -3
            peer 172.16.0.55
                shutdown
                log-sync-interval -4
                no shutdown
            exit
            no shutdown
        exit
```

Configure PE-2 as a boundary clock using PE-1 as its parent clock and the same set of 1588 parameters. In this example, PE-2 uses a loopback address of PE-1 for communication.

```
*A:PE-2#
        configure system ptp
            shutdown
            profile ieee1588-2008
            clock-type boundary
            log-anno-interval -3
            peer 192.0.2.183 create
                shutdown
                log-sync-interval -4
                no shutdown
            exit
            no shutdown
        exit
```

On PE-1, validate the status of the PTP topology by checking the unicast sessions. Also validate the PTP process has elected GM-1 as both the parentClock and the grandmaster clock.

```
*A:PE-1# show system ptp unicast
===============================================================================
IEEE 1588/PTP Unicast Negotiation Information
===============================================================================
Router
  IP Address      Dir Type     Rate       Duration State    Time
-------------------------------------------------------------------------------
Base
  192.0.2.182     Tx  Announce 8 pkt/s    300      Granted  05/30/2014 07:02:36
  192.0.2.182     Tx  Sync     16 pkt/s   300      Granted  05/30/2014 07:02:37
  192.0.2.182     Rx  DelayReq 16 pkt/s   300      Granted  05/30/2014 07:02:37
  192.0.2.182     Tx  DelayRsp 16 pkt/s   300      Granted  05/30/2014 07:02:37
  172.16.0.55     Rx  Announce 8 pkt/s    300      Granted  05/30/2014 07:02:42
  172.16.0.55     Rx  Sync     16 pkt/s   300      Granted  05/30/2014 07:02:43
  172.16.0.55     Tx  DelayReq 16 pkt/s   300      Granted  05/30/2014 07:02:43
  172.16.0.55     Rx  DelayRsp 16 pkt/s   300      Granted  05/30/2014 07:02:43
-------------------------------------------------------------------------------
PTP Peers              : 2
Total Packet Rate      : 112 packets/second
===============================================================================


*A:PE-1# show system ptp
===============================================================================
IEEE 1588/PTP Clock Information
===============================================================================
-------------------------------------------------------------------------------
Local Clock
-------------------------------------------------------------------------------
Clock Type        : boundary           PTP Profile        : IEEE 1588-2008
Domain            : 0                  Network Type       : sdh
Admin State       : up                 Oper State         : up
Announce Interval : 8 pkt/s            Announce Rx Timeout: 3 intervals
Peer Limit        : none (Base Router)
Clock Id          : 00233efffe69f250   Clock Class        : 248 (default)
Clock Accuracy    : unknown            Clock Variance     : ffff (not computed)
Clock Priority1   : 128                Clock Priority2    : 128
PTP Recovery State: locked             Last Changed       : 05/30/2014 07:05:17
Frequency Offset  : +50.305 ppb
-------------------------------------------------------------------------------
Parent Clock
-------------------------------------------------------------------------------
IP Address        : 172.16.0.55        Router             : Base
Parent Clock Id   : 8887868584838281   Remote PTP Port    : 1
GM Clock Id       : 8887868584838281   GM Clock Class     : 7
GM Clock Accuracy : within 250 ns      GM Clock Variance  : 0x6400 (3.7E-09)
GM Clock Priority1: 128                GM Clock Priority2 : 128
-------------------------------------------------------------------------------
Time Information
-------------------------------------------------------------------------------
Timescale           : PTP
Current Time        : 2014/05/30 15:07:01.1 (UTC)
```

```
Frequency Traceable : yes
Time Traceable      : yes
Time Source         : GPS
```

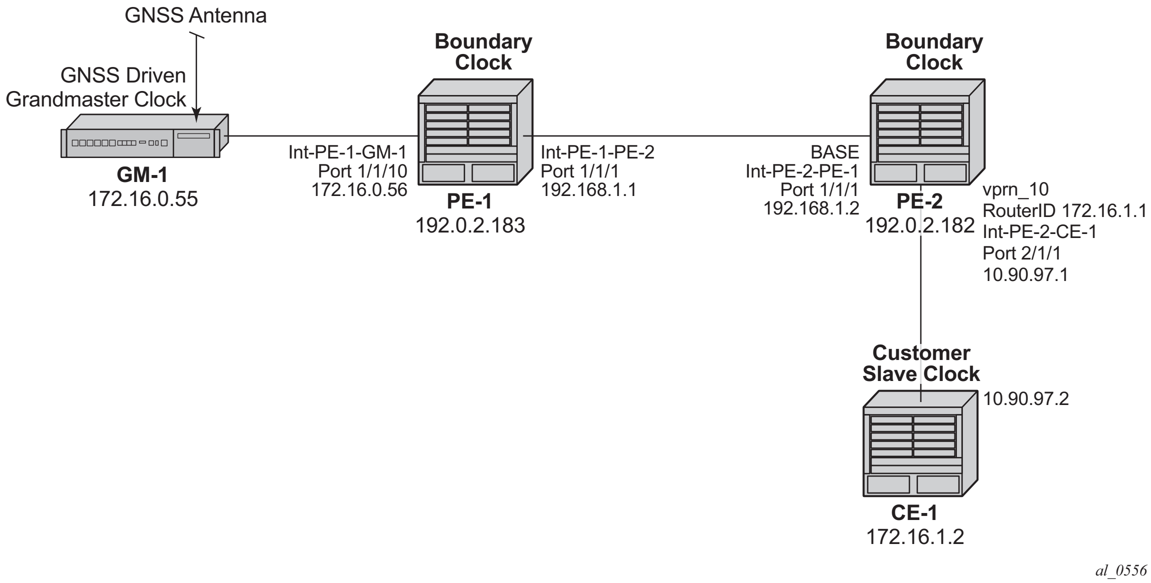
On PE-2, validate the PTP process has elected PE-1 as its parentClock and that the grandmaster clock is GM-1.

```
*A:PE-2# show system ptp
===============================================================================
IEEE 1588/PTP Clock Information
===============================================================================
-------------------------------------------------------------------------------
Local Clock
-------------------------------------------------------------------------------
Clock Type        : boundary           PTP Profile        : IEEE 1588-2008
Domain            : 0                  Network Type       : sdh
Admin State       : up                 Oper State         : up
Announce Interval : 8 pkt/s            Announce Rx Timeout: 3 intervals
Peer Limit        : none (Base Router)
Clock Id          : 00233efffe808250   Clock Class        : 248 (default)
Clock Accuracy    : unknown            Clock Variance     : ffff (not computed)
Clock Priority1   : 128                Clock Priority2    : 128
PTP Recovery State: locked             Last Changed       : 05/30/2014 10:02:36
Frequency Offset  : -9.345 ppb
-------------------------------------------------------------------------------
Parent Clock
-------------------------------------------------------------------------------
IP Address        : 192.0.2.183        Router             : Base
Parent Clock Id   : 00233efffe69f250   Remote PTP Port    : 1
GM Clock Id       : 8887868584838281   GM Clock Class     : 7
GM Clock Accuracy : within 250 ns      GM Clock Variance  : 0x6400 (3.7E-09)
GM Clock Priority1: 128                GM Clock Priority2 : 128
-------------------------------------------------------------------------------
Time Information
-------------------------------------------------------------------------------
Timescale           : PTP
Current Time        : 2014/05/30 15:09:26.5 (UTC)
Frequency Traceable : yes
Time Traceable      : yes
Time Source         : GPS
===============================================================================
```

Boundary Clock with VPRN Access

The system supports access to the 1588 process through Base routing, IES, and VPRN contexts. This permits the system 1588 topology to be created and managed in one context with access for edge distribution through other contexts. For example, building on top of the base routing distribution shown in the previous example, access can be given to the 1588 process on PE-2 via a VPRN existing on that node. This allows the VPRN customer to have access to the high accuracy time available within the system in the customer edge equipment connecting into that node.

Figure 10 **Boundary Clocks with Edge VPRN Access**



For the example shown in Figure 10, it is assumed that a VPRN service is already configured and operational on PE-2 providing connectivity between PE-2 and CE-1:

```
*A:PE-2#
        configure service vprn 10 customer 1 create
            router-id 176.16.1.1
            autonomous-system 64496
            route-distinguisher 64496:10
            interface "int-PE-2-CE-1" create
                address 10.90.97.1/30
                sap 2/1/1 create
                exit
            exit
            no shutdown
        exit
```

To enable access to the PTP process via VPRN 10 in PE-2, PTP must be enabled within the VPRN context. To ensure that no more than 10 external clocks access the system PTP through this VPRN at any one time, a peer-limit may be defined.

```
*A:PE-2#
        configure service vprn 10
            peer-limit 10
            ptp no shutdown
        exit
```

To confirm PTP access with the VPRN, the PTP information with the VPRN context can be queried. Either of the following two commands can be used:

```
*A:PE-2# show system ptp unicast router 10
```

or

```
*A:PE-2# show service id 10 ptp unicast
```

These two commands provide the same information as shown below.

```
*A:PE-2# show system ptp unicast router 10
===============================================================================
IEEE 1588/PTP Unicast Negotiation Information
===============================================================================
Router
  IP Address      Dir Type     Rate       Duration State    Time
-------------------------------------------------------------------------------
10
  172.16.1.2      Tx  Announce 1 pkt/2 s  300      Granted  05/30/2014 12:40:53
  172.16.1.2      Tx  Sync     64 pkt/s   300      Granted  05/30/2014 12:40:59
  172.16.1.2      Rx  DelayReq 64 pkt/s   300      Granted  05/30/2014 12:40:59
  172.16.1.2      Tx  DelayRsp 64 pkt/s   300      Granted  05/30/2014 12:40:59
-------------------------------------------------------------------------------
PTP Peers               : 1
Total Packet Rate       : 192 packets/second
===============================================================================
```

Port Based Timestamping

As described above, optimal performance is achieved when the 1588 port based timestamping (PBT) feature is used. This feature is not available on all hardware and the interfaces for PTP should be planned in advance if this feature is to be used.

Since 1588 messages ingress and egress the node through router interfaces, the configuration of the 1588 PBT feature is enabled within the router interface context. In the previous examples, if 1588 PBT is to be enabled on all the PTP interfaces the following commands are required.

On PE-1, enable 1588 PBT on the interface toward GM-1 and PE-2.

```
*A:PE-1#
        configure
            router
                interface "int-PE-1-PE-2"
                    ptp-hw-assist
                exit
                interface "int-PE-1-GM-1"
                    ptp-hw-assist
                exit
            exit
```

On PE-2, enable 1588 PBT on the interface toward PE-1 and CE-1.

```
*A:PE-2#
        configure
            router
                interface "int-PE-2-PE-1"
                    ptp-hw-assist
                exit
            exit
        exit
        configure service vprn 10 customer 1
            interface "int-PE-2-CE-1"
                ptp-hw-assist
            exit
        exit
```

To verify 1588 PBT is active on the 1588 messages to the peers, check the timestamp point for the specific peer. It now indicates *port* rather than cpm.

On PE-1 for the CE-1 communication:

```
*A:PE-1# show system ptp peer 172.16.0.55
===============================================================================
IEEE 1588/PTP Peer Information
===============================================================================
Router            : Base
IP Address        : 172.16.0.55        Announce Direction : rx
Admin State       : up                 G.8265.1 Priority  : n/a
Sync Interval     : 16 pkt/s
Local PTP Port    : 2                  PTP Port State     : slave
Clock Id          : 8887868584838281   Remote PTP Port    : 1
GM Clock Id       : 8887868584838281   GM Clock Class     : 7
GM Clock Accuracy : within 250 ns      GM Clock Variance  : 0x6400 (3.7E-09)
GM Clock Priority1: 128                GM Clock Priority2 : 128
Steps Removed     : 0                  Parent Clock       : yes
Tx Timestamp Point: port               Rx Timestamp Point : port
Last Tx Port      : 5/1/1              Last Rx Port       : 5/1/1
===============================================================================
```

On PE-1 the communication with the PE-2 will still be CPM timestamping since the port has not been configured to watch for the ‘system’ loopback address.

```
*A:PE-1# show system ptp peer 192.0.2.182
===============================================================================
IEEE 1588/PTP Peer Information
===============================================================================
Router            : Base
IP Address        : 192.0.2.182        Announce Direction : tx
Admin State       : n/a                G.8265.1 Priority  : n/a
Sync Interval     : n/a
Local PTP Port    : 3                  PTP Port State     : master
Clock Id          : 00233efffe808250   Remote PTP Port    : 4
Tx Timestamp Point: cpm                Rx Timestamp Point : cpm
Last Tx Port      : 5/1/2              Last Rx Port       : 5/1/2
===============================================================================
```

In order to configure the **system** loopback address for PTP, enter the following on PE-1:

```
*A:PE-1#
        configure
            system security
                source-address application ptp "system"
                exit
            exit
```

Now the timestamp point on PE-1 will be the port.

```
*A:PE-1# show system ptp peer 192.0.2.182
===============================================================================
IEEE 1588/PTP Peer Information
===============================================================================
Router            : Base
IP Address        : 192.0.2.182        Announce Direction : tx
Admin State       : n/a                G.8265.1 Priority  : n/a
Sync Interval     : n/a
Local PTP Port    : 3                  PTP Port State     : master
Clock Id          : 00233efffe808250   Remote PTP Port    : 4
Tx Timestamp Point: port               Rx Timestamp Point : port
Last Tx Port      : 5/1/2              Last Rx Port       : 5/1/2
===============================================================================
```

Repeat this configuration of system address for the base routing context on PE-2

```
*A:PE-2#
        configure
            system security
                source-address application ptp "system"
                exit
            exit
```

Now the timestamp point on PE-2 will be the port.

```
*A:PE-2# show system ptp peer 192.0.2.183
===============================================================================
IEEE 1588/PTP Peer Information
===============================================================================
Router            : Base
IP Address        : 192.0.2.183        Announce Direction : rx
Admin State       : up                 G.8265.1 Priority  : n/a
Sync Interval     : 16 pkt/s
Local PTP Port    : 4                  PTP Port State     : slave
Clock Id          : 00233efffe69f250   Remote PTP Port    : 3
GM Clock Id       : 8887868584838281   GM Clock Class     : 6
GM Clock Accuracy : within 100 ns      GM Clock Variance  : 0x6400 (3.7E-09)
GM Clock Priority1: 128                GM Clock Priority2 : 128
Steps Removed     : 1                  Parent Clock       : yes
Tx Timestamp Point: port               Rx Timestamp Point : port
Last Tx Port      : 1/1/2              Last Rx Port       : 1/1/2
===============================================================================
```

On PE-2, a loopback address must assigned for PTP communication as follows:

```
*A:PE-2#
configure service vprn 10
    interface "ptp_loopback"
        address 172.16.1.1/32
        loopback
    exit
    source-address
        application ptp "ptp_loopback"
            exit
        exit
```

1588 as NTP Local Clock (server)

If the system is configured as a boundary clock or slave clock then the time recovered from the 1588 slave port can be used as the source of system time on the node. This allows for higher accuracy and better stability in the timebase when compared to NTP. To enable this, PTP must be made the preferred server in the NTP context in the node.

Note that if the system is acting as an NTP server or peer to other NTP clocks, then turning on this feature will impact the existing NTP topology. The system shall advertise itself as an NTP Stratum 1 server to external clients and peers. Given the much higher accuracies achievable with PTP time distribution, this change in topology does not degrade the time in the clients and peers.

```
*A:PE-1#
        configure system time ntp
            server ptp prefer
        exit
```

To validate PTP is now being used for NTP time and system time, use the following command:

```
*A:PE-1# show system ntp all
===============================================================================
NTP Status
===============================================================================
Configured         : Yes                Stratum              : 1
Admin Status       : up                 Oper Status          : up
Server Enabled     : No                 Server Authenticate  : No
Clock Source       : ptp
Auth Check         : Yes
Current Date & Time: 2014/05/30 17:53:11 UTC
===============================================================================
===============================================================================
NTP Active Associations
===============================================================================
State                     Reference ID    St Type  A  Poll Reach     Offset(ms)
    Remote
-------------------------------------------------------------------------------
chosen                    PTP             0  srvr  -  64   ......YY  0.000
    ptp
===============================================================================
===============================================================================
NTP Clients
===============================================================================
vRouter                                                    Time Last Request Rx
    Address
-------------------------------------------------------------------------------
===============================================================================
```

Conclusion

The systems provide support for IEEE 1588 frequency and time distribution for the synchronization applications of the mobile networks. They can be configured as frequency distribution grandmasters and slave clocks or time distribution boundary and slave clocks.

Synchronous Ethernet

This chapter provides information about Synchronous Ethernet (SyncE).

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS release 8.0.R7. The CLI in the current edition is based on SR OS release 14.0.R6. There are no software prerequisites for this configuration, however, the hardware requires the use of Synchronous Ethernet capable MDA-XPs/CMA-XPs or the IMMs.

In addition, Synchronous Ethernet is only supported on optical interfaces. It is not supported on 10/100/1000 base copper interfaces.

Summary

Synchronous Ethernet (SyncE) is the ability to provide PHY-level frequency distribution through an Ethernet port. It is one of the building blocks of Next Generation Networks (NGNs).

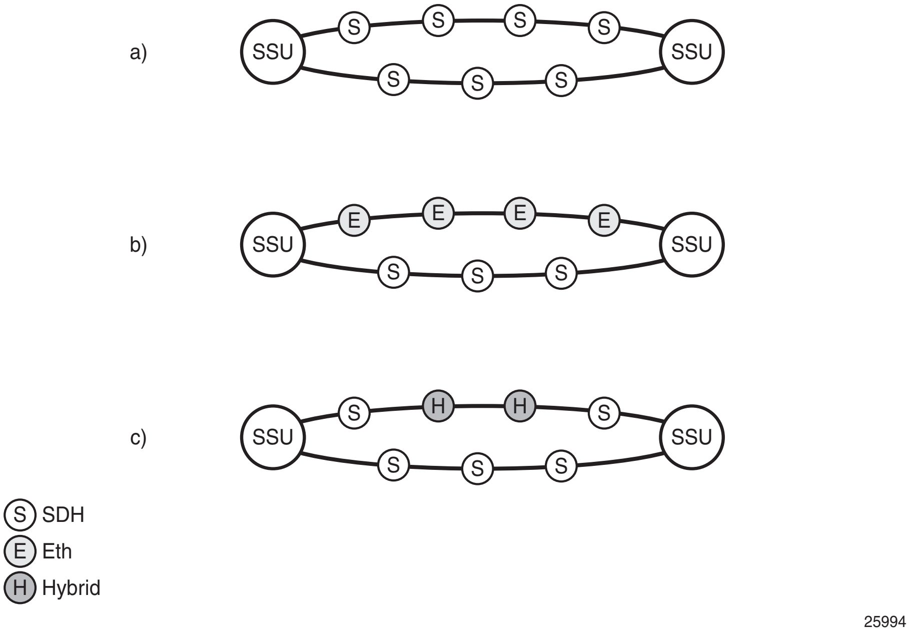
Overview

Synchronous Ethernet

Traditionally, Ethernet based networks employ the physical layer transmitter clock to be derived from an inexpensive +/-100ppm crystal oscillator and the receiver locks onto it. There is no need for long term frequency stability because the data is packetized and can be buffered. For the same reason, there is no need for consistency between the frequencies of different links. However, one could choose to derive the physical layer transmitter clock from a high quality frequency reference by replacing the crystal with a frequency source traceable to a primary reference clock. This would not affect the operation of any of the Ethernet layers, for which this change would be transparent. The receiver at the far end of the link would lock onto the physical layer clock of the received signal, and thus itself gain access to a highly accurate and stable frequency reference. Then, in a manner analogous to conventional hierarchical master-slave network synchronization, this receiver could lock the transmission clock of its other ports to this frequency reference and a fully time synchronous network could be established.

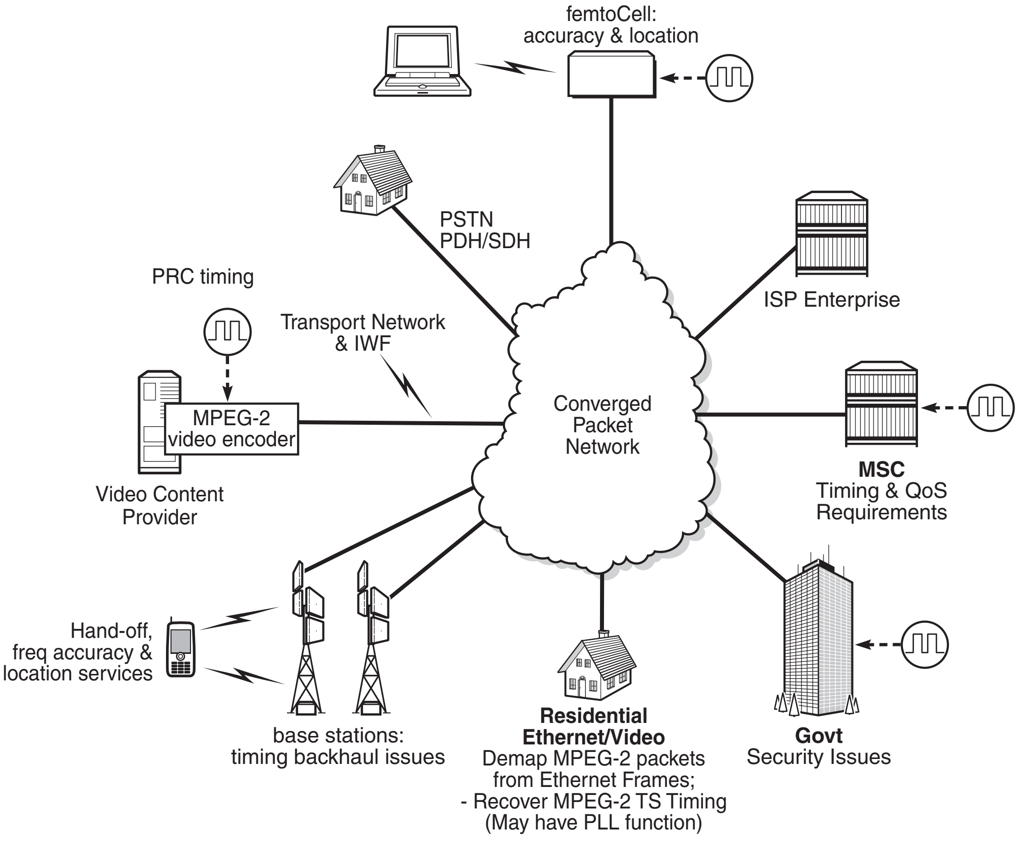
The advantage of using SyncE, as compared to methods relying on sending timing information in packets over an unlocked physical layer, is that SyncE is not influenced by impairments introduced by the higher levels of the networking technology (packet loss, packet delay variation). Therefore, the frequency accuracy and stability may be expected to exceed those of networks with unsynchronized physical layers. In addition, SyncE was designed to integrate into any existing SONET/SDH synchronization distribution architecture to allow for the easy migration from the traditional to the new synchronous interfaces. SyncE includes the concept of a hybrid switch which supports the interworking of synchronization distribution through SONET/SDH and the SyncE interfaces at the same time.

Figure 11 **SyncE Hypothetical Reference Network Architecture**



Many Tier 1 carriers are looking to migrate their synchronization infrastructure to a familiar and manageable model. In order to enable rapid migration of these networks, SyncE may be the easiest to deploy in order to ensure robust frequency synchronization.

Figure 12 **Packet Based Network Timing Infrastructure**

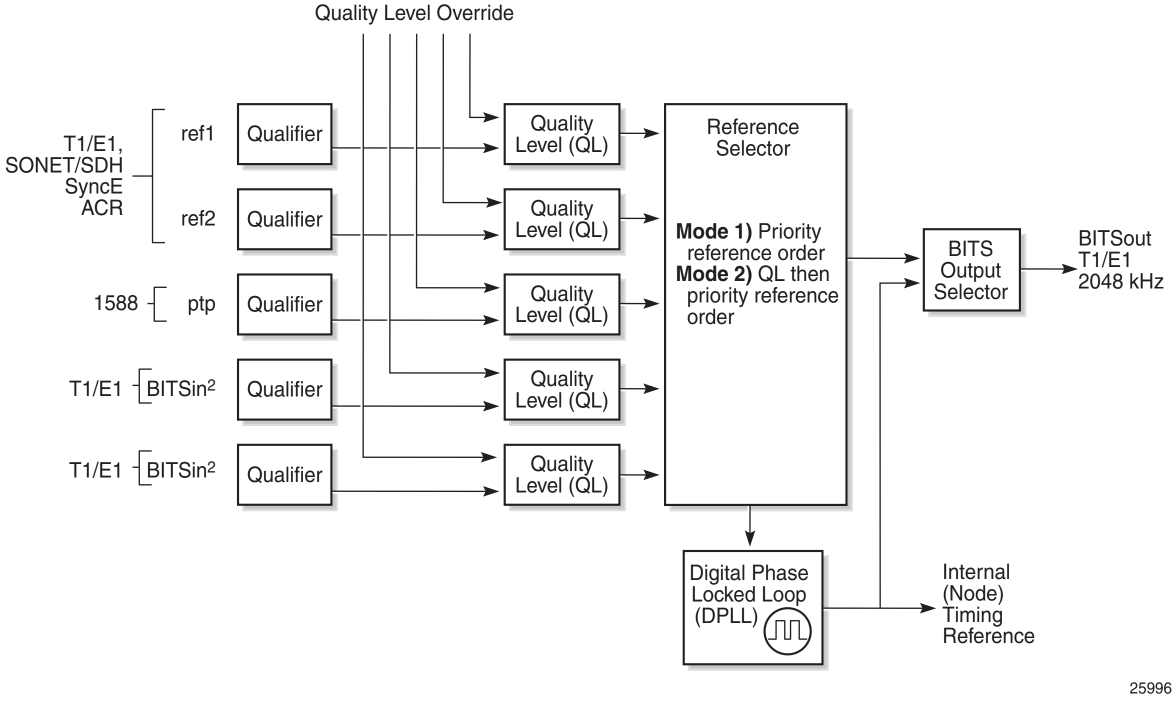


Central Synchronization Sub-System

The timing subsystem for the SR OS platforms has a central clock located on the Control Processor Module (CPM). The timing subsystem performs many of the duties of the network element clock as defined by Telcordia GR-1244 and ITU-T G.781.

The system can select from up to three (7950 XRS) or four (7450 ESS and 7750 SR) timing inputs to train the local oscillator. The priority order of these references must be specified. This is a simple ordered list of inputs: {BITS [Building Integrated Timing Source], ref1, ref2, PTP [Precision Time Protocol]}. The CPM clock output has the ability to drive the clocking for all line cards in the system. The SR OS platforms support selection of the node reference using Quality Level (QL) indications.

Figure 13 **CPM Clock Synchronization Reference Selection**



The recovered clock is able to derive its timing from any of the following:

- OC3/STM1, OC12/STM4, OC48/STM16, OC192/STM64 ports (7450 ESS and 7750 SR only)
- T1/E1 CES channel (adaptive clocking) (7750 SR only)
- SyncE ports
- T1/E1 ports (7750 SR only)
- BITS port on a channelized OC3/STM1 CES CMA (7750 SR-c12 only)
- BITS port on the CPM, CFM, or CCM module
- 10GE ports in WAN PHY mode
- IEEE 1588v2 slave port (PTP) (7450 ESS and 7750 SR only)

On 7750 SR-12 and 7750 SR-7 systems with redundant CPMs, the system has two BITS input ports (one per CPM). On the 7750 SRc-4 systems, there are two BITS input ports on the chassis front plate. These BITS input ports provide redundant synchronization inputs from an external BITS/SSU. However, the 7750 SR-c12 does not support BITS input port redundancy or BITS out.

All settings of the signal characteristics for the BITS input apply to both ports. When the active CPM considers the BITS input as a possible reference, it will consider first the BITS input port on the active CPM followed the BITS input port on the standby CPM in that relative priority order. This relative priority order is in addition to the user definable **ref-order**. For example, a ref-order of ‘bits-ref1-ref2-ptp’ would actually be BITS in (active CPM) followed by BITS in (standby CPM) followed by ref1 followed by ref2 followed by PTP. When **ql-selection** is enabled, then the QL of each BITS input port is viewed independently. The higher QL source is chosen.

On the 7750 SR-c4 platform CFM, there are two BITS input ports and two BITS output ports on this one module. These two ports are provided for BITS redundancy for the chassis. All settings of the signal characteristics for the BITS input apply to both ports. This includes the **ql-override** setting. When the CFM considers the BITS input as a possible reference, it will consider first the BITS input port “bits1” followed the BITS input port “bits2” in that relative priority order. This relative priority order is in addition to the user definable **ref-order**. For example, a ref-order of ‘bits-ref1-ref2’ would actually be “bits1” followed by “bits2” followed by ref1 followed by ref2. When **ql-selection** is enabled, then the QL of each BITS input port is viewed independently. The higher QL source is chosen.

The BITS output ports deliver a unfiltered recovered line clock from a SR/ESS port directly to a dedicated timing device in the facility (BITS or Standalone Synchronization Equipment (SASE) device). The signal selected will be one of ref1 or ref2. It cannot be the BITS input port signal nor can it be the output of the central clock.

When QL selection mode is disabled, then the reversion setting controls when the central clock can re-select a previously failed reference.

Table 1 shows the selection followed for two references in both revertive and non-revertive modes.

Table 1 **Revertive, Non-Revertive Timing Reference Switching Operation**

Status of Reference A	Status of Reference B	Active Reference Non-revertive Case	Active Reference Revertive Case
OK	OK	A	A
Failed	OK	B	B
OK	OK	B	A
OK	Failed	A	A
OK	OK	A	A
Failed	Failed	holdover	holdover

Table 1 **Revertive, Non-Revertive Timing Reference Switching Operation (Continued)**

Status of Reference A	Status of Reference B	Active Reference Non-revertive Case	Active Reference Revertive Case
OK	Failed	A	A
Failed	Failed	holdover	holdover
Failed	OK	B	B
Failed	Failed	holdover	holdover
OK	OK	A or B	A

Synchronization Status Messages (SSM)

SSM provides a mechanism to allow the synchronization distribution network to both determine the quality level of the clock sourcing a given synchronization trail and to allow a network element to select the best of multiple input synchronization trails. Synchronization Status messages have been defined for various transport protocols including SONET/SDH, T1/E1, and SyncE, for interaction with office clocks, such as BITS or SSUs (synchronization supply unit) and embedded network element clocks.

SSM allows equipment to autonomously provision and reconfigure (by reference switching) their synchronization references, while helping to avoid the creation of timing loops. These messages are particularly useful to allow synchronization reconfigurations when timing is distributed in both directions around a ring.

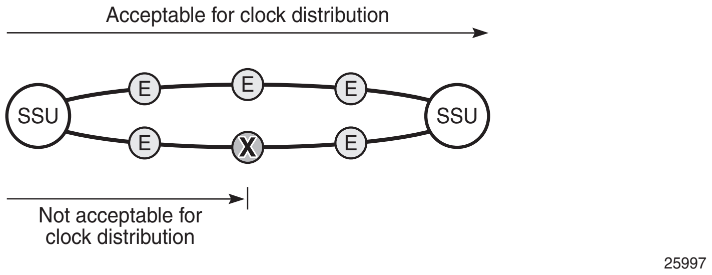
In SyncE, the SSM is provided through the Ethernet Synchronization Messaging Channel (ESMC). This mechanism uses Ethernet OAM PDU to exchange the Quality Level values over the SyncE link.

SyncE Chains

Transmission of a reference clock through a chain of Ethernet equipment requires that all of the equipment support SyncE.

A single piece of equipment not capable of SyncE breaks the chain as shown in Figure 14. Ethernet frames will still get through but downstream devices will recognize that the signal is out of pull-in range so they can not use it for reference.

Figure 14 **Network Considerations for Ethernet Timing Distribution**


Acceptable for clock distribution
SSU
E
E
E
SSU
E
X
E
Not acceptable for clock distribution
25997

Configuration

Configuration 1 - QL-Selection Mode Disabled

The following example shows the configuration options for SyncE when ql-selection mode is disabled. Generally, North American SONET networks do not use the automatic reference selection mechanisms. If SyncE is being added into such a network, it would likely have ql-selection set to disabled.

```
*A:PE-1# configure card 1 mda 1
  - mda <mda-slot>
  - no mda <mda-slot>

 <mda-slot>           : [1..2]

      access          + Configure access MDA parameters
      atm             + Configure ATM MDA parameters
      clock-mode      - Configure clock mode and timestamp frequency
      egress          + Configure egress MDA parameters
      egress-xpl      + Configure egress MDA XPL interface error parameters
 [no] fail-on-error   - Configure the behavior of the MDA state when an error is
                        detected
 [no] hi-bw-mcast-src - Enable/disable allocation of resources for high bandwidth
                        multicast streams
      ingress         + Configure ingress MDA parameters
      ingress-xpl     + Configure ingress MDA XPL interface error parameters
 [no] mda-type        - Provisions/de-provisions an MDA to/from the device
                        configuration for the slot
      named-pool-mode + Enable/Disable named pool mode
      network         + Configure network MDA parameters
 [no] shutdown        - Administratively shut down an mda
 [no] sync-e          - Enable/Disable Synchronous Ethernet
```

SyncE is enabled on MDA 1 of card 1 as follows:

```
*A:PE-1# configure card 1 mda 1 sync-e
```

After syncE is enabled, the configuration of MDA 1 is as follows

```
*A:PE-1# configure card 1 mda 1
*A:PE-1>config>card>mda#             info detail
----------------------------------------------
            mda-type m4-10gb-xp-xfp
            sync-e
            named-pool-mode
                ingress
                    no named-pool-policy
                exit
                egress
                    no named-pool-policy
                exit
            exit
            ingress
            exit
            ingress-xpl
                threshold 1000
                window 60
            exit
            egress
                no hsmda-pool-policy
                hsmda-agg-queue-burst
                    no low-burst-multiplier
                    no high-burst-increase
                exit
            exit
            egress-xpl
                threshold 1000
                window 60
            exit
            no fail-on-error
            network
                ingress
                    pool default
                        no amber-alarm-threshold
                        no red-alarm-threshold
                        resv-cbs default
                        slope-policy "default"
                    exit
                    queue-policy "default"
                exit
                egress
                    pool default
                        no amber-alarm-threshold
                        no red-alarm-threshold
                        resv-cbs default
                        slope-policy "default"
                    exit
                exit
            exit
            access
                ingress
                    pool default
---snip---
```

The synchronous interface timing can be configured with the following parameters:

```
*A:PE-1# configure system sync-if-timing
  - sync-if-timing

     abort           - Discard the changes that have been made to sync interface
                       timing during a session
      begin           - Switch to edit mode for sync interface timing - use commit to
                        save or abort to discard the changes made in a session
      bits            + Configure parameters for the Building Integrated Timing
                        Supply (BITS)
      commit          -
 Save the changes made to sync interface timing during a session
      ptp             + Configure parameters for Precision Timing Protocol (PTP)
                        timing reference
 [no] ql-minimum      - Configure the minimum quality level of the input
 [no] ql-selection    - Enable/disable reference selection based on quality-level
 [no] ref-order       - Priority order of timing references
      ref1            + Configure parameters for the first timing reference
      ref2            + Configure parameters for the second timing reference
 [no] revert          - Revert/do not revert to a higher priority re-validated
                        reference source
 [no] wait-to-restore - Configure the wait-to-restore timer
```

The synchronous interface timing configuration parameters for the first timing reference ref1 are the following:

```
*A:PE-1# configure system sync-if-timing ref1
  - ref1

 [no] ql-override     - Override the quality level of a timing reference
 [no] shutdown        - Administratively shutdown the timing reference
 [no] source-port     - Configure the source port for the first timing reference
```

The synchronous interface timing for ref1 with source port 1/1/2 is configured as follows:

```
configure
    system
        sync-if-timing
            begin
            ref-order bits ref1             # default setting
            ref1
                source-port 1/1/2
                no shutdown
            exit
            bits
                interface-type ds1 esf      # default setting
                input
                    no shutdown
                exit
            exit
            revert
            commit
```

The detailed settings for the synchronous interface timing are as follows:

```
*A:PE-1>config>system>sync-if-timing# info detail
----------------------------------------------
            no ql-minimum
            no ql-selection
            ref-order bits ref1 ref2 ptp
            ref1
                source-port 1/1/2
                no shutdown
                no ql-override
            exit
            ref2
                shutdown
                no source-port
                no ql-override
            exit
            bits
                interface-type ds1 esf
                no ql-override
                input
                    no shutdown
                exit
                output
                    shutdown
                    line-length 110
                    no ql-minimum
                    source line-ref
                    no squelch
                exit
            exit
            ptp
                shutdown
                no ql-override
            exit
            revert
            no wait-to-restore
----------------------------------------------
*A:PE-1>config>system>sync-if-timing#
```

The following output displays the associated show information.

```
*A:PE-1# show system sync-if-timing

===============================================================================
System Interface Timing Operational Info
===============================================================================
System Status CPM A                : Master Locked
    Reference Input Mode           : Revertive
    Quality Level Selection        : Disabled
    Reference Selected             : ref1
    System Quality Level           : unknown
    Current Frequency Offset (ppm) : +0
    Input Minimum Quality Level    : none
    Wait to Restore Timer          : Disabled

Reference Order                    : bits ref1 ref2 ptp

Reference Mate CPM
    Qualified For Use              : No
```

```
        Not Qualified Due To       :     LOS
    Selected For Use               : No
        Not Selected Due To        :     not qualified

Reference Input 1
    Admin Status                   : up
    Rx Quality Level               : unknown
    Quality Level Override         : none
    Qualified For Use              : Yes
    Selected For Use               : Yes
    Source Port                    : 1/1/2

Reference Input 2
    Admin Status                   : down
    Rx Quality Level               : unknown
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
    Source Port                    : None

Reference BITS A
    Input Admin Status             : up
    Rx Quality Level               : failed
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     LOS
    Selected For Use               : No
        Not Selected Due To        :     not qualified
    Interface Type                 : DS1
    Framing                        : ESF
    Line Coding                    : B8ZS
    Line Length                    : 0-110ft
    Output Admin Status            : down
    Output Minimum Quality Level   : none
    Output Source                  : line reference
    Output Reference Selected      : none
    Output Squelch                 : Disabled
    Tx Quality Level               : N/A

Reference PTP
    Admin Status                   : down
    Rx Quality Level               : failed
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
===============================================================================
*A:PE-1#
```

Configuration 2 - QL Selection Mode Enabled

The following example shows the configuration options for SyncE when ql-selection mode is enabled.

This is the normal case for European SDH networks.

SyncE is enabled as follows:

```
*A:PE-1# configure card 1 mda 1 sync-e
```

On port 1/1/2, the Synchronization Status Message (SSM) channel is configured to SDH, as follows:

```
*A:PE-1# configure port 1/1/2 ethernet ssm
  - ssm

 [no] code-type       - Set the SSM channel to either use sonet or sdh
 [no] shutdown        - Enable/Disable SSM
 [no] tx-dus          - Enable/disable always transmit 0xF (dus/dnu) in SSM
                        messaging channel


configure port 1/1/2 ethernet ssm code-type sdh
configure port 1/1/2 ethernet ssm no shutdown
```

The synchronization interface timing is configured as follows with timing reference ref1:

```
configure
    system
        sync-if-timing
            begin
            ql-selection
            ref-order bits ref1             # default setting
            ref1
                source-port 1/1/2
                no shutdown
            exit
            bits
                interface-type e1 pcm31crc  # for Europe
                ql-override prc             # for Europe
                input
                    no shutdown
                exit
            exit
            revert
            commit
```

The European QL-codes are the following: prc, ssu-a, ssu-b, sec, eec1. For North America, the QL-codes are: prs, stu, st2, tnc, st3e, st3, eec2. In this configuration example, Primary Reference Clock (PRC) is chosen.

```
*A:PE-1>config>system>sync-if-timing# info detail
----------------------------------------------
            no ql-minimum
            ql-selection
            ref-order bits ref1 ref2 ptp
            ref1
                source-port 1/1/2
                no shutdown
                no ql-override
            exit
            ref2
                shutdown
                no source-port
                no ql-override
            exit
            bits
                interface-type e1 pcm31crc
                ssm-bit 8
                ql-override prc
                input
                    no shutdown
                exit
                output
                    shutdown
                    no ql-minimum
                    source line-ref
                    no squelch
                exit
            exit
            ptp
                shutdown
                no ql-override
            exit
            revert
            no wait-to-restore
----------------------------------------------
```

The following output displays the associated show information.

```
*A:PE-1# show system sync-if-timing

===============================================================================
System Interface Timing Operational Info
===============================================================================
System Status CPM A                : Master Holdover
    Reference Input Mode           : Revertive
    Quality Level Selection        : Enabled
    Reference Selected             : none
    System Quality Level           : st3
    Current Frequency Offset (ppm) : +0
    Input Minimum Quality Level    : none
    Wait to Restore Timer          : Disabled

Reference Order                    : bits ref1 ref2 ptp

Reference Mate CPM
    Qualified For Use              : No
        Not Qualified Due To       :     LOS
    Selected For Use               : No
        Not Selected Due To        :     not qualified

Reference Input 1
    Admin Status                   : up
    Rx Quality Level               : failed
    Quality Level Override         : none
    Qualified For Use              : Yes
    Selected For Use               : No
        Not Selected Due To        :     ssm quality
    Source Port                    : 1/1/2

Reference Input 2
    Admin Status                   : down
    Rx Quality Level               : unknown
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
    Source Port                    : None

Reference BITS A
    Input Admin Status             : up
    Rx Quality Level               : failed
    Quality Level Override         : prc
    Qualified For Use              : No
        Not Qualified Due To       :     LOS
    Selected For Use               : No
        Not Selected Due To        :     not qualified
    Interface Type                 : E1
    Framing                        : PCM31 CRC
    Line Coding                    : HDB3
    Line Length                    : 8
    Output Admin Status            : down
    Output Minimum Quality Level   : none
    Output Source                  : line reference
    Output Reference Selected      : none
    Output Squelch                 : Disabled
    Tx Quality Level               : N/A

Reference PTP
    Admin Status                   : down
    Rx Quality Level               : failed
    Quality Level Override         : none
    Qualified For Use              : No
        Not Qualified Due To       :     disabled
    Selected For Use               : No
        Not Selected Due To        :     disabled
===============================================================================
*A:PE-1#
```

Conclusion

With the world rapidly transitioning to IP/MPLS-based NGNs with Ethernet as the transport medium of choice, there is an increasing need to enhance services and capabilities while still leveraging existing infrastructure, thereby easing the transition while continuing to increase revenue and reduce the Total Cost of Ownership (TCO). In areas such as mobile backhaul, TDM CES etc., these requirements create a need for SONET/SDH-like frequency synchronization capability in the inherently asynchronous Ethernet network.

SyncE, natively supported on the Nokia SR OS routers, is an ITU-T standardized PHY-level way of transmitting frequency synchronization across Ethernet packet networks that fulfills that need in a reliable, secure, scalable, efficient, and cost-effective manner. It allows service providers to keep existing revenue streams alive and create new ones while simplifying the network design and reducing TCO.

System Management

In This Section

This section provides configuration information for the following topics:

- Distributed CPU Protection
- Event Handling System
- SR OS NETCONF Server Basics

Distributed CPU Protection

This chapter describes Distributed CPU Protection (DCP) configurations.

Topics in this chapter include:

Applicability

This chapter was originally written for SR OS release 11.0R1. The CLI in the current edition corresponds to release 15.0.R1.

Overview

SR OS provides several rate limiting mechanisms to protect the CPM/CFM processing resources of the router:

- CPU Protection: A centralized rate limiting function that operates on the CPM to limit traffic destined to the CPUs.
- Distributed CPU Protection: A control traffic rate limiting protection mechanism for the CPM/CFM that operates on the line cards (hence ‘distributed’). CPU protection protects the CPU of the node that it is configured on from a DOS attack by limiting the amount of traffic coming in from one of its ports and destined to the CPM (to be processed by its CPU) using a combination of the configurable limits.

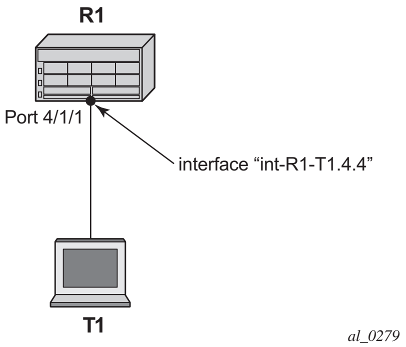
The goal of this chapter is to familiarize the reader with the configuration and use of DCP. A simple and controlled setup is used to illustrate how the protection behaves and how to use the tools provided for the feature.

External testing equipment (“tester”) is used to send control traffic of various protocols at various rates to the router in order to exercise DCP. Log events and show routines are examined to explain the indications that the router provides to an operator.

Configuration

The test topology is shown in Figure 15. A 10Gb Ethernet link is used between the tester and the router.

Figure 15 **Test Topology**



Step 1. The basic configuration of the MDA, port, interface and a security event log on the router is as follows.

```
configure
    card 4
        card-type iom3-xp-b
        mda 1
            mda-type m4-10gb-xp-xfp
            no shutdown
        exit
    exit
exit
```

```
configure
    port 4/1/1
        ethernet
        exit
        no shutdown
    exit
exit
```

```
configure
    router Base
        interface "int-R1-T1.4.4"
            address 192.168.40.1/24
            description "to tester T1, port 4.4"
            port 4/1/1
            no shutdown
        exit
    exit
exit
```

```
configure
    log
        log-id 15
            from security
            to memory 1024
            no shutdown
        exit
    exit
exit
```

This chapter was originally developed on a 7750 SR-c12 platform but it is equally applicable to other platforms such as the 7750 SR-7/12. If other platforms, such as the 7750 SR-7/12 that support centralized CPU Protection, are used to explore DCP then the centralized CPU Protection should be disabled (for the purposes of this chapter) so that it does not interfere with reproducing the same results as described below. In a normal production network, CPU Protection and DCP are complimentary and can be used together. To disable centralized CPU Protection for the purposes of reproducing the results below please ensure that:

- **protocol-protection** is disabled.
- All rates in all polices (including any default polices) are configure to **max**.

Step 2. In order to activate DCP a policy is created and assigned to the interface.

The first policy that is used in this chapter is used to simply count protocol packets to see that they are indeed flowing from the tester to the router and being extracted and identified.

The *dcp-policy-count* policy is configured as follows:

```
configure
    system
        security
            dist-cpu-protection
                policy "dcp-policy-count" create
                    description "Static policers with rate 0 for counting packets"
                    static-policer "sp-arp" create
                        description "static policer for ARP"
                        rate packets 0 within 1
                    exit
                    static-policer "sp-icmp" create
                        description "static policer for ICMP"
                        rate packets 0 within 1
                    exit
                    static-policer "sp-igmp" create
                        description "static policer for IGMP"
                        rate packets 0 within 1
                    exit
                    protocol arp create
                        enforcement static "sp-arp"
                    exit
                    protocol icmp create
                        enforcement static "sp-icmp"
                    exit
                    protocol igmp create
                        enforcement static "sp-igmp"
                    exit
                exit
            exit
        exit
    exit
exit
```

For the *dcp-policy-count policy* configuration:

- The policy contains three static policers: s*p-arp*, *sp-icmp* and *sp-igmp*. These policers are then used by the three configured protocols that are part of the policy: *arp*, *icmp* and *igmp*.
- The list of protocols that are applicable to DCP are as follows: arp, dhcp, http-redirect, icmp, igmp, mld, ndis, pppoe-pppoa, all-unspecified, mpls-ttl, bfd-cpm, bgp, eth-cfm, isis, ldp, ospf, pim and rsvp. The all-unspecified protocol is a special “catch-all”. See the 7750 SR OS System Management Guide for more details.
- This policy instantiates three permanent (static) policers for every object (for example, interface) that the policy is associated with.
- The three protocols each reference their own static policer so each protocol will be independently rate limited. A single static policer can also be used to rate limit multiple protocols but that capability is not used in this chapter.
- The rate is set to 0 which means all packets will be considered as non-conformant to the policer. This configuration is used to provide counters of protocol packets. The DCP counters provide the count of packets exceeding the policing parameters since the given policer was previously declared as conformant or newly instantiated. A rate of zero ensures that the policer will never be declared as conformant and hence will never reset the counters.
- The exceed-action is not configured and takes the default value of *none.* The *log-events* parameter is not configured and is enabled by default. That means the policer will notify the operator when the first packet arrives but will not discard or mark any packets.

Step 3. Assign the *dcp-policy-count* to the interface:

```
*A:R1# configure router interface "int-R1-T1.4.4"
*A:R1>config>router>if# dist-cpu-protection "dcp-policy-count"
```

Step 4. Examine some log and status on the router to get a baseline (no traffic is flowing from the tester to the router at this point). Notice that the CPU utilization is fairly low with an overall Idle of 94% and no task groups at more than 2.5% capacity usage. Future example output from this show routine will be snipped to only show relevant and interesting lines.

```
*A:R1# show system cpu
```

```
===============================================================================
CPU Utilization (Sample period: 1 second)
===============================================================================
Name                                   CPU Time       CPU Usage        Capacity
                                         (uSec)                           Usage
-------------------------------------------------------------------------------
BFD                                          60          ~0.00%          ~0.00%
BGP                                      30,892           0.34%           0.62%
BGP PE-CE                                     0           0.00%           0.00%
CALLTRACE                                 5,210           0.05%           0.51%
CFLOWD                                    5,128           0.05%           0.51%
Cards & Ports                            39,591           0.44%           0.95%
DHCP Server                                  35          ~0.00%          ~0.00%
ETH-CFM                                   4,584           0.05%           0.46%
HQoS Algorithm                                0           0.00%           0.00%
HQoS Statistics                               0           0.00%           0.00%
ICC                                       1,225           0.01%           0.12%
IGMP/MLD                                  1,080           0.01%           0.10%
IMSI Db Appl                                258          ~0.00%           0.01%
IOM                                           0           0.00%           0.00%
IP Stack                                 56,965           0.63%           0.51%
IS-IS                                    51,342           0.57%           0.60%
ISA                                      16,173           0.18%           0.55%
LDP                                      31,118           0.34%           0.55%
Logging                                      53          ~0.00%          ~0.00%
MBUF                                          0           0.00%           0.00%
MCS                                         536          ~0.00%           0.04%
MPLS/RSVP                                 8,915           0.09%           0.57%
MSCP                                          0           0.00%           0.00%
MSDP                                          0           0.00%           0.00%
Management                               18,039           0.20%           0.73%
OAM                                      12,422           0.13%           0.48%
OSPF                                    118,279           1.32%           0.58%
OpenFlow                                  1,037           0.01%           0.01%
PIM/L2Mcast                                   0           0.00%           0.00%
PKI                                         272          ~0.00%           0.02%
PTP                                          71          ~0.00%          ~0.00%
RIP                                           0           0.00%           0.00%
RTM/Policies                                  0           0.00%           0.00%
Redundancy                               10,321           0.11%           0.63%
SNMP Daemon                                   0           0.00%           0.00%
Security                                      0           0.00%           0.00%
Services                                  8,775           0.09%           0.52%
Stats                                         0           0.00%           0.00%
Subscriber Mgmt                           6,439           0.07%           0.14%
System                                   94,262           1.05%           2.28%
Traffic Eng                                   0           0.00%           0.00%
VRRP                                      1,942           0.02%           0.12%
WEB Redirect                                 95          ~0.00%          ~0.00%
-------------------------------------------------------------------------------
Total                                 8,936,439         100.00%
   Idle                               8,411,320          94.12%
   Usage                                525,119           5.87%
Busiest Core Utilization                 79,550           8.01%
===============================================================================
*A:R1#
```

The DCP feature is reporting no violations for interfaces on card 4.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

There are no security log events.

```
*A:R1# show log log-id 15
===============================================================================
Event Log 15
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=1024   next event=1  (not wrapped)]
*A:R1#
```

The detailed DCP status for the interface shows all three policers are currently in the conform state.

```
*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection

===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-policy-count

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
Policer-Name        : sp-arp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : arp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-icmp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : icmp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-igmp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : igmp
```

```
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

Step 5. Configure the tester to send ARP, ICMP and IGMP traffic to the router using the following rates:

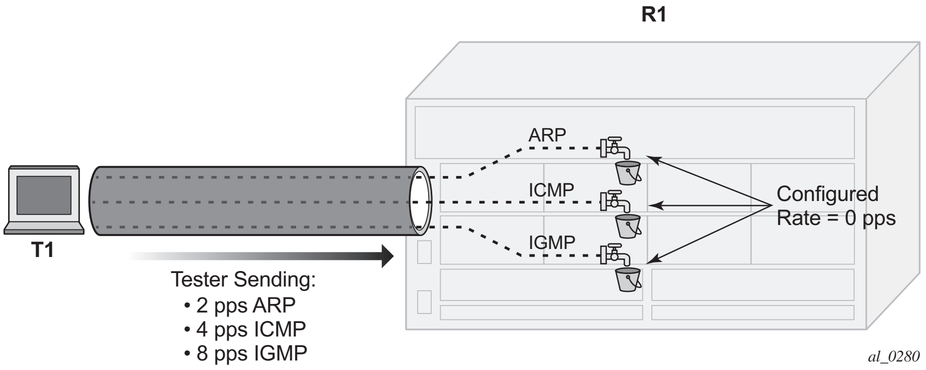
- – ARP: 2 packets per second (pps)
- – ICMP: 4 pps
- – IGMP: 8 pps

Here are some tips for how to configure the tester to send protocol packets that will be recognized by the router:

- – ARP:
  - Set the MAC destination address to FF-FF-FF-FF-FF-FF
  - Use an ARP Request format
- – ICMP:
  - Use an ICMP type of 8 (echo request, such as **ping**).
  - Set the MAC destination address equal to the MAC address of the receiving port. The MAC address of port 4/1/1 can be seen in the output of show port 4/1/1 as the configured address.
  - Set the IP destination address to 192.168.40.1 and the IP source address to 192.168.40.2.
- – IGMP:
  - Set the MAC destination address equal to the MAC address of the receiving port. The MAC address of port 4/1/1 can be seen in the output of show port 4/1/1 as the configured address.
  - Set the IP destination address to 224.0.0.2 and the IP source address to 0.0.0.0.
  - Set the IGMP version to 2, make the IGMP message type a Membership Query to Group 0.

Also ensure that the tester interleaves the three streams of protocol packets such that it schedules them independently in an interleaved fashion, not serially.

Figure 16 **Count Traffic with DCP Policy Count**



Step 6. Notice that DCP now reports some violations of the policy against the interface.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
int-R1-T1.4.4                   sp-arp                          [S] none
int-R1-T1.4.4                   sp-icmp                         [S] none
int-R1-T1.4.4                   sp-igmp                         [S] none
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

After a few seconds, the DCP exceed-count values can be seen incrementing.

Note the following details:

- Exceed-Count is non-zero. This will continue incrementing and will never reset since the rate configured in the DCP policy is zero.
- The Policer-State is Exceed. The policers have detected that the protocol is non-conformant to the configured rate.
- Detec. Time Remain stays at 29 seconds. This countdown timer is automatically reset to 30 seconds every time a policer is detected as non-conformant (which will be continually when the rate is set to 0 and packets of that protocol are being received).

```
*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection
```

```
===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-policy-count

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
Policer-Name        : sp-arp
Card/FP             : 4/1               Policer-State       : Exceed
Protocols Mapped    : arp
Exceed-Count        : 263
Detec. Time Remain  : 29 seconds        Hold-Down Remain.   : none

Policer-Name        : sp-icmp
Card/FP             : 4/1               Policer-State       : Exceed
Protocols Mapped    : icmp
Exceed-Count        : 525
Detec. Time Remain  : 29 seconds        Hold-Down Remain.   : none

Policer-Name        : sp-igmp
Card/FP             : 4/1               Policer-State       : Exceed
Protocols Mapped    : igmp
Exceed-Count        : 1050
Detec. Time Remain  : 29 seconds        Hold-Down Remain.   : none
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

Step 7. Keep the tester running.

Now a DCP policy that enforces protocol rates using static policers will be applied to the interface. First, the policy is created:

```
configure
    system
        security
            dist-cpu-protection
                    description "Static policers for arp, icmp and igmp"
                    static-policer "sp-arp" create
                        rate packets 10 within 1
                        exceed-action discard
                    exit
                    static-policer "sp-icmp" create
                        rate packets 20 within 1
                        exceed-action discard
                    exit
                    static-policer "sp-igmp" create
                        rate packets 10 within 1
                        exceed-action discard
                    exit
                    protocol arp create
                        enforcement static "sp-arp"
                    exit
                    protocol icmp create
                        enforcement static "sp-icmp"
                    exit
                    protocol igmp create
                        enforcement static "sp-igmp"
                    exit
                exit
            exit
        exit
    exit
exit
```

For the dcp-static-policy-1 policy configuration, note that a few parameters are different than in the previously created dcp-policy-count policy:

- The rates are set to low (but non-zero) values.
- The exceed-action is configured such that packets are dropped once the rate is exceeded.

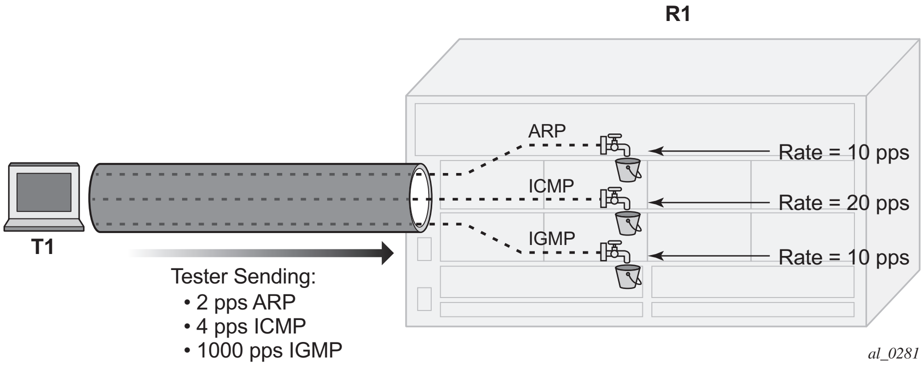
Now assign the policy to the test interface:

```
*A:R1# configure router interface "int-R1-T1.4.4"
                        dist-cpu-protection "dcp-static-policy-1"


*A:R1# show system security dist-cpu-protection policy "dcp-static-policy-
1" association

===============================================================================
Distributed CPU Protection Policy
===============================================================================
Policy Name : dcp-static-policy-1
Description : Static policers for arp, icmp and igmp

-------------------------------------------------------------------------------
Associations
-------------------------------------------------------------------------------

SAP associations
-------------------------------------------------------------------------------
  None

Managed SAP associations
-------------------------------------------------------------------------------
  None

Interface associations
-------------------------------------------------------------------------------
Router-Name : Base
  int-R1-T1.4.4
```

```
-------------------------------------------------------------------------------
Number of interfaces : 1
===============================================================================
*A:R1#
```

Step 8. Increase the rate of IGMP packets that the tester is sending to 1000pps (keep ARP and ICMP at 2pps and 4pps).

Figure 17 **Limit Traffic with dcp-static-policy-1**



Step 9. Notice that the system has identified a violation of the DCP rates for the IGMP policer.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
int-R1-T1.4.4                   sp-igmp                         [S] none
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

After a few minutes, the violation will be indicated as a log event. This delay is due to the design of DCP. In order to support large scale operation of DCP, and also to avoid overload conditions, a polling process is used to monitor state changes in the policers and to gather violations. This means there can be a delay between when an event occurs in the data plane and when the relevant state change or event notification occurs towards an operator, but in the meantime the policers are still operating and protecting the control plane.

```
*A:R1# show log log-id 15
===============================================================================
Event Log 15
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=1024   next event=2  (not wrapped)]


1 2017/04/27 09:47:53.21 CEST WARNING: SECURITY #2066 Base DCPUPROT
"Non conformant network_if "int-R1-
T1.4.4" on fp 4/1 detected at 04/27/2017 09:47:07.
 Policy "dcp-static-policy-1". Policer="sp-igmp"(static). Excd count=411"
*A:R1#


*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection

===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-static-policy-1

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
Policer-Name        : sp-arp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : arp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-icmp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : icmp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-igmp
Card/FP             : 4/1               Policer-State       : Exceed
Protocols Mapped    : igmp
Exceed-Count        : 640151
Detec. Time Remain  : 29 seconds        Hold-Down Remain.   : none
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

The status of DCP on the interface also shows the igmp policer as being in an Exceed state:

The CPU utilization of the IGMP task group is not impacted since DCP is discarding packets that are non-conformant to the configure rate.

```
*A:R1# show system cpu

===============================================================================
CPU Utilization (Sample period: 1 second)
===============================================================================
Name                                       CPU Time       CPU Usage       Capacity
                                             (uSec)                           Usage
-------------------------------------------------------------------------------
BFD                                             160          ~0.00%          0.01%
--- snipped ---
IGMP/MLD                                      1,795           0.02%          0.09%
--- snipped ---
Stats                                             0           0.00%          0.00%
Subscriber Mgmt                               5,661           0.06%          0.09%
System                                       77,189           0.86%          1.23%
Traffic Eng                                       0           0.00%          0.00%
VRRP                                          1,679           0.01%          0.09%
WEB Redirect                                     83          ~0.00%         ~0.00%
-------------------------------------------------------------------------------
Total                                     8,936,280         100.00%
   Idle                                   8,420,519          94.22%
   Usage                                    515,761           5.77%
Busiest Core Utilization                     78,908           7.94%
===============================================================================
*A:R1#
```

Step 10. Remove the DCP policy from the interface and see the capacity usage going up for the IGMP task group.

```
*A:R1# configure router interface "int-R1-T1.4.4" no dist-cpu-protection
*A:R1#


*A:R1# show system cpu

===============================================================================
CPU Utilization (Sample period: 1 second)
===============================================================================
Name                                       CPU Time       CPU Usage       Capacity
                                             (uSec)                           Usage
-------------------------------------------------------------------------------
BFD                                             191          ~0.00%          0.01%
--- snipped ---
ICC                                           1,332           0.01%          0.13%
IGMP/MLD                                     78,184           0.87%          7.78%
IMSI Db Appl                                    249          ~0.00%         ~0.00%
IOM                                               0           0.00%          0.00%
IP Stack                                    183,448           2.05%          9.16%
IS-IS                                        49,866           0.55%          0.56%
--- snipped ---
Subscriber Mgmt                               8,153           0.09%          0.14%
System                                      144,738           1.62%          6.11%
Traffic Eng                                       0           0.00%          0.00%
```

```
VRRP                                      2,396           0.02%           0.15%
WEB Redirect                                 83          ~0.00%          ~0.00%
-------------------------------------------------------------------------------
Total                                 8,925,786         100.00%
   Idle                               8,123,698          91.01%
   Usage                                802,088           8.98%
Busiest Core Utilization                170,131          17.15%
===============================================================================
*A:R1#
```

Step 11. Increase the rate of IGMP traffic from the tester to 5000 pps. See the CPU utilization increase further.

```
*A:R1# show system cpu

===============================================================================
CPU Utilization (Sample period: 1 second)
===============================================================================
Name                                   CPU Time       CPU Usage        Capacity
                                         (uSec)                           Usage
-------------------------------------------------------------------------------
BFD                                         158          ~0.00%           0.01%
--- snipped ---
ICC                                       1,061           0.01%           0.10%
IGMP/MLD                                398,106           4.44%          39.99%
IMSI Db Appl                                142          ~0.00%          ~0.00%
IOM                                           0           0.00%           0.00%
IP Stack                                648,378           7.24%          43.68%
IS-IS                                    59,623           0.66%           0.65%
--- snipped ---
Subscriber Mgmt                           7,308           0.08%           0.13%
System                                  364,124           4.06%          28.73%
Traffic Eng                                   0           0.00%           0.00%
VRRP                                      2,156           0.02%           0.12%
WEB Redirect                                117          ~0.00%           0.01%
-------------------------------------------------------------------------------
Total                                 8,951,453         100.00%
   Idle                               7,114,732          79.48%
   Usage                              1,836,721          20.51%
Busiest Core Utilization                590,342          59.35%
===============================================================================
*A:R1#
```

Step 12. Reinstall the DCP policy to the interface and see the CPU utilization drop.

```
*A:R1# configure router interface "int-R1-T1.4.4"
                        dist-cpu-protection "dcp-static-policy-1"


*A:R1# show system cpu

===============================================================================
CPU Utilization (Sample period: 1 second)
===============================================================================
Name                                   CPU Time       CPU Usage        Capacity
                                         (uSec)                           Usage
-------------------------------------------------------------------------------
BFD                                          72          ~0.00%          ~0.00%
```

```
--- snipped ---
ICC                                       1,000           0.01%           0.10%
IGMP/MLD                                  2,149           0.02%           0.12%
IMSI Db Appl                                166          ~0.00%          ~0.00%
IOM                                           0           0.00%           0.00%
IP Stack                                 60,407           0.67%           0.55%
IS-IS                                    50,966           0.57%           0.58%
--- snipped ---
Subscriber Mgmt                           5,847           0.06%           0.09%
System                                   96,233           1.07%           2.23%
Traffic Eng                                   0           0.00%           0.00%
VRRP                                      2,338           0.02%           0.15%
WEB Redirect                                 94          ~0.00%          ~0.00%
-------------------------------------------------------------------------------
Total                                 8,925,256         100.00%
   Idle                               8,396,016          94.07%
   Usage                                529,240           5.92%
Busiest Core Utilization                 81,620           8.22%
===============================================================================
*A:R1#
```

Step 13. Stop the tester from sending packets, wait a few minutes and then note the status of the system.

There are no longer any violations of any enforcement policers on any interfaces on card 1.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

The IGMP policer is indicated as conformant in the log events.

```
*A:R1# show log log-id 15

===============================================================================
Event Log 15
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=1024   next event=4  (not wrapped)]

3 2017/04/27 10:02:53.25 CEST WARNING: SECURITY #2072 Base DCPUPROT
"Network_if "int-R1-
T1.4.4" on fp 4/1 newly conformant at 04/27/2017 10:02:04. Policy
 "dcp-static-policy-1". Policer="sp-igmp"(static). Excd count=227756"
```

```
--- snipped ---

*A:R1#
```

The interface DCP details show all policers as conformant.

```
*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection

===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-static-policy-1

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
Policer-Name        : sp-arp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : arp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-icmp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : icmp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none

Policer-Name        : sp-igmp
Card/FP             : 4/1               Policer-State       : Conform
Protocols Mapped    : igmp
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

An optional hold-down can be used in the configuration of the exceed-action of the policers in order to apply the exceed-action for a defined period (even if the policer goes conformant again during that period). The hold-down could be used, for, to discard all packets associated with a policer for one hour after a violation is detected.   An “indefinite” period is also supported which enforces discard or marking until the operator clears the policer with the **tools perform security dist-cpu-protection release-hold-down** command.

Step 14. The next scenario explored in this chapter is the use of DCP dynamic enforcement.

In order to use dynamic enforcement policers, a number of dynamic policers must be allocated to the DCP pool for the particular card being used.

```
*A:R1# configure card 4 fp dist-cpu-protection dynamic-enforcement-policer-
pool 1000
*A:R1#
```

The number allocated should be greater than the maximum number of dynamic policers expected to be in use on the card at one time. A conservative (large) number could be selected at first, and then the following show command can give data to help tune the pool to a smaller size over time:

```
*A:R1# show card 4 fp 1 dist-cpu-protection

===============================================================================
Card : 4 Forwarding Plane(FP) : 1
===============================================================================
Dynamic Enforcement Policer Pool : 1000
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Statistics Information
-------------------------------------------------------------------------------
Dynamic-Policers Currently In Use      : 0
Hi-WaterMark Hit Count                 : 0
Hi-WaterMark Hit Time                  : 04/27/2017 10:08:24 UTC
Dynamic-Policers Allocation Fail Count : 0
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

If the dynamic-enforcement-policer-pool is too small then when a local-monitoring-policer detects violating traffic, the dynamic enforcement policers will not be able to be instantiated. A log event will warn the operator when the pool is nearly exhausted.

A sample dynamic enforcement policy is created as follows:

```
configure
    system
        security
```

```
            dist-cpu-protection
                policy "dcp-dynamic-polcy-1" create
                    description "Dynamic policing policy"
                    local-monitoring-policer "local-mon" create
                        description "Monitor for arp, icmp, igmp and all-
unspecified"
                        rate packets 100 within 10
                    exit
                    protocol arp create
                        enforcement dynamic "local-mon"
                        dynamic-parameters
                            rate packets 20 within 10
                            exceed-action discard
                        exit
                    exit
                    protocol icmp create
                        enforcement dynamic "local-mon"
                        dynamic-parameters
                            rate packets 20 within 10
                            exceed-action discard
                        exit
                    exit
                    protocol igmp create
                        enforcement dynamic "local-mon"
                        dynamic-parameters
                            rate packets 20 within 10
                            exceed-action discard
                        exit
                    exit
                    protocol all-unspecified create
                        enforcement dynamic "local-mon"
                        dynamic-parameters
                            rate packets 100 within 10
                            exceed-action discard
                        exit
                    exit
                exit
            exit
        exit
    exit
exit
```

For the *dcp-dynamic-policy-1* policy configuration:

- The policy contains no static policers. Per-protocol enforcement policers will be instantiated dynamically but only if triggered by a violation of the local-monitoring-policer.
- A local-monitoring-policer is configured for the policy. The configured rate determines the rate of arriving protocol packets at which the policy will trigger the automatic instantiation of dynamic per-protocol policers for the interface.
- Four protocols are configured and they are all associated with the local-monitoring-policer. The all-unspecified protocol will include all other extracted control packets on the interface.

- Each protocol has its own configured dynamic rates that will be used by the dynamic enforcement policers if they are instantiated. Note these rates are lower than previous scenarios (the **within** parameter is 10 seconds instead of 1 second).
- When this DCP policy is associated with an interface, only a single policer (the local-monitoring-policer) will be instantiated (statically/permanently). The per-protocol dynamic policers are only instantiated when there is a violation of the local-monitoring-policer.

The policy is then associated with the interface:

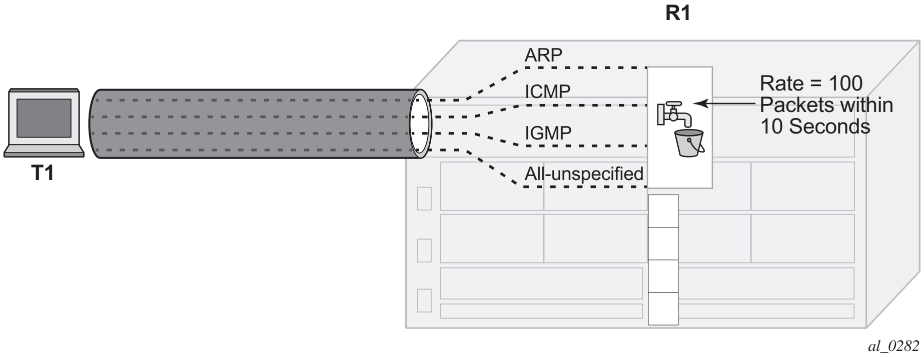
```
*A:R1# configure router interface "int-R1-T1.4.4"
                            dist-cpu-protection "dcp-dynamic-polcy-1"
*A:R1#
```

Step 15. Configure the tester to send:

- 1pps of ARP
- 4pps of ICMP
- 1000pps of IGMP

Start the tester.

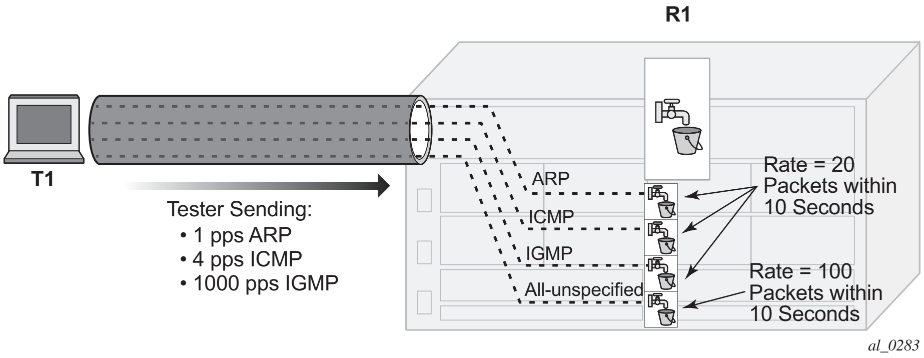
Figure 18 **Dynamic Policing – Local Monitor**



In Figure 18, the dynamic policers have not been instantiated yet.

Step 16. The local-monitoring-policer will become non-conforming since the aggregate arrival rate of arp+icmp+igmp+all-unspecified packets is greater than the configured local-monitoring-policer rate of 100 packets within 10 seconds. Dynamic enforcement policers will then be instantiated.

Figure 19 **Dynamic Policers Instantiated**



The ICMP and IGMP dynamic policers will see violations since their dynamic rates are being exceeded.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
int-R1-T1.4.4                   icmp                            [D] none
int-R1-T1.4.4                   igmp                            [D] none
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

The ARP and all-unspecified dynamic policers were instantiated but will be counting down their detection time (if this show command is issued within 30 seconds of the attack starting).

```
*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection

===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-dynamic-polcy-1

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
Policer-Name        : local-mon
Card/FP             : 4/1                Policer-State       : Exceed
Protocols Mapped    : arp, icmp, igmp, all-unspecified
Exceed-Count        : 1249
All Dyn-Plcr Alloc. : True
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
Protocol(Dyn-Plcr)  : arp
Card/FP             : 4/1                Protocol-State      : Conform
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds          Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : icmp
Card/FP             : 4/1                Protocol-State      : Exceed
Exceed-Count        : 72
Detec. Time Remain  : 26 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : igmp
Card/FP             : 4/1                Protocol-State      : Exceed
Exceed-Count        : 56190
Detec. Time Remain  : 29 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : all-unspecified
Card/FP             : 4/1                Protocol-State      : Conform
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds          Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

After 30 seconds have passed, the “Detec. Time Remain” for ARP and all-unspecified will simply read 0 (zero).

After a few minutes the log events will be collected indicating a non-conformance was seen.

```
*A:R1# show log log-id 15

===============================================================================
Event Log 15
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=1024   next event=10  (not wrapped)]

9 2017/04/27 10:22:53.32 CEST WARNING: SECURITY #2067 Base DCPUPROT
"Non conformant network_if "int-R1-
T1.4.4" on fp 4/1 detected at 04/27/2017 10:18:37.
 Policy "dcp-dynamic-polcy-1". Policer="icmp"(dynamic). Excd count=2"

8 2017/04/27 10:22:53.32 CEST WARNING: SECURITY #2067 Base DCPUPROT
```

```
"Non conformant network_if "int-R1-
T1.4.4" on fp 4/1 detected at 04/27/2017 10:18:30.
 Policy "dcp-dynamic-polcy-1". Policer="igmp"(dynamic). Excd count=80"

--- snipped ---

*A:R1#
```

Step 17. Stop the tester.

The dynamic policer detection timers will start counting down since they are no longer seeing violating packets.

```
*A:R1# show router interface "int-R1-T1.4.4" dist-cpu-protection

===============================================================================
Interface "int-R1-T1.4.4" (Router: Base)
===============================================================================
Distributed CPU Protection Policy :  dcp-dynamic-polcy-1

-------------------------------------------------------------------------------
Statistics/Policer-State Information
===============================================================================
-------------------------------------------------------------------------------
Static Policer
-------------------------------------------------------------------------------
No entries found
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Local-Monitoring Policer
-------------------------------------------------------------------------------
Policer-Name        : local-mon
Card/FP             : 4/1               Policer-State       : Exceed
Protocols Mapped    : arp, icmp, igmp, all-unspecified
Exceed-Count        : 5072
All Dyn-Plcr Alloc. : True
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Dynamic-Policer (Protocol)
-------------------------------------------------------------------------------
Protocol(Dyn-Plcr)  : arp
Card/FP             : 4/1               Protocol-State      : Conform
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : icmp
Card/FP             : 4/1               Protocol-State      : Exceed
Exceed-Count        : 482
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : igmp
Card/FP             : 4/1               Protocol-State      : Exceed
Exceed-Count        : 326409
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True

Protocol(Dyn-Plcr)  : all-unspecified
```

```
Card/FP             : 4/1               Protocol-State      : Conform
Exceed-Count        : 0
Detec. Time Remain  : 0 seconds         Hold-Down Remain.   : none
Dyn-Policer Alloc.  : True
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

After 30 seconds there are no more violators.

```
*A:R1# tools dump security dist-cpu-
protection violators enforcement interface card 4
===============================================================================
Distributed Cpu Protection Current Interface Enforcer Policer Violators
===============================================================================
Interface                       Policer/Protocol                    Hld Rem
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Violators on Slot-4 Fp-1
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
[S]-Static [D]-Dynamic [M]-Monitor
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

The dynamic policer pool Hi-WaterMark for card 1 fp 1 shows 4 since the highest number of dynamic policers allocated at any one time on the card/fp was 4.

```
*A:R1# show card 4 fp 1 dist-cpu-protection

===============================================================================
Card : 4 Forwarding Plane(FP) : 1
===============================================================================
Dynamic Enforcement Policer Pool : 1000
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
Statistics Information
-------------------------------------------------------------------------------
Dynamic-Policers Currently In Use      : 0
Hi-WaterMark Hit Count                 : 4
Hi-WaterMark Hit Time                  : 04/27/2017 10:10:34 UTC
Dynamic-Policers Allocation Fail Count : 0
-------------------------------------------------------------------------------
===============================================================================
*A:R1#
```

A few minutes later the log events indicate that the flood has ended.

```
*A:R1# show log log-id 15

===============================================================================
Event Log 15
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=1024   next event=12  (not wrapped)]
```

```
11 2017/04/27 10:27:53.34 CEST WARNING: SECURITY #2073 Base DCPUPROT
"Network_if "int-R1-
T1.4.4" on fp 4/1 newly conformant at 04/27/2017 10:24:27. Policy
 "dcp-dynamic-polcy-1". Policer="igmp"(dynamic). Excd count=326409"

10 2017/04/27 10:27:53.34 CEST WARNING: SECURITY #2073 Base DCPUPROT
"Network_if "int-R1-
T1.4.4" on fp 4/1 newly conformant at 04/27/2017 10:24:27. Policy
 "dcp-dynamic-polcy-1". Policer="icmp"(dynamic). Excd count=482"

--- snipped ---

*A:R1#
```

Conclusion

Distributed CPU Protection (DCP) offers a powerful rate limiting function for control protocol traffic that is extracted from the data path and sent to the CPM.

This chapter has demonstrated how to configure DCP on an interface and what indications SR OS provides to the operator during a potential attack or misconfiguration.

DCP can also be deployed in scenarios where per-SAP-per-protocol rate limiting is useful, such as for subscriber management in a subscriber per-VLAN scenario. A DCP policy can be assigned to an MSAP policy on a Broadband Network Gateway, for example, to limit traffic related to certain protocols and to discard certain protocols. When deployed in a subscriber management scenario, DCP can help isolate SAPs (subscribers) from each other and even isolate protocols from each other within an individual SAP (subscriber). Many of the same concepts introduced in this chapter are applicable when DCP is deployed in a subscriber management application.

Event Handling System

This chapter provides information about Event Handling Systems (EHS).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 13.0.R3. The CLI in the current edition is based on SR OS release 15.0.R5.

SR OS release 13.0.R1 introduced Event Handling System (EHS).

SR OS release 14.0.R4 introduced EHS script enhanced capabilities, such as static variables, advanced syntax (shell scripting commands), and so on. The examples in this chapter do not include these enhancements,

Overview

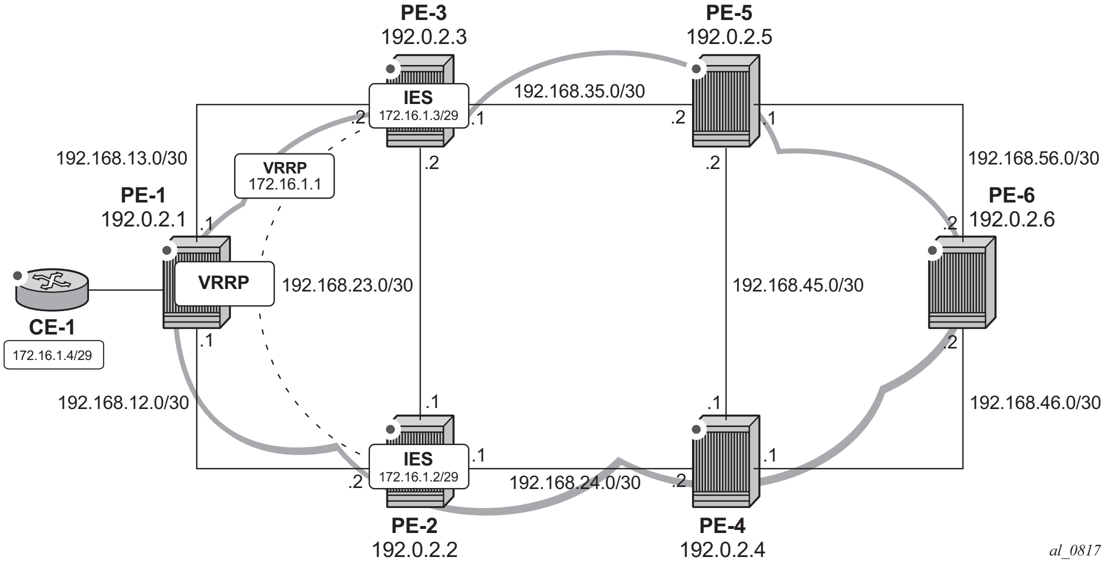
The Event Handling System (EHS) in SR OS allows operators to configure user-defined actions defined in CLI scripts that the router executes in response to an event. The event is referred to as the trigger, where the trigger can be all or part of any event message generated by the event-control framework. The user-defined action is controlled by the script-control function. This script-control function references one or more scripts that are able to execute any command available in CLI when the trigger event occurs.

This feature allows for customized automated event management based on specific operator requirements.

Configuration

The topology shown in Figure 20 provides an example of an EHS configuration. All routers within the example topology participate in the same IS-IS Level-2 area and run LDP. All routers are BGP speakers and form part of Autonomous System 64496, exchanging routes for IPv4 address family only.

Figure 20 **Example Topology**



PE-1 has a CE router connected (CE-1) that is indexed into a VPLS service. This VPLS has spoke-SDPs to an IES instance on both PE-2 and PE-3, which provide a redundant default gateway to CE-1 using the Virtual Router Redundancy Protocol (VRRP). The subnet used for this redundant gateway connectivity between PE-2 and PE-3 is 172.16.1.0/29. The configuration at PE-3 is shown in the following output. The configuration at PE-2 is similar; the exception being IP addressing and VRRP priority, which is 254.

```
# on PE-3
configure
    service
        ies 1 customer 1 create
            interface "redundant-interface" create
                address 172.16.1.3/29
                ip-mtu 1500
                vrrp 1
                    backup 172.16.1.1
                    priority 253
                    ping-reply
                exit
                spoke-sdp 31:1 create
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

The objective is to ensure that both upstream and downstream traffic are always routed through the same PE router. That is, if PE-3 is VRRP Master, it will attract upstream traffic from CE-1 using the VRRP virtual IP/MAC. At the same time, PE-3 should also attract the downstream traffic destined toward CE-1. Having both upstream and downstream traffic transit through the same PE router, simplifies troubleshooting, QoS configuration, and reconciliation of ingress/egress statistics.

In normal operation, PE-2 is the VRRP master and advertises the BGP prefix 172.16.1.0/29 with a local preference of 100 (default value). Similarly, PE-3 is the VRRP backup and advertises the BGP prefix 172.16.1.0/29 with a local preference of 50, using a BGP export policy (route-policy). Therefore, upstream and downstream traffic normally transit through PE-2.

```
*A:PE-3# show router vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
redundant-interface              1     No  Up   Backup      253       1
                                 IPv4      Up   n/a         253        No
  Backup Addr: 172.16.1.1
-------------------------------------------------------------------------------
Instances : 1
===============================================================================
*A:PE-3#
```

```
*A:PE-3# show router bgp routes 172.16.1.0/29 hunt | match expression
                                            "Network|Nexthop|To|Local Pref"
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.3
Res. Nexthop   : 192.0.2.3
Local Pref.    : 50                     Interface Name : system
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.3
To             : 192.0.2.6
Res. Nexthop   : n/a
Local Pref.    : 50                     Interface Name : NotAvailable
*A:PE-3#
```

When PE-3 transitions from backup to master, it must modify its local preference attribute for prefix 172.16.1.0/29 to a value of 150 to attract downstream traffic destined toward CE-1. Similarly, when PE-3 reverts to backup, it must advertise the prefix with a local preference of 50.

Script Control

The first step in configuring event handling is to configure a script containing the CLI commands to be executed when the event is triggered. This script can be stored locally on the compact flash, or it can be stored off-node at a defined remote URL, where it can be accessed using FTP or TFTP. When the script is stored locally on the compact flash and the router is equipped with redundant CPMs, the script must be manually saved on the same compact flash on both CPMs, because it is not synchronized automatically.

The first requirement is to modify the local preference of the prefix 172.16.1.0/29 to 150 on transition to VRRP master. The script, which in this example is held locally on CF3:/, therefore contains the following commands (where the policy-statement, redundant-interface, is the name of the export policy used to advertise the 172.16.1.0/29 prefix):

```
*A:PE-3# file type cf3:/vrrp-master.txt
File: vrrp-master.txt
-------------------------------------------------------------------------------
exit all
configure router policy-options
begin
policy-statement redundant-interface
entry 10
action accept
local-preference 150
exit
exit
exit
commit
exit all

===============================================================================
*A:PE-3#
```

There is no syntax checking when the script file is created; instead, the script will fail with a command error. Also, transactional CLI (for example the edit command) cannot be used in the script, and will fail with a command error.

Within the **system>script-control** context, the script is assigned a name and reference is made to its location. It is then put in the no shutdown state. When the script has been defined, a **script-policy** is configured that calls the previously configured script. The script-policy also specifies a location and filename for a results file that records the successful or unsuccessful conclusion of each script run and each command executed during that run. Each time the script is run, the results are recorded in a file with the name specified for results, followed by an underscore and the date and time when the script was run. A results file must be specified in order for the script to successfully run. The results file can be on the local compact flash, or a remote URL can be specified. As with the script, the script-policy must also be put in the no shutdown state.

```
configure
    system
        script-control
            script "vrrp-master-script"
                location "cf3:/vrrp-master.txt"
                no shutdown
            exit
            script-policy "vrrp-master-policy"
                results "cf3:/script-results.txt"
                script "vrrp-master-script"
                max-completed 4
                expire-time 3600
                lifetime forever
                no shutdown
            exit
        exit
```

The optional lifetime command specifies the maximum time that the script may run. The **max-completed** command specifies the maximum number of script run history status entries to be retained. An optional **expire-time** command specifies the maximum time that the system keeps the run history status (default is 1 h). The system maintains the script run history table, which has a maximum size of 255 entries. Entries are removed from this table when the max-completed or expire-time thresholds are crossed. If the table reaches the maximum value, subsequent script launch requests are not run until older run history entries expire (due to expire-time), or entries are manually cleared. To manually clear entries, the following command is used:

```
clear system script-control script-policy completed <script-policy-name>
```

The script run history status information can be viewed using the following command:

```
*A:PE-3# show system script-control script-policy "vrrp-master-policy"

===============================================================================
Script-policy Information
===============================================================================
Script-policy                : vrrp-master-policy
Script-policy Owner          : TiMOS CLI
Administrative status        : enabled
Operational status           : enabled
Script                       : vrrp-master-script
Script owner                 : TiMOS CLI
Script source location       : cf3:/vrrp-master.txt
Script results location      : cf3:/script-results.txt
Max running allowed          : 1
Max completed run histories  : 4
Max lifetime allowed         : 248d 13:13:56 (21474836 seconds)
Completed run histories      : 1
Executing run histories      : 0
Initializing run histories   : 0
Max time run history saved   : 0d 01:00:00 (3600 seconds)
Script start error           : N/A
Last change                  : 2017/10/26 13:25:14 CEST
Max row expire time          : never
Last application             : event-script
Last auth. user account      : not-specified

===============================================================================
Script Run History Status Information
-------------------------------------------------------------------------------
Script Run #1
-------------------------------------------------------------------------------
Start time    : 2017/10/26 13:31:12 CEST
End time      : 2017/10/26 13:31:12 CEST
Elapsed time  : 0d 00:00:00             Lifetime      : 0d 00:00:00
State         : terminated              Run exit code : noError
Result time   : 2017/10/26 13:31:12 CEST
Keep history  : 0d 00:58:53
Error time    : never
Results file  : cf3:/script-results.txt_20171026-113112-UTC.451977.out
Run exit      : Success
Error         : N/A
Application   : event-script            Auth. user ac*: not-specified
* indicates that the corresponding row element may have been truncated.
===============================================================================
*A:PE-3#
```

Event Handler

The second step in configuring event handling is to assign actions to be performed as a result of the trigger-event. These actions are typically one or more configured scripts defined as entries in an action-list. In the following output, the event-handler is assigned the name event-handler-1, and the action-list consists of a single entry. This entry calls the previously configured script-policy vrrp-master-policy (which in turn references the previously defined script vrrp-master-script). If multiple actions are required based on a single event-trigger, they can be configured in the action-list with subsequent entries, which are run in sequence (up to 1500 action-list entries are supported).

For this example, only a single entry is required; therefore, there is a one to one relationship between the event-handler and the action-list entry. Both the entry within the action-list and the handler should be put in the no shutdown state.

```
configure
    log
        event-handling
            handler "event-handler-1"
```

```
                action-list
                    entry 10
                        script-policy "vrrp-master-policy"
                        no shutdown
                    exit
                exit
                no shutdown
            exit
        exit
```

Event Trigger

The final step in configuring event handling is to configure the event-trigger. The event-trigger defines the event that triggers the running of the script. The event-trigger is based on any event generated by the event-control framework, and can match against the application and event number (event_id). Log filters can also be used to match against specific events using the subject and/or message fields. Regular expressions can be used where required. EHS will not use any message that is suppressed through event-control configuration, or any event message that is throttled.

The general format for an event in an event log is as follows:

```
nnnn YYYY/MM/DD HH:MM:SS.SS Zone <severity>:<application> # <event_id> <router-
name> <subject> description
```

Where:

```
nnnn                    The log entry sequence number
YYYY/MM/DD              The UTC date stamp for the log entry:
                           YYYY - Year
                           MM - Month
                           DD - Date
HH:MM:SS.SS             The UTC time stamp for the event
                           HH - Hours (24 hour format)
                           MM - Minutes
                           SS.SS - Seconds
TZONE                   The timezone
<severity>              The severity level name of the event
<application>           The application generating the log message
<event_id>              The application’s event ID number for the event
<subject>               The subject/affected object for the event
<message>               A textual description of the event
```

In the example, the following event message is generated when PE-3 becomes VRRP Master:

```
58 2017/10/26 13:28:34.401 CEST MINOR: VRRP #2001 Base Becoming Master
"VRRP virtual router instance 1 on interface redundant-interface
(primary address 172.16.1.3) changed state to master"
```

Therefore, the event-trigger configuration is based on an application of VRRP and an event number of 2001 (vrrptrapNewMaster). In the following snippet, vrrp 2001 is configured as the event. The trigger-entry is defined as 1, and in this example, there is only one trigger event. Up to 1500 trigger-entries can be included, each of which can act as a potential trigger event. The trigger-entry also references the previously configured event-handler-1. (Recall that the event-handler references the script-control, which in turn references the script that should be run.)

```
configure
    log
        event-trigger
            event "vrrp" 2001
                trigger-entry 1
                    event-handler "event-handler-1"
                    log-filter 1
                    no shutdown
                exit
                no shutdown
            exit
        exit
    exit
```

Finally, there is a reference to log-filter 1. Without more explicit filtering, event handling will be triggered on any event with the application of VRRP and event number 2001. There may be multiple VRRP instances running on this router, but the requirement is that event handling should only be triggered when the VRRP instance running on redundant-interface transitions to master at PE-3. Therefore, log-filter 1 is used to define a more explicit match using the message field, which contains an explicit reference to the interface. Both the trigger-entry and the event-handler should be put in the no shutdown state.

```
configure
    log
        filter 1
            default-action drop
            entry 10
                action forward
                match
                    message eq pattern "interface redundant-interface (primary
                               address 172.16.1.3) changed state to master"
                exit
            exit
        exit
```

The configuration of the example event handling for the failure event (PE-3 transitions to VRRP master) is now complete. By shutting down the spoke-SDP between PE-1 and PE-2, it is possible to simulate a failure event where the VRRP message path is broken. Therefore, four events are generated.

- The first indicates that PE-3 has become VRRP master for the interface named redundant-interface.

- The second indicates that EHS handler event-handler-1 was invoked by a CLI user.
- The third indicates that a script file has initiated an attempt to execute CLI commands contained in script file vrrp-master.txt.
- The fourth indicates that the attempt to execute those CLI commands was successful (in release 14.0.R5, result “none” is displayed for a successful script run; in release 13.0.R3, the result is “success”).

```
60 2017/10/26 13:31:12.451 CEST MINOR: VRRP #2001 Base Becoming Master
"VRRP virtual router instance 1 on interface redundant-interface
(primary address 172.16.1.3) changed state to master"

61 2017/10/26 13:31:12.451 CEST MINOR: SYSTEM #2069 Base EHS script
"Ehs handler :"event-handler-1" with the description : "" was invoked
by the cli-user account "not-specified"."

62 2017/10/26 13:31:12.461 CEST MAJOR: SYSTEM #2052 Base CLI 'exec'
"A CLI user has initiated an 'exec' operation to process the commands
in the SROS CLI file cf3:/vrrp-master.txt"

63 2017/10/26 13:31:12.472 CEST MAJOR: SYSTEM #2053 Base CLI 'exec'
"The CLI user initiated 'exec' operation to process the commands in
the SROS CLI file cf3:/vrrp-master.txt has completed with the result
of success"
```

A successful script run shows the commands contained in the script, followed by an indication that the commands were executed.

```
*A:PE-3# file type script-results.txt_20171026-113112-UTC.451977.out
File: script-results.txt_20171026-113112-UTC.451977.out
-------------------------------------------------------------------------------
exit all
configure router policy-options
begin
policy-statement redundant-interface
entry 10
action accept
local-preference 150
exit
exit
exit
commit
exit all
Executed 12 lines in 0.0 seconds from file cf3:/vrrp-master.txt

===============================================================================
*A:PE-3#
```

The following output confirms that PE-3 is VRRP master

```
*A:PE-3# show router vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
redundant-interface              1     No  Up   Master       253       1
                                 IPv4      Up   n/a         253        No
  Backup Addr: 172.16.1.1
-------------------------------------------------------------------------------
Instances : 1
===============================================================================
*A:PE-3#
```

Also, the local preference attribute for prefix 172.16.1.0/29 has changed to a value of 150. The result of this action is that PE-3 will now be the transit router for both upstream and downstream traffic.

```
*A:PE-
3# show router bgp routes 172.16.1.0/29 hunt | match expression "Network|Nexthop|To|
Local Pref"
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.3
Res. Nexthop   : 192.0.2.3
Local Pref.    : 150                    Interface Name : system
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.3
To             : 192.0.2.6
Res. Nexthop   : n/a
Local Pref.    : 150                    Interface Name : NotAvailable
*A:PE-3#
```

The event-handler indicates that the referenced script was triggered and run using the command shown in the following output. The Action-List Entry Execution Statistics window provides statistics on the number of times an action (script) was queued to run, and the number of times an error was experienced, both during launch and due to a non-operational admin status. The remainder of the fields in the output are self-explanatory.

```
*A:PE-3# show log event-handling handler "event-handler-1"

===============================================================================
Event Handling System - Handlers
===============================================================================

===============================================================================
Handler           : event-handler-1
===============================================================================
Description       : (Not Specified)
Admin State       : up                             Oper State : up

-------------------------------------------------------------------------------
Handler Execution Statistics
  Success         : 1
  Err No Entry    : 0
  Err Adm Status  : 0
Total             : 1
```

```
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Handler Action-List Entry
-------------------------------------------------------------------------------
Entry-id         : 10
Description      : (Not Specified)
Admin State      : up                                Oper State : up
Script
  Policy Name    : vrrp-master-policy
  Policy Owner   : TiMOS CLI
Min Delay        : 0
Last Exec        : 10/26/17 13:31:12 CEST
-------------------------------------------------------------------------------
Handler Action-List Entry Execution Statistics
  Success        : 1
  Err Min Delay  : 0
  Err Launch     : 0
  Err Adm Status : 0
Total            : 1
===============================================================================
*A:PE-3#
```

The example includes an event-trigger and script to meet the requirements of a fail-forward where PE-3 becomes VRRP master. Now, configuration is needed for when PE-3 reverts to VRRP backup. Without another event-trigger and script, PE-3 will continue to advertise the prefix 172.16.1.0/29 with a local preference of 150 and upstream/downstream traffic will be asymmetric through PE-1/PE-3 respectively.

As before, a script is required. Because PE-2 advertises the prefix with a local preference of 100 (default), PE-3 needs to advertise the same prefix with a lower value (50 in the following output), so that PE-2 is the preferred next hop.

```
*A:PE-3# file type cf3:/vrrp-backup.txt
File: vrrp-backup.txt
-------------------------------------------------------------------------------
exit all
config router policy-options
begin
policy-statement redundant-interface
entry 10
action accept
local-preference 50
exit
exit
exit
commit
exit all

===============================================================================
*A:PE-3#
```

The script must then be configured within the script-control context, and subsequently referenced in a script-policy as vrrp-backup-policy.

```
configure
```

```
    system
        script-control
            script "vrrp-backup-script"
                location "cf1:/vrrp-backup.txt"
                no shutdown
            exit
            script-policy "vrrp-backup-policy"
                results "cf1:/script-revert-results.txt"
                script "vrrp-backup-script"
                max-completed 4
                lifetime forever
                no shutdown
            exit
        exit
```

The event-handler acts as the interface between the configured script-policy and event-trigger. Therefore, a second event-handler is configured with an action-list consisting of a single entry referencing the newly configured vrrp-backup-policy.

```
configure
    log
        event-handling
            handler "event-handler-2"
                action-list
                    entry 10
                        script-policy "vrrp-backup-policy"
                        no shutdown
                    exit
                exit
                no shutdown
            exit
        exit
```

Finally, the event-trigger is configured. To revert to VRRP Backup, the application is VRRP and the event number is 2006 (tmnxVrrpBecameBackup). The configuration is filtered on the message field, as before, using log-filter 2, so that it is specific to the interface named redundant-interface.

```
configure
    log
        filter 2
            default-action drop
            entry 10
                action forward
                match
                    message eq pattern "interface redundant-interface changed
                                       state to backup"
                exit
            exit
        exit


configure
    log
        event-trigger
            event "vrrp" 2006
```

```
                trigger-entry 1
                    event-handler "event-handler-2"
                    log-filter 2
                    no shutdown
                exit
                no shutdown
            exit
        exit
    exit
```

The configuration of the example event handling for the revertive failure event (PE-3 transitions to VRRP backup) is now complete. By re-enabling the spoke-SDP between PE-1 and PE-2, the VRRP message path is restored, and PE-2 again becomes the VRRP master. The following events are generated. The first indicates that PE-3 has become VRRP backup for the interface named redundant-interface. The second indicates that EHS handler event-handler-2 was invoked by a CLI user. The third indicates that a script file has initiated an attempt to execute CLI commands contained in script file vrrp-backup.txt. The fourth indicates that the attempt to execute those CLI commands was successful.

```
64 2017/10/26 13:37:26.113 CEST MINOR: VRRP #2006 Base Becoming Backup
"VRRP virtual router instance 1 on interface redundant-interface changed
state to backup - current master is 172.16.1.2"

65 2017/10/26 13:37:26.113 CEST MINOR: SYSTEM #2069 Base EHS script
"Ehs handler :"event-handler-2" with the description : "" was invoked
by the cli-user account "not-specified"."

66 2017/10/26 13:37:26.129 CEST MAJOR: SYSTEM #2052 Base CLI 'exec'
"A CLI user has initiated an 'exec' operation to process the commands
in the SROS CLI file cf3:/vrrp-backup.txt"

67 2017/10/26 13:37:26.144 CEST MAJOR: SYSTEM #2053 Base CLI 'exec'
"The CLI user initiated 'exec' operation to process the commands
in the SROS CLI file cf3:/vrrp-backup.txt has completed with the
result of success"
```

The following outputs confirm that PE-3 is VRRP backup, and that the local preference attribute for prefix 172.16.1.0/29 has changed to a value of 50. The result of this action is that PE-2 will now be the transit router for both upstream and downstream traffic.

```
*A:PE-3# show router vrrp instance

===============================================================================
VRRP Instances
===============================================================================
Interface Name                   VR Id Own Adm  State       Base Pri   Msg Int
                                 IP        Opr  Pol Id      InUse Pri  Inh Int
-------------------------------------------------------------------------------
redundant-interface              1     No  Up   Backup       253       1
                                 IPv4      Up   n/a         253        No
  Backup Addr: 172.16.1.1
-------------------------------------------------------------------------------
```

```
Instances : 1
===============================================================================
*A:PE-3#

*A:PE-3# show router bgp routes 172.16.1.0/29 hunt | match expression
                                              "Network|Nexthop|To|Local Pref"
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.2
Res. Nexthop   : 192.168.23.1
Local Pref.    : 100                    Interface Name : int-PE-3-PE-2
Network        : 172.16.1.0/29
Nexthop        : 192.0.2.3
To             : 192.0.2.6
Res. Nexthop   : n/a
Local Pref.    : 50                     Interface Name : NotAvailable
*A:PE-3#
```

Conclusion

EHS allows operators to configure user-defined actions on the router when an event occurs. The event trigger can be anything that is generated by the event-control framework, and explicit filtering is possible using regular expressions. A user-defined action typically runs a script that allows any CLI commands to be executed. Multiple actions are permitted, running multiple scripts if required.

SR OS NETCONF Server Basics

This chapter provides information about SR OS NETCONF Server Basics.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

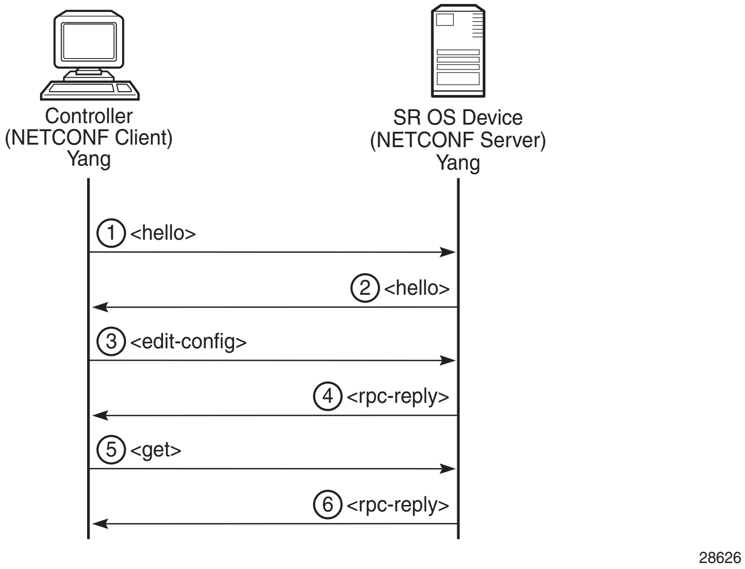
Applicability

The information and configuration in this chapter are based on SR OS release 16.0.R4.

Overview

The SR OS Network Configuration Protocol (NETCONF) server can communicate with a NETCONF client, that is, exchange hello messages, receive requests, and reply with responses. Before communicating with the SR OS NETCONF server, some SR OS configurations are prerequisites, and others are optional. This chapter describes the basic configurations needed for a seamless interaction with the SR OS NETCONF server. Figure 21 shows the NETCONF client-server communication between the controller and the SR OS node.

Figure 21 **NETCONF client-server communication**



Configuration

The following steps describe the procedure to configure a NETCONF server on SR OS.

- Because NETCONF uses SSH for transport, enable the SSH server in SR OS:

```
configure system security ssh no server-shutdown
```

- Enable the NETCONF server:

```
configure system netconf no shutdown
```

- Enable the YANG modules to use with NETCONF; for example, the Nokia modules:

```
configure
    system
        management-interface
            yang-modules
                no base-r13-modules
                no nokia-combined-modules
                nokia-modules
            exit
```

Note: The Nokia combined modules and the Nokia modules cannot both be set to true at the same time.

Note: Keep the base-r13 modules set to false.

- Configure an "nc_user" user with administrative privileges (**access netconf**):

```
configure
    system
        security
            user "nc-user"
                password <password>
                access netconf
                console
                    member "administrative"
                exit
            exit
```

- Optionally, enable NETCONF auto-config-save, which auto-saves the data (that is, makes it persistent) after each successful NETCONF commit:

```
configure system netconf auto-config-save
```

- Optionally, enable the NETCONF user to **lock** a datastore through NETCONF:

```
configure system security profile "administrative" netconf base-op-authorization lock
```

- Optionally, enable the NETCONF user to kill an open NETCONF session:

```
configure system security profile "administrative" netconf base-op-authorization kill-session
```

- Optionally, disable advertising the "writeable-running" capability in the NETCONF server hello message. With that, there will be no need to lock the writeable-running datastore by controllers:

```
configure system netconf capabilities no writable-running
```

- Save the configurations:

```
admin save
```

Conclusion

This chapter describes general SR OS NETCONF server configurations.

Interface Configuration

In This Section

This section provides interface configuration information for the following topics:

- Multi-Chassis APS and Pseudowire Redundancy Interworking
- Multi-Chassis LAG and Pseudowire Redundancy Interworking
- Port Cross-Connect (PXC)

Multi-Chassis APS and Pseudowire Redundancy Interworking

This chapter describes multi-chassis APS and pseudowire redundancy interworking.

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS release 6.0.R2, but the CLI in the current edition is based on SR OS release 14.0.R5.The configuration in this chapter includes the use of the ATM ports. See the Release Notes for information about support of ATM (and other) MDAs on various platforms as well as MC-APS restrictions.

Overview

MC-APS

MC-APS is an extension to the APS feature to provide not only link redundancy but also node level redundancy. It can protect against nodal failure by configuring the working circuit of an APS group on one node while configuring the protect circuit of the same APS group on a different node.

The two nodes connect to each other with an IP link that is used to establish a signaling path between them. The relevant APS groups in both the working and protection routers must have the same group ID and working circuit, and the protect circuit must have compatible configurations (such as the same speed, framing, and port-type). Signaling is provided using the direct connection between the two service routers. A heartbeat protocol can be used to add robustness to the interaction between the two routers.

Signaling functionality includes support for:

- APS group matching between service routers.
- Verification that one side is configured as a working circuit and the other side is configured as the protect circuit. In case of a mismatch, a trap (incompatible-neighbor) is generated.
- Change in working circuit status is sent from the working router to keep the protection router in sync.
- Protection router, based on K1/K2 byte data, member circuit status, and external request, selects the active circuit and informs the working router to activate or de-activate the working circuit.

Pseudowire Redundancy

Pseudowire (PW) redundancy provides the ability to protect a pseudowire with a pre-provisioned pseudowire and to switch traffic over to the secondary standby pseudowire in case of a SAP and/or network failure condition. Normally, pseudowires are redundant by the virtue of the SDP redundancy mechanism. For instance, if the SDP is an RSVP LSP and is protected by a secondary standby path and/or by Fast-Reroute paths, the pseudowire is also protected.

However, there are a few applications in which SDP redundancy does not protect the end-to-end pseudowire path when there are two different destination 7x50 PE nodes for the same VLL service. The main use case is the provisioning of dual-homing of a CPE or access node to two 7x50 PE nodes located in different POPs. The other use case is the provisioning of a pair of active and standby BRAS nodes, or active and standby links to the same BRAS node, to provide service resiliency to broadband service subscribers.

Example Topology

The setup in this section contains two access nodes and 4 PE nodes. The access nodes can be any ATM switches that support 1+1 bi-directional APS. Figure 22 shows the physical topology of the setup. Figure 24 shows the use of both MC-APS in the access network and pseudowire redundancy in the core network to provide a resilient end-to-end VLL service.

Figure 22 MC-APS Network Topology

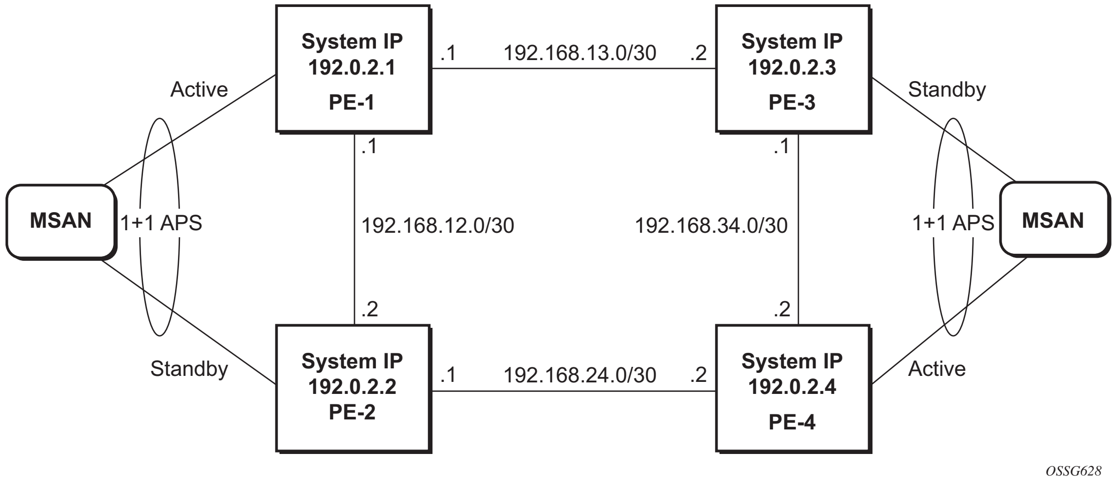


Figure 23 Access Node and Network Resilience (Part 1)

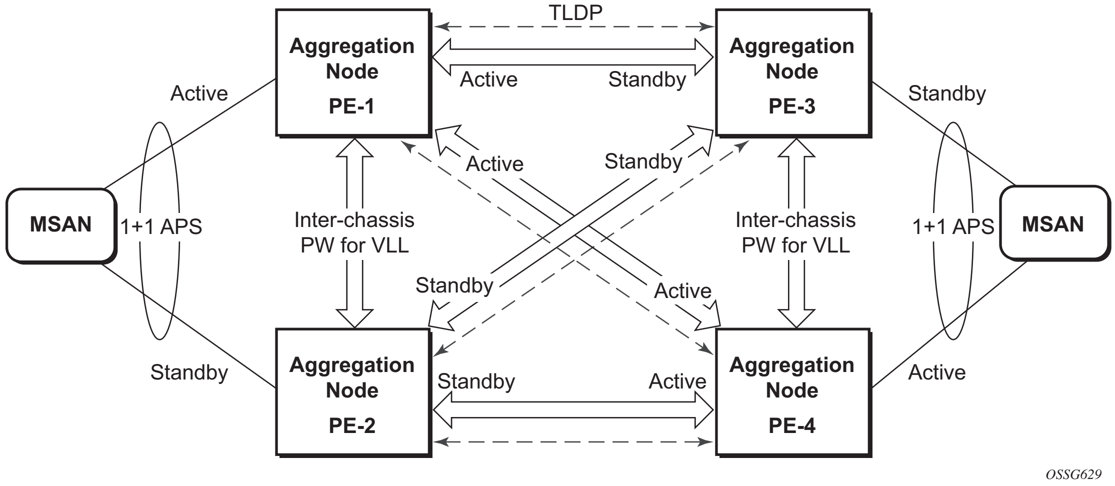
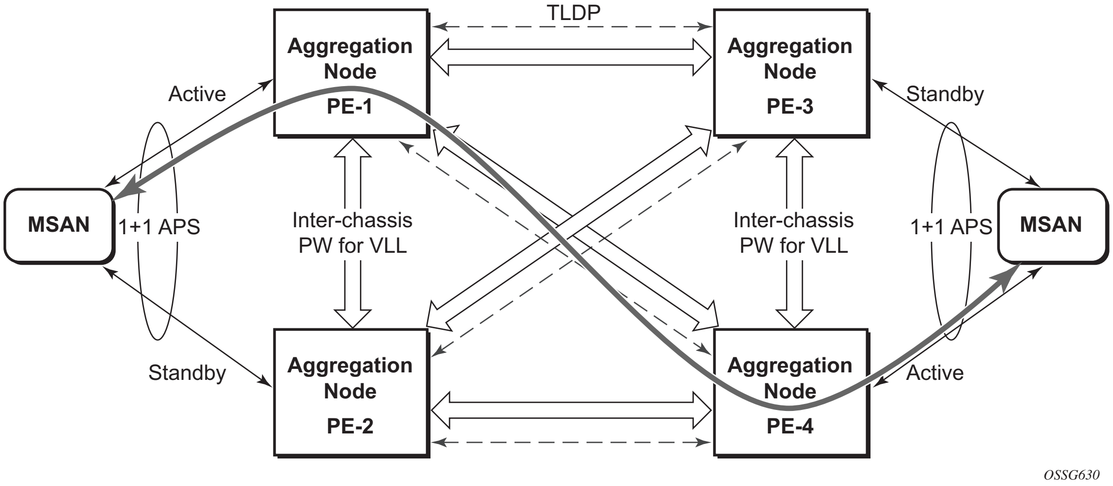


Figure 24 **Access Node and Network Resilience (Part 2)**



Configuration

The following configuration should be completed on the PEs before configuring MC-APS:

- Cards, MDAs and ports
- Interfaces
- IGP configured and converged
- MPLS
- SDPs configured between all PE routers

For the IGP, OSPF or IS-IS can be used. MPLS or GRE can be used for the transport tunnels. For MPLS, LDP or RSVP protocols can be used for signaling MPLS labels. In this example, OSPF and LDP are used. The following commands can be used to check if OSPF has converged and to make sure the SDPs are up (for example, on PE-1):

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h02m12s  0
       system                                                       0
192.0.2.2/32                                  Remote  OSPF      00h01m17s  10
```

```
       192.168.12.2                                                 10
192.0.2.3/32                                  Remote  OSPF      00h01m05s  10
       192.168.13.2                                                 10
192.0.2.4/32                                  Remote  OSPF      00h01m08s  10
       192.168.12.2                                                 20
192.168.12.0/30                               Local   Local     00h02m13s  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     00h02m12s  0
       int-PE-1-PE-3                                                0
192.168.24.0/30                               Remote  OSPF      00h01m17s  10
       192.168.12.2                                                 20
192.168.34.0/30                               Remote  OSPF      00h01m05s  10
       192.168.13.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 8
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show service sdp
============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
12     0       1556    192.0.2.2        Up   Up          MPLS    L     TLDP
13     0       1556    192.0.2.3        Up   Up          MPLS    L     TLDP
14     0       1556    192.0.2.4        Up   Up          MPLS    L     TLDP
----------------------------------------------------------------------------
Number of SDPs : 3
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-1#
```

Step 1. APS configuration on MSANs

The access nodes can be any ATM switches that support 1+1 bi-directional APS. Here is an example on 7670RSP (Routing Switching Platform).

```
RSP> configure
RSP> port 1-6-1-1
RSP> options protection type 1+1
RSP> options protection switching bidirect
RSP> options protection
                                                                    (Standby)
#                  Type        Status           Name
1-6-1-1            STM1_IR8    OK
Protection Group Contains:
  Protection Port     : 1-6-1-1     (Standby)
  Working Port        : 1-5-1-1
Protection Type       : 1+1
Switching Type        : Non-Revertive
Switching Mode        : Bi-directional
```

```
Wait-To-Restore Timer :  5 minute(s)
```

Step 2. MC-APS configuration on PE-1 and PE-2

Assuming the link between MSAN and PE-1 is working circuit and the link between MSAN and PE-2 is protection circuit.

Configure APS on the PE-1 port. Specify the system IP address of neighbor node (PE-2) and working-circuit.

```
*A:PE-1# configure
    port 1/2/1
        sonet-sdh
        exit
        no shutdown


*A:PE-1# configure
    port aps-1
        aps
            neighbor 192.0.2.2
            working-circuit 1/2/1
        exit
        sonet-sdh
            path
                atm
                exit
                no shutdown
            exit
        exit
        no shutdown
```

Configure APS on the PE-2 port. Specify the system IP address of neighbor node (PE-1) and protect-circuit instead of working-circuit.

```
*A:PE-2# configure
    port 1/2/1
        sonet-sdh
        exit
        no shutdown


*A:PE-2# configure port aps-1
        aps
            neighbor 192.0.2.1
            protect-circuit 1/2/1
        exit
        sonet-sdh
            path
                atm
                exit
                no shutdown
            exit
        exit
        no shutdown
```

The following parameters can be configured under APS optionally.

- advertise-interval — This command specifies the time interval, in 100s of milliseconds, between 'I am operational' messages sent by both protect and working circuits to their neighbor for multi-chassis APS.
- hold-time — This command specifies how much time can pass, in 100s of milliseconds, without receiving an advertise packet from the neighbor before the multi-chassis signaling link is considered not operational.
- revert-time — This command configures the revert-time timer to determine how long to wait before switching back to the working circuit after that circuit has been restored into service.
- switching-mode — This command configures the switching mode for the APS port which can be bi-directional or uni-directional.

Step 3. Verify the APS status on PE-1.

```
*A:PE-1# show port aps-1

===============================================================================
SONET/SDH Interface
===============================================================================
Description        : APS Group
Interface          : aps-1                  Speed                : oc3
Admin Status       : up                     Oper Status          : up
Physical Link      : Yes                    Loopback Mode        : none
Single Fiber Mode  : No
Clock Source       : node                   Framing              : sonet
Last State Change  : 10/25/2016 13:45:58    Port IfIndex         : 1358987264
Configured Address : 02:15:ff:00:02:49
Hardware Address   : 02:15:ff:00:02:49
Last Cleared Time  : N/A
J0 String          : 0x01                   Section Trace Mode   : byte
Rx S1 Byte         : 0x00 (stu)             Rx K1/K2 Byte        : 0x00/0x00
Tx S1 Byte         : 0x0a (st3)             Tx DUS/DNU           : Disabled
Rx J0 String (Hex) : 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
Cfg Alarm          : loc lrdi lb2er-sf slof slos
Alarm Status       :
BER SD Threshold   : 6                      BER SF Threshold     : 3
Hold time up       : 500 milliseconds       Reset On Path Down   : Disabled
Hold time down     : 200 milliseconds

Transceiver Data

Transceiver Status : not-equipped

===============================================================================
Port Statistics
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
Packets                                                0                      0
Discards                                               0                      0
Unknown Proto Discards                                 0
===============================================================================
*A:PE-1#
```

Step 4. Verify the MC-APS status and parameters on PE-1 and PE-2

Detailed parameters of the APS configuration on PE-1 can be verified, as follows. The admin/oper status of APS group 1 shows up/up. K1/K2 byte shows N/A as APS 1+1 exchanges that information through the protection circuit.

The admin/oper status of the working circuit (the link between MSAN and PE-1) is up/up.

```
*A:PE-1# show aps detail

===============================================================================
APS Group: aps-1
===============================================================================
Description        : APS Group
Group Id           : 1                    Active Circuit       : 1/2/1
Admin Status       : Up                   Oper Status          : Up
Working Circuit    : 1/2/1                Protection Circuit   : N/A
Switching-mode     : Bi-directional       Switching-arch       : 1+1(sig-only)
Annex B            : No
Revertive-mode     : Non-revertive        Revert-time (min)    :
Rx K1/K2 byte      : N/A
Tx K1/K2 byte      : N/A
Current APS Status : OK
Multi-Chassis APS  : Yes
Neighbor           : 192.0.2.2
Control link state : Up
Advertise Interval : 1000 msec            Hold Time            : 3000 msec
Mode mismatch Cnt  : 0                    Channel mismatch Cnt : 0
PSB failure Cnt    : 0                    FEPL failure Cnt     : 0
-------------------------------------------------------------------------------
 APS Working Circuit - 1/2/1
-------------------------------------------------------------------------------
Admin Status       : Up                   Oper Status          : Up
Current APS Status : OK                   No. of Switchovers   : 0
Last Switchover    : None                 Switchover seconds   : 0
Signal Degrade Cnt : 0                    Signal Failure Cnt   : 0
Last Switch Cmd    : N/A                  Last Exercise Result : N/A
Tx L-AIS           : None
===============================================================================
*A:PE-1#
```

Detailed parameters of the APS configuration on PE-2 can be verified, as follows. The admin/oper status of APS group 1 shows up/up. Both Rx and Tx of the K1/K2 byte are in the status of 0x00/0x05 (No-Req on Protect) as there is no failure or force-switchover request.

The admin/oper status of the protection circuit (the link between MSAN and PE-2) is up/up.

```
*A:PE-2# show aps detail
===============================================================================
APS Group: aps-1
===============================================================================
Description        : APS Group
Group Id           : 1                    Active Circuit       : N/A
Admin Status       : Up                   Oper Status          : Up
```

```
Working Circuit    : N/A                 Protection Circuit   : 1/2/1
Switching-mode     : Bi-directional      Switching-arch       : 1+1(sig-only)
Annex B            : No
Revertive-mode     : Non-revertive       Revert-time (min)    :
Rx K1/K2 byte      : 0x00/0x05 (No-Req on Protect)
Tx K1/K2 byte      : 0x00/0x05 (No-Req on Protect)
Current APS Status : OK
Multi-Chassis APS  : Yes
Neighbor           : 192.0.2.1
Control link state : Up
Advertise Interval : 1000 msec           Hold Time            : 3000 msec
Mode mismatch Cnt  : 0                   Channel mismatch Cnt : 0
PSB failure Cnt    : 0                   FEPL failure Cnt     : 1
-------------------------------------------------------------------------------
 APS Working Circuit - Neighbor
-------------------------------------------------------------------------------
Admin Status       : N/A                 Oper Status          : N/A
Current APS Status : OK                  No. of Switchovers   : 0
Last Switchover    : None                Switchover seconds   : 0
Signal Degrade Cnt : 0                   Signal Failure Cnt   : 1
Last Switch Cmd    : No Cmd              Last Exercise Result : Unknown
Tx L-AIS           : None
-------------------------------------------------------------------------------
 APS Protection Circuit - 1/2/1
-------------------------------------------------------------------------------
Admin Status       : Up                  Oper Status          : Up
Current APS Status : OK                  No. of Switchovers   : 0
Last Switchover    : None                Switchover seconds   : 0
Signal Degrade Cnt : 0                   Signal Failure Cnt   : 0
Last Switch Cmd    : No Cmd              Last Exercise Result : Unknown
Tx L-AIS           : None
===============================================================================
*A:PE-2#
```

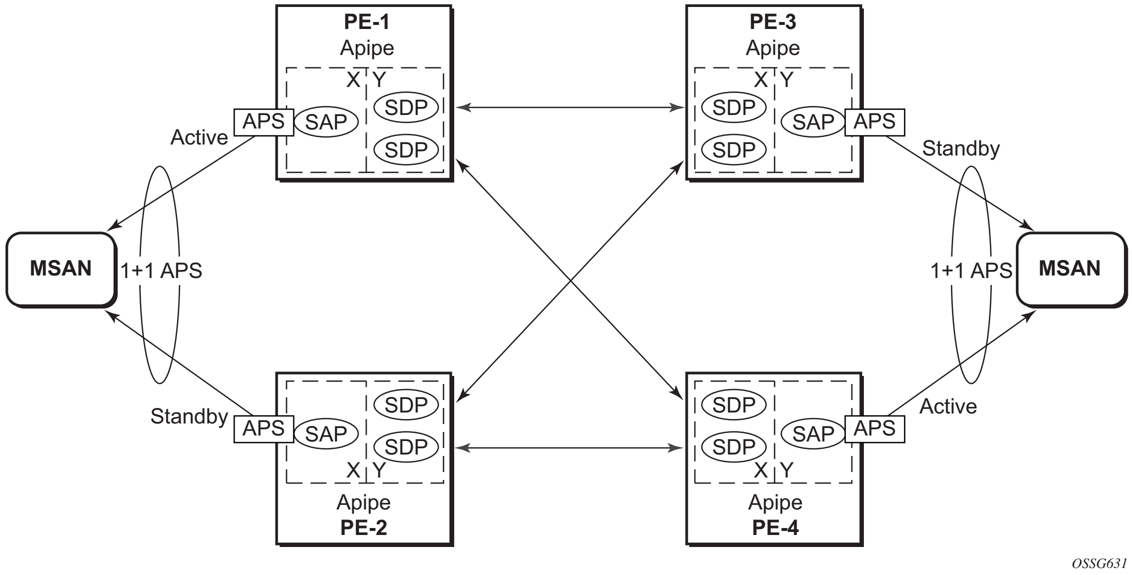
Step 5. MC-APS configuration on PE-3 and PE-4

The MC-APS configuration on PE-3 and PE-4 is similar to the configuration on PE-1 and PE-2. Configure the working circuit on PE-4 and the protection circuit on PE-3.

Step 6. Pseudowire configuration

Configure an Apipe service on every PE and create endpoints X and Y. Associate the SAPs and spoke SDPs with the endpoints, as shown in Figure 25.

Figure 25 **Association of SAPs/SDPs and Endpoints**



```
*A:PE-1# configure
    service
        apipe 1 customer 1 create
            endpoint "X" create
            exit
            endpoint "Y" create
            exit
            sap aps-1:0/32 endpoint "X" create
            exit
            spoke-sdp 13:1 endpoint "Y" create
            exit
            spoke-sdp 14:1 endpoint "Y" create
            exit
            no shutdown
        exit
```

Syntax aps-1:0/32 specifies the APS group and VPI/VCI of the ATM circuit (aps-id:vpi/vci).

Likewise, an Apipe service, with endpoints, SAPs and spoke SDPs must be configured on the other PE routers.

Step 7. Pseudowire verification

The Apipe service is up in PE-1 (MC-APS working circuit), as follows:

```
*A:PE-1# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Apipe     Up   Up   1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
```

```
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The Apipe service is down in PE-2 (MC-APS protect circuit), as follows:

```
*A:PE-2# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Apipe     Up   Down 1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

The Apipe service is down in PE-3 (MC-APS protect circuit), as follows:

```
*A:PE-3# show service service-using
==============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Apipe     Up   Down 1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

The Apipe service is up in PE-4 (MC-APS working circuit), as follows:

```
*A:PE-4# show service service-using
===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Apipe     Up   Up   1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-4#
```

Note: After configuring ICB spoke-SDPs, the Apipe will be up on all PEs.

Step 8. Verify SDP status

The status of SDP 23:1 on PE-2 can be verified as follows.

Peer Pw Bits shows the status of the pseudowire on the peer node. In this example, both the local node (PE-2) as the remote node (PE-3) are sending the **lacIngressFault lacEgressFault** and **pwFwdingStandby** flags. This is because the Apipe service on these nodes is down because the MC-APS is in protection status.

```
*A:PE-2# show service id 1 sdp 23:1 detail

===============================================================================
Service Destination Point (Sdp Id : 23:1) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 23:1  -(192.0.2.3)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 23:1                     Type              : Spoke
Spoke Descr     : (Not Specified)
Split Horiz Grp    : (Not Specified)
VC Type            : AAL5SDU                  VC Tag            : 0
Admin Path MTU     : 0                        Oper Path MTU     : 1556
Delivery           : MPLS
Far End            : 192.0.2.3
Tunnel Far End     : 192.0.2.3                LSP Types         : LDP

Admin State        : Up                       Oper State        : Up
Acct. Pol          : None                     Collect Stats     : Disabled
Ingress Label      : 262139                   Egress Label      : 262138
Ingr Mac Fltr-Id   : n/a                      Egr Mac Fltr-Id   : n/a
Ingr IP Fltr-Id    : n/a                      Egr IP Fltr-Id    : n/a
Admin ControlWord  : Preferred                Oper ControlWord  : True
Admin BW(Kbps)     : 0                        Oper BW(Kbps)     : 0
BFD Template       : None
BFD-Enabled        : no                       BFD-Encap         : ipv4
Last Status Change : 10/25/2016 13:48:30      Signaling         : TLDP
Last Mgmt Change   : 10/25/2016 13:48:22
Endpoint           : Y                        Precedence        : 4
PW Status Sig      : Enabled
Class Fwding State : Down
Flags              : None
Local Pw Bits      : lacIngressFault lacEgressFault pwFwdingStandby
Peer Pw Bits       : lacIngressFault lacEgressFault pwFwdingStandby
Peer Fault Ip      : None
Peer Vccv CV Bits  : lspPing bfdFaultDet
Peer Vccv CC Bits  : pwe3ControlWord mplsRouterAlertLabel

Ingress Qos Policy : (none)                   Egress Qos Policy : (none)
Ingress FP QGrp    : (none)                   Egress Port QGrp  : (none)
Ing FP QGrp Inst   : (none)                   Egr Port QGrp Inst: (none)

KeepAlive Information :
Admin State        : Disabled                 Oper State        : Disabled
Hello Time         : 10                       Hello Msg Len     : 0
Max Drop Count     : 3                        Hold Down Time    : 10
```

```
---snip---
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

In case of failure, the access link can be protected by MC-APS. An MPLS network failure can be protected by pseudowire redundancy. Node failure can be protected by the combination of MC-APS and pseudowire redundancy.

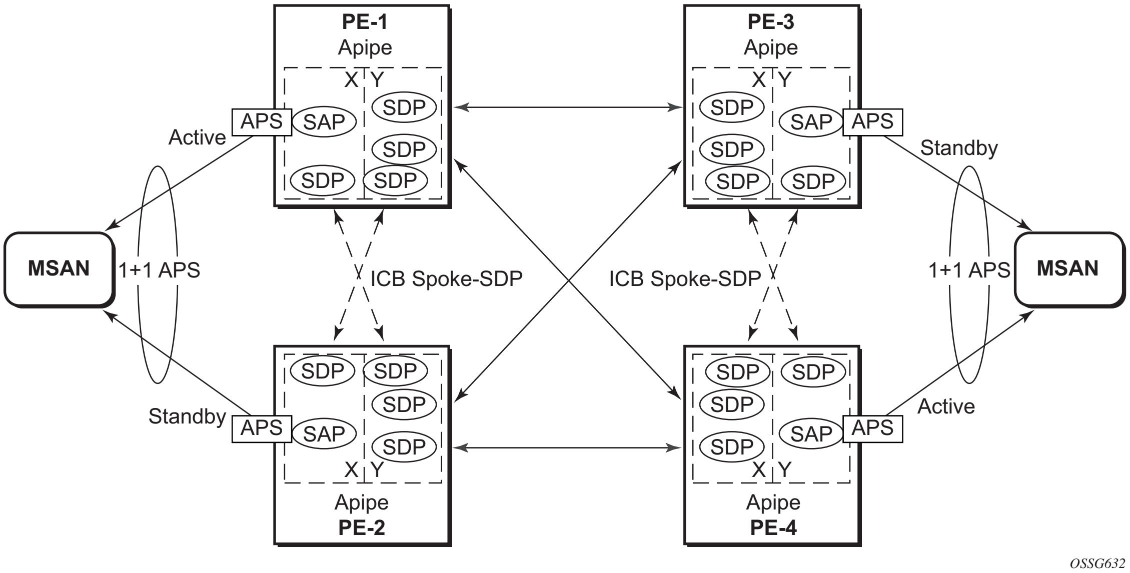
Step 9. Inter-Chassis Backup (ICB) pseudowire configuration.

Configuring Inter-Chassis Backup (ICB) is optional. It can reduce traffic impact by forwarding traffic on ICB spoke SDPs during MC-APS switchover. The ICB spoke SDP cannot be added to the endpoint if the SAP is not part of an MC-APS (or MC-LAG) instance. Conversely, a SAP which is not part of a MC-APS (or MC-LAG) instance cannot be added to an endpoint which already has an ICB spoke SDP. Forwarding between ICBs is blocked on the same node. The user has to explicitly indicate the spoke SDP is actually an ICB at creation time. Figure 5 shows some setup examples where ICBs are required.

After configuring ICB spoke SDPs, the Apipe will be in admin/oper up/up status on all PE routers.

ICB SDPs are configured and associated to endpoints, as shown in Figure 26.

Figure 26 **ICB Spoke SDPs and Association with the Endpoints**



Two ICB spoke SDPs must be configured in the Apipe service on each PE router, one in each endpoint. The same SDP IDs can be used for the ICBs since the far-end will be the same. However, the vc-id must be different. The ICB spoke SDPs must cross, meaning one end should be associated with endpoint X and the other end (on the other PE) should be associated with endpoint Y.

An ICB is always the last forwarding resort. Only one spoke SDP will be forwarding. If there is an ICB and an MC-APS SAP in an endpoint, the ICB will only forward if the SAP goes down. If an ICB resides in an endpoint together with other spoke SDPs the ICB will only forward if there is no other active spoke SDP.

The following shows the additional configuration for ICB on each PE:

```
*A:PE-1# configure
    service
        apipe 1
            spoke-sdp 12:1 endpoint "X" icb create
            exit
            spoke-sdp 12:2 endpoint "Y" icb create
            exit


*A:PE-2# configure
    service
        apipe 1
            spoke-sdp 21:1 endpoint "Y" icb create
            exit
            spoke-sdp 21:2 endpoint "X" icb create
            exit


*A:PE-3# configure
    service
        apipe 1
            spoke-sdp 34:1 endpoint "X" icb create
            exit
            spoke-sdp 34:2 endpoint "Y" icb create
            exit


*A:PE-4# configure
    service
        apipe 1
            spoke-sdp 43:1 endpoint "Y" icb create
            exit
            spoke-sdp 43:2 endpoint "X" icb create
            exit
```

Step 10. Verification of active objects for each endpoint

The following command shows which objects are configured for each endpoint and which is the active object at this moment:

```
*A:PE-1# show service id 1 endpoint

===============================================================================
Service 1 endpoints
===============================================================================
Endpoint name                : X
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Tx Active                    : aps-1:0/32
Tx Active Up Time            : 0d 00:02:58
Revert Time Count Down       : N/A
Tx Active Change Count       : 1
Last Tx Active Change        : 10/25/2016 13:48:15
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
SAP      : aps-1:0/32                               Oper Status: Up
Spoke-sdp: 12:1 Prec:4 (icb)                        Oper Status: Up
===============================================================================
Endpoint name                : Y
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Tx Active (SDP)              : 14:1
Tx Active Up Time            : 0d 00:02:35
Revert Time Count Down       : N/A
Tx Active Change Count       : 2
Last Tx Active Change        : 10/25/2016 13:48:38
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 12:2 Prec:4 (icb)                        Oper Status: Up
Spoke-sdp: 13:1 Prec:4                              Oper Status: Up
Spoke-sdp: 14:1 Prec:4                              Oper Status: Up
===============================================================================
===============================================================================
*A:PE-1#
```

On PE-1, both the SAP and the spoke SDP 14:1 are active. The other objects do not forward traffic.

Step 11. Other types of setups

The following figures show other setups that combine MC-APS and pseudowire redundancy.

Figure 27 Additional Setup Example 1 (Part 1)

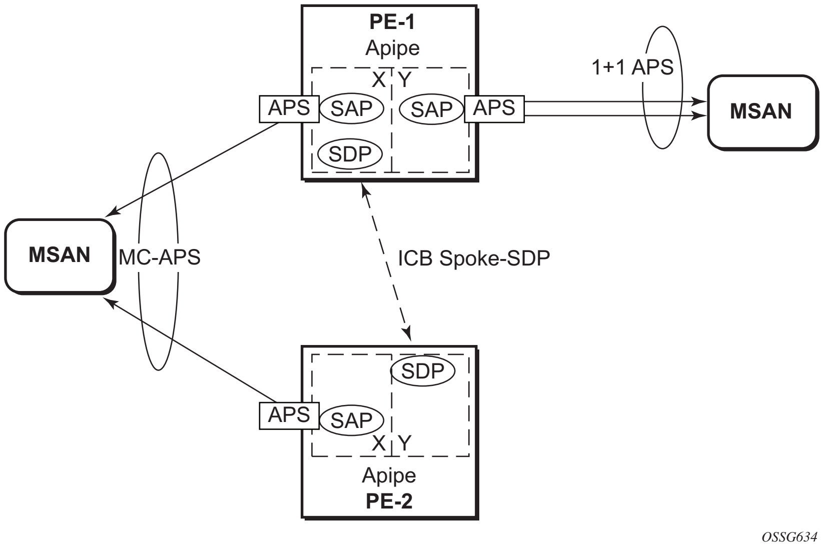


Figure 28 Additional Setup Example 1 (Part 2)

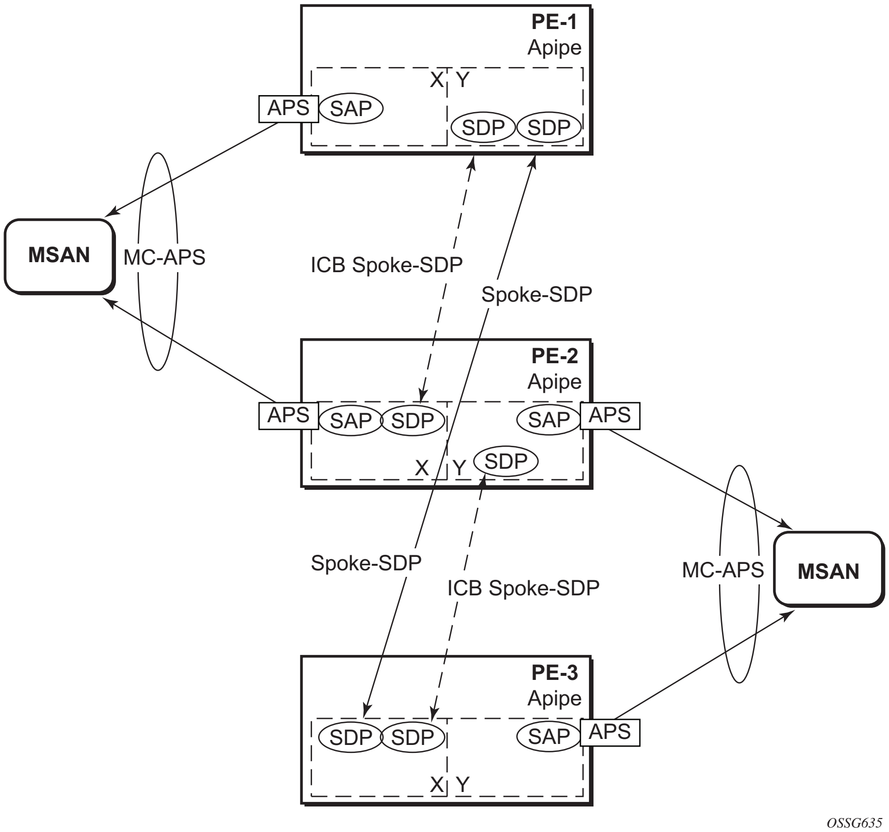


Figure 29 Additional Setup Example 2 (Part 1a)

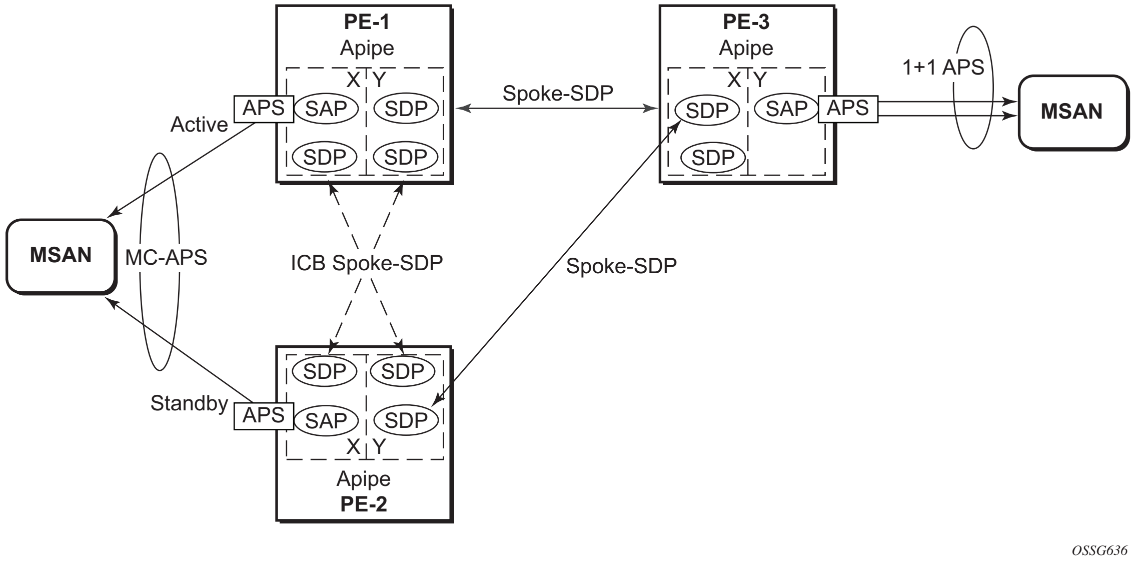


Figure 30 Additional Setup Example 2 (Part 1b)

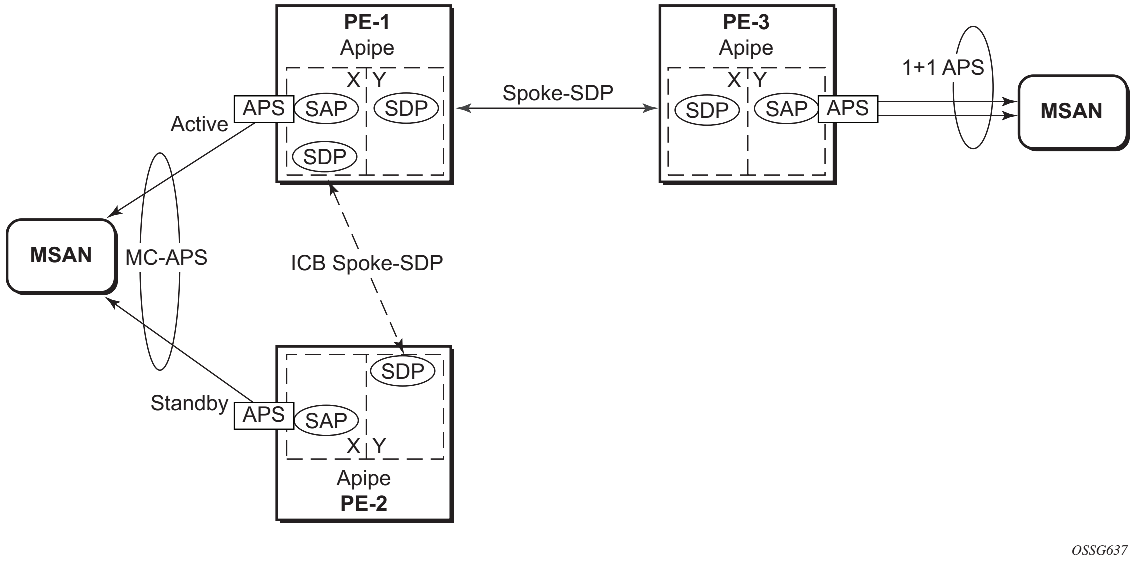
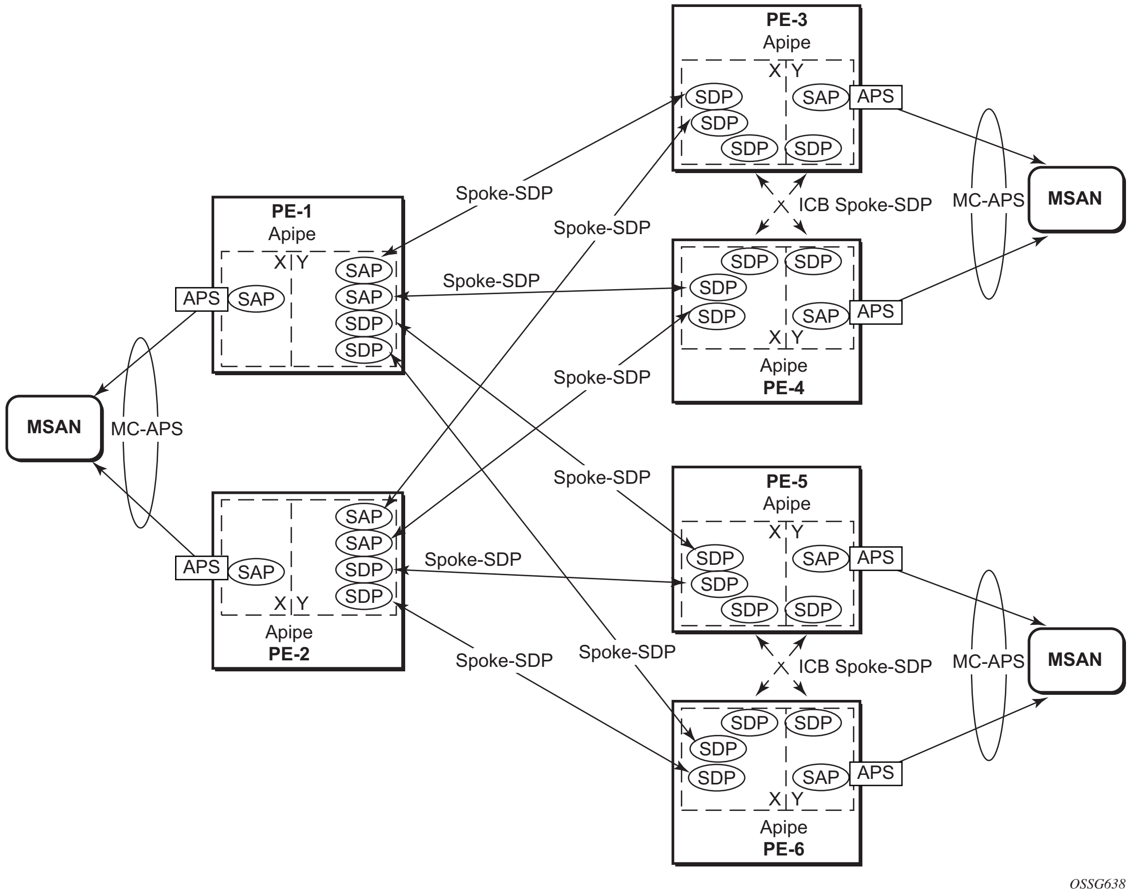


Figure 31 **Additional Setup Example 2 (Part 2)**



Forced Switchover

MC-APS convergence can be forced with the **tools perform aps** command:

```
*A:PE-1# tools perform aps force
  - force <aps-id> {protect|working} [number <number>]

 <aps-id>              : aps-<group-id>
                           aps            - keyword
                           group-id       - [1..128]
 <protect|working>     : keyword
 <number>              : [1-2]
```

After the forced switchover, it is important to clear the forced switchover:

```
*A:PE-1# tools perform aps clear
  - clear <aps-id> {protect|working} [number <number>]
```

```
<aps-id>              : aps-<group-id>
                          aps           - keyword
                          group-id      - [1..128]
<protect|working>     : protect|working
<number>              : [1-2]
```

Conclusion

In addition to Multi-Chassis LAG, Multi-Chassis APS provides a solution for both network redundancy and access node redundancy. It supports ATM VLL and Ethernet VLL with ATM SAP. Access links and PE nodes are protected by APS and the MPLS network is protected by pseudowire redundancy/FRR. With this feature, Nokia can provide resilient end-to-end solutions.

Multi-Chassis LAG and Pseudowire Redundancy Interworking

This chapter provides information about Multi-Chassis Link Aggregation (MC-LAG) and pseudowire redundancy interworking.

Topics in this chapter include:

Applicability

MC-LAG is supported only on Ethernet MDAs, and this only for access ports, because the LAG group must be in access mode.

This chapter was initially written for SR OS release 7.0.R5. However, the CLI in the current edition is based on SR OS release 16.0.R1.

Overview

MC-LAG

MC-LAG is an extension to the LAG feature to provide not only link redundancy but also node-level redundancy. This feature provides a Nokia added value solution which is not defined in any IEEE standard.

A proprietary messaging system between redundant-pair nodes supports coordinating the LAG switchover.

Multi-chassis LAG supports LAG switchover coordination: one node connected to two redundant-pair peer nodes with the LAG. During the LACP negotiation, the redundant-pair peer nodes act like a single node using active/stand-by signaling to ensure that only links of one peer node are used at a time.

Pseudowire Redundancy

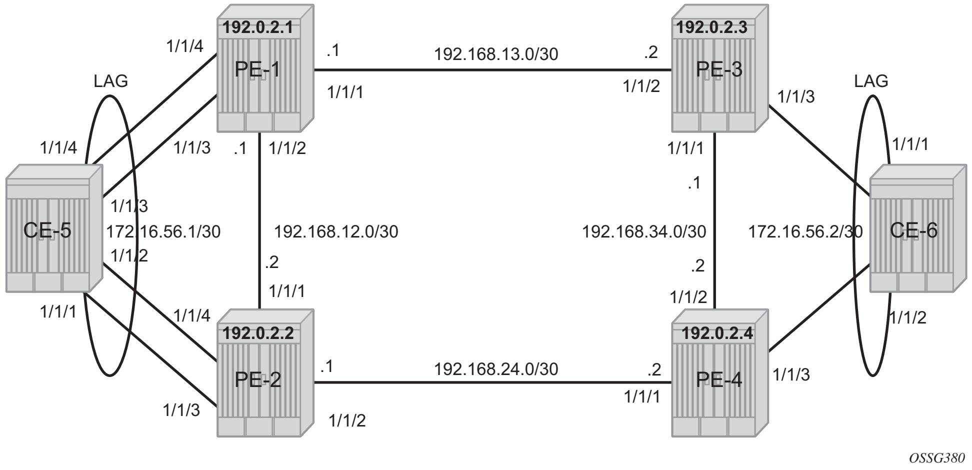
Pseudowire (PW) redundancy provides the ability to protect a pseudowire with a secondary pre-provisioned pseudowire and to switch traffic over to the secondary standby pseudowire in case of a SAP and/or network failure condition. Normally, pseudowires are redundant by the virtue of the SDP redundancy mechanism. For instance, if the SDP relies on an RSVP LSP that is protected by a secondary standby path and/or by Fast-Reroute paths, the pseudowire is also protected.

However, there are a few applications in which SDP redundancy does not protect the end-to-end pseudowire path, for example when there are two different destination 7x50 PE nodes for the same VLL service.

The main use case for PW redundancy is a scenario where dual homed CPEs or access nodes connected to two 7x50 PE nodes are located in different POPs. The other use case is the scenario where service resiliency for broadband service subscribers is required, for example when a pair of active and standby BRAS nodes are provisioned, or where active and standby links to the same BRAS node are provisioned.

Example Topology

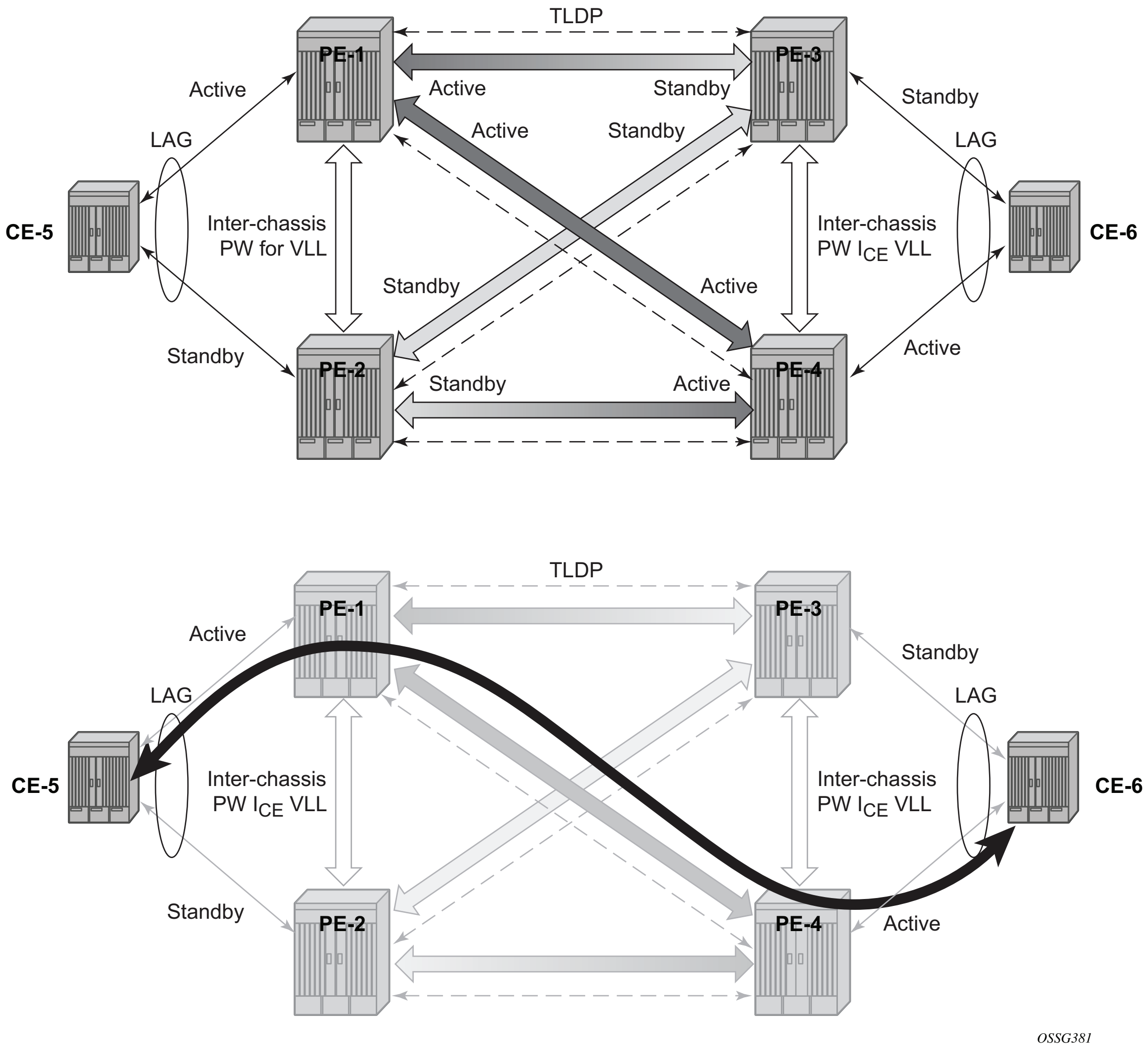
Figure 32 **MC-LAG Example Topology**



This section describes a setup which contains two CEs and four PEs. The CEs can be any routing/switching device that support the OUT_OF_SYNC signaling as described in IEEE Standard 802.3-2005 section 3 section 43.6.1. Figure 32 shows the physical topology of the setup.

Figure 33 shows the use of both MC-LAG in the access network and pseudowire redundancy in the core network to provide a resilient end-to-end VLL service between CE-5 and CE-6.

Figure 33 **Network Resiliency**



When an SDP is in standby, it sends the pseudowire status bit pwFwdingStandby to its peer.

Configuration

It is assumed that the following base configuration has been implemented on the PEs:

- Cards, MDAs and ports
- Interfaces
- IGP configured and converged
- MPLS
- SDPs configured between all PE routers

Either OSPF or IS-IS can be used as the IGP. Both LDP or RSVP can be used for signaling the transport MPLS labels. Alternatively, GRE can be used for the transport tunnels. If the SDPs are using LDP, RSVP, or GRE is irrelevant. In this example, OSPF and LDP are used.

The following command is used to check if OSPF has converged (for example, on PE-1):

```
*A:PE-1# show router route-table

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Local   Local     00h16m53s  0
       system                                                        0
192.0.2.2/32                                   Remote  OSPF      00h14m26s  10
       192.168.12.2                                                  100
192.0.2.3/32                                   Remote  OSPF      00h14m21s  10
       192.168.13.2                                                  100
192.0.2.4/32                                   Remote  OSPF      00h14m21s  10
       192.168.12.2                                                  200
192.168.12.0/30                                Local   Local     00h16m53s  0
       int-PE-1-PE-2                                                 0
192.168.13.0/30                                Local   Local     00h16m53s  0
       int-PE-1-PE-3                                                 0
192.168.24.0/30                                Remote  OSPF      00h14m26s  10
       192.168.12.2                                                  200
192.168.34.0/30                                Remote  OSPF      00h14m21s  10
       192.168.13.2                                                  200
--------------------------------------------------------------------------------
No. of Routes: 8
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:PE-1#
```

The following command shows that the SDPs are up:

```
*A:PE-1# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
12     0       1556    192.0.2.2        Up   Up          MPLS    L     TLDP
13     0       1556    192.0.2.3        Up   Up          MPLS    L     TLDP
14     0       1556    192.0.2.4        Up   Up          MPLS    L     TLDP
----------------------------------------------------------------------------
Number of SDPs : 3
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-1#
```

MC-LAG for Epipe Services

Step 1 - MC-LAG configuration on CEs.

The LAG configuration on the CEs is only included for completeness; any CE device could be used.

Auto-negotiation must be switched off or set to limited on all ports that will be included into the LAG in order to guarantee a specific port speed.

Note: Disabling autonegotiation on Gigabit ports is not allowed because the IEEE 802.3 specification for Gigabit Ethernet requires autonegotiation to be enabled for far end fault detection.

Configure LACP on the LAG with at least one side of the LAG in **active** mode.

```
*A:CE-5# configure port 1/1/[1..4] ethernet autonegotiate limited
*A:CE-5# configure port 1/1/[1..4] no shutdown

# on CE-5
configure
    lag 1
        port 1/1/1 1/1/2 1/1/3 1/1/4
        lacp active
        no shutdown
    exit
exit
```

Step 2 - LAG configuration on PEs.

The PE ports connected to the CEs must be configured as access ports because they will be used in the redundant pseudowire service. The LAG must also be configured in access mode.

The LAG encapsulation type (null | dot1q | qinq) must match the port encapsulation type of the LAG members.

Auto-negotiation must be switched off or configured to limited.

Configure LACP on the LAG. At least 1 side of the LAG (PE or CE) must be configured in **active** mode.

```
*A:PE-1# configure port 1/1/[3..4] ethernet mode access
*A:PE-1# configure port 1/1/[3..4] ethernet autonegotiate limited
*A:PE-1# configure port 1/1/[3..4] no shutdown
```

```
# on PE-1
configure
    lag 1
        mode access
        port 1/1/3 1/1/4
        lacp active
        no shutdown
    exit
exit
```

Step 3 - MC-LAG configuration on PE-1 and PE-2

The redundant PEs must act as one virtual node toward the CE. They have to be able to communicate the same LACP parameters to the CE side.

The following parameters uniquely identify a LAG instance:

- lacp-key
- system-id
- system-priority

These three parameters must be configured with the same value on both redundant PEs.

Multi-chassis redundancy requires a peering session (which operates by an IP connection using UDP destination port 1025) that is configured toward the redundant PE system address to which MC-LAG redundancy is enabled, as follows. The peering session can be configured with MD5 authentication.

```
# on PE-1
configure
    redundancy
        multi-chassis
            peer 192.0.2.2 create
                authentication-key "441dO/0RgDhHgzYwpOCTK9zbKjv4GZ/z" hash2
                mc-lag
                    lag 1 lacp-key 1 system-id 00:00:00:00:00:01 system-priority 100
                    no shutdown
                exit
                no shutdown
            exit

# on PE-2
configure
    redundancy
        multi-chassis
            peer 192.0.2.1 create
                authentication-key "441dO/0RgDg2CA0JlyzVNQBoRc327b1j" hash2
                mc-lag
                    lag 1 lacp-key 1 system-id 00:00:00:00:00:01 system-priority 100
                    no shutdown
                exit
                no shutdown
            exit
```

Step 4 - MC-LAG verification.

Verify MC peers showing that the authentication and admin state are enabled.

```
*A:PE-1# show redundancy multi-chassis sync

===============================================================================
Multi-chassis Peer Table
===============================================================================
Peer
-------------------------------------------------------------------------------
Peer IP Address         : 192.0.2.2
Description             : (Not Specified)
Authentication          : Enabled
Source IP Address       : 192.0.2.1
Admin State             : Enabled
Warm standby            : No
Remote warm standby     : No
-------------------------------------------------------------------------------
Sync: Not-configured
-------------------------------------------------------------------------------
===============================================================================
===============================================================================
*A:PE-1#
```

Step 5 - Verify MC-LAG peer status and LAG parameters.

```
*A:PE-1# show redundancy multi-chassis mc-lag peer 192.0.2.2

===============================================================================
Multi-Chassis MC-Lag Peer 192.0.2.2
===============================================================================
Last State chg  : 07/11/2018 12:33:32
Admin State     : Up                  Oper State           : Up
KeepAlive       : 10 deci-seconds     Hold On Ngbr Failure : 3
-------------------------------------------------------------------------------
Lag Id Lacp   Remote Source Oper   System Id          Sys   Last State Changed
       Key    Lag Id MacLSB MacLSB                    Prio
-------------------------------------------------------------------------------
1      1      1      Def    n/a    00:00:00:00:00:01  100   07/11/2018 12:33:33
-------------------------------------------------------------------------------
Number of LAGs : 1
===============================================================================
*A:PE-1#
```

There is a fixed keepalive timer of 1 second. The **hold-on-neighbor-failure multiplier** command indicates the interval that the standby node will wait for packets from the active node before assuming a redundant-neighbor failure. The **hold-on-neighbor-failure multiplier** command is configurable in the **config>redundancy>multi-chassis>peer>mc-lag** context. The standby node will also assume a redundant-neighbor failure when there is no route available to the redundant-neighbor.

```
# on PE-1
configure
    redundancy
        multi-chassis
            peer 192.0.2.2
                mc-lag
                    hold-on-neighbor-failure 10
                exit
            exit
        exit
    exit
```

In this example, the *lag-id* is 1 on both redundant PEs. This is not mandatory. If the *lag-id* on PE-2 is, for example 2, the following should be configured on PE-1:

```
# on PE-1
configure
    redundancy
        multi-chassis
            peer 192.0.2.2
                mc-lag
                    lag 1 remote-lag 2 lacp-key 1 system-id 00:00:00:00:00:01
                exit
            exit
        exit
    exit
```

Step 6 - Verify MC-LAG

```
*A:PE-1# show lag 1
===============================================================================
Lag Data
===============================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
-------------------------------------------------------------------------------
1              up      up      No       0         2        active
===============================================================================
*A:PE-1#
```

```
*A:PE-2# show lag 1
===============================================================================
Lag Data
===============================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
-------------------------------------------------------------------------------
1              up      down    No       0         0        standby
===============================================================================
*A:PE-2#
```

In this case, the LAG on PE-1 is active (operationally up) whereas the LAG on PE-2 is standby (operationally down).

The default selection criteria is highest number of links and priority. In this example, the number of links and the priority of the links is the same on both redundant PEs. Whichever PE’s LAG gets the operational status **up** first, will be the active LAG.

LAG ports of one PE could be preferred over the other PE by configuring port priority. For example, the following command lowers the priority of the LAG ports on PE-1, thus giving this LAG higher preference. The default priority is 32768, but it is modified to a value of 10, as follows:

```
*A:PE-1# configure lag 1 port 1/1/3 1/1/4 priority 10
```

The selection criteria can be configured as highest-count, highest-weight or best-port (the default is highest count).

```
*A:PE-1# configure lag 1 selection-criteria
  - selection-criteria [best-port|highest-count|highest-weight] [slave-to-partner]
                                                [subgroup-hold-time <hold-time>]
  - no selection-criteria

 <best-port|highest*> : keywords
 <slave-to-partner>   : keyword
 <hold-time>          : [0..2000] tenths of a second | infinite
```

If highest-weight is configured, the sum of the weights of the LAG members is considered. The weight of an individual LAG member is calculated as priority 65535 (the default is 32768).

Step 7 - Verify detailed MC-LAG status on PE-1

```
*A:PE-1# show lag 1 detail

===============================================================================
LAG Details
===============================================================================
Description        : N/A
-------------------------------------------------------------------------------
Details
-------------------------------------------------------------------------------
Lag-id              : 1                     Mode                 : access
Adm                 : up                    Opr                  : up
Thres. Exceeded Cnt : 2                     Port Threshold       : 0
Thres. Last Cleared : 07/11/2018 12:42:16   Port Thres. Action   : down
Dynamic Cost        : false
Encap Type          : null
Configured Address  : 02:0b:ff:00:01:41     Lag-IfIndex          : 1342177281
Hardware Address    : 02:0b:ff:00:01:41     Adapt Qos (access)   : distribute
Hold-time Down      : 0.0 sec               Port Type            : standard
Per-Link-Hash       : disabled
Include-Egr-Hash-Cfg: disabled              Forced               : -
Per FP Ing Queuing  : disabled              Per FP Egr Queuing   : disabled
Per FP SAP Instance : disabled
Access Bandwidth    : N/A                   Access Booking Factor: 100
Access Available BW : 0
Access Booked BW    : 0
LACP                : enabled               Mode                 : active
LACP Transmit Intvl : fast                  LACP xmit stdby      : enabled
Selection Criteria  : highest-count         Slave-to-partner     : disabled
MUX control         : coupled
Subgrp hold time    : 0.0 sec               Remaining time       : 0.0 sec
Subgrp selected     : 1                     Subgrp candidate     : -
Subgrp count        : 1
System Id           : 02:0b:ff:00:00:00     System Priority      : 32768
Admin Key           : 32768                 Oper Key             : 1
Prtr System Id      : 02:13:ff:00:00:00     Prtr System Priority : 32768
Prtr Oper Key       : 32768
Standby Signaling   : lacp
Port weight speed   : 0 gbps                Number/Weight Up     : 2
Weight Threshold    : 0
Weight Thres. Action: down

MC Peer Address     : 192.0.2.2             MC Peer Lag-id       : 1
MC System Id        : 00:00:00:00:00:01     MC System Priority   : 100
MC Admin Key        : 1                     MC Active/Standby    : active
MC Lacp ID in use   : true                  MC extended timeout  : false
MC Selection Logic  : local master decided
MC Config Mismatch  : no mismatch

-------------------------------------------------------------------------------
Port-id        Adm     Act/Stdby Opr     Primary   Sub-group     Forced  Prio
-------------------------------------------------------------------------------
1/1/3          up      active    up      yes       1             -       10
1/1/4          up      active    up                1             -       10

-------------------------------------------------------------------------------
Port-id        Role      Exp   Def   Dist  Col   Syn   Aggr  Timeout  Activity
-------------------------------------------------------------------------------
```

```
1/1/3           actor     No     No     Yes    Yes    Yes    Yes    Yes       Yes
1/1/3           partner   No     No     Yes    Yes    Yes    Yes    Yes       Yes
1/1/4           actor     No     No     Yes    Yes    Yes    Yes    Yes       Yes
1/1/4           partner   No     No     Yes    Yes    Yes    Yes    Yes       Yes
===============================================================================
*A:PE-1#
```

After changing the LAG port priorities from default (32768) to 10, the LAG on PE-1 is in up/up state and the ports are in up/active/up status. This show command also displays actor and partner bits set in the LACP messages.

Step 8 - MC-LAG configuration on PE-3 and PE-4.

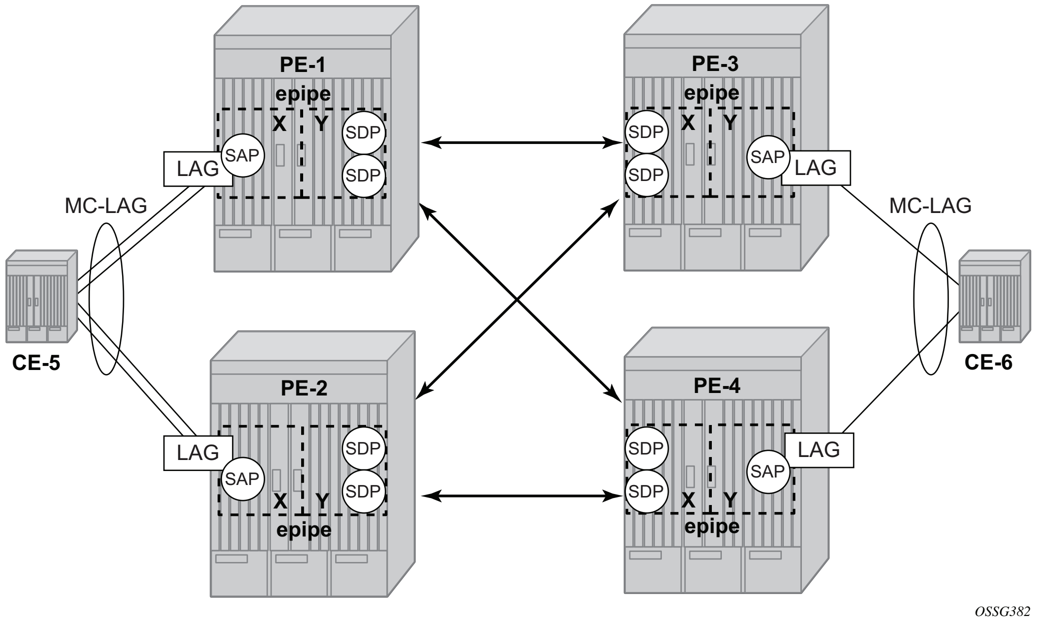
The MC-LAG configuration on PE-3 and PE-4 is similar to the configuration on PE-1 and PE-2. In this case, the priority of the LAG port on PE-4 is lowered to obtain the behavior in Figure 33 where the LAGs on PE-1 and PE-4 are active.

Step 9 - Pseudowire configuration.

Configure an Epipe service on every PE and create endpoints **x** and **y** (the endpoint names can be any text string). Traffic can only be forwarded between two endpoints, for example, it is not possible for objects associated with the same endpoint to forward traffic to each other.

Associate the SAPs and spoke SDPs with the endpoints as shown in Figure 34.

Figure 34 **Association of SAPs/SDPs and Endpoints**



```
# on PE-1
configure
    service
        epipe 1 name "epipe-1" customer 1 create
            endpoint "X" create
            exit
            endpoint "Y" create
            exit
            sap lag-1 endpoint "X" create
            exit
            spoke-sdp 13:1 endpoint "Y" create
            exit
            spoke-sdp 14:1 endpoint "Y" create
            exit
            no shutdown
        exit
    exit
exit
```

Likewise, an Epipe service, endpoints, SAPs and spoke SDPs must be configured on the other PE routers.

Step 10 - Pseudowire verification.

```
*A:PE-1# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Up   1          epipe-1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

```
*A:PE-2# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Down 1          epipe-1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

```
*A:PE-3# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Down 1          epipe-1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-3#
```

```
*A:PE-4# show service service-using

===============================================================================
Services
===============================================================================
ServiceId    Type      Adm  Opr  CustomerId Service Name
-------------------------------------------------------------------------------
1            Epipe     Up   Up   1          epipe-1
2147483648   IES       Up   Down 1          _tmnx_InternalIesService
2147483649   intVpls   Up   Down 1          _tmnx_InternalVplsService
-------------------------------------------------------------------------------
Matching Services : 3
-------------------------------------------------------------------------------
===============================================================================
*A:PE-4#
```

The Epipe service on PE-2 and PE-3 is down and up on PE-1 and PE-4. This reflects the standby behavior shown in Figure 33. However, after configuring ICB spoke SDPs (described later in this document), the Epipe will be in up/up status on all PE routers.

Step 11 - Verify SDP status

Local pseudowire bits indicate the status of the pseudowire on the PE node. These pseudowire bits will be sent to the peer. Peer pseudowire bits indicate the status of the pseudowire on the peer, as sent by the peer. The following example is taken on PE-2:

```
*A:PE-2# show service id 1 sdp 23:1 detail

===============================================================================
Service Destination Point (Sdp Id : 23:1) Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 23:1  -(192.0.2.3)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 23:1                    Type               : Spoke
Spoke Descr     : (Not Specified)
```

```
VC Type            : Ether                   VC Tag            : n/a
Admin Path MTU     : 0                       Oper Path MTU     : 1556
Delivery           : MPLS
Far End            : 192.0.2.3               Tunnel Far End    :
Oper Tunnel Far End: 192.0.2.3
LSP Types          : LDP
Hash Label         : Disabled                Hash Lbl Sig Cap  : Disabled
Oper Hash Label    : Disabled
Entropy Label      : Disabled

Admin State        : Up                      Oper State        : Up
MinReqd SdpOperMTU : 1514
Acct. Pol          : None                    Collect Stats     : Disabled
Ingress Label      : 524283                  Egress Label      : 524282
Ingr Mac Fltr-Id   : n/a                     Egr Mac Fltr-Id   : n/a
Ingr IP Fltr-Id    : n/a                     Egr IP Fltr-Id    : n/a
Ingr IPv6 Fltr-Id  : n/a                     Egr IPv6 Fltr-Id  : n/a
Admin ControlWord  : Not Preferred           Oper ControlWord  : False
Admin BW(Kbps)     : 0                       Oper BW(Kbps)     : 0
BFD Template       : None
BFD-Enabled        : no                      BFD-Encap         : ipv4
Last Status Change : 07/11/2018 12:51:04     Signaling         : TLDP
Last Mgmt Change   : 07/11/2018 12:50:53
Endpoint           : Y                       Precedence        : 4
PW Status Sig      : Enabled
Force Vlan-Vc      : Disabled                Force Qinq-Vc     : none
Class Fwding State : Down
Flags              : None
Local Pw Bits      : lacIngressFault lacEgressFault pwFwdingStandby
Peer Pw Bits       : lacIngressFault lacEgressFault pwFwdingStandby
Peer Fault Ip      : None
Peer Vccv CV Bits  : lspPing bfdFaultDet
Peer Vccv CC Bits  : mplsRouterAlertLabel

--- snipped ---

-------------------------------------------------------------------------------
Segment Routing
-------------------------------------------------------------------------------
ISIS               : disabled
OSPF               : disabled
TE-LSP             : disabled
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-2#
```

In this example, the remote side of the SDP is sending lacIngressFault lacEgressFault pwFwdingStandby flags. This is because the Epipe service on PE-3 is down because the MC-LAG is in standby/down status.

Link and node protection can be tested. The access links are protected by the MC-LAG, the PE routers are protected by the combination of MC-LAG/pseudowire redundancy. The SDPs can be protected by FRR in the case of RSVP-TE or LDP.

Revertive behavior is expected when different MC-LAG port priorities are configured or if the number of MC-LAG ports is different on the MC-LAG peers: convergence takes place when the active PE fails and convergence takes place again when that PE is online again.

In case of revertive behavior, MC-LAG convergence might take less time than the setup of the spoke SDPs, thus creating a temporary black-hole. To avoid this situation, it is best to configure **hold-time up** on the LAG ports. In that case, the ports are kept in a down state for a configured period of time after the node has rebooted. This is done to ensure that the SDPs are operationally up when the MC-LAG convergence takes place. The **hold-time up** is expressed in seconds.

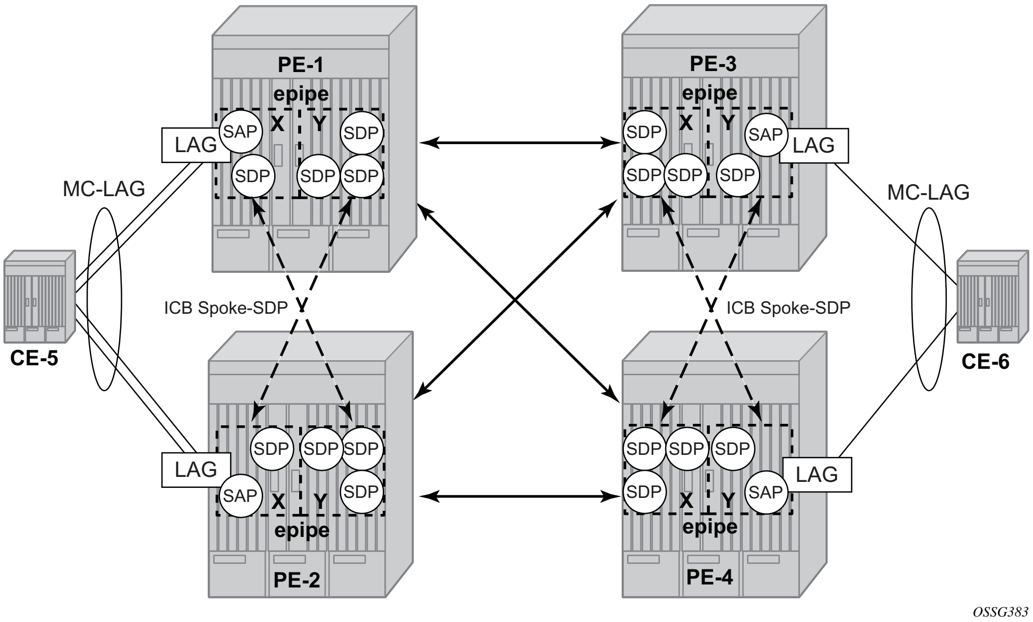
```
*A:PE-1# configure port 1/1/3 ethernet hold-time up 50
*A:PE-1# configure port 1/1/4 ethernet hold-time up 50
```

Step 12 - Inter-Chassis Backup (ICB) pseudowire configuration.

In this setup, the configuration of ICBs is optional. It can be used to speed up convergence by forwarding in-flight packets during MC-LAG transition. Figure 36 shows some setup examples where ICBs are required. ICBs cannot be configured at endpoints where the other object is a standard SAP, only MC-LAG SAPs and pseudowires are allowed with ICBs.

ICB SDPs and associated to endpoints as shown in Figure 35.

Figure 35 **ICB Spoke SDPs and Their Association with the Endpoints**



Two ICB spoke SDPs must be configured in the Epipe service on each PE router, one in each endpoint. Different SDP IDs can be used for the ICBs (as opposed to the regular pseudowires), but this is not necessary, because the far-end will be the same. The *vc-id* must be different however.

The ICB spoke SDPs must cross, one end should be associated with endpoint **x** and the other end (on the other PE) should be associated with endpoint **y**. After configuring the ICB spoke SDPs, the Epipe service will be up/up on all four PE routers.

Only one spoke SDP will be forwarding. If there is an ICB and a MC-LAG SAP in an endpoint, the ICB will only forward if the SAP goes down. If an ICB resides in an endpoint together with other spoke SDPs, the ICB will only forward if there is no other active spoke SDP.

The following output shows the additional Epipe service configuration on each PE:

```
# on PE-1
configure
    service
        epipe 1
            spoke-sdp 12:1 endpoint "X" icb create
            exit
            spoke-sdp 12:2 endpoint "Y" icb create
            exit
```

```
# on PE-2
configure
    service
        epipe 1
            spoke-sdp 21:1 endpoint "Y" icb create
            exit
            spoke-sdp 21:2 endpoint "X" icb create
            exit
```

```
# on PE-3
configure
    service
        epipe 1
            spoke-sdp 34:1 endpoint "X" icb create
            exit
            spoke-sdp 34:2 endpoint "Y" icb create
            exit
```

```
# on PE-4
configure
    service
        epipe 1
            spoke-sdp 43:1 endpoint "Y" icb create
            exit
            spoke-sdp 43:2 endpoint "X" icb create
            exit
```

Step 13 - Verification of active objects for each endpoint.

The following command shows which objects are configured for each endpoint and which is the active object at this moment:

```
*A:PE-1# show service id 1 endpoint

===============================================================================
Service 1 endpoints
===============================================================================
Endpoint name                : X
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Standby Signaling Master     : false
Standby Signaling Slave      : false
Tx Active                    : lag-1
Tx Active Up Time            : 0d 00:09:47
Revert Time Count Down       : N/A
Tx Active Change Count       : 1
Last Tx Active Change        : 07/11/2018 12:49:52
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
SAP      : lag-1                                   Oper Status: Up
Spoke-sdp: 12:1 Prec:4 (icb)                       Oper Status: Up
===============================================================================
Endpoint name                : Y
Description                  : (Not Specified)
Creation Origin              : manual
Revert time                  : 0
Act Hold Delay               : 0
Standby Signaling Master     : false
Standby Signaling Slave      : false
Tx Active (SDP)              : 14:1
Tx Active Up Time            : 0d 00:06:02
Revert Time Count Down       : N/A
Tx Active Change Count       : 2
Last Tx Active Change        : 07/11/2018 12:53:37
-------------------------------------------------------------------------------
Members
-------------------------------------------------------------------------------
Spoke-sdp: 12:2 Prec:4 (icb)                       Oper Status: Up
Spoke-sdp: 13:1 Prec:4                             Oper Status: Up
Spoke-sdp: 14:1 Prec:4                             Oper Status: Up
===============================================================================
===============================================================================
*A:PE-1#
```

On PE-1, the SAP and the spoke SDP 14:1 are active. The other objects do not forward traffic.

Step 14 - Other types of setups.

Figure 36 and Figure 37 show other setups that combine MC-LAG and pseudowire redundancy.

Figure 36 Additional Setup Example 1

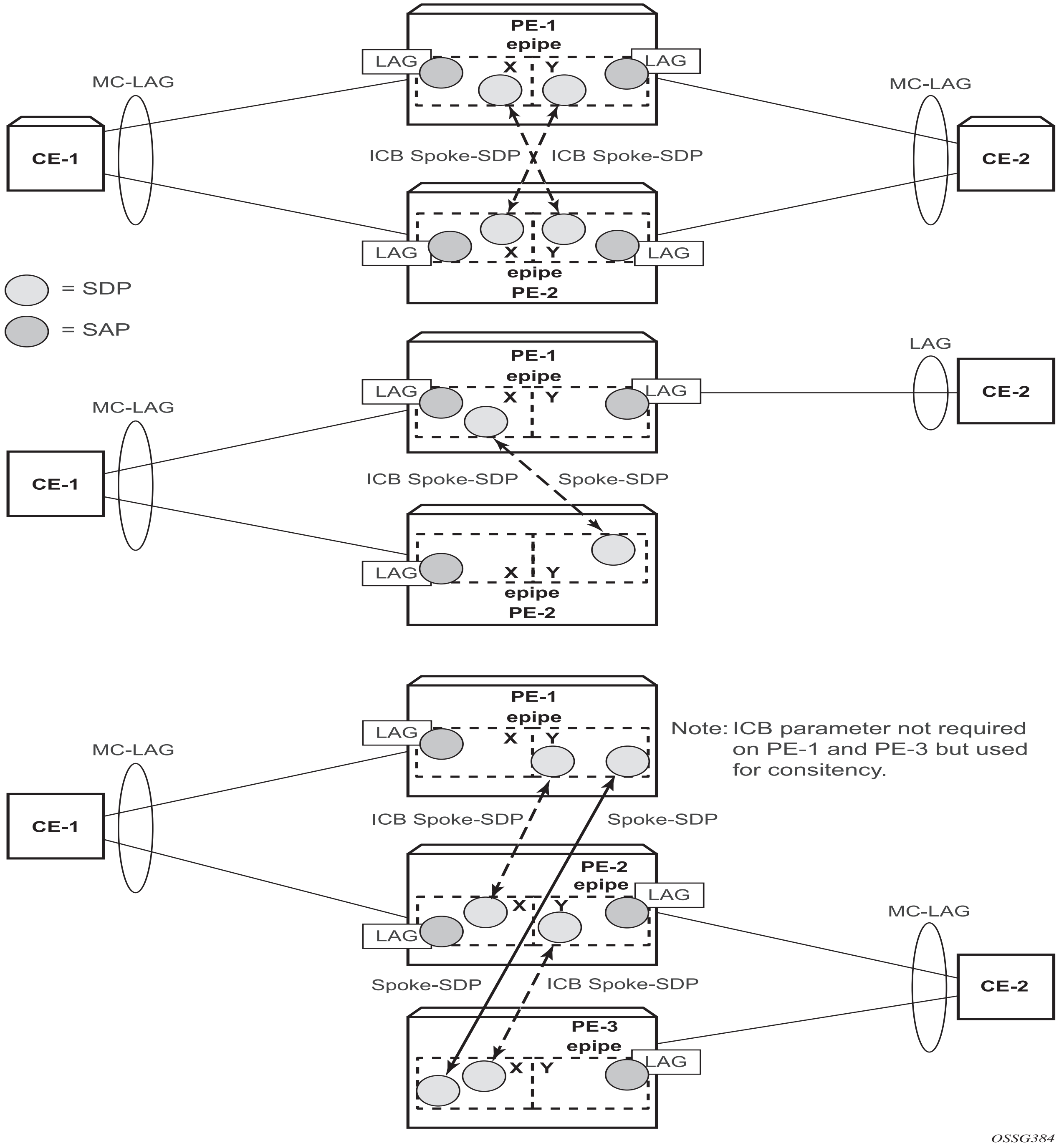
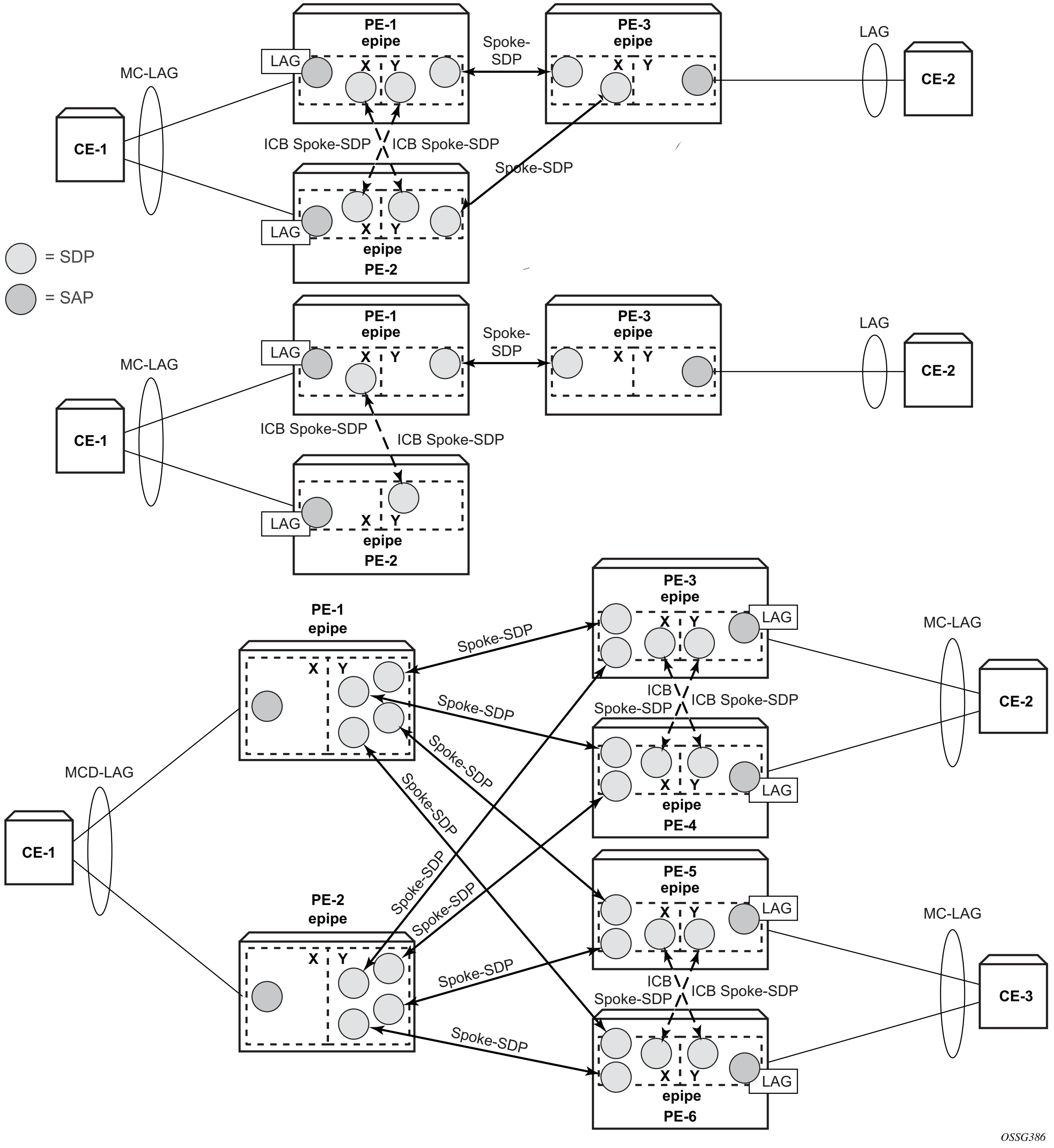


Figure 37 Additional Setup Example 2



MC-LAG for VPLS Services

MC-LAG can also be configured for VPLS services. When the MC-LAG converges, the PE that transitions to standby state for the MC-LAG will send out an LDP address withdrawal message to all peers configured in the VPLS service. Both types of SDPs (spoke and mesh) support this feature. The PE peers will then flush all the MAC addresses learned via the PE that sent the LDP MAC address withdrawal message.

Because a VPLS service is a multipoint service, pseudowire redundancy is not required. The MC-LAG redundancy configuration is identical.

Forced Switchover

MC-LAG convergence can be forced with **tools perform lag** command:

```
*A:PE-1# tools perform lag force
  - force all-mc {active|standby}
  - force lag-id <lag-id> [sub-group <sub-group-id>] {active|standby}
  - force peer-mc <ip-address> {active|standby}

 <lag-id>             : [1..800]
 <sub-group-id>       : [1..16]
 <all-mc>             : keyword
 <ip-address>         : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
 <active|standby>     : keywords

*A:PE-1# tools perform lag force lag-id 1 standby


*A:PE-1# show lag 1
===============================================================================
Lag Data
===============================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
-------------------------------------------------------------------------------
1              up      down    No       0         0        standby
===============================================================================
*A:PE-1#
```

After the forced switchover, it is important to clear the forced switchover:

```
*A:PE-1# tools perform lag clear-force
  - clear-force all-mc
  - clear-force lag-id <lag-id> [sub-group <sub-group-id>]
  - clear-force peer-mc <ip-address>
```

```
 <lag-id>             : [1..800]
 <sub-group-id>       : [1..16]
 <all-mc>             : keyword
 <ip-address>         : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D


*A:PE-1# tools perform lag clear-force lag-id 1


*A:PE-1# show lag 1

===============================================================================
Lag Data
===============================================================================
Lag-id         Adm     Opr     Weighted Threshold Up-Count MC Act/Stdby
-------------------------------------------------------------------------------
1              up      up      No       0         2        active
===============================================================================
*A:PE-1#
```

Conclusion

MC-LAG is a Nokia added value redundancy feature that offers fast access link convergence in Epipe and VPLS services for CE devices that support standard LACP. PE node convergence for VPLS services is enhanced by using LDP address withdrawal messages to flush the FDB on the PE peers. PE node convergence for Epipes is guaranteed by using pseudowire redundancy.

Port Cross-Connect (PXC)

This chapter provides information about Port Cross-Connect (PXC).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 14.0.R5.

Overview

The Port Cross-Connect (PXC) feature allows for a port, or number of ports, to be logically looped to themselves. The purpose of looping a port in this manner is to provide an "anchor point" function, such that traffic may ingress the node through any interface/port and be redirected to that anchor point.

When traffic is passed through the egress data path of the PXC, it can be used for additional packet processing that cannot be supported on the ingress data path, such as the removal of an encapsulation header. When traffic is looped back to the ingress data path of the PXC, it is processed as if it were the conventional service termination point. This essentially decouples the Input/Output (I/O) port through which packets ingress the node from the I/O port that implements the service termination. This decoupling removes the previous constraint for pseudowire-port (pw-port) whereby the I/O port through which packets ingress and egress the node was bound and could not be changed during, for example, a reconvergence event.

PXC provides two modes of operation: Distributed Versatile Service Module (DVSM) mode and Application Specific (AS) mode.

- The DVSM mode provides functionality like that of the VSM2 card, enabling the user to create an internal loopback through the card. This allows for back-to-back configurations similar to a VLAN cross-connect.

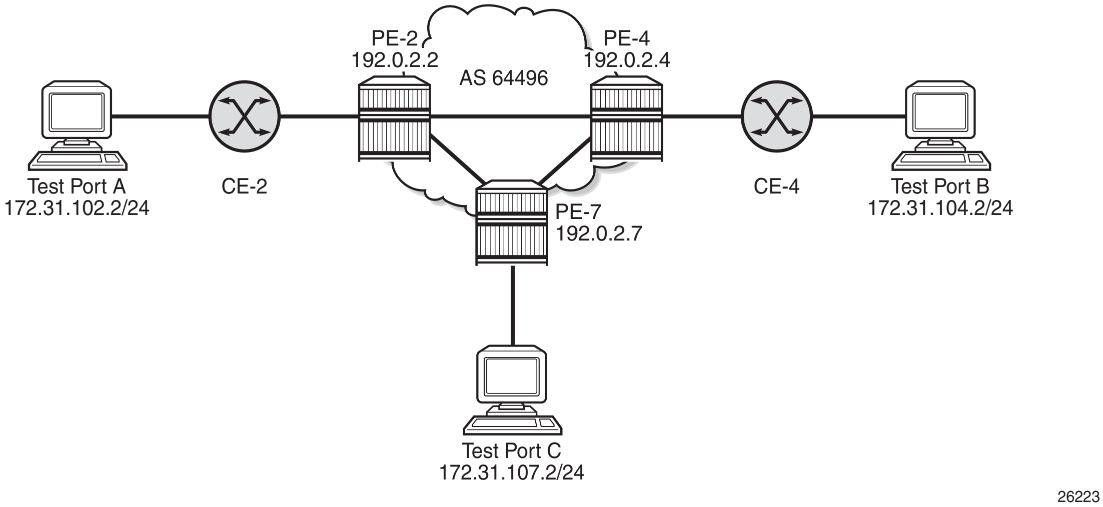
- The AS mode creates a Forwarding Path Extension (FPE) context through which the system can automatically create cross-connects to simplify user provisioning. Use-case examples for AS mode include PW port for business VPN services, VXLAN termination on a non-system interface, ESM over Pseudowire, and GRE tunnel termination.

This chapter describes the generic principles of PXC, combined with examples of both DVSM mode and AS mode.

Example Topology

The topology shown in Figure 38 is used within this chapter to illustrate the use of PXC. PE-2, PE-4, and PE-7 form part of Autonomous System 64496 and run IS-IS level 2 together with LDP for the MPLS control plane. PE-2, PE-4, and PE-7 also peer in IBGP for the VPN-IPv4 address family. Test ports are connected to all PEs (in the case of PE-2 and PE-4, via CE routers) for the purpose of validating IP connectivity.

Figure 38 Example Topology



PE-7 will host the PXC and is equipped with an FP3-based 20 x 10GE IMM in slot 1 for this purpose, as shown in the following output:

```
*A:PE-7# show card 1
===============================================================================
Card 1
===============================================================================
Slot      Provisioned Type                         Admin Operational   Comments
              Equipped Type (if different)         State State
```

```
-------------------------------------------------------------------------------
1         imm-2pac-fp3                             up    up
===============================================================================
```

Configuration

PXC Configuration

A PXC can consist of a single non-redundant port, or for redundancy and increased capacity, can consist of multiple ports that form member links of a Link Aggregation Group (LAG). Both options are described here.

Non-Redundant PXC

The non-redundant PXC is created within the **port-xc** context and can be numbered from 1 to 64. A port must be assigned to the PXC before it is put into a **no shutdown** state, and that port must be in a **shutdown** state when it is assigned. There is no requirement for any kind of optical transceiver to be inserted in the port assigned to the PXC; it is only a logical loopback. When the port is assigned to the PXC, it cannot be used for any other purpose besides a PXC-based service assignment (for example, a regular SAP could not be configured on this port).

```
configure
    port-xc
        pxc 1 create
            description "PXC-dVSM mode non-redundant"
            port 1/2/1
            no shutdown
        exit
    exit
```

After the PXC has been put into a **no shutdown** state, two PXC sub-ports are automatically created by the system. The PXC sub-ports are identified by *.a* and *.b* suffixes of the parent PXC (in this example, pxc-1) and are created in hybrid mode with an MTU of 9208 bytes, both of which are non-configurable. The 9208-byte MTU represents the SR OS default of 9212 bytes minus four bytes to allow for an internal VLAN tag that is used to identify each back-to-back sub-port. Finally, the encapsulation is set to dot1q, which is the default for hybrid ports. Q-in-Q encapsulation is also supported. It is also possible to configure dot1q encapsulation on one PXC sub-port and Q-in-Q encapsulation on the opposing PXC sub-port if, for example, there is a requirement to expose more VLAN tags on one side of the loop than the other side of the loop.

```
*A:PE-7# show port pxc 1
===============================================================================
Ports on Port Cross Connect 1
===============================================================================
Port          Admin Link Port    Cfg  Oper LAG/ Port Port Port   C/QS/S/XFP/
Id            State      State   MTU  MTU  Bndl Mode Encp Type   MDIMDX
-------------------------------------------------------------------------------
pxc-1.a       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
pxc-1.b       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
===============================================================================
```

After the PXC creation, the PXC sub-port CLI configuration is automatically generated and can be accessed in the same way as a conventional physical port, using the syntax "port pxc-n.l" where "n" represents the assigned PXC number and "l" represents the sub-port letter (a or b). As shown in the previous output, the sub-ports are in an admin down state following automatic creation and need to be manually put into a **no shutdown** state, as follows:

```
*A:PE-7# configure port pxc-1.a no shutdown
*A:PE-7# configure port pxc-1.b no shutdown
```

The physical port assigned to the PXC must also now be put into a **no shutdown** state in order for the PXC to become operational:

```
*A:PE-7# configure port 1/2/1 no shutdown
```

The command in the following output can then be used to verify the operational state of the PXC:

```
*A:PE-7# show port-xc pxc 1

===============================================================================
Port Cross-Connect Information
===============================================================================
PXC    Admin     Oper      Port        Description
Id     State     State     Id
-------------------------------------------------------------------------------
1      Up        Up        1/2/1       PXC non-redundant
===============================================================================
```

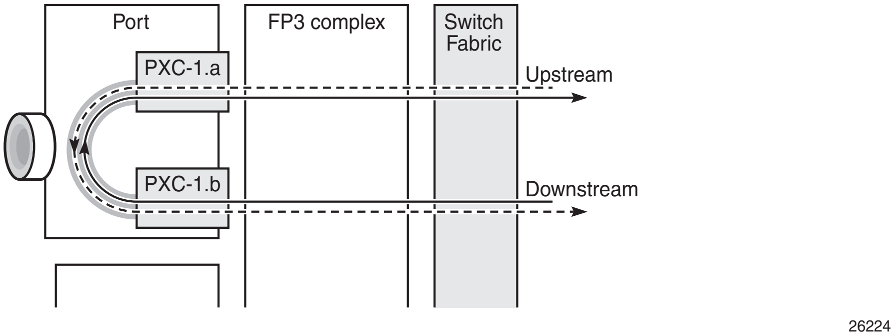
Similarly, the operational state of each of the sub-ports can be verified as follows. The physical link is indicated as being present even though there is no transceiver installed in this port.

```
*A:PE-7# show port pxc-1.a
```

```
===============================================================================
Ethernet Interface
===============================================================================
Description        : Port cross-connect
Interface          : pxc-1.a                     Oper Speed       : 10 Gbps
Link-level         : Ethernet                    Config Speed     : N/A
Admin State        : up                          Oper Duplex      : full
Oper State         : up                          Config Duplex    : N/A
Physical Link      : Yes                         MTU              : 9208
Single Fiber Mode  : No                          Min Frame Length : 64 Bytes
IfIndex            : 1090523137                  Hold time up     : 0 seconds
Last State Change  : 10/27/2016 09:47:18         Hold time down   : 0 seconds
Last Cleared Time  : N/A
Phys State Chng Cnt: 0
---snip---
```

Figure 39 shows a representation of the non-redundant PXC configuration. Both upstream and downstream traffic will pass twice through the FP data-path and port. For example, upstream traffic passes through the FP complex and PXC-1.b. The traffic is then looped back to PXC-1.a, and back into the FP complex. Similarly, downstream traffic passes through the FP complex to PXC-1.a. It is then looped back to PXC-1.b and back into the FP complex.

Figure 39 **Non-Redundant PXC**



When using a PXC, the physical port effectively simulates two (sub-)ports, which creates two egress traffic paths: one upstream and one downstream. When the receive side of the PXC port receives those paths, it needs to distinguish between them, and this is where the internal additional VLAN tag is used.

The difference between this PXC configuration and a conventional port not looped or configured as PXC is as follows. With a conventional port, ingress traffic passes through the port and ingress data-path of the FP complex only once, and egress traffic passes through the egress data-path of the FP complex and port only once.

Redundant PXC

For a redundant PXC, the fundamental building blocks are identical to those of the non-redundant PXC, but there are a few additional configuration steps required to construct the LAGs to which the redundant PXC ports belong.

The redundant PXC example consists of two ports: 1/2/2 and 1/2/3 in the following output. In this case, the redundant PXC ports belong to the same IMM, but different IMMs can be used for increased redundancy. Two PXCs are created and each one is assigned one of the redundant PXC ports. Both PXCs are put into a **no shutdown** state.

```
configure
    port-xc
        pxc 2 create
            description "PXC redundant"
            port 1/2/2
            no shutdown
        exit
        pxc 3 create
            description "PXC redundant"
            port 1/2/3
            no shutdown
        exit
    exit
```

As with the non-redundant PXC, when the PXC has been put into a **no shutdown** state, two PXC sub-ports with .a and .b suffixes are automatically created by the system for each PXC port:

```
*A:PE-7# show port pxc [2..3]

===============================================================================
Ports on Port Cross Connect 2
===============================================================================
Port          Admin Link Port    Cfg  Oper LAG/ Port Port Port   C/QS/S/XFP/
Id            State      State   MTU  MTU  Bndl Mode Encp Type   MDIMDX
-------------------------------------------------------------------------------
pxc-2.a       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
pxc-2.b       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
===============================================================================

===============================================================================
Ports on Port Cross Connect 3
===============================================================================
Port          Admin Link Port    Cfg  Oper LAG/ Port Port Port   C/QS/S/XFP/
Id            State      State   MTU  MTU  Bndl Mode Encp Type   MDIMDX
-------------------------------------------------------------------------------
pxc-3.a       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
pxc-3.b       Down  Yes  Link Up 9208 9208    - hybr dotq xgige
===============================================================================
```

The PXC sub-ports, together with the physical port, must then all be put into a **no shutdown** state:

```
*A:PE-7# configure port pxc-2.a no shutdown
*A:PE-7# configure port pxc-2.b no shutdown
*A:PE-7# configure port pxc-3.a no shutdown
*A:PE-7# configure port pxc-3.b no shutdown
*A:PE-7# configure port 1/2/2 no shutdown
*A:PE-7# configure port 1/2/3 no shutdown
```

After the associated components have been put into a no shutdown state, the operational state of the PXCs can be verified:

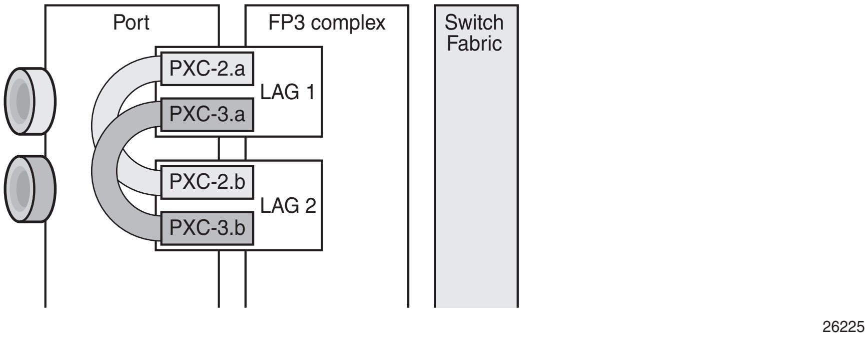
```
*A:PE-7# show port-xc pxc [2..3]

===============================================================================
Port Cross-Connect Information
===============================================================================
PXC    Admin     Oper      Port        Description
Id     State     State     Id
-------------------------------------------------------------------------------
2      Up        Up        1/2/2       PXC redundant
===============================================================================

===============================================================================
Port Cross-Connect Information
===============================================================================
PXC    Admin     Oper      Port        Description
Id     State     State     Id
-------------------------------------------------------------------------------
3      Up        Up        1/2/3       PXC redundant
===============================================================================
```

The PXC sub-ports are then associated with two LAGs to essentially form an internal back-to-back LAG. To do this, both sub-ports with the .a suffix belong to one LAG instance, and both sub-ports with the .b suffix belong to the other LAG instance. Like any other LAG member links, PXC sub-ports in a LAG must be configured with the same physical attributes, such as speed and duplex. Both LAG instances are configured with **mode hybrid** to match the mode of the physical ports. Setting the mode to **hybrid** automatically sets the **encap-type** to **dot1q**.

```
configure
    lag 1
        mode hybrid
        encap-type dot1q
        port pxc-2.a
        port pxc-3.a
        no shutdown
    exit
    lag 2
        mode hybrid
        encap-type dot1q
        port pxc-2.b
        port pxc-3.b
        no shutdown
    exit
```

Figure 40 shows a representation of the redundant PXC with LAG. Both upstream and downstream traffic will pass twice through the FP data-path and port.

Figure 40 **PXC Redundant Mode with LAG**



When the LAGs are configured and the associated PXC sub-ports assigned as member links, the operational status can be verified. Note that at the LAG level, each of the configured LAG instances is not aware that it is internally connected to another LAG instance, even though the member sub-ports are logically looped. It would be possible, for example, to put LAG 1 into an admin shutdown state and not affect the operational state of LAG 2. LACP is not supported for PXC LAG; however, it is possible to run the 802.3ah Ethernet in the First Mile (EFM) at PXC sub-port level, if required.

```
*A:PE-7# show lag 1 detail

===============================================================================
LAG Details
===============================================================================
Description         : N/A
-------------------------------------------------------------------------------
Details
-------------------------------------------------------------------------------
Lag-id              : 1                     Mode                 : hybrid
Adm                 : up                    Opr                  : up
Thres. Exceeded Cnt : 5                     Port Threshold       : 0
Thres. Last Cleared : 10/27/2016 14:12:57   Threshold Action     : down
Dynamic Cost        : false                 Encap Type           : dot1q
Configured Address  : 00:07:ff:00:01:41     Lag-IfIndex          : 1342177281
Hardware Address    : 00:07:ff:00:01:41     Adapt Qos (access)   : distribute
Hold-time Down      : 0.0 sec               Port Type            : standard
Per-Link-Hash       : disabled
Include-Egr-Hash-Cfg: disabled
Per FP Ing Queuing  : disabled              Per FP Egr Queuing   : disabled
Per FP SAP Instance : disabled
Access Bandwidth    : not-applicable        Access Booking Factor: 100
Access Available BW : 0
Access Booked BW    : 0
```

```
LACP                 : disabled
Standby Signaling    : lacp
Port weight speed    : 0 gbps              Number/Weight Up     : 2
Weight Threshold     : 0                   Threshold Action     : down

-------------------------------------------------------------------------------
Port-id        Adm     Act/Stdby Opr      Primary    Sub-group      Forced  Prio
-------------------------------------------------------------------------------
pxc-2.a        up      active    up       yes        1              -       32768
pxc-3.a        up      active    up                  1              -       32768
===============================================================================
```

DVSM Mode

DVSM mode enables the creation of a back-to-back cross-connect. This back-to-back connection can be network-to-network, access-to-access, or a combination such as network-to-access. To provide an example of using DVSM mode, PE-4 in Figure 38 functions as a Layer 2 backhaul device, and PE-7 housing the PXC functions as the Layer 3 service edge. A pseudowire is extended from PE-4 to PE-7, where it is terminated in a VPRN, providing point-to-point connectivity between CE-4 and PE-7.

VLAN 100 is extended from CE-4 to PE-4, where it is indexed into an Epipe service. The SAP is service-delimiting; therefore, the VLAN is removed before frames are encapsulated into the pseudowire. The Epipe then has a single non-redundant spoke-SDP to PE-7 with VC-ID 11. The service configuration on PE-4 is as follows:

```
configure
    service
        sdp 2007 mpls create
            far-end 192.0.2.7
            ldp
            no shutdown
        exit
        epipe 11 customer 1 create
            sap 1/1/4:100 create
                no shutdown
            exit
            spoke-sdp 2007:11 create
                no shutdown
            exit
            no shutdown
        exit
```

At PE-7, the configuration of the corresponding end of the Epipe service is shown in the following output. This service consists of a single spoke-SDP toward PE-4 with VC-ID 11 to match the VC-ID advertised by PE-4, and a single SAP toward the PXC port. The syntax takes the form "pxc-n.l:vlan" where "n" is the PXC identifier, "l" is the sub-port letter (in this case .a), and "vlan" represents the VLAN identifer of the SAP.

As shown in the following output, the Epipe service uses PXC 1, which is the non-redundant PXC port. This is only an example; it could similarly use the redundant PXC port, in which case the SAP syntax would be the conventional LAG syntax (for example, lag-1:100, lag-2:100). Also note that although VLAN 100 is used both at PE-4's Epipe SAP and PE-7's Epipe PXC SAP, there is no correlation or dependence between the two. Both VLAN tags are service-delimiting and are subsequently stripped before the Ethernet frame is encapsulated into the pseudowire payload, so any valid VLAN value could be used at either point. The service configuration on PE-7 is as follows:

```
configure
    service
        sdp 2004 mpls create
            far-end 192.0.2.4
            ldp
            no shutdown
        exit
        epipe 11 customer 1 create
            sap pxc-1.a:100 create
                no shutdown
            exit
            spoke-sdp 2004:11 create
                no shutdown
            exit
            no shutdown
        exit
```

The VPRN configuration at the corresponding side of the PXC port is shown in the following output. The VPRN has two interfaces: the first is toward a directly connected test port used to verify IP connectivity, and the second ("to-CE-4") is toward CE-4 and has a SAP with a PXC syntax. The PXC syntax represents the same PXC and VLAN identifiers as the preceding Epipe configuration, but the PXC sub-port is .b, to represent the "other side" of the PXC logical loopback. Therefore, the VLAN values must match to create the back-to-back connection. Although not shown in the output (for brevity), a BGP session is configured between PE-7 and CE-4 for route exchange. The remainder of the VPRN parameters are generic and are not explained here.

```
PE-7
configure
    service
        vprn 10 customer 1 create
            vrf-import "vrf10-import"
            vrf-export "vrf10-export"
            autonomous-system 64496
            route-distinguisher 64496:10
            auto-bind-tunnel
                resolution any
            exit
            interface "Test-Port-C" create
                address 172.31.107.1/24
                sap 1/1/3:100 create
                exit
            exit
            interface "to-CE-4" create
                address 192.168.100.1/30
                sap pxc-1.b:100 create
                exit
            exit
            bgp
                group "EBGP"
            ---snip---
            no shutdown
        exit
```

PXC Port Dimensioning

When the VPRN service at PE-7 is put into a **no shutdown** state, the EBGP session to CE-4 is established. The relevant routes are exchanged between CE-4 and PE-7 and traffic can be exchanged between test ports B (connected to CE-4) and C (connected to PE-7). Initially, traffic is sent from test port B toward port C at a rate of 100 packets/s. Traffic is intentionally sent in only one direction (in this example) to emphasize a point regarding PXC port dimensioning and capacity planning, as follows.

The PXC in use by the Epipe/VPRN service is PXC 1, which uses physical port 1/2/1. The following output shows a snapshot of a monitor command against the physical port. Although traffic is only being sent in a single direction (test port B behind CE-4 toward test port C connected to PE-7), the input/output rate of packets per second is the same at 100 packets/s. This is because the physical port consists of two PXC sub-ports that are looped. In this example, traffic is output from pxc-1.a when traffic is sent from the Epipe SAP into the PXC port, and traffic is input at pxc-1.b when traffic is received by the VPRN SAP from the PXC port. Because both upstream/ingress traffic and downstream/egress traffic will be seen as output packets using the available capacity of the physical port, this needs to be considered when capacity is being planned.

```
*A:PE-7# monitor port 1/2/1 rate interval 3

===============================================================================
Monitor statistics for Port 1/2/1
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Octets                                             51600                  51600
Packets                                              100                    100
Errors                                                 0                      0
Bits                                              412800                 412800
Utilization (% of port capacity)                   ~0.00                  ~0.00
```

QoS Continuity

The application of ingress/egress SAP QoS policies is fundamentally the same for a PXC-based SAP as it is for a conventional SAP. However, there is a difference with regard to how ingress Forwarding Class (FC) mappings are maintained throughout the PXC in DVSM mode. On a conventional SAP, ingress packets are classified and mapped to an FC. That FC mapping is maintained (as part of the fabric header) when the packet transits through the system and is ultimately used to define the egress queue and egress marking, such as MPLS EXP bits or dot1p bits.

However, the PXC sub-ports are subtly different. Consider SAP ingress traffic entering the VPRN at PE-7 from the locally connected test port C destined toward test port B at CE-4. At the ingress to PE-7, this traffic is mapped to FC Expedited Forwarding (EF) and forwarded into the PXC port through SAP pxc-1.b:100. When the traffic is forwarded out of the (PXC) SAP, the fabric header is removed as if it were a conventional SAP, and therefore, the information conveying the FC mapping is lost. When the traffic arrives at the opposing PXC sub-port SAP (in this case, pxc-1.a:100), a further FC classification is undertaken, and without some non-default configuration, traffic will be classified as FC Best Effort (BE). Therefore, it is a requirement to use non-default ingress/egress QoS policies through the PXC port in order to maintain FC continuity. A relatively simple way to do to this is through the use of dot1p markings.

To illustrate how this FC continuity is achieved, and in general how QoS is applied to PXC ports, an example of the relevant policies applied to PE-7's egress traffic toward CE-4 is used.

The first of the following outputs provides an example of the SAP-egress QoS policy applied at the VPRN PXC SAP (pxc-1.b:100). There are three classes in use: BE, Assured-Forwarding (AF), and EF. These FCs are remapped to queues 1, 2, and 3, respectively, and each queue is mapped to a parent H-QoS scheduler. Because the FCs must be maintained through the PXC loop, dot1p markings are used to distinguish between them. FC EF uses dot1p 5, FC AF uses dot1p 3, and FC BE uses dot1p 1. The SAP egress QoS policy is configured on PE-7 as follows:

```
configure
    qos
        sap-egress 2 create
            queue 1 create
                parent "aggregate-rate" level 2 weight 10
            exit
            queue 2 best-effort create
                parent "aggregate-rate" level 2 weight 40 cir-level 2
                rate 5000 cir max
            exit
            queue 3 expedite create
                parent "aggregate-rate" cir-level 3
                rate 2000 cir 2000
            exit
```

```
                fc af create
                    queue 2
                    dot1p 3
                exit
                fc be create
                    queue 1
                    dot1p 1
                exit
                fc ef create
                    queue 3
                    dot1p 5
                exit
            exit
```

The configuration of the Tier 1 scheduler "aggregate-rate" referenced by the child queues in the preceding SAP-egress QoS policy is shown in the following output. The scheduler in turn references a **port-scheduler-policy** using the command **port-parent**. Parenting to a port-scheduler is optional, but allows for inclusion of Preamble and Inter-Frame Gap (IFG) in the QoS scheduling algorithm, which is otherwise not included by a conventional H-QoS scheduler. The **port-scheduler-policy** "port-scheduler" is not referenced directly by the Tier 1 scheduler, but rather the port-scheduler is inherited by any child queues on the port to which the port-scheduler is applied. In this case, the **port-scheduler-policy** "port-scheduler" is applied to the PXC sub-port pxc-1.b as follows:

```
PE-7
configure
    qos
        port-scheduler-policy "port-scheduler" create
        exit
        scheduler-policy "egress-hqos-scheduler" create
            tier 1
                scheduler "aggregate-rate" create
                    port-parent
                    rate 1
                exit
            exit
        exit
    exit
    port pxc-1.b
        ethernet
            egress-scheduler-policy "port-scheduler"
        exit
        no shutdown
    exit
```

Finally, the SAP-egress QoS policy is applied to the PXC sub-port SAP within the VPRN interface context. The H-QoS scheduler is also attached and an override of the rate configured. In summary, the SAP-egress QoS policy configuration looks exactly like that used on a conventional SAP, other than the dot1p markings used for FC continuity, which may not always be used or required.

```
PE-7
```

```
configure
    service
        vprn 10 customer 1 create
            interface "to-CE-4" create
                address 192.168.100.1/30
                sap pxc-1.b:100 create
                    egress
                        scheduler-policy "egress-hqos-scheduler"
                        scheduler-override
                            scheduler "aggregate-rate" create
                                rate 20000
                            exit
                        exit
                        qos 2
                    exit
                exit
            exit
```

On the opposing side of the PXC loop, the dot1p markings imposed by the VPRN SAP egress are used to reclassify traffic back to its original FC mapping. The following output shows the SAP-ingress QoS policy applied at the Epipe PXC sub-port SAP (pxc-1.a:100). As shown in this output, dot1p 5 is mapped to FC EF, dot1p 3 is mapped to FC AF, and dot1p 1 is mapped to FC BE, thereby retaining the FC mappings through the PXC port.

```
PE-7
configure
    qos
        sap-ingress 11 create
            queue 1 create
            exit
            queue 2 best-effort create
                rate max cir max
            exit
            queue 3 expedite create
                rate max cir max
            exit
            fc "af" create
                queue 2
            exit
            fc "be" create
                queue 1
            exit
            fc "ef" create
                queue 3
            exit
            dot1p 1 fc "be"
            dot1p 3 fc "af"
            dot1p 5 fc "ef"
        exit
```

```
PE-7
configure
    service
        epipe 11 customer 1 create
            sap pxc-1.a:100 create
```

```
                ingress
                    qos 11
                exit
                no shutdown
            exit
        exit
```

The previous configuration show the required QoS policies for downstream traffic (VPRN egress to Epipe ingress). Corresponding QoS policies must also be configured for upstream traffic (Epipe egress to VPRN ingress). For brevity, they are not shown here.

AS Mode

AS mode creates an FPE context that is used to provide information to the system about which PXC ports or LAGs are paired, so that the configuration process can be simplified by automatic provisioning of cross-connects. To illustrate the use of AS mode, the redundant PXC (formed of LAG 1 and 2) configured earlier in this chapter is used. However, redundancy is not a requirement. Non-redundant PXC ports can also be used with AS mode.

For AS mode, a similar setup to the DVSM example is used, with Epipe termination into a VPRN. This provides a generic view of the applicability of AS mode, but also allows a direct comparison between the DVSM and AS mode approaches. Again, PE-4 in Figure 38 functions as a Layer 2 backhaul device and PE-7 hosts the PXC functions as the Layer 3 service edge. A pseudowire is extended from PE-4 to PE-7 where it will be terminated in a VPRN, providing point-to-point connectivity between CE-4 and PE-7.

The following output illustrates the configuration of the Epipe service at PE-4. CE-4 uses Q-in-Q encapsulation on the PE-CE link to PE-4 with SVLAN tag 100 and CVLAN tag 1024. At PE-4, it is indexed into an Epipe service using a q.* SAP to make the CVLAN tag transparent (part of the payload). As the spoke-SDP toward PE-7 is also configured with **force-vlan-vc-forwarding**, both SVLAN and CVLAN tags will be encapsulated in the pseudowire payload.

```
PE-4
configure
    service
        epipe 13 customer 1 create
            sap 1/1/4:100.* create
                no shutdown
            exit
            spoke-sdp 2007:13 create
                force-vlan-vc-forwarding
                no shutdown
            exit
            no shutdown
        exit
```

As in the previous configuration example, LAG 1 and LAG 2 are used for PXC redundancy. LAG 1 has the PXC sub-ports pxc-2.a and pxc-3.a as member links, while LAG 2 has the PXC sub-ports pxc-2.b and pxc-3.b as member links. For AS mode, the next requirement is to configure the FPE construct and assign the paired LAG instances to that FPE. When entering the **fwd-path-ext** context, the **sdp-id-range** must be configured before any **fpe** instances can be created. The **sdp-id-range** allocates a block of SDP identifiers to be used for the automatic cross-connects between service applications and the FPE. Up to 128 SDP identifiers can be allocated in the range 1 to 17407.

After the **sdp-id-range** is configured, the **fpe** instance is created and the user enters the **fpe** context. The **path** command is used to assign redundant or non-redundant PXC objects to the FPE. In the case of a non-redundant FPE, the **path** command would refer to a **pxc** instance. In the case of a redundant FPE, the **path** syntax requires that each of the paired LAG instances is assigned to cross-connect "a" or cross-connect "b". Each FPE has two fundamental components, known as the transit side and the terminating side. The transit side is the side where additional traffic preprocessing is carried out, such as header removal or manipulation. It can be considered as the side closest to the network. The terminating side is the side where the preprocessed traffic is terminated in a service. When an FPE is used, the system automatically assigns cross-connect "a" to the transit side, and cross-connect "b" to the terminating side. In the following example, the command **path xc-a lag-1 xc-b lag-2** assigns LAG 1 to cross-connect "a" and LAG 2 to cross-connect "b". This means that LAG 1 is the transit side while LAG 2 is the terminating side.

The application of the FPE also needs to be configured. In this example, **pw-port** is selected to allow for support of pseudowire-SAP (including Enhanced Subscriber Management (ESM) over pseudowire). The other available option is **vxlan-termination**, which is beyond the scope of this chapter.

```
PE-7
configure
    fwd-path-ext
        sdp-id-range from 17280 to 17407
        fpe 1 create
            path xc-a lag-1 xc-b lag-2
            pw-port
        exit
    exit
```

After the LAG instance is assigned to the FPE, it can no longer be used for other general purposes, such as IP interfaces and/or SAPs. Any attempt to do so is blocked in CLI. The operational state of the FPE can be verified as shown in the following output. It is also useful to be able to identify the services and pw-ports that are mapped to an FPE. This can be obtained using the "show fwd-path-ext fpe <number> associations" command.

```
*A:PE-7# show fwd-path-ext fpe 1

===============================================================================
FPE Id: 1
===============================================================================
Description        : (Not Specified)
Path               : lag 1, lag 2
Pw Port            : Enabled                           Oper     : up
Vxlan Termination  : Disabled                          Oper     : down
===============================================================================
```

The next step is to configure a pseudowire-port (pw-port) that will be used for terminating services. The creation of the **pw-port** creates a new context in which the only required configuration is to define the encapsulation type as dot1q or qinq. In this instance, the **pw-port** will support **encap-type qinq**.

```
PE-7
configure
    pw-port 1 create
        encap-type qinq
    exit
```

The operational state of the pw-port is captured as a reference at this point, so that a comparison can be made later in the configuration process.

```
*A:PE-7# show pw-port 1

===================================================================
PW Port Information
===================================================================
PW Port   Encap          SDP        IfIndex            VC-Id
-------------------------------------------------------------------
1         qinq                      1526726657
===================================================================
```

At PE-7, the requirement now is to link the spoke-SDP from PE-4 to the configured pw-port (pw-port 1) via the FPE. To do this, an Epipe service must be used that is configured for multi-segment pseudowire working, using the creation-time attribute **vc-switching**. The Epipe service consists of a single spoke-SDP toward PE-4 with a VC-ID matching that signaled by PE-4 (VC-ID 13). The second endpoint within the Epipe service uses the command **pw-port 1 fpe 1** to reference the previously configured pw-port and FPE objects. This command essentially creates an internal cross-connect between the Epipe service and the pw-port via the configured FPE object.

```
PE-7
configure
    service
        epipe 13 customer 1 vc-switching create
            pw-port 1 fpe 1 create
                no shutdown
            exit
            spoke-sdp 2004:13 create
```

```
                no shutdown
            exit
            no shutdown
        exit
```

The following output shows the SDPs belonging to the preceding vc-switched Epipe service configured. The first SDP with identifier 2004:13 is the pseudowire toward PE-4 with VC-ID 13. The second SDP has identifier 17280:1 allocated from the preconfigured **sdp-id-range**, and has a type of Fpe. In the configuration of **fpe 1**, the **path** command assigned LAG 1 to cross-connect "a" (**xc-a**) and LAG 2 to cross-connect "b" (**xc-b**). Also, cross-connect "b" is always automatically assigned to the terminate side of the FPE. Therefore, the Far End address is shown as fpe_1.b, in order to terminate the service.

```
*A:PE-7# show service id 13 sdp

===============================================================================
Services: Service Destination Points
===============================================================================
SdpId            Type     Far End addr    Adm     Opr       I.Lbl     E.Lbl
-------------------------------------------------------------------------------
2004:13          Spok     192.0.2.4       Up      Up        262126    262122
17280:1          Fpe      fpe_1.b         Up      Up        262128    262127
-------------------------------------------------------------------------------
Number of SDPs : 2
-------------------------------------------------------------------------------
===============================================================================
```

With the vc-switching Epipe service configured and operational, the state of the pw-port can again be shown in the following output. Before the configuration of the vc-switching Epipe, the pw-port had no SDP identifier or VC-ID. Now both entries exist; automatically created by the system when p**w-port 1 fpe 1** was configured as an endpoint within the vc-switching Epipe. The SDP identifier of 17281 is allocated from the preconfigured **sdp-id-range**.

```
*A:PE-7# show pw-port 1

===================================================================
PW Port Information
===================================================================
PW Port   Encap          SDP        IfIndex             VC-Id
-------------------------------------------------------------------
1         qinq           17281      1526726657          100001
===================================================================
```

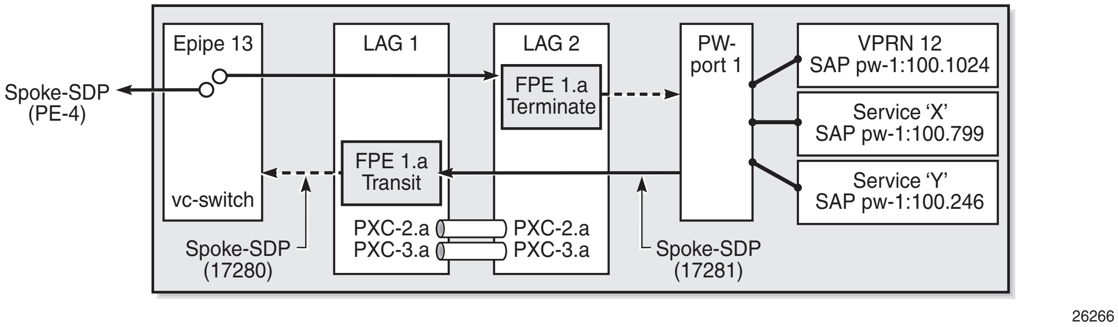
The output for SDP 17281 shows that the Far End is fpe_1.a (transit), the Delivery (Del) is MPLS, the LSP type is FPE (F), and that no signaling (Sig) is used for this internal SDP, as follows:

```
*A:PE-7# show service sdp 17281

============================================================================
Service Destination Point (Sdp Id : 17281)
=============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
-----------------------------------------------------------------------------
17281  0       8678    fpe_1.a          Up   Up          MPLS    F     None
=============================================================================
```

In SR OS, the combination of SDP ID and VC-ID is always associated with a service. When using AS mode, the system automatically creates an internal VPLS service with ID 2147383649 and a name of _tmns_InternalVplsService. This VPLS includes all internal SDPs dynamically created for binding pw-ports to the transit side of the corresponding FPE. The VPLS is an internal construct that does not affect forwarding.

Figure 41 shows the components of the FPE from vc-switching Epipe to pw-port.

Figure 41 **AS mode with Redundant FPE**



Next, bind the VPRN service to the pw-port with the relevant VLAN delimiters. CE-4 is using SVLAN tag 100 and CVLAN tag 1024 and both VLANs are encapsulated inside the pseudowire as payload. The following VPRN configuration has two interfaces: the first is toward a directly connected test port used to verify IP connectivity, and the second is toward CE-4 and has a SAP with a pw-port syntax. The SAP pw-1:100.1024 represents pw-port 1 with Q-in-Q encapsulation using SVLAN tag 100 and CVLAN tag 1024 as service delimiters. Although not shown (for brevity), a BGP session is configured between PE-7 and CE-4 for route exchange. The remainder of the VPRN parameters are generic and are not explained here.

```
PE-7
configure
    service
        vprn 12 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:12
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:12
            interface "Test-Port-C" create
```

```
                address 172.31.107.1/24
                sap 1/1/3:100 create
                exit
            exit
            interface "to-CE-4" create
                address 192.168.100.1/30
                sap pw-1:100.1024 create
                exit
            exit
            bgp
                group "EBGP"
            ---snip---
            no shutdown
            exit
        exit
```

FPE Port Dimensioning

After the VPRN service at PE-7 is put into a **no shutdown** state, the EBGP session to CE-4 is established. The relevant routes are exchanged between CE-4 and PE-7 and traffic can be exchanged between test ports B (behind CE-4) and C (connected to PE-7). Initially, traffic is sent unidirectionally downstream from test port C (connected to PE-7) toward port B (connected to CE-4) at a rate of 100 packets/s. To provide a level of entropy for the generated traffic, 100 destination IP addresses are used in the range 172.31.104.2 through 172.31.104.101, and 100 source IP addresses are used in the range 172.31.107.2 through 172.31.107.101.

The following output shows a snapshot of a monitor command against LAG 2 (xc-b, or terminating side) incorporating both physical ports. First, note that the input and output rate of packets per second are equal at 100 packets/s, which is not intuitive for a unidirectional traffic flow. This is because the LAG statistics are essentially a copy of the physical port statistics and the physical port consists of two PXC sub-ports that are looped. Logically, this unidirectional traffic flow is forwarded in a single upstream direction from pxc-2.a/pxc-3.a to pxc-2.b/pxc-3.b. Physically, the unidirectional traffic is transmitted by ports 1/2/2 and 1/2/3, then received by the same ports through the loop. Second, note that traffic is load-balanced over both member links (PXC sub-ports) of the LAG. This is because conventional LAG load-balancing mechanisms are used for the FPE LAG, which in the case of a VPRN SAP-to-network relies on source/destination IP address (with optional Layer 4, which is not currently configured).

```
*A:PE-7# monitor lag 2 rate interval 3

===============================================================================
Monitor statistics for LAG ID 2
===============================================================================
Port-id    Input       Input       Output      Output      Input      Output
           Bytes       Packets     Bytes       Packets     Errors     Errors
-------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
At time t = 18 sec (Mode: Rate)
-------------------------------------------------------------------------------
1/2/2!    22041       41          22041       41          0          0
  % Util  ~0.00       --          ~0.00       --          --         --
1/2/3!    32159       59          32159       59          0          0
  % Util  ~0.00       --          ~0.00       --          --         --
-------------------------------------------------------------------------------
Totals    54200       100         54200       100         0          0
  % Util  ~0.00       --          ~0.00       --          --         --
! indicates that the port is assigned to a port-xc.
```

Traffic is then generated unidirectionally upstream from test port B (connected to CE4) toward port C (connected to PE7) at a rate of 100 packets/s. Again, to provide a level of entropy for the generated traffic, 100 destination IP addresses are used in the range 172.31.107.2 through 172.31.107.101, and 100 source IP addresses are used in the range 172.31.104.2 through 172.31.104.101. The input/output rates of packets per second are the same, as previously explained. Again, traffic is load-balanced over both member links (PXC sub-ports). This is because hashing of traffic through a vc-switched Epipe service uses source/destination IP information (and optional Layer 4 information, which is not currently configured).

```
*A:PE-7# monitor lag 2 rate interval 3

===============================================================================
Monitor statistics for LAG ID 2
===============================================================================
Port-id   Input       Input       Output      Output      Input      Output
          Bytes       Packets     Bytes       Packets     Errors     Errors
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
At time t = 9 sec (Mode: Rate)
-------------------------------------------------------------------------------
1/2/2!    23848       44          23848       44          0          0
  % Util  ~0.00       --          ~0.00       --          --         --
1/2/3!    30352       56          30352       56          0          0
  % Util  ~0.00       --          ~0.00       --          --         --
-------------------------------------------------------------------------------
Totals    54200       100         54200       100         0          0
  % Util  ~0.00       --          ~0.00       --          --         --
! indicates that the port is assigned to a port-xc.
```

QoS Continuity

When using AS mode, the FPE construct creates internal cross-connects between the vc-switching Epipe and the pw-port. These internal cross-connects function as MPLS tunnels that transit through internal network interfaces on the PXC sub-ports. The internal network interfaces use the default network policy 1 for egress marking and ingress classification/FC mapping. Like all default QoS policies, this network policy cannot be modified (or deleted). Also, it is not possible to use a non-default network policy, because there is no router interface to which the non-default policy can be attached.

The internal cross-connects also use the default network-queue policy named "default". While this policy also cannot be modified, it is possible to configure and apply a non-default network-queue policy (including a port-scheduler policy, if required) at PXC sub-port level. An example of how this would be applied is shown in the following output. Where redundant PXC ports are used in an LAG instance, the queue-policy must be applied to the primary link of the LAG, which is then automatically applied to all other member links. (The primary link of the LAG can be identified using the command "show lag n port".)

```
configure
    port pxc-2.a
        ethernet
            network
                queue-policy "non-default"
            exit
        exit
        no shutdown
    exit
```

To demonstrate QoS continuity through the FPE, the following is established:

- **Downstream**: Traffic is generated from test port C (connected to PE-7) toward test port B (connected to CE-4) with DiffServ marking EF at a rate of 100 packets/s. At PE-7 SAP ingress, this traffic is mapped into FC EF.
- **Upstream**: Traffic is generated from test port B (connected to CE-4) toward test port C (connected to PE-7) with DiffServ marking EF at a rate of 100 packets/s. At PE-4, a SAP-ingress QoS policy is used to map the traffic into FC EF.
- The default network QoS policy 1 is used on all network interfaces at PE-4 and PE-7. On egress, this policy marks FC EF as MPLS EXP 5. On ingress, MPLS EXP 5 is mapped to FC EF.
- The default network queue-policy "default" is used on all network interfaces at PE-4 and PE-7. This maps FC EF traffic to queue 6 at ingress and egress.

First, QoS continuity for downstream traffic is validated. The following output shows the relatively simple SAP-egress QoS policy that is applied to the egress of the VPRN interface (pw-port) toward CE-4. No classification of traffic and mapping to FCs are present in the policy, because the classification and mapping have already taken place on the SAP ingress at PE-7 (the SAP facing the test port C).

```
PE-7
configure
    qos
        sap-egress 12 create
            queue 1 create
                parent "aggregate-rate" level 2 weight 10
            exit
            queue 2 best-effort create
                parent "aggregate-rate" level 2 weight 40 cir-level 2
                rate 5000 cir max
            exit
            queue 3 expedite create
                parent "aggregate-rate" cir-level 3
                rate 2000 cir 2000
            exit
            fc af create
                queue 2
            exit
            fc be create
                queue 1
            exit
            fc ef create
                queue 3
            exit
        exit
```

The configuration of the Tier 1 scheduler "aggregate-rate" referenced by the child queues in the preceding SAP-egress QoS policy is as follows. The Tier 1 scheduler references a **port-scheduler policy** using the command **port-parent**. Parenting to a port-scheduler is optional, but allows for inclusion of Preamble and IFG in the QoS scheduling algorithm, which otherwise are not included. The Tier 1 scheduler does not directly reference the **port-scheduler-policy** by name, but rather inherits any port-scheduler configured on the port to which the child queues are mapped. In this example, the port-scheduler-policy "port-scheduler" is applied to PXC sub-port pxc-2.b (terminating side). This is the primary link of LAG 2 and ensures that the same port-scheduler policy is automatically applied to other member ports.

```
PE-7
configure
    qos
        port-scheduler-policy "port-scheduler" create
        exit
        scheduler-policy "egress-hqos-scheduler" create
            tier 1
                scheduler "aggregate-rate" create
                    port-parent
                    rate 1
                exit
            exit
       exit
    exit
    port pxc-2.b
        ethernet
            egress-scheduler-policy "port-scheduler"
        exit
        no shutdown
    exit
```

Finally, the SAP-egress QoS policy is applied to the pw-port SAP within the VPRN. The egress H-QoS scheduler is also attached and an override of the rate is configured.

```
PE-7
configure
    service
        vprn 12 customer 1 create
            interface "to-CE-4" create
                sap pw-1:100.1024 create
                    egress
                        scheduler-policy "egress-hqos-scheduler"
                        scheduler-override
                            scheduler "aggregate-rate" create
                                rate 25000
                            exit
                        exit
                        qos 12
                    exit
                exit
            exit
```

When traffic is generated downstream toward CE-4 in FC EF at a rate of 100 packets/s, the first point of verification is the VPRN pw-port SAP egress. The following output is a **monitor** of the SAP showing that traffic is correctly mapped to queue 3.

```
*A:PE-7# monitor service id 12 sap pw-1:100.1024 rate

===============================================================================
Monitor statistics for Service 12 SAP pw-1:100.1024
===============================================================================
-------------------------------------------------------------------------------
At time t = 11 sec (Mode: Rate)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap per Queue Stats
-------------------------------------------------------------------------------
                        Packets                 Octets                  % Port
                                                                        Util.
---snip---
Egress Queue 3
For. In/InplusProf    : 0                       0                       0.00
For. Out/ExcProf      : 100                     51600                   0.04
Dro. In/InplusProf    : 0                       0                       0.00
Dro. Out/ExcProf      : 0                       0                       0.00
```

Monitoring of network interfaces does not show queue statistics (and is not supported on PXC sub-ports), but a verification of the sub-port statistics on the transit side (LAG 1) shows that packets are incrementing in ingress queue 6 on both sub-ports, as follows:

```
*A:PE-7# show port pxc-2.a detail | match "Ingress Queue  6" post-lines 4
Ingress Queue  6             Packets                Octets
     In Profile forwarded  :    711                     382518
     In Profile dropped    :    0                       0
     Out Profile forwarded :    0                       0
     Out Profile dropped   :    0                       0
*A:PE-7# show port pxc-3.a detail | match "Ingress Queue  6" post-lines 4
Ingress Queue  6             Packets                Octets
     In Profile forwarded  :    404                     217352
     In Profile dropped    :    0                       0
     Out Profile forwarded :    0                       0
     Out Profile dropped   :    0                       0
```

The last point of verification is the network egress interface toward PE-4. Again, a check at the physical port level shows that packets are incrementing in egress queue 6. Therefore, we can conclude that QoS/FC continuity is maintained in the downstream direction.

```
*A:PE-7# show port 1/1/1 detail | match "Egress Queue  6" post-lines 4
Egress Queue  6              Packets                Octets
     In/Inplus Prof fwded  :    2394                    1297548
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    0                       0
     Out/Exc Prof dropped  :    0                       0
```

Next, the upstream QoS continuity is verified. PE-4 is marking traffic generated by test port B to FC EF, which in turn is marked as MPLS EXP 5 by PE-4's default network QoS policy. The following output taken at PE-7 shows that packets are incrementing in ingress queue 6 of the network interface toward PE-4 and confirms that traffic is correctly marked as FC EF at ingress.

```
*A:PE-7# show port 1/1/1 detail | match "Ingress Queue  6" post-lines 4
Ingress Queue  6             Packets                Octets
     In Profile forwarded  :    3458                    1874236
     In Profile dropped    :    0                       0
     Out Profile forwarded :    0                       0
     Out Profile dropped   :    0                       0
```

The next point of verification is the egress side of the PXC sub-ports (pxc-2.a and pxc-3.a) forming the transit side (LAG 1). The sub-port statistics verify that packets are incrementing in egress queue 6 of both sub-ports (as traffic is being load-balanced).

```
*A:PE-7# show port pxc-2.a detail | match "Egress Queue  6" post-lines 4
Egress Queue  6              Packets                Octets
     In/Inplus Prof fwded  :    12441                   6693258
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    0                       0
     Out/Exc Prof dropped  :    0                       0
*A:PE-7# show port pxc-3.a detail | match "Egress Queue  6" post-lines 4
Egress Queue  6              Packets                 Octets
     In/Inplus Prof fwded  :    12893                   6936434
     In/Inplus Prof dropped:    0                       0
     Out/Exc Prof fwded    :    0                       0
     Out/Exc Prof dropped  :    0                       0
```

PXC sub-ports operate in hybrid mode. When the upstream traffic arrives on the PXC sub-ports that form the terminating side of the FPE (pxc-2.b and pxc-3.b), it is mapped to the pw-port SAP-ingress queues, bypassing the ingress network QoS policy and associated ingress network queues. As a result, the MPLS EXP-to-FC mapping cannot be fulfilled and traffic requires reclassification and remapping to the correct FC by the SAP-ingress QoS policy. The following output shows the SAP-ingress QoS policy applied to the pw-port SAP within the VPRN. Because the EXP-to-FC mapping could not be completed, FC reclassification is required in order to map traffic to its original FC before transiting the FPE. In this example, DSCP is used. Also, FC EF is mapped to queue 3.

```
PE-7
configure
    qos
      sap-ingress 12 create
            queue 1 create
                parent "aggregate-rate" level 2 weight 10
            exit
            queue 2 best-effort create
                parent "aggregate-rate" level 2 weight 40 cir-level 2
                rate 5000 cir max
            exit
            queue 3 expedite create
                parent "aggregate-rate" cir-level 3
                rate 2000 cir 2000
            exit
            queue 11 multipoint create
                rate max cir max
            exit
            fc "af" create
                queue 2
            exit
            fc "be" create
                queue 1
            exit
            fc "ef" create
                queue 3
            exit
            dscp af31 fc "af"
            dscp be fc "be"
            dscp ef fc "ef"
        exit
```

For completeness, the configuration of the Tier 1 scheduler "aggregate-rate" referenced by the child queues in the preceding SAP-ingress QoS policy is as follows. Unlike the egress counterpart, there is no parenting to a port-scheduler because this is an egress function only.

```
PE-7
configure
    qos
        scheduler-policy "ingress-hqos-scheduler" create
            tier 1
                scheduler "aggregate-rate" create
                    rate 1
                exit
            exit
        exit
    exit
```

The SAP-ingress QoS policy is applied to the pw-port SAP within the VPRN, together with the ingress H-QoS scheduler. An override of the scheduler rate is also applied.

```
PE-7
configure
    service
        vprn 12 customer 1 create
            interface "to-CE-4" create
                sap pw-1:100.1024 create
                    ingress
                        scheduler-policy "ingress-hqos-scheduler"
                        scheduler-override
                            scheduler "aggregate-rate" create
                                rate 25000
                            exit
                        exit
                        qos 12
                    exit
                exit
            exit
```

With the SAP-ingress policy applied, a monitor output of the SAP in the following output verifies that the packets are being received in queue 3 at a rate of 100 packets/s. This verifies the FC continuity in the upstream direction, noting that reclassification and remapping of FC is required at SAP ingress.

```
*A:PE-7# monitor service id 12 sap pw-1:100.1024 rate

===============================================================================
Monitor statistics for Service 12 SAP pw-1:100.1024
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time      : 11/07/2016 15:32:26
```

```
                        Packets                 Octets
---snip---

Ingress Queue 3 (Unicast) (Priority)
Off. HiPrio          : 0                      0                      0.00
Off. LowPrio         : 100                    51647                  0.04
Dro. HiPrio          : 0                      0                      0.00
Dro. LowPrio         : 0                      0                      0.00
For. InProf          : 0                      0                      0.00
For. OutProf         : 100                    51647                  0.04
```

OAM Continuity

The FPE pw-port functionality may be used by redundant routers to provide resilient service termination for a Layer 2 backhaul node implementing a mechanism such as active/standby pseudowire. In SR OS, an active/standby pseudowire is modeled as an Epipe or VPLS service with an endpoint object containing two spoke-SDPs. This form of redundancy relies on the propagation of the Pseudowire Status TLV within an LDP Notification message to convey the operational status of the pseudowires and thereby indicate which one of the pseudowires is active and which one is standby.

The FPE construct uses the concept of a multi-segment pseudowire, implementing Switching-PE (S-PE) functionality to instantiate dynamic cross-connects through the FPE. To verify that LDP status signaling is maintained through this S-PE function, the following is established:

- The Epipe service at PE-4 used for Layer 2 backhaul to the FPE is modified to include an **endpoint** object referenced by two spoke-SDPs.
- The first spoke-SDP has a far end of PE-2 and is configured as **precedence primary**, so becomes the active pseudowire.
- The second spoke-SDP has a far end of PE-7 and is configured with the default precedence 4, so becomes the standby pseudowire.
- Because the endpoint object is configured for **standby-signaling-master**, PE-4 will signal a status of standby toward PE-7.

For completeness, the configuration of the Epipe service at PE-4 is as follows:

```
configure
    service
       epipe 13 customer 1 create
            endpoint "redundant-Layer3" create
                standby-signaling-master
            exit
            sap 1/1/4:100.* create
                no shutdown
            exit
            spoke-sdp 2002:13 endpoint "redundant-Layer3" create
                precedence primary
                no shutdown
            exit
            spoke-sdp 2007:13 endpoint "redundant-Layer3" create
                no shutdown
            exit
            no shutdown
        exit
```

As shown in the following output, PE-4 has the spoke-SDP to PE-7 (sdp 2007:13) as administratively and operationally up, but is signaling a status of standby (pwFwdingStandby).

```
*A:PE-
4# show service id 13 sdp 2007:13 detail | match expression "Local Pw Bits|Peer Pw B
its|Admin State"
Admin State         : Up                          Oper State        : Up
Local Pw Bits       : pwFwdingStandby
Peer Pw Bits        : None
```

At PE-7, the signaled status is acknowledged at the far end of the pseudowire in the Peer Pw Bits field.

```
*A:PE-
7# show service id 13 sdp 2004:13 detail | match expression "Admin State|Local Pw Bi
ts|Peer Pw Bits"
Admin State         : Up                          Oper State        : Up
Local Pw Bits       : None
Peer Pw Bits        : pwFwdingStandby
```

Typically, an S-PE would propagate the status TLV received from one pseudowire segment into the opposing pseudowire segment in order to provide end-to-end status signaling. However, when using FPE, the SR OS Service Manager process correlates between a pseudowire and its corresponding pw-port SAPs, so can take the necessary actions based upon the operational state of each. Therefore, it is not necessary for the S-PE to propagate the status TLV from one segment to another. This is illustrated in the following output at PE-7, which shows the second segment of the multi-segment pseudowire toward the terminating side fpe_1.b. As described, the status bits are not copied between single segments and all local/peer pseudowire bits remain unset.

```
*A:PE-
7# show service id 13 sdp 17280:1 detail | match expression "Admin State|Peer Pw Bit
s|Local Pw Bits"
Admin State         : Up                          Oper State        : Up
Local Pw Bits       : None
Peer Pw Bits        : None
```

The pw-port 1 used throughout in this example is internally bound to SDP 17281, as shown in the first of the following outputs. The second output shows that this SDP is operationally down with the flag "stitchingSvcTxDown".

```
*A:PE-7# show pw-port 1

===================================================================
PW Port Information
===================================================================
PW Port   Encap          SDP        IfIndex             VC-Id
-------------------------------------------------------------------
1         qinq           17281      1526726657          100001
===================================================================
```

```
*A:PE-7# show service sdp 17281 detail | match "SDP: 17281 Pw-port: 1" post-lines 8
SDP: 17281 Pw-port: 1
-------------------------------------------------------------------------------
VC-Id                : 100001              Admin Status       : up
Encap                : qinq                Oper Status        : down
VC Type              : ether

Admin Ingress label  : 262127              Admin Egress label : 262128
Oper Flags           : stitchingSvcTxDown
Monitor Oper-Group   : (Not Specified)
```

At service level, the first of the following two outputs shows the state of the SAP bound to pw-port 1. As shown, the operational state is down with an indication that this is due to the port being operationally down. The second output shows that this SAP status is propagated to IP interface level because the interface "to-CE-4" is also shown as operationally down.

```
*A:PE-7# show service id 12 sap pw-
1:100.1024 detail | match expression "Admin State|Flags"
Admin State        : Up                       Oper State        : Down
Flags              : PortOperDown
```

```
*A:PE-7# show router 12 interface "to-CE-4"

===============================================================================
Interface Table (Service: 12)
===============================================================================
Interface-Name                   Adm       Opr(v4/v6)  Mode    Port/SapId
   IP-Address                                                  PfxState
-------------------------------------------------------------------------------
to-CE-4                           Up        Down/Down   VPRN    pw-1:100.1024
   192.168.100.1/30                                            n/a
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
```

To verify a failover, the state of the active/standby pseudowire is transitioned by failing the active pseudowire between PE-4 and PE-2. This causes PE-4 to declare the pseudowire to PE-7 active, which clears the standby status bits. This action causes the SDP (17281) bound to pw-port 1 to become operationally up, followed by pw-port 1 and its associated SAPs, followed by the VPRN IP interface "to-CE-4".

```
11 2016/11/10 16:48:33.98 UTC MINOR: SVCMGR #2306 Base
"Status of SDP Bind 17281:100001 in service 2147483649 (customer 1) changed to
admin=up oper=up flags="


12 2016/11/10 16:48:33.98 UTC MINOR: SVCMGR #2313 Base
"Status of SDP Bind 2004:13 in service 13 (customer 1) peer PW status bits changed
to none"

13 2016/11/10 16:48:33.98 UTC MAJOR: SVCMGR #2210 Base
"Processing of an access port state change event is finished and the status of all
affected SAPs on port pw-1 has been updated."

14 2016/11/10 16:48:33.98 UTC WARNING: SNMP #2005 vprn12 to-CE-4
"Interface to-CE-4 is operational"
```

This example in the AS mode section illustrated how notification of a downstream failure is propagated through the components of the PXC in AS mode and reflected in the status of the pw-port (and its associated services). Also, if a pw-port fails due to a PXC failure (for example, the physical port fails), it is just as important that the operational state is propagated externally. In the case of pseudowire backhaul (as in the example), this would be achieved by setting the LDP pseudowire status bits to psnIngressFault and psnEgressFault toward the far end.

Conclusion

This chapter demonstrates the principles of PXC configuration. The PXC can be used to provide a relatively simple back-to-back cross-connect operation in DVSM mode, or it can be used in AS mode to provide an integrated path through the FPE with automated cross-connects used to simplify the provisioning process. In both DVSM mode and AS mode, the PXC can be configured as redundant or non-redundant. A relatively simple use-case of terminating an Epipe into a VPRN has been demonstrated for both modes.

There are a large number of use-cases where frame/packet preprocessing is required before service termination. The workaround for these use-cases has previously been a physical external loop, but can now be resolved logically and internally through use of the PXC.

Router Configuration

In This Section

This section provides configuration information for the following topics:

6PE Next-Hop Resolution

This chapter provides information about 6PE Next-Hop Resolution.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

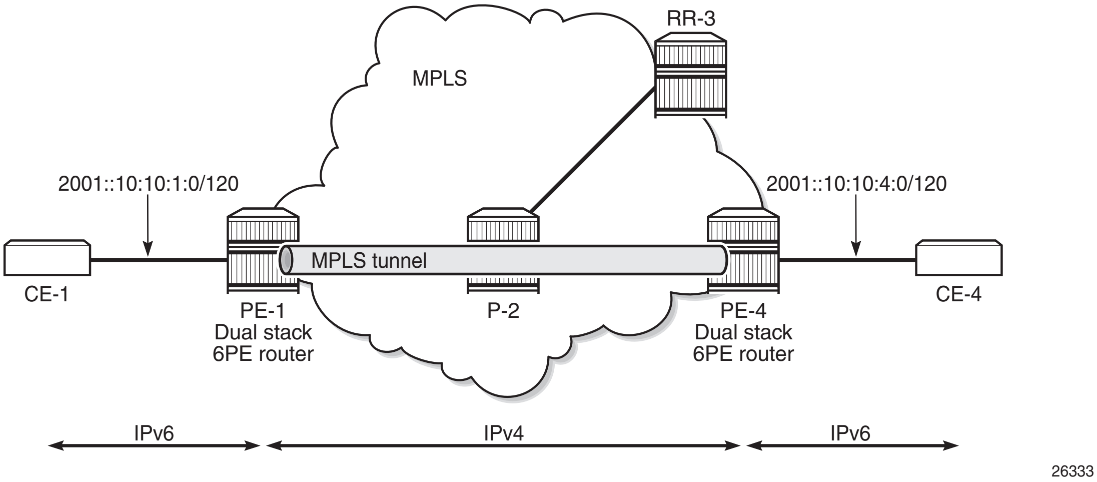
The information and configuration in this chapter are based on SR OS Release 14.0.R7. In releases earlier than 14.0.R1, only Label Distribution Protocol Label Switched Paths (LDP LSPs) could be used to resolve IPv6 Provider Edge (6PE) next hops. The following additional options for 6PE next-hop resolution are supported in SR OS Release 14.0.R1, and later:

- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) LSPs
- Segment Routing-Traffic Engineering (SR-TE)
- SR-OSPF
- SR-ISIS
- BGP labeled IPv4 routes

Overview

IPv6 Provider Edge (6PE) enables IPv6 communication between IPv6 domains over an IPv4 Multi-Protocol Label Switching (MPLS) cloud. IPv6 packets are forwarded in an MPLS tunnel from one dual-stack 6PE router to another, as shown in Figure 42.

Figure 42 **IPv6 Provider Edge (6PE)**



The 6PE route next-hop resolution is configured using the following command:

```
*A:PE-1# configure router bgp next-hop-resolution label-route-transport-
tunnel family label-ipv6 resolution
  - resolution {any|filter|disabled}
```

With 6PE next-hop resolution set to **any**, the tunnels are selected based on availability and preference. The 6PE next hops can be resolved to the following six options, as per the following order of preference:

1. RSVP-TE
2. SR-TE
3. LDP
4. SR-OSPF
5. SR-ISIS
6. BGP (labeled IPv4 routes)

For LDP to be used, it is sufficient to enable LDP on the interfaces in the MPLS network.

For RSVP-TE to be used, an RSVP-TE LSP to the 6PE next-hop destination must be available or configured. For segment routing to be used, an SR-signaled path to the 6PE next-hop destination must be available or configured. For BGP labeled routes to be used, the 6PE next hop must have been learned via a BGP peering carrying labeled unicast routes and placed in the active route table.

With 6PE next-hop resolution set to filter, a subset of protocols is required, and LDP is automatically added to the protocol list in the resolution filter. The following example shows that when one tries to create a resolution filter that includes the BGP protocol only, the resolution filter includes LDP and BGP. The first info command shows that initially no resolution filter had been defined.

```
*A:PE-1# configure router bgp next-hop-resolution label-route-transport-
tunnel family label-ipv6
*A:PE-1>config>router>bgp>next-hop-res>lbl-rt-tunn>family# info
----------------------------------------------
----------------------------------------------
*A:PE-1>config>router>bgp>next-hop-res>lbl-rt-tunn>family# resolution-filter bgp
*A:PE-1>config>router>bgp>next-hop-res>lbl-rt-tunn>family# info
----------------------------------------------
                    resolution-filter
                        ldp
                        bgp
                    exit
                    resolution filter
----------------------------------------------
```

If the 6PE next hop can be resolved to an LDP tunnel, this tunnel is preferred to a BGP tunnel.

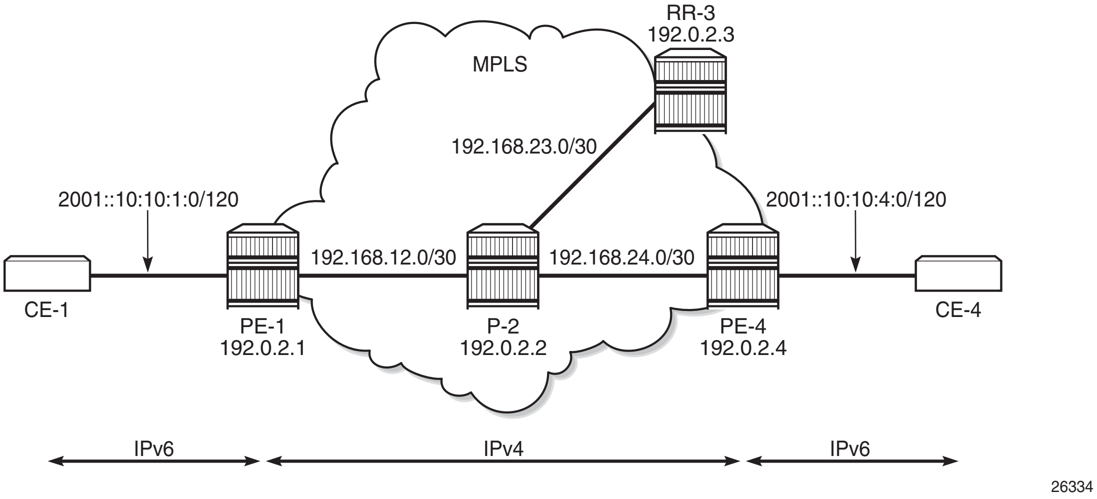
It is possible to explicitly exclude LDP from the list, as follows:

```
*A:PE-1>config>router>bgp>next-hop-res>lbl-rt-tunn>family# resolution-filter no ldp
*A:PE-1>config>router>bgp>next-hop-res>lbl-rt-tunn>family# info
----------------------------------------------
                    resolution-filter
                        no ldp
                        bgp
                    exit
                    resolution filter
----------------------------------------------
```

Configuration

Figure 43 shows the example topology with two dual-stack 6PE routers (PE-1 and PE-4), a core router (P-2), and a route reflector (RR-3). IPv4 is used in the core network; IPv6 is used between the CEs and the PEs.

Figure 43 **Example Topology**



The initial configuration on the nodes is as follows:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP in the core IPv4 network (alternatively, OSPF can be used)
- LDP enabled on the interfaces between the PEs and P-2, but not toward RR-3
- MPLS and RSVP enabled on the interfaces between the PEs and P-2, but not toward RR-3

BGP Configuration

BGP is configured for the label-ipv6 address family on PE-1, PE-4, and RR-3, but not on P-2. The BGP configuration on both PEs defines how the 6PE next hops will be resolved: the resolution filter contains three options (LDP, RSVP, and SR-ISIS). The BGP configuration is identical on PE-1 and PE-4.

```
configure
    router
        autonomous-system 64496
        bgp
            split-horizon
            next-hop-resolution
                label-route-transport-tunnel
                    family label-ipv6
                        resolution-filter
                            ldp
                            rsvp
                            sr-isis
                        exit
                        resolution filter
                    exit
                exit
            exit
            group "iBGP"
                export "export-6pe"
                peer-as 64496
                neighbor 192.0.2.3
                    family label-ipv6
                exit
            exit
```

The export policy "export-6pe" exports the IPv6 prefixes that are local to the PE, for example, on PE-1: 2001::10:10:1:0/120, and is defined as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "export-6pe"
                entry 10
                    from
                        protocol direct
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The BGP configuration on RR-3 does not include any export policy or any next-hop resolution settings, as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            split-horizon
            group "iBGP"
                cluster 3.3.3.3
                peer-as 64496
                neighbor 192.0.2.1
                    family label-ipv6
                exit
                neighbor 192.0.2.4
                    family label-ipv6
                exit
            exit
```

IES Configuration

On PE-1, an IES is configured with IPv6 addresses on the interface toward CE-1, as follows:

```
configure
    service
        ies 1 customer 1 create
            description "6PE"
            interface "int-PE-1-CE-1" create
                ipv6
                    address 2001::10:10:1:1/120
                exit
                sap 1/2/1:1 create
                exit
            exit
            no shutdown
```

The configuration on PE-4 is similar; the IPv6 address on interface "int-PE-4-CE-4" is different: 2001::10:10:4:1/120.

A BGP labeled tunnel, which is active in the routing table, is established between the PEs, as follows:

```
*A:PE-1# show router route-table ipv6

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001::10:10:1:0/120                           Local   Local     00h02m46s  0
       int-PE-1-CE-1                                               0
2001::10:10:4:0/120                           Remote  BGP_LABEL 00h02m15s  170
       192.0.2.4 (tunneled)                                         20
-------------------------------------------------------------------------------
No. of Routes: 2
```

CE-1 can send IPv6 packets with source address 2001::10:10:1:2 to destination address 2001::10:10:4:2 on CE-4, as follows:

```
*A:PE-1# ping router 10 2001::10:10:4:2 source 2001::10:10:1:2
PING 2001::10:10:4:2 56 data bytes
64 bytes from 2001::10:10:4:2 icmp_seq=1 hlim=62 time=1.41ms.
64 bytes from 2001::10:10:4:2 icmp_seq=2 hlim=62 time=1.33ms.
64 bytes from 2001::10:10:4:2 icmp_seq=3 hlim=62 time=1.39ms.
64 bytes from 2001::10:10:4:2 icmp_seq=4 hlim=62 time=1.49ms.
64 bytes from 2001::10:10:4:2 icmp_seq=5 hlim=62 time=1.30ms.
---- 2001::10:10:4:2 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 1.30ms, avg = 1.38ms, max = 1.49ms, stddev = 0.066ms
```

6PE Next Hop Resolved to an LDP Tunnel

On PE-1, the route for prefix 2001::10:10:4:0/120 uses a tunnel to 6PE next hop 192.0.2.4, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001::10:10:4:0/120                           Remote  BGP_LABEL 00h00m44s  170
       192.0.2.4 (tunneled)                                         20
-------------------------------------------------------------------------------
No. of Routes: 1
```

LDP is enabled on the interfaces between the PEs and P-2, which is sufficient for 6PE next-hop resolution to an LDP tunnel. RSVP-TE tunnels have a higher priority, but no MPLS LSPs have been configured yet on the PEs. The tunnel table on PE-1 shows that the only tunnel to 6PE next hop 192.0.2.4 is an LDP tunnel, as follows:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.4/32      ldp       MPLS  65538     9        192.168.12.2   20
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

Alternatively, the following show command can be used: the only tunnel on slot 1 (card 1) to 6PE next hop 192.0.2.4 is an LDP tunnel:

```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop            Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              LDP         -
  262141                 192.168.12.2       1/1/1
```

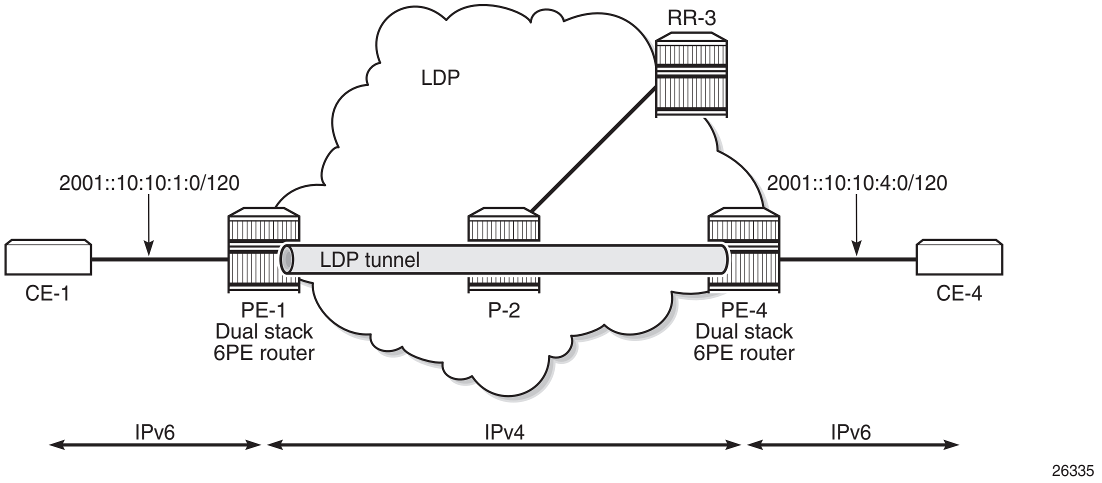
```
-------------------------------------------------------------------------------
Total Entries : 1
```

The extended route information for IPv6 prefix 2001::10:10:4:0/120 shows that the 6PE next hop 192.0.2.4 is resolved to an LDP tunnel:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix                : 2001::10:10:4:0/120
  Protocol                 : BGP_LABEL
  Age                      : 00h00m37s
  Preference               : 170
  Indirect Next-Hop        : 192.0.2.4
    Label                  : 2
    QoS                    : Priority=n/c, FC=n/c
    Source-Class           : 0
    Dest-Class             : 0
    ECMP-Weight            : N/A
    Resolving Next-Hop     : 192.0.2.4 (LDP tunnel)
      Metric               : 20
      ECMP-Weight          : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

Figure 44 shows that the 6PE next hop is resolved to an LDP tunnel. No other tunnels are available in the IPv4 core network.

Figure 44   6PE Next Hop Resolved to an LDP Tunnel



6PE Next Hop Resolved to an RSVP-TE Tunnel

MPLS and RSVP are enabled on the interfaces between the PEs and P-2. On both PEs, an RSVP-TE LSP is configured toward the peer PE; for example, on PE-1:

```
configure
    router
        mpls
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-4"
                to 192.0.2.4
                primary "dyn"
                exit
                no shutdown
            exit
```

The configuration is similar on PE-4. No additional configuration is required on P-2.

The following output shows that two tunnels are available to 6PE next hop 192.0.2.4/32: an LDP tunnel and an RSVP-TE tunnel:

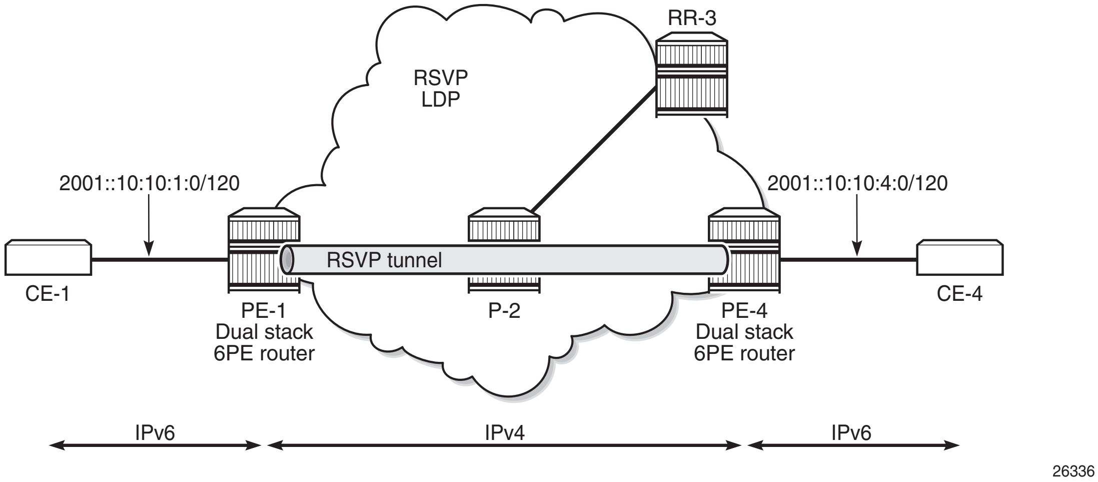
```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop           Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              LDP         -
  262141                 192.168.12.2      1/1/1
192.0.2.4/32                                              RSVP        1
  262140                 192.168.12.2      1/1/1
-------------------------------------------------------------------------------
Total Entries : 2
```

For 6PE next-hop resolution, RSVP-TE tunnels are preferred to any other tunnel type in the tunnel table, so the BGP next hop 192.0.2.4 will be resolved to an RSVP-TE tunnel, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001::10:10:4:0/120
  Protocol              : BGP_LABEL
  Age                   : 00h00m11s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.4
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.4 (RSVP tunnel:1)
      Metric            : 20
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

Figure 45 shows that the 6PE next hop 192.0.2.4 is resolved to an RSVP-TE tunnel, even though an LDP tunnel is available too.

Figure 45 **6PE Next Hop Resolved to an RSVP-TE Tunnel**



6PE Next Hop Resolved to an SR-ISIS Tunnel

Segment routing is enabled for IS-IS on PE-1, P-2, and PE-4. The configuration is similar on each of these nodes; the only difference is the index on the system interface. The SR-ISIS configuration on PE-1 is as follows:

```
configure
    router
        mpls-labels
            sr-labels start 20000 end 20099
        exit
        isis
            advertise-router-capability area
            interface "system"
                ipv4-node-sid index 1
            exit
            segment-routing
                prefix-sid-range start-label 20000 max-index 99
                no shutdown
            exit
        exit
```

For more information about SR-ISIS, see chapter Segment Routing with IS-IS Control Plane.

The following output shows that three tunnels are available toward 6PE next hop 192.0.2.4/32:

```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                 NextHop            Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              LDP         -
  262142                192.168.12.2       1/1/1
192.0.2.4/32                                              RSVP        1
  262136                192.168.12.2       1/1/1
192.0.2.4/32                                              SR-ISIS-0   -
  20004                 192.168.12.2       1/1/1
-------------------------------------------------------------------------------
Total Entries : 3
```

The preference for RSVP-TE tunnels is the highest; therefore, the 6PE next hop 192.0.2.4 is resolved to the RSVP-TE tunnel, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001::10:10:4:0/120
  Protocol              : BGP_LABEL
  Age                   : 01h04m19s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.4
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.4 (RSVP tunnel:1)
      Metric            : 20
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

To verify that LDP tunnels are preferred over SR-ISIS tunnels, the RSVP-TE LSPs are put in a shutdown state, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-4" shutdown
*A:PE-4# configure router mpls lsp "LSP-PE-4-PE-1" shutdown
```

The following output shows that two tunnels are available toward 6PE next hop 192.0.2.4/32: an LDP tunnel and an SR-ISIS tunnel.

```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop            Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              LDP          -
  262142                 192.168.12.2       1/1/1
192.0.2.4/32                                              SR-ISIS-0    -
  20004                  192.168.12.2       1/1/1
-------------------------------------------------------------------------------
Total Entries : 2
```

For 6PE next-hop resolution, the LDP tunnel is preferred over the SR-ISIS tunnel, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001::10:10:4:0/120
  Protocol              : BGP_LABEL
  Age                   : 00h00m08s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.4
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.4 (LDP tunnel)
      Metric            : 20
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

When LDP is disabled on interface "int-PE-1-P-2" on PE-1, the only remaining tunnel is an SR-ISIS tunnel, as follows:

```
*A:PE-1# configure router ldp interface-parameters interface "int-PE-1-P-2" shutdown
```

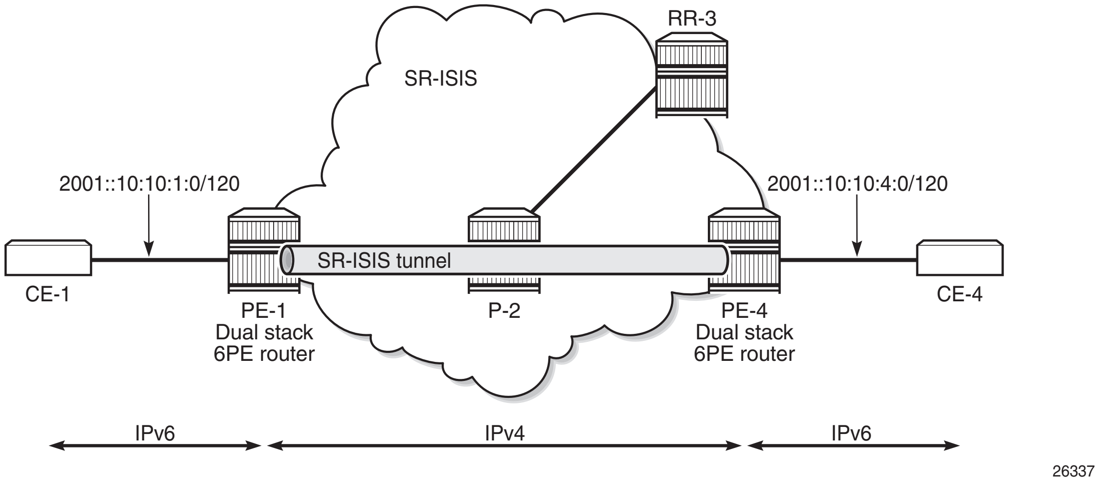
```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display
Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop           Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              SR-ISIS-0   -
  20004                  192.168.12.2      1/1/1
-------------------------------------------------------------------------------
Total Entries : 1
```

The 6PE next hop 192.0.2.4 is resolved to an SR-ISIS tunnel, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001::10:10:4:0/120
  Protocol              : BGP_LABEL
  Age                   : 00h00m08s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.4
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.4 (SR-ISIS:0 tunnel)
      Metric            : 20
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

Figure 46 shows that the 6PE next hop 192.0.2.4 is resolved to an SR-ISIS tunnel after the RSVP-TE LSPs are disabled and LDP is disabled on the interfaces between the PEs and P-2. No other tunnels are available.

Figure 46 **6PE Next Hop Resolved to an SR-ISIS Tunnel**



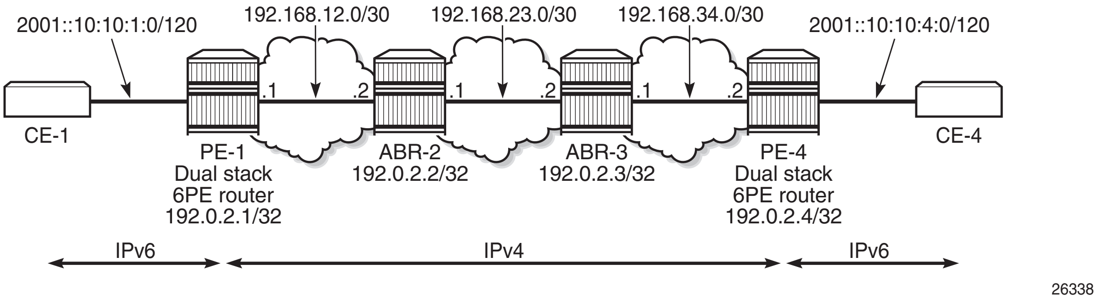
6PE Next-Hop Resolution to a BGP IPv4 Tunnel

The preceding example cannot be extended with BGP labeled IPv4 tunnels. The reason is that for BGP to work, some underlying MPLS signaling protocol is required, such as RSVP-TE or LDP. Because BGP tunnels have a very low preference, they will not be used when an LDP or RSVP-TE tunnel is available to the 6PE next hop.

This section shows a seamless MPLS example where 6PE next hops are resolved to BGP labeled IPv4 routes, because no LDP tunnel is available to the 6PE next hop in a different IGP topology (in this example, LDP is configured, not RSVP-TE). For a description of this seamless MPLS implementation, see chapter Seamless MPLS: Isolated IGP/LDP Domains and Labeled BGP.

Figure 47 shows the example topology for seamless MPLS with two aggregation networks and one core network.

Figure 47 **Example Topology for Seamless MPLS**

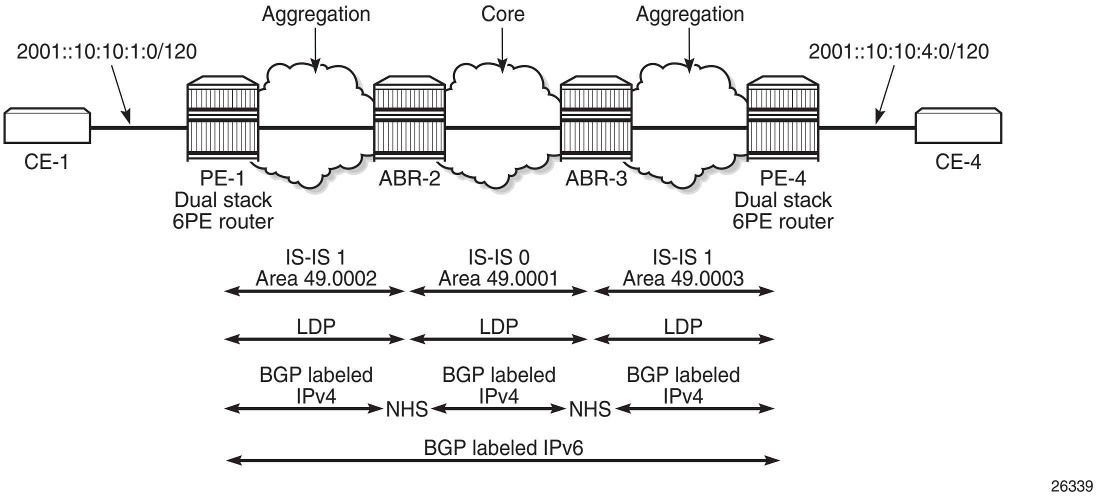


Different IS-IS instances are configured: IS-IS instance 0 is configured in the core, whereas IS-IS instance 1 is configured in the aggregation networks. On the Area Border Routers (ABRs) ABR-2 and ABR-3, two instances of IS-IS are configured: IS-IS instance 0 for the core and IS-IS instance 1 for the aggregation network. PE-1 and PE-4 will only learn routes to destinations within their respective aggregation networks; ABRs learn routes within one aggregation network and the core network. LDP is configured on all interfaces, but PE-1 will not have an LDP binding for prefix 192.0.2.4/32, as shown in the following output. Therefore, 6PE next hop 192.0.2.4 cannot be resolved to an LDP tunnel.

```
*A:PE-1# show router ldp bindings active prefixes

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC, e - Label ELC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop             262143      --
  --                                          --

192.0.2.2/32                                Push              --      262143
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
```

Figure 48 shows the configured protocols for this example: IS-IS instances, LDP, BGP labeled IPv4 with the ABRs as route reflector with **next-hop-self** (NHS) option, and BGP labeled IPv6 peering between PE-1 and PE-4.

Figure 48 **Configured Protocols for Seamless MPLS**



The following initial configuration on ABR-2 includes two IS-IS instances in different areas. IS-IS instance 0 with area ID 49.0001 is configured in the core network; IS-IS instance 1 with area ID 49.0002 is configured in the aggregation network. LDP is configured on each router interface.

```
configure
    router
        interface "int-ABR-2-PE-1"
            address 192.168.12.2/30
            port 1/1/2
        exit
        interface "int-ABR-2-ABR-3"
            address 192.168.23.1/30
            port 1/1/3
        exit
        interface "system"
            address 192.0.2.2/32
        exit
        isis 0
            level-capability level-2
            area-id 49.0001
            interface "system"
            exit
            interface "int-ABR-2-PE-1"
                interface-type point-to-point
            exit
        exit
        isis 1
            level-capability level-2
            area-id 49.0002
            interface "system"
            exit
            interface "int-ABR-2-ABR-3"
                interface-type point-to-point
            exit
        exit
        ldp
            interface-parameters
                interface "int-ABR-2-PE-1"
                exit
                interface "int-ABR-2-ABR-3"
                exit
            exit
        exit
```

The configuration is similar on the other nodes. Only the ABRs have two IS-IS instances configured; the PEs only have one IS-IS instance.

BGP needs to be configured for the label-ipv4 and label-ipv6 address families:

- The label-ipv4 address family is used with the ABRs as RR in the aggregation network. Each ABR is configured with the **next-hop-self** option. BGP label-ipv4 peering is between the ABRs without RR.
- The label-ipv6 address family is used between PE-1 and PE-4. The BGP session can only be established after the BGP labeled IPv4 routes have been exchanged between PE-1 and PE-4.

BGP is configured on PE-1 as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            split-horizon
            next-hop-resolution
                label-route-transport-tunnel
                    family label-ipv6
                        resolution-filter
                            bgp
                        exit
                        resolution filter
                    exit
                exit
            exit
            group "iBGPv4"
                export "export-sys"
                peer-as 64496
                neighbor 192.0.2.2
                    family label-ipv4
                exit
            exit
            group "iBGPv6"
                export "export-6pe"
                peer-as 64496
                neighbor 192.0.2.4
                    family label-ipv6
                exit
            exit
```

The configuration is similar on PE-4, but the neighbor IP addresses are different.

The resolution filter will include LDP as well as BGP, because it is added automatically. However, no LDP tunnel will be available from PE-1 to PE-4, or vice versa; therefore, BGP labeled IPv4 will be used.

The "export-sys" policy exports the IPv4 system address of the PE and is defined as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "system"
                prefix 192.0.2.0/24 longer
            exit
            policy-statement "export-sys"
                entry 10
                    from
                        protocol direct
                        prefix-list "system"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The "export-6pe" policy exports the local labeled IPv6 routes and is the same in the preceding examples:

```
configure
    router
        policy-options
            begin
            policy-statement "export-6pe"
                entry 10
                    from
                        protocol direct
                        family label-ipv6
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The BGP configuration on ABR-2 has two different groups for BGP labeled IPv4 peering: one toward the aggregation network-with the ABR as RR-and one toward the core, as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            advertise-inactive
            split-horizon
            group "iBGPv4-agg"
                next-hop-self
                cluster 2.2.2.2
                peer-as 64496
                neighbor 192.0.2.1
                    family label-ipv4
                exit
            exit
            group "iBGPv4-core"
                next-hop-self
                peer-as 64496
                neighbor 192.0.2.3
                    family label-ipv4
                exit
            exit
```

The configuration is similar on ABR-3, but the neighbor IP addresses and the cluster ID are different. An RR can be configured for group "iBGPv4-core" in the core network; for example, on ABR-3.

The ABRs are configured with the **next-hop-self** option for both groups. The 6PE next hop 192.0.2.4 will have next hop ABR-2 on PE-1, which can be resolved to an LDP tunnel. On ABR-2, 6PE next hop 192.0.2.4 will have ABR-3 as next hop, which can be resolved to an LDP tunnel. On ABR-3, the 6PE next hop 192.0.2.4 can be resolved to an LDP tunnel (no active BGP route to 192.0.2.4/32 on ABR-3 as the route via IS-IS is preferred).

The **advertise-inactive** option is required for ABR-2 to export a BGP route for prefix 192.0.2.1/32, which is not active on ABR-2, because an IS-IS route is available for this prefix and IS-IS routes are preferred over BGP routes.

The IES configuration is the same as in the preceding example.

When the labeled IPv4 routes are exchanged between PE-1 and PE-4, the BGP labeled session using IPv6 peering can be established between PE-1 and PE-4, as follows:

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
```

```
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
Def. Instance  64496        5    0 00h00m08s 1/1/1 (Lbl-IPv4)
                            7    0
192.0.2.4
Def. Instance  64496        4    0 00h00m02s 1/1/1 (Lbl-IPv6)
                            5    0
-------------------------------------------------------------------------------
*A:PE-1#
```

For IPv6 prefix 2001::10:10:4:0/120 on PE-1, 6PE next hop 192.0.2.4 is resolved to a BGP tunnel, as follows:

```
*A:PE-1# show router route-table 2001::10:10:4:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001::10:10:4:0/120
  Protocol              : BGP_LABEL
  Age                   : 01h24m01s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.4
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.4 (BGP tunnel)
      Metric            : 1000
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
```

The BGP labeled IPv4 route to 192.0.2.4 has different next hops in different nodes, because both ABRs set the **next-hop-self** option. On PE-1, the BGP labeled IPv4 route for prefix 192.0.2.4 has next hop 192.0.2.2 and uses an LDP tunnel to reach ABR-2 within the aggregation network, as follows:

```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop                  Intf/Tunnel
```

```
-------------------------------------------------------------------------------
192.0.2.4/32                                              BGP         -
  262135                 192.0.2.2         LDP
-------------------------------------------------------------------------------
Total Entries : 1
```

On ABR-2, the BGP labeled route to 192.0.2.4/32 has next hop 192.0.2.3 and uses an LDP tunnel in the core network to reach ABR-3, as follows:

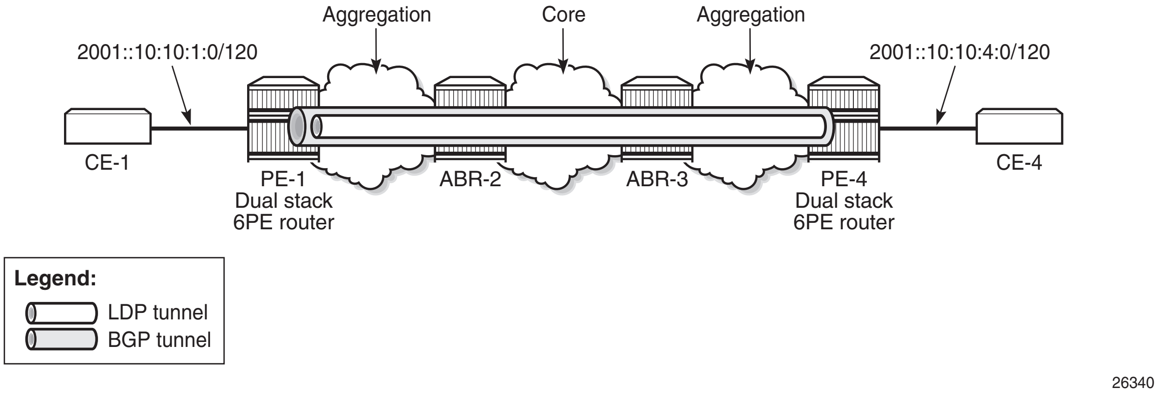
```
*A:ABR-2# show router fp-tunnel-table 1 192.0.2.4/32
===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop           Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              BGP         -
  262136                 192.0.2.3         LDP
-------------------------------------------------------------------------------
Total Entries : 1
```

On ABR-3, no BGP labeled IPv4 route is active for prefix 192.0.2.4 because IS-IS routes are preferred to BGP routes. An LDP tunnel is used toward PE-4 in the aggregation network, as follows:

```
*A:ABR-3# show router fp-tunnel-table 1 192.0.2.4/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                  NextHop           Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.4/32                                              LDP         -
  262143                 192.168.34.2      1/1/1
-------------------------------------------------------------------------------
Total Entries : 1
```

Figure 49 shows the BGP and LDP tunnels used for 6PE next hop 192.0.2.4/32.

Figure 49 **BGP Labeled IPv4 Tunnel for 192.0.2.4/32 Using LDP Tunnels**



Conclusion

The 6PE next hops can be resolved to different types of MPLS tunnels: RSVP-TE, LDP, SR-ISIS, SR-OSPF, SR-TE, and BGP labeled IPv4 tunnels, each with a different preference.

Aggregate Route Indirect Next-Hop Option

This chapter provides information about aggregate route indirect next-hop option configurations.

Topics in this chapter include:

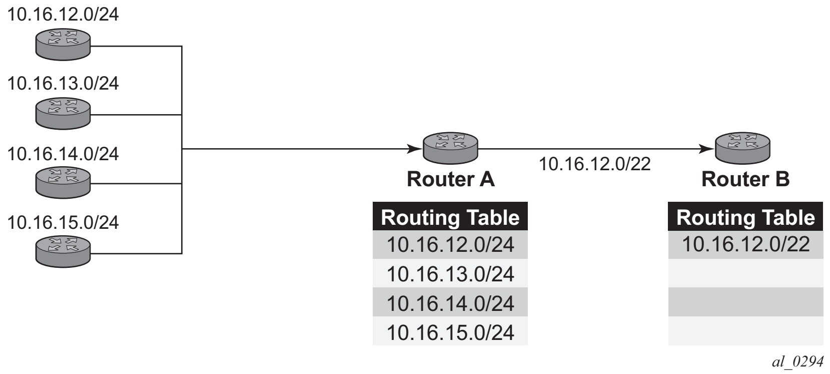
Applicability

This chapter was initially written for SR OS release 11.0.R1. The CLI in this edition corresponds to release 14.0.R2.

Overview

The 7x50s have for many releases supported the ability to configure IPv4 and IPv6 aggregate routes. A configured aggregate route that has the best preference for the prefix is added to the routing table (activated) when it has at least one contributing route and removed when there are no longer any more contributing routes. A contributing route is any route installed in the forwarding table that is a more-specific match of the aggregate. (10.16.12.0/24 is a contributing route to the aggregate route 10.16.12.0/22, but for this same aggregate 10.16.0.0/16 and 10.0.0.0/8 are not contributing routes).

Figure 50 **Aggregate Routes**



In Figure 50, Router A could choose to advertise all the four routes or one aggregate route. By aggregating the four routes, fewer updates are sent on the link between routers A and B, router B needs to maintain a smaller routing table resulting in better convergence and router B saves on computational resources by evaluating fewer entries in its routing table.

Release 11.0.R1 added the ability to configure an indirect hop for aggregate routes. The indirect next hop specifies where packets will be forwarded if they match the aggregate route, but not a more specific route in the IP forwarding table.

Different network operators have different requirements on how to forward a packet that matches an aggregate route but not any of the more-specific routes in the forwarding table that activated the aggregate. In general, there are three different options:

a. The packet can be forwarded according to the next-most specific route, ignoring the aggregate route. This can lead to routing loops in some topologies.

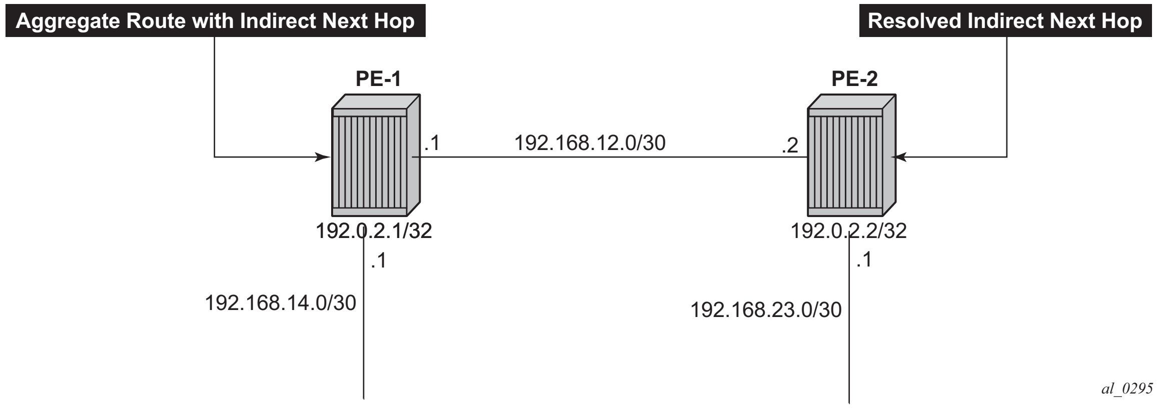
b. The packet can be discarded.

c. The packet can be forwarded toward an indirect next-hop address that is configured by the operator. The indirect next-hop could be the address of a threat management server that analyzes the packets it receives for security threats. This option requires the aggregate route to be installed in the forwarding table with a resolved next-hop interface determined from a route lookup of the indirect next-hop address.

Configuration

The example topology is shown in Figure 51.

Figure 51 **Example topology**



Initial Configuration

The nodes have the following basic configuration:

- Cards, MDAs
- Ports
- Router interfaces

The router interfaces on PE-1 are configured as follows:

```
*A:PE-1# configure router
      interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-4"
            address 192.168.14.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
```

The configuration on the other nodes is similar. The IP addresses are shown in Figure 51. In this example, static routes will be configured. There is no need for an IGP, but could be configured.

Aggregate Route with Indirect Next Hop Option

This feature adds a keyword **indirect** and an associated IP address parameter to the aggregate command in these configuration contexts:

- **— config>router**
- **— config>service>vprn**

The aggregate route configuration commands are as follows:

```
configure [router | vprn <vprn-id> ] aggregate
  - no aggregate <ip-prefix/ip-prefix-length>
  - aggregate <ip-prefix/ip-prefix-length> [summary-only] [as-set]
    [aggregator <as-number:ip-address>] [black-hole [generate-icmp]]
    [community <comm-id>][description <description>]
  - aggregate <ip-prefix/ip-prefix-length> [summary-only] [as-set]
    [aggregator <as-number:ip-address>] [community <comm-id>]
    [indirect <ip-address>][description <description>]

 <ip-prefix/ip-pref*> : ipv4-prefix    - a.b.c.d (host bits must be 0)
                        ipv4-prefix-le - [0..32]
                        ipv6-prefix    - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
                        ipv6-prefix-le - [0..128]
 <summary-only>       : keyword
 <as-set>             : keyword
 <as-number:ip-addr*> : as-number      - [1..4294967295]
                        ip-address     - a.b.c.d
 <black-hole>         : keyword
 <comm-id>            : [72 chars max]
                        <2 byte-asnumber:comm-val>|<well-known-comm>
                        2 byte-asnumber - [0..65535]
                        comm-val        - [0..65535]
                        well-known-comm - no-export|no-export-subconfed|
                                          no-advertise
 <ip-address>         : [64 chars max]
 <generate-icmp>      : keyword
 <description-string> : [80 chars max]
```

Parameters:

- **indirect** — This indicates that the aggregate route has an indirect address. The indirect option is mutually exclusive with the black-hole option. To change the next-hop type of an aggregate route (for example from black-hole to indirect) the route must be deleted and then re-added with the new next-hop type (however other configuration attributes can generally be changed dynamically).

- <ip-address> — Installing an aggregate route with an indirect next-hop is supported for both IPv4 and IPv6 prefixes. However if the aggregate prefix is IPv6 the indirect next-hop must be an IPv6 address and if the aggregate prefix is IPv4 the indirect next-hop must be an IPv4 address.

An indirect next-hop address of an aggregate route may be resolved by any of the following route types:

- Direct/local route
- Static route with regular next-hop, black-hole next-hop or an indirect next-hop
- OSPFv2 or RIP IPv4 route (applicable only to IPv4 aggregate routes)
- LDP shortcut route (applicable only to IPv4 aggregate routes])
- OSPFv2 or IS-IS shortcut route (IPv4 route with an LDP/RSVP or RSVP tunnel next-hop) (applicable only to IPv4 aggregate routes)
- OSPFv3 or IS-IS route
- BGP route resolved by an IGP route
- BGP route resolved by a BGP route
- BGP labeled-IPv4 route resolved by an LDP or RSVP tunnel (applicable only to IPv4 aggregate routes
- 6PE route resolved by an LDP tunnel or static route with black-hole next-hop (applicable only to IPv6 aggregate routes)
- BGP-VPN route resolved by a BGP labeled-IPv4 route, LDP tunnel, or RSVP tunnel (applicable only to aggregate routes configured in a VPRN context)

If an indirect next-hop is not resolved, the aggregate route will show up as black-hole.

The actual configuration of the aggregate route 10.16.12.0/22 is as follows:

```
*A:PE-
1# configure router aggregate 10.16.12.0/22 community 64496:64498 indirect 192.168.1
1.11
```

This creates an aggregate route, but there are no contributing routes that are more specific defined yet. Therefore the aggregate route remains inactive.

```
*A:PE-1# show router aggregate
===============================================================================
Legend: G - generate-icmp enabled
===============================================================================
Aggregates (Router: Base)
===============================================================================
Prefix                                          Aggr IP-Address   Aggr AS
   Summary                                         AS Set          State
     NextHop                                         Community     NextHopType
-------------------------------------------------------------------------------
10.16.12.0/22                                   0.0.0.0           0
   False                                           False           Inactive
```

```
    192.168.11.11                                   64496:64498       Indirect
-------------------------------------------------------------------------------
No. of Aggregates: 1
===============================================================================
*A:PE-1#
```

The inactive aggregate route does not appear in the routing table:

```
*A:PE-1# show router route-table
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h01m57s  0
       system                                                       0
192.168.12.0/30                               Local   Local     00h01m52s  0
       int-PE-1-PE-2                                                0
192.168.14.0/30                               Local   Local     00h01m52s  0
       int-PE-1-PE-4                                                0
-------------------------------------------------------------------------------
No. of Routes: 3
```

Configure Contributing Routes

The aggregate route remains inactive as long as there is no contributing route which is more specific than the aggregate route. The following contributing routes are statically configured on PE-1:

```
*A:PE-1# configure router
        static-route-entry 10.16.12.0/24
            next-hop 192.168.14.2
                no shutdown
            exit
        exit
        static-route-entry 10.16.13.0/24
            next-hop 192.168.14.2
                no shutdown
            exit
        exit
        static-route-entry 10.16.14.0/24
            next-hop 192.168.14.2
                no shutdown
            exit
        exit
        static-route-entry 10.16.15.0/24
            next-hop 192.168.14.2
                no shutdown
            exit
        exit
```

As a result, the aggregate route becomes active:

```
*A:PE-1# show router aggregate
===============================================================================
Legend: G - generate-icmp enabled
===============================================================================
Aggregates (Router: Base)
===============================================================================
Prefix                                          Aggr IP-Address   Aggr AS
   Summary                                         AS Set          State
     NextHop                                         Community     NextHopType
-------------------------------------------------------------------------------
10.16.12.0/22                                   0.0.0.0           0
   False                                           False           Active
    192.168.11.11                                   64496:64498       Indirect
-------------------------------------------------------------------------------
No. of Aggregates: 1
===============================================================================
*A:PE-1#
```

The active aggregate route is added to the route table, as well as the contributing routes:

```
*A:PE-1# show router route-table
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.16.12.0/22                                 Blackh* Aggr      00h00m00s  130
       Black Hole                                                   0
10.16.12.0/24                                 Remote  Static    00h00m00s  5
       192.168.14.2                                                 1
10.16.13.0/24                                 Remote  Static    00h00m00s  5
       192.168.14.2                                                 1
10.16.14.0/24                                 Remote  Static    00h00m00s  5
       192.168.14.2                                                 1
10.16.15.0/24                                 Remote  Static    00h00m00s  5
       192.168.14.2                                                 1
192.0.2.1/32                                  Local   Local     00h04m10s  0
       system                                                       0
192.168.12.0/30                               Local   Local     00h04m05s  0
       int-PE-1-PE-2                                                0
192.168.14.0/30                               Local   Local     00h04m05s  0
       int-PE-1-PE-4                                                0
-------------------------------------------------------------------------------
No. of Routes: 8
```

The aggregate route is black-holed since the next hop is not resolved. There is no route to 192.168.11.0/24.

Configure Route to Indirect Next Hop

A static route is configured on PE-1to the indirect next hop, as follows:

```
*A:PE-1# configure router
        static-route-entry 192.168.11.0/24
            next-hop 192.168.12.2
                no shutdown
            exit
        exit
```

In the route table, the aggregate route is no longer black-holed. The next hop for the indirect next hop is 192.168.12.2 (PE-2).

```
*A:PE-1# show router route-table
================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
10.16.12.0/22                                 Remote  Aggr      00h02m05s  130
       192.168.12.2                                                  0
10.16.12.0/24                                 Remote  Static    00h02m05s  5
       192.168.14.2                                                  1
10.16.13.0/24                                 Remote  Static    00h02m05s  5
       192.168.14.2                                                  1
10.16.14.0/24                                 Remote  Static    00h02m05s  5
       192.168.14.2                                                  1
10.16.15.0/24                                 Remote  Static    00h02m05s  5
       192.168.14.2                                                  1
192.0.2.1/32                                  Local   Local     00h06m15s  0
       system                                                        0
192.168.11.0/24                               Remote  Static    00h02m05s  5
       192.168.12.2                                                  1
192.168.12.0/30                               Local   Local     00h06m10s  0
       int-PE-1-PE-2                                                 0
192.168.14.0/30                               Local   Local     00h06m10s  0
       int-PE-1-PE-4                                                 0
--------------------------------------------------------------------------------
No. of Routes: 9
```

In this example, PE-2 is the resolved indirected next hop and it has a route for prefix 10.16.12.0/22:

```
*A:PE-2# configure router
        static-route-entry 10.16.12.0/22
            next-hop 192.168.23.2
                no shutdown
            exit
        exit
```

The route table on PE-2 looks as follows:

```
*A:PE-2# show router route-table
================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
```

```
-------------------------------------------------------------------------------
10.16.12.0/22                                 Remote  Static    00h00m00s  5
       192.168.23.2                                                 1
192.0.2.2/32                                  Local   Local     00h13m44s  0
       system                                                       0
192.168.12.0/30                               Local   Local     00h13m44s  0
       int-PE-2-PE-1                                                0
192.168.23.0/30                               Local   Local     00h13m44s  0
       int-PE-2-PE-3                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
```

Conclusion

Aggregate routes offer several advantages, the key being reduction in the routing table size and overcoming routing loops, among other things. Aggregate routes with indirect next hop option helps in faster network convergence by decreasing the number of route table changes. This example shows how to configure aggregate routes with indirect next hop option.

Bi-Directional Forwarding Detection

This chapter provides information about bi-directional forwarding (BFD) detection.

Topics in this chapter include:

Applicability

This chapter was originally written for SR OS release 8.0.R4. The CLI now corresponds to release 15.0.R7.

Overview

Bi-Directional Forwarding Detection (BFD) is a light-weight protocol which provides rapid path failure detection between two systems. It has been published as a series of RFCs (RFC 5880, *Bidirectional Forwarding Detection (BFD)*, to RFC 5884, *Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)*.

If a system running BFD stops receiving BFD messages on an interface, it will determine that there has been a failure in the path and notify other protocols associated with the interface. BFD is useful in situations where two nodes are interconnected through either an optical Dense Wavelength Division Multiplexing (DWDM) or Ethernet network. In both cases, the physical network has numerous extra devices which are not part of the Layer 3 network and therefore, the Layer 3 nodes are incapable of detecting failures which occur in the physical network on spans to which the Layer 3 devices are not directly connected.

BFD protocol provides rapid link continuity checking between network devices, and the state of BFD can be propagated to IP routing protocols to drastically reduce convergence time in cases where a physical network error occurs in a transport network.

RFC 5880 defines two modes of operation for BFD:

- Asynchronous mode (supported) — Uses periodic BFD control messages to test the path between systems
- Demand mode (not supported)

In addition to the two operational modes, an echo function is defined (SR OS routers only supports response sending, this is looping back received BFD messages to the original sender).

BFD is running between two peers and supported for the following scenarios:

- BFD for ISIS
- BFD for OSPF
- BFD for PIM
- BFD for Static route
- BFD for RSVP

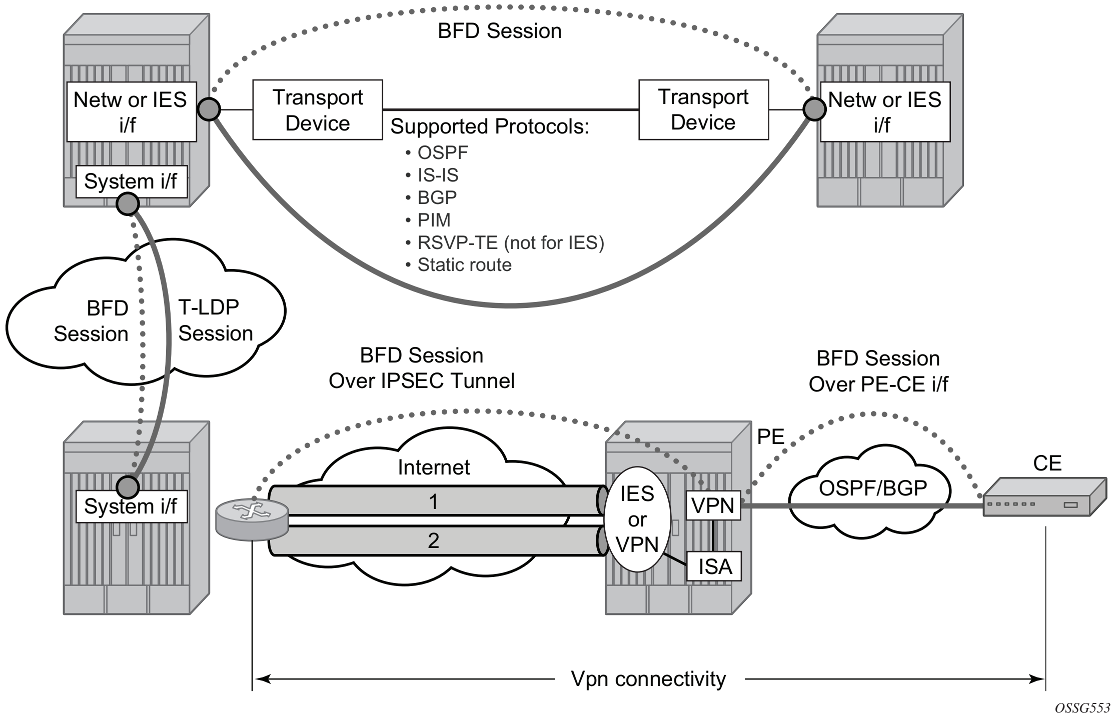
  BFD for I-LDP
- BFD for T-LDP

  BFD for MPLS-TP
- BFD for OSPF CE-PE adjacencies
- BFD for IPSec
- BFD for VRRP

  BFD for SRRP

The scenarios shown in Figure 52 are described in this chapter.

Figure 52 **BFD Multi-Scenarios**



Configuration

BFD packets are processed both locally (processed on the IOM CPU) and centrally (processed on the CPM).

The CPM is able to centrally generate the BFD packets at a sub second interval as low as 10 ms. However, it should be noted that the BFD state machine is still implemented in software. BFD packet generation can be selectively delegated to CPM hardware as needed. This is applicable when sub-second operations or exceeding the IOM scaling limits is required.

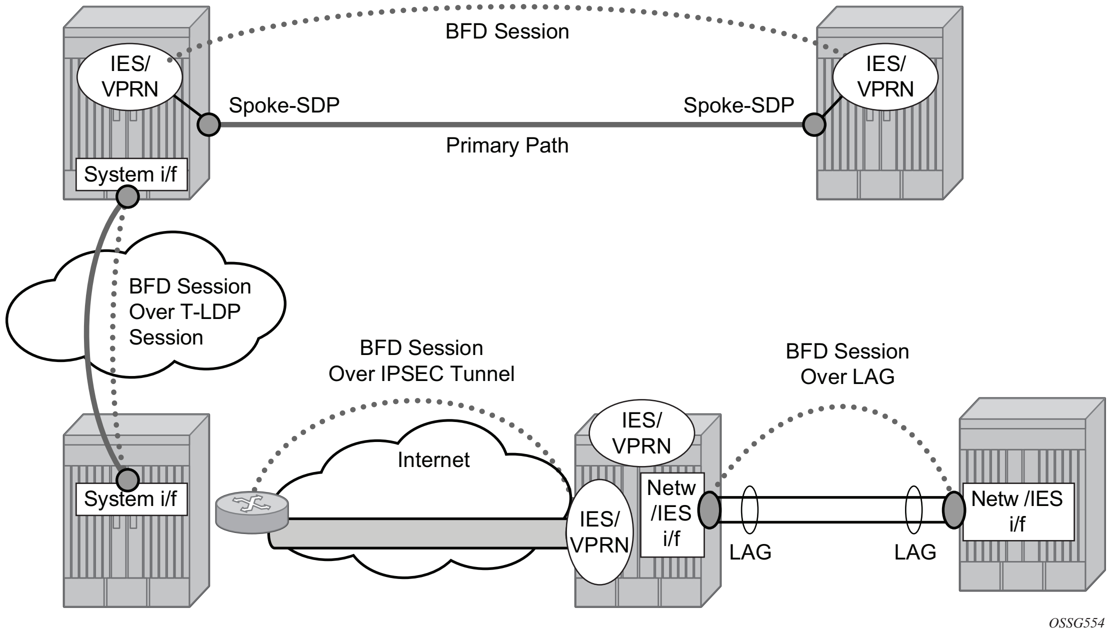
The following applications require BFD to run centrally on the SF/CPM and a centralized session will be created independently of the type explicitly declared by the user:

- BFD for IES/VPRN over Spoke SDP
- BFD for LAG and VSM Interfaces
- Protocol associations using loopback and system interfaces (e.g. BFD for T-LDP)

- BFD for IPSec sessions
- BFD sessions associated with multi-hop peering (BGP)

Figure 53 shows the most relevant scenarios where centralized BFD sessions are used.

Figure 53 **BFD Centralized Sessions**



On the other end, when the two peers are directly connected, the BFD session is local by default, but the user can choose what session type (local or centralized) to implement.

As general rule, the following steps are required to configure and enable a BFD session when peers are directly connected:

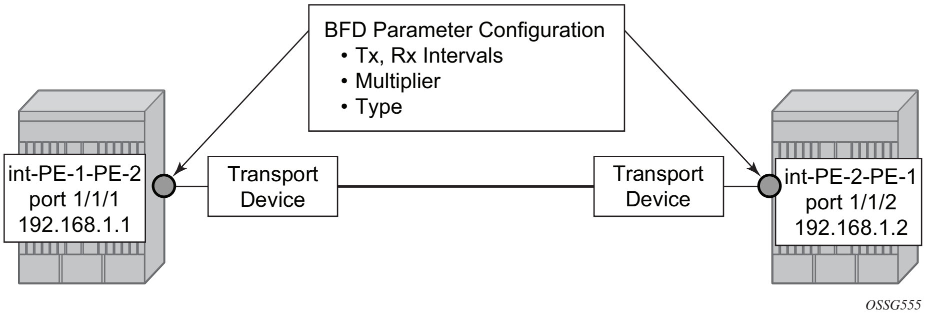
1. Configure BFD parameters on the peering interfaces.
2. Check that the Layer 3 protocol, that is to be bound to BFD, is up and running.
3. Enable BFD under the Layer 3 protocol interface.

Since most of the following procedures share the same first step, it is described only once in the next section and then referred to in subsequent sections.

BFD Base Parameter Configuration and Troubleshooting

The reference topology for the generic configuration of BFD over two local peers is shown in Figure 54.

Figure 54 **BFD Interface Configuration**



To configure BFD between two peers, the user should firstly enable base level BFD on interfaces between PE-1 and PE-2.

On PE-1:

```
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.1.1/30
            port 1/1/1
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        interface "int-PE-2-PE-1"
            address 192.168.1.2/30
            port 1/1/2
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
    exit
exit
```

The following **show** commands are used to verify the BFD configuration on the router interfaces on PE-1 and PE-2.

On PE-1:

```
*A:PE-1# show router bfd interface

===============================================================================
BFD Interface
===============================================================================
Interface name                     Tx Interval    Rx Interval    Multiplier
-------------------------------------------------------------------------------
int-PE-1-PE-2                      100            100            3
-------------------------------------------------------------------------------
No. of BFD Interfaces: 1
===============================================================================
*A:PE-1#
```

On PE-2:

```
*A:PE-2# show router bfd interface

===============================================================================
BFD Interface
===============================================================================
Interface name                     Tx Interval    Rx Interval    Multiplier
-------------------------------------------------------------------------------
int-PE-2-PE-1                      100            100            3
-------------------------------------------------------------------------------
No. of BFD Interfaces: 1
===============================================================================
*A:PE-2#
```

Note that, BFD being an asynchronous protocol, it is possible to configure different Tx and Rx intervals on the two peers. This is because BFD Rx/Tx interval values are signaled in the BFD packets while establishing the BFD session.

The configurable BFD parameters are as follows:

```
bfd <transmit-interval> [receive <receive-interval>] [multiplier <multiplier>]
                                  [echo-receive <echo-interval>] [type <cpm-np>]


 <transmit-interval>  : [10..100000] in milliseconds
 <receive-interval>   : [10..100000] in milliseconds
 <multiplier>         : [3..20]
 <echo-interval>      : [100..100000] in milliseconds
 <cpm-np>             : keyword - use CPM network processor
```

It is possible to force the BFD session to be centrally managed by the CPM hardware.

Regarding the echo function, it is possible to set the minimum echo receive interval, in milliseconds, for the BFD session. The default value is 100 ms.

If a BFD session is running, it is possible to modify its parameters, but to change its type, the session must be previously shut down manually. This causes the upper layer protocols bound to it to be brought down as well.

```
configure
    router
        interface "int-PE-1-PE-2"
            bfd 100 receive 100 multiplier 3
        exit
    exit
exit
```

Forcing a centralized session in the case of directly connected peers can be useful when:

- Lower Tx and Rx intervals are desired (up to 10 ms instead of 100 ms supported by local sessions)
- There are no more available local (IOM) sessions
- Max limit of 500 packets per second per IOM has been reached

The instructions illustrated in following paragraphs are required to complete the configuration and enable BFD.

The BFD session should come up. To verify it, execute a **show router bfd session** command (bound to OSPF in the following example).

```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-2-PE-1                                        Up         1218       1216
  192.168.1.1                                         3          100        100
  ospf2                                             iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

If the command gives a negative output, troubleshoot it by first checking that the protocol that is bound to it is up: for instance, check the OSPF neighbor adjacency as shown in following example.

```
*A:PE-1# show router ospf neighbor
```

```
===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
   Area-Id
-------------------------------------------------------------------------------
int-PE-1-PE-2                    192.0.2.2       Full       1    0       35
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 1
===============================================================================
*A:PE-1#
```

Then check whether a BFD resource limit has been reached (maximum number of local/centralized sessions or maximum number of packet per second per IOM).

If the overloaded limit is the maximum supported number of sessions, the cause is shown in log 99 (maxSessionsPerSlot).

In this case, when one of the running sessions is manually removed or goes down, then the additional configured session will come up. If the IOM limit is reached, it is possible to bring up the session by changing the session type to centralized.

To check if the IOM CPU is able to start more local BFD sessions, execute a **show router bfd session summary** command.

If the **show router bfd session** command reports that the BFD session is down, then check the BFD peer's configuration and state.

The **show router bfd session src <*ip-address*> detail** command can help debugging the BFD session. The sent and received counters are not supported for cpm-np type sessions.

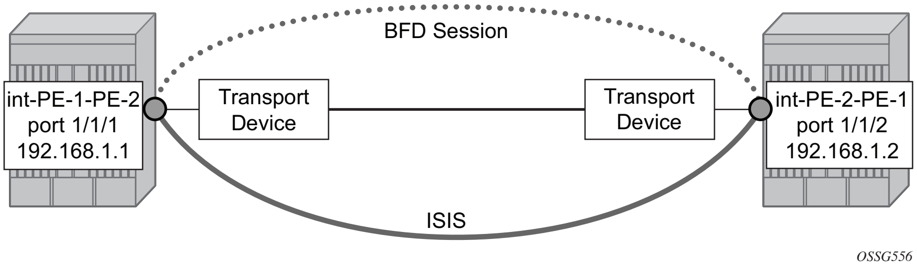
```
*A:PE-1# show router bfd session src 192.168.1.1 detail

===============================================================================
BFD Session
===============================================================================
Remote Address : 192.168.1.2
Admin State    : Up                       Oper State       : Up
Protocols      : ospf2
Rx Interval    : 100                      Tx Interval      : 100
Multiplier     : 3                        Echo Interval    : 0
Recd Msgs      : 2964                     Sent Msgs        : 2976
Up Time        : 0d 00:04:56              Up Transitions   : 1
Down Time      : None                     Down Transitions : 0
                                          Version Mismatch : 0

Forwarding Information

Local Discr    : 106                      Local State      : Up
Local Diag     : 0 (None)                 Local Mode       : Async
Local Min Tx   : 100                      Local Mult       : 3
```

```
Last Sent       : 03/07/2018 13:20:49      Local Min Rx      : 100
Type            : iom
Remote Discr    : 104                      Remote State      : Up
Remote Diag     : 0 (None)                 Remote Mode       : Async
Remote Min Tx   : 100                      Remote Mult       : 3
Last Recv       : 03/07/2018 13:20:49      Remote Min Rx     : 100
===============================================================================
===============================================================================
*A:PE-1#
```

BFD for IS-IS

The goal of this section is to configure BFD on a network interlink between two 7750 nodes that are IS-IS peers. The topology used is shown in Figure 55.

Figure 55 **BFD for ISIS**



For the base BFD configuration, please refer to BFD Base Parameter Configuration and Troubleshooting.

Apply BFD to the IS-IS Interfaces.

On PE-1:

```
configure
    router
        isis
            interface "int-PE-1-PE-2"
                bfd-enable ipv4
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        isis
            interface "int-PE-2-PE-1"
```

```
                bfd-enable ipv4
            exit
        exit
    exit
exit
```

Finally, verify that the BFD session is operational between PE-1 and PE-2.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-PE-2                                        Up           76         75
  192.168.1.2                                         3          100        100
  isis                                              iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-2:

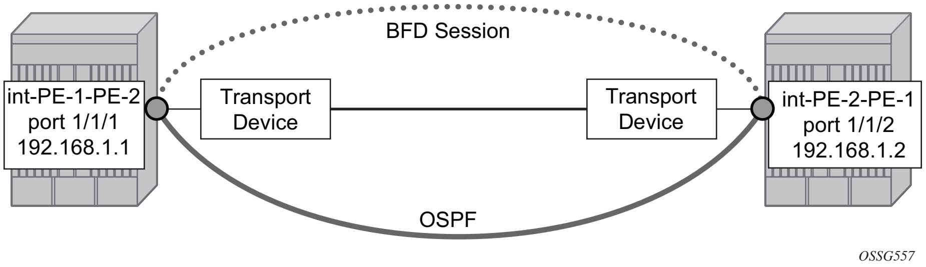
```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-2-PE-1                                        Up          165        166
  192.168.1.1                                         3          100        100
  isis                                              iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

BFD for OSPF

The goal of this section is to configure BFD on a network interlink between two 7750 nodes that are OSPF peers.

For this scenario, the topology is shown in Figure 56.

Figure 56 **BFD for OSPF**



For the base BFD configuration, refer to BFD Base Parameter Configuration and Troubleshooting.

Apply BFD on the OSPF Interfaces.

On PE-1:

```
configure
    router
        ospf
            area 0
                interface "int-PE-1-PE-2"
                    bfd-enable
                exit
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        ospf
            area 0
                interface "int-PE-2-PE-1"
                    bfd-enable
                exit
            exit
        exit
    exit
exit
```

Verify that the BFD session is operational between PE-1 and PE-2.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-PE-2                                        Up          616        604
  192.168.1.2                                         3          100        100
  ospf2                                             iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-2:

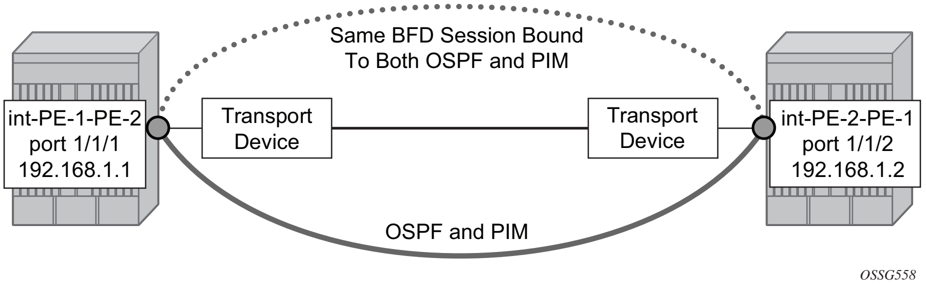
```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-2-PE-1                                        Up         1218       1216
  192.168.1.1                                         3          100        100
  ospf2                                             iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

BFD for PIM

Since the implementation of PIM uses an Interior Gateway Protocol (IGP) in order to determine its Reverse Path Forwarding (RPF) tree, BFD configuration to support PIM will require BFD configuration of both the IGP protocol and the PIM protocol. Let's assume that IGP protocol is OSPF and that the starting configuration is as described in the previous section.

In this paragraph, configure and enable BFD for PIM on the same interfaces that were previously configured with BFD for OSPF, in reference to the topology shown in Figure 57.

Figure 57 **BFD for OSPF and PIM**



Since BFD has been already configured on the router interfaces, let's start by applying BFD on the PIM Interface.

On PE-1:

```
configure
    router
        pim
            interface "int-PE-1-PE-2"
                bfd-enable ipv4
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        pim
            interface "int-PE-2-PE-1"
                bfd-enable ipv4
            exit
        exit
    exit
exit
```

The final step is to verify whether the BFD Session is operational between PE-1 and PE-2 for PIM.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-PE-2                                        Up         8953       8941
  192.168.1.2                                         3          100        100
  ospf2 pim                                         iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

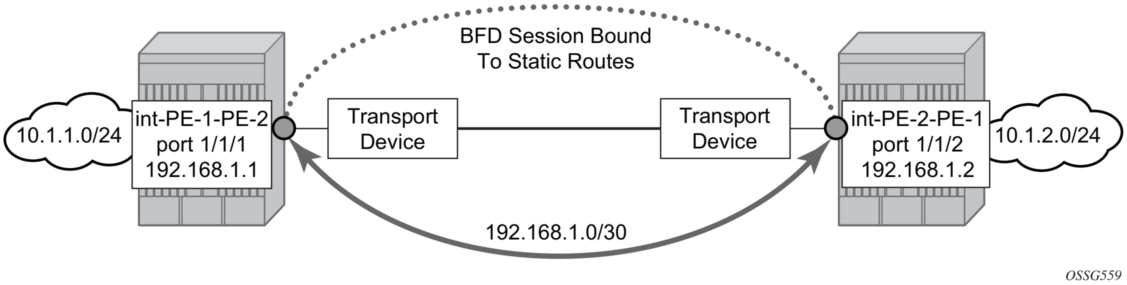
On PE-2:

```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-2-PE-1                                        Up         8850       8849
  192.168.1.1                                         3          100        100
  ospf2 pim                                         iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

BFD for Static Routes

The following procedures will go through the necessary steps to configure the base level BFD configuration and then apply BFD to the static routes between PE-1 and PE-2, using the topology shown in Figure 58.

Figure 58 **BFD for Static Routes**



First, create the static routes for the remote networks both in PE-1 and PE-2.

On PE-1:

```
configure
    router
        static-route-entry 10.1.2.0/24
            next-hop 192.168.1.2
                no shutdown
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        static-route-entry 10.1.1.0/24
            next-hop 192.168.1.1
                no shutdown
            exit
        exit
    exit
exit
```

Next, verify that static routes are populated in the routing table.

On PE-1:

```
*A:PE-1# show router route-table protocol static

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.2.0/24                                   Remote  Static    00h01m49s  5
       192.168.1.2                                                  1
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

On PE-2:

```
*A:PE-2# show router route-table protocol static

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.1.0/24                                   Remote  Static    00h01m21s  5
       192.168.1.1                                                  1
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

The next step is to configure the base level BFD on PE-1 and PE-2.

Refer to paragraph BFD Base Parameter Configuration and Troubleshooting.

Then apply BFD to the static routing entries using the BFD interfaces as next-hop.

On PE-1:

```
configure
    router
       static-route-entry 10.1.2.0/24
            next-hop 192.168.1.2
                bfd-enable
                no shutdown
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
       static-route-entry 10.1.1.0/24
            next-hop 192.168.1.1
                bfd-enable
                no shutdown
            exit
        exit
    exit
exit
```

Note that BFD cannot be enabled if the next hop is indirect or the **black-hole** keyword is specified.

Finally, show the BFD session status.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-PE-2                                        Up        14653      14641
  192.168.1.2                                         3          100        100
  static                                            iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-2:

```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
```

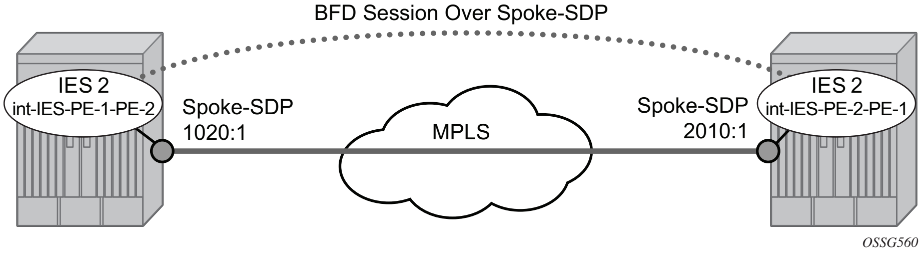
```
int-PE-2-PE-1                                                 Up         1476        1477
  192.168.1.1                                                  3          100         100
  static                                                      iom          N/A         N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

BFD for IES

The goal of this section is to configure BFD for one IES service over a spoke SDP.

The IES service is configured in both 7750 nodes, PE-1 and PE-2, and their interfaces are connected by spoke SDPs. The topology is shown in Figure 59.

Figure 59 **BFD for IES over Spoke SDP**



In this scenario, BFD is run between the IES interfaces independent of the SDP/LSP paths.

The first step is to configure the IES service on both nodes.

On PE-1:

```
configure
    service
        ies 2 customer 1 create
            interface "int-IES-PE-1-PE-2" create
                address 192.168.3.1/30
                spoke-sdp 1020:1 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
```

```
    service
        ies 2 customer 1 create
            interface "int-IES-PE-2-PE-1" create
                address 192.168.3.2/30
                spoke-sdp 2010:1 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The next step is to add the IES interfaces to the OSPF area domain.

On PE-1:

```
configure
    router
        ospf
            area 0
                interface "int-IES-PE-1-PE-2"
                exit
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        ospf
            area 0
                interface "int-IES-PE-2-PE-1"
                exit
            exit
        exit
    exit
exit
```

Then verify that OSPF and the services are up using show commands on both routers.

On PE-1:

```
*A:PE-1# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : IES
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
```

```
Last Status Change: 03/07/2018 13:45:29
Last Mgmt Change  : 03/07/2018 13:45:09
Admin State       : Up                   Oper State        : Up
SAP Count         : 0                    SDP Bind Count    : 1

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:1020:1 S(192.0.2.2)                  Spok         0       1556    Up   Up
===============================================================================
*A:PE-1#

*A:PE-1# show router ospf neighbor

===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
   Area-Id
-------------------------------------------------------------------------------
int-PE-1-PE-2                    192.0.2.2       Full       1    0       36
   0.0.0.0
int-IES-PE-1-PE-2                192.0.2.2       Full       1    0       37
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 2
===============================================================================
*A:PE-1#
```

On PE-2:

```
*A:PE-2# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                    Vpn Id            : 0
Service Type      : IES
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                    Creation Origin   : manual
Last Status Change: 03/07/2018 13:45:29
Last Mgmt Change  : 03/07/2018 13:45:29
Admin State       : Up                   Oper State        : Up
SAP Count         : 0                    SDP Bind Count    : 1

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sdp:2010:1 S(192.0.2.1)                  Spok         0       1556    Up   Up
===============================================================================
*A:PE-2#
```

```
*A:PE-2# show router ospf neighbor

================================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
================================================================================
Interface-Name                   Rtr Id          State       Pri  RetxQ   TTL
   Area-Id
--------------------------------------------------------------------------------
int-PE-2-PE-1                    192.0.2.1       Full        1    0       39
   0.0.0.0
int-IES-PE-2-PE-1                192.0.2.1       Full        1    0       39
   0.0.0.0
--------------------------------------------------------------------------------
No. of Neighbors: 2
================================================================================
*A:PE-2#
```

Then configure BFD on the IES interfaces.

On PE-1:

```
configure
    service
        ies 2
            interface "int-IES-PE-1-PE-2"
                bfd 100 receive 100 multiplier 3
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    service
        ies 2
            interface "int-IES-PE-2-PE-1"
                bfd 100 receive 100 multiplier 3
            exit
        exit
    exit
exit
```

Finally, enable BFD on the interfaces under OSPF area 0.

On PE-1:

```
*A:PE-1# configure router ospf area 0.0.0.0 interface "int-IES-PE-1-PE-2" bfd-enable
```

On PE-2:

```
*A:PE-2# configure router ospf area 0.0.0.0 interface "int-IES-PE-2-PE-1" bfd-enable
```

In case of BFD over spoke SDP, a centralized BFD session is created even if a physical link exists between the two nodes. In fact, the next output shows that BFD session type is cpm-np. This is because the spoke SDP is terminated at the CPM. This is also true for BFD running over LAG bundles.

The **central** type is used when BFD packets are completely generated and processed by software on the CPM. The **cpm-np** type is used when BFD packets are generated and processed with hardware assistance on the CPM.

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-IES-PE-1-PE-2                                    Up          N/A        N/A
  192.168.3.2                                         3          100        100
  ospf2                                          cpm-np          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-IES-PE-2-PE-1                                    Up          N/A        N/A
  192.168.3.1                                         3          100        100
  ospf2                                          cpm-np          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

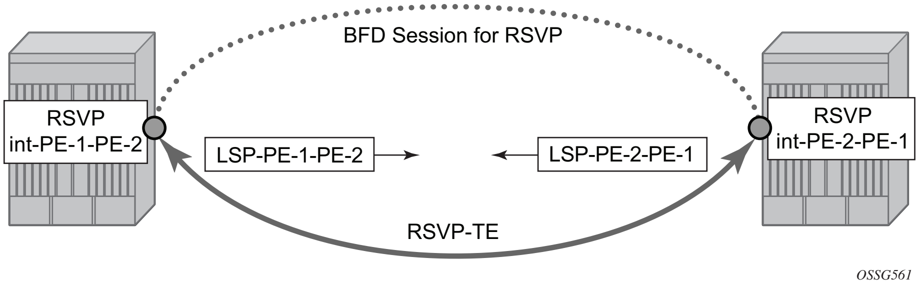
In the case of centralized BFD sessions, transmitted and received packet counters are not shown.

BFD for RSVP

The goal of this section is to configure BFD between two RSVP interfaces configured in two 7750 nodes.

For this scenario, the topology is shown in Figure 60.

Figure 60 **BFD for RSVP**



To enable the BFD session between the two RSVP peers, the user should follow these steps:

First, configure BFD on interfaces between PE-1 and PE-2 as described in BFD Base Parameter Configuration and Troubleshooting.

Next, configure MPLS, creating the path, the LSP and the interfaces within MPLS (and RSVP).

On PE-1:

```
configure
    router
        mpls
            interface "system"
                no shutdown
            exit
            interface "int-PE-1-PE-2"
                no shutdown
            exit
        exit
        rsvp
            interface "system"
                no shutdown
            exit
            interface "int-PE-1-PE-2"
                no shutdown
            exit
            no shutdown
        exit
        mpls
            path "dyn"
```

```
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                cspf
                primary "dyn"
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        mpls
            interface "system"
                no shutdown
            exit
            interface "int-PE-2-PE-1"
                no shutdown
            exit
        exit
        rsvp
            interface "system"
                no shutdown
            exit
            interface "int-PE-2-PE-1"
                no shutdown
            exit
            no shutdown
        exit
        mpls
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-2-PE-1"
                to 192.0.2.1
                cspf
                primary "dyn"
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

Next, verify that the RSVP session is up.

```
*A:PE-1# show router rsvp session

===============================================================================
RSVP Sessions
```

```
===============================================================================
From            To              Tunnel LSP   Name                           State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.2       192.0.2.1       1      8192  LSP-PE-2-PE-1::dyn           Up
192.0.2.1       192.0.2.2       1      18944 LSP-PE-1-PE-2::dyn           Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-1#
```

Then, apply BFD on the RSVP Interfaces.

On PE-1:

```
configure
    router
        rsvp
            interface "int-PE-1-PE-2"
                bfd-enable
            exit
        exit
    exit
exit
```

On PE-2:

```
configure
    router
        rsvp
            interface "int-PE-2-PE-1"
                bfd-enable
            exit
        exit
    exit
exit
```

Finally, verify that the BFD session is operational between PE-1 and PE-2.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-PE-2                                        Up          265        256
  192.168.1.2                                         3          100        100
  rsvp                                          iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-2:

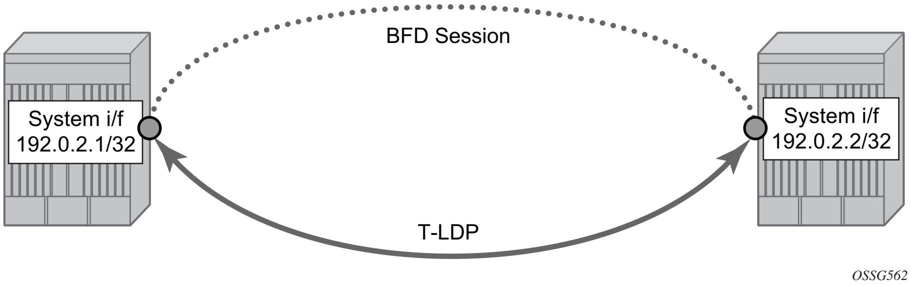
```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                       Multipl     Tx Intvl   Rx Intvl
  Protocols                                         Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-2-PE-1                                        Up          558        558
  192.168.1.1                                         3          100        100
  rsvp                                              iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

BFD for T-LDP

BFD tracking of an LDP session associated with a T-LDP adjacency allows for faster detection of the liveliness of the session by registering the transport address of an LDP session with a BFD session.

The goal of this paragraph is to configure BFD for T-LDP, referring to the scheme shown in Figure 61.

Figure 61 **BFD for T-LDP**



The parameters used for the BFD session are configured under the loopback interface corresponding to the LSR-ID (by default, the LSR-ID matches the system interface address).

```
configure
    router
        interface "system"
            bfd 300 receive 3000 multiplier 3
        exit
    exit
exit
```

By enabling BFD for a selected targeted session, the state of that session is tied to the state of the underlying BFD session between the two nodes.

When using BFD over other links with the ability to reroute, such as spoke-SDPs, the interval and multiplier values configuring BFD should be set to allow sufficient time for the underlying network to re-converge before the associated BFD session expires. A general rule of thumb should be that the expiration time (interval * multiplier) is three times the convergence time for the IGP network between the two endpoints of the BFD session.

Before enabling BFD, ensure that the T-LDP session is up.

On PE-1:

```
*A:PE-1# show router ldp session

===============================================================================
LDP IPv4 Sessions
===============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
-------------------------------------------------------------------------------
192.0.2.2:0         Both      Established   1232      1237      0d 00:45:16
-------------------------------------------------------------------------------
No. of IPv4 Sessions: 1
===============================================================================

================================================================================
LDP IPv6 Sessions
================================================================================
Peer LDP Id
 Adj Type            State          Msg Sent       Msg Recv       Up Time
--------------------------------------------------------------------------------
No Matching Entries Found
================================================================================
*A:PE-1#
```

On PE-2:

```
*A:PE-2# show router ldp session

===============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.1:0         Both      Established   1260      1258      0d 00:46:39
------------------------------------------------------------------------------
No. of IPv4 Sessions: 1
==============================================================================

===============================================================================
LDP IPv6 Sessions
===============================================================================
Peer LDP Id
 Adj Type            State          Msg Sent       Msg Recv       Up Time
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-2#
```

Then, enable the BFD session.

```
configure
    router
        ldp
            targeted-session
                peer 192.0.2.2
                    bfd-enable
                exit
            exit
        exit
    exit
exit
```

The loopback interface can be used to source BFD sessions to many peers in the network.

Finally, check that the BFD session is up.

On PE-1:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
system                                               Up          N/A        N/A
  192.0.2.2                                           3         3000       3000
  ldp                                            cpm-np          N/A        N/A
```

```
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

On PE-2:

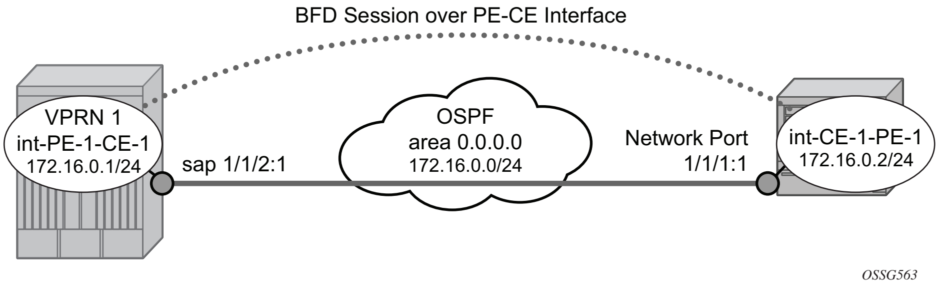
```
*A:PE-2# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
system                                               Up          N/A        N/A
  192.0.2.1                                           3         3000       3000
  ldp                                            cpm-np          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-2#
```

When the T-LDP session comes up, a centralized BFD session is always created (cpm-np) even if the local interface has a direct link to the peer.

BFD for OSPF PE-CE Adjacencies

This feature extends BFD support to OSPF within a VPRN context when OSPF is used as the PE-CE protocol. In this section, the topology used is shown in Figure 62.

Figure 62 **BFD for OSPF PE-CE I/F**



First, configure the VPRN service interface int-PE-1-CE-1 on PE-1 with BFD parameters.

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 1:1
            vrf-target target:1:1
            interface "int-PE-1-CE-1" create
                address 172.16.0.1/24
                bfd 100 receive 100 multiplier 3
                sap 1/1/2:1 create
                exit
            exit
            ospf
                area 0.0.0.0
                    interface "int-PE-1-CE-1"
                        no shutdown
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

Next, configure the router interface on CE-1 and add it to the OSPF area 0 domain.

```
configure
    router
        interface "int-CE-1-PE-1"
            address 172.16.0.2/24
            port 1/1/1:1
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
        ospf
            area 0
                interface int-CE-1-PE-1
                exit
            exit
        exit
    exit
exit
```

Then, ensure that OSPF adjacency is up.

On PE-1:

```
*A:PE-1# show router 1 ospf neighbor

===============================================================================
Rtr vprn1 OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
```

```
   Area-Id
-------------------------------------------------------------------------------
int-PE-1-CE-1                    192.0.2.5       Full       1    0       34
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 1
===============================================================================
*A:PE-1#
```

On CE-1:

```
*A:CE-1# show router ospf neighbor

===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
   Area-Id
-------------------------------------------------------------------------------
int-CE-1-PE-1                    192.0.2.1       Full       1    0       35
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 1
===============================================================================
*A:CE-1#
```

Then, enable BFD on the PE-1-CE-1 interface on PE-1.

```
configure service vprn 1 ospf area 0 interface int-PE-1-CE-1 bfd-enable
```

Enable BFD on the CE-1-PE-1 interface on CE-1.

```
configure router ospf area 0 interface int-CE-1-PE-1 bfd-enable
```

Finally, check that the BFD sessions are up in both PE-1 and CE-1.

```
*A:PE-1# show router 1 bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-PE-1-CE-1                                        Up          259        228
  172.16.0.2                                          3          100        100
  ospf2                                             iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:PE-1#
```

```
*A:CE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
int-CE-1-PE-1                                        Up          719        719
  172.16.0.1                                          3          100        100
  ospf2                                             iom          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:CE-1#
```

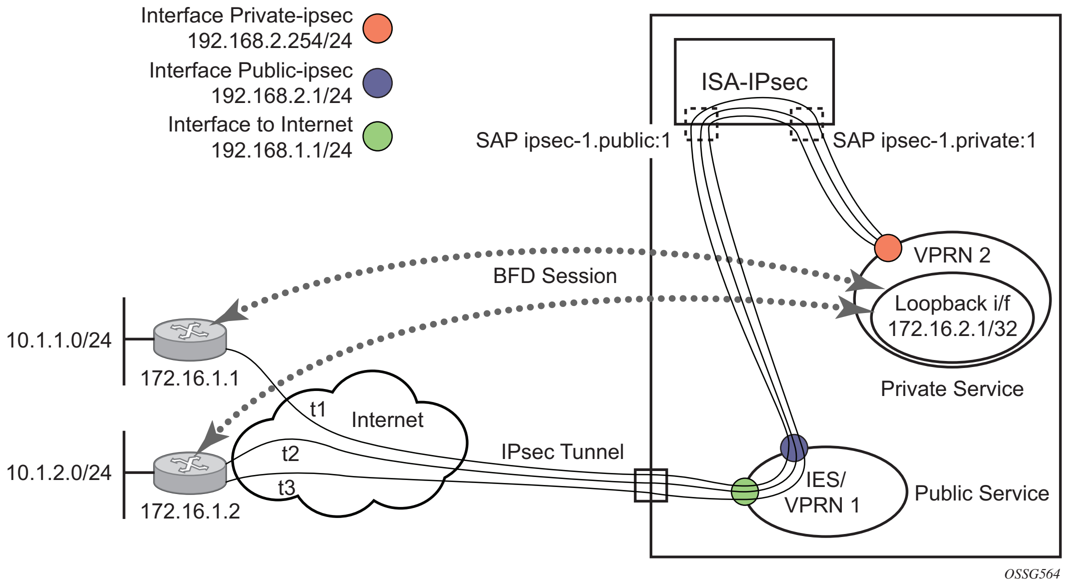
BFD within IPSec Tunnels

The ability to assign a BFD session to a given static LAN-to-LAN IPSec tunnel that provides heart-beat mechanism for fast failure detection has been introduced in Release.8.0.

IPSec needs a Multi-service Integrated Service Adapter (MS-ISA) installed.

The topology used is shown in Figure 63.

Figure 63 **BFD Sessions within IPSec Tunnels**



The first step is to configure MS-ISA card as **type isa-tunnel**.

```
configure
    card 1
        card-type iom3-xp
        mda 1
            mda-type m4-10gb-xp-xfp
        exit
        mda 2
            mda-type isa-tunnel
        exit
        no shutdown
    exit
exit
```

Next, instantiate the tunnels t1, t2 and t3 from the private service (in this example, VPRN 2) to the peers passing through the public service (in this example VPRN 1, but it could be instead an IES).

Since the configuration of IPSec tunnels is out of the scope of this section, only relevant command lines are reported to configure the interfaces shown in Figure 63.

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 1:1
            interface "toInternet" create
                address 192.168.1.1/24
                sap 1/1/2 create
                exit
            exit
            interface "public-ipsec" create
                address 192.168.2.1/24
                sap tunnel-1.public:1 create
                exit
            exit
            no shutdown
        exit
    exit
exit

configure
    service
        vprn 2 customer 1 create
            ipsec
                security-policy 1 create
                    entry 10 create
                        local-ip 192.168.3.1/32
                        remote-ip any
                    exit
                exit
            exit
            route-distinguisher 1:2
            interface "private-ipsec" tunnel create
                sap tunnel-1.private:1 create
                    ipsec-tunnel "t1" create
                        security-policy 1
                        local-gateway-address 192.168.2.254 peer 172.16.1.1
                                                            delivery-service 1
                        dynamic-keying
                            pre-shared-key "vpn1secret"
                        exit
                    exit
                    ipsec-tunnel "t2" create
                        security-policy 1
                        local-gateway-address 192.168.2.254 peer 172.16.1.2
                                                            delivery-service 1
                        dynamic-keying
                            pre-shared-key "vpn1secret"
                        exit
                    exit
                    ipsec-tunnel "t3" create
                        security-policy 1
                        local-gateway-address 192.168.2.254 peer 172.16.1.2
                                                            delivery-service 1
                        dynamic-keying
                            pre-shared-key "vpn1secret"
                        exit
                    exit
                exit
            exit
            interface "loop" create
                address 172.16.2.1/32
                loopback
            exit
            static-route 10.1.1.0/24 ipsec-tunnel "t1"
            static-route 10.1.2.0/24 ipsec-tunnel "t2" metric 1
            static-route 10.1.2.0/24 ipsec-tunnel "t3" metric 5
            no shutdown
        exit
    exit
exit
```

Then configure the BFD parameters within loopback interface loop (refer to BFD Base Parameter Configuration and Troubleshooting).

```
configure
    service
        vprn 2
            interface "loop"
                bfd 100 receive 100 multiplier 3
            exit
        exit
    exit
exit
```

And finally enable BFD within the tunnels.

```
configure
    service
        vprn 2
            interface "private-ipsec" tunnel create
                sap tunnel-1.private:1 create
                    ipsec-tunnel "t1" create
                        bfd-enable service 2 interface "loop" dst-ip 172.16.1.1
                    exit
                    ipsec-tunnel "t2" create
                        bfd-designate
                        bfd-enable service 2 interface "loop" dst-ip 172.16.1.2
                    exit
                    ipsec-tunnel "t3" create
                        bfd-enable service 2 interface "loop" dst-ip 172.16.1.2
                    exit
                exit
            exit
        exit
    exit
exit
```

The BFD-enable parameters are as follows:

- **service** <*service-id*> — Specifies the service-id where the BFD session resides.
- **interface** <*interface-name*> — Specifies the name of the interface used by the BFD session.
- **dst-ip** <*ip-address*> — Specifies the destination address to be used for the BFD session.

The following statements are to be taken into consideration to correctly configure BFD in this environment:

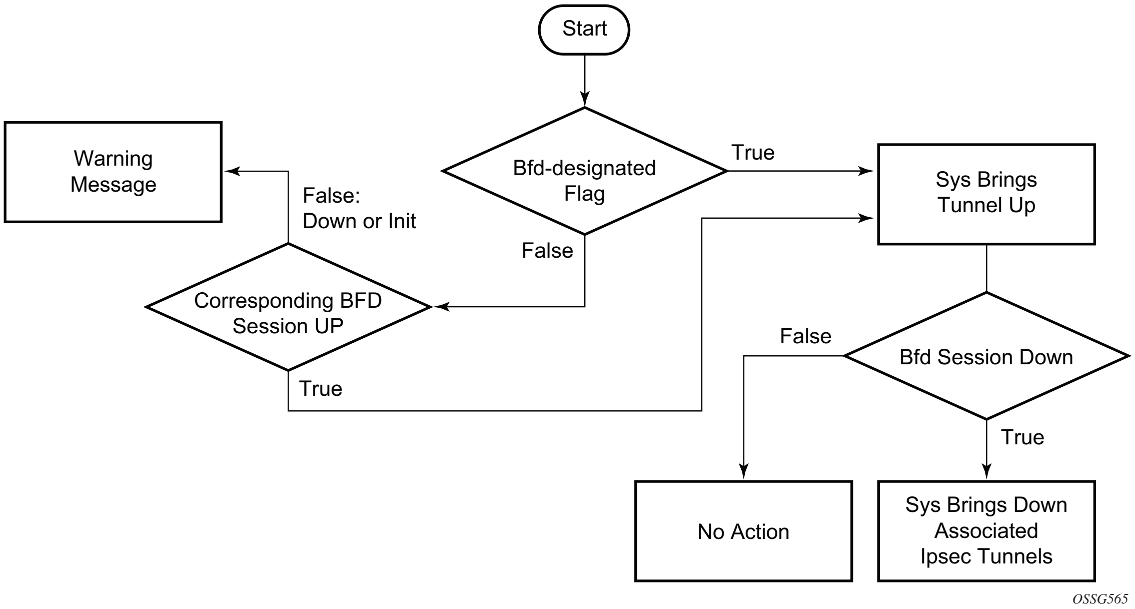
- BFD over IPSec sessions are centralized, managed by the hardware on the CPM.

- Only BFD over static lan-to-lan tunnel is supported in Release 8.0 (not dynamic).
- Only one BFD session is allowed between a given source/destination address pair.
- Each tunnel can be associated to only one BFD session but multiple tunnels can be associated to the same BFD session.
- In case multiple tunnels share the same BFD session, one IPSec tunnel carries BFD traffic: the BFD-DESIGNATED tunnel.

Referring to Figure 63 and to the preceding configuration, the tunnels t2 and t3 share the same BFD-session. Tunnel t2 is the bfd-designated tunnel, the BFD session runs within it and the other tunnel t3 shares its BFD session. If the BFD session goes down, the system will bring down both the designated tunnel t2 and the associated tunnel t3.

The state machine in Figure 64 shows the decision process in case of shared BFD sessions.

Figure 64 **Logic for Shared BFD Sessions**

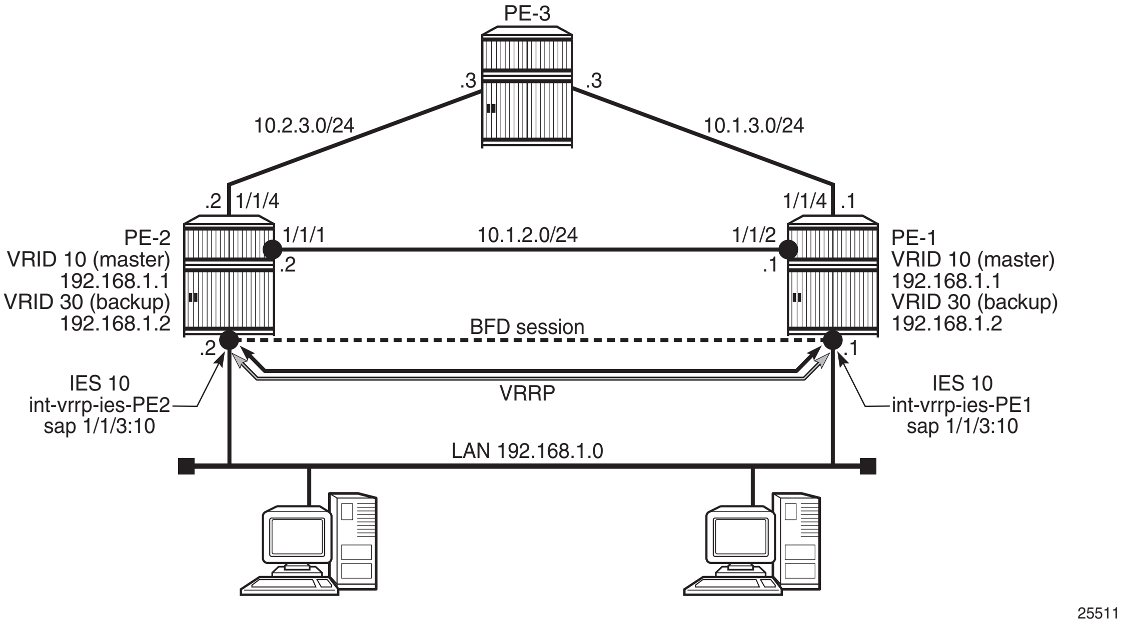


BFD for VRRP

This feature assigns a BFD session to provide a heart-beat mechanism for the VRRP instance. There can be only one BFD session assigned to any given VRRP instance, but there can be multiple VRRP sessions using the same BFD session.

In this section, the topology is shown in Figure 65.

Figure 65 **BFD for VRRP**



First, create the LAN subnet. Two PE routers are connected by IES or VPRN services (in following examples IES 10 is created in both routers).

On PE-1:

```
configure
    service
        ies 10 customer 1 create
            interface "int-vrrp-ies-PE-1" create
                address 192.168.1.1/24
                sap 1/1/3:10 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
    service
        ies 10 customer 1 create
            interface "int-vrrp-ies-PE-2" create
                address 192.168.1.2/24
                sap 1/1/3:10 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

Verify that the IES services are operational (**show service service-using**) and verify that you can ping the remote interface IP address.

Next, configure the VRRP parameters for both PE-1 and PE-2, enable VRRP on the IES interface that connects to the 192.168.1.0/24 subnet.

In this section, the configurations are shown for the VRRP owner mode for master but any other scenario for VRRP can be configured (non owner mode for master).

In the following examples, two VRRP instances are created on the 192.168.1.0/24 subnet:

```
     VRID = 10  Master (owner)  = PE-1
                Backup          = PE-2
                VRRP IP =  192.168.1.1
     VRID = 30  Master (owner)  = PE-2
                Backup          = PE-1
                VRRP IP =  192.168.1.2
```

Host 1 is configured with default gateway = 192.168.1.1, and host 2 is configured with default gateway = 192.168.1.2.

On PE-1:

```
configure
    service
        ies 10 customer 1
            interface "int-vrrp-ies-PE-1"
                vrrp 10 owner
                    backup 192.168.1.1
                exit
                vrrp 30
                    backup 192.168.1.2
                    ping-reply
                    telnet-reply
                    ssh-reply
                exit
            exit
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
    service
```

```
        ies 10 customer 1 create
            interface "int-vrrp-ies-PE-2"
                vrrp 10
                    backup 192.168.1.1
                    ping-reply
                    telnet-reply
                    ssh-reply
                exit
                vrrp 30 owner
                    backup 192.168.1.2
                exit
            exit
            no shutdown
        exit
    exit
exit
```

To bind the VRRP instances with a BFD session, add the following command under any VRRP instance: **bfd-enable** *service-id* **interface** *interface-name* **dst-ip** *ip-address*.

The IES service-id must be declared where the interface is configured.

On PE-1:

```
configure
    service
        ies 10 customer 1
            interface "int-vrrp-ies-PE-1"
                vrrp 10 owner
                    bfd-enable 10 interface "int-vrrp-ies-PE-1" dst-ip 192.168.1.2
                exit
                vrrp 30
                    bfd-enable 10 interface "int-vrrp-ies-PE-1" dst-ip 192.168.1.2
                exit
            exit
            no shutdown
        exit
    exit
exit
```

On PE-2:

```
configure
    service
        ies 10 customer 1
            interface "int-vrrp-ies-PE-2"
                vrrp 10
                    bfd-enable 10 interface "int-vrrp-ies-PE-2" dst-ip 192.168.1.1
                exit
                vrrp 30 owner
                    bfd-enable 10 interface "int-vrrp-ies-PE-2" dst-ip 192.168.1.1
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The parameters used for the BFD are set by the BFD command under the IP interface.

Unlike the previous scenarios, the user can enter the preceding commands , enabling the BFD session, even if the specified interface (int-vrrp-ies-PE-1) has not been configured with BFD parameters.

If it has not been configured yet, the BFD session will be initiated only after the following configuration.

On PE-1:

```
configure service ies 10 interface "int-vrrp-ies-PE-1" bfd 1000
                                             receive 1000 multiplier 10
```

On PE-2:

```
configure service ies 10 interface "int-vrrp-ies-PE-2" bfd 1000
                                             receive 1000 multiplier 10
```

Finally, verify that the BFD session is up (for instance on PE-1):

```
*A:PE-1# show router bfd session src 192.168.1.1 detail

===============================================================================
BFD Session
===============================================================================
Remote Address : 192.168.1.2
Admin State    : Up                       Oper State       : Up
Protocols      : vrrp
Rx Interval    : 1000                     Tx Interval      : 1000
Multiplier     : 10                       Echo Interval    : 0
Recd Msgs      : 120                      Sent Msgs        : 558
Up Time        : 0d 00:00:09              Up Transitions   : 2
Down Time      : None                     Down Transitions : 1
                                          Version Mismatch : 0

Forwarding Information

Local Discr    : 1                        Local State      : Up
Local Diag     : 0 (None)                 Local Mode       : Async
Local Min Tx   : 1000                     Local Mult       : 10
Last Sent      : 03/07/2018 14:58:04      Local Min Rx     : 1000
Type           : iom
Remote Discr   : 3                        Remote State     : Up
Remote Diag    : 0 (None)                 Remote Mode      : Async
Remote Min Tx  : 1000                     Remote Mult      : 10
Last Recv      : 03/07/2018 14:58:05      Remote Min Rx    : 1000
===============================================================================
===============================================================================
*A:PE-1#
```

This session is shared by all the VRRP instances configured between the specified interfaces.

When BFD is configured in a VRRP instance, the following command gives details of BFD related to every instance:

```
*A:PE-1# show router vrrp instance interface "int-vrrp-ies-PE-1"

===============================================================================
VRRP Instances for interface "int-vrrp-ies-PE-1"
===============================================================================
-------------------------------------------------------------------------------
VRID 10
-------------------------------------------------------------------------------
Owner              : Yes                 VRRP State        : Master
Primary IP of Master: 192.168.1.1 (Self)
Primary IP         : 192.168.1.1         Standby-Forwarding: Disabled
VRRP Backup Addr   : 192.168.1.1
Admin State        : Up                  Oper State        : Up
Up Time            : 03/07/2018 14:42:13 Virt MAC Addr     : 00:00:5e:00:01:0a
Auth Type          : None
Config Mesg Intvl  : 1                   In-Use Mesg Intvl : 1
Base Priority      : 255                 In-Use Priority   : 255
Init Delay         : 0                   Init Timer Expires: 0.000 sec
Creation State     : Active

-------------------------------------------------------------------------------
BFD Interface
-------------------------------------------------------------------------------
Service ID         : 10
Interface Name     : int-vrrp-ies-PE-1
Src IP             : 192.168.1.1
Dst IP             : 192.168.1.2
Session Oper State : connected

-------------------------------------------------------------------------------
Master Information
-------------------------------------------------------------------------------
Primary IP of Master: 192.168.1.1 (Self)
Addr List Mismatch : No                  Master Priority   : 255
Master Since       : 03/07/2018 14:42:13

-------------------------------------------------------------------------------
Masters Seen (Last 32)
-------------------------------------------------------------------------------
Primary IP of Master   Last Seen            Addr List Mismatch     Msg Count
-------------------------------------------------------------------------------
192.168.1.1            03/07/2018 14:42:13  No                             0

-------------------------------------------------------------------------------
Statistics
-------------------------------------------------------------------------------
Become Master      : 1                   Master Changes    : 1
Adv Sent           : 1059                Adv Received      : 0
Pri Zero Pkts Sent : 0                   Pri Zero Pkts Rcvd: 0
Preempt Events     : 0                   Preempted Events  : 0
Mesg Intvl Discards : 0                  Mesg Intvl Errors : 0
Addr List Discards : 0                   Addr List Errors  : 0
```

```
Auth Type Mismatch  : 0                   Auth Failures     : 0
Invalid Auth Type   : 0                   Invalid Pkt Type  : 0
IP TTL Errors       : 0                   Pkt Length Errors : 0
Total Discards      : 0

-------------------------------------------------------------------------------
VRID 30
-------------------------------------------------------------------------------
Owner               : No                  VRRP State        : Backup
Primary IP of Master: 192.168.1.2 (Other)
Primary IP          : 192.168.1.1         Standby-Forwarding: Disabled
VRRP Backup Addr    : 192.168.1.2
Admin State         : Up                  Oper State        : Up
Up Time             : 03/07/2018 14:42:13 Virt MAC Addr     : 00:00:5e:00:01:1e
Auth Type           : None
Config Mesg Intvl   : 1                   In-Use Mesg Intvl : 1
Master Inherit Intvl: No
Base Priority       : 100                 In-Use Priority   : 100
Policy ID           : n/a                 Preempt Mode      : Yes
Ping Reply          : Yes                 Telnet Reply      : Yes
SSH Reply           : Yes                 Traceroute Reply  : No
Init Delay          : 0                   Init Timer Expires: 0.000 sec
Creation State      : Active

-------------------------------------------------------------------------------
BFD Interface
-------------------------------------------------------------------------------
Service ID          : 10
Interface Name      : int-vrrp-ies-PE-1
Src IP              : 192.168.1.1
Dst IP              : 192.168.1.2
Session Oper State  : connected

-------------------------------------------------------------------------------
Master Information
-------------------------------------------------------------------------------
Primary IP of Master: 192.168.1.2 (Other)
Addr List Mismatch  : No                  Master Priority   : 255
Master Since        : 03/07/2018 14:42:22
Master Down Interval: 3.609 sec (Expires in 3.050 sec)

-------------------------------------------------------------------------------
Masters Seen (Last 32)
-------------------------------------------------------------------------------
Primary IP of Master   Last Seen             Addr List Mismatch     Msg Count
-------------------------------------------------------------------------------
192.168.1.1            03/07/2018 14:42:16   No                             0
192.168.1.2            03/07/2018 14:59:52   No                          1052

-------------------------------------------------------------------------------
Statistics
-------------------------------------------------------------------------------
Become Master       : 1                   Master Changes    : 2
Adv Sent            : 6                   Adv Received      : 1052
Pri Zero Pkts Sent  : 0                   Pri Zero Pkts Rcvd: 0
Preempt Events      : 0                   Preempted Events  : 1
Mesg Intvl Discards : 0                   Mesg Intvl Errors : 0
Addr List Discards  : 0                   Addr List Errors  : 0
Auth Type Mismatch  : 0                   Auth Failures     : 0
```

```
Invalid Auth Type   : 0                    Invalid Pkt Type  : 0
IP TTL Errors       : 0                    Pkt Length Errors : 0
Total Discards      : 0
===============================================================================
*A:PE-1#
```

Finally, for troubleshooting: it could be that the BFD session between the two IP interfaces is up but (in one or both peers) the command **show router vrrp instance interface** *interface-name* gives the following output regarding BFD for one or more VRIDs.

```
*A:PE-1# show router vrrp instance interface "int-vrrp-ies-PE-1"

===============================================================================
VRRP Instances for interface "int-vrrp-ies-PE-1"
===============================================================================
-------------------------------------------------------------------------------
VRID 10
-------------------------------------------------------------------------------
Owner               : Yes                  VRRP State        : Master
Primary IP of Master: 192.168.1.1 (Self)
Primary IP          : 192.168.1.1          Standby-Forwarding: Disabled
VRRP Backup Addr    : 192.168.1.1
Admin State         : Up                   Oper State        : Up
Up Time             : 03/07/2018 14:42:13 Virt MAC Addr     : 00:00:5e:00:01:0a
Auth Type           : None
Config Mesg Intvl   : 1                    In-Use Mesg Intvl : 1
Base Priority       : 255                  In-Use Priority   : 255
Init Delay          : 0                    Init Timer Expires: 0.000 sec
Creation State      : Active

-------------------------------------------------------------------------------
BFD Interface
-------------------------------------------------------------------------------
Service ID          : None
Interface Name      : int-vrrp-ies-PE-1
Src IP              :
Dst IP              : 192.168.1.2
Session Oper State  : notConfigured

--- snipped ---

===============================================================================
*A:PE-1#
```

To fix this, check that BFD has been correctly configured for the VRRP instances.

For instance, in the following example, the cause of the misconfiguration is that the IES service-id is not declared in the **bfd-enable** command:

```
configure
    service
        ies 10
            interface int-vrrp-ies-PE-1
                vrrp 10
                    bfd-enable 10 interface int-vrrp-ies-PE-1 dst-ip 192.168.1.2
                exit
            exit
        exit
    exit
exit
```

Conclusion

BFD is a light-weight protocol which provides rapid path failure detection between two systems and it is useful in situations where the physical network has numerous intervening devices which are not part of the Layer 3 network.

BFD is linked to a protocol state. For BFD session to be established, the prerequisite condition is that the protocol to which the BFD is linked must be operationally active. Once the BFD session is established, the state of the protocol to which BFD is tied to is then determined based on the BFD session's state. This means that if the BFD session goes down, the corresponding protocol will be brought down.

In this section several scenarios where BFD could be implemented have been described, including the configuration, show output, and troubleshooting hints.

Hybrid OpenFlow Switch

This chapter provides information about Hybrid OpenFlow Switch.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 14.0.R5.

Overview

OpenFlow is defined by the Open Networking Foundation and provides a standard interface between the control layer and forwarding layer of a Software Defined Networking (SDN) architecture. The control layer has northbound interfaces to the application layer and translates the requirements from this layer into low-level control protocols on its southbound interfaces toward the forwarding layer. Made up of SDN controllers, the control layer provides an abstraction between the application layer and the forwarding layer. The forwarding layer may consist of physical and/or virtual network elements.

An OpenFlow controller operates at the control layer while an OpenFlow switch operates at the forwarding layer, and the OpenFlow protocol is used for communication between them. The term Hybrid OpenFlow switch refers to switches or routers that fully integrate both OpenFlow operation and conventional Ethernet switching or IP routing. Conversely, OpenFlow-only switches support only OpenFlow operations. SR OS platforms operate as Hybrid OpenFlow switches.

An OpenFlow switch may have one or more flow tables, each of which contains one or more flow entries. A flow is a sequence of packets that matches a specific entry in a flow table. When a packet is processed by a flow table, it is matched against flow entries that contain match fields and a priority to uniquely identify each entry. Match fields consist of criteria to match against a packet, such as ingress port/VLAN, source/destination IP address, protocol, or source/destination port.

The sequence with which a packet is parsed through a flow table that consists of multiple flow entries depends on the priority of each flow entry. The highest priority flow entry is processed first, and if no match is found, the packet continues to the next highest flow entry until a packet is either matched by a flow entry or all flow entries are parsed and no match is found. Priority 0 is reserved for the table-miss flow entry, which is used when a packet does not match any other flow entries in the flow table. In this case, the packet could be forwarded, dropped, or sent to the OpenFlow controller using a Packet-In message.

Each flow entry consists of one or more OpenFlow Protocol Instruction Types (OFPITs) that collectively form an instruction set. The instruction type defines the type of action to be taken, such as Write-Action, Write-Metadata, or Clear-Action. Each instruction type contains an OpenFlow Protocol Action Type (OFPAT), and the group of actions associated with a flow entry is referred to as an action set. These actions may be to manipulate a packet, or rate-limit packets matching this flow entry, or to output to a specific port, where port may be physical, logical (such as an MPLS or VXLAN tunnel), or a reserved port (such as the control channel with the OpenFlow controller).

Each flow entry is also associated with a 64-bit opaque cookie value assigned by the OpenFlow controller. However, this cookie value is not used for packet lookup or processing. Its purpose is to enable the controller to filter flow statistics and for flow modification/deletion. In SR OS, the cookie value is also used to distinguish between flow entries associated with the base routing instance and those associated with services, as described in more detail later in this chapter.

OpenFlow Protocol

The OpenFlow channel is the interface that connects the OpenFlow switch to a controller and runs over TCP port 6653. By default, the OpenFlow channel is a single TCP connection, but auxiliary connections are also supported. These auxiliary connections may be used for general OpenFlow messages, but are intended to allow for parallel processing of statistics requests and Packet-In messages. Auxiliary connections use the same destination IP address and destination port as the main channel, but are uniquely identified by having a different combination of the switch Datapath ID and an Auxiliary ID on the OpenFlow switch.

The Datapath ID is an 8-byte value used to uniquely identify the switch. To construct it, SR OS uses a concatenation of the OpenFlow switch instance ID (2 bytes) and the chassis MAC (6 bytes). Because the Datapath ID is a switch-wide parameter, it is common to all connections from a switch, but the Auxiliary ID is unique for each channel. In SR OS, the primary channel uses an Auxiliary ID of zero, while auxiliary channels use a unique non-zero value. The Datapath ID and Auxiliary ID are exchanged during the connection setup. After the OpenFlow session is established and Hello messages exchanged, the controller requests a list of supported features from the switch (using an OFPT_FEATURES_REQUEST message). The response from the switch (OFPT_FEATURES_REPLY) contains (among other things) the Datapath ID and Auxiliary ID.

The OpenFlow protocol supports three message types: controller-to-switch, asynchronous, and symmetric.

- Controller-to-switch messages are initiated by the controller and are used to manage or inspect the state of the OpenFlow switch.
- Asynchronous messages are initiated by the switch and are used to notify the controller of network events and changes to switch state.
- Symmetric messages are initiated by either switch or controller and are sent in an unsolicited manner.

OpenFlow specifies the use of a number of different messages within its operation and the use of these messages is constrained to the message type to which they are associated. Table 1 lists the various OpenFlow messages associated with each message type, with a brief description of its usage. Some of the messages will be referred to throughout this chapter, with examples of how and when they are used.

Table 2 **OpenFlow Messages**

Message Type	Message	Description
Controller-to-switch	Feature	[OFPT_FEATURES_REQUEST/REPLY] Used by controller to query capabilities of the switch. Typically used on session establishment.
	Configuration	[OFPT_GET_CONFIG_REQUEST/REPLY, OFPT_SET_CONFIG] Used to set and query configuration parameters in the switch.
	Modify-State	[OFPT_FLOW/PORT/TABLE_MOD] Used to add, delete, and modify flow entries in the OpenFlow tables.
	Read-State	Used to collect information such as configuration, statistics, and capabilities from the switch.
	Packet-Out	[OFPT_PACKET_OUT] Used by the controller to send packets out of a specific port on the switch, and to forward packets received in Packet-In messages.
	Barrier	[OFPT_BARRIER_REQUEST/REPLY] Used to ensure that messages prior to the barrier are processed before any messages after the barrier. Allows for ordering of message processing.
	Role-Request	[OFPT_ROLE_REQUEST/REPLY] Used to set the role of the OpenFlow channel. Can be Master, Slave, or Equal. When multiple controllers are used, only one can be set to Master.
	Asynchronous-Configuration	[OFPT_GET_ASYNC_REQUEST/REPLY, OFPT_SET_ASYNC] Used by the controller to set a filter on the asynchronous messages that it needs to receive.

Table 2 **OpenFlow Messages (Continued)**

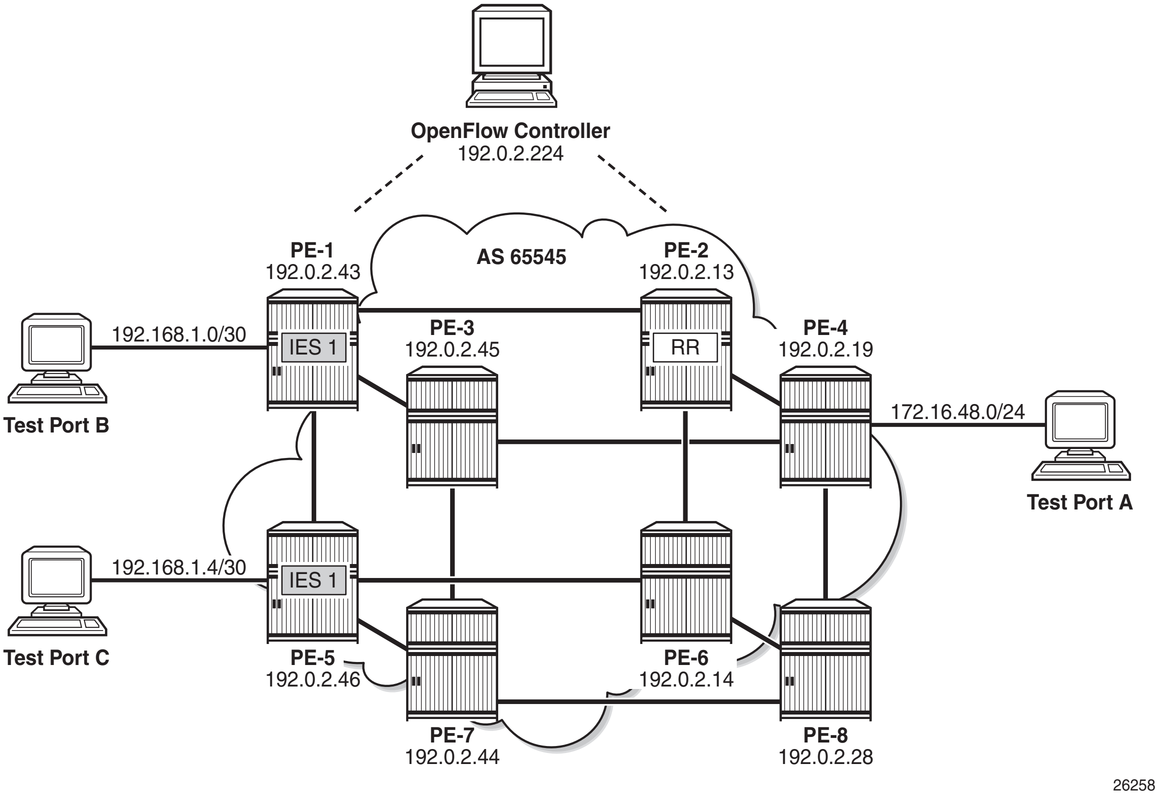
Message Type	Message	Description
Asynchronous	Packet-In	[OFPT_PACKET_IN] Used to transfer a packet to the controller (for example, a table-miss flow entry).
	Flow-Removed	[OFPT_FLOW_REMOVED] Used to notify the controller that a flow entry has been removed from the flow table.
	Port-Status	[OFPT_PORT_STATUS] Used to notify the controller of a change in the configuration or status of a port.
	Error	[OFPT_ERROR] Used to notify the controller of an error.
Symmetric	Hello	[OFPT_HELLO] Exchanged between controller and switch during session startup.
	Echo	[OFPT_ECHO_REQUEST/REPLY] Used to maintain the liveliness of the OpenFlow channel.
	Experimenter	[OFPT_EXPERIMENTER] Provides a standard way to offer additional proprietary functionality within the standard message space.

OpenFlow messages have a standard header that includes the version of the protocol. SR OS supports OpenFlow specification 1.3.1, which requires the use of OpenFlow protocol version 4. Although the OpenFlow protocol defines the standard through which controllers and switches should communicate, it also allows for additional functionality to be implemented, using Experimenter messages and fields. SR OS uses Experimenter fields as additional match criteria and action types.

Configuration

Figure 66 shows an example topology to demonstrate the use of OpenFlow. PE routers PE-1 through PE-8 form part of AS 65545 and run IS-IS and RSVP. All PE routers are IBGP clients of a Route Reflector situated at PE-2 for the IPv4 and VPN-IPv4 address families. An OpenFlow Controller is at address 192.0.2.224, which is reachable from AS 65545. Test port A is connected to PE-4, and test ports B and C are connected to PE-1 and PE-5, respectively. These test ports will be configured to advertise routes and source/sink traffic within the base and service routing contexts to verify OpenFlow operation. More information about specific configurations will be provided within the relevant parts of this chapter.

Figure 66 **Example Topology**



OpenFlow Switch Configuration

OpenFlow specification 1.3.1 allows for multiple flow tables within an OpenFlow switch that are sequentially numbered starting at zero. A function referred to as pipeline processing subsequently matches packets, first against flow entries of flow table 0, but allows for instructions to optionally direct a packet to another flow table, where the process is repeated. Up to eight Hybrid OpenFlow switch instances can be supported per system. Each switch instance supports a single flow table: table 0.

Flow entries pushed from an OpenFlow controller are dynamically embedded within ingress IP filters provisioned on the system. Within the OpenFlow specification, there is no provision for enabling context-specific flow entries. That is, it is not possible to enable a flow entry explicitly within the base routing context, or enable a flow entry explicitly within a service or VPN context. To overcome this, and provide maximum flexibility without the requirement for proprietary extensions, SR OS makes intelligent use of the 64-bit cookie value that is associated with every flow entry in a Modify Flow Entry (OFPT_FLOW_MOD) message. The high-order 32 bits of the value are subdivided into two parts. Bits 63 to 60 are used to determine whether the flow entry is applicable to a filter on an IES or router interface in the base routing context (also referred to as the Global Routing Table [GRT]), or a System filter, or a filter applied within a VPRN or VPLS service context. For the latter, bits 59 to 32 are then used to define the service ID value. This use of the cookie value in this manner is referred to as a multi-service OpenFlow switch instance.

Table 3 **FLOW_MOD Cookie Value**

Bits 63 to 60	Bits 59 to 32	Bits 31 to 0	SR OS context
0000	0	Arbitrary	Used by filters in base router
1000	0	Arbitrary	Used by System filter policy
1100	Service ID value	Arbitrary	Used by filters policies within specified service

Note: The use of a System filter allows for a common rule set defined in an IP filter with a scope of System to be embedded in multiple interface filters, reducing configuration requirements and increasing system scale. The use of System filters is not described within this chapter.

The following output shows the configuration required to define the Hybrid OpenFlow switch and to establish connectivity with the OpenFlow controller:

```
configure
    open-flow
        of-switch "ofs-1"
```

```
                aux-channel-enable
                controller 192.0.2.224:6653
                flowtable 0
                    switch-defined-cookie
                    max-size 4096
                exit
                logical-port-status rsvp-te
                no shutdown
            exit
        exit
```

The **of-switch** command allows for the creation of a switch instance and requires a name of 1 to 32 characters. This creates a new **of-switch** context under which the characteristics of the switch are defined. The **controller** command requires the destination IP address and port of the OpenFlow controller to be entered, separated by a colon. Port 6653 is the standard IANA assigned port for OpenFlow. When this connection is successfully established, it creates the primary channel (with Auxiliary-ID 0) only.

The output shows the configuration of a single controller, but it is possible to configure multiple controllers for redundancy; each controller may create/modify/delete flows entries in the flow table of this switch instance. Also, the OpenFlow switch can use both in-band (base routing context) and out-of-band (management routing context) to establish connectivity with the controller, with preference given to out-of-band if a valid route exists.

The **aux-channel-enable** command establishes the auxiliary channels. When enabled, this command creates an auxiliary channel for statistics with Auxiliary-ID 1, and an auxiliary channel for Packet-In messages with Auxiliary-ID 2. Although these auxiliary channels are assigned an explicit purpose, the switch will still accept any generic OpenFlow messages over these auxiliary channels and will respond in return on the same channel. The **flowtable** command modifies the characteristics of flow table 0. The **max-size** command configures a limit on the number of flow entries that can be populated within each flow-table. Flow-table entries are created in hardware on the line-card datapath and consume Content-Addressable Memory (CAM) entries; therefore, placing a limit on how much of that resource is used by OpenFlow may be needed. The **switch-defined-cookie** command enables the use of a multi-instance OpenFlow switch. This is the recommended approach for deploying an OpenFlow switch in SR OS; it allows for creation of service-specific flow entries, and offers an increased number of traffic actions. Finally, the **logical-port-status rsvp-te** command instructs the switch to report configuration and/or state changes to RSVP-TE logical ports to the controller, which is achieved using asynchronous Port-Status (OFPT_PORT_STATUS) messages.

When the OpenFlow switch is put into a **no shutdown** state, its operational state can be verified with the command shown in the following output:

```
*B:PE-4# show open-flow of-switch "ofs-1" status
```

```
===============================================================================
Open Flow Switch Information
===============================================================================
Switch Name        : ofs-1
Data Path ID       : 00030ca40202d401    Admin Status       : Up
Echo Interval      : 10 seconds          Echo Multiple      : 3
Logical Port Type  : rsvp-te
Buffer Size        : 0                   Num. of Tables     : 1
Description        : (Not Specified)
Capabilities Supp. : flow-stats table-stats port-stats
Aux Channel Enabled: True
===============================================================================
```

The output shows the switch Datapath ID, which together with the Auxiliary ID uniquely identifies a (primary/auxiliary) channel between switch and controller. The output also shows the logical port types in use as being RSVP-TE, support for a single flow table, and a buffer size of 0. The buffer size is used when Packet-In messages are used for a table-miss.

The OpenFlow specification provides an option for a switch to truncate the packet and send only a portion of the packet to the controller in a Packet-In message, together with a buffer-ID, while the remainder of the packet is buffered. When the controller subsequently responds with a Packet-Out message containing a corresponding buffer-ID, the packet is retracted from buffer, re-assembled, and forwarded through the port specified in the Packet-Out message. Rather than buffering, SR OS sends the complete packet to the controller in a Packet-In message, so requires no buffer. Also, SR OS sends only the first packet of a flow in a Packet-In message; any subsequent packets of that flow are dropped at ingress. This avoids overwhelming the controller with table-miss packets, and equally offers a level of protection to the CPM. The expectation is that the controller should create a new flow entry for that flow.

The following output shows the status of the OpenFlow channel to the controller:

```
*B:PE-4# show open-flow of-switch "ofs-1" controller 192.0.2.224:6653 detail

===============================================================================
Open Flow Controller Information
===============================================================================
IP Address            : 192.0.2.224       Port              : 6653
Role                  : equal
Generation ID         : 0

-------------------------------------------------------------------------------
Open Flow Channel Information - Channel ID(1)
-------------------------------------------------------------------------------
Channel ID        : 1                 Version           : 4
Connection Type   : primary           Operational Status: Up
Auxiliary ID      : 0
Source Address    : 192.0.2.19        Source Port       : 56261
Operational Flags : socket-state-established hello-received hello-transmitted
                    handshake
Async Fltr Packet In
```

```
 (Master or Equal): table-miss apply-action
 (Slave)          : (Not Specified)
Async Fltr Port Status
 (Master or Equal): port-add port-delete port-modify
 (Slave)          : port-add port-delete port-modify
Async Fltr Flow Rem
 (Master or Equal): idle-time-out hard-time-out flow-mod-delete group-delete
 (Slave)          : (Not Specified)
Echo Time Expiry  : 0d 00:00:01       Hold Time Expiry  : 0d 00:00:21
Conn. Uptime      : 0d 06:09:59       Conn. Retry       : 0d 00:00:00

-------------------------------------------------------------------------------
Open Flow Channel Stats - Channel ID(1)
-------------------------------------------------------------------------------
Packet Type      Transmitted Packets  Received Packets     Error Packets
-------------------------------------------------------------------------------
Hello            1                    1                    0
Error            1                    0                    0
Echo Request     1722                 508                  0
Echo Reply       508                  1722                 0
Experimenter     0                    0                    0
Feat. Request    0                    1                    0
Feat. Reply      1                    0                    0
---snip---
```

The complete output would show the details of all the OpenFlow channels between the switch and the controller. As the **aux-channel-enable** command is configured, there are three channels in total, but only the primary channel (Auxiliary ID 0) is shown for brevity.

Each controller is assigned a role, which can be master, slave, or equal, with the default being equal. The role determines what access the controller has to the switch and also what asynchronous messages the switch should forward to the controller:

- Equal role: The controller has full access to the switch and is considered equal to other controllers in the same role. All controllers should receive asynchronous messages from the switch.
- Slave role: The controller has read-only access to the switch. Controllers do not receive asynchronous messages from the switch apart from Port-Status messages.
- Master role: The controller has full access to the switch, but only one controller can have the role of master.

If a controller changes its role to master using a Role-Request (OFPT_ROLE_REQUEST) message, the switch modifies all other connections to the role of slave. To ensure that the switch has the latest information on a controller mastership election, controllers coordinate the assignment of a Generation ID, also shown in the output. The Generation ID is a monotonically increasing 64-bit counter; therefore, any OFPT_ROLE_REQUEST message received with a role of master or slave with Generation ID of a lower value than one previously received is ignored.

The version, connection type, and Auxiliary ID have been previously described.

The output shows asynchronous filters (Async Fltr), dependent on the role that the controller is playing. A controller may use Asynchronous Configuration (OFPT_SET_ASYNC) messages to set a filter on the asynchronous messages that it receives from the switch. In the absence of an OFPT_SET_ASYNC message from the controller, the switch sets an initial configuration of asynchronous messages for Packet-In, Port-Status, and Flow-Removal messages and this configuration is shown. The remainder of the output (again truncated) shows detailed statistics for all message types sent and received over this channel.

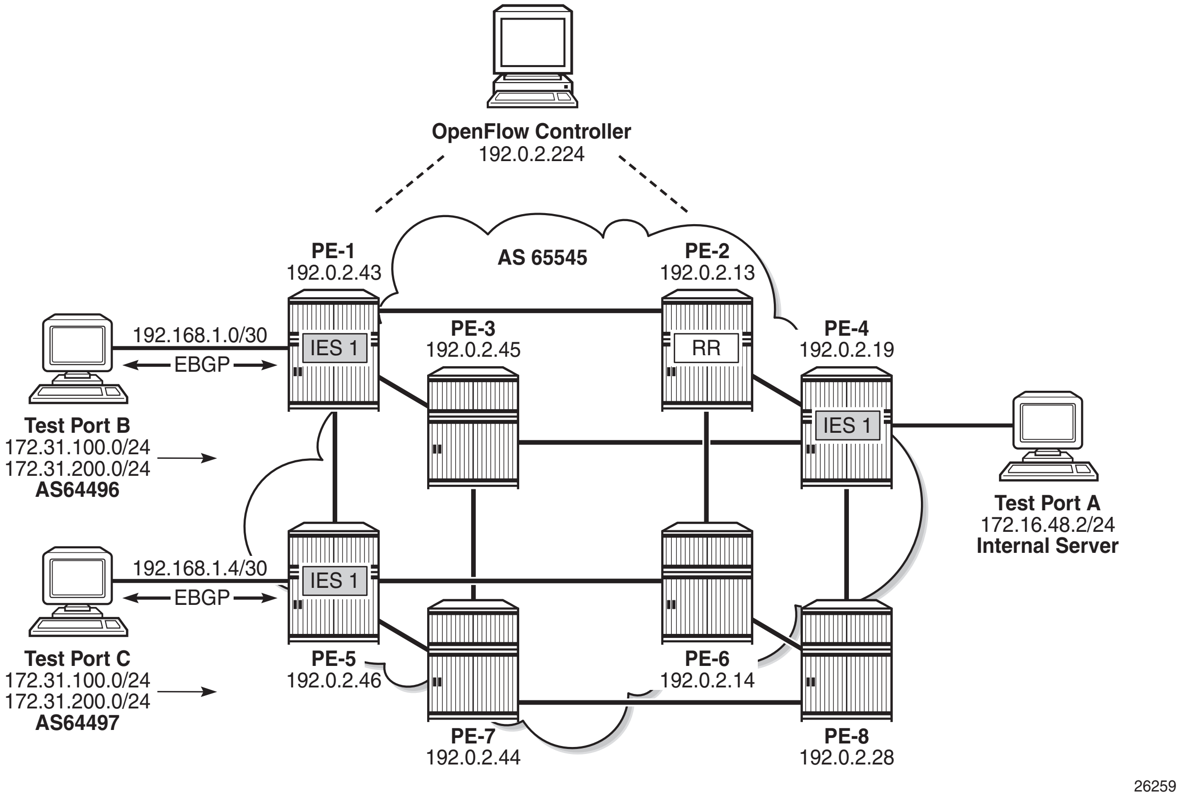
Dynamic Flow Entry Creation

With the basic switch configured and a channel established to the controller, the next step is to configure one or more IP filters that allow for dynamic embedding of OpenFlow flow entries. The following section will describe flow entries created in the base routing instance (also referred to as GRT), followed by entries created within a service instance.

Base Routing Instance

To generate rules within the base routing instance, the example topology is configured as shown in Figure 67. Test ports B and C simulate external peers located in AS 64496 and 64497, respectively. Both external peers advertise prefixes 172.31.100.0/24 and 172.31.200.0/24 to AS 64496, which are propagated internally within AS 65545. PE-4 hosts an internal server on subnet 172.16.48.0/24, which is advertised to the external peers. All three test ports are indexed to IES 1 at the corresponding PE router. BGP is configured within AS 65545 such that next-hops are resolved to shortcut tunnels using RSVP.

Figure 67 **OpenFlow Operation in Base Routing Context**



An IP filter is configured using the **embed-filter open-flow** command to allow for dynamic embedding of flow entries by an OpenFlow switch instance. In this example, the OpenFlow switch is the previously configured ofs-1. IP filters allow dynamically embedded OpenFlow filter entries to co-exist with static filter entries and other dynamic filter entries created by Flowspec or RADIUS. Therefore, an **offset** is defined that specifies the start point for dynamically created OpenFlow entries. This ensures that the OpenFlow flow entries can be isolated from other dynamic and static filter entries; FlowSpec filter entries must be created after static entries. In this example, the offset is 100.

```
configure
    filter
        ip-filter 10 create
            description "OpenFlow Basic GRT Filter"
            embed-filter open-flow "ofs-1" offset 100
        exit
    exit
```

The filter is applied as an ingress filter at PE-4 on the SAP connecting test port A, as follows:

```
configure
    service
        ies 1 customer 1 create
            interface "Test-Port-A" create
                address 172.16.48.1/24
                sap 3/1/4:10 create
                    ingress
                        filter ip 10
                    exit
                exit
```

Before any flow entries are initiated from the controller, a single entry with ID 65535 (maximum) is automatically populated in the embedding filter. This entry is inserted by OpenFlow when the **embed-filter open-flow** command is configured in the filter context and represents the table-miss entry. When OpenFlow uses filters, it ignores any **default-action** that may be configured in the filter so that filters can be chained. However, a table-miss action must exist and this is represented by entry 65535.

The source/destination addresses are 0.0.0.0/0 and the default primary action is forward (also referred to as fall-through). This primary action is configurable using the **no-match-action** command within the **flowtable 0** context. Other actions include **packet-in** or **drop**. When **packet-in** is configured and a packet of a flow matches entry 65535 (table-miss), SR OS sends only the first packet of that flow to the controller in a Packet-In message, while subsequent packets of that same flow are dropped.

```
*B:PE-4# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id          : 10                            Applied        : Yes
Scope              : Template                      Def. Action    : Drop
System filter      : Unchained
Radius Ins Pt      : n/a
CrCtl. Ins Pt      : n/a
RadSh. Ins Pt      : n/a
PccRl. Ins Pt      : n/a
Entries            : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description        : OpenFlow Basic GRT Filter
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry              : 65535
Origin             : Inserted by open-flow (no-match-action)
Description        : (Not Specified)
Log Id             : n/a
Src. IP            : 0.0.0.0/0
Src. Port          : n/a
Dest. IP           : 0.0.0.0/0
Dest. Port         : n/a
Protocol           : Undefined                     Dscp           : Undefined
ICMP Type          : Undefined                     ICMP Code      : Undefined
Fragment           : Off                           Src Route Opt  : Off
Sampling           : Off                           Int. Sampling  : On
IP-Option          : 0/0                           Multiple Option: Off
```

```
TCP-syn             : Off                          TCP-ack        : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Forward
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

===============================================================================
```

An OpenFlow IP filter is also automatically created by the system with a filter ID of _tmnx_ofs_<name>:<number>, where <name> is the name of the OpenFlow switch instance and <number> is a numerical integer. This is shown in the following output as _tmnx_ofs_ofs-1:8. This system-created filter ID contains all of the active flow entries dynamically created by the OpenFlow switch ofs-1 for the base (GRT) context, effectively acting as a repository for that routing context.

Any filter that is subsequently configured to dynamically embed GRT OpenFlow filter entries from the same OpenFlow switch instance will inherit all of the current entries contained in this filter. That is, if a new filter is configured to embed GRT OpenFlow entries from ofs-1, all of the flow-entries contained in _tmnx_ofs_ofs-1:8 will be automatically cloned into that new filter. There is no requirement for any active flow entries to be re-sent by the controller in order to populate this new filter. This approach allows filters to be enabled for OpenFlow embedding before or after flow entries have been received by the OpenFlow switch, thereby removing any order dependency.

```
*B:PE-4# show filter ip filter-type openflow
===============================================================================
Openflow IP Filters                                              Total:     1
===============================================================================
Filter-Id                     Description
-------------------------------------------------------------------------------
_tmnx_ofs_ofs-1:8             Filter for OFS 'ofs-1' for grt context
```

OpenFlow Filtering in Action

Before initiating any flow entries from the controller, the following traffic flows are sourced from test port A connected to PE-4.

- A UDP-based flow with a destination IP address of 172.31.100.1/24 at a rate of 1000 packets/s.
- A UDP-based flow with a destination address of 172.31.200.1/24, again at a rate of 1000 packets/s.

Both test port B and C are advertising the preceding prefixes, which are advertised internally by PE-1 and PE-5, respectively. At PE-4, the preferred next-hop for these prefixes is PE-1 (192.0.2.43).

```
*B:PE-4# show router bgp routes 172.31.0.0/16 longer
===============================================================================
 BGP Router ID:192.0.2.19       AS:65545       Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.31.100.0/24                                    100         0
      192.0.2.43                                         None        -
      64496 64500
u*>i  172.31.200.0/24                                    100         0
      192.0.2.43                                         None        -
      64496 64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
```

BGP next-hops are resolved to RSVP shortcut tunnels. For this test, there are two RSVP LSPs, one to PE-1 and one to PE-5, and they are viewed as logical ports by the OpenFlow switch. Because PE-1 is the preferred next-hop for the advertised prefixes, this resolves to the LSP named PE-4-PE-1-RSVP.

```
*B:PE-4# show open-flow of-switch "ofs-1" port
===============================================================================
Open Flow Port Stats
===============================================================================
Port ID        Type           Transmitted Packets      Transmitted Bytes
Port Name
-------------------------------------------------------------------------------
1073741833     logical        0                        0
PE-4-PE-1-RSVP
1073741834     logical        0                        0
PE-4-PE-5-RSVP
===============================================================================
```

The following output shows, as expected, that PE-1 is egressing traffic at a rate of 2000 packets/s toward test port B, representing the sum of the two 1000 packets/s test streams.

```
B:PE-1# monitor service id 1 sap 5/1/3:10 rate
===============================================================================
Monitor statistics for Service 1 SAP 5/1/3:10
===============================================================================
---snip---
-------------------------------------------------------------------------------
At time t = 11 sec (Mode: Rate)
-------------------------------------------------------------------------------
                        Packets                 Octets                  % Port
```

```
                                                                        Util.
Egress Queue 1
For. In/InplusProf    : 0                        7                      ~0.00
For. Out/ExcProf      : 2000                     1023907                0.08
Dro. In/InplusProf    : 0                        0                      0.00
Dro. Out/ExcProf      : 0                        0                      0.00
```

The controller initiates an OFPT_FLOW_MOD message containing an OFPFC_ADD command to the switch to create a new flow entry. The flow entry is viewed using the command shown in the following output:

```
*B:PE-4# tools dump open-flow of-switch "ofs-1"
===============================================================================
Switch: ofs-1
===============================================================================
Table     : 0                           Flow Pri  : 0
Cookie    : 0x0000000000000000          CookieType: grt
Controller: :::0
Filter Hnd: 0xC30000080000FFFF
Filter    : _tmnx_ofs_ofs-1:8 entry 65535

In Port   : *
VID       : *                           Outer VID : *
EthType   : *
Src IP    : *
Dst IP    : *
IP Proto  : *                           DSCP      : *
Src Port  : *                           Dst Port  : *
ICMP Type : *                           ICMP Code : *
Label     : *
IPv6ExtHdr: (Not Specified)

Action    : Fall-through

Flow Flags: IPv4/6 [!E] [RO] [DEF]
Up Time   : 0d 00:18:47                 Add TS    : 405757580
Mod TS    : 0                           Stats TS  : 405870240
#Packets  : 1638207                     #Bytes    : 838761984
-------------------------------------------------------------------------------
Table     : 0                           Flow Pri  : 1635
Cookie    : 0x0000000000000100          CookieType: grt
Controller: 192.0.2.224:6653
Filter Hnd: 0x830000080000F99C
Filter    : _tmnx_ofs_ofs-1:8 entry 63900

In Port   : *
VID       : *                           Outer VID : *
EthType   : 0x0800
Src IP    : *
Dst IP    : 172.31.100.0/24
IP Proto  : *                           DSCP      : *
Src Port  : *                           Dst Port  : *
ICMP Type : *                           ICMP Code : *
Label     : *

Action    : Forward LspId 10
            Lsp PE-4-PE-5-RSVP
```

```
Flow Flags: IPv4 [FR]
Up Time   : 0d 00:01:57                 Add TS    : 405858646
Mod TS    : 0                           Stats TS  : 405870241
#Packets  : 115951                      #Bytes    : 59366912
--------------------------------------------------------------------------------
Number of flows: 2
================================================================================
```

The first flow entry shown is the table-miss entry with an action of fall-through (or forward). The second entry contains the new flow entry.

The cookie associated with the message has a value of 0x0000000000000100 and, as shown in Table 2, because the high-order bits are set to zero, the cookie represents a flow entry that is used by filters within the base routing instance (shown as Global Routing Table or GRT). The filter used by this flow entry is _tmnx_ofs_ofs-1:8. This is the system-created OpenFlow IP filter for OpenFlow switch ofs-1 and contains all active GRT flow entries for that switch. Any filters that embed GRT OpenFlow entries from switch instance ofs-1 will automatically inherit all the active flow entries contained within this filter. In this example, the flow entries in _tmnx_ofs_ofs-1:8 are inherited only by IP filter 10. In addition, IP filter 10 will include ingress packets/bytes matched for each entry.

The priority field indicates a value of 1635 and, as previously described, determines the order with which flow entries are processed. Because OpenFlow states that the highest priority should be processed first and SR OS processes packets starting with the lowest numeric entry ID within a filter, the formula [65535 - flow_priority + embedding offset] is used to convert the cookie priority into a filter entry ID. This yields a filter entry ID of 63900. When a packet matches an entry in the filter, the packet is subject to the action defined in that entry, and is not subject to further filter entry processing.

The OpenFlow match fields specify an Ethertype of IPv4 (0x0800) and a destination prefix of 172.31.100.0/24, which are converted directly into filter entry match criteria. The OpenFlow instruction type is Write_Actions (OFPIT_WRITE_ACTIONS) in order to create the new flow, and has an action type of Output (OFPAT_OUTPUT). The output is directed to a (logical) port, which is the LSP PE-4-PE-5-RSVP.

The Modify Flow Entry (OFPT_FLOW_MOD) message contains a field for flags that are associated with each flow entry. These flags, together with some internal flags, are indicated in the Flow Flags field. Their meanings are described in Table 3.

Table 4 **FLOW_MOD Flags**

Flag	Meaning	Description
!E	Not evictable	Entry cannot be removed.
RO	Read only	Entry cannot be modified.

Table 4 **FLOW_MOD Flags (Continued)**

Flag	Meaning	Description
DEF	Default	
FR	SEND_FLOW_REM	If set, the switch must send a Flow-Removed message when the flow entry is deleted.
CO	CHECK_OVERLAP	If set, the switch must check that there are no conflicting entries with the same priority before inserting it into the flow entry table. An error is returned if a conflict exists.
RC	RESET_COUNTS	Reset flow packet and byte counts.
!PC	NO_PKT_COUNTS	When set, the switch does not need to keep track of the flow packet count.
!BC	NO_BYT_COUNTS	When set, the switch does not need to keep track of the byte count.

The dynamic OpenFlow flow entry redirecting traffic destined for prefix 172.31.100.0/24 is now in place as entry 63900 within IP filter _tmnx_ofs_ofs-1:8, and subsequently IP filter 10. The following two outputs show a monitor command run against PE-1's SAP toward test port B and PE-5's SAP toward test port C. Both SAPs are equally spreading the load of the two test streams of 1000 packets/s. Traffic for prefix 172.31.200.0/24 is routed toward PE-1 based on a route-table lookup. Traffic for prefix 172.31.100.0/24 is forwarded to PE-5 as a result of the OpenFlow redirect.

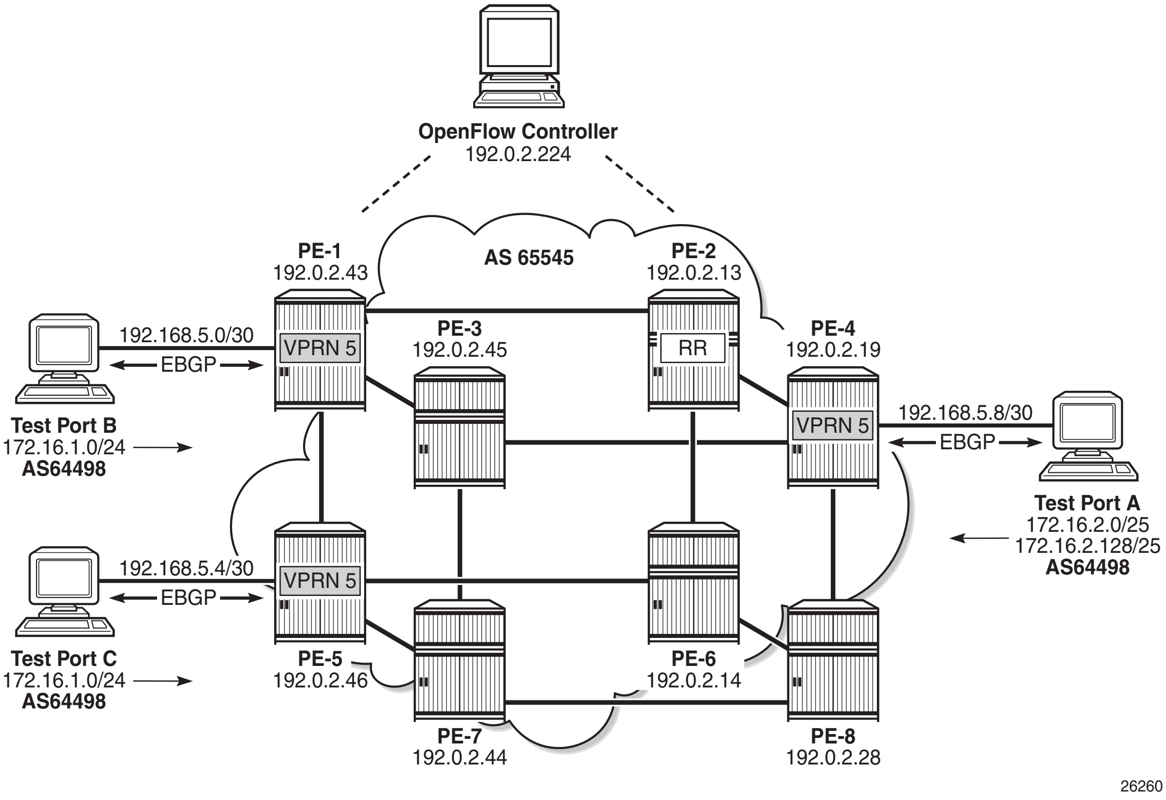
```
*A:PE-5# monitor service id 1 sap lag-1:10 rate
===============================================================================
Monitor statistics for Service 1 SAP lag-1:10
===============================================================================
---snip---
                        Packets                 Octets                  % Port
                                                                        Util.
Egress Queue 1
For. In/InplusProf    : 0                       0                       0.00
For. Out/ExcProf      : 1000                    512186                  0.04
Dro. In/InplusProf    : 0                       0                       0.00
Dro. Out/ExcProf      : 0                       0                       0.00


B:PE-1# monitor service id 1 sap 5/1/3:10 rate
===============================================================================
Monitor statistics for Service 1 SAP 5/1/3:10
===============================================================================
---snip---
                        Packets                 Octets                  % Port
                                                                       Util.
Egress Queue 1
For. In/InplusProf    : 0                       7                       ~0.00
For. Out/ExcProf      : 1000                   512186                 0.04
Dro. In/InplusProf    : 0                      0                      0.00
Dro. Out/ExcProf      : 0                      0                      0.00
```

FLOW_MOD messages allow for flow entries to be associated with hard and idle timeouts, which are not currently used by SR OS. Although timeout values can be passed by a controller in a FLOW_MOD message, they are effectively ignored. As a result, dynamic flow entries remain in place as filter entries until removed by the controller, or the OpenFlow switch instance is placed in a **shutdown** state. If the LSP transitions to an operationally down state while the redirect flow entry is still active, the switch will notify the controller of the change of state using a Port-Status message, and traffic will be subject to a forward action. If the LSP becomes operational again, the flow entry becomes active again.

Service Routing Instance

To generate rules within a VPRN routing instance, the example topology is configured as shown in Figure 68. Test ports A, B, and C belong to VPRN 5. Test ports B and C simulate CE routers in a dual-homed site, advertising prefix 172.16.1.0/24 in EBGP to PE-1 and PE-5, respectively. Test port A simulates a CE router at a different site, advertising prefixes 172.16.2.0/25 and 172.16.2.128/25 in EBGP to PE-4. The PE to CE (WAN) links are also advertised into VPN-IPv4 by the respective PE routers, to provide complete visibility of the VPN.

Figure 68 **Example Topology for OpenFlow within a Service Routing Context**



An IP filter is configured using the **embed-filter open-flow** command to allow for dynamic embedding of flow entries by an OpenFlow switch instance. In this example, the OpenFlow switch remains as ofs-1. The command also specifies **service 5** to make this filter applicable to interfaces within that service instance. Thereafter, this filter can only be deployed in the configured service. An **offset** is also defined to specify the start point for dynamically created OpenFlow entries and allow them to remain isolated from other dynamic and static filter entries.

```
configure
    filter
        ip-filter 20 create
            description "OpenFlow Service Filter"
            embed-filter open-flow "ofs-1" service 5 offset 100
        exit
```

The **embed-filter** command has the option to configure a service ID or a SAP ID. The former is applicable to embedding filters applied in VPRN or VPLS services. The latter is applicable only to VPLS services. It requires that the embedding filter has the scope of exclusive (as opposed to the default scope of template) and can only be deployed on the SAP specified in the argument.

The filter is applied at PE-4 on the SAP connecting test port A, as follows:

```
configure
    service
        vprn 5 customer 1 create
            interface "Test-Port-A" create
                address 192.168.5.9/30
                sap 3/1/4:5 create
                    ingress
                        filter ip 20
                    exit
                exit
            exit
```

As with the example of the base routing context, before any flow entries are initiated from the controller, a single entry with ID 65535 (maximum) is automatically populated in the filter, representing the table-miss entry. As before, when OpenFlow uses filters, it ignores any **default-action** that may be configured in the filters, so that filters can be chained. However, a table-miss action must exist and this is represented by entry 65535, as follows:

```
B:PE-4# show filter ip 20
===============================================================================
IP Filter
===============================================================================
Filter Id           : 20                           Applied        : Yes
Scope               : Template                     Def. Action    : Drop
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : OpenFlow Service Filter
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 65535
Origin              : Inserted by open-flow (no-match-action)
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 0.0.0.0/0
Src. Port           : n/a
Dest. IP            : 0.0.0.0/0
Dest. Port          : n/a
Protocol            : Undefined                    Dscp           : Undefined
ICMP Type           : Undefined                    ICMP Code      : Undefined
Fragment            : Off                          Src Route Opt  : Off
Sampling            : Off                          Int. Sampling  : On
IP-Option           : 0/0                          Multiple Option: Off
TCP-syn             : Off                          TCP-ack        : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Forward
Ing. Matches        : 8193635 pkts (4194988287 bytes)
Egr. Matches        : 0 pkts
```

```
===============================================================================
```

An OpenFlow IP filter, _tmnx_ofs_ofs-1:16, is also automatically created by the system and contains all of the flow entries dynamically created by the OpenFlow switch ofs-1 for service ID 5. This filter acts as a repository for active flow entries specific to that service context and its purpose has been previously described. If a new filter is configured to embed OpenFlow entries for service ID 5, the entries from _tmnx_ofs_ofs-1:16 will be cloned into that new filter.

```
B:PE-4# show filter ip filter-type openflow
===============================================================================
Openflow IP Filters                                              Total:     2
===============================================================================
Filter-Id                     Description
-------------------------------------------------------------------------------
_tmnx_ofs_ofs-1:15            Filter for OFS 'ofs-1' for grt context
_tmnx_ofs_ofs-1:16            Filter for OFS 'ofs-1' for service [5] context
===============================================================================
```

OpenFlow Filtering in Action

To validate flow entries initiated by the controller, the following traffic flows are sourced from test port A connected to PE-4:

- A UDP-based flow with a source address of 172.16.2.1/25 and a destination address of 172.16.1.1/24 at a rate of 1000 packets/s.
- A UDP-based flow with a source address of 172.16.2.129/25 and a destination address of 172.16.1.1/24, again at a rate of 1000 packets/s.

Both test port B and C are advertising the preceding prefixes, which are advertised internally in VPN-IPv4 by PE-1 and PE-5, respectively. At PE-4, the preferred next-hop for 172.16.1.0/24 within VPRN 5 is PE-1 (192.0.2.43), as follows:

```
B:PE-4# show router 5 route-table 172.16.1.0/24
===============================================================================
Route Table (Service: 5)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.0/24                                 Remote  BGP VPN   01h47m36s  170
       192.0.2.43 (tunneled:RSVP:9)                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

PE-1 is egressing traffic at a rate of 2000 packets/s toward test port B, representing the sum of the two 1000 packets/s test streams, as follows:

```
B:PE-1# monitor service id 1 sap 5/1/3:10 rate
===============================================================================
Monitor statistics for Service 1 SAP 5/1/3:10
===============================================================================
---snip---
-------------------------------------------------------------------------------
At time t = 22 sec (Mode: Rate)
-------------------------------------------------------------------------------
                        Packets                 Octets                  % Port
                                                                        Util.
Egress Queue 1
For. In/InplusProf    : 0                       7                       ~0.00
For. Out/ExcProf      : 2000                    1023907                 0.08
Dro. In/InplusProf    : 0                       0                       0.00
Dro. Out/ExcProf      : 0                       0                       0.00
```

An OFPT_FLOW_MOD message containing an OFPFC_ADD command is initiated by the controller and can be viewed using the command in the following output:

```
B:PE-4# tools dump open-flow of-switch "ofs-1"

===============================================================================
Switch: ofs-1
===============================================================================
Table     : 0                           Flow Pri  : 0
Cookie    : 0x0000000000000000          CookieType: grt
Controller: :::0
Filter Hnd: 0xC300000F0000FFFF
Filter    : _tmnx_ofs_ofs-1:15 entry 65535

In Port   : *
VID       : *                           Outer VID : *
EthType   : *
Src IP    : *
Dst IP    : *
IP Proto  : *                           DSCP      : *
Src Port  : *                           Dst Port  : *
ICMP Type : *                           ICMP Code : *
Label     : *
IPv6ExtHdr: (Not Specified)

Action    : Fall-through

Flow Flags: IPv4/6 [!E] [RO] [DEF]
Up Time   : 0d 05:01:44                 Add TS    : 422384764
Mod TS    : 0                           Stats TS  : 424195136
#Packets  : 27501425                    #Bytes    : 14080300394
-------------------------------------------------------------------------------
Table     : 0                           Flow Pri  : 1535
Cookie    : 0xC000000500000038          CookieType: service 5
Controller: 192.0.2.224:6653
Filter Hnd: 0x830000100000FA00
Filter    : _tmnx_ofs_ofs-1:16 entry 64000
```

```
In Port   : *
VID       : *                           Outer VID : *
EthType   : 0x0800
Src IP    : 172.16.2.128/25
Dst IP    : *
IP Proto  : *                           DSCP      : *
Src Port  : *                           Dst Port  : *
ICMP Type : *                           ICMP Code : *
Label     : *

Action    : Forward On Nhop(Indirect)
            Nhop: 192.168.5.6

Flow Flags: IPv4 [FR]
Up Time   : 0d 00:00:44                 Add TS    : 424190707
Mod TS    : 0                           Stats TS  : 424195137
#Packets  : 44301                       #Bytes    : 22682112
-------------------------------------------------------------------------------
Number of flows: 2
===============================================================================
```

The first flow entry with cookie value 0x0000000000000000 is the table-miss entry with a fall-through or forward action. The second entry with cookie value 0xC000000500000038 contains the new flow entry. The high-order bits of the cookie are set to 0xC (or 1100), which (as shown in Table 2) means that this represents a flow entry that is used by filters used within a service instance. Bits 59 to 32 encode the service instance, which in this case is 5.

The filter used by this second flow entry is _tmnx_ofs_ofs-1:16, which is the system-created OpenFlow filter for OpenFlow switch ofs-1, and contains all active flows entries initiated by that switch for service ID 5. Any filters embedding OpenFlow flow entries from ofs-1 in service ID 5 will clone all of the entries contained in _tmnx_ofs_ofs-1:16. In this example, the entries in _tmnx_ofs_ofs-1:16 are cloned into IP filter 20. IP filter 20 will also include ingress packets/bytes matched for each entry.

The priority field indicates a value of 1535 and, as previously described, determines the order in which flow entries are processed, using the formula [65535 - flow_priority + embedding offset].

The OpenFlow Match fields specify an Ethertype of IPv4 (0x0800) for source prefix 172.16.2.128/25, and are mapped directly into filter entry match criteria. The OpenFlow instruction type is Write_Actions (OFPIT_WRITE_ACTIONS) in order to create the new flow entry, and has an action type of Forward to Next-Hop IP Address. Because OpenFlow has no standard action type of Forward to Next-Hop IP Address, an Experimenter (OFPAT_EXPERIMENTER) is used for this purpose, which encompasses the use of both direct and indirect next-hops. In this example, an indirect next-hop of 192.168.5.6 is used.

The preferred next-hop for traffic destined to prefix 172.16.1.0/24 is PE-1. The indirect next-hop address of 192.168.5.6 represents the (simulated) CE WAN address of test port C, and is known in the routing table of VPRN 5 with a next-hop of PE-5 (192.0.2.46), as follows:

```
B:PE-4# show router 5 route-table 192.168.5.6
===============================================================================
Route Table (Service: 5)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.168.5.4/30                                Remote  BGP VPN   19h19m22s  170
       192.0.2.46 (tunneled:RSVP:10)                                0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

The dynamic OpenFlow flow entry redirecting traffic with a source address of 172.16.2.128/25 is now in place as entry 64000 within IP filter _tmnx_ofs_ofs-1:16, and cloned into IP filter 20. The effect of this OpenFlow flow entry on the two test streams is as follows:

- Traffic sourced from prefix 172.16.2.0/25 to prefix 172.16.1.0/24 is routed in accordance with the VPRN 5 routing table, with a next-hop of PE-1.
- Traffic sourced from prefix 172.16.2.128/25 to prefix 172.16.1.0/24 is subject to policy-based routing and, rather than being routed directly toward the destination prefix known via PE-1, is forwarded to an indirect next-hop of 192.168.5.6, known via PE-5.

This is validated in the following two outputs, which show a monitor command run against PE-1's SAP toward test port B and PE-5's SAP toward test port C. The outputs show that each SAP is egressing 1000 packets/s:

```
B:PE-1# monitor service id 5 sap 5/1/3:5 rate
===============================================================================
Monitor statistics for Service 5 SAP 5/1/3:5
===============================================================================
---snip---
                        Packets                 Octets                  % Port
                                                                        Util.
Egress Queue 1
For. In/InplusProf    : 0                       8                       ~0.00
For. Out/ExcProf      : 1000                    512000                  0.04
Dro. In/InplusProf    : 0                       0                       0.00
Dro. Out/ExcProf      : 0                       0                       0.00


A:PE-5# monitor service id 5 sap lag-1:5 rate
```

```
===============================================================================
Monitor statistics for Service 5 SAP lag-1:5
===============================================================================
---snip---
                        Packets                 Octets                  % Port
                                                                        Util.
Egress Queue 1
For. In/InplusProf    : 0                       0                       0.00
For. Out/ExcProf      : 1000                    512000                  0.04
Dro. In/InplusProf    : 0                       0                       0.00
Dro. Out/ExcProf      : 0                       0                       0.00
```

As previously described, dynamic flow entries will remain in place as filter entries until removed by the controller or the OpenFlow switch instance is put in a **shutdown** state.

Supported Redirect Actions

Table 4 lists the redirect actions supported in SR OS together with the applicability and associated action types. Experimenter encodings are described in user guides. Unless otherwise stated, all instruction types are WRITE_ACTION/APPLY_ACTION.

Table 5 **Supported Redirect Actions**

Action	Applicability	Action Type	Remarks
Redirect to IP Next-Hop	IPv4/IPv6 traffic ingressing an IP interface	OFPAT_EXPERIMENTER (ALU_AXN_REDIRECT_TO_NEXTHOP)	Next-hop can be direct or indirect
Redirect to Routing Context (GRT or VRF)		OFPAT_OUTPUT <logical_port> <logical_port> encoding: Bits 31-28=0100, bits 27-24=0001, bits 23-0=VPRN service ID or 0 for GRT	
Redirect to Next-Hop and VRF/GRT Routing Context		Action 1: OFPAT_EXPERIMENTER (ALU_AXN_REDIRECT_TO_NEXTHOP)	Next-hop must be indirect
		Action 2: OFPAT_OUTPUT <logical_port> <logical_port> encoding: Bits 31-28=0100, bits 27-24=0001, bits 23-0=VPRN service ID or 0 for GRT	
Redirect to LSP		OFPAT_OUTPUT <logical_port> <logical_port> encoding: Bits 31-28=0100, bits 27-24=0000, bits 23-0=RSVP-TE Tunnel ID	

Table 5 **Supported Redirect Actions (Continued)**

Action	Applicability	Action Type	Remarks
Redirect to SAP	Traffic ingressing a VPLS interface	Action 1: OFPAT_OUTPUT <port> <port> encoding: OXM_OF_IN_PORT: TmnxPortID for Ethernet port or LAG	TmnxPortId encoding in TIMETRA-CHASSIS-MIB (port) or LAG TIMETRA-TC-MIB (LAG)
		Action 2: OFPAT_SET_FIELD <vlan_encoding> VLAN encoding: OXM_OF_VLAN_ID (null, dot1Q, or inner QinQ tag) Optional EXPERIMENTER OFL_OUT_VLAN_ID (outer QinQ tag)	
Redirect to SDP		OFPAT_EXPERIMENTER (ALU_AXN_REDIRECT_TO_SDP)	Possible to match against entire SAP using OXM_OF_IN_PORT encoding TmnxPortID, OXM_OF_VLAN_ID (null tag, dot1Q tag, inner Q-in-Q tag) and optional EXPERIMENTER OFL_OUT_VLAN_ID (outer Q-in-Q tag)

Resource Consumption

Dynamic OpenFlow flow entries are embedded in filters as filter entries, and as such, consume CAM entries in the same way as statically configured filter entries and/or other dynamic filter entries, such as those created by BGP FlowSpec or RADIUS. When a flow entry is created and dynamically embedded as a filter entry, it will consume one or more ingress ACL/QoS entries from the line card to which the filter is attached. If a flow entry is embedded in multiple filters, an ingress ACL/QoS entry will be consumed for each filter. If a flow entry is embedded in a single filter with a default scope of template, and this filter is attached to multiple SAPs on the same line card, only a single entry is consumed.

As with conventional ACL resource consumption, a standard four- or five-tuple match will consume a single entry. Defining a range of ports, for example, will consume multiple entries, as follows:

```
B:PE-4# tools dump system-resources 3
Resource Manager info at 049 d 12/01/16 09:10:18.148:
Hardware Resource Usage for Slot #3, CardType imm12-10gb-sf+, Cmplx #0:
                                 |   Total   | Allocated |    Free
  -------------------------------|-----------|-----------|------------
---snip---
         Ingress ACL/QoS Entries |      65536|          5|      65531
---snip---
```

Debugging

A number of OpenFlow debug commands are available. For troubleshooting and interoperability purposes, detailed packet-level debug commands are available for all OpenFlow message types. Also, the ability to debug OpenFlow switch errors is useful. An example is provided in the following output:

```
debug
    open-flow
        of-switch "ofs-1"
            error
            packet flow-mod detail
        exit
    exit
exit
```

Conclusion

OpenFlow has a number of use-cases in the WAN. The dynamic insertion of flow entries from a controller can be used for flow placement in an SDN environment implementing some business logic. Equally, it could be used to implement security measures, or off-ramping of traffic to a DDoS scrubbing center.

This chapter described how to configure and deploy Hybrid OpenFlow in SR OS. It described how to configure the OpenFlow switch, and how filter entries are dynamically embedded in GRT filters and service filters. These examples are intended to provide an overview of functionality.

LFA Policies Using OSPF as IGP

This chapter provides information about LFA policies using OSPF as IGP.

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS release 12.0.R4, but the CLI in this edition corresponds to release 14.0.R2.

Overview

Loop Free Alternate (LFA) is a local control plane feature. When multiple LFAs exist, RFC 5286, *Basic Specification for IP Fast Reroute: Loop-Free Alternates*, chooses the LFA providing the best coverage of the failure cases. In general, this means that node LFA has preference above link LFA. In some deployments, however, this can lead to suboptimal LFA. For example, an aggregation router (typically using lower bandwidth links) protecting a core node/link (typically using high bandwidth links) is potentially undesirable.

For this reason, the operator wants to have more control in the LFA next-hop selection algorithm. This is achieved by the introduction of LFA Shortest Path First (SPF) policies.

LFA policies can work in combination with IP fast reroute (FRR) and/or LDP FRR.

Implementation

The SROS LFA policy implementation is built around the concept of route-next-hop (NH) templates which are applied to IP interfaces. A route-next-hop template specifies criteria which influence the selection of an LFA backup NH for either:

- a set of prefixes in a prefix-list or
- a set of prefixes which resolve to a specific primary NH

See http://tools.ietf.org/html/draft-litkowski-rtgwg-lfa-manageability for further information. Two powerful methods which can be used as criteria inside a route-next-hop template are IP admin groups and IP shared risk link groups (SRLGs). IP SRLG and IP admin group criteria are applied before running the LFA NH algorithm. IP admin groups and IP SRLGs work in a similar way as the well-known MPLS admin groups and MPLS SRLGs.

For example, when one or more IP admin groups/SRLGs are applied to an IP interface, the same MPLS admin group/SRLG rules apply:

- IP interfaces which do not include one or more of the admin groups defined in the **include** statements are pruned before computing the LFA next-hop.
- IP interfaces which belong to admin groups which have been explicitly excluded using the **exclude** statement are pruned before computing the LFA next-hop.
- IP interfaces which belong to the SRLGs used by the primary NH of a prefix are pruned before computing the LFA next-hop.

For more information about MPLS admin groups, see chapter RSVP Point-to-Point LSPs; for SRLGs, see chapter Shared Risk Link Groups for RSVP-Based LSP.

For compatibility reasons with the existing MPLS, admin groups and SRLGs, a single set of admin groups and SRLGs are defined within the **configure router if-attribute** context .

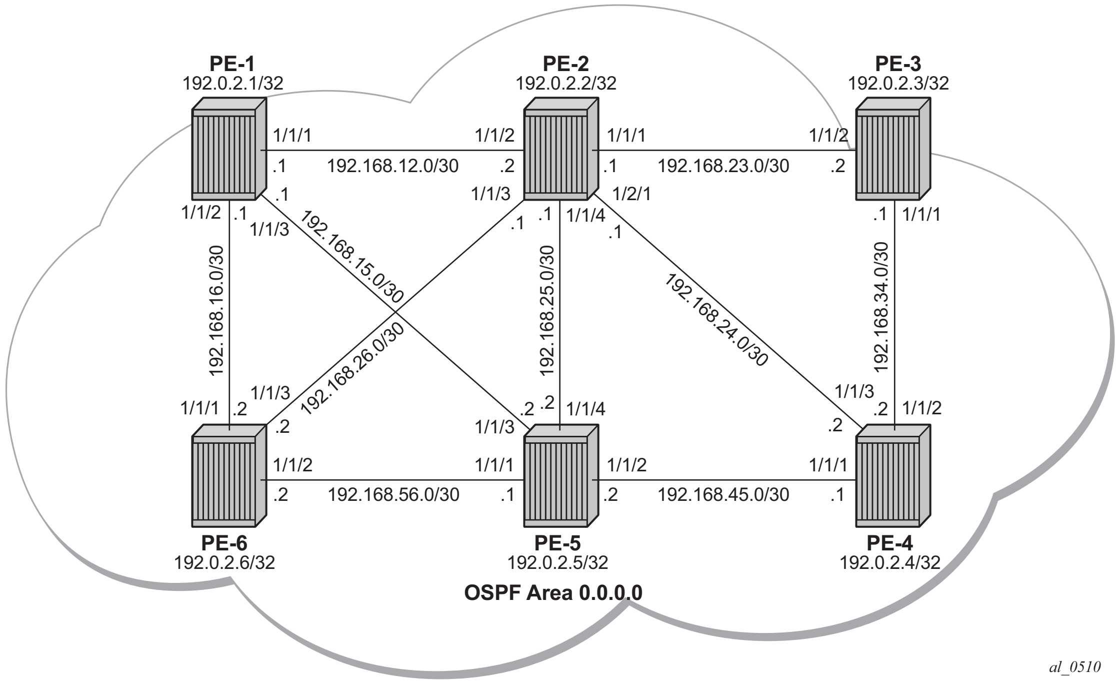
Once one or more admin groups or SRLGs have been defined, theyare applied to an MPLS interface and/or an IP interface.

In the SR OS implementation, IP admin groups and SRLGs are locally significant, meaning they are not advertised by the IGP.

The well-known MPLS admin groups and SRLGs are advertised in TE link TLVs and sub-TLVs when the traffic-engineering option is enabled in the IGP protocol.

Other selection criteria which can be configured inside a route-next-hop template are protection type preference and NH type preference. More details on these parameters are provided later in this chapter.

Figure 69 **Example Topology**



Configuration

Step 1. Configuring an IP/MPLS network with LDP FRR enabled on PE-7.

Since the focus is not on how to set up an IP/MPLS network, only summary bullets are provided.

– The system and IP interface addresses are configured according to Figure 69.

– OSPF area 0 is selected as the interior gateway protocol (IGP) to distribute routing information between all PEs. All OSPF interfaces are set up as **type point-to-point** to avoid running the designated router/backup designated router (DR/BDR) election process. All links have the default metric, which is 10 in this example, except for interface “int-PE-2-PE-5” on PE-2, which is configured with a metric of 20.

– Enable link LDP on all interfaces. This establishes a full mesh of LDP LSPs between all PE system interfaces. As an example, the tunnel-table on PE-2 looks as follows:

```
*A:PE-2# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
```

```
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        192.168.12.1   10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.23.2   10
192.0.2.4/32      ldp       MPLS  65539     9        192.168.24.2   10
192.0.2.5/32      ldp       MPLS  65540     9        192.168.12.1   20
192.0.2.6/32      ldp       MPLS  65541     9        192.168.26.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

The LDP LSP metric follows the IGP cost.

– Enable LDP FRR on PE-2. This is a two-fold configuration command: first the IGP needs to be triggered to do LFA NH computation, and secondly, FRR needs to be enabled within the LDP context. This is configured as follows:

```
*A:PE-2# configure router ospf loopfree-alternate
*A:PE-2# show router ospf status | match LFA
LFA                           : Enabled
Remote-LFA                    : Disabled
```

Remote LFA can be enabled for segment routing, but this is beyond the scope of this chapter.

```
*A:PE-2# configure router ldp fast-reroute
*A:PE-2# show router ldp status | match FRR
FRR                 : Enabled              Mcast Upstream FRR   : Disabled
```

Multicast upstream FRR is for multicast LDP and is beyond the scope of this chapter.

After issuing these two CLI commands, the software pre-computes both a primary and a backup Next-hop Label Forwarding Entry (NHLFE) for each LDP FEC in the network and downloads them into the IOM/IMM. The primary NHLFE corresponds to the label of the FEC received from the primary NH as per standard LDP resolution of the FEC prefix in the Routing Table Manager (RTM). The backup NHLFE corresponds to the label received for the same FEC from an LFA NH. The **show router route-table alternative** command adds an LFA flag to the associated alternative NH for a specific destination prefix. Other useful IGP related show commands are **show router ospf lfa-coverage** and **show router ospf routes alternative detail**.

```
*A:PE-2# show router route-table alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      00h11m32s  10
       192.168.12.1                                                 10
       192.168.26.2 (LFA)                                           20
192.0.2.2/32                                  Local   Local     00h11m44s  0
       system                                                       0
192.0.2.3/32                                  Remote  OSPF      00h11m18s  10
       192.168.23.2                                                 10
       192.168.24.2 (LFA)                                           20
192.0.2.4/32                                  Remote  OSPF      00h11m12s  10
       192.168.24.2                                                 10
       192.168.23.2 (LFA)                                           20
192.0.2.5/32                                  Remote  OSPF      00h11m02s  10
       192.168.12.1                                                 20
       192.168.24.2 (LFA)                                           20
192.0.2.6/32                                  Remote  OSPF      00h10m54s  10
       192.168.26.2                                                 10
       192.168.12.1 (LFA)                                           20
192.168.12.0/30                               Local   Local     00h11m44s  0
       int-PE-2-PE-1                                                0
192.168.15.0/30                               Remote  OSPF      00h11m32s  10
       192.168.12.1                                                 20
       192.168.26.2 (LFA)                                           30
192.168.16.0/30                               Remote  OSPF      00h11m32s  10
       192.168.12.1                                                 20
       192.168.26.2 (LFA)                                           30
192.168.23.0/30                               Local   Local     00h11m44s  0
       int-PE-2-PE-3                                                0
192.168.24.0/30                               Local   Local     00h11m44s  0
       int-PE-2-PE-4                                                0
192.168.25.0/30                               Local   Local     00h11m44s  0
       int-PE-2-PE-5                                                0
192.168.26.0/30                               Local   Local     00h11m44s  0
       int-PE-2-PE-6                                                0
192.168.34.0/30                               Remote  OSPF      00h11m18s  10
       192.168.23.2                                                 20
       192.168.24.2 (LFA)                                           30
192.168.45.0/30                               Remote  OSPF      00h11m12s  10
       192.168.24.2                                                 20
       192.168.23.2 (LFA)                                           30
192.168.56.0/30                               Remote  OSPF      00h10m54s  10
       192.168.26.2                                                 20
       192.168.12.1 (LFA)                                           30
-------------------------------------------------------------------------------
No. of Routes: 16
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Displaying the Label Forwarding Information Base (LFIB) on PE-2 shows the available alternate NHs; displayed with the BU flag.

```
*A:PE-2# show router ldp bindings active prefixes ipv4
```

```
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Push              --      262142BU
192.168.26.2                                1/1/3

192.0.2.1/32                                Swap            262142    262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap            262142    262142BU
192.168.26.2                                1/1/3

192.0.2.2/32                                Pop             262143      --
  --                                          --

192.0.2.3/32                                Push              --      262143
192.168.23.2                                1/1/1

192.0.2.3/32                                Push              --      262140BU
192.168.24.2                                1/2/1

192.0.2.3/32                                Swap            262141    262143
192.168.23.2                                1/1/1

192.0.2.3/32                                Swap            262141    262140BU
192.168.24.2                                1/2/1

192.0.2.4/32                                Push              --      262143
192.168.24.2                                1/2/1

192.0.2.4/32                                Push              --      262140BU
192.168.23.2                                1/1/1

192.0.2.4/32                                Swap            262140    262143
192.168.24.2                                1/2/1

192.0.2.4/32                                Swap            262140    262140BU
192.168.23.2                                1/1/1

192.0.2.5/32                                Push              --      262139
192.168.12.1                                1/1/2
```

```
192.0.2.5/32                                Push              --      262139BU
192.168.24.2                                1/2/1

192.0.2.5/32                                Swap            262139    262139
192.168.12.1                                1/1/2

192.0.2.5/32                                Swap            262139    262139BU
192.168.24.2                                1/2/1

192.0.2.6/32                                Push              --      262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Push              --      262138BU
192.168.12.1                                1/1/2

192.0.2.6/32                                Swap            262138    262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Swap            262138    262138BU
192.168.12.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 21
===============================================================================
*A:PE-2#
```

Finally, a synchronization timer is enabled between the IGP and LDP protocol when LDP FRR is enabled. From the moment that the interface for the previous primary NH is restored, the IGP may re-converge back to that interface before LDP has completed the FEC exchange with its neighbor over that interface. This may cause LDP to de-program the LFA NH from the FEC and blackhole the traffic. In this example a timer of 10 seconds is used. This is configured as follows:

```
*A:PE-x# configure router interface <itf-name> ldp-sync-timer 10
```

When this timer is set, on restoring a failed interface, the IGP advertises this link into the network with an infinite metric for the duration of this timer. When the failed link is restored, the **ldp-sync-timer** is started, and LDP adjacencies are brought up over the restored link and a label exchange is completed between the peers. After the **ldp-sync-timer expires**, the normal metric is advertised into the network again.

At this point, everything is in place to start creating LFA policies to influence the calculated LFA NHs.

Step 2. Create a route-next-hop policy template.

This is a mandatory step in the context of LFA policies. The route-next-hop template name is 32 characters at maximum. Creating a route-next-hop policy is done in the following way:

```
*A:PE-x# configure router route-next-hop-policy template <template name>
```

Commands within a **route-next-hop** policy template follow the **begin-abort-commit** model. After a **commit**, the IGP re-evaluates the template and schedules a new LFA SPF to re-compute the LFA NH for the prefixes associated with this template.

Step 3. Configure admin group constraints in route-next-hop policy.

This is an optional step in the context of LFA policies. Firstly, configure a group-name and a group-value for each admin group locally on the router. This is configured as follows:

```
*A:PE-x# configure router if-attribute admin-group <group-name> value <group-value>
```

Second, configure the admin group membership of the IP interface(s) (network, IES or VPRN). Up to five admin groups can be applied to an IP interface in one command but the command can be applied multiple times. The configured IP admin group membership applies to all levels/areas the interface is participating in, as follows:

```
*A:PE-x# configure router interface <itf-name> if-attribute admin-group <group-
name> [ <group-name> ... (upto 5 max)]
*A:PE-x# configure service vprn <svc-id> interface <itf-name> if-attribute admin-
group <group-name> [ <group-name> ... (upto 5 max)]
*A:PE-x# configure service ies <svc-id> interface <itf-name> if-attribute admin-
group <group-name> [ <group-name> ... (upto 5 max)]
```

Third, add the IP admin group constraints to the route-next-hop policy template one by one. The **include-group** statement instructs the LFA SPF selection algorithm to select a subset of LFA NHs among the links which belong to one or more of the specified admin groups. A link which does not belong to at least one of the admin groups is excluded. The **pref** option is used to provide a relative preference for the admin group selection. A lower preference value means that LFA SPF will first attempt to select an LFA backup NH which is a member of the corresponding admin group. If none is found, then the admin group with the next higher preference value is evaluated. If no preference is configured, then it is the least preferred (default preference value is 255).

When evaluating multiple **include-group** statements having the same preference, any link which belongs to one or more of the included admin groups can be selected as an LFA next-hop. There is no relative preference based on how many of those included admin groups the link is a member.

The **exclude-group** command simply prunes all links belonging to the specified admin group before making the LFA backup NH selection for a prefix. If the same group name is part of both **include** and **exclude** statements, the exclude statement will takes precedence. In other words, the **exclude** statement can be viewed as having an implicit preference value of 0.

This is configured as follows:

```
*A:PE-x# configure router route-next-hop-policy template <template-name> exclude-
group <group-name>
*A:PE-x# configure router route-next-hop-policy template <template-name> include-
group <group-name> [pref <preference>]
```

Step 4. Configure SRLG constraints in route-next-hop policy.

This is an optional step in the context of LFA policies. Firstly, configure a group-name and group-value, of each SRLG group locally on the router. This is configured as follows:

```
*A:PE-x# configure router if-attribute srlg-group <group-name> value <group-
value> [penalty-weight <penalty-weight>]
```

Second, configure the SRLG group membership of the IP interfaces (network, IES or VPRN). Up to five SRLG groups can be applied to an IP interface in one command but the command can be applied multiple times. The configured IP SRLG group membership is applied in all levels/areas the interface is participating in. This is configured as follows:

```
*A:PE-x# configure router interface <itf-name> if-attribute srlg-group <group-
name> [ <group-name> ... (upto 5 max)]
*A:PE-x# configure service vprn <svc-id> interface <itf-name> if-attribute srlg-
group <group-name> [ <group-name> ... (upto 5 max)]
*A:PE-x# configure service ies <svc-id> interface <itf-name> if-attribute srlg-
group <group-name> [ <group-name> ... (upto 5 max)]
```

Third, add IP SRLG group constraints to the route-next-hop policy template. When this command is applied to a prefix, the LFA SPF attempts to select an LFA NH which uses an outgoing interface that does not participate in any of the SRLGs of the outgoing interface used by the primary NH. This is configured as follows:

```
*A:PE-x# configure router route-next-hop-policy template <template-name> srlg-
enable
```

Step 5. Configure the protection type in route-next-hop policy.

This is an optional step in the context of LFA policies. With the use of LFA policies, the user can also select if link protection or node protection is preferred for IP prefixes and LDP FEC prefixes protected by a backup LFA NH. By default, node protection is chosen. The implementation falls back to link protection if no LFA NH is found for node protection. This is configured as follows:

```
*A:PE-x# configure router route-next-hop-policy template <template-name> protection-
type {link|node}
```

Step 6. Configure the NH preference type in route-next-hop policy.

This is an optional step in the context of LFA policies. With the use of LFA policies, the user can also select if tunnel backup NH or IP backup NH is preferred for IP prefixes and LDP FEC prefixes protected by a backup LFA NH. By default, IP backup NH is chosen. The implementation falls back to the other type (tunnel) if no LFA NH of the preferred type is found. This is configured as follows:

```
*A:PE-x# configure router route-next-hop-policy template <template-name> nh-
type {ip|tunnel}
```

Step 7. Apply the route-next-hop policy template to an IP interface.

When the route-next-hop policy is applied to an IP interface, all prefixes using this interface as primary NH take the selection criteria specified in Step 3, Step 4, Step 5 and Step 6 into account. This is configured as follows:

```
*A:PE-x# configure router ospf area interface lfa-policy-map route-nh-
template <template-name>
*A:PE-x# configure router ospf3 area interface lfa-policy-map route-nh-
template <template-name>
*A:PE-x# configure service vprn ospf area interface lfa-policy-map route-nh-
template <template-name>
*A:PE-x# configure service vprn ospf3 area interface lfa-policy-map route-nh-
template <template-name>
```

LFA Policy Examples

All of the examples focus on providing another LFA NH for LDP FEC prefix 192.0.2.1/32 and 192.0.2.6/32 (the system IP addresses of PE-1 and PE-6), with PE-2 being the Point of Local Repair (PLR).

See Example Topology for the example topology.

The default LFA NH (without policy) for LDP FEC prefix192.0.2.1/32 is as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
```

```
===============================================================================
Prefix                                      Op             IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push             --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Push             --      262142BU
192.168.26.2                                1/1/3

192.0.2.1/32                                Swap           262142    262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap           262142    262142BU
192.168.26.2                                1/1/3

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-2#
```

The default LFA NH forLDP FEC prefix 192.0.2.6/32 is as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op             IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push             --      262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Push             --      262138BU
192.168.12.1                                1/1/2

192.0.2.6/32                                Swap           262138    262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Swap           262138    262138BU
192.168.12.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-2#
```

This default LFA NH can be changed by adding specific selection criteria inside a route-next-hop policy template.

Example 1: LFA Policy with Admin Group Constraint

The objective is to force the LFA NH for both LDP FEC prefixes to use the path between PE-2 and PE-5.

Define admin group ‘red’ with value ‘1’ and apply it to the IP interfaces PE-2 to PE-1 and PE-2 to PE-6.

```
*A:PE-2# configure router if-attribute admin-group "red" value 1
*A:PE-2# configure router interface "int-PE-2-PE-1" if-attribute admin-group "red"
*A:PE-2# configure router interface "int-PE-2-PE-6" if-attribute admin-group "red"
```

Define a route-next-hop policy template ‘LFA_NH_exclRed’, which excludes IP admin group ‘red’.

```
*A:PE-2# configure
    router
        route-next-hop-policy
            begin
            template "LFA_NH_exclRed"
                exclude-group "red"
            exit
            commit
```

The policy is applied ot the OSPF interfaces toward PE-1 and PE-6, as follows:

```
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-1" lfa-policy-
map route-nh-template "LFA_NH_exclRed"
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-6" lfa-policy-
map route-nh-template "LFA_NH_exclRed"
```

From the moment that the route-next-hop policy template ‘LFA_NH_exclRed’ is applied to the OSPF interfaces toward PE-1 and PE-6, the LFA NHs for both LDP FEC prefixes change. They now both point to the IP interface from PE-2 to PE-5 as LFA backup NH:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Push              --      262142BU
192.168.25.2                                1/1/4

192.0.2.1/32                                Swap            262142    262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap            262142    262142BU
192.168.25.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-2#

*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Push              --      262138BU
192.168.25.2                                1/1/4

192.0.2.6/32                                Swap            262138    262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Swap            262138    262138BU
192.168.25.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-2#
```

Example 2: LFA Policy with SRLG Constraint

The objective is to force the LFA NH for both LDP FEC prefixes to use the path from PE-2 to PE-5.

Define SRLG group ‘blue’ with value ‘2’ and apply it to the IP interfaces PE-2 to PE-5 and PE-2 to PE-6.

```
*A:PE-2# configure router if-attribute srlg-group "blue" value 2

*A:PE-2# configure router interface "int-PE-2-PE-1" if-attribute srlg-group "blue"
*A:PE-2# configure router interface "int-PE-2-PE-6" if-attribute srlg-group "blue"
```

Define a route-next-hop policy template ‘LFA_NH_SRLG’, where SRLG is enabled, as follows:

```
*A:PE-2# configure
    router
        route-next-hop-policy
            begin
            template "LFA_NH_SRLG"
                srlg-enable
            exit
            commit
```

The policy is applied to the OSPF interface toward PE-1 and PE-6, as follows:

```
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-1" lfa-policy-
map route-nh-template "LFA_NH_SRLG"
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-6" lfa-policy-
map route-nh-template "LFA_NH_SRLG"
```

Only one LFA policy mapping is allowed on an OSPF interface at a time. The new LFA policy mapping replaces the previous one.

The LFA NHs for both LDP FEC prefixes will both point now to the interface from PE-2 to PE-5 as LFA backup NH, as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.1/32

================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
================================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Push              --      262142BU
192.168.25.2                                1/1/4

192.0.2.1/32                                Swap            262142    262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap            262142    262142BU
192.168.25.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================


*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.6/32
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Push              --      262138BU
192.168.25.2                                1/1/4

192.0.2.6/32                                Swap            262138    262143
192.168.26.2                                1/1/3

192.0.2.6/32                                Swap            262138    262138BU
192.168.25.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-2#
```

The LFA policy mapping is removed from the OSPF interfaces as follows:

```
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-1" no lfa-policy-map
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-6" no lfa-policy-map
```

Example 3: LFA Policy with NH-type Constraint

The objective is to force the LFA NH for IP prefix 192.0.2.6/32 to use an RSVP tunnel.

Enable IP FRR as follows:

```
*A:PE-2# configure router ip-fast-reroute
```

Set up an RSVP LSP tunnel toward 192.0.2.6 with a strict MPLS path going over PE-2 to PE-4 to PE-5 to PE-6.

Note: Since an RSVP LSP is set up between PE-2 and PE-6, MPLS/RSVP protocols also need to be enabled on all the corresponding IP interfaces along the MPLS path.

```
*A:PE-2# configure
    router
        mpls
            interface "int-PE-2-PE-4"
            exit
            path "path-PE-2-PE-4-PE-5-PE-6"
                hop 10 192.168.24.2 strict
                hop 20 192.168.45.2 strict
                hop 30 192.168.56.2 strict
                no shutdown
            exit
            lsp "LSP-PE-2-PE-6-strict"
                to 192.0.2.6
                primary "path-PE-2-PE-4-PE-5-PE-6"
                exit
                no shutdown
            exit
            no shutdown
```

Enable RSVP shortcut within the IGP on PE-2 and indicate that the newly created RSVP LSP is a possible shortcut candidate for LFA backup NH only.

```
*A:PE-2# configure router ospf rsvp-shortcut
*A:PE-2# configure router mpls lsp "LSP-PE-2-PE-6-strict" igp-shortcut lfa-only
```

Displaying the tunnel-table on PE-2 shows that an LDP LSP and an RSVP LSP is available toward PE-6:

```
*A:PE-2# show router tunnel-table 192.0.2.6

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.6/32      rsvp      MPLS  1         7        192.168.24.2   16777215
192.0.2.6/32      ldp       MPLS  65541     9        192.168.26.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

The RSVP tunnel with tunnel ID 1 corresponds to the RSVP LSP “LSP-PE-2-PE-6-strict”, as can be seen as follows:

```
*A:PE-2# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-PE-2-PE-6-strict               192.0.2.6        1       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 1
===============================================================================
*A:PE-2#
```

By default, the preferred NH type is IP, not tunnel. Therefore, the RSVP tunnel will not be used for the LFA backup, as can be seen in the following output:

```
*A:PE-2# show router route-table alternative 192.0.2.6/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.6/32                                  Remote  OSPF      00h03m21s  10
       192.168.26.2                                                  10
       192.168.12.1 (LFA)                                            20
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Define a route-next-hop policy template **"LFA_NH_Tunnel"**, where nh-type is set to **tunnel**.

```
*A:PE-2# configure
    router
        route-next-hop-policy
            begin
            template "LFA_NH_Tunnel"
                nh-type tunnel
            exit
            commit
```

Apply the policy template to the interface toward PE-6, as follows:

```
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-6" lfa-policy-
map route-nh-template "LFA_NH_Tunnel"
```

The LFA NH uses the RSVP tunnel. The reference to the RSVP tunnel-ID (1) in the following show output corresponds with the tunnel-ID shown in the preceding **show router tunnel-table 192.0.2.6** output:

```
*A:PE-2# show router route-table alternative 192.0.2.6/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.6/32                                  Remote  OSPF      00h10m35s  10
       192.168.26.2                                                 10
       192.0.2.6 (LFA) (tunneled:RSVP:1)                            65535
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

```
*A:PE-2# show router fib 1 nh-table-usage

===============================================================================
FIB Next-Hop Summary
===============================================================================
IPv4/IPv6                     Active                   Available
-------------------------------------------------------------------------------
IP Next-Hop                   9                        16383
Tunnel Next-Hop               1                        993279
ECMP Next-Hop                 0                        512000
ECMP Tunnel Next-Hop          0                        261120
===============================================================================
*A:PE-2#
```

Example 4: Exclude Prefix from LFA Policy

The objective is to force no LFA NH for LDP FEC prefix 192.0.2.1/32 where PE-2 is the PLR.

The IP/LDP FRR implementation in SR OS allows to exclude an IGP interface, IGP area (OSPF), or IGP level (IS-IS) from the LFA SPF computation. The user also has the ability to exclude specific prefixes from the LFA SPF by using well-known prefix-lists and policy statements.

This is configured as follows:

```
*A:PE-2# configure
    router
        policy-options
            begin
            prefix-list "lo0-PE-1"
                prefix 192.0.2.1/32 exact
            exit
            policy-statement "LFA_Exclude_PE-1"
                entry 10
                    from
                        prefix-list "lo0-PE-1"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The configured policy statement is applied to the IGP protocol, as follows:

```
*A:PE-2# configure router ospf loopfree-alternate-exclude prefix-
policy "LFA_Exclude_PE-1"
```

From the moment that it is applied, the existing LFA NH entries for LDP FEC prefix 192.0.2.5/32 disappear instantly (compare with the preceding Example 1):

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap            262142    262143
192.168.12.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-2#
```

Conclusion

In production MPLS networks where IP FRR and/or LDP FRR are deployed, it is possible that the existing calculated LFA NHs are not always taking the most optimal or desirable paths.

With LFA policies, operators have better control on the way in which LFA backup NHs are computed.

Different selection criteria can be part of the route-next-hop policy: IP admin groups, IP SRLG groups, protection type preference and NH type preference.

PBR/PBF Redundancy

This chapter provides information about PBR/PBF Redundancy.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 14.0.R7. Secondary actions in IPv4, IPv6, and MAC Access Control List (ACL) filter policies are supported in SR OS Release 14.0.R1, and later.

Overview

PBR and PBF

Policy-Based Forwarding (PBF) and Policy-Based Routing (PBR) are used to make forwarding decisions based on filter policies defined by the network administrator. PBF is L2 traffic steering, whereas PBR is L3 traffic steering. For ordinary routing, the destination IP address is looked up in the routing table; for ordinary forwarding in a VPLS, the destination MAC address is looked up in the Forwarding Database (FDB). However, with PBR, routing decisions are based on IP filters that use more criteria, such as source and destination IP address, port number, DSCP value, and so on. Packets can take paths that differ from the next hop path specified by the routing table. PBF forwarding decisions can be made based on IP filters, but also on MAC filters that use criteria such as source and destination MAC address, inner and outer VLAN tag, dot1p priority, and so on.

The benefits of PBR/PBF are the following:

- The forwarding decision can be based on multiple attributes of a packet, not only its destination address

- Different QoS treatment can be provided, based on additional criteria
- Cost saving: time-sensitive traffic can be sent over higher-speed links at a higher cost, while bulk file transfers are sent over lower-speed links at a lower cost
- Load sharing: traffic can be load balanced across multiple and unequal paths

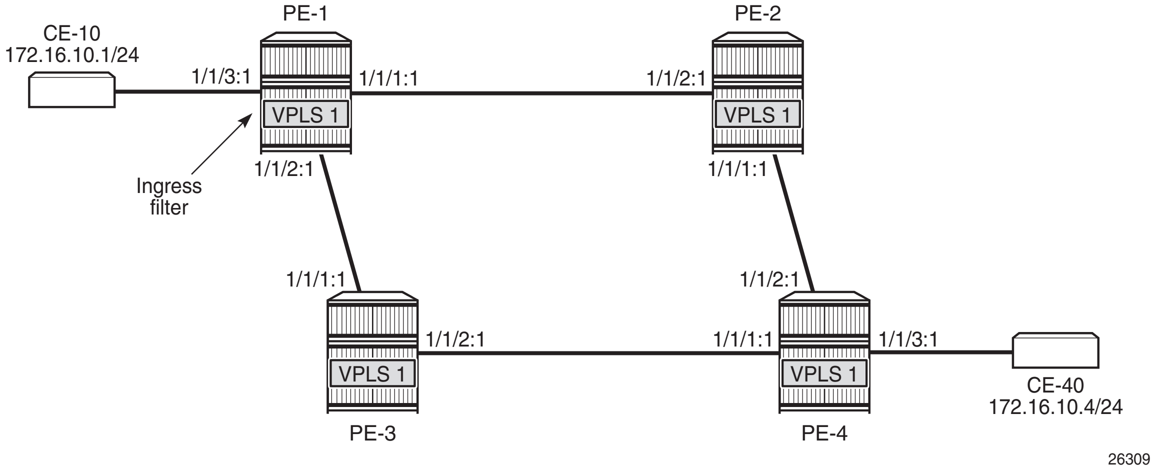
In most situations, PBR/PBF works on inbound unicast packets; therefore, a filter is applied at the ingress of access or network interfaces. In this chapter, examples will be shown for IPv4 filters and MAC filters applied on SAP ingress. IPv6 filters are also supported, but the examples in this chapter are based on IPv4. Filters are also supported on the egress, but that is beyond the scope of this chapter.

An IPv4 filter contains one or more entries, which can be configured with the following command:

```
*A:PE-1# configure filter ip-filter 10 create
*A:PE-1>config>filter>ip-filter# entry 10
  - entry <entry-id> [create]
  - no entry <entry-id>
 <entry-id>           : [1..65535]
 <create>             : keyword - mandatory while creating an entry.

 [no] action          + Configure action for the filter entry
 [no] description     - Description for this filter entry
 [no] egress-pbr      - Enable egress PBR
 [no] filter-sample   - Enable/Disable Cflowd sampling
 [no] interface-disa* - Disable/Enable Cflowd sampling on the interfaces
 [no] log             - Configure log for the filter entry
 [no] match           + Configure match criteria for this IP-filter entry
 [no] pbr-down-actio* - Configure action that overrides default PBR/PBF down action.
                        'no pbr-down-action-override' preserves default PBR/PBF down
                        action, which varies for different actions.
 [no] sticky-dest     - Set stickiness of PBR/PBF destinations and hold-time-up
                        for stickiness to take effect
```

Figure 70 shows the example topology with VPLS 1 configured in the PEs. PBF is applied in VPLS 1 on PE-1.

Figure 70 **PBF in VPLS 1 on PE-1**



The following configuration creates an IPv4 filter that forwards all packets matching the source and destination IPv4 addresses, 172.16.10.1/24 and 172.16.10.4/24 respectively, to SAP 1/1/1:1. When SAP 1/1/1:1 is operationally down, the default behavior is to drop the packet. Not every IPv4/v6 filter needs to have match criteria defined, but in this case, only packets with the configured IPv4 SA and IPv4 DA are affected, whereas the other packets are forwarded per the FDB in VPLS 1 on PE-1.

```
configure
    filter
        ip-filter 10 create
        default-action drop            ## default
            entry 10 create
                match
                    dst-ip 172.16.10.4/24
                    src-ip 172.16.10.1/24
                exit
                action
                    forward sap 1/1/1:1
                exit
            exit
       exit
```

In a similar way, an entry in a MAC filter can be configured with the following command:

```
*A:PE-1# configure filter mac-filter 20 entry 10
  - entry <entry-id> [create]
  - no entry <entry-id>

 <entry-id>           : [1..65535]
 <create>             : keyword - mandatory while creating an entry.

 [no] action          + Configure action for the filter entry
 [no] description     - Description for this filter entry
```

```
[no] log             - Configure log for the filter entry
[no] match           + Configure match criteria for this mac filter entry
[no] pbr-down-actio* - Configure action that overrides default PBR/PBF down action.
                       'no pbr-down-action-override' preserves default PBR/PBF down
                       action, which varies for different actions.
[no] sticky-dest     - Set stickiness of PBF destinations and hold-time-up
                       for stickiness to take effect
```

The following MAC filter forwards all frames with source MAC SA aa:aa:01:10:10:10 to SAP 1/1/1:1:

```
configure
    filter
        mac-filter 20 create
            entry 10 create
                match
                    src-mac aa:aa:01:10:10:10 ff:ff:ff:ff:ff:ff
                exit
                action
                    forward sap 1/1/1:1
                exit
            exit
        exit
```

Instead of defining a specific MAC address, a range of MAC addresses can be defined using a mask. The default mask is all 1s, ff:ff:ff:ff:ff:ff, which corresponds to an exact match of the configured MAC address.

When the primary SAP 1/1/1:1 is down, the default action is drop. However, PBR/PBF redundancy can be configured, as described in the following section.

PBR/PBF Redundancy

PBR/PBF redundancy is supported for MAC filters, IPv4 filters, and IPv6 filters. Within each entry in the IP/MAC filter, a secondary action can be configured; for example, for entry 10 in IPv4 filter 10, as follows:

```
configure
    filter
        ip-filter 10 create
            entry 10 create
                match
                    dst-ip 172.16.10.4/24
                    src-ip 172.16.10.1/24
                exit
                action
                    forward sap 1/1/1:1
                exit
                action secondary
                    forward sap 1/1/2:1
                exit
            exit
```

This IPv4 filter only affects packets with IPv4 SA 172.16.10.1/24 and IPv4 DA 172.16.10.4/24. When the primary action SAP 1/1/1:1 is operationally up, the primary action is executed; when SAP 1/1/1:1 is operationally down, the secondary action is executed, until SAP 1/1/1:1 is operationally up again. When both SAPs are down, the default behavior is to drop the packet.

In the preceding example, both actions forward packets to a SAP, but the PBR/PBF target can also be an SDP binding or-for PBR-a next-hop IP address in a VPRN. Table 6 shows the allowed primary and secondary forwarding action combinations within a filter entry.

Table 6 **Primary and Secondary Forwarding Actions**

Primary Forwarding Action	Secondary Forwarding Action
sap <sap-id>	sap <sap-id>
sap <sap-id>	sdp <sdp-id:vc-id>
sdp <sdp-id:vc-id>	sdp <sdp-id:vc-id>
sdp <sdp-id:vc-id>	sap <sap-id>
next-hop <ipv4/ipv6-address> router <router-instance>	next-hop <ipv4-ipv6-address> router <router-instance>
next-hop indirect <ipv4/ipv6-address> router <router-instance>	next-hop indirect <ipv4/ipv6-address> router <router-instance>

The IPv4 filter is applied on the ingress of SAP 1/1/3:1 in VPLS 1 on PE-1, as follows:

```
configure
    service
        vpls 1 customer 1 create
            sap 1/1/1:1 create
            exit
            sap 1/1/2:1 create
            exit
            sap 1/1/3:1 create
                ingress
                    filter ip 10
                exit
            exit
            spoke-sdp 12:1 create
            exit
            spoke-sdp 13:1 create
            exit
            no shutdown
        exit
```

When the primary action SAP 1/1/1:1 is operationally up (PBR Target Status: Up), the primary action is executed (Downloaded Action: Primary), as follows:

```
*A:PE-1# show filter ip 10

===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 1
Description         : (Not Specified)
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 10
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.10.1/24
Src. Port           : n/a
Dest. IP            : 172.16.10.4/24
Dest. Port          : n/a
Protocol            : Undefined                      Dscp           : Undefined
ICMP Type           : Undefined                      ICMP Code      : Undefined
Fragment            : Off                            Src Route Opt  : Off
Sampling            : Off                            Int. Sampling  : On
IP-Option           : 0/0                            Multiple Option: Off
TCP-syn             : Off                            TCP-ack        : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Primary
Dest. Stickiness    : None                           Hold Remain    : 0
Ing. Matches        : 500 pkts (53000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

When the primary action SAP 1/1/1:1 is operationally down, the secondary action is executed. When SAP 1/1/1:1 is down, packets are forwarded to secondary action SAP 1/1/2:1 instead. However, when the primary action SAP 1/1/1:1 is operationally up again, the primary action is executed. This revertive behavior can be disabled by configuring stickiness in the filter entry, as follows:

```
*A:PE-1# configure filter ip-filter 10 entry 10 sticky-dest
  - no sticky-dest
  - sticky-dest <hold-time-up>
  - sticky-dest no-hold-time-up

 <hold-time-up>       : 0..65535 seconds
```

When both the primary action SAP 1/1/1:1 and the secondary action SAP 1/1/2:1 are down, the default action is drop, unless the **pbr-down-action-override** <*filter-action*> parameter is configured. When the configured filter action is **forward**, the packets can be forwarded to another object in the service that is up, for example, to another SAP or to an SDP binding, per the packet's destination address. This means that in a VPLS (PBF), the MAC DA is looked up in the FDB; in a VPRN (PBR), the IP DA is looked up in the routing table. The configuration of the **pbr-down-action-override** parameter is as follows. No specific SAPs or SDP bindings need to be defined.

```
*A:PE-1# configure filter ip-filter 10 entry 10 pbr-down-action-override
  - no pbr-down-action-override
  - pbr-down-action-override <filter-action>

 <filter-action>      : drop|forward|filter-default-action


*A:PE-1# configure filter ip-filter 10 entry 10 pbr-down-action-override forward
```

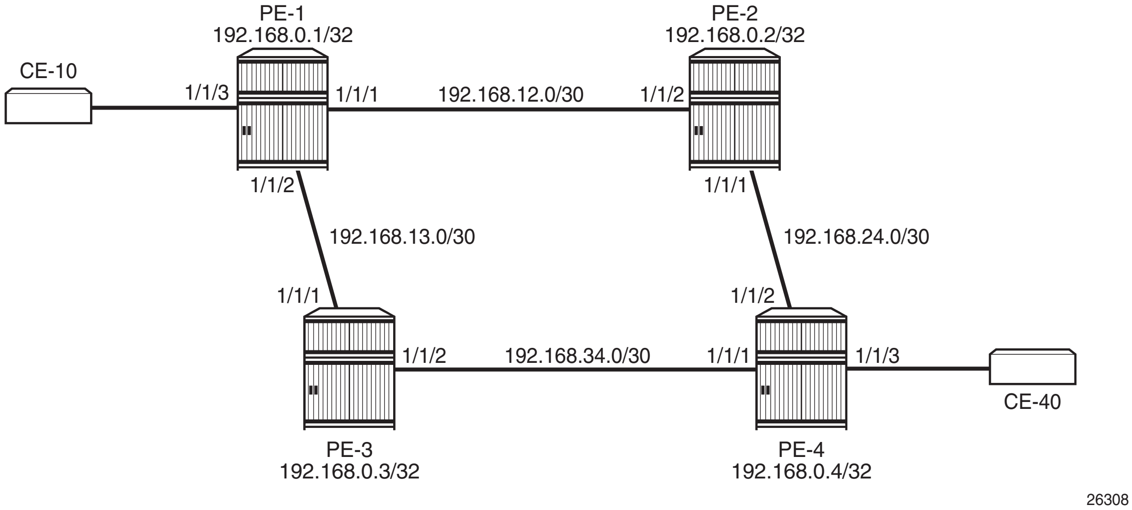
Configuration

In this section, the following examples are described:

- PBF in a VPLS using an IPv4 filter
- PBF in a VPLS using a MAC filter
- PBR in a VPRN using an IPv4 filter

Figure 71 shows the example topology with four PEs and two CEs.

Figure 71 **Example Topology**



The initial configuration is as follows:

- Cards, MDAs, ports (all ports are in hybrid mode with dot1q encapsulation)
- Router interfaces
- IS-IS as IGP between the PEs (alternatively, OSPF could be configured as IGP)
- LDP between the PEs
- CE-10 is emulated using VPRN 10 on PE-1 with a hairpin to loop the traffic back to the PE; CE-40 is emulated using VPRN 10 on PE-4.

PBF in a VPLS Using an IP Filter

The configuration will be shown for PE-1. The following cases will be described:

1. Initial situation: primary action is executed.
2. Primary action SAP 1/1/1:1 is put in a shutdown state. The secondary action in the entry in the IPv4 filter is executed.
3. Both primary and secondary action SAPs 1/1/1:1 and 1/1/2:1 are put in a shutdown state. The default action is drop.
4. Both primary and secondary action SAPs 1/1/1:1 and 1/1/2:1 are put in a shutdown state. The **pbr-down-action-override** parameter is configured with action forward.
5. The secondary action SAP 1/1/2:1 is put in a no shutdown state. The secondary action is executed.

6. The primary action SAP 1/1/1:1 is put in a no shutdown state. The primary action is executed.
7. Stickiness is configured with a hold timer of 60 seconds. At timer expiry, stickiness takes effect. If SAP 1/1/1:1 is up at timer expiry, the primary action is programmed; otherwise, if SAP 1/1/2:1 is up, the secondary action is programmed.
8. Stickiness is configured without a hold timer and takes effect immediately.

Configure VPLS 1 with IPv4 Filter on SAP Ingress

IPv4 filter 10 has one entry with primary action to forward to SAP 1/1/1:1 and secondary action to forward to SAP 1/1/2:1. No match criteria are defined. When all action forward SAPs are operationally down, the default action is drop. No stickiness is configured.

```
configure
    filter
        ip-filter 10 create
            entry 10 create
                action
                    forward sap 1/1/1:1
                exit
                action secondary
                    forward sap 1/1/2:1
                exit
            exit
```

VPLS 1 on PE-1 is configured with three SAPs and two spoke-SDPs, as follows. IPv4 filter 10 is configured on the ingress of SAP 1/1/3:1 and applies to traffic originating from CE-10.

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        sdp 13 mpls create
            far-end 192.0.2.3
            ldp
            no shutdown
        exit
        vpls 1 customer 1 create
            sap 1/1/1:1 create
            exit
            sap 1/1/2:1 create
            exit
            sap 1/1/3:1 create
                ingress
```

```
                    filter ip 10
                exit
            exit
            spoke-sdp 12:1 create
            exit
            spoke-sdp 13:1 create
            exit
            no shutdown
        exit
```

When all SAPs are up, all packets from CE-10 enter SAP 1/1/3:1 and are forwarded to primary action SAP 1/1/1:1. The response message will follow the reverse path. This can be shown as follows. No other traffic is sent and the number of packets received or sent on port 1/1/1 will only slightly exceed the number of packets sent on the SAP, because of signaling between the PEs for IS-IS and LDP. The port statistics are cleared for ports 1/1/1 through 1/1/3 on PE-1. CE-10 sends a series of 1000 ICMP echo requests and, afterward, the port statistics on PE-1 are verified.

```
*A:PE-1# clear port 1/1/[1..3] statistics
```

```
*A:PE-1# ping router 10 172.16.10.4 rapid count 1000
---snip---
---- 172.16.10.4 PING Statistics ----
1000 packets transmitted, 1000 packets received, 0.00% packet loss
round-trip min = 1.55ms, avg = 1.71ms, max = 2.60ms, stddev = 0.176ms
```

```
*A:PE-1# show port 1/1/[1..3] statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1011          107190            1010          107024
===============================================================================
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        10            1061              10            1061
===============================================================================
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                      1000          106000            1000          106000
===============================================================================
*A:PE-1#
```

All traffic is forwarded from ingress SAP 1/1/3:1 to SAP 1/1/1:1 and the reply messages from SAP 1/1/1:1 to SAP 1/1/3:1. No packets are forwarded via SAP 1/1/2:1.

When the primary action SAP 1/1/1:1 is operationally up, the primary action is executed, as follows:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Primary
Dest. Stickiness    : None                           Hold Remain    : 0
Ing. Matches        : 1000 pkts (106000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-1#
```

Primary Action PBR Target Down

The primary action SAP 1/1/1:1 is put in a shutdown state. Therefore, the primary action cannot be executed, and the secondary action is executed instead. When CE-10 sends 1000 ICMP echo requests, all packets are forwarded to SAP 1/1/2:1.

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 shutdown
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
```

```
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Secondary
Dest. Stickiness    : None                         Hold Remain    : 0
Ing. Matches        : 2000 pkts (212000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

Secondary Action PBR Target Down

The secondary action SAP 1/1/2:1 is put in a shutdown state, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/2:1 shutdown
```

Both SAP 1/1/1:1 and SAP 1/1/2:1 are in a shutdown state. Neither the primary nor the secondary action in entry 10 of IPv4 filter 10 can be executed. Therefore, the default action is executed, which is drop; see the following output (PBR Down Action: Drop (entry-default)). When CE-10 sends ICMP echo requests, they are all dropped.

```
*A:PE-1# ping router 10 172.16.10.4 rapid count 1000
PING 172.16.10.4 56 data bytes
---snip---
---- 172.16.10.4 PING Statistics ----
1000 packets transmitted, 0 packets received, 100% packet loss
```

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                           Applied        : Yes
Scope               : Template                     Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Down
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Primary
Dest. Stickiness    : None                         Hold Remain    : 0
Ing. Matches        : 3000 pkts (318000 bytes)
```

```
Egr. Matches        : 0 pkts

===============================================================================
```

PBR Down Action Override

Both SAPs remain in a shutdown state. The default PBR down action is drop, but that can be overruled by configuring the **pbr-down-action-override** parameter, as follows:

```
*A:PE-1# configure filter ip-filter 10 entry 10 pbr-down-action-override forward
```

With this configuration added in entry 10 of IPv4 filter 10, the PBR down action will be forward. No specific next-hop needs to be defined. The forwarding is based on the destination address. When CE-10 sends 1000 ICMP echo requests, the traffic is forwarded, as follows:

```
*A:PE-1# ping router 10 172.16.10.4 rapid count 1000
---snip---
---- 172.16.10.4 PING Statistics ----
1000 packets transmitted, 1000 packets received, 1 duplicate
round-trip min = 1.58ms, avg = 1.72ms, max = 2.70ms, stddev = 0.174ms
```

The statistics in the detailed output for spoke-SDP 12:1 in VPLS 1 shows that these packets have been sent over this spoke-SDP:

```
*A:PE-1# show service id 1 sdp 12:1 detail | match Statistics post-lines 5
Statistics           :
I. Fwd. Pkts.      : 1001                     I. Dro. Pkts.     : 0
I. Fwd. Octs.      : 98098                    I. Dro. Octs.     : 0
E. Fwd. Pkts.      : 1006                     E. Fwd. Octets    : 98360
-------------------------------------------------------------------------------
```

The PBR down action for entry 10 in IPv4 filter 10 is forward, as defined by the **pbr-down-action-override** parameter, and the PBR downloaded action is forward, as follows:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                          Applied        : Yes
Scope               : Template                    Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
```

```
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Down
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Forward
Dest. Stickiness    : None                         Hold Remain    : 0
Ing. Matches        : 3000 pkts (318000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-1#
```

Secondary Action Up - Revertive Behavior

The primary action SAP 1/1/1:1 remains in a shutdown state, whereas secondary action SAP 1/1/2:1 is re-enabled, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/2:1 no shutdown
```

The secondary action in entry 10 of IPv4 filter 10 is executed (Downloaded Action: Secondary), as follows:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                           Applied        : Yes
Scope               : Template                     Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Secondary
Dest. Stickiness    : None                         Hold Remain    : 0
Ing. Matches        : 5000 pkts (530000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-1#
```

Primary Action Up - Revertive Behavior

As well as the secondary action SAP, also the primary action SAP 1/1/1:1 is re-enabled, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 no shutdown
```

The default PBR/PBF behavior is revertive; therefore, the primary action is executed: the packets are forwarded to SAP 1/1/1:1, as follows:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                            Applied        : Yes
Scope               : Template                      Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Primary
Dest. Stickiness    : None                          Hold Remain    : 0
Ing. Matches        : 6000 pkts (636000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

Stickiness in IP Filter with Hold Timer

When the primary action SAP becomes up, traffic will be forwarded to this SAP instantaneously, unless stickiness applies. Stickiness can be defined on the IPv4/v6 filter entry level to override this revertive behavior. The following command enables stickiness at timer expiry with a hold remain timer of-in this case-60 seconds for entry 10 in IPv4 filter 10:

```
*A:PE-1# configure filter ip-filter 10 entry 10 sticky-dest 60
```

The hold remain timer starts counting down when stickiness is configured and at least one PBR target is up. If the primary action SAP 1/1/1:1 remains operationally up for the configured 60 seconds, the primary action will be active, and at timer expiry, stickiness applies. However, if SAP 1/1/1:1 goes down and then up again before timer expiry, the secondary action remains active until the hold remain timer expires, as shown in the following example.

The hold remain timer has not expired. The primary action SAP 1/1/1:1 is put in a shutdown state, so the secondary action is active, as follows. The hold remain timer keeps counting down.

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 shutdown

*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Secondary
Dest. Stickiness    : 60                             Hold Remain    : 49
Ing. Matches        : 7000 pkts (742000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

The primary action SAP 1/1/1:1 is restored and the secondary action is active until the hold remain timer expires, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 no shutdown

*A:PE-1# show filter ip 10

===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
```

```
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Secondary
Dest. Stickiness    : 60                           Hold Remain    : 29
Ing. Matches        : 8000 pkts (848000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

In the preceding output, the hold remain time is 29 seconds. When the hold remain timer expires and the primary action SAP 1/1/1:1 is up, the primary action is activated again and stickiness applies from then onward, as follows:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                           Applied        : Yes
Scope               : Template                     Def. Action    : Drop
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Primary
Dest. Stickiness    : 60                           Hold Remain    : 0
Ing. Matches        : 8000 pkts (848000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

The hold remain timer stays at zero. When the primary action cannot be activated, the secondary action is activated and will remain activated even when the primary action SAP 1/1/1:1 is up again. However, when the secondary action SAP 1/1/2:1 is down, the primary action can be activated again.

The hold remain timer starts counting down when it is first configured, or reconfigured with a different value, and at least one of the PBR/PBF targets is up. The hold remain timer also starts counting down after both the primary and the secondary PBR/PBF targets have been down, for example, after a reboot, and at least one of them transitions to the up status. The secondary action might be available first, even though the primary action is preferred. This situation is automatically resolved when the timer expires: the primary action will be activated if available when the hold remain timer expires.

Force Primary Action

Stickiness can be enabled without any delay, as follows:

```
*A:PE-1# configure filter ip-filter 10 entry 10 sticky-dest no-hold-time-up
```

```
*A:PE-1# configure filter
*A:PE-1>config>filter# info
----------------------------------------------
        ip-filter 10 create
            entry 10 create
                action
                    forward sap 1/1/1:1
                exit
                action secondary
                    forward sap 1/1/2:1
                exit
                pbr-down-action-override forward
                sticky-dest 0
            exit
        exit
----------------------------------------------
```

Initially, the primary action is executed, but when the primary action SAP 1/1/1:1 is put in a shutdown state, the secondary action is executed, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 shutdown
```

```
*A:PE-1# show filter ip 10
================================================================================
IP Filter
================================================================================
Filter Id           : 10                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Down
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Secondary
Dest. Stickiness    : 0                              Hold Remain    : 0
Ing. Matches        : 9000 pkts (954000 bytes)
Egr. Matches        : 0 pkts

================================================================================
*A:PE-1#
```

The secondary action is active and will remain active as long as the secondary action SAP 1/1/2:1 is up. The hold remain timer is not enabled (== value 0). When the primary action SAP 1/1/1:1 is operationally up again, the secondary action remains active, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/1:1 no shutdown
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                          Applied        : Yes
Scope               : Template                    Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Secondary
Dest. Stickiness    : 0                           Hold Remain    : 0
Ing. Matches        : 10000 pkts (1060000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

The following **tools** command forces activation of the primary action in entry 10 of IPv4-filter 10:

```
*A:PE-1# tools perform filter ip-filter 10 entry 10 activate-primary-action
```

The result is that the primary action is executed again, as shown in the following output:

```
*A:PE-1# show filter ip 10
===============================================================================
IP Filter
===============================================================================
Filter Id           : 10                          Applied        : Yes
Scope               : Template                    Def. Action    : Drop
---snip---
Entry               : 10
---snip---
Primary Action      : Forward (SAP)
  Next Hop          : 1/1/1:1
  Service Id        : 1
  PBR Target Status : Up
Secondary Action    : Forward (SAP)
  Next Hop          : 1/1/2:1
  Service Id        : 1
  PBR Target Status : Up
PBR Down Action     : Forward (pbr-down-action-override)
Downloaded Action   : Primary
Dest. Stickiness    : 0                            Hold Remain    : 0
Ing. Matches        : 11000 pkts (1166000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
```

This **tools** command can also be used in combination with a running sticky-destination hold remain timer. In that case, the hold remain timer will stop counting down and the primary action immediately reverts.

PBF in a VPLS Using a MAC Filter

PBF in a VPLS can use a MAC filter instead of an IPv4 filter, but not both. The following MAC filter is defined on PE-1:

```
configure
    filter
        mac-filter 20 create
            entry 10 create
                match
                    src-mac aa:aa:01:10:10:10 ff:ff:ff:ff:ff:ff
                exit
                action
                    forward sap 1/1/1:1
                exit
                action secondary
                    forward sap 1/1/2:1
                exit
                pbr-down-action-override forward
                sticky-dest 0
            exit
        exit
```

MAC filter 20 cannot be applied next to IPv4 filter 10 on the ingress direction of SAP 1/1/3:1 in VPLS 1; therefore, an error message is raised, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:1 ingress filter mac 20
MINOR: SVCMGR #1631 There is another filter already defined for the SAP
```

The filter that was applied must be removed first, then the MAC filter can be applied, as follows:

```
*A:PE-1# configure service vpls 1 sap 1/1/3:1 ingress no filter
*A:PE-1# configure service vpls 1 sap 1/1/3:1 ingress filter mac 20
```

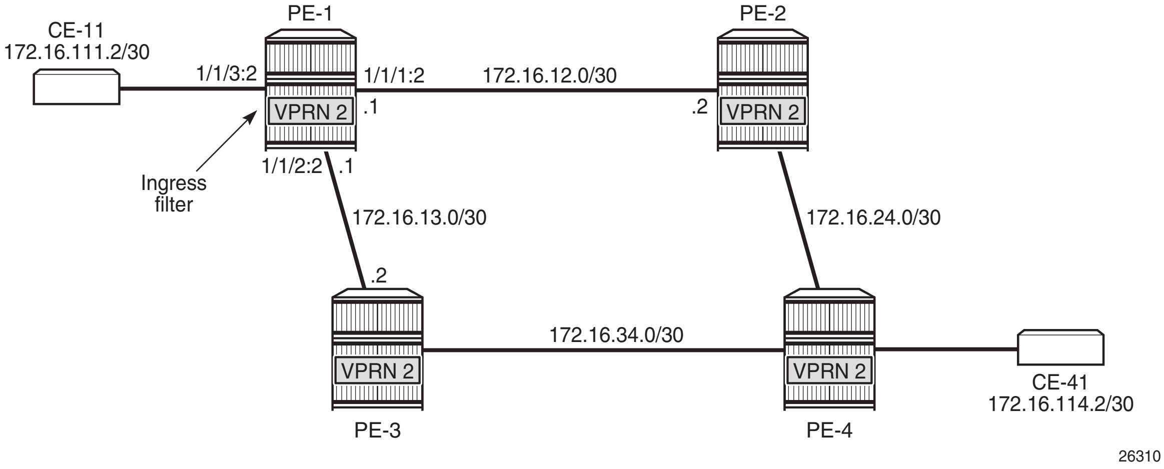
When all SAPs in the VPLS are up, the primary action is activated, as follows:

```
*A:PE-1# show filter mac 20
===============================================================================
Mac Filter
===============================================================================
Filter Id            : 20                            Applied        : Yes
Scope                : Template                      Def. Action    : Drop
Entries              : 1                             Type           : normal
Description          : (Not Specified)
-------------------------------------------------------------------------------
Filter Match Criteria : Mac
-------------------------------------------------------------------------------
Entry                : 10                            FrameType      : Ethernet
Description          : (Not Specified)
Log Id               : n/a
Src Mac              : aa:aa:01:10:10:10 ff:ff:ff:ff:ff:ff
Dest Mac             : Undefined
Dot1p                : Undefined                     Ethertype      : Undefined
DSAP                 : Undefined                     SSAP           : Undefined
Snap-pid             : Undefined                     ESnap-oui-zero : Undefined
Primary Action       : Forward (SAP)
  Next Hop           : 1/1/1:1
  Service Id         : 1
  PBR Target Status  : Up
Secondary Action     : Forward (SAP)
  Next Hop           : 1/1/2:1
  Service Id         : 1
  PBR Target Status  : Up
PBR Down Action      : Forward (pbr-down-action-override)
Downloaded Action    : Primary
Dest. Stickiness     : 0                             Hold Remain    : 0
Ing. Matches         : 1000 pkts (106000 bytes)
Egr. Matches         : 0 pkts

===============================================================================
```

PBR in a VPRN Using an IP Filter

Figure 72 shows the example topology used with VPRN 2 configured on each PE and the CEs configured as VPRN 11 on PE-1 and PE-4.

Figure 72 **PBR in a VPRN**



The configuration of VPRN 2 on PE-1 is as follows:

```
configure
    service
        vprn 2 customer 1 create
            route-distinguisher 64496:2
            interface "int-PE-1-CE-11_VPRN2" create
                address 172.16.111.1/30
                sap 1/1/3:2 create
                exit
            exit
            interface "int-PE-1-PE-2_VPRN2" create
                address 172.16.12.1/30
                sap 1/1/1:2 create
                exit
            exit
            interface "int-PE-1-PE-3_VPRN2" create
                address 172.16.13.1/30
                sap 1/1/2:2 create
                exit
            exit
            no shutdown
        exit
```

The configuration of VPRN 2 on the remaining PEs is similar, except that static route entries are configured for subnets 172.16.111.0/30 (toward CE-11) and 172.16.114.0/30 (toward CE-41). VPRN 2 in PE-1 has the following IPv4 filter applied to SAP 1/1/3:2 toward CE-11:

```
configure
    filter
        ip-filter 30 create
            entry 10 create
                action
                    forward next-hop 172.16.12.2 router 2
                exit
                action secondary
                    forward next-hop 172.16.13.2 router 2
                exit
            exit
configure service vprn 2 interface "int-PE-1-CE-11_VPRN2" sap 1/1/3:2
                    ingress filter ip 30
```

The primary action forwards packets from CE-11 to next-hop 172.16.12.2, which is an interface in VPRN 2 on PE-2; the secondary action forwards to next-hop 172.16.13.2, an interface in VPRN 2 on PE-3. When all interfaces are up, the primary action is executed and traffic from CE-11 to CE-41 is forwarded from VPRN 2 on PE-1 to VPRN 2 on PE-2 (next-hop 172.16.12.2), as follows:

```
*A:PE-1# show filter ip 30
===============================================================================
IP Filter
===============================================================================
Filter Id           : 30                             Applied        : Yes
Scope               : Template                       Def. Action    : Drop
---snip---
Primary Action      : Forward (Next Hop VRF)
  Next Hop          : 172.16.12.2
  Router            : 2
  PBR Target Status : Up
  Extended Action   : None
Secondary Action    : Forward (Next Hop VRF)
  Next Hop          : 172.16.13.2
  Router            : 2
  PBR Target Status : Up
  Extended Action   : None
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Primary
Dest. Stickiness    : None                           Hold Remain    : 0
Ing. Matches        : 1000 pkts (106000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-1#
```

The output includes an additional line per action: both the primary and the secondary action in PBR can have DSCP remarking as extended action, but that is not configured in this example. It can be configured using the following command; for example, for the primary action, as follows:

```
*A:PE-1# configure filter ip-filter 30 entry 10 action extended-action
*A:PE-1# configure filter ip-filter 30 entry 10 action secondary extended-action
  - extended-action
  - no extended-action

      remark          - Activate dscp remarking for packets matching the entry
```

When the primary action cannot be activated, the secondary action is activated, as follows:

```
*A:PE-1# configure service vprn 2 interface "int-PE-1-PE-2_VPRN2" sap 1/1/1:2
shutdown
```

```
*A:PE-1# show filter ip 30
===============================================================================
IP Filter
===============================================================================
Filter Id           : 30                           Applied        : Yes
Scope               : Template                     Def. Action    : Drop
---snip---
Primary Action      : Forward (Next Hop VRF)
  Next Hop          : 172.16.12.2
  Router            : 2
  PBR Target Status : Down
  Extended Action   : None
Secondary Action    : Forward (Next Hop VRF)
  Next Hop          : 172.16.13.2
  Router            : 2
  PBR Target Status : Up
  Extended Action   : None
PBR Down Action     : Drop (entry-default)
Downloaded Action   : Secondary
Dest. Stickiness    : None                         Hold Remain    : 0
Ing. Matches        : 1000 pkts (106000 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-1#
```

When both PBR targets are down, the default action is drop, because the IPv4 filter does not have the **pbr-down-action-override** parameter configured. Stickiness is not enabled in this filter. The configuration of the IPv4/v6 filters is similar for PBR and PBF.

In the preceding PBR example, the primary and secondary next-hop router is the same VRF (VPRN 2), but it can be any mix of VRFs, such as primary next-hop router 100 and secondary next-hop router 200.

PBR can also steer traffic to the base routing instance; for example, with the following IP filter:

```
configure
    filter
        ip-filter 40 create
            entry 10 create
                action
                    forward next-hop 192.0.2.2 router "Base"
                exit
                action secondary
                    forward next-hop 192.0.2.3 router "Base"
                exit
            exit
        exit
```

Conclusion

Operators can define two targets for L2 and L3 traffic steering (PBF and PBR): primary and secondary. The primary target is used when both targets are up; the secondary target is used when the primary is down. However, when stickiness is enabled, it is possible that the secondary action is executed, even when the primary action PBR target reverts to up. When both targets are down, the default action is drop, unless the **pbr-down-action-override** parameter is configured. Both 1+1 redundancy and N+1 redundancy are supported.

Rate Limit Filter Action

This chapter provides information about Rate Limit Filter Action.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

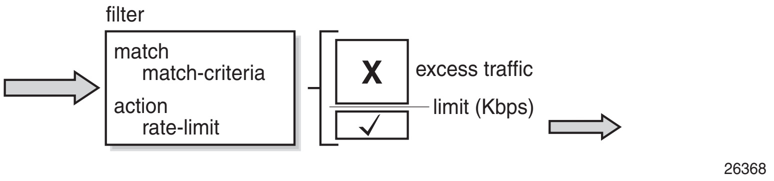
Applicability

This chapter is applicable to SR OS routers and is based on SR OS Release 14.0.R7.

Overview

Filter-based rate limiting can be used by operators for security reasons to protect their network resources or mitigate DDoS attacks; see Figure 73.

Figure 73 **Filter Based Rate Limiting**



SR OS supports filter-based rate limiting on ingress (Release 14.0.R1) and on egress (Release 14.0.R4) for IPv4, IPv6, and MAC filter policies

The rate-limit value is configurable in kilobits per second and applicable to traffic matching the filter condition. Packets matching the filter condition are dropped when the traffic rate is above the configured policer rate value and forwarded when the traffic rate is below the configured policer rate value.

QoS Interaction

On ingress, if the MAC or IPv4/IPv6 filter action indicates that traffic must be rate limited, this traffic is redirected to a rate-limiting filter policer before delivery to the switching fabric. Traffic not matching the MAC or IP filter will pass through the regular packet processing chain, and can be limited through SAP-ingress policies. Control traffic that is extracted to the CPM is not rate limited. Rate-limiting filter policies can coexist with the cflowd, log, and mirror features.

On egress, control and data traffic matching an egress rate-limiting filter policy bypasses egress QoS policing, but the usual egress QoS queueing still applies.

Rate-Limiting with Single or Multiple FlexPaths

Filter-based rate limiting can be applied to Layer 2 and Layer 3 services, and is supported on following items, including but not limited to:

- SAPs
- Network interface
- Spoke-SDPs
- group interfaces
- ESM subscribers

Filter-based rate limiting can also be used when the underlying infrastructure uses link aggregation.

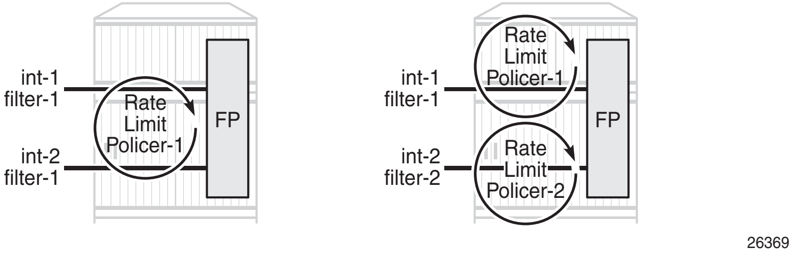
If multiple interfaces use the same rate-limiting filter policy on the same FP, the system will allocate a single rate-limiter resource to the FP; a common aggregate rate limit is applied to those interfaces.

If multiple interfaces use the same rate-limiting filter policy on different FPs, the system will allocate a rate-limiter resource for each FP; an independent rate limit applies to each FP.

The example to the left in Figure 74 has two interfaces with the same filter applied, and terminated on the same FP. Therefore, there is only one policer, and the aggregate traffic is topped at the rate defined in the filter. The example to the right has two interfaces with different filters, again terminated on the same FP. Because the interfaces have distinct filters, two different rate-limiting policers are created, which could (but not necessarily) define the same rate.

The actual packet length is used for the rate limit, not factoring in the encapsulation.

Figure 74 **Rate Limit Filters and FlexPaths**



Use caution when applying filter-based rate limiting to SAPs on group interfaces, because group interfaces can host many ESM subscribers, which could defeat per-subscriber and per-ESM host rate limiting.

Syntax

The following syntax defines an IPv4/IPv6 filter or a MAC filter with rate-limiting action:

```
A:7750-A>config>filter# info
    ip-filter | ipv6-filter | mac-filter <filter-id> create
        entry <entry-id> create
            match
                ** match criteria, e.g.: IP/Port **
            action
                rate-limit {<value-Kps> | max}
        exit
```

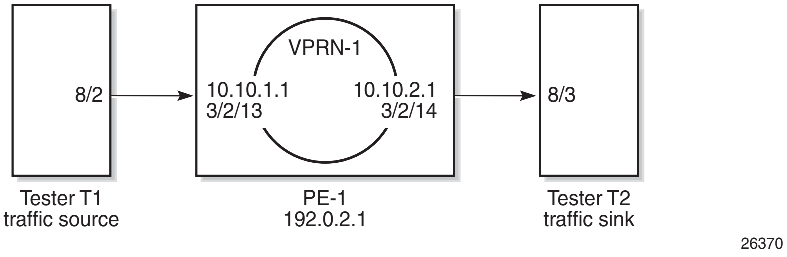
All regular IP and MAC match criteria are supported with the **action rate-limit**.

Configuration

Figure 75 shows the example configuration. Traffic is sourced on Tester T1, port 8/2, passes through VPRN-1, and is received on port 8/3 of Tester T2.

Ingress IPv4 filtering applies at the ingress SAP in VPRN-1. Ingress IPv6 filtering and ingress MAC filtering are similar to ingress IPv4 filtering and are not shown in this chapter.

Figure 75 **Example Configuration**



The configuration of VPRN-1 on PE-1 is as follows:

```
# R1
configure
    service
        vprn 1 customer 1 create
            description "rate limit action for ip filter"
            route-distinguisher 65536:1
            interface "int-TST-1" create
                address 10.10.1.1/24
                sap 3/2/13 create
                ingress
                    filter ip 1
                exit
                no shutdown
                exit
            exit
            interface "int-TST-2" create
                address 10.10.2.1/24
                sap 3/2/14 create
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The filter configuration is as follows:

```
configure
    filter
        ip-filter 1 create
            filter-name "ip-filter-2M"
            default-action forward
            description "IP filter test for rate limit action"
            entry 10 create
                match
                    dst-ip 10.10.2.2/32
                    src-ip 10.10.1.2/32
                exit
                action
                    rate-limit 2048
                exit
            exit
        exit
    exit
exit
```

A stream of UDP packets with a fixed size of 128 bytes is sent out of Tester T1 at a rate of 1000 packets/sec, accounting for a data rate of 128 x 8 x 1000 = 1.024 Mbit/s. At this rate, all packets pass through because the actual rate is lower than the rate-limit, as follows:

```
*A:PE1# monitor filter ip 1 entry 10 rate repeat 1
===============================================================================
Monitor statistics for IP filter 1 entry 10
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
Ing. Matches        : 14170 pkts (1813760 bytes)
Egr. Matches        : 0 pkts
Ing. Rate-limiter
  Offered           : 14160 pkts (1812480 bytes)
  Forwarded         : 14160 pkts (1812480 bytes)
  Dropped           : 0 pkts
Egr. Rate-limiter
  Offered           : 0 pkts
  Forwarded         : 0 pkts
  Dropped           : 0 pkts
-------------------------------------------------------------------------------
At time t = 10 sec (Mode: Rate)
-------------------------------------------------------------------------------
Ing. Matches        : 1001 pkts (128090 bytes)
Egr. Matches        : 0 pkts
Ing. Rate-limiter
  Offered           : 1002 pkts (128218 bytes)
  Forwarded         : 1002 pkts (128218 bytes)
  Dropped           : 0 pkts
Egr. Rate-limiter
  Offered           : 0 pkts
  Forwarded         : 0 pkts
  Dropped           : 0 pkts
===============================================================================
*A:PE1#
```

Increasing the actual rate to 3000 packets/s without changing the frame size corresponds to a data rate of 128 x 8 x 3000 = 3.072 Mbit/s, so part of the traffic is dropped as 3.072 Mbit/s > 2.048 Mbit/s, as follows:

```
*A:PE1# monitor filter ip 1 entry 10 rate repeat 1
===============================================================================
Monitor statistics for IP filter 1 entry 10
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
Ing. Matches        : 3222085 pkts (412426880 bytes)
Egr. Matches        : 0 pkts
```

```
Ing. Rate-limiter
  Offered           : 3222046 pkts (412421888 bytes)
  Forwarded         : 2147991 pkts (274942848 bytes)
  Dropped           : 1074055 pkts (137479040 bytes)
Egr. Rate-limiter
  Offered           : 0 pkts
  Forwarded         : 0 pkts
  Dropped           : 0 pkts

-------------------------------------------------------------------------------
At time t = 10 sec (Mode: Rate)
-------------------------------------------------------------------------------
Ing. Matches        : 3000 pkts (383974 bytes)
Egr. Matches        : 0 pkts
Ing. Rate-limiter
  Offered           : 3004 pkts (384473 bytes)
  Forwarded         : 2002 pkts (256307 bytes)
  Dropped           : 1001 pkts (128166 bytes)
Egr. Rate-limiter
  Offered           : 0 pkts
  Forwarded         : 0 pkts
  Dropped           : 0 pkts
===============================================================================
*A:PE1#
```

When sending traffic at a rate of 1000 packets/s with a 256 bytes packet-size and monitoring at entry-point SAP 3/2/13 over 20 s intervals, then 20,000 packets are received on interface int-TST-1 accounting for 5,120,000 bytes, as follows:

```
*A:PE1# monitor service id 1 sap 3/2/13 interval 20
===============================================================================
Monitor statistics for Service 1 SAP 3/2/13
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
                        Packets                 Octets
CPM Ingress           : 25                      1614
Forwarding Engine Stats
Dropped               : 128590687               8338701952
Received Valid        : 331812178               23060748680
Off. HiPrio           : 0                       0
Off. LowPrio          : 311643389               20030922920
Off. Uncolor          : 0                       0
Off. Managed          : 0                       0
Queueing Stats(Ingress QoS Policy 1)
Dro. HiPrio           : 0                       0
Dro. LowPrio          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 311643389               20030922920

--- snipped ---

-------------------------------------------------------------------------------
```

```
At time t = 20 sec (Mode: Delta)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
                        Packets                 Octets
CPM Ingress           : 0                       0
Forwarding Engine Stats
Dropped               : 0                       0
Received Valid        : 19901                   5094656
Off. HiPrio           : 0                       0
Off. LowPrio          : 0                       0
Off. Uncolor          : 0                       0
Off. Managed          : 0                       0

--- snipped ---

-------------------------------------------------------------------------------
At time t = 40 sec (Mode: Delta)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
                        Packets                 Octets
CPM Ingress           : 0                       0
Forwarding Engine Stats
Dropped               : 0                       0
Received Valid        : 20000                   5120000
Off. HiPrio           : 0                       0
Off. LowPrio          : 0                       0
Off. Uncolor          : 0                       0
Off. Managed          : 0                       0

--- snipped ---

^C
*A:PE1#
```

When sending at a rate of 3000 packets/sec a with a 256 bytes packet-size and monitoring at exit-point SAP 3/2/14 over 20 s intervals, then 10,000 packets are sent out of interface int-TST-2 accounting for 2,560,000 bytes, as follows:

```
*A:PE1# monitor service id 1 sap 3/2/14 interval 20
===============================================================================
Monitor statistics for Service 1 SAP 3/2/14
===============================================================================
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
                        Packets                 Octets
CPM Ingress           : 3544                    212716
Forwarding Engine Stats
```

```
Dropped               : 0                       0
Received Valid        : 312516277               20001041728
Off. HiPrio           : 0                       0
Off. LowPrio          : 312516277               20001041728
Off. Uncolor          : 0                       0
Off. Managed          : 0                       0
Queueing Stats(Ingress QoS Policy 1)
Dro. HiPrio           : 0                       0
Dro. LowPrio          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 312516277               20001041728
Queueing Stats(Egress QoS Policy 1)
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0
For. In/InplusProf    : 10360173                1590874396
For. Out/ExcProf      : 311585647               20027227432
-------------------------------------------------------------------------------
Sap per Queue Stats
-------------------------------------------------------------------------------
                        Packets                 Octets
Ingress Queue 1 (Unicast) (Priority)
Off. HiPrio           : 0                       0
Off. LowPrio          : 312516277               20001041728
Dro. HiPrio           : 0                       0
Dro. LowPrio          : 0                       0
For. InProf           : 0                       0
For. OutProf          : 312516277               20001041728
Egress Queue 1
For. In/InplusProf    : 10360173                1590874396
For. Out/ExcProf      : 311585647               20027227432
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0
-------------------------------------------------------------------------------
At time t = 20 sec (Mode: Delta)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
                        Packets                 Octets
CPM Ingress           : 0                       0

--- snipped ---

Queueing Stats(Egress QoS Policy 1)
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0
For. In/InplusProf    : 10016                   2564096
For. Out/ExcProf      : 0                       0

--- snipped ---

-------------------------------------------------------------------------------
At time t = 40 sec (Mode: Delta)
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Sap Statistics
-------------------------------------------------------------------------------
Last Cleared Time     : N/A
```

```
                        Packets                 Octets
CPM Ingress           : 0                       0

--- snipped ---

Queueing Stats(Egress QoS Policy 1)
Dro. In/InplusProf    : 0                       0
Dro. Out/ExcProf      : 0                       0
For. In/InplusProf    : 10005                   2561280
For. Out/ExcProf      : 0                       0

--- snipped ---

^C
*A:PE1#
```

Conclusion

Rate-limiting filter actions can be used by network operators for security purposes to protect network resources and can also be used to mitigate DDoS attacks.

Weighted ECMP for 6PE over RSVP-TE LSPs

This chapter provides information about Weighted ECMP for 6PE over RSVP-TE LSPs.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R7. Weighted ECMP for IPv6 provider edge routers (6PE) over RSVP-TE LSPs is supported in SR OS Release 15.0.R6, and later.

Chapter Weighted ECMP for VPRN over RSVP-TE and SR-TE LSPs is recommended reading.

Overview

Equal Load Balancing

In this chapter, equal cost multipath (ECMP) refers to spraying traffic flows over multiple RSVP-TE LSPs within an ECMP set. ECMP spraying consists of hashing the relevant fields in the packet header and selecting the tunnel next-hop based on the modulo operation of the output of the hash and the number of RSVP-TE LSPs present in the ECMP set. The maximum number of RSVP-TE LSPs in the ECMP set is defined by the **ecmp** command.

Only RSVP-TE LSPs with the same lowest LSP metric can be part of the ECMP set. If the number of such RSVP-TE LSPs exceeds the maximum number of RSVP-TE LSPs allowed in the ECMP set as defined by the **ecmp** command, the RSVP-TE LSPs with the lowest tunnel IDs are selected first. By default, all RSVP-TE LSPs in the ECMP set have the same weight, and traffic flows are spread evenly over all RSVP-TE LSPs in the ECMP set, regardless of the bandwidth of the active path in the RSVP-TE LSPs. By default, ECMP is enabled and set to 1.

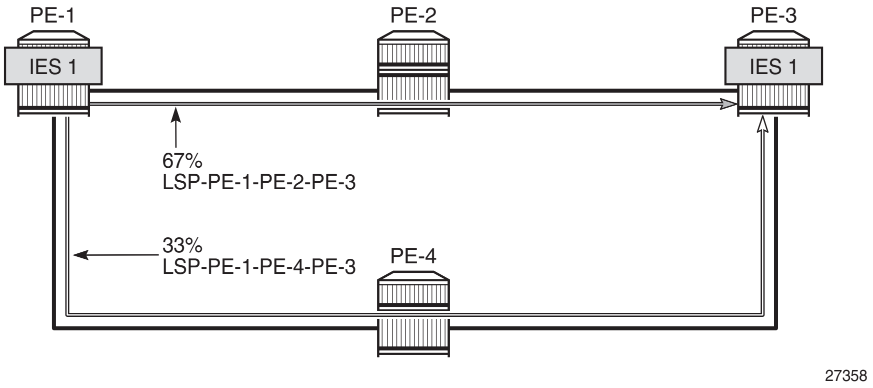
The following command enables ECMP with load balancing over two paths in the base router:

```
configure router ecmp 2
```

Unequal Load Balancing

Weighted ECMP sprays traffic flows over RSVP-TE LSPs proportionally to the **load-balancing-weight** <*weight*> value configured on each RSVP-TE LSP in the ECMP set. Figure 76 shows that PE-1 forwards two thirds of the traffic flows on LSP-PE-1-PE-2-PE-3 with weight 2 and one third on LSP-PE-1-PE-4-PE-3 with weight 1.

Figure 76 **Weighted ECMP in AS 64496**



The LSP load-balancing weight can be configured in an LSP template or on an RSVP-TE LSP. By default, the load-balancing weight equals zero, in which case regular ECMP applies.

Weighted load balancing can be performed only when all the next-hops are associated with the same neighbor and all the RSVP-TE LSPs are configured with a non-zero load-balancing weight. If one or more RSVP-TE LSPs in the ECMP set toward a specific next-hop do not have a load-balancing weight configured, regular ECMP spraying is used.

The following command is used to configure the weight in an LSP template:

```
*A:PE-1# configure router mpls lsp-template "LSPtemplate1" load-balancing-weight
  - no load-balancing-weight
  - load-balancing-weight <weight>

 <weight>             : [0..4294967295] Default - 0
```

The following command is used to configure the weight on an LSP:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2-PE-3" load-balancing-weight
  - load-balancing-weight <weight>
  - no load-balancing-weight

 <weight>             : [0..4294967295] Default - 0
```

The LSP load-balancing weight on LSP-PE-1-PE-2-PE-3 is configured with a value of 2, as follows:

```
configure
    router
        mpls
            lsp "LSP-PE-1-PE-2-PE-3"
                to 192.0.2.3
                cspf
                metric 100
                load-balancing-weight 2
                primary "path-PE-1-PE-2-PE-3"
                exit
                no shutdown
            exit
```

Weighted ECMP for 6PE over RSVP-TE LSPs is enabled in the BGP next-hop resolution context as follows:

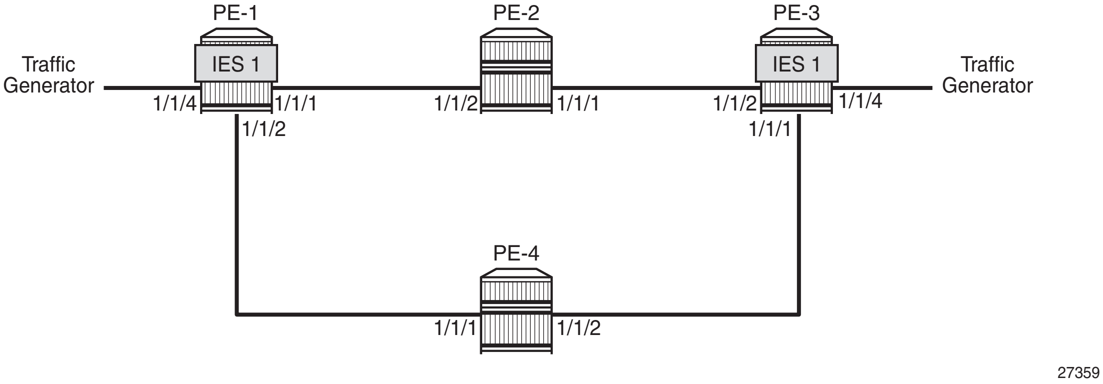
```
configure router bgp next-hop-resolution weighted-ecmp
```

The **weighted-ecmp** option controls load balancing to the same next-hop only.

Configuration

Figure 77 shows the example topology with four PEs. IES 1 is configured on PE-1 and PE-3. A traffic generator is connected to IES 1 SAP 1/1/4 on PE-1 and IES 1 SAP 1/1/4 on PE-3. The traffic generator will generate multiple IPv6 traffic flows with random IP addresses and TCP/UDP port numbers and these flows will be sprayed over different RSVP-TE LSPs between PE-1 and PE-3.

Figure 77 **Example Topology for 6PE over RSVP-TE LSPs**



Initial Configuration

The initial configuration on the PEs includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP (alternatively, OSPF can be used) with traffic engineering enabled
- MPLS and RSVP enabled on all router interfaces
- MPLS paths with strict hops from PE-1 to PE-3 and vice versa: one via PE-2 and the other via PE-4. The LSP via PE-2 gets a load-balancing weight of 2, whereas the LSP via PE-4 gets a load-balancing weight of 1. Both LSPs have the same metric.

The initial configuration on PE-1 is as follows. The configuration on PE-3 is similar.

```
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-4"
            address 192.168.14.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        isis
            area-id 49.0001
            traffic-engineering
            interface "system"
            exit
            interface "int-PE-1-PE-2"
                interface-type point-to-point
            exit
            interface "int-PE-1-PE-4"
                interface-type point-to-point
            exit
            no shutdown
        exit
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-4"
            exit
            path "path-PE-1-PE-2-PE-3"
                hop 10 192.168.12.2 strict
                hop 20 192.168.23.2 strict
                no shutdown
            exit
            path "path-PE-1-PE-4-PE-3"
                hop 10 192.168.14.2 strict
                hop 20 192.168.34.1 strict
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2-PE-3"
                to 192.0.2.3
                cspf
                metric 100
                load-balancing-weight 2
                primary "path-PE-1-PE-2-PE-3"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-PE-4-PE-3"
                to 192.0.2.3
                cspf
                metric 100
                load-balancing-weight 1
                primary "path-PE-1-PE-4-PE-3"
                exit
                no shutdown
            exit
            no shutdown
        exit
        rsvp
            no shutdown
        exit
```

With the preceding configuration, MPLS and RSVP are enabled on all interfaces, including the system interface, which is added automatically.

Weighted ECMP for 6PE over RSVP-TE LSPs

BGP will be configured for the label-IPv6 address family and the next-hop resolution will be set to RSVP; see the *6PE Next-Hop Resolution* chapter.

In this example, the traffic generator sends IPv6 traffic to the SAP in IES 1. The IPv6 packets will be tunneled through the IPv4 network between PE-1 and PE-3. The service configuration on PE-1 is as follows. The configuration on PE-3 is similar.

```
configure
    service
        ies 1 customer 1 create
            description "6PE"
            interface "int-PE-1-CE-1" create
                ipv6
                    address 2001:db8::11:1/120
                exit
                sap 1/1/4 create
                exit
            exit
            no shutdown
        exit
```

On PE-1, the following BGP configuration defines next-hop resolution with weighted ECMP and the resolution filter only allows RSVP-TE LSPs. BGP is configured for the label-IPv6 address family and BGP multipath is configured in the BGP context. The configuration on PE-3 is similar.

```
configure
    router
        autonomous-system 64496
        bgp
            ibgp-multipath
            split-horizon
            next-hop-resolution
                weighted-ecmp
                labeled-routes
                    transport-tunnel
                        family label-ipv6
                            resolution-filter
                                rsvp
                            exit
                            resolution filter
                        exit
                    exit
                exit
            exit
            group "iBGP"
                export "export-6PE"
                peer-as 64496
                neighbor 192.0.2.3
                    family label-ipv6
                exit
            exit
```

On PE-1 and PE-3, the following export policy is configured:

```
configure
    router
        policy-options
            begin
            policy-statement "export-6PE"
                entry 10
                    from
                        protocol direct
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The following command enables ECMP in the base router.

```
configure router ecmp 2
```

On PE-1, the route table in the base router shows that the remote prefix 2001:db8::33:0/120 has flag [2], meaning that the next-hop 192.0.2.3 occurs twice for this prefix, as follows:

```
*A:PE-1# show router route-table 2001:db8::33:0/120

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
2001:db8::33:0/120 [2]                        Remote  BGP_LABEL 00h00m28s  170
       192.0.2.3 (tunneled)                                       100
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
---snip---
```

The route table on PE-3 shows a similar route with flag [2] for prefix 2001:db8::11:0/120.

On PE-1, the following detailed route table info (using keyword **extensive**) for prefix 2001:db8::33:0/120 shows that RSVP-TE tunnel 1 and RSVP-TE tunnel 2 are used to reach the next-hop 192.0.2.3. Both RSVP-TE tunnels have metric 100, but the weight of RSVP-TE tunnel 1 is twice as much as the weight of RSVP tunnel 2, so the load on RSVP-TE LSP 1 is twice as high as the load on RSVP LSP 2.

```
*A:PE-1# show router route-table 2001:db8::33:0/120 extensive

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix             : 2001:db8::33:0/120
  Protocol              : BGP_LABEL
  Age                   : 00h16m13s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.3
    Label               : 2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.3 (RSVP tunnel:1)
      Metric            : 100
      ECMP-Weight       : 2
    Resolving Next-Hop  : 192.0.2.3 (RSVP tunnel:2)
      Metric            : 100
      ECMP-Weight       : 1
-------------------------------------------------------------------------------
No. of Destinations: 1
```

The following tunnel table on PE-1 shows that RSVP-TE tunnel 1 has PE-2 as next-hop (192.168.12.2) and RSVP-TE tunnel 2 has next-hop PE-4 (192.168.14.2):

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.3/32      rsvp      MPLS  1         7        192.168.12.2   100
192.0.2.3/32      rsvp      MPLS  2         7        192.168.14.2   100
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

Traffic Verification

The traffic generator sends IPv6 traffic flows to SAP 1/1/4 of IES 1 on PE-1. The packets will be tunneled over the available RSVP-TE LSPs present in the ECMP set. The traffic will be load balanced unevenly: two thirds of the traffic flows will be tunneled via PE-2 (port 1/1/2) while one third of the traffic flows will be tunneled via PE-4 (port 1/1/1). The load on the ports is as follows:

```
*A:PE-1# monitor port 1/1/1 1/1/2 1/1/4 rate interval 3 repeat 3

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                               126                3786813
Packets                                                1                   3670
Errors                                                 0                      0
Bits                                                1008               30294504
Utilization (% of port capacity)                   ~0.00                  30.88

Port 1/1/2
-------------------------------------------------------------------------------
Octets                                                86                1761381
Packets                                                1                   1707
Errors                                                 0                      0
Bits                                                 688               14091048
Utilization (% of port capacity)                   ~0.00                  14.36

Port 1/1/4
-------------------------------------------------------------------------------
Octets                                           5505024                      0
Packets                                             5376                      0
Errors                                                 0                      0
Bits                                            44040192                      0
Utilization (% of port capacity)                   44.90                   0.00

-------------------------------------------------------------------------------
```

Conclusion

Operators can control how 6PE traffic is load balanced unequally over multiple RSVP-TE LSPs by defining a load-balancing weight value on each LSP.

Unicast Routing Protocols

In This Section

This section provides configuration information for the following topics:

Associating Communities with Static and Aggregate Routes

This chapter provides information about associating communities with static and aggregate routes configurations.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 11.0.R3, but the CLI in this edition corresponds to release 14.0.R2. There are no prerequisites for this configuration.

Introduction

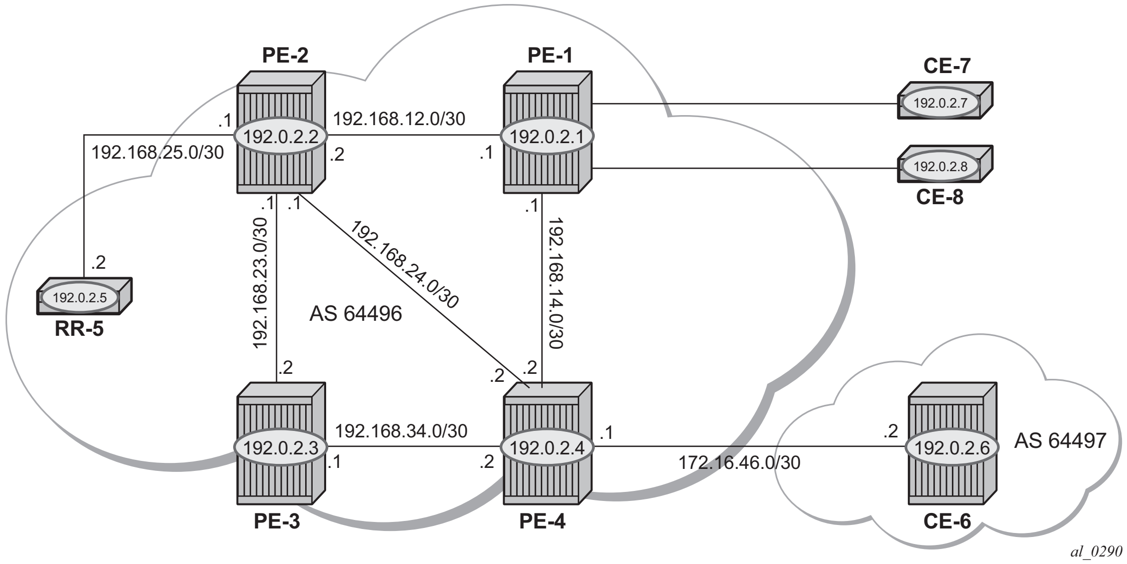
Border gateway protocol (BGP) communities are optional, transitive attributes attached to BGP route prefixes to carry additional information about that route prefix. A number of route prefixes can have the same community attached such that it can be matched by a route policy. As a result, the presence of a community value can be used to influence and control route policy.

A BGP community is a 32-bit value that is written as two separate 16 bit numbers separated by a colon. The first number usually represents the autonomous system (AS) number that defines or originates the community whilst the second is set by the network administrator.

Knowledge of RFC 4271 (BGP-4) and RFC 1997 (BGP Communities Attribute) is assumed throughout this document, as well as knowledge of multi-protocol BGP (MP-BGP) and RFC 4364 (BGP/MPLS IP VPNs).

Overview

Figure 78 **Example Topology**



The example topology is displayed in Figure 78. The setup uses 7750 Service Router (SR) nodes. PE-1 to PE-4 and the Route Reflector (RR-5) are located in the same Autonomous System (AS); AS64496. CE-6 is in a separate AS 64497 and peers using eBGP with its directly connected neighbor, PE-4.

The objectives are:

- To configure static-routes in a VPRN in PE-1 with various community values – including well-known communities – export them to other PEs within the same AS, and then via eBGP to CE-6. During this process, the community values for each route will be examined to ensure that the transitive nature of the attribute is maintained.
- To associate a community with an aggregate route that represents a larger number of composite prefixes. The aggregate will be advertised in place of the composite prefixes.

The following configuration tasks should be completed as a pre-requisite:

- Full mesh IS-IS or OSPF between all of the PE routers and the route reflector.
- iBGP between the RR and all PEs.
- eBGP between PE-4 and CE-6.
- Link-layer LDP between each PE.

Associating Communities with Static and Aggregate Routes

It is possible to add a single community value to a static and aggregate route without using a route policy.

The community value can be in the 4-byte format comprising of a 2-byte AS value, followed by a 2 byte decimal value, separated by a colon. It can also be the name of a well-known standard community; no-export, no-advertise, no-export-subconfed.

Any community added can be matched using a route policy.

The purpose of this example is to provision static and aggregate IPv4 route prefixes and associate a community with each route. These routes are then redistributed into the BGP protocol and advertised to other BGP speakers.

This is shown for IPv4 routes within a VPRN. Well-known, standard communities will also be configured to show that the correct behavior is observed.

Configuration

The first step is to configure an iBGP session between each of the PEs and the Route Reflector (RR). The address family negotiated between peers is vpn-ipv4.

The configuration for all PEs is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                family vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.5
                exit
            exit
```

The IP addresses can be derived from Figure 78.

The configuration for RR-5 is:

```
configure
    router
        autonomous-system 64496
        bgp
            cluster 0.0.0.1
            group “RR-clients"
                family vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
```

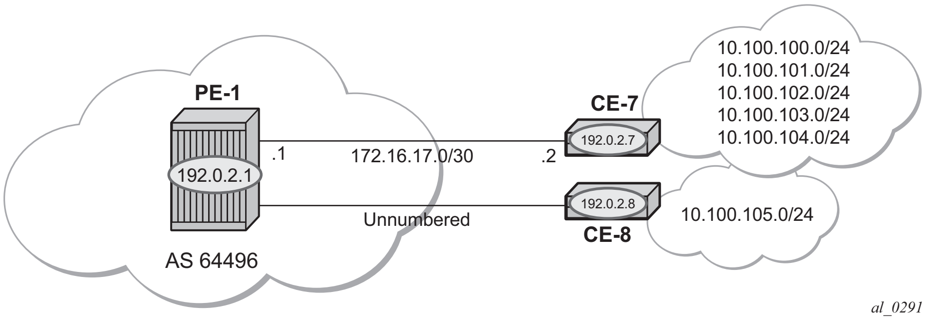
On RR-5, show that BGP sessions with each PE are established, and have correctly negotiated the VPN IPv4 address family capability.

```
*A:RR-5# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.5        AS:64496       Local AS:64496
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 4
Total BGP Paths         : 6           Total Path Memory           : 1104
Total IPv4 Remote Rts   : 0           Total IPv4 Rem. Active Rts  : 0
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total IPv6 Remote Rts   : 0           Total IPv6 Rem. Active Rts  : 0
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0

Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total VPN Local Rts     : 0
Total VPN-IPv4 Rem. Rts : 0           Total VPN-IPv4 Rem. Act. Rts: 0
Total VPN-IPv6 Rem. Rts : 0           Total VPN-IPv6 Rem. Act. Rts: 0
Total VPN-IPv4 Bkup Rts : 0           Total VPN-IPv6 Bkup Rts     : 0

Total VPN Supp. Rts     : 0           Total VPN Hist. Rts         : 0
Total VPN Decay Rts     : 0

Total L2-VPN Rem. Rts   : 0           Total L2VPN Rem. Act. Rts   : 0
Total MVPN-IPv4 Rem Rts : 0           Total MVPN-IPv4 Rem Act Rts : 0
Total MDT-SAFI Rem Rts  : 0           Total MDT-SAFI Rem Act Rts  : 0
Total MSPW Rem Rts      : 0           Total MSPW Rem Act Rts      : 0
Total RouteTgt Rem Rts  : 0           Total RouteTgt Rem Act Rts  : 0
Total McVpnIPv4 Rem Rts : 0           Total McVpnIPv4 Rem Act Rts : 0
Total MVPN-IPv6 Rem Rts : 0           Total MVPN-IPv6 Rem Act Rts : 0
Total EVPN Rem Rts      : 0           Total EVPN Rem Act Rts      : 0
Total FlowIpv4 Rem Rts  : 0           Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0           Total FlowIpv6 Rem Act Rts  : 0
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
```

```
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
               64496       3    0 00h00m05s 0/0/0 (VpnIPv4)
                           3    0
192.0.2.2
               64496       3    0 00h00m05s 0/0/0 (VpnIPv4)
                           3    0
192.0.2.3
               64496       3    0 00h00m05s 0/0/0 (VpnIPv4)
                           3    0
192.0.2.4
               64496       3    0 00h00m05s 0/0/0 (VpnIPv4)
                           3    0
-------------------------------------------------------------------------------
*A:RR-5#
```

VPRN: IPv4

Figure 79 **CE Connections for Next-Hops**



The VPRN configuration for PE-1 is as follows:

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-1-CE-7" create
                address 172.16.17.1/30
                sap 1/2/1:1.0 create
                exit
            exit
            interface "loop1" create
                address 192.0.2.100/32
                loopback
            exit
            interface "int-PE-1-CE-8" create
                unnumbered "loop1"
                sap 1/2/2:1.0 create
            exit
        exit
        no shutdown
```

For unnumbered interfaces, an IP address is borrowed from a loopback interface, for example from the system interface, see MPLS chapter Unnumbered Interfaces in RSVP-TE and LDP.

LDP is used as the label-switching protocol for next-hop resolution.

PE-4 is configured with an interface toward CE-6 that supports eBGP. The following export policy is configured:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-VPN-accept"
                entry 10
                    from
                        protocol bgp-vpn
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

The configuration of the VPRN service is as follows:

```
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-4-CE-6" create
                address 172.16.46.1/30
                sap 1/2/1:1 create
                exit
            exit
            bgp
```

```
                group "VPRN1-external"
                    export "BGP-VPN-accept"
                    peer-as 64497
                    neighbor 172.16.46.2
                    exit
                exit
            exit
            no shutdown
```

Static Routes with Communities

A static route has a number of next-hop options: direct connected IP address, black-hole, indirect IP address, and interface-name.

Figure 79 shows a pair of Customer Edge (CE) routers connected to PE1. The link to CE-7 is a numbered link. The link to CE-8 is an unnumbered link. The loopback interface address is used as a reference address for the unnumbered Ethernet interface.

Beyond CE-7 are a number of /24 subnets. Static routes to these individual subnets are created on PE-1 using a static route with a next-hop type of “interface address” or an “indirect address”. The indirect address is learned using a static route.

Beyond CE-8 is a single /24 subnet. A static route to this subnet is created with an interface-name as the next-hop.

There are a number of well-known, standard communities:

- no-export: the route is not advertised to any external peer. This route should be present in the route tables of all BGP speakers in the originating AS, but not in those in neighboring ASs.
- no-advertise: the route is not advertised to any peer. This route should not be present in any router as BGP-learned route.

The requirement for each subnet is:

- 10.100.100.0/24 must not be advertised outside of the AS. This must be associated with the standard, well-known community no-export. The community value is encoded as 65535:65281 (0xFFFFFF01), but the CLI requires the keyword **no-export**.

```
configure service vprn 1
            static-route-entry 10.100.100.0/24
                next-hop 172.16.17.2
                    community no-export
                    no shutdown
                exit
```

- 10.100.101.0/24 must be advertised with a community of 64496:101

```
configure service vprn 1
            static-route-entry 10.100.101.0/24
                next-hop 172.16.17.2
                    community 64496:101
                    no shutdown
                exit
```

- 10.100.102.0/24 must not be advertised to any BGP peer. This must be associated with the standard, well-known community **no-advertise**. The community value is encoded as 65535:65282 (0xFFFFFF02), but the CLI requires the keyword **no-advertise**.

```
configure service vprn 1
            static-route-entry 10.100.102.0/24
                next-hop 172.16.17.2
                    community no-advertise
                    no shutdown
                exit
```

- 10.100.103.0/24 must be advertised with a community of 64496:103 and a route tag of 10.

```
configure service vprn 1
            static-route-entry 10.100.103.0/24
                next-hop 172.16.17.2
                    community 64496:103
                    tag 10
                    no shutdown
                exit
            exit
```

- 10.100.104.0/24 must be advertised with a community of 64496:104. It is reachable via 192.0.2.7 which, in turn, is reachable via 172.16.17.2. For more information about this configuration, see chapter Associating Communities with Static and Aggregate Routes. This is using a static route which does not need to be advertised – hence it is associated with the **no-advertise** community.

```
configure service vprn 1
            static-route-entry 10.100.104.0/24
                indirect 192.0.2.7
                    community 64496:104
                    no shutdown
                exit
            exit
            static-route-entry 192.0.2.7/32
                next-hop 172.16.17.2
                    community no-advertise
                    no shutdown
                exit
            exit
```

- 10.100.105.0/24 must be advertised with a community of 64496:105. It is reachable via the unnumbered interface to CE-8.

```
configure service vprn 1
            static-route-entry 10.100.105.0/24
                next-hop "int-PE-1-CE-8"
                    community 64496:105
                    no shutdown
                exit
            exit
```

On PE-1, configure static routes that match the static routes from Figure 79, and the preceding conditions.

The default behavior of a VPRN is to export all static and connected routes into a BGP labeled route with the appropriate route-target extended community configured in the vrf-target statement. A single community string can be added using the static-route community commands shown above. If multiple communities are required, then a VRF-export policy should be used. This is outside the scope of this chapter.

Examine the BGP table of PE-1 to establish that routes have been exported correctly in VPN IPv4 toward RR-5.

```
*A:PE-1# show router bgp neighbor 192.0.2.5 advertised-routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
i     64496:1:10.100.100.0/24                            100         None
-------------------------------------------------------------------------------
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:10.100.101.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:10.100.103.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:10.100.104.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:10.100.105.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:172.16.17.0/30                             100         None
      192.0.2.1                                          None        262139
      No As-Path
i     64496:1:192.0.2.100/32                             100         None
      192.0.2.1                                          None        262139
      No As-Path
```

```
-------------------------------------------------------------------------------
Routes : 7
===============================================================================
*A:PE-1#
```

There are only seven exported routes. The route prefixes associated with the **no-advertise** community are not present, as expected.

Examining the BGP table of PE-4 shows the presence of the expected routes, with the correct community values.

The prefix 10.100.100.0/24 is a member of community **no-export**. This is not advertised to PE-4.

In release 14.0.R1, the **show router bgp routes** command has been restructured to support more consistent parameter ordering. In most cases, a family filter must be specified before all other filtering parameters, except for the prefix.

```
*A:PE-4# show router bgp routes 10.100.100.0/24 vpn-ipv4 detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.100.100.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262139
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-4-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : no-export target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Add Paths Send : Default
Source Class   : 0                      Dest Class     : 0
Last Modified  : 01h16m07s
```

```
VPRN Imported  :  1
---snip---
```

The following command shows all members of the community no-report:

```
*A:PE-4# show router bgp routes vpn-ipv4 community no-export
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:10.100.100.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

Because the community no-export is encoded as community 65535:65281, the same output can be retrieved as follows:

```
*A:PE-4# show router bgp routes vpn-ipv4 community 65535:65281
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:10.100.100.0/24                            100         None
      192.0.2.1                                          None        262139
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

The prefix 10.100.101.0/24 is a member of community 64496:101. This is correctly advertised to PE-4.

```
*A:PE-4# show router bgp routes 10.100.101.0/24 vpn-ipv4 detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.100.101.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262139
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-4-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : 64496:101 target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h34m23s
VPRN Imported  :  1
---snip---
```

The prefix 10.100.103.0/24 is a member of community 64496:103. This is correctly advertised to PE-4.

```
*A:PE-4# show router bgp routes 10.100.103.0/24 vpn-ipv4 detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Original Attributes
```

```
Network        : 10.100.103.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262139
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-4-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : 64496:103 target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h26m24s
VPRN Imported  :  1
---snip---
```

The prefix 10.100.104.0/24 is a member of community 64496:104. This is correctly advertised to PE-4.

```
*A:PE-4# show router bgp routes 10.100.104.0/24 vpn-ipv4 detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.100.104.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262139
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-4-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : 64496:104 target:64496:1
Cluster        : 0.0.0.1
```

```
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h20m45s
VPRN Imported  :  1
---snip---
```

The prefix 10.100.105.0/24 is a member of community 64496:105. This is correctly advertised to PE-4.

```
*A:PE-4# show router bgp routes 10.100.105.0/24 vpn-ipv4 detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.100.105.0/24
Nexthop        : 192.0.2.1
Route Dist.    : 64496:1                VPN Label      : 262139
Path Id        : None
From           : 192.0.2.5
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : int-PE-4-PE-1
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : 64496:105 target:64496:1
Cluster        : 0.0.0.1
Originator Id  : 192.0.2.1              Peer Router Id : 192.0.2.5
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: N/A
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 01h18m11s
VPRN Imported  :  1
---snip---
```

Examine the route table of VPRN 1 on PE-4 – looking specifically at the BGP-learned routes, the same seven routes are present as valid routes.

```
*A:PE-4# show router 1 route-table protocol bgp-vpn

===============================================================================
Route Table (Service: 1)
===============================================================================
      Next Hop[Interface Name]                                    Metric
Dest Prefix[Flags]                            Type    Proto     Age        Pref
-------------------------------------------------------------------------------
10.100.100.0/24                               Remote  BGP VPN   01h54m30s  170
       192.0.2.1 (tunneled)                                         0
10.100.101.0/24                               Remote  BGP VPN   01h46m55s  170
       192.0.2.1 (tunneled)                                         0
10.100.103.0/24                               Remote  BGP VPN   01h37m47s  170
       192.0.2.1 (tunneled)                                         0
10.100.104.0/24                               Remote  BGP VPN   01h30m18s  170
       192.0.2.1 (tunneled)                                         0
10.100.105.0/24                               Remote  BGP VPN   01h26m58s  170
       192.0.2.1 (tunneled)                                         0
172.16.17.0/30                                Remote  BGP VPN   01h54m30s  170
       192.0.2.1 (tunneled)                                         0
192.0.2.100/32                                Remote  BGP VPN   01h54m30s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 7
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-4#
```

Examine the route table of CE-6 – looking specifically at the BGP-learned routes, six routes are present as valid routes, as expected.

```
*A:CE-6# show router route-table protocol bgp

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.100.101.0/24                               Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
10.100.103.0/24                               Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
10.100.104.0/24                               Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
10.100.105.0/24                               Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
172.16.17.0/30                                Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
192.0.2.100/32                                Remote  BGP       00h04m31s  170
       172.16.46.1                                                  0
```

```
--------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:CE-6#
```

The prefix 10.100.100.0/24 is not received from PE-4 as it is a member of the **no-export** community.

```
*A:CE-6# show router bgp routes 10.100.100.0/24 detail

================================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

================================================================================
BGP IPv4 Routes
================================================================================
No Matching Entries Found
================================================================================
*A:CE-6#
```

Static route 10.100.101.0/24 is received with the correct community 64496:101.

```
*A:CE-6# show router bgp routes community 64496:101
================================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

================================================================================
BGP IPv4 Routes
================================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
--------------------------------------------------------------------------------
u*>i  10.100.101.0/24                                    None        None
      172.16.46.1                                        None        -
      64496
--------------------------------------------------------------------------------
Routes : 1
================================================================================
*A:CE-6#
```

Static route 10.100.103.0/24 is received with the correct community 64496:103.

```
*A:CE-6# show router bgp routes community 64496:103

===============================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.100.103.0/24                                    None        None
      172.16.46.1                                        None        -
      64496
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CE-6#
```

Static route 10.100.104.0/24 is received with the correct community 64496:104.

```
*A:CE-6# show router bgp routes community 64496:104
===============================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
 Legend -
===============================================================================
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.100.104.0/24                                    None        None
      172.16.46.1                                        None        -
      64496
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CE-6#
```

Static route 10.100.105.0/24 is received with the correct community 64496:105.

```
*A:CE-6# show router bgp routes community 64496:105
===============================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
===============================================================================
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
 Legend -
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.100.105.0/24                                    None        None
      172.16.46.1                                        None        -
      64496
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:CE-6#
```

Aggregate Routes with Communities

An aggregate route can be configured to represent a larger number of prefixes. For example, a set of prefixes 10.101.0.0/24 to 10.101.7.0/24 can be represented as a single aggregate prefix of 10.101.0.0/21.

This is due to the fact that the third octet in the range 0 to 7 can be represented by the 8 bits 00000000 to 00000111. The first 5 bits of this octet are common, along with the previous 2 octets, giving a prefix where the first 21 bits are common. Therefore, the aggregate can be written as 10.101.0.0/21.

In order to illustrate the configuration of an aggregate, consider following.

Figure 80 **CE-7 Connectivity**

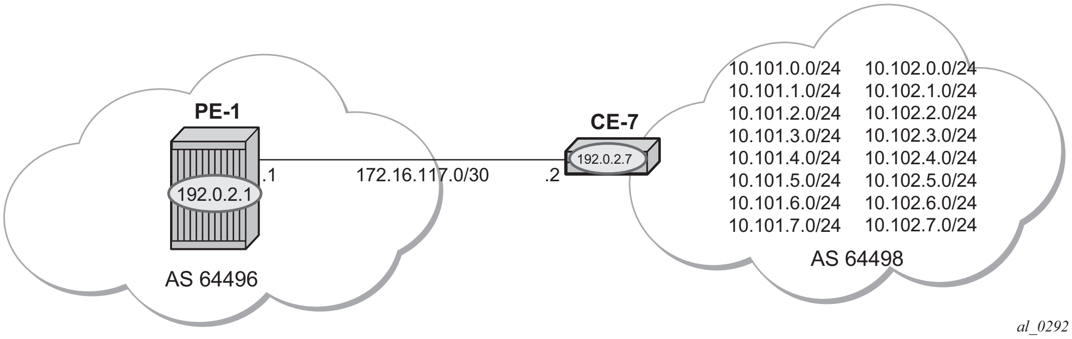


Figure 80 shows a CE router (CE-7), in AS 64498, that advertises a series of contiguous prefixes via BGP.

- 10.101.0.0/24 to 10.101.7.0/24
- 10.102.0.0/24 to 10.102.7.0/24

Instead of advertising all of these prefixes out of the VPRN towards an external CE individually, an aggregate route can be configured that summarizes each set of eight prefixes and a community can be directly associated with each aggregate route.

The configuration for a VPRN on PE-1, including the external BGP configuration is as follows:

```
*A:PE-1# configure
    service
        vprn 2 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:2
            interface "int-PE-1-CE-7_2nd" create
                address 172.16.117.1/30
                sap 1/2/1:2.0 create
                exit
            exit
            bgp
                group "external"
                    peer-as 64498
                    neighbor 172.16.117.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The neighbor relationship shows:

```
*A:PE-1# show router 2 bgp neighbor

===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 172.16.117.2
Description          : (Not Specified)
Group                : external
-------------------------------------------------------------------------------
Peer AS              : 64498            Peer Port            : 49673
Peer Address         : 172.16.117.2
Local AS             : 64496            Local Port           : 179
Local Address        : 172.16.117.1
Peer Type            : External         Dynamic Peer         : No
```

```
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : IPv4
Remote Family        : IPv4
Hold Time            : 90               Keep Alive           : 30
Min Hold Time        : 0
Active Hold Time     : 90               Active Keep Alive    : 30
Cluster Id           : None
Preference           : 170              Num of Update Flaps  : 0
Recd. Paths          : 1
IPv4 Recd. Prefixes  : 16               IPv4 Active Prefixes : 16
IPv4 Suppressed Pfxs : 0                VPN-IPv4 Suppr. Pfxs : 0
VPN-IPv4 Recd. Pfxs  : 0                VPN-IPv4 Active Pfxs : 0
Mc IPv4 Recd. Pfxs.  : 0                Mc IPv4 Active Pfxs. : 0
Mc IPv4 Suppr. Pfxs  : 0                IPv6 Suppressed Pfxs : 0
IPv6 Recd. Prefixes  : 0                IPv6 Active Prefixes : 0
VPN-IPv6 Recd. Pfxs  : 0                VPN-IPv6 Active Pfxs : 0
VPN-IPv6 Suppr. Pfxs : 0
Mc IPv6 Recd. Pfxs.  : 0                Mc IPv6 Active Pfxs. : 0
Mc IPv6 Suppr. Pfxs  : 0                L2-VPN Suppr. Pfxs   : 0
L2-VPN Recd. Pfxs    : 0                L2-VPN Active Pfxs   : 0
MVPN-IPv4 Suppr. Pfxs: 0                MVPN-IPv4 Recd. Pfxs : 0
MVPN-IPv4 Active Pfxs: 0                MDT-SAFI Suppr. Pfxs : 0
MDT-SAFI Recd. Pfxs  : 0                MDT-SAFI Active Pfxs : 0
Flow-IPv4 Suppr. Pfxs: 0                Flow-IPv4 Recd. Pfxs : 0
Flow-IPv4 Active Pfxs: 0                Rte-Tgt Suppr. Pfxs  : 0
Rte-Tgt Recd. Pfxs   : 0                Rte-Tgt Active Pfxs  : 0
Backup IPv4 Pfxs     : 0                Backup IPv6 Pfxs     : 0
Mc Vpn Ipv4 Recd. Pf*: 0                Mc Vpn Ipv4 Active P*: 0
Mc Vpn Ipv4 Suppr. P*: 0
Backup Vpn IPv4 Pfxs : 0                Backup Vpn IPv6 Pfxs : 0
Input Queue          : 0                Output Queue         : 0
i/p Messages         : 6                o/p Messages         : 6
i/p Octets           : 228              o/p Octets           : 232
i/p Updates          : 1                o/p Updates          : 1
MVPN-IPv6 Suppr. Pfxs: 0                MVPN-IPv6 Recd. Pfxs : 0
MVPN-IPv6 Active Pfxs: 0
Flow-IPv6 Suppr. Pfxs: 0                Flow-IPv6 Recd. Pfxs : 0
Flow-IPv6 Active Pfxs: 0
Evpn Suppr. Pfxs     : 0                Evpn Recd. Pfxs      : 0
Evpn Active Pfxs     : 0
MS-PW Suppr. Pfxs    : 0                MS-PW Recd. Pfxs     : 0
MS-PW Active Pfxs    : 0
TTL Security         : Disabled         Min TTL Value        : n/a
Graceful Restart     : Disabled         Stale Routes Time    : n/a
Restart Time         : n/a
Advertise Inactive   : Disabled         Peer Tracking        : Disabled
Advertise Label      : None
Auth key chain       : n/a
Disable Cap Nego     : Disabled         Bfd Enabled          : Disabled
Flowspec Validate    : Disabled         Default Route Tgt    : Disabled
Aigp Metric          : Disabled         Split Horizon        : Disabled
Damp Peer Oscillatio*: Disabled         Update Errors        : 0
GR Notification      : Disabled         Fault Tolerance      : Disabled
Rem Idle Hold Time   : 00h00m00s
Next-Hop Unchanged   : None
Local Capability     : RtRefresh MPBGP 4byte ASN
Remote Capability    : RtRefresh MPBGP 4byte ASN
```

```
Routes Resolve To St*: Disabled
Local AddPath Capabi*: Disabled
Remote AddPath Capab*: Send - None
                     : Receive - None
Import Policy        : None Specified / Inherited
Export Policy        : None Specified / Inherited
Origin Validation    : N/A
EBGP Link Bandwidth  : n/a
IPv4 Rej. Pfxs       : 0                IPv6 Rej. Pfxs       : 0
VPN-IPv4 Rej. Pfxs   : 0                VPN-IPv6 Rej. Pfxs   : 0
Mc IPv4 Rej. Pfxs    : 0                Mc IPv6 Rej. Pfxs    : 0
MVPN-IPv4 Rej. Pfxs  : 0                MVPN-IPv6 Rej. Pfxs  : 0
Flow-IPv4 Rej. Pfxs  : 0                Flow-IPv6 Rej. Pfxs  : 0
L2-VPN Rej. Pfxs     : 0                MDT-SAFI Rej. Pfxs   : 0
Rte-Tgt Rej. Pfxs    : 0                MS-PW Rej. Pfxs      : 0
Mc Vpn Ipv4 Rej. Pfxs: 0                Evpn Rej. Pfxs       : 0
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

The following output shows that 16 BGP routes are received by PE-1.

```
*A:PE-1# show router 2 bgp routes
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.101.0.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.1.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.2.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.3.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.4.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.5.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
```

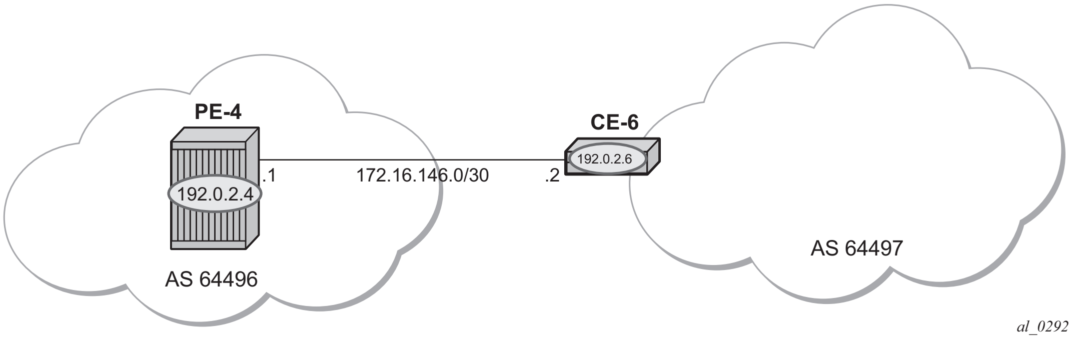
```
u*>i  10.101.6.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.101.7.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.0.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.1.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.2.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.3.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.4.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.5.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.6.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
u*>i  10.102.7.0/24                                      None        None
      172.16.117.2                                       None        -
      64498
-------------------------------------------------------------------------------
Routes : 16
===============================================================================
*A:PE-1#
```

PE-4 also has a VPRN 2 instance configured, so that it will receive the imported BGP routes. The configuration for PE-4 is:

```
*A:PE-4# configure
    service
        vprn 2 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:2
            interface "int-PE-4-CE-6_2nd" create
                address 172.16.146.1/30
                sap 1/2/1:2 create
                exit
            exit
            bgp
                group "VPRN2-external"
                    peer-as 64497
                    neighbor 172.16.146.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
```

Figure 81 shows the connectivity between PE-4 and CE-6. PE-4 will only forward a summarizing aggregate route toward CE-6.

Figure 81 **CE-6 Connectivity**



PE-4 receives labeled BGP route prefixes from PE-1 via the route reflector and installs them in the FIB for router instance 2:

```
*A:PE-4# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.101.0.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.1.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.2.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.3.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.4.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.5.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.6.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.101.7.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.102.0.0/24                                 Remote  BGP VPN   00h01m07s   170
       192.0.2.1 (tunneled)                                         0
10.102.1.0/24                                 Remote  BGP VPN   00h01m07s   170
```

```
       192.0.2.1 (tunneled)                                          0
10.102.2.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
10.102.3.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
10.102.4.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
10.102.5.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
10.102.6.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
10.102.7.0/24                                 Remote   BGP VPN   00h01m07s  170
       192.0.2.1 (tunneled)                                          0
172.16.117.0/30                               Remote   BGP VPN   00h02m41s  170
       192.0.2.1 (tunneled)                                          0
172.16.146.0/30                               Local    Local     00h02m42s  0
       int-PE-4-CE-6_2nd                                             0
-------------------------------------------------------------------------------
No. of Routes: 18
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-4#
```

The CE-6 configuration for an interface toward PE-4 is as follows:

```
*A:CE-6# configure
    service
        ies 2 customer 1 create
            interface "int-CE-6-PE-4_2nd" create
                address 172.16.146.2/30
                sap 1/1/1:2 create
                exit
            exit
            no shutdown
```

The BGP configuration of CE-6:

```
*A:CE-6# configure
    router
        bgp
            group "external-toVPRN2onPE-4"
                peer-as 64496
                neighbor 172.16.146.1
                exit
            exit
            no shutdown
```

The BGP neighbor state for PE-4:

```
*A:PE-4# show router 2 bgp neighbor 172.16.146.2

===============================================================================
BGP Neighbor
```

```
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 172.16.146.2
Description          : (Not Specified)
Group                : VPRN2-external
-------------------------------------------------------------------------------
Peer AS              : 64497            Peer Port            : 51154
Peer Address         : 172.16.146.2
Local AS             : 64496            Local Port           : 179
Local Address        : 172.16.146.1
Peer Type            : External         Dynamic Peer         : No
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : IPv4
Remote Family        : IPv4
Hold Time            : 90               Keep Alive           : 30
Min Hold Time        : 0
Active Hold Time     : 90               Active Keep Alive    : 30
Cluster Id           : None
Preference           : 170              Num of Update Flaps  : 0
Recd. Paths          : 0
IPv4 Recd. Prefixes  : 0                IPv4 Active Prefixes : 0
IPv4 Suppressed Pfxs : 0                VPN-IPv4 Suppr. Pfxs : 0
VPN-IPv4 Recd. Pfxs  : 0                VPN-IPv4 Active Pfxs : 0
Mc IPv4 Recd. Pfxs.  : 0                Mc IPv4 Active Pfxs. : 0
Mc IPv4 Suppr. Pfxs  : 0                IPv6 Suppressed Pfxs : 0
IPv6 Recd. Prefixes  : 0                IPv6 Active Prefixes : 0
VPN-IPv6 Recd. Pfxs  : 0                VPN-IPv6 Active Pfxs : 0
VPN-IPv6 Suppr. Pfxs : 0
Mc IPv6 Recd. Pfxs.  : 0                Mc IPv6 Active Pfxs. : 0
Mc IPv6 Suppr. Pfxs  : 0                L2-VPN Suppr. Pfxs   : 0
L2-VPN Recd. Pfxs    : 0                L2-VPN Active Pfxs   : 0
MVPN-IPv4 Suppr. Pfxs: 0                MVPN-IPv4 Recd. Pfxs : 0
MVPN-IPv4 Active Pfxs: 0                MDT-SAFI Suppr. Pfxs : 0
MDT-SAFI Recd. Pfxs  : 0                MDT-SAFI Active Pfxs : 0
Flow-IPv4 Suppr. Pfxs: 0                Flow-IPv4 Recd. Pfxs : 0
Flow-IPv4 Active Pfxs: 0                Rte-Tgt Suppr. Pfxs  : 0
Rte-Tgt Recd. Pfxs   : 0                Rte-Tgt Active Pfxs  : 0
Backup IPv4 Pfxs     : 0                Backup IPv6 Pfxs     : 0
Mc Vpn Ipv4 Recd. Pf*: 0                Mc Vpn Ipv4 Active P*: 0
Mc Vpn Ipv4 Suppr. P*: 0
Backup Vpn IPv4 Pfxs : 0                Backup Vpn IPv6 Pfxs : 0
Input Queue          : 0                Output Queue         : 0
i/p Messages         : 160              o/p Messages         : 36
i/p Octets           : 577              o/p Octets           : 558
i/p Updates          : 0                o/p Updates          : 0
MVPN-IPv6 Suppr. Pfxs: 0                MVPN-IPv6 Recd. Pfxs : 0
MVPN-IPv6 Active Pfxs: 0
Flow-IPv6 Suppr. Pfxs: 0                Flow-IPv6 Recd. Pfxs : 0
Flow-IPv6 Active Pfxs: 0
Evpn Suppr. Pfxs     : 0                Evpn Recd. Pfxs      : 0
Evpn Active Pfxs     : 0
MS-PW Suppr. Pfxs    : 0                MS-PW Recd. Pfxs     : 0
MS-PW Active Pfxs    : 0
TTL Security         : Disabled         Min TTL Value        : n/a
Graceful Restart     : Disabled         Stale Routes Time    : n/a
Restart Time         : n/a
Advertise Inactive   : Disabled         Peer Tracking        : Disabled
Advertise Label      : None
Auth key chain       : n/a
Disable Cap Nego     : Disabled         Bfd Enabled          : Disabled
Flowspec Validate    : Disabled         Default Route Tgt    : Disabled
Aigp Metric          : Disabled         Split Horizon        : Disabled
Damp Peer Oscillatio*: Disabled         Update Errors        : 0
GR Notification      : Disabled         Fault Tolerance      : Disabled
Rem Idle Hold Time   : 00h00m00s
Next-Hop Unchanged   : None
Local Capability     : RtRefresh MPBGP 4byte ASN
Remote Capability    : RtRefresh MPBGP 4byte ASN
Routes Resolve To St*: Disabled
Local AddPath Capabi*: Disabled
Remote AddPath Capab*: Send - None
                     : Receive - None
Import Policy        : None Specified / Inherited
Export Policy        : None Specified / Inherited
Origin Validation    : N/A
EBGP Link Bandwidth  : n/a
IPv4 Rej. Pfxs       : 0                IPv6 Rej. Pfxs       : 0
VPN-IPv4 Rej. Pfxs   : 0                VPN-IPv6 Rej. Pfxs   : 0
Mc IPv4 Rej. Pfxs    : 0                Mc IPv6 Rej. Pfxs    : 0
MVPN-IPv4 Rej. Pfxs  : 0                MVPN-IPv6 Rej. Pfxs  : 0
Flow-IPv4 Rej. Pfxs  : 0                Flow-IPv6 Rej. Pfxs  : 0
L2-VPN Rej. Pfxs     : 0                MDT-SAFI Rej. Pfxs   : 0
Rte-Tgt Rej. Pfxs    : 0                MS-PW Rej. Pfxs      : 0
Mc Vpn Ipv4 Rej. Pfxs: 0                Evpn Rej. Pfxs       : 0
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-4#
```

In order to advertise a summarizing aggregate route with an associated community string, an aggregate route is required. In this case, the 10.101.x.0/24 group of prefixes will be associated with community 64496:101. The 10.102.x.0/24 group of prefixes will be associated with the standard community n**o-export**, so that it will not be advertised to any external peer.

The configuration required on PE-4 is as follows:

```
configure
    service
        vprn 2
            aggregate 10.101.0.0/21 community 64496:101
            aggregate 10.102.0.0/21 community no-export
        exit
```

An export policy is required on PE-4 to allow the advertising of the aggregate route. No community is applied using this policy.

```
configure
    router
        policy-options
            begin
            policy-statement "PE-4-VPN-Agg"
                entry 10
                    from
                        protocol aggregate
                    exit
                    action accept
                    exit
                exit
            commit
            exit
```

This is applied as an export policy within the group context of the BGP configuration of the VPRN.

```
configure
    service
        vprn 2
            bgp
                group "VPRN2-external"
                    export "PE-4-VPN-Agg"
                exit
```

The aggregate route 10.101.0.0/21 is received at CE-6 via BGP. The community that was associated with this prefix is seen: 64496:101. The route is seen as an aggregate, with PE-4 as the aggregating router (192.0.2.4). The “Atomic Aggregate” attribute is present, meaning that PE-4 has not advertised any details of the AS Paths of the composite routes.

```
*A:CE-6# show router bgp routes 10.101.0.0/21 hunt
===============================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 10.101.0.0/21
Nexthop        : 172.16.146.1
Path Id        : None
From           : 172.16.146.1
Res. Nexthop   : 172.16.146.1
Local Pref.    : None                   Interface Name : int-CE-6-PE-4_2nd
Aggregator AS  : 64496                  Aggregator     : 192.0.2.4
Atomic Aggr.   : Atomic                 MED            : None
AIGP Metric    : None
Connector      : None
Community      : 64496:101
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.4
```

```
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : External
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h02m07s
---snip---
```

The aggregate route 10.102.0.0/21 is not received at CE-6, as PE-4 does not advertise it, due to the fact that it is associated with the “no-export” community.

```
*A:CE-6# show router bgp routes 10.102.0.0/21 hunt
===============================================================================
 BGP Router ID:192.0.2.6        AS:64497       Local AS:64497
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
No Matching Entries Found
===============================================================================
*A:CE-6#
```

Conclusion

Community strings can be added to static and aggregate routes. This example shows the configuration of communities with both static and aggregate routes, together with the associated show outputs which can be used to verify and troubleshoot them.

BGP Add-Path

This chapter provides information about BGP Add-Path.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS Release 14.0.R7.

Overview

When a BGP router learns multiple paths for the same prefix, it selects one route as its best path and advertises only this route to its BGP peers. The BGP add-path feature allows advertising the best n paths for the same prefix, where n is configurable. If the set of n paths includes multiple paths with the same BGP next hop, only the best route with a specific next hop is advertised and the other paths are suppressed.

The BGP add-path feature increases path visibility in the autonomous system (AS), because more routes are stored in the Routing Information Base (RIB). BGP add-path has the following benefits:

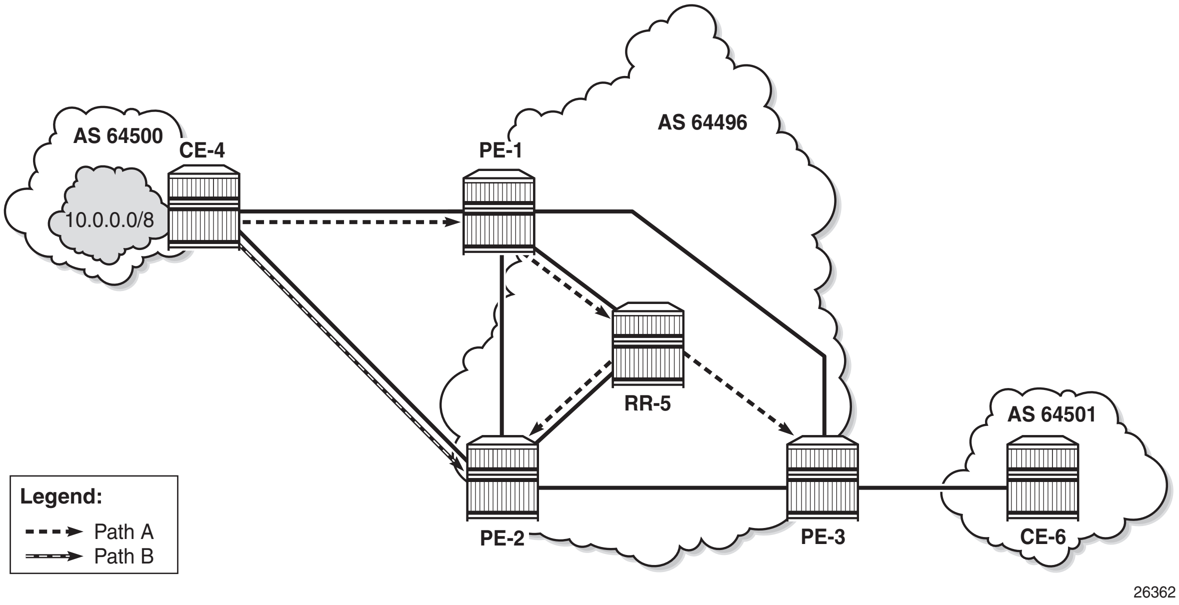
- Faster convergence after failure
- Enhanced load-sharing
- Reduced routing churn

These benefits are described in the following sections.

Faster Convergence after Failure

Figure 82 shows a network that does not support add-path. CE-4 advertises two paths for prefix 10.0.0.0/8 to its eBGP neighbors: PE-1 and PE-2. PE-1 has an import policy that sets the local preference (LP) of path A to 200; PE-2 keeps the default LP of 100 for path B. Therefore, path A that is advertised to PE-1 is preferred in AS 64496. The route reflector RR-5 advertises the preferred path A to PE-2 and PE-3. PE-2 suppresses the advertisement of its external path (B) to RR-5, because path A is preferred. Traffic from CE-6 to CE-4 is sent via PE-3 and PE-1.

Figure 82 **RR Advertises Best Path Only – Path A Preferred over Path B**



When the link between CE-4 and PE-1 fails, the following steps take place for reconvergence:

Step 1. PE-1 sends a BGP update withdrawing path A to RR-5.

Step 2. RR-5 receives and propagates the withdrawal to its other clients: PE-2 and PE-3.

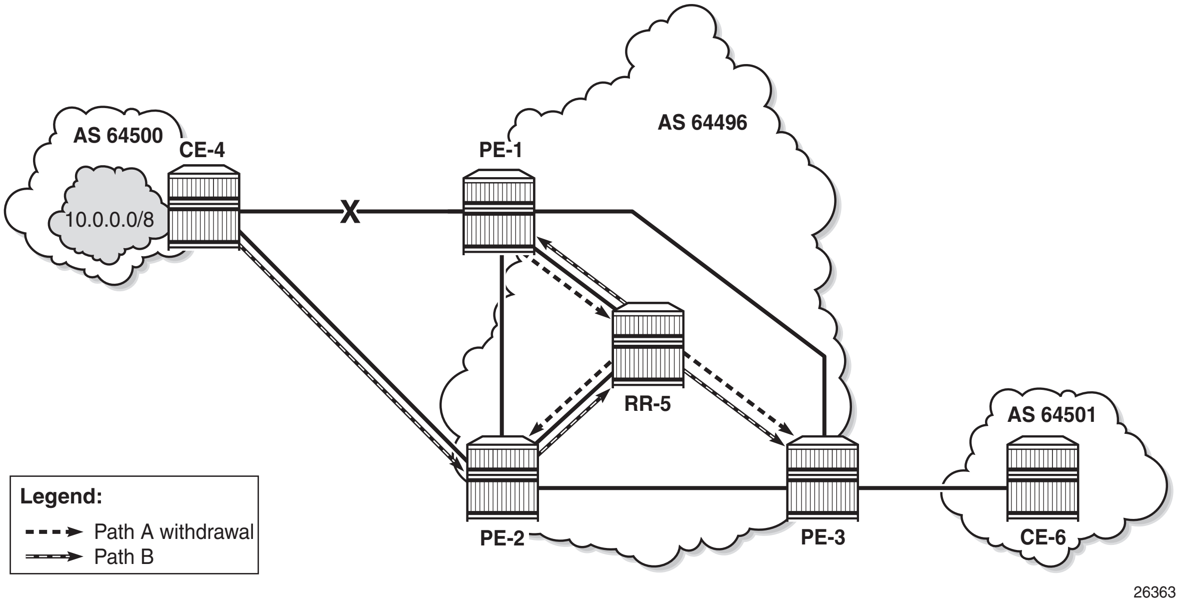
Step 3. PE-2 receives the withdrawal of path A and reruns the BGP decision process. PE-2 selects path B as its best route and advertises path B to RR-5.

Step 4. RR-5 receives the BGP update for path B and reruns its BGP decision process. RR-5 selects path B as its best path and advertises path B to its other clients: PE-1 and PE-3.

Step 5. PE-1 and PE-3 rerun their BGP decision process and determine that path B is the best path. Traffic can flow from CE-6 to CE-4 via PE-3 and PE-2.

Figure 83 shows the BGP updates sent to withdraw path A and advertise path B.

Figure 83 **Reconvergence after Path Failure (without Add-path)**



If the propagation time of a BGP update message between RR-5 and any of its clients is X, the convergence time is four times X, plus processing, transmission, and queuing delays.

With the use of add-path on all BGP routers in AS 64496, the convergence time can be reduced considerably, because PE-3 has more than one path for prefix 10.0.0.0/8 in its RIB-IN before the failure takes place. When there are no failures, PE-2 decides that path A is best, and PE-2 also advertises its second-best path (B)—which is its best external path—to RR-5. With add-path enabled, the RR has knowledge of two paths for prefix 10.0.0.0/8 and advertises both to its clients. PE-3 receives two routes for prefix 10.0.0.0/8, reruns the BGP decision process, and updates its forwarding table based on the results. The following options are possible:

- Path A is the best path, whereas path B is maintained in the RIB-IN. The FIB entry for destination 10.0.0.0/8 points at path {**A**} only.

- When BGP FRR is enabled as described in chapter BGP Fast Reroute, path A is the best path and path B is the second-best path. The FIB entry for destination 10.0.0.0/8 points to path {**A**,B}. If path A is available, it is used for all traffic to the destination; if path A is unavailable but path B is available, then all traffic to the destination is directed to path B. In this case, path B is effectively a pre-computed, pre-installed backup path for the destination.
- When Equal Cost Multi-Path (ECMP) and BGP multipath are enabled and the paths have an equal cost, both paths A and B represent the best path. The FIB entry for destination 10.0.0.0/8 points to multipath entry {**A,B**}. When both paths are available, traffic to the destination is load-shared across paths A and B. If only one path is available, traffic is directed to that available path.

Figure 84 shows the BGP update messages prior to any failures. RR-5 receives path A from PE-1 and path B from PE-2, whereas it advertises path B to PE-1, path A to PE-2, and both path A and path B to PE-3. Path B has the default LP 100, whereas path A gets LP 200 as per import policy on PE-1. However, in case of ECMP, both paths keep the default LP 100.

Figure 84 **Advertised Paths when BGP Add-path is Enabled in PEs and RR**

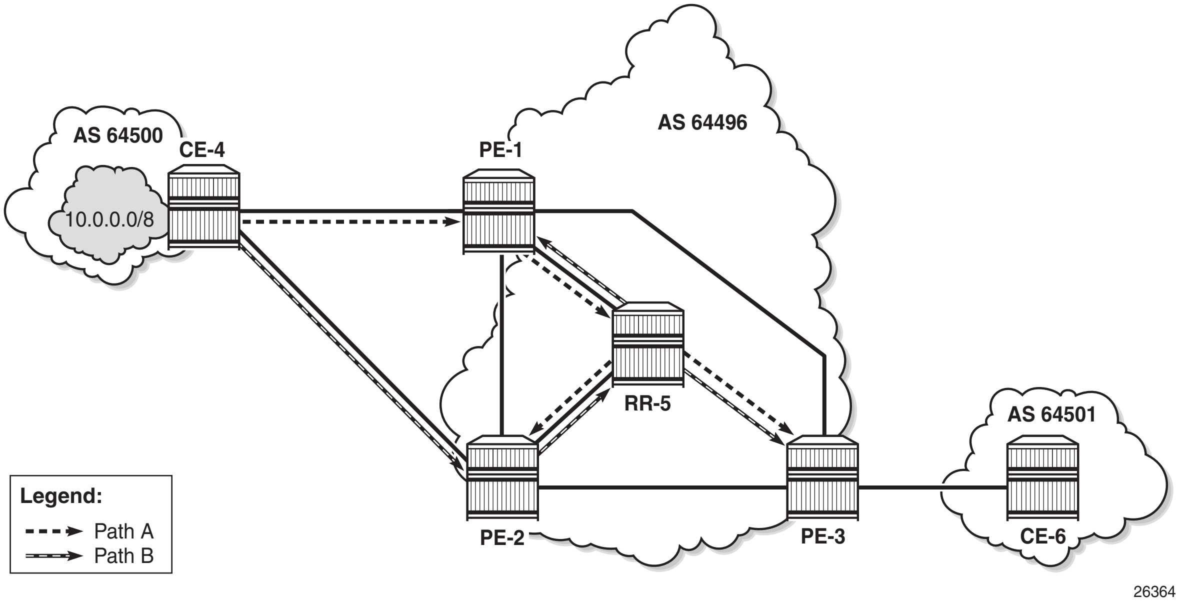


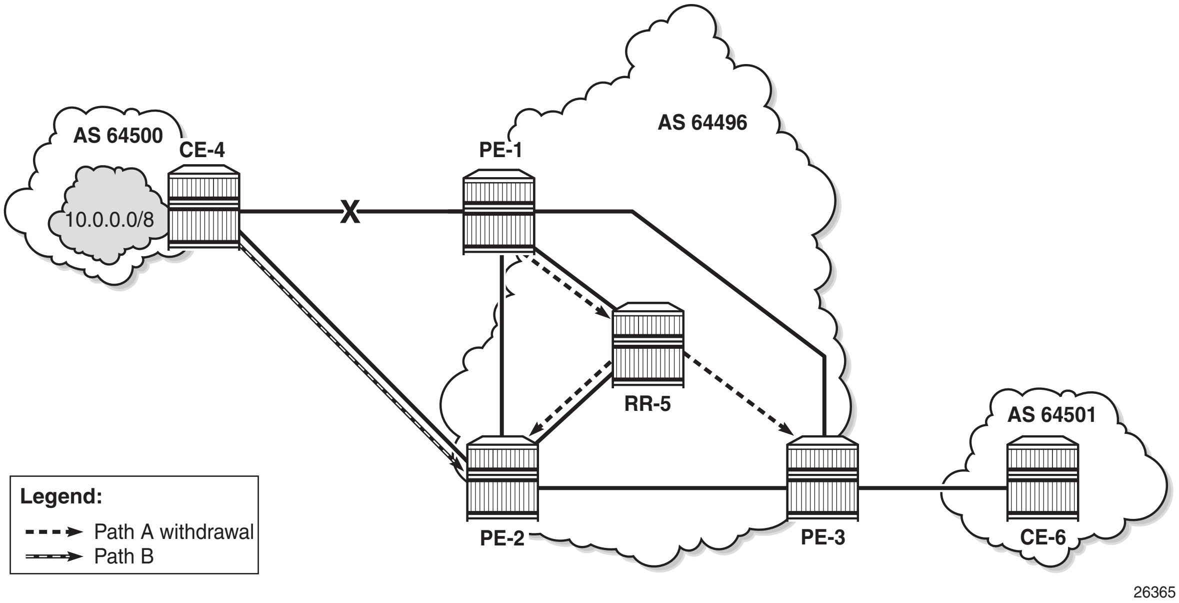
Figure 85 shows the BGP update messages that are sent after a link failure between CE-4 and PE-1. With add-path, fewer steps are required for convergence:

Step 1. PE-1 sends a BGP update message withdrawing path A.

Step 2. RR-5 receives the withdrawal and propagates it to its clients PE-2 and PE-3.

Step 3. PE-2 and PE-3 receive the withdrawal, rerun the BGP decision process, and update the forwarding entry for destination 10.0.0.0/8: path B is best.

Figure 85 **Reconvergence after Path Failure when BGP Add-path is Enabled**



The convergence time with add-path is much shorter than without add-path. If X is the propagation time of a BGP update message between RR and any of the PEs, then the convergence time is the time required for the BGP update from PE-1 to RR-5 (X) plus the time required for the BGP update propagation from RR-5 to the other PEs (X), in addition to delays for processing, transmission, and queuing. The convergence with add-path is twice as fast as without add-path.

For some types of failures, the convergence can be even faster:

- When PE-1 becomes unreachable, the next-hop tracking by PE-3 will invalidate path A before the BGP withdrawal message is received from RR-5.
- If PE-3 implements BGP FRR and path A has been marked as unusable, PE-3 can switch traffic destined to 10.0.0.0/8 to path B.
- When Bidirectional Forwarding Detection (BFD) is enabled on the eBGP sessions and on the IGP protocol, the failure is detected faster and BGP convergence can be sped up when BGP FRR is enabled.

Enhanced Load-Sharing

When paths A and B are equal in cost or preference, and ECMP and BGP multipath are enabled on all PEs, load-sharing can be done for traffic with destination 10.0.0.0/8. With BGP add-path, both paths A and B are advertised to the PEs. PE-3 runs the BGP decision process and determines that paths A and B are both best paths to destination 10.0.0.0/8, so paths A and B are combined into one multipath forwarding entry: {A,B}.

The benefits of load-sharing for traffic to destination 10.0.0.0/8 are the following:

- More even bandwidth utilization of the links in AS 64496
- More even bandwidth utilization for traffic across peering points PE-1 and PE-2 with AS 64500
- Faster reaction to some failures; for example, the BGP next hop for one of the paths becomes unreachable in the IGP and next hop tracking is enabled.

Reduced Routing Churn

Routing churn refers to repeated advertisements and withdrawals of a prefix and path. Some degree of routing churn is normal and expected in most networks. However, it should be contained as much as possible to avoid overloading router CPUs. Routing churn can be caused by:

- Flapping links (links that repeatedly transition between up and down state)
- Route oscillation (networks that use RRs or AS confederations and BGP path selection relies on Multi Exit Discriminator (MED) and IGP cost comparisons)

Add-path helps to reduce routing churn by constraining the effect of some failures to the local AS where they occur. For example, the link between CE-4 and PE-1 could repeatedly cycle up and down due to a misconfiguration. When the link goes down, a BGP withdrawal message is sent by PE-1 to RR-5 and from RR-5 to the other RR clients (PE-2 and PE-3). PE-3 will withdraw and advertise path A to its eBGP peer CE-6 in AS 64501, but path B is constantly advertised to CE-6 (when add-path has been negotiated between PE-3 and CE-6).

Without add-path, PE-2 would be affected by the instability in AS 64496 and there would be periods of time when AS 64501 has no paths to destination 10.0.0.0/8 (between the withdrawal of path A and the advertisement of path B).

Add-path Implementation

BGP add-path is configured in the base routing instance, for iBGP or eBGP, per address family at different levels: in the global BGP context, per group, and per neighbor. The following address families are supported: IPv4, IPv6, label-IPv4, label-IPv6, VPN-IPv4, and VPN-IPv6, as follows:

```
*A:PE-1# configure router bgp add-paths
  - add-paths
  - no add-paths

 [no] ipv4            - Configure ipv4 ADD-PATH limits
 [no] ipv6            - Configure ipv6 ADD-PATH limits
 [no] label-ipv4      - Configure label-ipv4 ADD-PATH limits
 [no] label-ipv6      - Configure label-ipv6 ADD-PATH limits
 [no] vpn-ipv4        - Configure vpn-ipv4 ADD-PATH limits
 [no] vpn-ipv6        - Configure vpn-ipv6 ADD-PATH limits
```

Up to 16 paths are configurable per address family per peer (send-limit):

```
*A:PE-1# configure router bgp add-paths ipv4
  - ipv4 send <send-limit>
  - ipv4 send <send-limit> receive [none]
  - no ipv4

 <send-limit>         : [1-16] | none
```

Only the number of advertised routes per prefix is controlled, not the number of received routes. All routes advertised by an add-path peer are accepted; otherwise, routing loops might occur. If a BGP speaker is configured with <send-limit> *n*, but has more than *n* paths available in the LOC-RIB, it selects the *n* best paths with unique BGP next hops following the Add-*n* path selection algorithm described in *draft-ietf-idr-add-paths-guidelines*. Also, the send limit *n* can be overridden, for specific prefixes, using route policies.

When BGP add-path is configured for an address family, the BGP capability will be announced to the BGP peer as part of the BGP open message, as follows:

```
*A:PE-1# debug router bgp open

57 2017/02/21 09:57:36.11 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.5 - Send (Active) BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.1: Opt Length 22
   Opt Para: Type CAPABILITY: Length = 20: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x1
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf0
     Cap_Code ADD-PATH: Length 4
       Bytes: 0x0 0x1 0x1 0x3
```

```
"

58 2017/02/21 09:57:36.11 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.5 - Received BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.5: Opt Length 22
   Opt Para: Type CAPABILITY: Length = 20: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x1
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf0
     Cap_Code ADD-PATH: Length 4
       Bytes: 0x0 0x1 0x1 0x3
"
```

The BGP add-path capability code value typically consists of one or more blocks of four bytes; two octets for the Address Family Identifier (AFI), one octet for the Subsequent Address Family Identifier (SAFI), and one octet for send/receive. In this example, AFI/SAFI bytes point to an IPv4 address family and send/receive value "3" means that the sender is able to receive and send multiple paths from/to its BGP peer.

In BGP update messages, a 4-octet path identifier (ID) is added to the Network Layer Reachability Information (NLRI) field. The combination of both prefix and path ID identifies a BGP path. SR OS allocates path IDs sequentially on a per address family basis, not per prefix. The path ID is only locally significant, which means that when a BGP speaker re-advertises a route with path IDs, it must generate its own path ID. PE-1 sends the following BGP update for prefix 10.0.0.0/8 with LP 200 to RR-5 with path ID 8.

```
*A:PE-1# debug router bgp update
62 2017/02/21 09:57:37.13 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.5
"Peer 1: 192.0.2.5: UPDATE
Peer 1: 192.0.2.5 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 27
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.1
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 200
    NLRI: Length = 6
        10.0.0.0/8 Path-ID 8
"
```

The BGP route for prefix 10.0.0.0/8 with LP 200 has path ID 8 on RR-5, as follows:

```
*A:RR-5# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.5        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      192.0.2.1                                          8           -
      64500
*i    10.0.0.0/8                                         100         None
      192.0.2.2                                          3           -
      64500
-------------------------------------------------------------------------------
Routes : 2
```

When routers have negotiated to advertise (and receive) routes with path identifiers, all BGP updates (advertisements or withdrawals) without path identifier will be rejected. There will be an NLRI parsing error—because the BGP update has an incorrect length—and a notification will be sent.

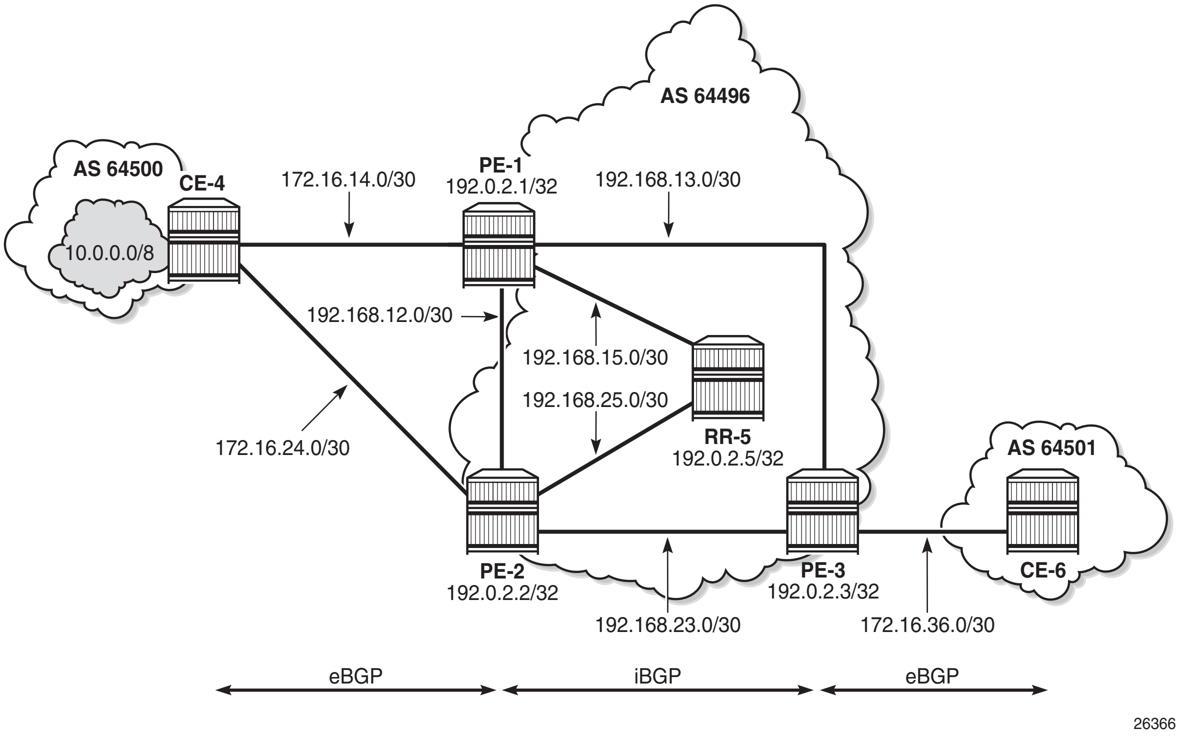
Configuration

The following configuration examples are in this section:

- BGP without add-path
- BGP with add-path for address family IPv4: no BGP FRR, no ECMP
- BGP with add-path for address family IPv4 and BGP FRR enabled
- BGP with add-path for address family IPv4 and ECMP enabled
- BGP with add-path for address family VPN-IPv4 and BGP FRR enabled
- BGP with add-path for address family VPN-IPv4 and ECMP enabled

Figure 86 shows the example topology with CE-4 in AS 64500 advertising route 10.0.0.0/8 to its eBGP peers PE-1 and PE-2 in AS 64496. PE-1 has an import policy that sets the LP for this route to 200, whereas PE-2 keeps the default local preference of 100. RR-5 is RR for all PEs in AS 64496. CE-6 in AS 64501 peers with PE-3 in AS 64496 and can send traffic to CE-4 in AS 64500.

Figure 86 **Example Topology**



Initial Configuration

The initial configuration on all nodes includes:

- Cards, MDAs, ports
  - – Ports between CEs and PEs are hybrid ports with dot1q encapsulation.
  - – Ports between PEs and RR in AS 64496 are network ports (null encapsulation).
- Router interfaces
- IS-IS as IGP on all interfaces within AS 64496 (alternatively, OSPF can be used)
- LDP on all interfaces between the PEs in AS 64496, but not toward RR-5

BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports prefix 10.0.0.0/8 to both eBGP peers, as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "eBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 172.16.14.1
                exit
                neighbor 172.16.24.1
                exit
            exit
        exit
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The BGP configuration on CE-6 is similar, except for the export policy.

PE-1 peers with CE-4 in AS 64500 and RR-5 in AS 64496. An import policy is configured to set the LP to 200 for all routes received from CE-4, as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "eBGP"
                import "import-bgp-LP200"
                peer-as 64500
                neighbor 172.16.14.2
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 64496
                neighbor 192.0.2.5
                exit
            exit
        policy-options
            begin
            policy-statement "import-bgp-LP200"
                default-action accept
                    local-preference 200
                exit
            exit
            commit
```

The BGP configuration on PE-2 and PE-3 is similar, but there is no import policy.

The BGP configuration on RR-5 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "iBGP"
                cluster 5.5.5.5
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
```

PE-1 advertises a route for prefix 10.0.0.0/8 with LP 200 to RR-5. RR-5 propagates this route to its other clients: PE-2 and PE-3. When PE-2 learns this route, it does not advertise its own route for 10.0.0.0/8 with LP 100 to RR-5 anymore. PE-3 only learns the route for prefix 10.0.0.0/8 with LP 200, as follows:

```
*A:PE-3# show router bgp routes
================================================================================
 BGP Router ID:192.0.2.3       AS:64496        Local AS:64496
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

================================================================================
BGP IPv4 Routes
================================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
--------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      192.0.2.1                                          None        -
      64500
--------------------------------------------------------------------------------
Routes : 1
```

Reconvergence without Add-path

A failure of the link between CE-4 and PE-1 is simulated as follows:

```
*A:CE-4# configure router interface "int-CE-4-PE-1" shutdown
```

The following four BGP update messages are received or sent by RR-5.

RR-5 receives the following withdrawal message from PE-1:

```
14 2017/02/21 12:26:44.56 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 2
        10.0.0.0/8
    Total Path Attr Length = 0
"
```

RR-5 propagates this withdrawal to its other clients, for example to PE-2, as follows:

```
15 2017/02/21 12:26:44.55 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 2
        10.0.0.0/8
    Total Path Attr Length = 0
"
```

When PE-2 receives this withdrawal, it reruns the BGP decision process and decides that its route for prefix 10.0.0.0/8 with LP 100 is the best route. PE-2 advertises this route to RR-5; it is received by RR-5 as follows:

```
17 2017/02/21 12:26:44.55 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 27
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.2
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    NLRI: Length = 2
        10.0.0.0/8
"
```

RR-5 propagates this message to its other clients: PE-1 and PE-3. The following BGP update is sent to PE-3:

```
19 2017/02/21 12:26:45.21 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 41
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.2
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.2
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        5.5.5.5
    NLRI: Length = 2
        10.0.0.0/8
"
```

Again, PE-3 has only one route for prefix 10.0.0.0/8, but this time with next hop 192.0.2.2, as follows:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         100         None
      192.0.2.2                                          None        -
      64500
-------------------------------------------------------------------------------
Routes : 1
```

The configuration is restored as follows:

```
*A:CE-4# configure router interface "int-CE-4-PE-1" no shutdown
```

Add-path Enabled: No BGP FRR, No ECMP

Before add-path is enabled, the following information is displayed on PE-1 for BGP neighbor RR-5:

```
*A:PE-1# show router bgp neighbor 192.0.2.5 | match "Local AddPath" post-lines 2
Local AddPath Capabi*: Disabled
Remote AddPath Capab*: Send - None
                     : Receive - None
```

Add-path is enabled on PE-1 and PE-2 with a send path limit of two for groups "eBGP" and "iBGP" and no limit on the receive path limit, which is the default setting, as follows:

```
configure router bgp group "eBGP" add-paths ipv4 send 2
configure router bgp group "iBGP" add-paths ipv4 send 2
```

When the preceding show command is repeated on PE-1 or PE-2, the local BGP add-path capabilities are specified for address family IPv4: a maximum of two paths can be sent for a specific IPv4 prefix. The remote peer RR-5 does not have add-path enabled yet.

```
*A:PE-1# show router bgp neighbor 192.0.2.5 | match "Local AddPath" post-lines 3
Local AddPath Capabi*: Send - IPv4 (2)
                     : Receive - IPv4
Remote AddPath Capab*: Send - None
                     : Receive - None
```

Initially, add-path remains disabled on PE-3. On the RR, add-path is enabled for neighbors 192.0.2.1 and 192.0.2.2, but not for 192.0.2.3 yet. For neighbor 192.0.2.1, the **receive none** option implies that the add-path receive capability is not negotiated.

```
*A:RR-5# configure router bgp group "iBGP" neighbor 192.0.2.1 add-paths ipv4 send 2
receive none
*A:RR-5# configure router bgp group "iBGP" neighbor 192.0.2.2 add-paths ipv4 send 2
```

The following output shows that add-path is enabled locally on RR-5 and remotely on PE-1 for address family IPv4. RR-5 can send a maximum of two paths for a specific prefix toward PE-1 and PE-2; toward PE-3, add-path remains disabled.

```
*A:RR-5# show router bgp neighbor 192.0.2.1 | match "Local AddPath" post-lines 3
Local AddPath Capabi*: Send - IPv4 (2)
                     : Receive - None
Remote AddPath Capab*: Send - IPv4
                     : Receive - IPv4
*A:RR-5# show router bgp neighbor 192.0.2.2 | match "Local AddPath" post-lines 3
Local AddPath Capabi*: Send - IPv4 (2)
                     : Receive - IPv4
Remote AddPath Capab*: Send - IPv4
                     : Receive - IPv4
*A:RR-5# show router bgp neighbor 192.0.2.3 | match "Local AddPath" post-lines 2
Local AddPath Capabi*: Disabled
Remote AddPath Capab*: Send - None
                     : Receive - None
```

The **receive none** option indicates that RR-5 does not negotiate the add-path receive capability with its peer. PE-1 knows that peer 192.0.2.5 may send IPv4 routes with a path ID, but has no information about what this peer will receive:

```
*A:PE-1# show router bgp neighbor 192.0.2.5 | match "Local AddPath" post-lines 3
Local AddPath Capabi*: Send - IPv4 (2)
```

```
                     : Receive - IPv4
Remote AddPath Capab*: Send - IPv4
                     : Receive - None
```

With BGP add-path enabled, PE-2 will advertise its second-best route for prefix 10.0.0.0/8 with LP 100 to RR-5. PE-1, PE-2, and RR-5 will have two routes for prefix 10.0.0.0/8 in their RIB-IN, but only the route with LP 200 will be used. The following output shows the BGP routes on RR-5, but it resembles the output on PE-1 and PE-2:

```
*A:RR-5# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.5        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      192.0.2.1                                          2           -
      64500
*i    10.0.0.0/8                                         100         None
      192.0.2.2                                          2           -
      64500
-------------------------------------------------------------------------------
Routes : 2
```

Even though RR-5 has two routes for this prefix, it only advertises its best route to PE-3, because add-path is not enabled for this BGP session. Therefore, PE-3 only has the route for 10.0.0.0/8 with LP 200, as follows:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
```

```
      192.0.2.1                                          None        -
      64500
-------------------------------------------------------------------------------
Routes : 1
```

When add-path is enabled on the session between PE-3 and RR-5, the second route will also be advertised, as follows:

```
*A:PE-3# configure router bgp group "iBGP" add-paths ipv4 send 2
```

```
*A:RR-5# configure router bgp group "iBGP" neighbor 192.0.2.3 add-paths ipv4 send 2
```

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      192.0.2.1                                          5           -
      64500
*i    10.0.0.0/8                                         100         None
      192.0.2.2                                          6           -
      64500
-------------------------------------------------------------------------------
Routes : 2
```

BGP add-path is enabled, but BGP FRR or ECMP are disabled. The routing table on PE-3 only contains one entry for prefix 10.0.0.0/8:

```
*A:PE-3# show router route-table 10.0.0.0/8

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP       00h05m45s  170
       192.168.13.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
```

Reconverge with Add-Path: No BGP FRR, No ECMP

A link failure between CE-4 and PE-1 is simulated as follows:

```
*A:CE-4# configure router interface "int-CE-4-PE-1" shutdown
```

PE-1 sends a withdrawal message for route 10.0.0.0/8 with LP 200 to RR-5 and reruns the BGP decision process. RR-5 propagates this withdrawal message to its other clients that rerun the BGP decision process. As a result, the route for prefix 10.0.0.0/8 with LP 100 will be used on all nodes; for example, on PE-3:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         100         None
      192.0.2.2                                          6           -
      64500
-------------------------------------------------------------------------------
Routes : 1
```

The routing table contains a route to 10.0.0.0/8 with PE-2 as next hop, as follows:

```
*A:PE-3# show router route-table 10.0.0.0/8
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP       00h05m52s  170
       192.168.23.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The convergence with add-path enabled is twice as fast as without BGP add-path. With BGP add-path disabled, four sequential messages are sent:

1. PE-1 sends a withdrawal to RR-5.
2. RR-5 propagates withdrawal.
3. PE-2 advertises its route.
4. RR-5 propagates the route.

In the scenario with add-path, the last two messages are already sent before the failure happened. During convergence, only two withdrawal messages are sent: PE-1 sends a withdrawal to RR-5; RR-5 propagates this to its clients.

Add-path and BGP FRR

The convergence time can be further reduced by enabling BGP FRR, where the BGP decision process runs for the best route and the backup path before any failure happens, as described in chapter BGP Fast Reroute. BGP FRR is enabled on all PEs, as follows:

```
configure router bgp backup-path
```

Each PE has two routes for prefix 10.0.0.0/8 and when BGP FRR is enabled, both are used, but one is used as backup, indicated by the "b"-flag in the following output:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         200         None
      192.0.2.1                                          7           -
      64500
ub*i  10.0.0.0/8                                         100         None
      192.0.2.2                                          6           -
      64500
-------------------------------------------------------------------------------
Routes : 2
```

The following routing table on PE-3 shows the active route for 10.0.0.0/8 and adds an indication "B", indicating that a backup route is available:

```
*A:PE-3# show router route-table 10.0.0.0/8

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.0.0/8 [B]                                Remote  BGP       00h00m42s  170
       192.168.13.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The following output shows both the active and the backup route for prefix 10.0.0.0/8:

```
*A:PE-3# show router route-table 10.0.0.0/8 alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
10.0.0.0/8                                    Remote  BGP       00h02m12s  170
       192.168.13.1                                                 0
10.0.0.0/8 (Backup)                           Remote  BGP       00h02m12s  170
       192.168.23.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

In case of link failure between CE-4 and PE-1, the same BGP withdrawals will be sent from PE-1 to RR-5 and from RR-5 to PE-2 and PE-3. When PE-2 and PE-3 receive the withdrawal, the BGP decision process need not run again. The backup path is promoted to active immediately.

BGP FRR is disabled on the PEs as follows:

```
configure router bgp no backup-path
```

Add-Path and ECMP

On PE-1, the import policy is removed to have paths with equal cost:

```
*A:PE-1# configure router bgp group "eBGP" no import
```

ECMP is enabled on all PEs with a value of two, as follows:

```
configure router ecmp 2
```

On all PEs, BGP multipath is configured with a value of two in the BGP context, as follows:

```
configure router bgp multipath 2
```

For more information about BGP multipath, see chapter BGP Multipath.

All PEs have two routes for prefix 10.0.0.0/8 and both are active when ECMP is enabled; for example, for PE-3, as follows:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.0.0/8                                         100         None
      192.0.2.1                                          11          -
      64500
u*>i  10.0.0.0/8                                         100         None
      192.0.2.2                                          6           -
      64500
-------------------------------------------------------------------------------
Routes : 2


*A:PE-3# show router route-table 10.0.0.0/8

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
```

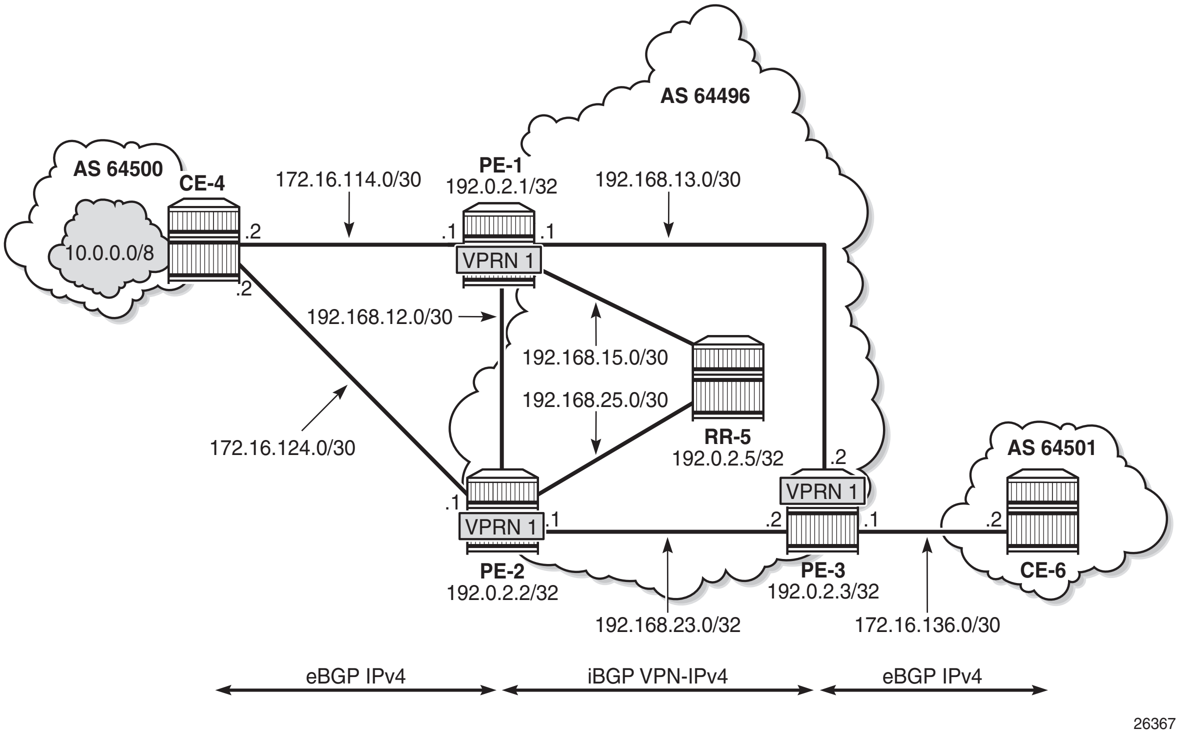
```
10.0.0.0/8                                      Remote   BGP        00h02m40s  170
       192.168.13.1                                                     0
10.0.0.0/8                                      Remote   BGP        00h02m40s  170
       192.168.23.1                                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
```

Traffic flows with destination 10.0.0.0/8 will be sprayed over the two active paths.

Add-path for Family VPN-IPv4 with BGP FRR

Figure 87 shows the example topology with VPRN1 configured on the PEs in AS 64496. CE-4 exports prefix 172.31.0.0/16 to VPRN 1 on PE-1 and PE-2.

Figure 87 **Example Topology with VPRNs**



VPRN 1 is configured on all PEs in AS 64496, but not on the RR. BGP FRR is enabled in the VPRN with the **enable-bgp-vpn-backup** option. The configuration of VPRN 1 is similar on all PEs; for example, for PE-1, the VPRN configuration is as follows:

```
configure
    service
```

```
        vprn 1 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution any
            exit
            enable-bgp-vpn-backup ipv4
            vrf-target target:64496:1
            interface "int-PE-1-CE-4_VPRN1" create
                address 172.16.114.1/30
                sap 1/1/3:1 create
                exit
            exit
            bgp
                split-horizon
                group "eBGP_1"
                    next-hop-self
                    import "import-bgp-LP200"
                    peer-as 64500
                    neighbor 172.16.114.2
                    exit
                exit
            exit
            export-inactive-bgp
            no shutdown
```

The import policy sets the LP to 200 for the routes received from CE-4. The configuration on PE-2 is similar, but without any import policy. Therefore, the path via PE-1 will be preferred over the path via PE-2.

The **export-inactive-bgp** option must be configured on PE-2, because the route for prefix 172.31.0.0/16 received by PE-2 from CE-4 is inactive, but should still be advertised as BGP VPN-IPv4 route to RR-5; see chapter *BGP Best-External in a VPRN*. In this example, the **export-inactive-bgp** option is configured on all PEs.

On the CEs, the configuration is either in the base routing instance-with additional router interfaces and BGP neighbors-or in a VPRN. In this example, the following VPRN is configured on CE-4:

```
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64500
            route-distinguisher 64500:1
            interface "int-CE-4-PE-1_VPRN1" create
                address 172.16.114.2/30
                sap 1/1/1:1 create
                exit
            exit
            interface "int-CE-4-PE-2_VPRN1" create
                address 172.16.124.2/30
                sap 1/1/2:1 create
                exit
            exit
            interface "test_connectedNW" create
```

```
                address 172.31.0.1/16
                loopback
            exit
            bgp
                split-horizon
                group "eBGP_1"
                    export "export_172.31.0.0/16"
                    peer-as 64496
                    neighbor 172.16.114.1
                    exit
                    neighbor 172.16.124.1
                    exit
                exit
            exit
            no shutdown
```

The export policy to export prefix 172.31.0.0/16 is defined as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "172.31.0.0/16"
                prefix 172.31.0.0/16 longer
            exit
            policy-statement "export_172.31.0.0/16"
                entry 10
                    from
                        prefix-list "172.31.0.0/16"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The configuration on CE-6 is similar, but no prefix is exported from CE-6.

For all BGP speakers in AS 64496, BGP must be configured for address family VPN-IPv4 as well as for IPv4, as follows:

```
configure router bgp group "iBGP" family ipv4 vpn-ipv4
```

BGP add-path cannot be enabled in the BGP context within a VPRN. However, BGP add-path can be enabled in the base routing instance for address family VPN-IPv4. This is done on all PEs at group level with the following command:

```
configure router bgp group "iBGP" add-paths vpn-ipv4 send 2
```

In this example, BGP add-path is enabled at neighbor level on RR-5, as follows:

```
configure router bgp group "iBGP" neighbor 192.0.2.1 add-paths vpn-ipv4 send 2
configure router bgp group "iBGP" neighbor 192.0.2.2 add-paths vpn-ipv4 send 2
configure router bgp group "iBGP" neighbor 192.0.2.3 add-paths vpn-ipv4 send 2
```

The BGP configuration for group "iBGP" on PE-1 is as follows:

```
*A:PE-1# configure router bgp group "iBGP"
*A:PE-1>config>router>bgp>group# info
----------------------------------------------
                family ipv4 vpn-ipv4
                next-hop-self
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                    vpn-ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
----------------------------------------------
```

With add-path enabled for address family vpn-ipv4, PE-1 and PE-2 will advertise their route for prefix 172.31.0.0/16 as VPN-IPv4 route to RR-5. RR-5 will advertise both routes to its other RR clients. PE-3 receives two VPN-IPv4 routes for prefix 172.31.0.0/16, as follows:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.0.0/16                              100         None
      192.0.2.1                                          20          262140
      64500
ub*>i 64496:1:172.31.0.0/16                              100         None
      192.0.2.2                                          13          262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

Both routes are used: the route via PE-1 is the active route and the route via PE-2 is used as a backup, as indicated by the "b" flag.

The routing table for VPRN 1 on PE-3 shows that there is a backup route for prefix 172.31.0.0/16, as indicated by "B" as follows:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
172.31.0.0/16 [B]                             Remote  BGP VPN   00h00m27s  170
       192.0.2.1 (tunneled)                                       0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

The active route and the alternative (backup) route are shown in the following output:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16 alternative

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
      Alt-NextHop                                              Alt-
                                                                Metric
-------------------------------------------------------------------------------
172.31.0.0/16                                 Remote  BGP VPN   00h03m06s  170
       192.0.2.1 (tunneled)                                       0
172.31.0.0/16 (Backup)                        Remote  BGP VPN   00h03m06s  170
       192.0.2.2 (tunneled)                                       0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
```

BGP FRR is disabled in VPRN 1 on the PEs, as follows:

```
configure service vprn 1 no enable-bgp-vpn-backup
```

Add-path for Family VPN-IPv4 with ECMP

The import policy is removed in VPRN 1 on PE-1 to make the cost of the paths via PE-1 and PE-2 equal, as follows:

```
*A:PE-1# configure service vprn 1 bgp group "eBGP_1" no import
```

ECMP is enabled in VPRN 1 on all PEs, as follows:

```
configure service vprn 1 ecmp 2
```

BGP multipath needs to be enabled in the base routing context, but that already happened.

With ECMP enabled, the two routes that are received on PE-3 from RR-5 are both active, as follows:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.0.0/16                              100         None
      192.0.2.1                                          20          262140
      64500
u*>i  64496:1:172.31.0.0/16                              100         None
      192.0.2.2                                          13          262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

ECMP is enabled with a value of two, so traffic flows in VPRN 1 on PE-3 with destination 172.31.0.0/16 are distributed over two paths: one via PE-1 and another via PE-2, as follows:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.31.0.0/16                                 Remote  BGP VPN    00h06m28s 170
       192.0.2.1 (tunneled)                                         0
172.31.0.0/16                                 Remote  BGP VPN    00h06m28s 170
       192.0.2.2 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 2
```

Conclusion

BGP add-path allows BGP speakers to advertise multiple distinct paths for the same prefix. The potential benefits of BGP add-path include reduced routing churn, faster convergence, and better load-sharing.

BGP Add-Path Policy Control

This chapter provides information about BGP Add-Path Policy Control.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4.

Overview

BGP Add-Path allows for advertising multiple paths per prefix for faster convergence, load sharing, and reduction of routing churn. See the BGP Add-Path chapter for more information.

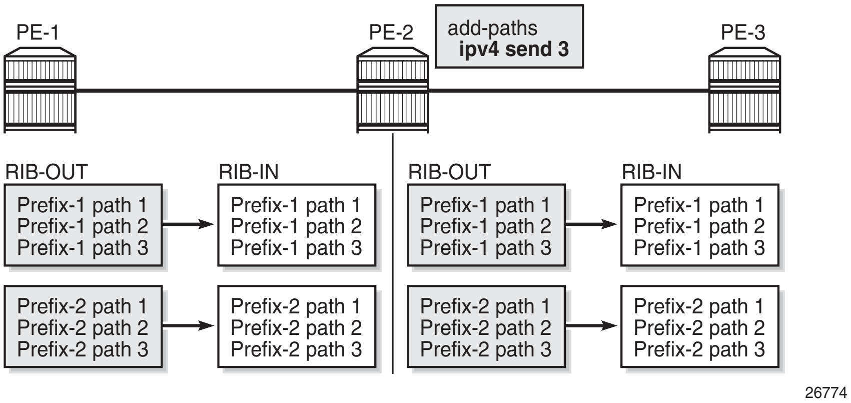
The BGP Add-Path Policy Control feature extends the functionality of BGP Add-Path, which was able to control the number of advertised paths per prefix per address family. This meant that all prefixes that belonged to an address family (such as IPv4, IPv6, and so on) were subject to the same sending limit imposed by the **send-limit** configured at the BGP instance, group, or neighbor level.

BGP Add-Path Policy Control adds the capability to configure the number of advertised paths on a per-prefix basis. The **add-paths-send-limit** route policy action allows overriding the **send-limit** in the BGP context for selected prefixes. This adds finer granularity to BGP Add-Path, where a global path limit is defined at the relevant BGP level and specific limits can be defined for exceptional prefixes at an import policy level.

A value between 1 and 16 is configurable for **add-paths-send-limit**.

Figure 88 shows a topology for BGP Add-Paths before policy control.

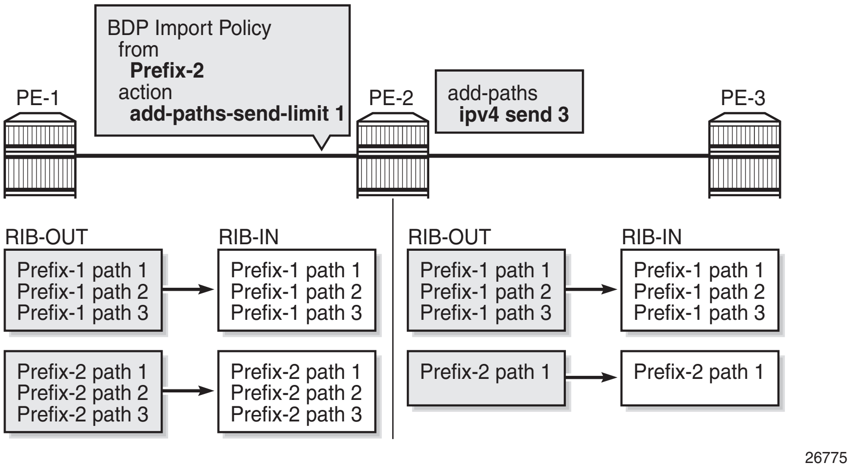
Figure 88 **BGP Add-Paths Before Policy Control**



In Figure 88, PE-2 receives two prefixes with three diverse paths from PE-1. PE-2 has a send-limit with a value of 3 configured at a BGP level that is applicable to PE-3. Therefore, PE-2 sends both prefixes with three different path IDs to PE-3.

Figure 89 shows a topology for BGP Add-Paths after policy control.

Figure 89 **BGP Add-Paths After Policy Control**



In Figure 89, a BGP-import policy is applied on PE-2. The policy selectively applies a send-limit of 1 on the paths received for Prefix-2. Therefore, PE-2 sends only one path for Prefix-2 to PE-3, while the BGP level send-limit of 3 still applies for Prefix-1.

The policy action is only applicable for BGP-import policy and has no effect on BGP-export policy, VRF-import policy, or VRF-export policy. The reason for this is that the policy needs to be applied on the routes accepted into the RIB-in, otherwise two or more paths may not be present.

The BGP-import policy does not match VPN-IP routes unless the **vpn-apply-import** command is configured in the BGP global base, group, or neighbor level.

Note: The route policy only controls the number of advertised paths, not the set of paths.

Configuration

The following configuration examples are in this section:

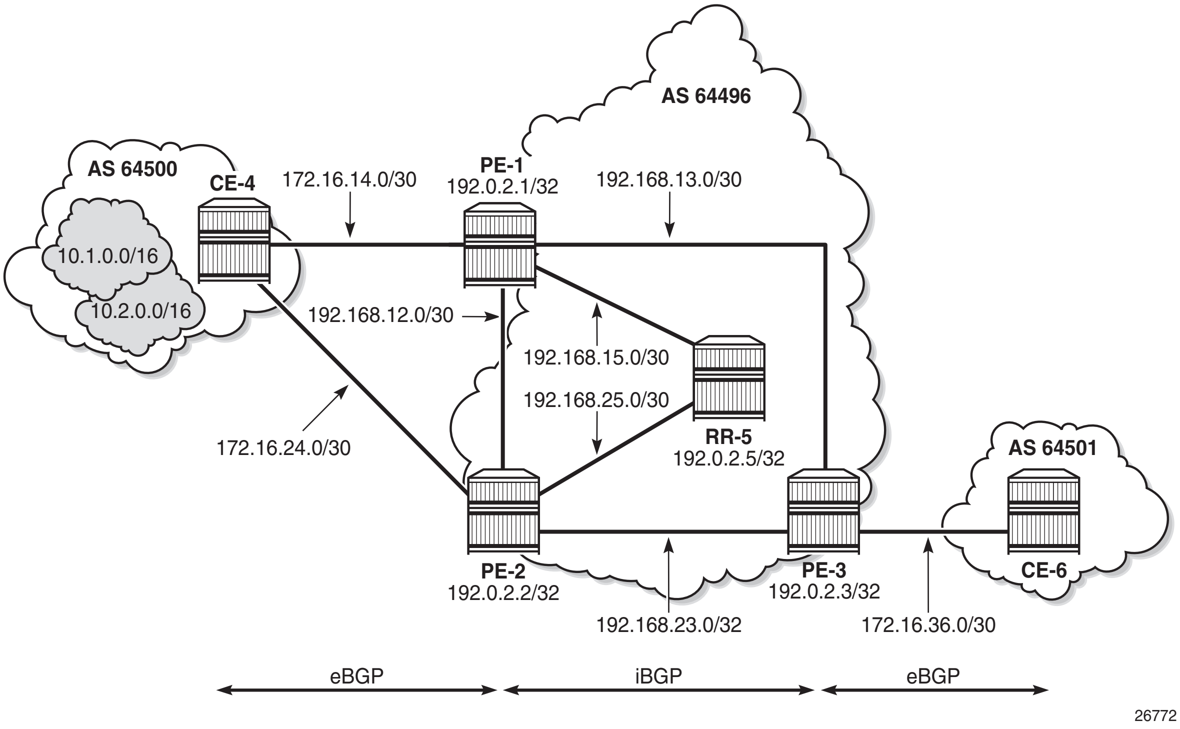
- BGP Add-Path for address family IPv4 without policy control
- BGP Add-Path for address family IPv4 with policy control
- BGP Add-Path for address family VPN-IPv4 with policy control

BGP Add-Path Policy Control Feature for Address Family IPv4

Figure 90 shows the example topology used for the BGP Add-Path Policy Control feature for the IPv4 address family. The topology used is similar to the one in the BGP Add-Path chapter, with the following characteristics:

- CE-4 in AS 64500 advertises both prefixes 10.1.0.0/16 and 10.2.0.0/16 to its eBGP peers PE-1 and PE-2 in AS 64496.
- RR-5 is route reflector for all PEs in AS 64496.
- Add-Path is configured on all PE routers and RR-5 with a send-limit of 2.
- CE-6 in AS 64501 peers with PE-3 in AS 64496 and can send traffic to CE-4 in 64500.

Figure 90 **Example Topology - IPv4**



Initial Configuration

The initial configuration on all nodes includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP on all interfaces within AS 64496 (alternatively, OSPF can be used)
- LDP on all interfaces between the PEs in AS 64496, but not toward RR-5. LDP will be used to create the transport tunnels that will bound to the VPRN services BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports prefixes 10.1.0.0/16 and 10.2.0.0/16 to both eBGP peers, as follows:in the VPN-IPv4 address family section.

BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports prefixes 10.1.0.0/16 and 10.2.0.0/16 to both eBGP peers, as follows:

```
# on CE-4
configure
    router
```

```
autonomous-system 64500
bgp
    min-route-advertisement 1
    rapid-withdrawal
    split-horizon
    group "eBGP"
        export "export-bgp"
        peer-as 64496
        neighbor 172.16.14.1
        exit
        neighbor 172.16.24.1
        exit
    exit
    no shutdown
exit
policy-options
    begin
    prefix-list "10.1.0.0/16"
        prefix 10.1.0.0/16 longer
    exit
    prefix-list "10.2.0.0/16"
        prefix 10.2.0.0/16 longer
    exit
    policy-statement "export-bgp"
        entry 10
            from
                prefix-list "10.1.0.0/16"
            exit
            action accept
            exit
        exit
        entry 20
            from
                prefix-list "10.2.0.0/16"
            exit
            action accept
            exit
        exit
    exit
    commit
exit
interface "int-loopback-1"
    address 10.1.1.1/16
    loopback
    no shutdown
exit
interface "int-loopback-2"
    address 10.2.1.1/16
    loopback
    no shutdown
exit
```

The BGP configuration on CE-6 is similar, except for the export policy.

PE-1 peers with CE-4 in AS 65400 and RR-5 in AS 64496. The BGP configuration on PE-1 is as follows:

```
configure
```

```
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "eBGP"
                peer-as 64500
                neighbor 172.16.14.2
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on PE-2 and PE-3 is similar to that of PE-1.

RR-5 acts as a route reflector to all the PEs in AS 64500 with a cluster ID of 5.5.5.5. The configuration on RR-5 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "iBGP"
                cluster 5.5.5.5
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
```

BGP Add-Path for Address Family IPv4 without Policy Control

RR-5 receives both the 10.1.0.0/16 and 10.2.0.0/16 prefixes with two paths from PE-1 and PE-2:

```
*A:RR-5# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.5        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          2           -
      64500
*i    10.1.0.0/16                                        100         None
      192.0.2.2                                          6           -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.1                                          1           -
      64500
*i    10.2.0.0/16                                        100         None
      192.0.2.2                                          5           -
      64500
-------------------------------------------------------------------------------
Routes : 4
```

RR-5 propagates these updates to its clients, for example to PE-3, as follows:

```
118 2017/05/19 17:32:52.79 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 41
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.2
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.2
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        5.5.5.5
    NLRI: Length = 14
        10.2.0.0/16 Path-ID 70
        10.1.0.0/16 Path-ID 68
"


119 2017/05/19 17:32:52.79 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
```

```
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 41
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64500 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.1
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.1
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        5.5.5.5
    NLRI: Length = 14
        10.2.0.0/16 Path-ID 69
        10.1.0.0/16 Path-ID 67
"
```

PE-3 receives both prefixes in its BGP routing table with two different paths (also, optionally, has ECMP and BGP multipath enabled as explained in the BGP Add-Path chapter):

```
A:PE-3# configure router ecmp 2
A:PE-3# configure router bgp multipath 2
A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          67          -
      64500
u*>i  10.1.0.0/16                                        100         None
      192.0.2.2                                          68          -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.1                                          69          -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.2                                          70          -
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
```

BGP Add-Path for Address Family IPv4 with Policy Control

The following policy is enabled on RR-5, which limits the number of advertised paths for prefix 10.2.0.0/16 to one:

```
# on RR-5
configure
    router
        policy-options
            begin
            prefix-list "10.2.0.0/16"
                prefix 10.2.0.0/16 longer
            exit
            policy-statement "import-add-path"
                entry 10
                    from
                        prefix-list "10.2.0.0/16"
                    exit
                    action accept
                        add-paths-send-limit 1
                    exit
                exit
            exit
            commit
        exit
configure router bgp group "iBGP" import "import-add-path"
```

RR-5 sends the following withdrawal message to PE-3:

```
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 7
        10.2.0.0/16 Path-ID 70
    Total Path Attr Length = 0
"
```

PE-3 deletes the route with Path-ID 70 for prefix 10.2.0.0/16:

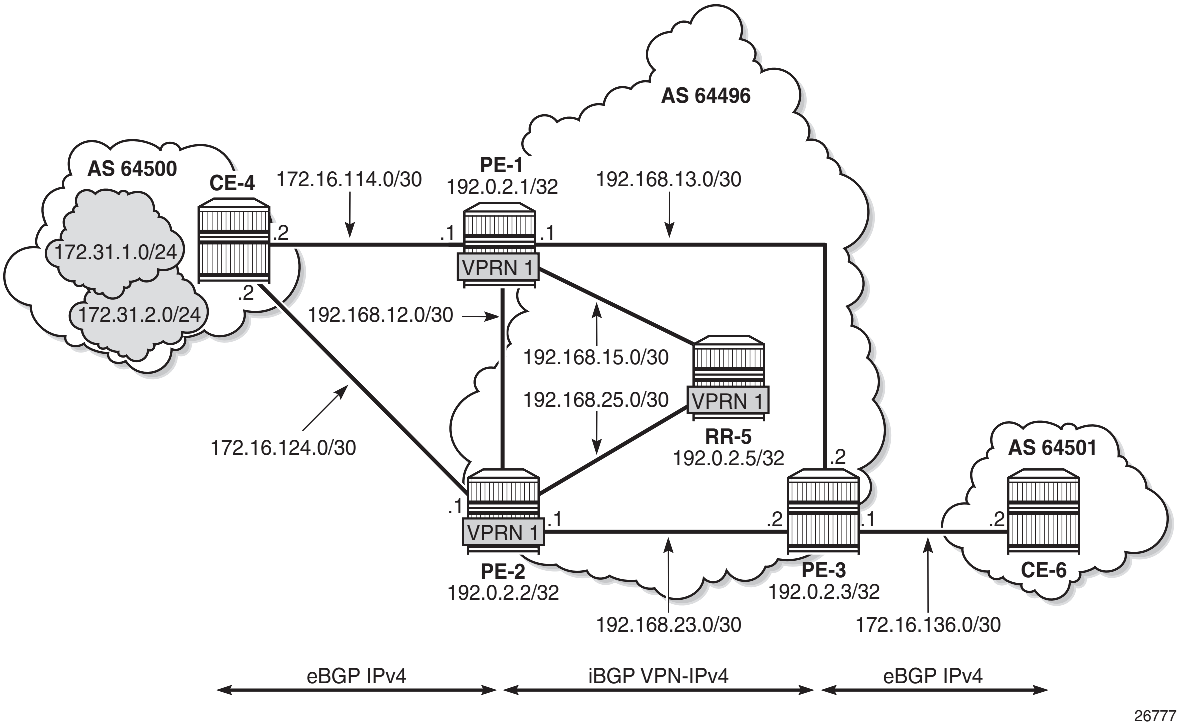
```
A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          67          -
      64500
```

```
u*>i  10.1.0.0/16                                         100        None
      192.0.2.2                                           68         -
      64500
u*>i  10.2.0.0/16                                         100        None
      192.0.2.1                                           69         -
      64500
-------------------------------------------------------------------------------
Routes : 3
```

BGP Add-Path for Address Family VPN-IPv4 with Policy Control

Figure 91 shows the example topology used for the BGP Add-Path Policy Control feature for VPN-IPv4 route family. The topology used is similar to the one used in the BGP Add-Path chapter. CE-4 exports both prefixes 172.31.1.0/24 and 172.31.2.0/24 to VPRN 1 on PE-1 and PE-2.

Figure 91 **Example Topology - VPN-IPv4**



VPRN 1 is configured on all PEs in AS 64496. The configuration of VPRN 1 is similar on all PEs; for example, for PE-1, the VPRN configuration is as follows:

```
configure
```

```
    service
        customer 1 create
            description "Default customer"
        exit
        vprn 1 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:1
            interface "int-PE-1-CE-4-VPRN1" create
                address 172.16.114.1/30
                sap 1/1/1:1 create
                exit
            exit
            bgp
                min-route-advertisement 1
                split-horizon
                group "eBGP-1"
                    peer-as 64500
                    neighbor 172.16.114.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

On the CEs, the configuration is either in the base routing instance, with additional router interfaces and BGP neighbors, or in a VPRN. In this example, the following VPRN is configured on CE-4:

```
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64500
            route-distinguisher 64500:1
            interface "int-CE-4-PE-1-VPRN1" create
                address 172.16.114.2/30
                sap 1/1/2:1 create
                exit
            exit
            interface "int-CE-4-PE-2-VPRN1" create
                address 172.16.124.2/30
                sap 1/1/1:1 create
                exit
            exit
            interface "loopback1-VPRN1" create
                address 172.31.1.1/24
                loopback
            exit
            interface "loopback2-VPRN1" create
                address 172.31.2.1/24
                loopback
            exit
            bgp
                min-route-advertisement 1
                split-horizon
                group "eBGP-1"
                    export "export-VPRN1"
                    peer-as 64496
                    neighbor 172.16.114.1
                    exit
                    neighbor 172.16.124.1
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

The export policy to export prefixes 172.31.1.0/24 and 172.31.2.0/24 is defined as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "172.31.0.0/16"
                prefix 172.31.0.0/16 longer
            exit
            policy-statement "export-VPRN1"
                entry 10
                    from
                        prefix-list "172.31.0.0/16"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The configuration on CE-6 is similar, but no prefix is exported from CE-6.

For all BGP speakers in AS 64496, BGP must be configured for address family VPN-IPv4 as well as for IPv4, as follows:

```
configure router bgp group "iBGP" family ipv4 vpn-ipv4
```

BGP Add-Path cannot be enabled in the BGP context within a VPRN. However, it can be enabled in the base routing instance for address family VPN-IPv4. This is done on all PEs and RR-5 at group level with the following command:

```
configure router bgp group "iBGP" add-paths vpn-ipv4 send 2
```

The BGP configuration for group iBGP on PE-1 is as follows:

```
*A:PE-1>config>router>bgp>group# info
----------------------------------------------
                family ipv4 vpn-ipv4
                next-hop-self
                peer-as 64496
                add-paths
```

```
                    ipv4 send 2 receive
                    vpn-ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
----------------------------------------------
```

With Add-Path enabled for address family VPN-IPv4, PE-1 and PE-2 will advertise their routes for prefixes 172.31.1.0/24 and 172.31.2.0/24 as VPN-IPv4 routes to RR-5. RR-5 will advertise both routes to its other RR clients. PE-3 receives two VPN-IPv4 routes for each of the prefixes 172.31.1.0/24 and 172.31.2.0/24, as follows:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4 longer
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.1                                          11          262140
      64500
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.2                                          22          262140
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.1                                          13          262140
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.2                                          24          262140
      64500
-------------------------------------------------------------------------------
Routes : 4
```

All routes are used due to the ECMP setting in VPRN 1:

```
*A:PE-3# configure service vprn 1 ecmp 2
```

Alternatively, BGP FRR can be enabled for VPRN 1, as explained in the BGP Add-Path chapter.

To limit the advertisement of prefix 172.31.2.0/24 to a single path, the following route policy is configured on RR-5:

```
configure
    router
        policy-options
```

```
begin
prefix-list "172.31.2.0/24"
    prefix 172.31.2.0/24 longer
exit
policy-statement "import-add-path"
    entry 20
        from
            prefix-list "172.31.2.0/24"
        exit
        action accept
            add-paths-send-limit 1
        exit
    exit
exit
commit
```

The policy entry for prefix 172.31.2.0/24 can be configured in a new policy-statement or be added to an existing BGP policy (used for the previous IPv4 Add-Path policy section, for example).

If this is a new policy-statement, apply the policy in the iBGP group context on RR-5:

```
*A:RR-5# configure router bgp group "iBGP" import "import-add-path"
```

At this point, PE-3 still has two paths for each of the prefixes:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4 longer
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.1                                          11          262140
      64500
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.2                                          22          262140
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.1                                          13          262140
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.2                                          24          262140
      64500
-------------------------------------------------------------------------------
Routes : 4
```

The following configuration is applied on RR-5 to make the BGP policy effective for VPN-IPV4 routes:

```
*A:RR-5# configure router bgp group "iBGP" vpn-apply-import
```

Upon application of this command, RR-5 sends the following withdrawal to PE-3:

```
116 2017/05/23 16:28:30.36 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.3
"Peer 1: 192.0.2.3: UPDATE
Peer 1: 192.0.2.3 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 26
    Flag: 0x90 Type: 15 Len: 22 Multiprotocol Unreachable NLRI:
        Address Family VPN_IPV4
        172.31.2.0/24 RD 64496:1 Label 0 Path-ID 24
"
```

PE-3 now has a single route for prefix 172.31.2.0/24 in its BGP routing table:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4 longer
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.1                                          11          262140
      64500
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.2                                          22          262140
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.1                                          13          262140
      64500
-------------------------------------------------------------------------------
Routes : 3
```

PE-3 has installed a single route for prefix 172.31.2.0/24 in its VPRN route table:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16 longer
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.31.1.0/24                                 Remote  BGP VPN   02h06m36s  170
```

```
       192.0.2.1 (tunneled)                                          0
172.31.1.0/24                                  Remote  BGP VPN   02h06m36s  170
       192.0.2.2 (tunneled)                                          0
172.31.2.0/24                                  Remote  BGP VPN   00h46m33s  170
       192.0.2.1 (tunneled)                                          0
--------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
```

Conclusion

The BGP Add-Path Policy Control feature allows BGP speakers to advertise multiple distinct paths for the same prefix. The potential benefits of using BGP Add-Path Policy Control are increased granularity and flexibility in advertising multiple paths to BGP neighbors.

BGP Conditional Route Advertisement

This chapter provides information about BGP Conditional Route Advertisement.

Topics in this chapter include:

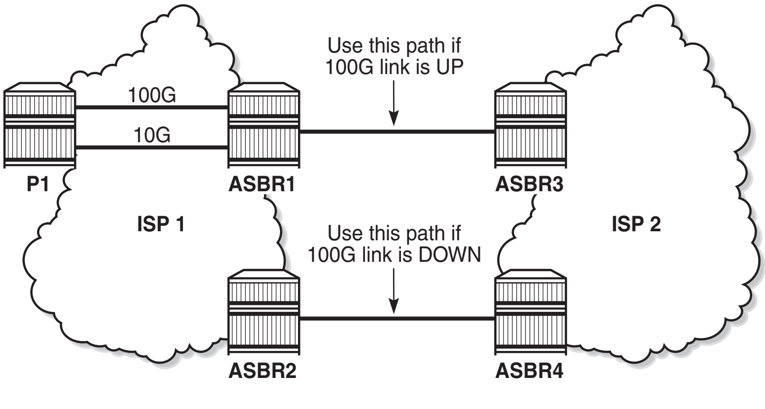
Applicability

The information and configuration in this chapter was originally based on SR OS Release 15.0.R4. The CLI in the current edition is based on SR OS Release 16.0.R1.

Overview

The BGP conditional route advertisement feature allows a router to control the advertisement of routes based on predetermined routes in the route table. Figure 92 shows an example in which this feature can bring flexibility in an ISP peering scenario.

Figure 92 **Conditional BGP Route Advertisement - ISP Peering**



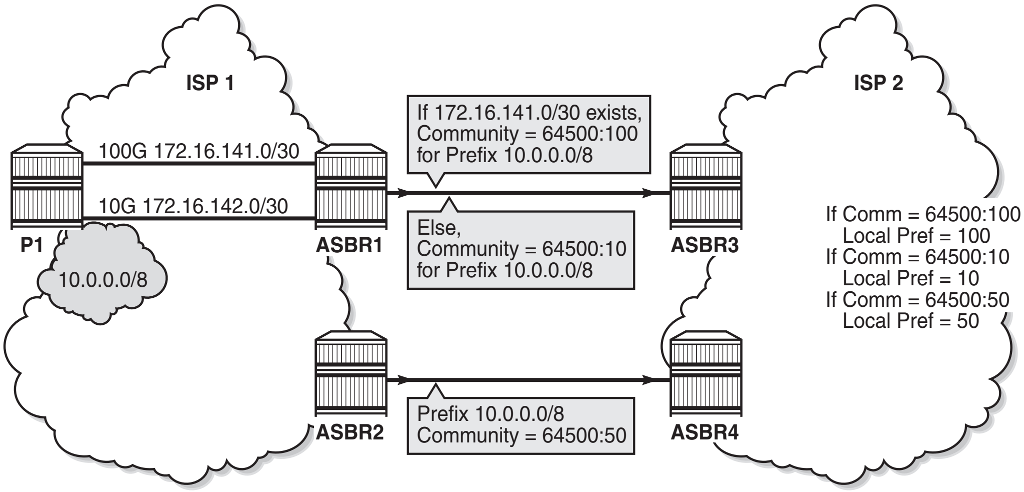
ISP 1 and ISP 2 have two peering points; a first between ASBR1 and ASBR3, and a second between ASBR2 and ASBR4. For redundancy, ISP 1 has two links between ASBR1 and the internal P1 router, one with 100 Gb/s and the other with 10 Gb/s capacity. According to the service agreement, ISP 1 instructs ISP 2 to send traffic using the upper path (between ASBR1 and ASBR3) only if the 100 Gb/s link between P1 and ASBR1 is up. If this is not the case, ISP 2 uses the lower path.

To implement the BGP conditional route advertisement feature, a conditional route policy entry is used. The route policy is as follows:

- Within a policy-statement entry, a conditional expression is created.
- The conditional expression tests for active IPv4 or IPv6 routes defined in a prefix list.
- If the expression is true, the action commands of the policy entry are applied.
- If the expression is false, the entire policy entry is skipped and processing continues with the next policy entry.
- Conditional expressions are only applicable when the route policy is used as a BGP export policy or a VRF export policy.

Figure 93 shows the implementation using the example in Figure 1.

Figure 93   Conditional BGP Route Advertisement Implementation Example



The prefix of the 100G interface between ASBR1 and P1 is 172.16.141.0/30. ASBR1 receives prefix 10.0.0.0/8 from P1 via BGP. Under standard conditions, the 100G interface is up and 172.16.141.0/30 exists in the route table and ASBR1 advertises 10.0.0.0/8 with a community value of 64500:100. ASBR2 advertises the same prefix with a community value of 64500:50. ASBR3 and ASBR4 in ISP 2 use an import policy that applies local preference values of 100 and 50 on the routes advertised by ASBR1 and ASBR2, respectively. As a result, the routers in ISP 2 prefer ASBR3 as an exit point for traffic flowing toward ISP 1.

If the 100G interface goes down, the prefix 172.16.141.0/30 is withdrawn from the route table and, as a result, ASBR1 starts advertising 10.0.0.0/8 with a community value of 64500:10. ASBR3 and ASBR4 adjust the local preference value for ASBR1 to 10 and, therefore, ASBR4 becomes the preferred exit point for routers in ISP 2.

The only conditional expression that can be contained in a policy-statement entry is a route-existence test defined by the **route-exists** keyword in the CLI. The command accepts two parameters: **all** and **none**:

- If neither the **all** nor the **none** parameter is used, the match logic is **any** - that is, the conditional expression is true if any exact match entry in the referenced prefix list has an active route in the route table associated with the policy.
- **all** - the conditional expression is true only if all the exact match entries in the referenced prefix list have an active route in the route table associated with the policy.
- **none** - the conditional expression is true only if none of the exact match entries in the referenced prefix list have an active route in the route table associated with the policy.

Configuration

The following configuration examples are covered in this section:

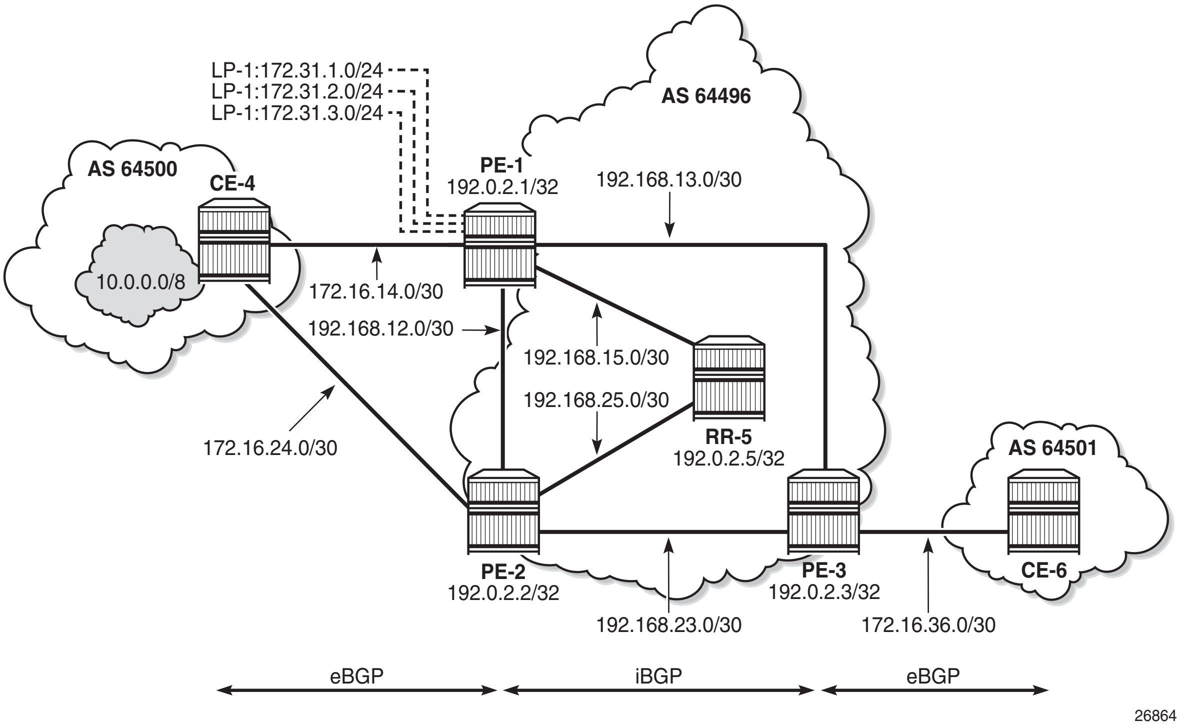
- BGP Conditional Route Advertisement using **any** prefix list match
- BGP Conditional Route Advertisement using **all** prefix list match
- BGP Conditional Route Advertisement using **none** prefix list match

Figure 94 shows the example topology for BGP conditional route advertisement with the following characteristics:

- CE-4 in AS 64500 advertises prefix 10.0.0.0/8 to its eBGP peers PE-1 and PE-2 in AS 64496.

- PE-1 has three loopback interfaces configured to demonstrate the use of conditional route advertisement: LP-1, LP-2, and LP-3.
- RR-5 is route reflector for all PEs in AS 64496.
- CE-6 in AS 64501 peers with PE-3 in AS 64496 and can send traffic to CE-4 in 64500.

Figure 94 **Example Topology**



Initial Configuration

The initial configuration on all nodes includes:

- Cards, MDAs, ports
- LAG configured for the link between CE-4 and PE-1 with two member links
- Router interfaces
- IS-IS as IGP on all interfaces within AS 64496 (alternatively, OSPF can be used)

BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports the prefix 10.0.0.0/8 to both eBGP peers, which includes the address of the *int-loopback-1* interface, as follows:

```
configure
    router
        interface "int-loopback-1"
            address 10.1.1.1/8
            loopback
            no shutdown
        exit
        autonomous-system 64500
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            policy-statement "policy-export-bgp"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        bgp
            rapid-withdrawal
            split-horizon
            group "eBGP"
                export "policy-export-bgp"
                peer-as 64496
                neighbor 172.16.14.1
                exit
                neighbor 172.16.24.1
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on CE-6 is similar, except for the loopback interface and export policy.

PE-1 peers with CE-4 in AS 65400 and RR-5 in AS 64496. The BGP configuration on PE-1 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            rapid-withdrawal
            split-horizon
            group "eBGP"
                peer-as 64500
                neighbor 172.16.14.2
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 64496
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on PE-2 and PE-3 is similar to that of PE-1.

RR-5 is the route reflector to all the PEs in AS 64500 with a cluster ID of 5.5.5.5. The configuration on RR-5 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            rapid-withdrawal
            split-horizon
            group "iBGP"
                cluster 5.5.5.5
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
```

Three loopback interfaces are configured in PE-1 to be used for route existence tests:

```
configure
    router
        interface "int-loopback-1"
            address 172.31.1.1/24
            loopback
            no shutdown
        exit
        interface "int-loopback-2"
            address 172.31.2.1/24
            loopback
            no shutdown
        exit
        interface "int-loopback-3"
            address 172.31.3.1/24
            loopback
            no shutdown
        exit
```

BGP Conditional Route Advertisement Using "any" Prefix List Match

In the initial condition, RR-5 receives the prefix 10.0.0.0/8 from PE-1 and PE-2 with no community values and the default local preference value of 100:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                         match "^Nexthop|community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : No Community Members
Nexthop        : 192.0.2.2
Local Pref.    : 100                    Interface Name : int-RR-5-PE-2
Community      : No Community Members
*A:RR-5#
```

The following policy is configured on PE-1 that adds the community 64500:100 to the 10.0.0.0/8 prefix advertised to RR-5 if any of the conditional prefixes in the prefix list are active in the route table:

```
configure
    router
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            prefix-list "prefix-conditional-routes"
                prefix 172.31.1.0/24 longer
                prefix 172.31.2.0/24 longer
                prefix 172.31.3.0/24 longer
            exit
            community "64500:10" members "64500:10"
            community "64500:100" members "64500:100"
            policy-statement "policy-bgp-community"
                entry 10
                    conditional-expression
                        route-exists "[prefix-conditional-routes]"
                    exit
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                        community add "64500:100"
                    exit
                exit
                entry 20
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                        community add "64500:10"
                    exit
                exit
            exit
            commit
        exit
```

Special attention is required on the policy syntax. The square brackets […] in the expression of the **route-exists** command are very important.

The following policy is configured on PE-2 that adds the community 64500:50 to the 10.0.0.0/8 prefix advertised to RR-5 without any conditions:

```
configure
    router
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            community "64500:50" members "64500:50"
            policy-statement "policy-bgp-community"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                        community add "64500:50"
                    exit
                exit
            exit
            commit
        exit
```

The policy is applied to the iBGP group on PE-1 and PE-2:

```
*A:PE-1# configure router bgp group "iBGP" export "policy-bgp-community"
```

```
*A:PE-2# configure router bgp group "iBGP" export "policy-bgp-community"
```

The prefix 10.0.0.0/8 is received on RR-5 with the respective community values and still with the default local preference values:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                         match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : 64500:100
Nexthop        : 192.0.2.2
Local Pref.    : 100                    Interface Name : int-RR-5-PE-2
Community      : 64500:50
*A:RR-5#
```

The following policy is configured on RR-5 to apply different local preference values based on the corresponding community value:

```
configure
```

```
    router
        policy-options
            begin
            community "64500:10" members "64500:10"
            community "64500:50" members "64500:50"
            community "64500:100" members "64500:100"
            policy-statement "policy-bgp-preference"
                entry 10
                    from
                        community "64500:100"
                    exit
                    action accept
                        local-preference 100
                    exit
                exit
                entry 20
                    from
                        community "64500:50"
                    exit
                    action accept
                        local-preference 50
                    exit
                exit
                entry 30
                    from
                        community "64500:10"
                    exit
                    action accept
                        local-preference 10
                    exit
                exit
            exit
            commit
        exit
```

The policy is applied on RR-5:

```
*A:RR-5# configure router bgp group "iBGP" import "policy-bgp-preference"
```

The following command output shows the correct local preference values are applied on the routes received from PE-1 and PE-2:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                         match "^Nextop |Community |Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : 64500:100
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
TieBreakReason : LocalPref
--- snipped ---
*A:RR-5#
```

RR-5 advertises the route with local preference of 100 to PE-3, with next hop PE-1:

```
*A:PE-3# show router bgp routes hunt brief |
                                        match "^Nexthop |Community |Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-PE-3-PE-1
Community      : 64500:100
*A:PE-3#
```

The first loopback interface is shutdown on PE-1, which results in the withdrawal of prefix 172.31.1.0/24 from the route table on PE-1:

```
*A:PE-1# configure router interface "int-loopback-1" shutdown
```

PE-1 still advertises the prefix 10.0.0.0/8 with the community 64500:100:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : 64500:100
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
TieBreakReason : LocalPref
*A:RR-5#
```

The second loopback interface is shutdown on PE-1, which results in the withdrawal of prefix 172.31.2.0/24 from the route on PE-1:

```
*A:PE-1# configure router interface "int-loopback-2" shutdown
```

PE-1 still advertises the prefix 10.0.0.0/8 with the community 64500:100:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : 64500:100
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
TieBreakReason : LocalPref
--- snipped ---
*A:RR-5#
```

The third and the last loopback interface is shutdown on PE-1, which results in the withdrawal of prefix 172.31.3.0/24 from the route table on PE-1:

```
*A:PE-1# configure router interface "int-loopback-3" shutdown
```

PE-1 now starts advertising the prefix 10.0.0.0/8 with the community 64500:10 and RR-5 applies local preference 10 for this route:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
Nexthop        : 192.0.2.1
Local Pref.    : 10                     Interface Name : int-RR-5-PE-1
Community      : 64500:10
TieBreakReason : LocalPref
*A:RR-5#
```

RR-5 advertises prefix 10.0.0.0/8 to PE-3 with the next-hop address of PE-2:

```
*A:PE-3# show router bgp routes 10.0.0.0/8 hunt |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-PE-3-PE-2
Community      : 64500:50
*A:PE-3#
```

BGP Conditional Route Advertisement Using "all" Prefix List Match

The loopback interfaces on PE-1 are re-enabled:

```
*A:PE-1# configure router interface int-loopback-[1..3] no shutdown
```

Note: Do not use quotes in the interface name when using **ranges**, because it will be treated as a new interface creation.

The policy on PE-1 is changed so that the prefix 10.0.0.0/8 is advertised with community 64500:100 only if all the prefixes in the prefix list are active:

```
configure
    router
        policy-options
            begin
            policy-statement "policy-bgp-community"
                entry 10
                    conditional-expression
                        route-exists "[prefix-conditional-routes] all"
                    exit
                exit
            exit
            commit
        exit
```

The first loopback interface is shutdown on PE-1, which results in the withdrawal of prefix 172.31.1.0/24 from the route table on PE-1:

```
*A:PE-1# configure router interface "int-loopback-1" shutdown
```

PE-1 now advertises the prefix 10.0.0.0/8 with the community 64500:10:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief | match
                                        "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
Nexthop        : 192.0.2.1
Local Pref.    : 10                     Interface Name : int-RR-5-PE-1
Community      : 64500:10
TieBreakReason : LocalPref
*A:RR-5#
```

RR-5 advertises prefix 10.0.0.0/8 to PE-3 with the next-hop address of PE-2:

```
*A:PE-3# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-PE-3-PE-2
Community      : 64500:50
*A:PE-3#
```

BGP Conditional Route Advertisement Using "none" Prefix List Match

The loopback interfaces on PE-1 are re-enabled:

```
*A:PE-1# configure router interface int-loopback-[1..3] no shutdown
```

The policy on PE-1 is changed so that the prefix 10.0.0.0/8 is advertised with community 64500:100 only if none of the prefixes in the prefix list are active:

```
configure
    router
        policy-options
            begin
            policy-statement "policy-bgp-community"
                entry 10
                    conditional-expression
                        route-exists "[prefix-conditional-routes] none"
                    exit
                exit
            exit
            commit
        exit
```

PE-1 advertises the prefix 10.0.0.0/8 with the community 64500:10, because all loopback interface prefixes are active:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
Nexthop        : 192.0.2.1
Local Pref.    : 10                     Interface Name : int-RR-5-PE-1
Community      : 64500:10
TieBreakReason : LocalPref
*A:RR-5#
```

The loopback interfaces are shut down one by one or together using a range with the following command on PE-1:

```
*A:PE-1# configure router interface int-loopback-[1..3] shutdown
```

PE-1 now advertises the prefix 10.0.0.0/8 with the community 64500:100:

```
*A:RR-5# show router bgp routes 10.0.0.0/8 hunt brief |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-RR-5-PE-1
Community      : 64500:100
Nexthop        : 192.0.2.2
Local Pref.    : 50                     Interface Name : int-RR-5-PE-2
Community      : 64500:50
TieBreakReason : LocalPref
*A:RR-5#
```

RR-5 advertises prefix 10.0.0.0/8 to PE-3 with the next-hop address of PE-1:

```
*A:PE-3# show router bgp routes 10.0.0.0/8 hunt |
                                        match "^Nexthop|Community|Pref" expression
Nexthop        : 192.0.2.1
Local Pref.    : 100                    Interface Name : int-PE-3-PE-1
Community      : 64500:100
*A:PE-3#
```

Conclusion

BGP conditional route advertisement allows the control of BGP updates based on routes in the route table. A conditional policy entry can be created that tests whether any, all, or none of the prefixes in a prefix list are active and executes the related policy actions.

BGP Fast Reroute

This chapter provides information about BGP Fast Reroute.

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS routers and is based on SR OS Release 14.0.R7

Overview

Border Gateway Protocol (BGP) is a key protocol for ISPs, supporting inter-Autonomous System (inter-AS) and intra-Autonomous System (intra-AS) applications with many address families. Additionally, ISPs need to maintain the service level agreements with their customers, even in case of network failures.

MPLS Fast Reroute (FRR) is often used to provide resiliency to intra-AS services, and relies on alternate label switched paths being established through the network. Traffic is switched to the alternate path in case of a failure of the primary path.

However, the traffic for inter-AS services crosses the boundaries of multiple ASs, so to provide resiliency, BGP FRR can be used. Before a network failure occurs, multiple paths must be received for a prefix to take advantage of this feature. When a prefix has a backup path and its primary paths fail, the affected traffic is rapidly diverted to the backup path without waiting for the control plane to reconverge. When many prefixes share the same primary paths, and in some cases also the same backup path, the time to divert traffic to the backup path is independent of the number of prefixes; this is also known as Prefix Independent Convergence (PIC). The traffic goes back to the primary paths when those paths are restored. Multiple primary paths can be active simultaneously when Equal Cost Multi Path (ECMP) applies.

Within SR OS, two BGP FRR functions are supported: Core PIC and Edge PIC. Core PIC describes a scenario where a link or node on the path to the BGP next-hop fails, but the BGP next-hop remains reachable; see Figure 95. Edge PIC describes a scenario where an edge node or edge link fails, which results in a change of the BGP next-hop; see Figure 96.

Figure 95 **Core PIC**

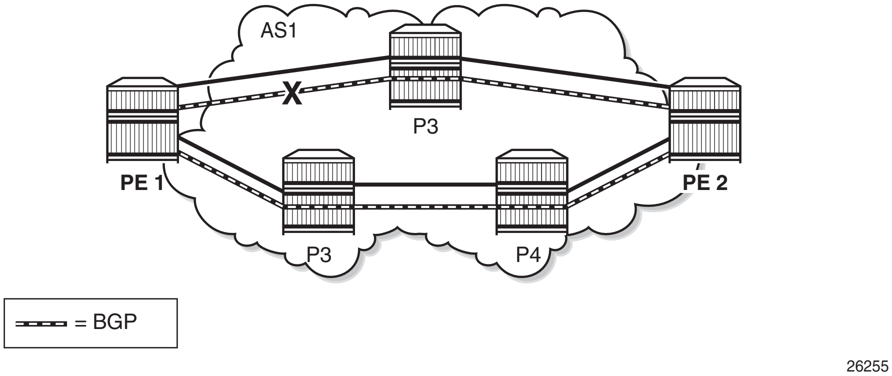
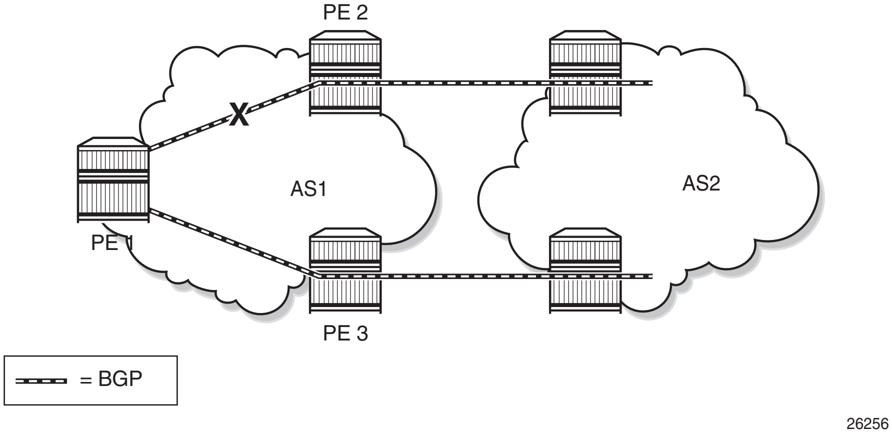


Figure 96 **Edge PIC**



Within SR OS, Core PIC is enabled by default and cannot be disabled. Therefore, this chapter will describe the use of Edge PIC.

BGP FRR is supported for different BGP address families in the base router context or within a specific VPRN context. This chapter will focus on the IPv4 address family within the base router context.

The following SR OS supported features can be used to allow BGP to maintain multiple paths through an autonomous system:

- BGP Best External
- BGP Add-Paths

Convergence goes through several phases, which also apply to BGP:

- detect the network failure
- distribute updated routing information, and update the network topology
- calculate new routes, and optionally change next-hops
- update the forwarding plane

Several mechanisms are available to enhance BGPs network convergence, such as:

- Bidirectional Forwarding Detection (BFD)
- Minimum Router Advertisement Interval (MRAI)
- BGP peer tracking

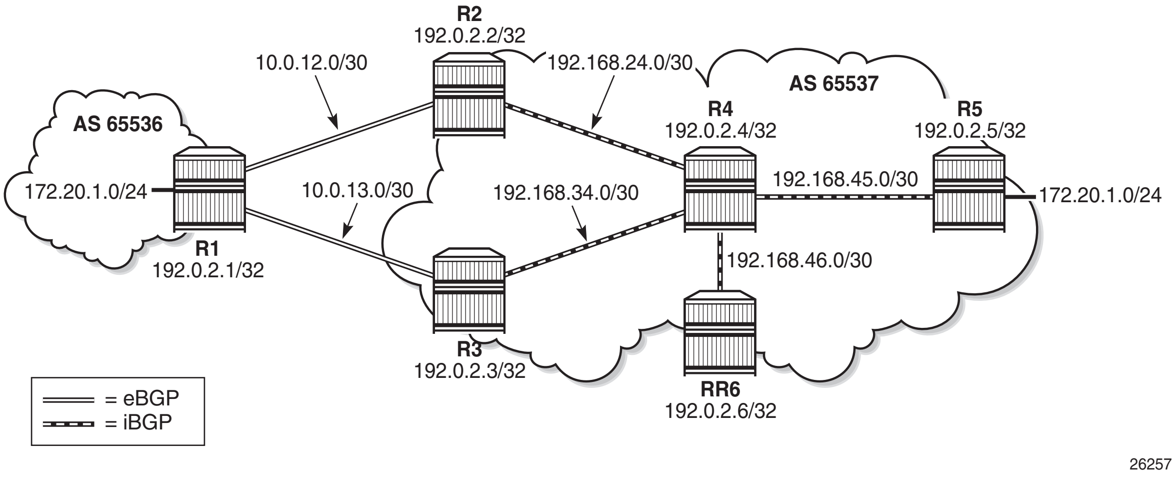
This chapter describes the use of BFD and MRAI for faster network convergence.

Configuration

The topology used in this chapter is shown in Figure 97, and has the following characteristics:

- iBGP sessions are established between AS 65537 routers using RR6 as route reflector with R2, R3, R4, and R5 as route reflector clients.
- eBGP sessions are established between R2 and R3 of AS 65537 and R1 of AS 65536.
- R1 advertises a BGP route for prefix 172.10.1.0/24 to R2 and R3.
- R2 changes the local preference to 150 for the route advertised to its route reflector RR6.
- R2 and R3 advertise a BGP route for prefix 172.20.1.0/24 to R1.

Figure 97 **BGP FRR Topology**



These characteristics enforce traffic for destination 172.10.1.0/24 to leave AS 65537 via R2. R2 (and also R5) learns the destination and the local preference via route reflector RR6. But because R3's own local preference is lower (default LP=100), it stops advertising prefix 172.10.1.0/24 toward RR6, so that R4 is aware of the path via R2 only.

The objective is for R4 to receive multiple copies of the 172.10.1.0/24 prefix with redundant next-hops, to provide for faster convergence under failure. Considering the characteristics previously listed for the topology, two features contribute for achieving this goal:

a. Using BGP Best External
b. Using BGP Add-Paths

The BGP Add-Paths feature is required in scenarios with route-reflectors, possibly combined with the BGP Best External feature. The BGP Best External feature can be used without BGP Add-Paths in scenarios when the BGP peers are in a full mesh.

As a result, multiple exit paths for prefix 172.10.1.0/24 leaving AS 65537 are available, improving convergence time on the iBGP peers because they only need to update their FIBs if they lose the primary route.

BGP Best External

R3 is configured with the BGP Best External feature, as follows:

```
# on R3
configure
    router
        bgp
            loop-detect discard-route
            split-horizon
            advertise-external ipv4
            group "eBGP_AS65536"
                export "AS65537_Export_External_Networks"
                peer-as 65536
                neighbor 192.168.13.1
                exit
            exit
            group "iBGP_AS65537"
                next-hop-self
                peer-as 65537
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
        exit
    exit
exit
```

In this output, advertise-external is activated for the IPv4 address family only. It can also be activated for the IPv6, label-IPv4, and label-IPv6 address families.

Although it is not necessary to also enable BGP Best External on R2, it is not uncommon to also configure this feature on other autonomous system border routers.

Therefore, R3 starts advertising prefix 172.10.1.0/24 toward the route reflector RR6, as follows:

```
*A:R3# show router bgp neighbor 192.0.2.6 advertised-routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     172.10.1.0/24                                      100         None
      192.0.2.3                                          None        -
      65536
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:R3#
```

The BGP Best External feature is sufficient for providing alternate paths in a fully meshed autonomous system, and could be used in conjunction with the BGP Add-Paths feature. The BGP Add-Paths feature is a requirement in scenarios with route reflectors.

BGP Add-Paths

R2, R3, R4 and RR6 are configured with the BGP Add-Paths feature. R5 does not require the Add-Paths feature, because the alternate path to 172.10.1.0/24 starts in R4.

The BGP configuration on R2 is as follows:

```
# on R2
configure
    router
        bgp
            loop-detect discard-route
            split-horizon
            group "eBGP_AS65536"
                local-preference 150
                export "AS65537_Export_External_Networks"
                peer-as 65536
                bfd-enable
                neighbor 192.168.12.1
                exit
            exit
            group "iBGP_AS65537"
                next-hop-self
                peer-as 65537
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The BGP configuration for R3 and R4 is very similar and is not shown here.

The BGP configuration on RR6 then is as follows:

```
# on RR6
configure
    router
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP_AS65537"
                cluster 6.6.6.6
                peer-as 65537
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The default behaviour of a route reflector is to only consider the best path. By enabling the Add-Paths feature on RR6 multiple paths are considered.

Both R2 and R3 start advertising this route to RR6, as follows:

```
*A:R2# show router bgp neighbor 192.0.2.6 advertised-routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     172.10.1.0/24                                      150         None
      192.0.2.2                                          2           -
      65536
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:R2#
*A:R3# show router bgp neighbor 192.0.2.6 advertised-routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                               LocalPref   MED
      Nexthop (Router)                                      Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     172.10.1.0/24                                         100         None
      192.0.2.3                                             2           -
      65536
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:R3#
```

For another example of the BGP Add-Paths feature, see the BGP Multipath chapter.

Backup Path

R4 is the place in the topology where an alternate path is created. The data plane part of the Edge PIC configuration is performed by enabling the **backup-path** command within the BGP context. In this way, BGP considers all alternate paths which are present through the BGP Best External and BGP Add-Paths feature. In the following, backup-paths are considered for the IPv4 address family only, but the IPv6, label-IPv4, and label-IPv6 address families are allowed too.

```
# R4
configure
    router
        bgp
            loop-detect discard-route
            split-horizon
            backup-path ipv4
            group "iBGP_AS65537"
                peer-as 65537
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
        exit
    exit
exit
```

In the default BGP behavior, without the **backup-path** command, two BGP routes exist. Both routes are valid, but only the first one is the best path (indicated by ">"), as follows:

```
*A:R4# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.4        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.10.1.0/24                                      150         None
      192.0.2.2                                          2           -
      65536
*i    172.10.1.0/24                                      100         None
      192.0.2.3                                          2           -
      65536
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:R4#
```

The routing table then is as follows:

```
*A:R4# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.10.1.0/24                                 Remote  BGP       00h17m19s  170
       192.168.24.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:R4#
```

With the **backup-path** command, again both BGP routes are valid; the first route is the best path, and now the second route is explicitly marked to be a backup path (indicated by "b"), as follows:

```
*A:R4# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.4        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
```

```
-------------------------------------------------------------------------------
u*>i  172.10.1.0/24                                      100         None
      192.0.2.2                                          8           -
      65536
ub*i  172.10.1.0/24                                      100         None
      192.0.2.3                                          1           -
      65536
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:R4#
```

Now the routing table is as follows. The "B" flag indicates that a BGP backup path is available.

```
*A:R4# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.10.1.0/24 [B]                             Remote  BGP       00h00m16s  170
       192.168.24.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:R4#
```

To show both routes, use the following command:

```
*A:R4# show router route-table protocol bgp alternative
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
172.10.1.0/24                                 Remote  BGP       00h01m36s  170
       192.168.24.1                                                 0
172.10.1.0/24 (Backup)                        Remote  BGP       00h01m36s  170
       192.168.34.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:R4#
```

The currently active next-hop in the forwarding path is 192.168.24.1, as follows:

```
*A:R4# show router fib 1 172.10.1.0/24 all
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                    Protocol            Installed
  NextHop
-------------------------------------------------------------------------------
172.10.1.0/24                                     BGP                 Y
  192.168.24.1 (int-R4-R2)
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:R4#
```

The active and standby next-hops are also programmed into the forwarding path, as follows:

```
*A:R4#  show router fib 1 172.10.1.0/24 extensive
===============================================================================
FIB Display (Router: Base)
===============================================================================
Dest Prefix             : 172.10.1.0/24
  Protocol              : BGP
  Installed             : Y
  Indirect Next-Hop     : 192.0.2.2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.24.1
      Interface         : int-R4-R2
      ECMP-Weight       : 1
  Indirect Next-Hop     : 192.0.2.3
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Backup-Path         : Yes
    Resolving Next-Hop  : 192.168.34.1
      Interface         : int-R4-R3
      ECMP-Weight       : 1
===============================================================================
Total Entries : 1
===============================================================================
*A:R4#
```

In summary, two paths are available out of R4 and leading to 172.10.1.0/24 in the remote AS, but only one is installed in the forwarding plane. The active route is R4-R2-R1; the backup route is R4-R3-R1. A **traceroute** command confirms the active path, as follows:

```
*A:R5# traceroute no-dns 172.10.1.1 source 172.20.1.1
traceroute to 172.10.1.1 from 172.20.1.1, 30 hops max, 40 byte packets
```

```
  1  192.168.45.1    0.722 ms  0.662 ms  0.646 ms
  2  192.168.24.1    1.22 ms  1.21 ms  1.21 ms
  3  172.10.1.1    3.09 ms  1.78 ms  1.74 ms
*A:R5#
```

Faster Convergence through BFD

As already described, BFD can help speed up BGP convergence, mainly when detecting network failure. In the following, BFD is enabled on the eBGP sessions, and on the IS-IS protocol.

The BFD parameters are defined at interface level, enabling BFD for an application is done in the application context. Because R1 only has eBFD sessions toward R2 and R3, it is enabled at the global BGP level, but it can also be enabled at the group or neighbor level.

```
# for R1
configure
    router
        interface "int-R1-R2"
            address 192.168.12.1/30
            port 1/1/1
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
        interface "int-R1-R3"
            address 192.168.13.1/30
            port 1/1/2
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
    exit
exit
configure
    router
        bgp
            loop-detect discard-route
            min-route-advertisement 1
            bfd-enable
            split-horizon
            backup-path ipv4
            group "eBGP_AS65537"
                export "Export_External_Networks"
                peer-as 65537
                neighbor 192.168.12.2
                exit
                neighbor 192.168.13.2
                exit
            exit
            no shutdown
        exit
    exit
exit
```

Because the BFD configuration for R2 and R3 is very similar, it is only shown for R2, as follows:

```
# for R2
configure
    router
        interface "int-R2-R1"
            address 192.168.12.2/30
            port 1/1/2
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
        interface "int-R2-R4"
            address 192.168.24.1/30
            port 1/1/1
            bfd 100 receive 100 multiplier 3
            no shutdown
        exit
        interface "system"
            address 192.0.2.2/32
            no shutdown
        exit
    exit
exit
```

BFD is enabled for group eBGP_AS65536 only, at group level, as follows:

```
configure
    router
        bgp
            group "eBGP_AS65536"
                export "AS65537_Export_External_Networks"
                peer-as 65536
                bfd-enable
                neighbor 192.168.12.1
                exit
            exit
            group "iBGP_AS65537"
                next-hop-self
                peer-as 65537
                neighbor 192.0.2.4
                exit
            exit
        exit
    exit
exit
```

BFD for IS-IS is enabled at the IS-IS interface level, and is enabled for IPv4 only, as follows.

```
configure
    router
        isis
            area-id 48.0001
            interface "system"
                no shutdown
            exit
            interface "int-R2-R4"
                interface-type point-to-point
                bfd-enable ipv4
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

Faster Convergence through MRAI

Adjusting the BGP MRAI also can help speed up network convergence, using the following command:

```
configure router bgp min-route-advertisement
  - min-route-advertisement <seconds>
  - no min-route-advertisement
 <seconds>            : [1..255]
```

Lowering the MRAI puts a higher load on the CPM, so a tradeoff must be made between convergence time and processing load.

For the example in this chapter, this timer would be set consistently to 1 second on all nodes throughout the network.

Switchover

To demonstrate a switchover scenario, a failure is introduced by shutting down port 1/1/1 on R1, as follows:

```
*A:R1# configure port 1/1/1 shutdown
```

The path through the network is R5-R4-R3-R1, as follows:

```
*A:R5# traceroute no-dns 172.10.1.1 source 172.20.1.1
traceroute to 172.10.1.1 from 172.20.1.1, 30 hops max, 40 byte packets
  1  192.168.45.1    0.698 ms  0.695 ms  0.698 ms
  2  192.168.34.1    1.21 ms  1.21 ms  1.15 ms
  3  172.10.1.1    1.73 ms  1.71 ms  1.70 ms
*A:R5#
```

On R4, traffic is now diverted to R3, and the BGP routes are as follows:

```
*A:R4# show router bgp routes
```

```
===============================================================================
 BGP Router ID:192.0.2.4        AS:65537       Local AS:65537
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.10.1.0/24                                      100         None
      192.0.2.3                                          1           -
      65536
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:R4#
```

The routing table is as follows:

```
*A:R4# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.10.1.0/24                                 Remote  BGP       00h01m41s  170
       192.168.34.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:R4#
```

The forwarding plane is reprogrammed to send traffic for the 172.10.1.0/24 subnet to R3, as follows:

```
*A:R4# show router fib 1 172.10.1.0/24 extensive
===============================================================================
FIB Display (Router: Base)
===============================================================================
Dest Prefix             : 172.10.1.0/24
  Protocol              : BGP
  Installed             : Y
  Indirect Next-Hop     : 192.0.2.3
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.34.1
      Interface         : int-R4-R3
      ECMP-Weight       : 1
===============================================================================
Total Entries : 1
===============================================================================
*A:R4#
```

Bringing port 1/1/1 on R1 up again will result in the path R5-R4-R2-R1 being reactivated. Switchback takes longer, because the external BGP session needs to be re-established, and routes have to be relearned.

Conclusion

BGP FRR provides ISPs the means to offer backup paths with fast switchover times when used in combination with short failure detection times and short advertisement intervals. By guaranteeing service in case of network failures, ISPs can provide enhanced service offerings to their customers.

BGP Fast Reroute Policy Control

This chapter provides information about BGP Fast Reroute Policy Control.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4.

Overview

BGP Fast Reroute (FRR) allows for precomputing multiple redundant BGP paths in the control plane and installing backup routes in the forwarding plane via indirection techniques. See the BGP Fast Reroute chapter for more information.

The BGP FRR Policy Control feature allows for selectively applying FRR for designated BGP prefixes. This allows an operator to develop separate service and redundancy models for different customers or services. It also allows for using data path resources required for BGP FRR in a more efficient way.

The BGP FRR Policy Control feature includes the **install-backup-path** policy action command. This command is supported in the following configuration contexts:

```
config>router>policy-options>policy-statement>default-action
config>router>policy-options>policy-statement>entry>action
```

The **install-backup-path** command is effective when configured in BGP-import or VRF-import policies. In cases where this command is configured in an import policy applied in the global BGP context, the command applies to the following types of routes:

- IPv4

- IPv6
- Label-IPv4
- 6PE
- VPN-IPv4 (only if **vpn-apply-import** is configured in BGP)
- VPN-IPv6 (only if **vpn-apply-import** is configured in BGP)

Figure 98 shows an example of community addition. Two prefixes, 10.1.0.0/16 and 10.2.0.0/16, are advertised by CE-4 to both of its peers, PE-1 and PE-2. The administrator of AS 64496 wants to apply FRR only for the 10.2.0.0/16 prefix that will eventually be advertised to and used on PE-3, and not for 10.1.0.0/16. To facilitate this procedure, an import policy is applied on both PE-1 and PE-2 for routes advertised by CE-4 in AS 64500. The import policy selects and adds a community value of "1:1" to the 10.2.0.0/16 prefix. No community is applied to 10.1.0.0/16.

Figure 98 **Community Addition on PE-1 and PE-2**

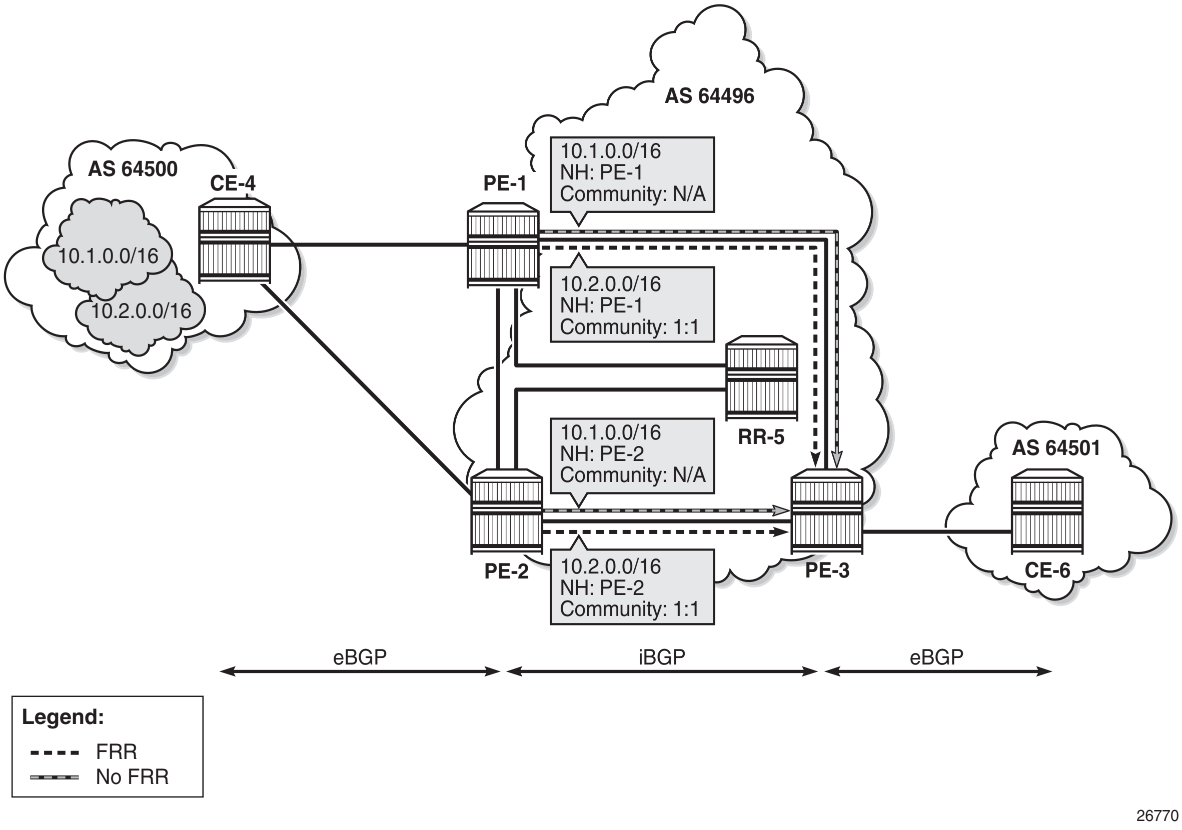
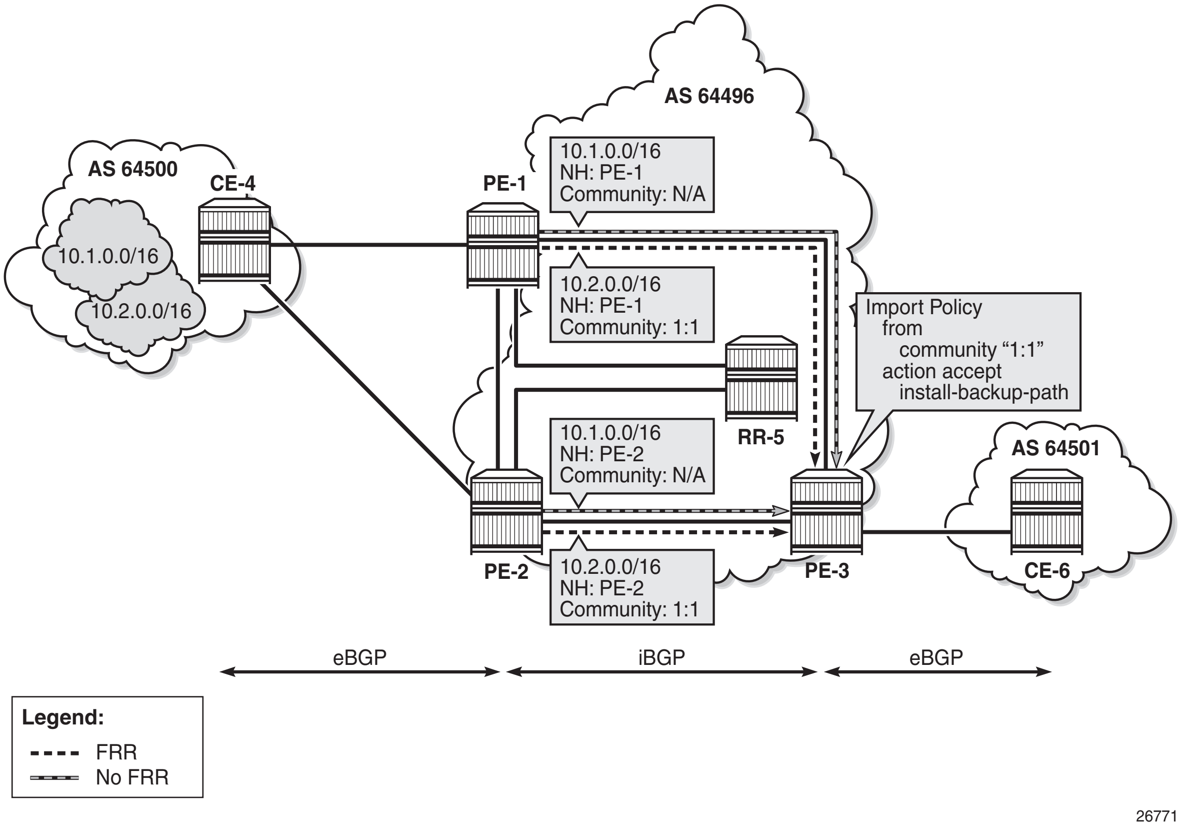


Figure 99 shows the FRR import policy applied on PE-3 for the routes received from PE-1 and PE-2. The policy matches routes with a community value of "1:1" and instructs the router to calculate and install a backup path for those matching routes.

Figure 99 **FRR Policy on PE-3**



Configuration

The following configuration examples are in this section:

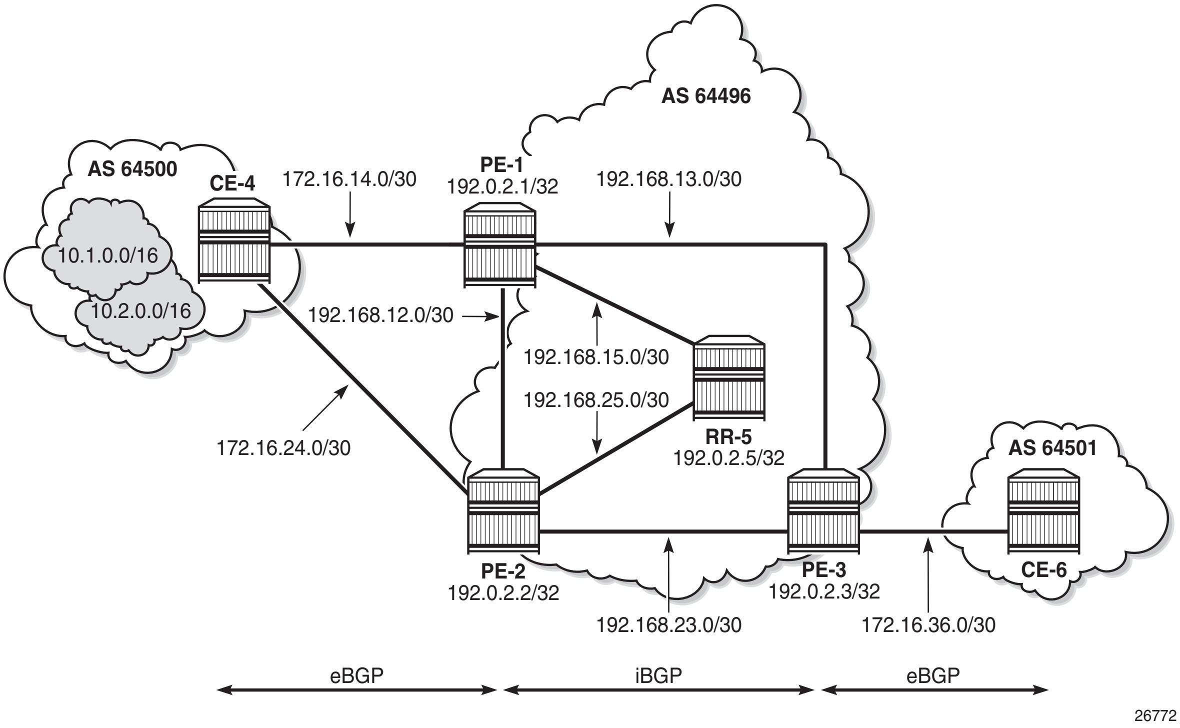
- BGP FRR for address family IPv4 without FRR policy
- BGP FRR for address family IPv4 with FRR policy
- BGP with FRR policy for address family VPN-IPv4 using global BGP policy and **vpn-apply-import**
- BGP with FRR policy for address family VPN-IPv4 using VRF-import policy

BGP FRR Policy Control Feature for Address Family IPv4

Figure 3 shows the example topology used for the BGP FRR Policy Control feature for the IPv4 address family. The topology is similar to the one in the BGP Add-Path chapter, with the following characteristics:

- CE-4 in AS 64500 advertises both prefixes 10.1.0.0/16 and 10.2.0.0/16 to its eBGP peers PE-1 and PE-2 in AS 64496.
- RR-5 is route reflector for all PEs in AS 64496.
- Add-Path is configured on all PE routers and RR-5 with a send-limit of 2.
- CE-6 in AS 64501 peers with PE-3 in AS 64496 and can send traffic to CE-4 in 64500.

Figure 100 **Example Topology - IPv4**



Initial Configuration

The initial configuration on all nodes includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP on all interfaces within AS 64496 (alternatively, OSPF can be used)
- LDP on all interfaces between the PEs in AS 64496, but not toward RR-5. LDP is used to create the transport tunnels that bound to the VPRN services in the VPN-IPv4 address family section.

BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports prefixes 10.1.0.0/16 and 10.2.0.0/16 to both eBGP peers, as follows:

```
# on CE-4
configure
    router
        autonomous-system 64500
        policy-options
            begin
            prefix-list "10.1.0.0/16"
                prefix 10.1.0.0/16 longer
            exit
            prefix-list "10.2.0.0/16"
                prefix 10.2.0.0/16 longer
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        prefix-list "10.1.0.0/16"
                    exit
                    action accept
                    exit
                exit
                entry 20
                    from
                        prefix-list "10.2.0.0/16"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        bgp
            rapid-withdrawal
            split-horizon
            group "eBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 172.16.14.1
                exit
                neighbor 172.16.24.1
                exit
            exit
            no shutdown
        exit
```

CE-4 also has configured the following loopback interfaces:

```
# on CE-4
configure
    router
        interface "int-loopback-1"
            address 10.1.1.1/16
            loopback
            no shutdown
        exit
        interface "int-loopback-2"
            address 10.2.1.1/16
            loopback
            no shutdown
        exit
```

The BGP configuration on CE-6 is similar, except for the export policy.

PE-1 peers with CE-4 in AS 65400 and RR-5 in AS 64496. The BGP configuration on PE-1 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            rapid-withdrawal
            split-horizon
            group "eBGP"
                peer-as 64500
                neighbor 172.16.14.2
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on PE-2 and PE-3 is similar to PE-1.

RR-5 acts as a route reflector to all the PEs in AS 64500 with a cluster ID of 5.5.5.5. The configuration on RR-5 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "iBGP"
                cluster 5.5.5.5
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
```

BGP FRR for Address Family IPv4 without FRR policy

PE-3 receives both prefixes from PE-1 and PE-2 via RR-5, but only uses the one from PE-1 (Nexthop: 192.0.2.1).

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          21          -
      64500
*i    10.1.0.0/16                                        100         None
      192.0.2.2                                          22          -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.1                                          23          -
      64500
*i    10.2.0.0/16                                        100         None
      192.0.2.2                                          24          -
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

The following configuration is applied on PE-3 to enable BGP FRR:

```
*A:PE-3# configure router bgp backup-path ipv4
```

PE-3 calculates and marks BGP routes from PE-2 as backup routes in the BGP routing table:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          21          -
      64500
ub*i  10.1.0.0/16                                        100         None
      192.0.2.2                                          22          -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.1                                          23          -
      64500
ub*i  10.2.0.0/16                                        100         None
      192.0.2.2                                          24          -
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

PE-3 installs BGP routes from PE-2 as backup routes in its route table:

```
*A:PE-3# show router route-table 10.0.0.0/8 longer alternative
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
10.1.0.0/16                                   Remote  BGP       00h00m38s  170
       192.168.13.1                                                 0
10.1.0.0/16 (Backup)                          Remote  BGP       00h00m38s  170
       192.168.23.1                                                 0
10.2.0.0/16                                   Remote  BGP       00h00m38s  170
       192.168.13.1                                                 0
```

```
10.2.0.0/16 (Backup)                          Remote  BGP       00h00m38s  170
       192.168.23.1                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

BGP FRR for Address Family IPv4 with FRR Policy

The global BGP FRR activation command enabled on PE-3 in the previous step is removed from the configuration:

```
*A:PE-3# configure router bgp no backup-path
```

The following command output on PE-3 shows no community values attached to the prefix 10.2.0.0/16 advertised by PE-1 and PE-2:

```
*A:PE-3# show router bgp routes 10.2.0.0/16 detail |
                                     match "^Nexthop |Community" expression
Nexthop        : 192.0.2.1
Community      : No Community Members
Nexthop        : 192.0.2.1
Community      : No Community Members
Nexthop        : 192.0.2.2
Community      : No Community Members
Nexthop        : 192.0.2.2
Community      : No Community Members
*A:PE-3#
```

The following policy is configured on PE-1 and PE-2 to add the BGP community "1:1" to the prefix 10.2.0.0/16 advertised by CE-4:

```
# on PE-1 and PE-2
configure
    router
        policy-options
            begin
            prefix-list "10.2.0.0/16"
                prefix 10.2.0.0/16 longer
            exit
            community "1:1" members "1:1"
            policy-statement "policy-bgp-community"
                entry 10
                    from
                        prefix-list "10.2.0.0/16"
                    exit
                    action accept
                        community add "1:1"
                    exit
                exit
            exit
            commit
```

The policy is applied as a BGP-import policy on PE-1 and PE-2 for the eBGP group:

```
configure router bgp group "eBGP" import "policy-bgp-community"
```

PE-3 now shows the community value associated with prefix 10.2.0.0/16 as applied and advertised by PE-1 and PE-2:

```
*A:PE-3# show router bgp routes 10.2.0.0/16 detail |
                                      match "^Nexthop |Community" expression
Nexthop        : 192.0.2.1
Community      : 1:1
Nexthop        : 192.0.2.1
Community      : 1:1
Nexthop        : 192.0.2.2
Community      : 1:1
Nexthop        : 192.0.2.2
Community      : 1:1
*A:PE-3#
```

The following policy is configured on PE-3 to selectively install a backup path only for prefixes with a community value equal to "1:1":

```
# on PE-3
configure
    router
        policy-options
            begin
            community "1:1" members "1:1"
            policy-statement "policy-bgp-frr-import"
                entry 10
                    from
                        community "1:1"
                    exit
                    action accept
                        install-backup-path
                    exit
                exit
            exit
            commit
```

The policy is applied on PE-3 to selectively install a backup path only for prefixes with a community value equal to "1:1":

```
*A:PE-3# configure router bgp group "iBGP" import "policy-bgp-frr-import"
```

The following command output shows PE-3 has calculated a BGP FRR path only for prefix 10.2.0.0/16 indicated by the "b" (backup) flag:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.0.0/16                                        100         None
      192.0.2.1                                          21          -
      64500
*i    10.1.0.0/16                                        100         None
      192.0.2.2                                          22          -
      64500
u*>i  10.2.0.0/16                                        100         None
      192.0.2.1                                          23          -
      64500
ub*i  10.2.0.0/16                                        100         None
      192.0.2.2                                          24          -
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

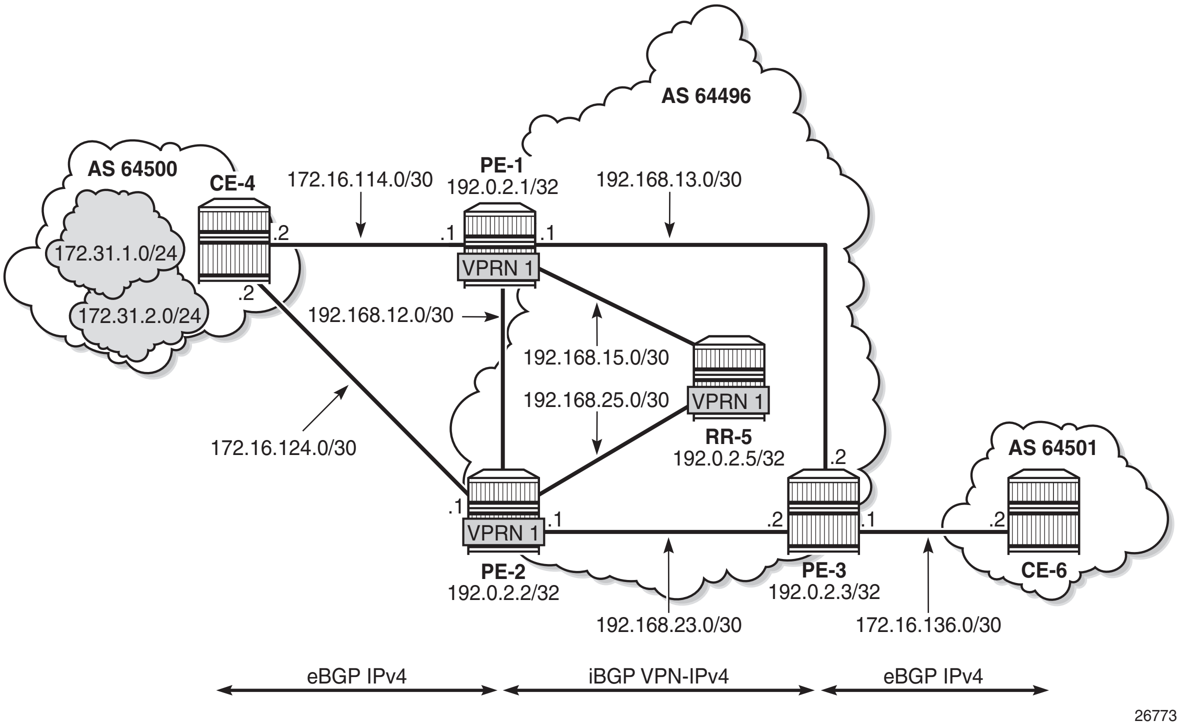
The following command output shows PE-3 has installed a backup route only for prefix 10.2.0.0/16 in its route table:

```
*A:PE-3# show router route-table 10.0.0.0/8 longer alternative
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
10.1.0.0/16                                   Remote  BGP       00h13m20s  170
       192.168.13.1                                                 0
10.2.0.0/16                                   Remote  BGP       00h04m00s  170
       192.168.13.1                                                 0
10.2.0.0/16 (Backup)                          Remote  BGP       00h04m00s  170
       192.168.23.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
```

BGP with FRR Policy for Address Family VPN-IPv4 using Global BGP Policy

Figure 101 shows the example topology used to illustrate the BGP FRR Policy Control feature for VPN-IPv4 route family. The topology used is similar to the one used in the BGP Add-Path chapter. CE-4 exports both prefixes 172.31.1.0/24 and 172.31.2.0/24 to VPRN 1 on PE-1 and PE-2.

Figure 101 **Example Topology - VPN-IPv4**



VPRN 1 is configured on all PEs in AS 64496. The configuration of VPRN 1 is similar on all PEs; for example, for PE-1, the VPRN configuration is as follows:

```
configure
    service
        customer 1 create
            description "Default customer"
        exit
        vprn 1 customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:1
            interface "int-PE-1-CE-4-VPRN1" create
                address 172.16.114.1/30
                sap 1/1/2:1 create
                exit
            exit
            bgp
                split-horizon
                group "eBGP-1"
                    peer-as 64500
                    neighbor 172.16.114.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

On the CEs, the configuration is either in the base routing instance, with additional router interfaces and BGP neighbors, or in a VPRN. In this example, the following VPRN is configured on CE-4:

```
configure
    service
        vprn 1 customer 1 create
            autonomous-system 64500
            route-distinguisher 64500:1
            interface "int-CE-4-PE-1-VPRN1" create
                address 172.16.114.2/30
                sap 1/1/1:1 create
                exit
            exit
            interface "int-CE-4-PE-2-VPRN1" create
                address 172.16.124.2/30
                sap 1/1/2:1 create
                exit
            exit
            interface "loopback1-VPRN1" create
                address 172.31.1.1/24
                loopback
            exit
            interface "loopback2-VPRN1" create
                address 172.31.2.1/24
                loopback
            exit
            bgp
                split-horizon
                group "eBGP-1"
                    export "export-VPRN1"
                    peer-as 64496
                    neighbor 172.16.114.1
                    exit
                    neighbor 172.16.124.1
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

The export policy to export prefixes 172.31.1.0/24 and 172.31.2.0/24 is defined as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "172.31.0.0/16"
                prefix 172.31.0.0/16 longer
            exit
            policy-statement "export-VPRN1"
                entry 10
                    from
                        prefix-list "172.31.0.0/16"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The configuration on CE-6 is similar, but no prefix is exported from CE-6.

For all BGP speakers in AS 64496, BGP must be configured for address family VPN-IPv4 as well as for IPv4, as follows:

```
configure router bgp group "iBGP" family ipv4 vpn-ipv4
```

BGP Add-Path cannot be enabled in the BGP context within a VPRN. However, it can be enabled in the base routing instance for address family VPN-IPv4. This is done on all PEs and RR-5 at group level with the following command:

```
configure router bgp group "iBGP" add-paths vpn-ipv4 send 2
```

The BGP configuration for group iBGP on PE-1 is as follows:

```
configure
    router
        bgp
            group "iBGP"
                family ipv4 vpn-ipv4
                next-hop-self
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                    vpn-ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
```

With Add-Path enabled for address family VPN-IPv4, PE-1 and PE-2 will advertise their routes for prefixes 172.31.1.0/24 and 172.31.2.0/24 as VPN-IPv4 routes to RR-5. RR-5 will advertise both routes to its other RR clients. PE-3 receives two VPN-IPv4 routes for each of the prefixes 172.31.1.0/24 and 172.31.2.0/24, as follows:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4 longer
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.1                                          36          524284
      64500
*>i   64496:1:172.31.1.0/24                              100         None
      192.0.2.2                                          37          524284
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.1                                          38          524284
      64500
*>i   64496:1:172.31.2.0/24                              100         None
      192.0.2.2                                          39          524284
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

The following policy is configured on PE-1 and PE-2 to include the community value "1:1" to prefix 172.31.2.0/24, as well as to the VPRN route target 64496:1 within entry 10. All the other routes are tagged with only the VPRN route target 64496:1 in entry 20.

```
# on PE-1 and PE-2
configure
    router
        policy-options
            begin
            prefix-list "172.31.2.0/24"
                prefix 172.31.2.0/24 longer
            exit
            community "1:1" members "1:1"
            community "target:64496:1" members "target:64496:1"
            policy-statement "policy-export-vprn1"
                entry 10
                    from
```

```
                        prefix-list "172.31.2.0/24"
                    exit
                    action accept
                        community add "1:1" "target:64496:1"
                    exit
                exit
                entry 20
                    from
                    exit
                    action accept
                        community add "target:64496:1"
                    exit
                exit
            exit
            commit
```

The policy is applied as a VRF-export policy in VPRN 1 on PE-1 and PE-2:

```
*A:PE-2# configure service vprn 1 vrf-export "policy-export-vprn1"
```

On PE-3, prefix 172.31.1.0/24 is received with the community value of the VPRN route target only:

```
*A:PE-3# show router bgp routes 172.31.1.0/24 vpn-ipv4 hunt | match "Comm"
Community      : target:64496:1
Community      : target:64496:1
*A:PE-3#
```

However, prefix 172.31.2.0/24 is received with both community values "1:1" and "target:64496:1" from PE-1 and PE-2:

```
*A:PE-3# show router bgp routes 172.31.2.0/24 vpn-ipv4 hunt | match "Comm"
Community      : 1:1 target:64496:1
Community      : 1:1 target:64496:1
*A:PE-3#
```

The following command is applied on PE-3 to make the policy named "policy-bgp-frr-import", configured in the previous section for IPv4 routes, effective also on VPN-IPv4 routes:

```
*A:PE-3# configure router bgp vpn-apply-import
```

PE-3 now has a BGP backup path only for prefix 172.31.2.0/24, as indicated by the "b" (backup) flag:

```
*A:PE-3# show router bgp routes 172.31.0.0/16 vpn-ipv4 longer
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:172.31.1.0/24                              100         None
      192.0.2.1                                          36          524284
      64500
*>i   64496:1:172.31.1.0/24                              100         None
      192.0.2.2                                          37          524284
      64500
u*>i  64496:1:172.31.2.0/24                              100         None
      192.0.2.1                                          38          524284
      64500
ub*>i 64496:1:172.31.2.0/24                              100         None
      192.0.2.2                                          39          524284
      64500
-------------------------------------------------------------------------------
Routes : 4
===============================================================================
*A:PE-3#
```

PE-3 has installed a backup route only for prefix 172.31.2.0/24 in its VPRN route table:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16 longer alternative

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
172.31.1.0/24                                 Remote  BGP VPN   00h01m16s  170
       192.0.2.1 (tunneled)                                         0
172.31.2.0/24                                 Remote  BGP VPN   00h01m16s  170
       192.0.2.1 (tunneled)                                         0
172.31.2.0/24 (Backup)                        Remote  BGP VPN   00h01m16s  170
       192.0.2.2 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

BGP with FRR Policy for Address Family VPN-IPv4 using VRF-Import Policy

The **vpn-apply-import** command enabled in the previous section is removed from the BGP configuration on PE-3:

```
*A:PE-3# configure router bgp no vpn-apply-import
```

PE-3 removes the backup path for prefix 172.31.2.0/24:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16 longer alternative
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
172.31.1.0/24                                 Remote  BGP VPN   00h00m00s  170
       192.0.2.1 (tunneled)                                         0
172.31.2.0/24                                 Remote  BGP VPN   00h00m00s  170
       192.0.2.1 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The following policy is configured to selectively apply FRR for prefixes with a matching community value equal to "1:1" and "target:64496:1" on PE-3:

```
configure
    router
        policy-options
            begin
            community "1:1" members "1:1"
            community "target:64496:1" members "target:64496:1"
            policy-statement "policy-import-vprn"
                entry 10
                    from
                        community expression "[target:64496:1] AND [1:1]"
                    exit
                    action accept
                        install-backup-path
                    exit
                exit
                default-action accept
                exit
            exit
            commit
```

The policy is applied as a VRF-import policy in VPRN 1 on PE-3:

```
*A:PE-3# configure service vprn 1 vrf-import "policy-import-vprn"
```

PE-3 again installs a backup path only for prefix 172.31.2.0/24 and not for 172.31.1.0/24:

```
*A:PE-3# show router 1 route-table 172.31.0.0/16 longer alternative

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
172.31.1.0/24                                 Remote  BGP VPN   00h00m29s  170
       192.0.2.1 (tunneled)                                         0
172.31.2.0/24                                 Remote  BGP VPN   00h00m29s  170
       192.0.2.1 (tunneled)                                         0
172.31.2.0/24 (Backup)                        Remote  BGP VPN   00h00m29s  170
       192.0.2.2 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

Conclusion

The BGP FRR Policy Control feature allows for selectively applying FRR for designated prefixes. The feature brings more flexibility and granularity to the BGP FRR implementation.

BGP FlowSpec for IPv4 and IPv6

This chapter provides information about BGP FlowSpec for IPv4 and IPv6.

Topics in this chapter include:

Applicability

The configuration and information in this chapter are based on SR OS Release 16.0.

Overview

The base BGP Flow Specification (FlowSpec) is defined in RFC 5575 (Flow Specification) and describes a method of encoding IPv4 flow specification information into Network Layer Reachability Information (NLRI). Draft-ietf-idr-flow-spec-v6 followed, and was considered an extension of RFC 5575 to include the IPv6 address family. The flow specification is an n-tuple consisting of one or more matching criteria, which can be applied to IP traffic. The FlowSpec NLRI is encoded into Multiprotocol BGP using MP_REACH_NLRI and MP_UNREACH_NLRI attributes.

As well as the flow specification defining match criteria, extended community attributes are defined to provide traffic filtering actions for the specified flow specification. Therefore, a FlowSpec route (MP_REACH_NLRI) contains a description of the traffic to be matched (using FlowSpec NLRI), and the filtering action to be taken with that traffic (using traffic filtering action extended communities). RFC 7674 provided an update to the original RFC 5575 specification to clarify the formatting of some of these traffic actions, notably redirect to VRF.

The use of FlowSpec is to dynamically distribute traffic filtering rules for mitigating distributed denial of service (DDoS) attacks. A router receiving a FlowSpec update can dynamically create IP filters to mitigate both intra-AS and inter-AS DDoS attacks. Mitigation is implemented by dropping traffic at the ingress point of the network (or nearest possible point toward the source of the DDoS attack) or by redirecting traffic to a separate routing context for forwarding (off-ramping) to a traffic-cleansing device. The ability to redirect traffic led to FlowSpec being considered for software defined networking (SDN)-driven applications or network re-optimization tools. In those cases, a subset of traffic needs to be forced (redirected) into a specific routing context or tunnel/label switched path (LSP) for network capacity optimization or to meet a service level agreement (SLA).

BGP FlowSpec uses AFI 1 (IPv4) or AFI 2 (IPv6) with SAFI 133 (IPv4 dissemination of flow specification rules) or SAFI 134 (VPNv4 dissemination of flow specification rules). Currently, only IPv4 and IPv6 are supported in SR OS.

The FlowSpec NLRI may consist of several components that form the flow specification. A packet only matches the flow specification when it matches all of the components in the NLRI. Table 1 lists the component types that are currently defined, their type values, and their support in SR OS. Flow specification components must follow strict ordering. If present in the specification, a component must precede any other component of higher type value.

Table 7 **FlowSpec Component Types**

Type Value	Component Type	SR OS Support
1	Destination Prefix	Yes
2	Source Prefix	Yes
3	IP Protocol	Partial (TCP/UDP only)
4	Port	Yes (specific number or single contiguous range)
5	Destination Port	Yes
6	Source Port	Yes
7	ICMP Type	Partial (single value only)
8	ICMP Code	Partial (single value only)
9	TCP Flags	SYN/ACK only
10	Packet Length	Yes
11	DiffServ Code Point	Partial (single value only)
12	Fragment	Partial (no support for matching DF bit, first-fragment or last-fragment).

The traffic filtering action for a flow specification uses a number of extended community attributes. The attributes standardized in RFC 5575 are listed in Table 2, with their types and support in SR OS. The traffic rate extended community specifies the rate in bytes per second, where a rate of zero specifies a drop action. The traffic action extended community consists of six bytes; only the two least significant bits of the last byte are currently defined. The terminal action (T-bit), when set to 1, indicates that subsequent filtering rules should be applied (like a next-entry action). When this bit is set to zero, and this action is applied, the evaluation of the traffic filter stops. The sample bit (S-bit), when set to 1, enables traffic sampling and logging for this flow specification. The *redirect-to-vrf* and mark traffic class extended communities are self-explanatory, with a route-target value being used to define the target redirect VRF.

Table 8 **FlowSpec Extended Community Attributes**

Type	Extended Community	SR OS Support
0x8006	Traffic Rate	Yes (0 = drop)
0x8007	Traffic Action S-bit	Yes (uses default filter log 101)
0x8007	Traffic Action T-bit	No
0x8008	Redirect to VRF	Yes
0x8009	Mark Traffic Class	Yes

FlowSpec routes are typically originated and contained within the administrative domain of an operator; particularly when used for DDoS mitigation purposes. This approach means applying ingress filters at the point where traffic enters the autonomous system (AS), such as an external peering point.

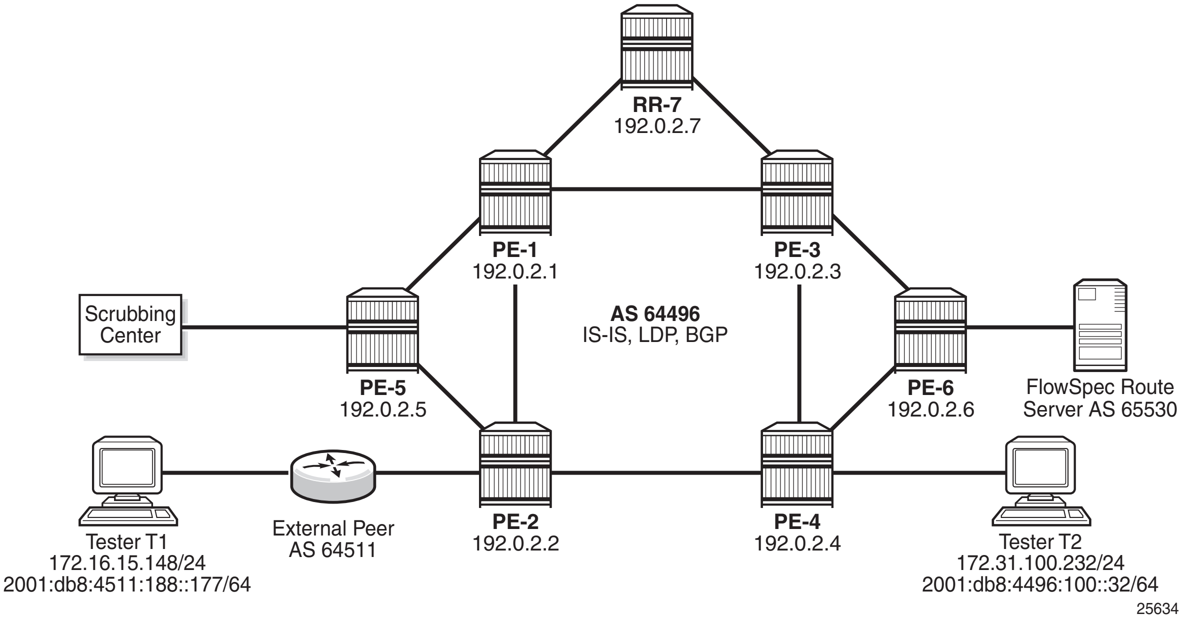
These filters should be instantiated as close as possible to the source of the attack traffic, even if that means applying filters within another operator’s domain. This means that FlowSpec routes must be exchanged between ASs, requiring a trust relationship between the ASs, and a method for validating FlowSpec routes exchanged across AS boundaries. This is covered in the BGP FlowSpec Route Validation chapter.

Example Topology

The example topology used in this chapter is shown in Figure 102. PE-1 through PE-6 and RR-7 participate in IS-IS Level-2 and LDP. All these devices are part of network AS 64496, with all PE routers peering in IBGP with the Route-Reflector RR-7 for address families IPv4, IPv6, VPN-IPv4, VPN-IPv6, Label-IPv4, Label-IPv6, Flow-IPv4, and Flow-IPv6.

By including the Label-IPv4 and Label-IPv6 address families, generating labeled routes, and resolving these labeled routes to LDP tunnels on all PEs in the topology, IPv4 and IPv6 traffic is tunneled in MPLS.

Figure 102 **Example Topology**



To demonstrate FlowSpec, the following items are connected to AS 64496:

- PE-2 is connected to an external peer in AS 64511, which advertises the IPv4 prefix 172.16.0.0/20 and the IPv6 prefix 2001:db8:4511::/48 in EBGP. Both prefixes are advertised within AS 64496 by PE-2 as labeled routes.
- PE-4 advertises IPv4 prefix 172.31.100.0/24 and IPv6 prefix 2001:db8:4496::/48 into IBGP, which PE-2 subsequently advertises in EBGP to AS 64511.

- Tester T1 is connected to the external peer in AS 64511 and sources and sinks traffic from IPv4 address 172.16.15.148 and IPv6 address 2001:db8:4511:188::177. Tester T2 is connected to PE-4 and sources and sinks traffic from IPv4 address 172.31.100.232 and IPv6 address 2001:db8:4496:100:32.
- PE-6 externally peers with a FlowSpec route server belonging to AS 65530.
- PE-5 connects to a DDoS scrubbing center with two interfaces:
  - A "dirty" interface for forwarding of mitigated traffic toward the scrubbing center for cleansing. This interface is connected to an off-ramp VPRN configured on PE-5 and PE-2. PE-5 has static IPv4/IPv6 default routes toward the scrubbing center, which are subsequently advertised into the off-ramp VPRN. This provides sufficient routing information to attract redirected traffic from PE-2 toward the scrubbing center for cleansing.
  - A "clean" interface for traffic received from the scrubbing center after it has been cleansed. This interface is connected to an IES service and is therefore routed toward its destination using the Global Routing Table (GRT).

Configuration

As an example of FlowSpec configuration, the following output shows the BGP configuration on PE-1. Similar configurations are applied to all other PE routers. All PE routers within AS 64496 peer as clients with RR-7 for the address families IPv4, IPv6, VPN-IPv4, VPN-IPv6, Label-IPv4, Label-IPv6, Flow-IPv4, and Flow-IPv6. The Label-IPv4 and Label-IPv6 address families are required for labeled routes, and the resolution filter enables IPv4 and IPv6 traffic to pass through the MPLS/LDP transport tunnels. The Flow-IPv4 and Flow-IPv6 address families are required for propagating the FlowSpec routes, and represent the only part of the BGP configuration required by FlowSpec.

```
configure
    router
        bgp
            loop-detect discard-route
            advertise-inactive
            split-horizon
            next-hop-resolution
                labeled-routes
                    transport-tunnel
                        family label-ipv4
                            resolution-filter
                                ldp
                            exit
                            resolution filter
                        exit
                        family label-ipv6
                            resolution-filter
                                ldp
                            exit
                            resolution filter
                        exit
                    exit
                exit
            exit
            group "IBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 flow-ipv4 flow-ipv6
                                                          label-ipv4 label-ipv6
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.7
                exit
            exit
            no shutdown
        exit
```

PE-2 peers with AS 64511 through an IES service interface using the IPv4 and IPv6 address families, with a dedicated BGP session for each family. This external peering point is the point where the IPv4 and IPv6 filters embedding the flowspec filters are applied. In the following output, these filters are applied in the SAP ingress context, to enable FlowSpec for IPv4 and IPv6, respectively. Such filters can also be enabled on spoke-SDPs within routed interfaces, and is supported within the base and VPRN routing instances.

```
configure
    service
         ies 1 name "FlowSpec-test-10" customer 1 create
            interface "to-AS64511" create
                address 192.168.2.1/30
                ipv6
                    address 2001:db8:2c0d:2121::1/120
                exit
                sap 1/1/4 create
                    ingress
                        filter ip 1
                        filter-ipv6 1
                    exit
                exit
            exit
            no shutdown
        exit
```

FlowSpec Operation

With FlowSpec enabled and configured as in previous section, FlowSpec routes can be advertised to dynamically trigger the instantiation of embedded filters. When valid FlowSpec routes are received, the FlowSpec filters are created. These FlowSpec filters must be referenced from the operator-defined IPv4 or IPv6 filters, for example as follows. These operator-defined filters must be applied to the interfaces in the ingress context for FlowSpec to work.

```
configure
    filter
        ip-filter 1 create
            default-action forward
            embed-filter flowspec router "Base"
        exit
        ipv6-filter 1 create
            default-action forward
            embed-filter flowspec router "Base"
        exit
```

This section demonstrates the use of FlowSpec for traffic black-holing and traffic redirection for both IPv4 and IPv6.

IPv4 FlowSpec

To validate the instantiation of ingress filters based on IPv4 FlowSpec routes, a bidirectional traffic stream is started between T1 (172.16.15.148) in AS 64511 and T2 (172.31.100.232) in AS 64496. In the T1 to T2 direction, the destination port is TCP port 4191.

An IPv4 FlowSpec route is generated to black-hole/drop traffic with a source address of 172.16.15.148 (T1) and a destination address of 172.31.100.232 (T2), for any destination ports in the range 4190-4199. The following output shows the route as received at PE-2.

```
1 2018/06/26 16:46:22.060 CEST MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 77
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family FLOW_IPV4
        NLRI len: 22
          dest_pref   172.31.100.232/32
          src_pref    172.16.15.148/32
          ip_proto    [ == 6 ]
          dest_port   [ >4190 ] and [ <4199 ]
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 65530 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.6
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        192.0.2.7
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        rate-limit: 0 kbps
"
```

The route is shown as an MP_REACH_NLRI for address family Flow-IPv4 (AFI 1 SAFI 133). The NLRI uses the source and destination prefixes, the IP protocol, and the destination-port components to describe the flow and create the filter match criteria. The traffic rate extended community is then used to define a rate of 0, which is the filter drop action.

SR OS does not support a non-zero traffic rate. Although the route will be accepted, no filters will be instantiated.

Unlike other address families, there is no strict requirement for the Next-Hop attribute to be present in the MP_REACH_NLRI. The Length of Next-Hop in the Address field can optionally be set to zero and should be ignored on receipt.

The received FlowSpec route can also be verified in the RIB, which provides a concise output of the flow attributes and traffic filtering function, as follows:

```
*A:PE-2# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 100             None
      65530

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.31.100.232/32
      Src Pref  : 172.16.15.148/32
      Ip Proto  : [ == 6 ]
      Dest Port : [ >4190 ] and [ <4199 ]
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The dynamically created FlowSpec IPv4 ingress filter is identified as *fSpec-0*, as follows. The origin indicates entry 256 has been added by BGP Flowspec.

```
*A:PE-2# show filter ip "fSpec-0"

===============================================================================
IP Filter
===============================================================================
Filter Id           : fSpec-0
Scope               : Embedded
Entries             : 1 (insert By Bgp)
Description         : IPv4 BGP FlowSpec filter for the Base router
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by BGP FlowSpec
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.15.148/32
Src. Port           : n/a
Dest. IP            : 172.31.100.232/32
Dest. Port          : 4191..4198
Protocol            : 6                            Dscp           : Undefined
ICMP Type           : Undefined                    ICMP Code      : Undefined
Fragment            : Off                          Src Route Opt  : Off
Sampling            : Off                          Int. Sampling  : On
IP-Option           : 0/0                          Multiple Option: Off
Tcp-flag            : (Not Specified)
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-2#
```

The configuration of filter 1 (embedding the *fSpec-0* filter) is as follows, and shows a count of ingress matches, which are dropped. This is verified with the loss of traffic in the direction from T1 to T2, but not in the reverse direction.

```
*A:PE-2# show filter ip 1

===============================================================================
IP Filter
===============================================================================
Filter Id           : 1                            Applied        : Yes
Scope               : Template                     Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
Filter Name         : 1
```

```
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry              : 256
Origin             : Inserted by embedded filter fSpec-0 entry 256
Description        : (Not Specified)
Log Id             : n/a
Src. IP            : 172.16.15.148/32
Src. Port          : n/a
Dest. IP           : 172.31.100.232/32
Dest. Port         : 4191..4198
Protocol           : 6                             Dscp           : Undefined
ICMP Type          : Undefined                     ICMP Code      : Undefined
Fragment           : Off                           Src Route Opt  : Off
Sampling           : Off                           Int. Sampling  : On
IP-Option          : 0/0                           Multiple Option: Off
Tcp-flag           : (Not Specified)
Option-pres        : Off
Egress PBR         : Disabled
Primary Action     : Drop
Ing. Matches       : 1067 pkts (136576 bytes)
Egr. Matches       : 0 pkts
===============================================================================
*A:PE-2#
```

When the route is withdrawn and PE-2 receives an MP_UNREACH_NLRI for the same FlowSpec NLRI, the dynamically created filter entries are removed and all associated hardware resources (TCAM entries) are released.

Instead of dropping traffic at the ingress point to the network, an alternative option is to redirect the mitigated traffic to a traffic-cleansing device, if this infrastructure exists. FlowSpec has the *redirect-to-vrf* extended community for this purpose, with the process of forwarding traffic toward a scrubbing center frequently referred to as off-ramping. At PE-2, a VPRN is configured to off-ramp traffic toward the scrubbing center connected to PE-5, as shown in the following output.

In the case of FlowSpec, traffic redirection is half-duplex. That is, traffic is forwarded from PE-2 toward PE-5, but not from PE-5 toward PE-2. This is because when the traffic has been cleansed, it re-enters the network at PE-5 within an IES, and is therefore routed toward its destination using the GRT. This process is frequently referred to as on-ramping. As a result of this half-duplex traffic flow, only a vrf-target import statement is required. There is no requirement to export any routes from PE-2.

```
# on PE-2
configure
    service
        vprn 2 name "FlowSpec-OffRamp-VRF" customer 1 create
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution any
            exit
            vrf-target import target:64496:2
            no shutdown
        exit
```

Off-ramping traffic also requires a VPRN service instance in PE-5 with a single SAP toward the scrubbing center, as shown in the following output. Static IPv4 and IPv4 default routes are configured with next hops of the scrubbing center and these are advertised into VPN-IPv4/VPN-IPv6 using route-policy. There is no requirement for PE-5 to import any BGP-VPN routes.

```
# on PE-5
configure
    service-name
        vprn 2 name "FlowSpec-OffRamp-VRF" customer 1 create
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution any
            exit
            vrf-export “vrf2-export”
            interface "OffRamp-to-Scrubbing-Center" create
                address 192.168.2.5/30
                ipv6
                    address 2001:db8:1b0c:2121::4/127
                exit
                sap 1/2/1 create
                exit
            exit
            static-route-entry 0.0.0.0/0
                next-hop 192.168.2.6
                    no shutdown
                exit
            exit
            static-route-entry ::/0
                next-hop 2001:db8:1b0c:2121::5
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

On-ramping the traffic back onto the network after cleansing the traffic is via IES 3, which is configured as follows. This way the cleansed traffic re-enters the network and is forwarded toward its destination using the GRT.

```
# on PE-5
configure
    service
        ies 3 name "FlowSpec-OnRamp-IES" customer 1 create
            interface "OnRamp" create
                address 192.168.2.9/30
                ipv6
                    address 2001:db8:1b0c:2121::6/127
                exit
                sap 1/2/2 create
                exit
            exit
            no shutdown
        exit
```

To validate the instantiation of the redirection filter, the same bidirectional traffic stream is started between T1 (172.16.15.148) in AS 64511 and T2 (172.31.100.232) in AS 64496. In the T1 to T2 direction, the destination port is TCP port 4191. When the IPv4 FlowSpec route is received at PE-2, the NLRI shows the same traffic match criteria previously used for the black-hole/drop scenario. The extended community has changed to *redirect-to-vrf* with a route-target value of *64496:2*, as shown in the following output.

```
3 2018/06/26 16:49:18.060 CEST MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 77
    Flag: 0x90 Type: 14 Len: 28 Multiprotocol Reachable NLRI:
        Address Family FLOW_IPV4
        NLRI len: 22
          dest_pref   172.31.100.232/32
          src_pref    172.16.15.148/32
          ip_proto    [ == 6 ]
          dest_port   [ >4190 ] and [ <4199 ]
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 65530 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.6
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        192.0.2.7
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        redirect-to-vrf:64496:2
"
```

The dynamically created FlowSpec IPv4 ingress filter is identified as *fSpec-0*, as follows. The filter match criteria for entry 256 indicate the primary action is *forward (VRF)*, and the forwarding router/service ID is service ID 2 (the off-ramp VPRN)

```
*A:PE-2# show filter ip "fSpec-0"

===============================================================================
IP Filter
===============================================================================
Filter Id           : fSpec-0
Scope               : Embedded
Entries             : 1 (insert By Bgp)
Description         : IPv4 BGP FlowSpec filter for the Base router
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by BGP FlowSpec
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.15.148/32
Src. Port           : n/a
Dest. IP            : 172.31.100.232/32
Dest. Port          : 4191..4198
```

```
Protocol            : 6                             Dscp           : Undefined
ICMP Type           : Undefined                     ICMP Code      : Undefined
Fragment            : Off                           Src Route Opt  : Off
Sampling            : Off                           Int. Sampling  : On
IP-Option           : 0/0                           Multiple Option: Off
Tcp-flag            : (Not Specified)
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Forward (VRF)
  Router            : 2
  Extended Action   : None
PBR Down Action     : Drop (entry-default)
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-2#
```

The configuration of filter 1 (embedding the *fSpec-0* filter) shows a count of ingress matches, and is as follows:

```
*A:PE-2# show filter ip 1

===============================================================================
IP Filter
===============================================================================
Filter Id           : 1                             Applied        : Yes
Scope               : Template                      Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
Filter Name         : 1
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by embedded filter fSpec-0 entry 256
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.15.148/32
Src. Port           : n/a
Dest. IP            : 172.31.100.232/32
Dest. Port          : 4191..4198
Protocol            : 6                             Dscp           : Undefined
ICMP Type           : Undefined                     ICMP Code      : Undefined
Fragment            : Off                           Src Route Opt  : Off
Sampling            : Off                           Int. Sampling  : On
IP-Option           : 0/0                           Multiple Option: Off
Tcp-flag            : (Not Specified)
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Forward (VRF)
  Router            : 2
  Extended Action   : None
```

```
PBR Down Action      : Drop (entry-default)
Ing. Matches         : 1101 pkts (140928 bytes)
Egr. Matches         : 0 pkts

===============================================================================
*A:PE-2#
```

Traffic is correctly received in the T1 to T2 direction, and also in the reverse direction. However, traffic in the T1 to T2 direction is redirected by PE-2 toward the scrubbing center attached to PE-5, before being forwarded to its destination at PE-4.

IPv6 FlowSpec

To validate the instantiation of ingress filters based on IPv6 FlowSpec routes, a bidirectional traffic stream is commenced between T1 (2001:db8:4511:188::177) in AS 64511 and T2 (2001:db8:4496:100::32) in AS 64496. In the T1 to T2 direction, the destination port is TCP port 4191.

An IPv6 FlowSpec route is generated to black-hole/drop traffic with a source address of 2001:db8:4511:188::177 (T1) and a destination address of 2001:db8:4496:100::32 (T2), for any destination ports in the range 4190-4199. The following output shows the route as received at PE-2.

```
5 2018/06/26 16:52:58.060 CEST MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 103
    Flag: 0x90 Type: 14 Len: 54 Multiprotocol Reachable NLRI:
        Address Family FLOW_IPV6
        NLRI len: 48
          dest_pref   2001:db8:4496:100::32/128 offset 0
          src_pref    2001:db8:4511:188::177/128 offset 0
          ip_proto    [ == 6 ]
          dest_port   [ >4190 ] and [ <4199 ]
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 65530 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.6
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        192.0.2.7
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        rate-limit: 0 kbps
"
```

The route is shown as an MP_REACH_NLRI for address family Flow-IPv6 (AFI 2 SAFI 133). As with the FlowSpec IPv4 example, the NLRI uses the source and destination prefixes, the IP protocol, and the destination-port components to describe the flow and create the filter match criteria. The traffic rate extended community is then used to define a rate of 0, which is equivalent to a filter drop action.

The dynamically created FlowSpec IPv6 ingress filter is identified as *fSpec-0*, as follows. The description indicates entry 256 has been added through BGP Flowspec.

```
*A:PE-2# show filter ipv6 "fSpec-0"

===============================================================================
IPv6 Filter
===============================================================================
Filter Id           : fSpec-0
Scope               : Embedded
Entries             : 1 (insert By Bgp)
Description         : IPv6 BGP FlowSpec filter for the Base router
-------------------------------------------------------------------------------
Filter Match Criteria : IPv6
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by BGP FlowSpec
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 2001:db8:4511:188::177/128
Src. Port           : n/a
Dest. IP            : 2001:db8:4496:100::32/128
Dest. Port          : 4191..4198
Next Header         : 6                          Dscp           : Undefined
ICMP Type           : Undefined                  ICMP Code      : Undefined
Sampling            : Off                        Int. Sampling  : On
Tcp-flag            : (Not Specified)
Fragment            : Off
HopByHop Opt        : Off                        Routing Type0  : Off
Auth Hdr            : Off                        ESP header     : Off
Flow-label          : n/a                        Flow-label Mask: n/a
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-2#
```

The configuration of filter 1 (embedding the *fSpec-0* filter) is as follows, and shows a count of ingress matches, which are dropped (primary action is drop). This is observed with the loss of traffic in the direction from T1 to T2, but not in the reverse direction.

```
*A:PE-2# show filter ipv6 1

===============================================================================
IPv6 Filter
===============================================================================
Filter Id           : 1                            Applied        : Yes
Scope               : Template                     Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
Filter Name         : 1
-------------------------------------------------------------------------------
Filter Match Criteria : IPv6
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by embedded filter fSpec-0 entry 256
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 2001:db8:4511:188::177/128
Src. Port           : n/a
Dest. IP            : 2001:db8:4496:100::32/128
Dest. Port          : 4191..4198
Next Header         : 6                            Dscp           : Undefined
ICMP Type           : Undefined                    ICMP Code      : Undefined
Sampling            : Off                          Int. Sampling  : On
Tcp-flag            : (Not Specified)
Fragment            : Off
HopByHop Opt        : Off                          Routing Type0  : Off
Auth Hdr            : Off                          ESP header     : Off
Flow-label          : n/a                          Flow-label Mask: n/a
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 403 pkts (51584 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-2#
```

The FlowSpec IPv6 route with the drop action is subsequently withdrawn, restoring the traffic flow between T1 and T2.

To off-ramp IPv6 traffic toward the scrubbing center, the same redirect infrastructure is used as in the IPv4 example:

- PE-2 and PE-5 use the same off-ramp VPRN (VPRN 2), which transports both VPN-IPv4 and VPN-IPv6 traffic.
- PE-5 uses the same on-ramp (IES). When traffic is returned from the scrubbing center, PE-5 routes packets toward their destination using the GRT.

An IPv6 FlowSpec route with a *redirect-to-vrf* extended community is then sourced by the FlowSpec route generator. When the route is received at PE-2, the NLRI shows the same traffic match criteria previously used for the IPv6 black-hole/drop scenario. The extended community has changed to *redirect-to-vrf* with a route-target value of *64496:2*, as shown in the following output.

```
7 2018/06/27 08:19:05.060 CEST MINOR: DEBUG #2001 Base Peer 1: 192.0.2.7
"Peer 1: 192.0.2.7: UPDATE
Peer 1: 192.0.2.7 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 103
    Flag: 0x90 Type: 14 Len: 54 Multiprotocol Reachable NLRI:
        Address Family FLOW_IPV6
        NLRI len: 48
          dest_pref   2001:db8:4496:100::32/128 offset 0
          src_pref    2001:db8:4511:188::177/128 offset 0
          ip_proto     [ == 6 ]
          dest_port    [ >4190 ] and [ <4199 ]
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 65530 >
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.6
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        192.0.2.7
    Flag: 0xc0 Type: 16 Len: 8 Extended Community:
        redirect-to-vrf:64496:2
"
```

The dynamically created FlowSpec IPv4 ingress filter is identified as *fSpec-0*, as follows. The filter match criteria for entry 256 indicate the primary action is *forward (VRF)*, and the forwarding router/service ID is service ID 2 (the off-ramp VPRN).

```
*A:PE-2# show filter ipv6 "fSpec-0"

===============================================================================
IPv6 Filter
===============================================================================
Filter Id           : fSpec-0
Scope               : Embedded
Entries             : 1 (insert By Bgp)
Description         : IPv6 BGP FlowSpec filter for the Base router
-------------------------------------------------------------------------------
Filter Match Criteria : IPv6
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by BGP FlowSpec
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 2001:db8:4511:188::177/128
Src. Port           : n/a
Dest. IP            : 2001:db8:4496:100::32/128
Dest. Port          : 4191..4198
Next Header         : 6                              Dscp           : Undefined
ICMP Type           : Undefined                      ICMP Code      : Undefined
Sampling            : Off                            Int. Sampling  : On
Tcp-flag            : (Not Specified)
Fragment            : Off
HopByHop Opt        : Off                            Routing Type0  : Off
Auth Hdr            : Off                            ESP header     : Off
Flow-label          : n/a                            Flow-label Mask: n/a
Egress PBR          : Disabled
Primary Action      : Forward (VRF)
  Router            : 2
  Extended Action   : None
PBR Down Action     : Drop (entry-default)
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

================================================================================
*A:PE-2#
```

The configuration of IPV6 filter 1 (embedding the *fSpec-0* filter) shows a count of ingress matches, and is as follows:

```
*A:PE-2# show filter ipv6 1

================================================================================
IPv6 Filter
================================================================================
Filter Id           : 1                             Applied        : Yes
Scope               : Template                      Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
Filter Name         : 1
--------------------------------------------------------------------------------
Filter Match Criteria : IPv6
--------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by embedded filter fSpec-0 entry 256
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 2001:db8:4511:188::177/128
Src. Port           : n/a
Dest. IP            : 2001:db8:4496:100::32/128
Dest. Port          : 4191..4198
Next Header         : 6                             Dscp           : Undefined
ICMP Type           : Undefined                     ICMP Code      : Undefined
Sampling            : Off                           Int. Sampling  : On
Tcp-flag            : (Not Specified)
Fragment            : Off
HopByHop Opt        : Off                           Routing Type0  : Off
Auth Hdr            : Off                           ESP header     : Off
Flow-label          : n/a                           Flow-label Mask: n/a
Egress PBR          : Disabled
Primary Action      : Forward (VRF)
  Router            : 2
  Extended Action   : None
PBR Down Action     : Drop (entry-default)
Ing. Matches        : 799 pkts (102272 bytes)
Egr. Matches        : 0 pkts

================================================================================
*A:PE-2#
```

Traffic is correctly received in the T1 to T2 direction, and also in the reverse direction. However, traffic in the T1 to T2 direction is redirected by PE-2 toward the scrubbing center attached to PE-5, before being forwarded to its destination at PE-4.

Resource Consumption

Similar to static filters consuming hardware resources, also dynamically instantiated FlowSpec filters consume hardware resources (TCAM entries) on the associated linecards. Therefore, resources must be checked and monitored to ensure that the system operates within its scaling boundaries.

Before the activation of any FlowSpec routes, there are two ingress ACL/QoS entries consumed for IPv4 and another two entries for IPv6, as shown in the following output.

```
*A:PE-2# tools dump system-resources | match
                    expression "Hardware|Ingress ACL/QoS Entries" post-lines 1
Hardware Resource Usage for Slot #1, CardType iom3-xp, Cmplx #0:
                                 |   Total   | Allocated |    Free
         Ingress ACL/QoS Entries |      65536|          2|      65534
        Ingress IPv6 ACL Entries |      28672|          2|      28670
*A:PE-2#
```

A FlowSpec IPv4 route matches on a source/destination IP address only. When the filter is dynamically instantiated, a single ingress ACL/QoS entry is consumed, as shown in the following output.

```
*A:PE-2# tools dump system-resources | match
                    expression "Hardware|Ingress ACL/QoS Entries" post-lines 1
Hardware Resource Usage for Slot #1, CardType iom3-xp, Cmplx #0:
                                 |   Total   | Allocated |    Free
         Ingress ACL/QoS Entries |      65536|          3|      65533
        Ingress IPv6 ACL Entries |      28672|          2|      28670
*A:PE-2#
```

Similarly, a single entry is consumed if the source/destination IP address is used with a source/destination port. However, if ranges are used for port definition, a number of entries are consumed. A FlowSpec IPv4 route matching the source/destination IP address and a destination port range of 4190 to 4199 consumes four entries, as shown in the following output.

```
*A:PE-2# tools dump system-resources | match
                           expression "Hardware|Ingress ACL/QoS Entries" post-lines 1
Hardware Resource Usage for Slot #1, CardType iom3-xp, Cmplx #0:
                                 |   Total   | Allocated |    Free
         Ingress ACL/QoS Entries |      65536|          6|      65530
        Ingress IPv6 ACL Entries |      28672|          2|      28670
*A:PE-2#
```

TCAM entries are not consumed on a per-interface basis. When TCAM entries are consumed on a linecard for a FlowSpec NLRI match criteria, the same criteria can be used for filtering across multiple IP interfaces on the same linecard without consuming additional TCAM entries.

Conclusion

FlowSpec IPv4 and IPv6 provide a dynamic way to activate (and tear down) ingress filters to mitigate against DDoS attacks. SR OS supports a wide range of match criteria (FlowSpec NLRI) coupled with the ability to either drop or redirect mitigated traffic. This offers flexibility not only in what traffic is matched, but also in traffic treatment, depending on the availability of a traffic-cleansing infrastructure.

The ability of FlowSpec to dynamically create and remove filters has some immediate benefits:

- Reduces the likelihood of configuration errors on one or more devices
- Allows for temporary use of hardware resources, which are released when the threat has passed
- Allows for a push configuration from a single point to a potentially large number of network devices, without having to visit each one to configure filters manually.

BGP FlowSpec Route Validation

This chapter provides information about BGP FlowSpec Route Validation.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R7. This chapter describes the BGP FlowSpec route validation as implemented in SR OS Release 15.0.R1, and later.

Overview

BGP FlowSpec refers to the use of BGP to distribute traffic flow specifications for IPv4 or IPv6 routes throughout a network. Flow specifications provide a means to quickly mitigate Distributed Denial of Service (DDoS) attacks. The BGP FlowSpec standard RFC5575 defines a method to define and advertise flow filters to upstream BGP peers via BGP Network Layer Reachability Information (NLRI). FlowSpec Component Types - in the BGP FlowSpec for IPv4 and IPv6 chapter provides the complete list of matching criteria, such as destination prefix, source prefix, IP protocol, destination port, source port, and so on. FlowSpec Extended Community Attributes - lists the actions, such as redirect and set traffic rate, and so on.

BGP flow specifications might be manipulated and sent with malicious intentions. By default, all flow specifications received from iBGP or eBGP peers are accepted with optional validation. In SR OS releases prior to 15.0.R1, the validity was checked only at the time when a FlowSpec route was received from the peer. In SR OS release 15.0.R1, and later, the FlowSpec routes that are in the routing information base (RIB) can become invalid at a later time, depending on the state of the unicast routes.

Draft-ietf-idr-bgp-FlowSpec-oid-03 describes validation procedures for BGP FlowSpec routes in specific route controller, route reflector, and route server scenarios. These recommendations, in combination with the original validation rules mentioned in RFC5575, are all supported in SR OS Release 15.0.R1, and later. The BGP FlowSpec route validation rules are as follows.

- Rule 1: Flowspec routes originated in the same Autonomous System (AS) as the receiving BGP speaker are always considered valid. This is the case when either of the following applies:
  - The AS_PATH and AS4_PATH attributes of the BGP FlowSpec route are empty.
  - The AS_PATH and AS4_PATH attributes of the BGP FlowSpec route do not contain AS_SET and AS_SEQUENCE segments.
- Rule 2: If Rule 1 does not apply, FlowSpec routes originated outside the local AS without a destination prefix subcomponent are always considered valid.
- Rule 3: If Rule 1 does not apply, FlowSpec routes originated outside the local AS with a destination prefix subcomponent are only considered valid if all the following is true:
  - The neighbor AS (the last non-confederation AS in its AS_PATH attribute) of the BGP Flowspec route matches the neighbor AS of the unicast IP route that is the best match of the destination prefix.
  - The neighbor AS of the BGP FlowSpec route matches the neighbor AS of all unicast IP routes that are longer matches of the destination prefix.
  - The best match unicast IP route and all longer match unicast IP routes must be BGP routes, so no static or IGP routes.

BGP FlowSpec route validation in the base router is enabled with the following command.

```
configure router bgp flowspec validate-dest-prefix
```

BGP FlowSpec route validation in a VPRN is enabled as follows.

```
configure service vprn <service-id> bgp flowspec validate-dest-prefix
```

When validate-dest-prefix is enabled, the validation checks must be repeated every time there is a change to the best route or any longer match route of the destination prefix.

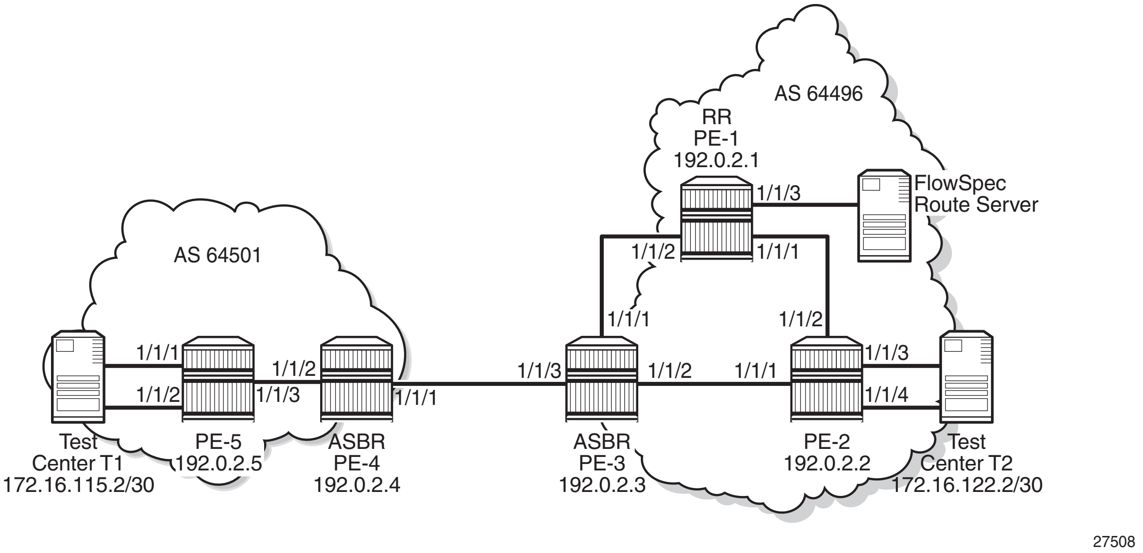
Configuration

In this section, BGP FlowSpec route validation for IPv4 routes in the base router is shown. The action will set the rate to zero, so the matching traffic is dropped. The following use cases will be shown:

- iBGP FlowSpec routes are valid when the AS_PATH attribute is empty. (Rule 1)
- eBGP FlowSpec routes are valid if the best match for the destination prefix is a BGP route toward the neighbor AS from which the BGP FlowSpec route was received (and all longer match unicast IP routes are also toward that AS). (Rule 3)
- eBGP FlowSpec routes are invalid if the best match for the destination prefix is not toward the AS from which the BGP FlowSpec route was received or when the route to the destination prefix is a static or an IGP route instead of a BGP route. (Rule 3)
- eBGP FlowSpec routes without destination prefix subcomponent are valid. (Rule 2)

Figure 103 shows the example topology with a FlowSpec route server in AS 64496 that will advertise iBGP FlowSpec routes to PE-1. Afterward, PE-1 will forward the valid FlowSpec routes to its BGP peers, and so on. Test center T1 in AS 64501 will generate traffic toward test center T2 in AS 64496. This traffic may be filtered by PE-5 when it receives a valid FlowSpec route with the correct matching criteria.

Figure 103 **Example Topology with FlowSpec Route Server in AS 64496**



The initial configuration in the PEs is as follows.

- Cards, MDAs, ports
- Router interfaces
- IGP routing protocol within each AS, but not between the autonomous system border routers (ASBRs) PE-3 and PE-4. It is possible to have OSPF in one AS and IS-IS in the other.

PE-1 is the route reflector (RR) in AS 64496 with clients PE-2 and PE-3. BGP is enabled for the IPv4 and flow-IPv4 address families between the PEs and between PE-1 and the FlowSpec route server. Initially, the FlowSpec route server is in AS 64496, but that will change in a later scenario. The BGP configuration on RR PE-1 is as follows.

```
configure
    router
        bgp
            split-horizon
            group "FlowSpec"
                family ipv4 flow-ipv4
                peer-as 64496
                neighbor 192.168.11.2
                exit
            exit
            group "iBGP"
                family ipv4 flow-ipv4
                cluster 192.0.2.1
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
        exit
```

The BGP configuration on PE-2 includes export policies for the system address 192.0.2.2/32 and the subnet toward the test center T2, 172.16.122.0/30, as follows. The configuration on PE-5 is similar, with export policies for the system address and for subnet 172.16.115.0/30.

```
configure
    router
        policy-options
            begin
            prefix-list "T2"
                prefix 172.16.122.0/28 longer
            exit
            prefix-list "sys"
                prefix 192.0.2.0/29 longer
            exit
            policy-statement "export-T2"
                entry 10
                from
                    protocol direct
                    prefix-list "T2"
                exit
                action accept
                exit
            exit
        exit
        policy-statement "export-sys"
            entry 10
                from
                    protocol direct
                    prefix-list "sys"
                exit
                action accept
                exit
            exit
        exit
        commit
    exit
    bgp
        split-horizon
        group "iBGP"
            family ipv4 flow-ipv4
            export "export-sys" "export-T2"
            peer-as 64496
            neighbor 192.0.2.1
            exit
        exit
```

On ASBR PE-3, the BGP configuration includes an iBGP group and an eBGP group. The BGP IPv4 routes for prefixes 192.0.2.2/32 and 172.16.122.0/30 are inactive within AS 64496, and the ASBR will advertise these inactive routes to its eBGP peer PE-4. The BGP configuration on PE-3 is as follows. The configuration is similar on PE-4.

```
configure
    router
        bgp
            split-horizon
            group "eBGP"
                family ipv4 flow-ipv4
                peer-as 64501
                neighbor 192.168.34.2
                    advertise-inactive
                exit
            exit
            group "iBGP"
                family ipv4 flow-ipv4
                next-hop-self
                peer-as 64496
                neighbor 192.0.2.1
                    advertise-inactive
                exit
            exit
        exit
```

PE-2 and PE-5 both advertise two BGP IPv4 routes: one for the system address and another for the subnet toward the test center. These BGP routes will not be used within the local AS, but they will be advertised by the ASBRs to the peer AS, where these BGP routes will be used. The BGP IPv4 routes on ASBR PE-4 are as follows.

```
*A:PE-4# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.4        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    172.16.115.0/30                                    100         None
      192.0.2.5                                          None        -
      No As-Path
u*>i  172.16.122.0/30                                    None        None
      192.168.34.1                                       None        -
      64496
u*>i  192.0.2.2/32                                       None        None
      192.168.34.1                                       None        -
      64496
*i    192.0.2.5/32                                       100         None
      192.0.2.5                                          None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 4
```

The BGP IPv4 routes on PE-5 are as follows.

```
*A:PE-5# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.5        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.16.122.0/30                                    100         None
      192.0.2.4                                          None        -
      64496
```

```
u*>i  192.0.2.2/32                                         100        None
      192.0.2.4                                            None       -
      64496
-------------------------------------------------------------------------------
Routes : 2
```

No flow specifications have been received and no traffic will be filtered. When traffic is generated by T1 with IP destination address (DA) 172.16.122.2 and IP source address (SA) 172.16.115.2, it is forwarded to T2.

Default Treatment of FlowSpec Routes

The FlowSpec route server announces a FlowSpec IPv4 route to PE-1 with destination prefix 172.16.122.2/30, source prefix 172.16.115.2/30, destination port 4191, source port greater than 1024 as matching criteria, and rate limit 0 kbps (drop) as action. By default, there is no validation check for FlowSpec routes. All FlowSpec routes are considered valid and used, even if no BGP route exists to the destination prefix. All FlowSpec routes are advertised to all PEs, within the AS and to neighbor ASs. On all PEs, the FlowSpec route status codes are valid, best, and used. For example, on PE5:

```
*A:PE-5# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.5        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 100             None
      64496

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.16.122.2/30
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

On all PEs, an embedded IPv4 filter "fSpec-0" will be auto-created for the base router, as follows.

```
*A:PE-5# show filter ip filter-type flowspec

===============================================================================
Flowspec IP Filters                                              Total:     1
===============================================================================
Filter-Id   Scope     Applied Description
-------------------------------------------------------------------------------
fSpec-0     Embedded  N/A     IPv4 BGP FlowSpec filter for the Base router
===============================================================================
*A:PE-5#
```

The details for this embedded filter are retrieved as follows.

```
*A:PE-5# show filter ip "fSpec-0"

===============================================================================
IP Filter
===============================================================================
Filter Id           : fSpec-0
Scope               : Embedded
Entries             : 1 (insert By Bgp)
Description         : IPv4 BGP FlowSpec filter for the Base router
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by BGP FlowSpec
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.115.2/30
Src. Port           : gt 1024
Dest. IP            : 172.16.122.2/30
Dest. Port          : eq 4191
Protocol            : 6                             Dscp            : Undefined
ICMP Type           : Undefined                     ICMP Code       : Undefined
Fragment            : Off                           Src Route Opt   : Off
Sampling            : Off                           Int. Sampling   : On
IP-Option           : 0/0                           Multiple Option: Off
TCP-syn             : Off                           TCP-ack         : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 0 pkts
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-5#
```

This embedded filter "fSpec-0" is created on all PEs, and no traffic is filtered when no IPv4 filter is configured referencing this embedded filter. For this reason, PE-5 has the following IPv4 filter configured and applied on the ingress direction of interface "int-PE-5-T1". The default action is forward; only traffic matching the embedded FlowSpec filter is dropped (rate limit 0 kbps).

```
configure
    filter
        ip-filter 1 create
            default-action forward
            embed-filter flowspec router "Base"
        exit
    info
    exit
    router
        interface "int-PE-5-T1"
            ingress
                filter ip 1
            exit
        exit
```

The following command on PE-5 shows that IPv4 filter 1 contains embedded filter "fSpec-0".

```
*A:PE-5# show filter ip 1 embedded

===============================================================================
IP Filter embedding
===============================================================================
In      Offset  From                   Inserted    Status
-------------------------------------------------------------------------------
1       0       fSpec-0                1/1         OK
===============================================================================
*A:PE-5#
```

Test center T1 generates TCP traffic with IP DA 172.16.122.2, IP SA 172.16.115.2, destination port 4191, and source port 1025. This traffic matches the FlowSpec criteria and will be discarded, because the FlowSpec action is to limit the rate to 0 kbps. The following monitor command on PE-5 shows that the traffic incoming at port 1/1/1 (interface int-PE-5-T1) is dropped instead of being forwarded to port 1/1/3 toward PE-3.

```
*A:PE-5# monitor port 1/1/1 1/1/3 rate interval 3 repeat 2

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
---snip---

At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                            544683                     27
Packets                                             4255                      0
---snip---

Port 1/1/3
-------------------------------------------------------------------------------
```

```
Octets                                                30                     30
Packets                                                0                      0
---snip---
```

The following command shows the IPv4 filter 1 with the filter match criteria. In this example, 67612 packets have matched the filter at the ingress and are dropped, because the primary action in the embedded FlowSpec filter is drop.

```
*A:PE-5# show filter ip 1

===============================================================================
IP Filter
===============================================================================
Filter Id           : 1                            Applied        : Yes
Scope               : Template                     Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by embedded filter fSpec-0 entry 256
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.115.2/30
Src. Port           : gt 1024
Dest. IP            : 172.16.122.2/30
Dest. Port          : eq 4191
Protocol            : 6                            Dscp           : Undefined
ICMP Type           : Undefined                    ICMP Code      : Undefined
Fragment            : Off                          Src Route Opt  : Off
Sampling            : Off                          Int. Sampling  : On
IP-Option           : 0/0                          Multiple Option: Off
TCP-syn             : Off                          TCP-ack        : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 67612 pkts (8654336 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-5#
```

FlowSpec Route Validation

On all PEs, FlowSpec route validation on the destination prefix is enabled within the base router context, as follows.

```
configure router bgp flowspec validate-dest-prefix
```

iBGP FlowSpec Routes

The FlowSpec route server is in AS 64496, so the AS_PATH attribute will be empty when it sends a FlowSpec IPv4 route to iBGP peer PE-1. For this reason, the FlowSpec route is considered valid. The following FlowSpec IPv4 route is received on PE-1 and the status codes are valid, best, and used:

```
*A:PE-1# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 100             None
      No As-Path

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.16.122.2/30
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

PE-1 will forward this valid route to its iBGP peers PE-2 and PE-3, which will also consider this FlowSpec route as valid.

eBGP FlowSpec Routes

Valid eBGP FlowSpec Routes with Destination Prefix

The FlowSpec IPv4 route is not only forwarded to the iBGP peers in AS 64496, but also by PE-3 in AS 64496 to its eBGP peer PE-4 in AS 64501. The eBGP FlowSpec route has a destination prefix subcomponent and it is valid on PE-4 because its neighbor AS (64496) matches the neighbor AS of the unicast IPv4 route that is the best match of destination prefix 172.16.122.2/30. It also matches the neighbor AS of all unicast IPv4 routes that are longer matches of the destination prefix. Also, the best match unicast IPv4 route is a BGP route. The following shows the FlowSpec IPv4 route received by PE-4 as valid, best, and used:

```
*A:PE-4# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 n/a             None
      64496

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.16.122.2/30
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

The following route table entry shows that the best match unicast IPv4 route for destination prefix 172.16.122.0/30 is a BGP route:

```
*A:PE-4# show router route-table 172.16.122.0/30

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
```

```
172.16.122.0/30                               Remote  BGP       00h19m55s  170
       192.168.34.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
```

The BGP IPv4 route for destination prefix 172.16.122.0/30 is as follows. The AS_PATH attribute only contains AS 64496, which is the AS where the FlowSpec IPv4 route originated.

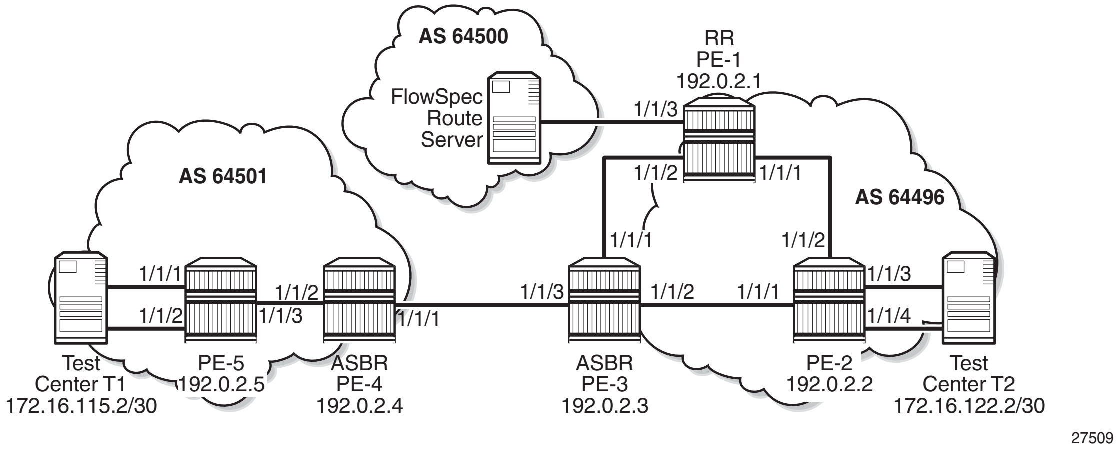
```
*A:PE-4# show router bgp routes 172.16.122.0/30
===============================================================================
 BGP Router ID:192.0.2.4        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.16.122.0/30                                    None        None
      192.168.34.1                                       None        -
      64496
-------------------------------------------------------------------------------
Routes : 1
```

PE-4 will then forward the valid FlowSpec IPv4 route to its iBGP peer PE-5, which will accept the FlowSpec IPv4 route as valid. As a result, an embedded filter "fSpec-0" will be auto-created. When test center T1 sends a traffic flow to T2 with matching criteria, the traffic will be dropped at the ingress port of interface "int-PE-5-T1" on PE-5.

Invalid eBGP FlowSpec Routes with Destination Prefix

Figure 104 shows an example topology with the FlowSpec route server in AS 64500 and the other nodes in the same ASs as before.

Figure 104 **Topology with FlowSpec Route Server in AS 64500**



The BGP configuration on RR PE-1 has been modified with a different peer AS in group "FlowSpec", as follows. FlowSpec validation remains enabled on all routers, so that part of the configuration need not be modified.

```
configure
    router
        bgp
            split-horizon
            group "FlowSpec"
                family ipv4 flow-ipv4
                peer-as 64500
                neighbor 192.168.11.2
                exit
            exit
            group "iBGP"
                family ipv4 flow-ipv4
                cluster 192.0.2.1
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
        exit
```

The FlowSpec route server advertises FlowSpec IPv4 routes to eBGP peer PE-1. When the FlowSpec route server advertises the preceding FlowSpec IPv4 route with IP DA 172.16.122.2/30, the receiving eBGP peer PE-1 will consider the FlowSpec IPv4 route invalid, because the FlowSpec IPv4 route was received from AS 64500 whereas IP prefix 172.16.122.2/30 is within AS 64496 and an IS-IS route to that prefix is available in the route table. The status codes in the following command on PE-1 show that the received FlowSpec IPv4 route is considered invalid.

```
*A:PE-1# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
i     --                  0.0.0.0                 n/a             None
      64500

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.16.122.2/30
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

The following route table on PE-1 shows that an IS-IS route is available toward destination prefix 172.16.122.0/30.

```
*A:PE-1# show router route-table 172.16.122.2

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.122.0/30                               Remote  ISIS      04h41m53s  18
       192.168.12.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 1
```

Invalid routes are not advertised to the BGP peers, so the other nodes will not receive this route. The following BGP summary on PE-1 shows that one FlowSpec IPv4 route was received from the FlowSpec route server, but it remains inactive and no FlowSpec IPv4 route is sent to PE-2 or PE-3.

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
Def. Instance  64496      113    0 00h54m56s 2/0/2 (IPv4)
                          115    0           0/0/0 (FlowIPv4)
192.0.2.3
Def. Instance  64496      113    0 00h54m56s 2/2/2 (IPv4)
                          115    0           0/0/0 (FlowIPv4)
192.168.11.2
Def. Instance  64500        9    0 00h00m36s 0/0/2 (IPv4)
                            8    0           1/0/0 (FlowIPv4)
-------------------------------------------------------------------------------
*A:PE-1#
```

The following command on PE-5 shows that IPv4 filter 1 does not have an embedded filter "fSpec-0".

```
*A:PE-5# show filter ip 1 embedded

===============================================================================
IP Filter embedding
===============================================================================
In      Offset  From                  Inserted    Status
-------------------------------------------------------------------------------
1       0       fSpec-0               0/0         OK
===============================================================================
*A:PE-5#
```

On PE-5, IPv4 filter 1 does not have an embedded filter "fSpec-0" and the default action of IPv4 filter 1 is forward, so the traffic from IP SA 172.16.115.2 to IP DA 172.16.122.2 with destination port 4191 and source port 1025 will be forwarded to T2.

Valid eBGP FlowSpec Routes without Destination Prefix

The FlowSpec route server advertises a FlowSpec IPv4 route for IP traffic with source prefix 172.16.115.2/30, destination port 4191, and source port greater than 1024. No destination prefix subcomponent is included, so the FlowSpec IPv4 route will be considered valid. The following command on PE-1 shows that the FlowSpec IPv4 route without destination prefix subcomponent is valid, best, and used, while an almost identical FlowSpec IPv4 route with destination prefix subcomponent is invalid.

```
*A:PE-1# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 n/a             None
      64500

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
i     --                  0.0.0.0                 n/a             None
      64500

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 172.16.122.2/30
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 2
```

The valid FlowSpec IPv4 route without destination prefix subcomponent will be advertised to the other PEs. The FlowSpec IPv4 route is valid, best, and used on PE-5, as follows.

```
*A:PE-5# show router bgp routes flow-ipv4
===============================================================================
 BGP Router ID:192.0.2.5        AS:64501       Local AS:64501
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV4 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  0.0.0.0                 100             None
      64496 64500

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Src Pref  : 172.16.115.2/30
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

Matching traffic originating from T1 will be discarded on PE-5, as follows.

```
*A:PE-5# monitor port 1/1/1 1/1/2 1/1/3 rate interval 3 repeat 2

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                            540459                     27
Packets                                             4222                      0
---snip---

Port 1/1/3
-------------------------------------------------------------------------------
Octets                                                 0                      0
Packets                                                0                      0
---snip---

*A:PE-5# show filter ip 1

===============================================================================
IP Filter
===============================================================================
Filter Id           : 1                            Applied        : Yes
Scope               : Template                     Def. Action    : Forward
System filter       : Unchained
Radius Ins Pt       : n/a
CrCtl. Ins Pt       : n/a
RadSh. Ins Pt       : n/a
PccRl. Ins Pt       : n/a
Entries             : 0/0/0/1 (Fixed/Radius/Cc/Embedded)
Description         : (Not Specified)
-------------------------------------------------------------------------------
Filter Match Criteria : IP
-------------------------------------------------------------------------------
Entry               : 256
Origin              : Inserted by embedded filter fSpec-0 entry 256
Description         : (Not Specified)
Log Id              : n/a
Src. IP             : 172.16.115.2/30
Src. Port           : gt 1024
Dest. IP            : 0.0.0.0/0
Dest. Port          : eq 4191
Protocol            : 6                            Dscp           : Undefined
ICMP Type           : Undefined                    ICMP Code      : Undefined
Fragment            : Off                          Src Route Opt  : Off
Sampling            : Off                          Int. Sampling  : On
```

```
IP-Option           : 0/0                          Multiple Option: Off
TCP-syn             : Off                          TCP-ack        : Off
Option-pres         : Off
Egress PBR          : Disabled
Primary Action      : Drop
Ing. Matches        : 321617 pkts (41166976 bytes)
Egr. Matches        : 0 pkts

===============================================================================
*A:PE-5#
```

Conclusion

Flow specifications received from iBGP or eBGP peers are by default accepted without validation. Flowspec routes with destination prefix subcomponent can be validated against BGP unicast routing.

BGP Graceful Restart and Long-Lived Graceful Restart

This chapter provides information about BGP Graceful Restart and Long-Lived Graceful Restart.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R8. Before Release 15.0.R1, SR OS supported BGP Graceful Restart (GR) helper procedures for the IPv4, IPv6, VPN IPv4, and VPN-IPv6 address families. In Release 15.0.R1, and later, SR OS additionally supports the label IPv4, label IPv6, L2 VPN, route target, IPv4 FlowSpec, and IPv6 FlowSpec address families. Long-Lived Graceful Restart (LLGR) is supported in SR OS Release 15.0.R4, and later, for the same address families as GR.

Overview

BGP was designed assuming that peer router failures should be reacted to immediately so that the forwarding state of the router can converge toward the current state of the network. However, BGP is often used to signal Network Layer Reachability Information (NLRIs) associated with configuration rather than forwarding, such as flow specifications, Route Target (RT) constraints, BGP Auto-Discovery (BGP-AD), and BGP-VPLS. GR can be applied when there is fate separation between the control plane and the forwarding plane, allowing a restart of the control plane without affecting forwarding.

Table 9 lists the supported address families for GR and LLGR in the base router and in a BGP instance in a VPRN.

Table 9 **Supported Address Families for GR and LLGR in Base Router and in VPRN**

Address Family	AFI/SAFI	BGP in Base Router	BGP in VPRN
IPv4 unicast	1/1	X	X
Labeled IPv4	1/4	X	X
VPN-IPv4	1/128	X	
RT Constraint	1/132	X	
FlowSpec IPv4	1/133	X	X
IPv6 unicast	2/1	X	X
Labeled IPv6	2/4	X	
VPN-IPv6	2/128	X	
FlowSpec IPv6	2/133	X	X
L2 VPN	25/65	X	

GR

GR can be applied in the general BGP context, in a BGP group, or per BGP neighbor. BGP GR can be applied for the base router or a VPRN. GR can be enabled as follows.

```
configure router bgp graceful-restart
configure router bgp group <groupName> graceful-restart
configure router bgp group <groupName> neighbor <neighborName> graceful-restart
configure service vprn <vprnId> bgp graceful-restart
configure service vprn <vprnId> bgp group <groupName> graceful-restart
configure service vprn <vprnId> bgp group <groupName> neighbor <neighborName>
                                                                      graceful-restart
```

The following shows the BGP configuration on the base router for multiple address families. GR is enabled with a stale routes time of 150 seconds and notifications will be sent. No restart time is configured explicitly; the default restart time is 300 seconds at group level and peer level; at BGP instance level, the default restart time is 120 seconds. LLGR is not configured.

```
configure
    router
        bgp
            min-route-advertisement 1
            split-horizon
```

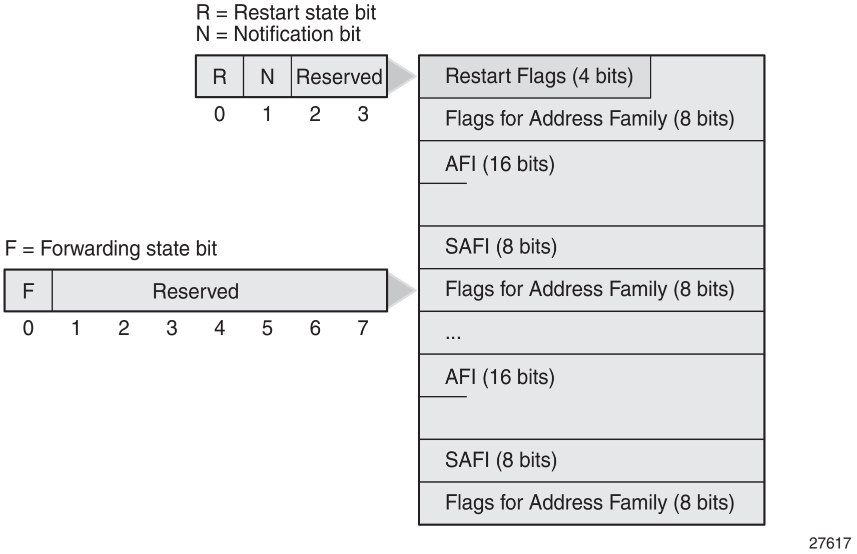
```
group "iBGP"
    family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
    graceful-restart
        stale-routes-time 150
        enable-notification
    exit
    peer-as 64496
    neighbor 192.0.2.1
    exit
exit
no shutdown
```

A BGP speaker can advertise a GR capability to indicate that it is able to preserve its forwarding state per address family (AF) during BGP restart. The GR capability can be used to inform the BGP peers that an end-of-RIB (EOR) message will be generated after all routing updates have been sent for an address family. An EOR message is a withdraw message for an address family without NLRI, as follows:

```
172 2018/04/05 11:49:58.321 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.1
"Peer 1: 192.0.2.1: UPDATE
Peer 1: 192.0.2.1 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 7
    Flag: 0x90 Type: 15 Len: 3 Multiprotocol Unreachable NLRI:
        Address Family VPN_IPV4
No NLRI present! attr len 3, Must be end-of-rib marker!
```

Figure 105 shows the GR capability with restart flags, restart time, and forwarding flags per address family. RFC4724 defines the GR BGP capability. The notification bit N is defined in *draft-ietf-idr-bgp-gr-notification-13*.

Figure 105 **BGP GR Capability**



- Restart flags:
    - The restart state bit R is used to avoid a possible deadlock when multiple BGP speakers peering with each other restart simultaneously and are waiting for the EOR. When set (R=1), the bit indicates that the BGP speaker has restarted and its peer must not wait for the EOR before advertising routing information.
    - The notification bit N indicates that the BGP speaker is willing to send and receive BGP notification messages in GR mode, including the BGP Notification Cease message, which is a hard-reset message causing a peer to terminate a BGP session.
    - The remaining two restart flag bits are reserved and must be set to 0.
- The restart time in seconds is the estimated time required to re-establish a BGP session after a restart. When the restart time expires before the BGP session is re-established, the GR helper stops helping and the (stale) routes received from the failed BGP speaker are removed.
- Flags for address family:
    - The forwarding state bit F indicates whether the forwarding state for routes with a certain AFI/SAFI are preserved during BGP restart. When set (F=1), the forwarding state is preserved. After a hard reset caused by a BGP Notification Cease message, the forwarding bit must be set to 1.
    - The remaining bits are reserved and must be 0.

A BGP speaker can advertise GR capability without any AFI/SAFI, indicating that the sender cannot preserve its forwarding state during BGP restart, but supports procedures for the receiving speaker.

Debugging is enabled for BGP Open messages, as follows:

```
debug router bgp open
```

The following BGP Open message received by PE-2 from PE-1 shows the GR capability for different address families and with a default start timer of 300 seconds. The restart bit R is false because no GR is taking place on peer PE-1. The notification bit N is set to true. The same AFI/SAFI information is presented in the GR capability as in the MP-BGP capabilities, because GR is always enabled for all configured AFI/SAFIs. LLGR is not enabled yet.

```
50 2018/04/20 13:25:41.971 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.1 - Received BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.1: Opt Length 84
   Opt Para: Type CAPABILITY: Length = 82: Data:
     Cap_Code GRACEFUL-RESTART: Length 30
       Bytes: 0x41 0x2c 0x0 0x1 0x1 0x0 0x0 0x2 0x1 0x0 0x0 0x1 0x80 0x0 0x0 0x2 0x8
0 0x0 0x0 0x19 0x41 0x0 0x0 0x1 0x85 0x0 0x0 0x2 0x85 0x0
     Cap_Code MP-BGP: Length 4
```

```
        Bytes: 0x0 0x1 0x0 0x1              # AFI 1/SAFI 1 = IPv4 unicast
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x2 0x0 0x1              # AFI 2/SAFI 1 = IPv6 unicast
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x1 0x0 0x80             # AFI 1/SAFI 128 = VPN-IPv4
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x2 0x0 0x80             # AFI 2/SAFI 128 = VPN-IPv6
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x19 0x0 0x41            # AFI 25/SAFI 65 = L2 VPN
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x1 0x0 0x85             # AFI 1/SAFI 133 = IPv4 FlowSpec
      Cap_Code MP-BGP: Length 4
        Bytes: 0x0 0x2 0x0 0x85             # AFI 2/SAFI 133 = IPv6 FlowSpec
      Cap_Code ROUTE-REFRESH: Length 0
      Cap_Code 4-OCTET-ASN: Length 4
        Bytes: 0x0 0x0 0xfb 0xf0
"
```

The first two octets in the GR capability are 0x41 0x2c (01000001 00101100 in binary). The first four bits-0100-represent the restart flags: R=0, N=1, and the remaining two bits are reserved and set to 0. The remaining twelve bits-000100101100-represent the restart time in seconds: 256+32+8+4=300.

The following four octets in the GR capability are 0x0 0x1 0x1 0x0 (00000000 00000001 00000001 00000000 in binary). The first two octets represent AFI 1 for IPv4, the third octet SAFI 1 for unicast, and the last octet represents the flags, with the forwarding bit F=0 and all other bits reserved and set to zero. The other bytes are for the other AFI/SAFIs that are configured in the example.

Debugging is enabled for GR, as follows:

```
debug router bgp graceful-restart
```

The following messages are in the debug trace on PE-2. The first message shows restart bit R false (no restart ongoing), notification bit N true (GR notifications are supported), restart time 300s (default value), and notification restart false (no GR notifications were sent).

```
51 2018/04/20 13:25:41.971 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer 1: 192.0.2.1: Restart Capability Receive: restart BIT FALSE: Graceful Notificat
ion BIT TRUE: Restart Time 300 secs: NOTIFICATION restart FALSE
"
```

The subsequent messages show the GR capabilities per address family with the value of the forwarding-preserved bit F, for example, for the IPv4 unicast address family, as follows. The forwarding-preserved bit is false.

```
52 2018/04/20 13:25:41.971 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer 1: 192.0.2.1: Restart Capability Receive: afi: AFI_IPV4 safi: SAFI_UNICAST forw
arding-preserved BIT FALSE
```

"

When routers have negotiated the GR capability for an address family and the BGP session drops, the BGP peers enter the GR helper state and do not immediately delete the routes of that address family received from the failed peer. The helpers mark these routes as stale and keep using them until the BGP session is restored, the BGP routes are refreshed, and an EOR message has been received for the AFI/ SAFIs.

However, if the BGP session with the restarting router is not restored before the configured restart time expires, the peer router stops helping and will send withdraw messages for the routes received from the restarting router. When the stale routes time expires, the router will withdraw all routes received from the restarting router. The restart time has an upper bound of 4095 seconds, so this mechanism is designed for relatively short outages in the order of minutes, not for hours. GR can deal with simple control plane restarts in terms of scope and severity.

```
*A:PE-1# configure router bgp graceful-restart restart-time
  - restart-time <seconds>
  - no restart-time

 <seconds>            : [0..4095]
```

LLGR

LLGR can handle failure scenarios where the repair takes several hours, such as a network where redundant route reflectors (RRs) fail simultaneously and the configuration-type BGP routes (that is, non-forwarding BGP routes) for FlowSpec, route target, and L2 VPNs can be preserved. BGP routes for forwarding can also be preserved longer. LLGR can be enabled for all address families that have GR enabled, or for a subset of these address families. LLGR allows a BGP session to stay down for hours or even days. The advertised stale time has an upper bound of 16777215 seconds and the default value is 86400 seconds. LLGR is configured in the GR context, which is in the general BGP context, per group, or per neighbor.

```
*A:PE-1# configure router bgp graceful-restart long-lived advertised-stale-time
  - advertised-stale-time <seconds>
  - no advertised-stale-time

 <seconds>            : [0..16777215]
```

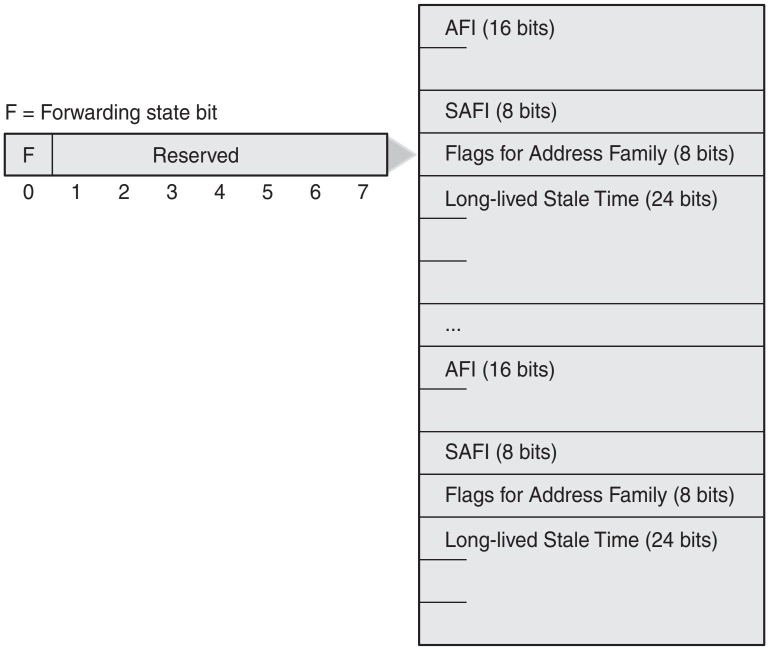
When GR is enabled, it automatically applies for all configured AFs; LLGR can be configured per AF, possibly with different LLGR-stale times, for example, for the L2 VPN address family in group "iBGP", as follows:

```
*A:PE-1# configure router bgp group "iBGP" graceful-restart long-lived family l2-
```

```
vpn advertised-stale-time 7200
```

Figure 106 shows the LLGR capability-as defined in draft-uttaro-idr-bgp-persistence-03-that adds a long-lived stale time per address family. The LLGR capability must be advertised in conjunction with the GR capability.

Figure 106 **LLGR Capability**


F = Forwarding state bit
F
Reserved
0
1
2
3
4
5
6
7
AFI (16 bits)
SAFI (8 bits)
Flags for Address Family (8 bits)
Long-lived Stale Time (24 bits)
...
AFI (16 bits)
SAFI (8 bits)
Flags for Address Family (8 bits)
Long-lived Stale Time (24 bits)
27618

GR and LLGR are configured in the "iBGP" group for all configured AFI/SAFIs, as follows. The default value of the **long-lived advertised-stale-time** is 86400 seconds.

```
configure
    router
        bgp
            group "iBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                graceful-restart
                    stale-routes-time 150
                    enable-notification
                    long-lived
                        advertised-stale-time 3600
                    exit
                exit
            exit
```

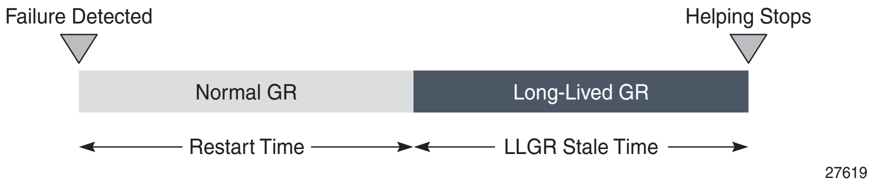
When LLGR is enabled, the BGP Open message contains a long-lived GR capability and a GR capability, with the supported AFI/SAFIs. The following BGP Open message is received by PE-2 from RR PE-1. GR and LLGR are supported for all the AFI/SAFIs in the BGP session.

```
280 2018/04/20 13:56:38.304 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.1 - Received BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.1: Opt Length 135
   Opt Para: Type CAPABILITY: Length = 133: Data:
     Cap_Code GRACEFUL-RESTART: Length 30
       Bytes: 0x41 0x2c 0x0 0x1 0x1 0x0 0x0 0x2 0x1 0x0 0x0 0x1 0x80 0x0 0x0 0x2 0x8
0 0x0 0x0 0x19 0x41 0x0 0x0 0x1 0x85 0x0 0x0 0x2 0x85 0x0
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x1
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x2 0x0 0x1
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x80
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x2 0x0 0x80
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x19 0x0 0x41
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x85
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x2 0x0 0x85
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf0
     Cap_Code LONG-LIVED-GR: Length 49
       Bytes: 0x0 0x1 0x1 0x0 0x0 0xe 0x10 0x0 0x2 0x1 0x0 0x0 0xe 0x10 0x0 0x1 0x80
 0x0 0x0 0xe 0x10 0x0 0x2 0x80 0x0 0x0 0xe 0x10 0x0 0x19 0x41 0x0 0x0 0xe 0x10 0x0 0
x1 0x85 0x0 0x0 0xe 0x10 0x0 0x2 0x85 0x0 0x0 0xe 0x10
"
```

The first seven octets in the LLGR capability-0x0 0x1 0x1 0x0 0x0 0xe 0x10-are for AF IPv4 unicast. The first two-0x0 0x1-represent AFI=1 for IPv4, the third-0x1-represents SAFI=1 for unicast, the fourth-0x0-indicates that the forwarding-preserved bit F=0 (and the other bits are reserved and must be zero). The next three octets represent the LLGR-stale time: 0x0 0xe 0x10 (00000000 00001110 00010000 in binary) is 2048 + 1024 + 512 + 16 = 3600 in decimal.

The LLGR-stale time in seconds specifies how long LLGR-stale routes for the AFI/SAFI may be retained, possibly added to the GR time. LLGR starts when GR terminates before the failed router has recovered, that is, when either the restart timer or the stale-routes timer expires (whichever expires first), as shown in Figure 107. LLGR ends when the advertised LLGR-stale time expires or when the failure is restored and all routes are re-advertised followed by an EOR message. When the AFI/SAFI is not listed in the GR capability, the restart time for GR is 0 seconds. The LLGR-stale time is defined by the **advertised-stale-time** option, which has a default value of 86400 seconds.

Figure 107 **GR and LLGR**



The forwarding-preserved bit F is configured with the following command. By default, all F bits are 0, indicating that the forwarding state was not preserved during the previous restart. The **forwarding-bits-set** command allows F bits for all AFI/SAFIs to be set to 1, or only the F bits for configuration-type (that is, non-forwarding) AFI/ SAFIs, such as L2 VPN, route target, IPv4 FlowSpec, and IPv6 FlowSpec.

```
 *A:PE-1# configure router bgp group "iBGP" graceful-restart long-lived forwarding-
bits-set
  - forwarding-bits-set {all|non-fwd}
  - no forwarding-bits-set
```

An address family is only protected with LLGR if the AFI/SAFI is in the advertised LLGR capability and in the received LLGR capability. In SR OS, LLGR can only be enabled when GR is enabled, so each advertised LLGR capability comes with a GR capability. If a peer advertises the LLGR capability without GR capability, the LLGR capability is ignored.

GR is used for short outages where the helpers pretend that everything is normal; LLGR is for longer outages where the helpers inform the other peers. Table 10 shows a comparison of the helper actions during GR and LLGR.

Table 10 **Helper Actions during GR and LLGR**

Helper actions during GR	Helper actions during LLGR
Mark GR-eligible routes from the failed peer as stale	Mark LLGR-eligible routes from the failed peer as LLGR-stale
Attempt to reconnect to the peer at periodic intervals	Attempt to reconnect to the peer at periodic intervals
	Depreference LLGR-stale routes so that they are less preferred than any valid non-LLGR-stale route
	If an LLGR-stale route remains the best path, inform the other peers by withdrawing the route or re-advertising the route with new attributes

A route is said to be depreferenced if it has its route selection preference reduced in reaction to some event. LLGR automatically depreferences LLGR-stale routes so that any valid non-LLGR-stale route for the same NLRI is more preferred. When advertising LLGR-stale routes to an LLGR-capable peer, LLGR adds the well-known **llgr-stale** BGP community to the routes, so that the LLGR-capable BGP peers can also depreference the LLGR-stale routes. The following option controls how LLGR-stale routes are advertised.

```
*A:PE-1# configure router bgp group "iBGP" graceful-restart long-lived advertise-
stale-to-all-neighbors
  - advertise-stale-to-all-neighbors
  - no advertise-stale-to-all-neighbors

 [no] without-no-exp* - Enable/Disable option to advertise stale routes to neighbors
                        without exporting
```

- The default is **no advertise-stale-to-all-neighbors**, in which case LLGR-aware routers re-advertise stale best routes to their LLGR-aware peers, with the addition of the well-known **llgr-stale** community. Toward BGP peers that did not advertise the LLGR capability, the stale routes are withdrawn.
- When **advertise-stale-to-all-neighbors** is configured combined with the default **no without-no-export**, the LLGR-stale routes are withdrawn toward eBGP peers that did not advertise the LLGR capability and re-advertised to all other peers with LLGR-stale community. Toward iBGP (and confederation-eBGP) peers that signaled the LLGR capability, the route is re-advertised with the well-known **llgr-stale** and **no-export** communities and the local preference is set to 0.
- When **advertise-stale-to-all-neighbors** is configured combined with **without-no-export**, the LLGR-stale routes are withdrawn toward eBGP peers that did not advertise the LLGR capability and re-advertised to all other peers with LLGR-stale community. Toward iBGP (and confederation-eBGP) peers that signaled the LLGR capability, the route is re-advertised with the LLGR-stale community, but without the no-export community. The local preference is set to 0.

Route policies can match, delete, or add the BGP well-known communities **llgr-stale** and **no-llgr**.

An iBGP peer not supporting LLGR normally does not receive route updates with LLGR-stale community, but if it does, it can only depreference them based on local preference 0.

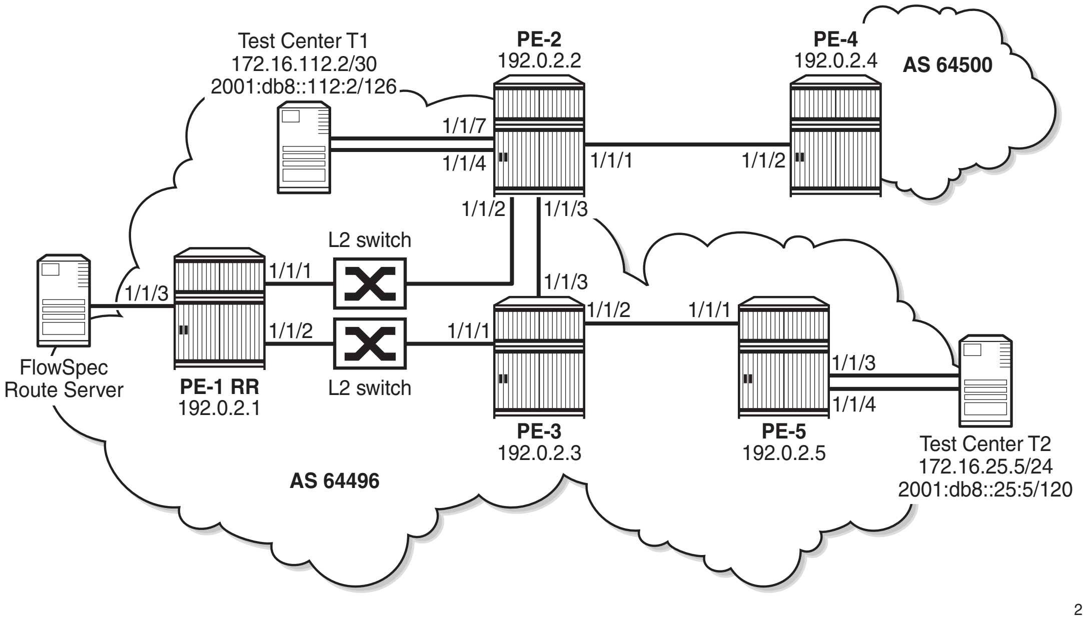
The LLGR-stale routes timer is not stopped when the BGP session with the failed peer is re-established; it only stops when the EOR is received for the AFI/SAFI. When the LLGR- stale routes time expires for an AFI/SAFI, the LLGR phase ends and all remaining LLGR-stale routes for that AFI/SAFI are deleted. However, stale routes will also be deleted before the LLGR stale-routes timer expires when the BGP session with the failed peer is re-established and either of the following applies:

- the GR or LLGR capability is missing
- the AFI/SAFI is missing from the LLGR capability
- the forwarding state bit F=0 for the AFI/SAFI

Configuration

Figure 108 shows the example topology with four routers in AS 64496. PE-1 combines the roles of a PE and an RR. A FlowSpec route server sends IPv4 and IPv6 FlowSpec routes to PE-1. Test centers T1 and T2 generate IPv4 and IPv6 traffic to each other, through the base router or a VPLS service. PE-4 is in AS 64500 and has an eBGP session with PE-2 in AS 64496.

Figure 108 **Example Topology**



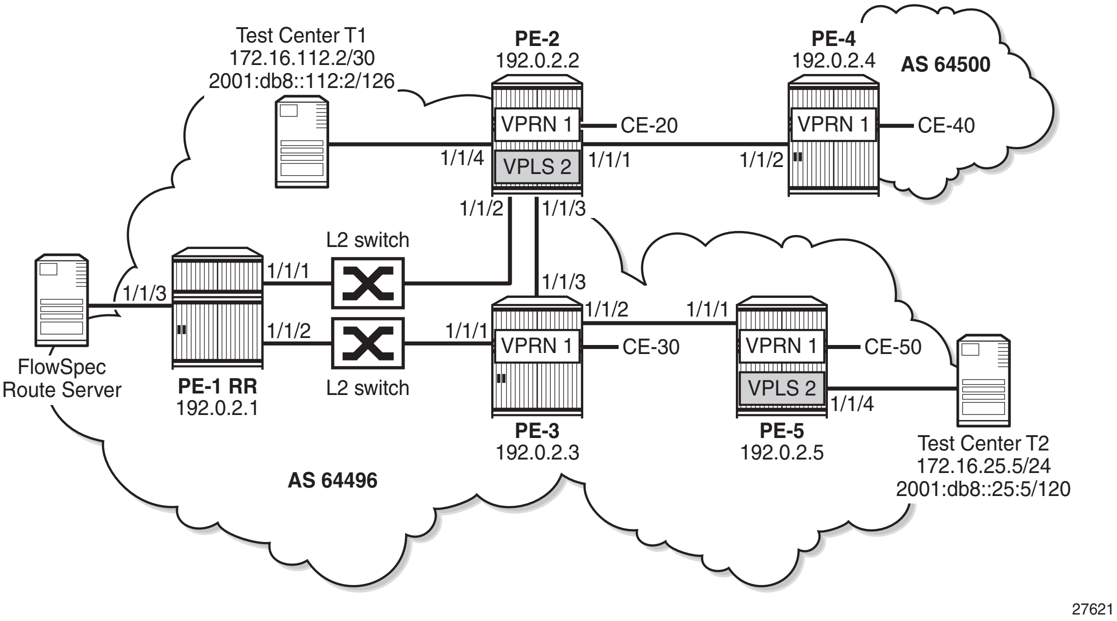
Initial Configuration

The initial configuration on the nodes includes:

- Cards, MDAs, ports
- Router interfaces with dual stack
- IS-IS on all interfaces of the routers in AS 64496 (alternatively, OSPF can be used)
- LDP on all interfaces in AS 64496, not between PE-2 and PE-4
- MPLS and RSVP on all interfaces in AS 64496, not between PE-2 and PE-4
- RSVP-TE LSPs between PE-2 and PE-5

Figure 5 shows the configured services on the PEs. VPRN 1 is configured on PE-2, PE-3, PE-4, and PE-5; VPLS 2 with BGP-AD on PE-2 and PE-5.

Figure 109 **VPRN 1 and VPLS 2 in the Example Topology**



The service configuration on PE-2 is as follows. The pseudowire (PW) template is required for BGP-AD in VPLS 2, as described in the *LDP VPLS Using BGP-Auto Discovery* chapter.

```
configure
    service
        pw-template 1 create
            split-horizon-group "vpls-shg"
            exit
        exit
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:1
            interface "int-VPRN1-PE-2-CE-20" create
                address 172.16.2.1/30
                ipv6
                    address 2001:db8::1:2:1/126
                exit
                sap 1/1/5:1 create
                exit
            exit
            no shutdown
        exit
        vpls 2 customer 1 create
            bgp
                route-distinguisher 64496:2
                route-target export target:64496:2 import target:64496:2
                pw-template-binding 1 import-rt "target:64496:2"
                exit
            exit
            bgp-ad
                vpls-id 64496:2
                vsi-id
                    prefix 192.0.2.2
                exit
                no shutdown
            exit
            sap 1/1/5:2 create
            exit
            sap 1/1/4 create
            exit
            no shutdown
        exit
```

For the exchange of the routes in the VPRN, the VPN IPv4 and VPN IPv6 address families need to be configured in BGP; for BGP-AD, the L2 VPN address family. BGP is configured on all PEs for the following address families: IPv4, IPv6, VPN-IPv4, VPN-IPv6, L2 VPN, IPv4 FlowSpec, and IPv6 FlowSpec. On RR PE-1, the initial BGP configuration is as follows. The "iBGP" group includes all the PEs in AS 64496, whereas the "FlowSpec" group includes the FlowSpec server only.

```
configure
    router
        bgp
            min-route-advertisement 1
            split-horizon
            group "iBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                cluster 192.0.2.1
                peer-as 64496
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.5
                exit
            exit
            group "FlowSpec"
                family ipv4 ipv6 flow-ipv4 flow-ipv6
                peer-as 64496
                neighbor 192.168.11.2
                exit
            exit
        exit
```

On PE-2, the prefixes toward the test center T1 are exported. BGP is configured as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "T1"
                prefix 172.16.112.0/28 longer
                prefix 2001:db8::112:0/124 longer
            exit
            policy-statement "export-T1"
                entry 10
                    from
                        protocol direct
                        prefix-list "T1"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        bgp
            min-route-advertisement 1
            split-horizon
            group "eBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                local-as 64496
                peer-as 64500
                neighbor 192.168.24.2
                exit
            exit
            group "iBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                export "export-T1"
                peer-as 64496
                neighbor 192.0.2.1
                exit
            exit
        exit
```

The configuration on PE-5 is similar, but without the "eBGP" group, and for the export, the prefixes from T2 are included, as follows.

```
configure
    router
        policy-options
            begin
            prefix-list "T2"
                prefix 172.16.225.0/28 longer
                prefix 2001:db8::225:0/124 longer
            exit
            policy-statement "export-T2"
                entry 10
                    from
                        protocol direct
                        prefix-list "T2"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        bgp
            min-route-advertisement 1
            split-horizon
            group "iBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                export "export-T2"
                peer-as 64496
                neighbor 192.0.2.1
                exit
            exit
        exit
```

On PE-3, the BGP configuration is similar, without the export policy.

The BGP configuration on PE-4 is as follows:

```
configure
    router
        bgp
            split-horizon
            group "eBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                local-as 64500
                peer-as 64496
                neighbor 192.168.24.1
                exit
            exit
```

BGP Routes

Under normal conditions, BGP routes of all the configured address families are advertised. The BGP summary on PE-5 shows the following number of received (Rcv), active (Act), and sent (Sent) BGP routes per address family for neighbor 192.0.2.1. Similar numbers occur on the other RR clients PE-2 and PE-3.

```
*A:PE-5# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496       69    0 00h01m25s 1/1/1 (IPv4)
                           21    0           1/0/1 (IPv6)
                                             4/3/2 (VpnIPv4)
                                             4/3/2 (VpnIPv6)
                                             1/1/1 (L2VPN)
                                             1/1/0 (FlowIPv4)
                                             1/1/0 (FlowIPv6)

-------------------------------------------------------------------------------
*A:PE-5#
```

On PE-2, the following BGP IPv4 route is valid, best, and used.

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.16.225.0/30                                    100         None
      192.0.2.5                                          None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

On PE-2, the following BGP L2 VPN route received from neighbor PE-1 (RR) is valid, best, and used.

```
*A:PE-2# show router bgp neighbor 192.0.2.1 received-routes l2-vpn
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP L2VPN Routes
===============================================================================
Flag  RouteType                   Prefix                                MED
      RD                          SiteId                                Label
      Nexthop                     VeId                   BlockSize   LocalPref
      As-Path                     BaseOffset             vplsLabelBa
                                                         se
-------------------------------------------------------------------------------
u*>i  AutoDiscovery               192.0.2.5              -           0
      64496:2                     -                                  -
      192.0.2.5                   -                      -           100
      No As-Path                  -                      -
-------------------------------------------------------------------------------
Routes : 1
```

On PE-2, the following active IPv6 FlowSpec route specifies that all traffic will be dropped (rate limit: 0 kbps) that matches the criteria: DA 2001:db8::225:2/126, SA 2001:db8::112:2/126, destination port 4191, and source port greater than 1024. This route is generated by the FlowSpec route server connected to PE-1.

```
*A:PE-2# show router bgp routes flow-ipv6
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP FLOW IPV6 Routes
===============================================================================
Flag  Network             Nexthop                 LocalPref       MED
      As-Path
-------------------------------------------------------------------------------
u*>i  --                  ::                      100             None
      No As-Path

      Community Action:  rate-limit: 0 kbps
      NLRI Subcomponents:
      Dest Pref : 2001:db8::225:2/126 offset 0
      Src Pref  : 2001:db8::112:2/126 offset 0
      Ip Proto  : [ == 6 ]
      Dest Port : [ == 4191 ]
      Src Port  : [ >1024 ]
-------------------------------------------------------------------------------
Routes : 1
```

The following sections describe:

- Default BGP behavior without GR
- GR
- LLGR

Default BGP Behavior without GR

The RR PE-1 is isolated from the other PEs by disabling the ports toward PE-2 and PE-3, as follows:

```
*A:PE-1# configure port 1/1/[1..2] shutdown
```

All BGP sessions with the BGP peers drop and the BGP peers remove the routes received from RR PE-1; for example, the list of IPv4 routes on PE-2 is empty. The same is true for the other configured address families.

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
No Matching Entries Found.
===============================================================================
*A:PE-2#
```

The following BGP summary on PE-2 shows that the session toward PE-4 is established, but the session toward PE-1 is down (state: Connect). A similar output is seen on the other PEs in AS 64496, because all BGP sessions toward the RR are down.

```
*A:PE-2# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496      108    0 00h01m10s Connect
                            8    0
192.168.24.2
Def. Instance  64500      137    0 01h05m02s 0/0/0 (IPv4)
                          211    0           0/0/0 (IPv6)
                                             1/1/2 (VpnIPv4)
                                             1/1/2 (VpnIPv6)
                                             0/0/1 (L2VPN)
                                             0/0/0 (FlowIPv4)
                                             0/0/0 (FlowIPv6)

-------------------------------------------------------------------------------
*A:PE-2#
```

The ports on PE-1 are re-enabled and the BGP routes are re-advertised.

```
*A:PE-1# configure port 1/1/[1..2] no shutdown
```

GR

On all PEs, GR is enabled with a stale routes time of 150 seconds and notification enabled, as follows. The default restart time is 300 seconds, but the stale routes will already be deleted when the stale-routes time expires after 150 seconds. LLGR is not enabled yet.

```
configure
    router
        bgp
            group "iBGP"
                graceful-restart
                    stale-routes-time 150
                    enable-notification
                exit
            exit
```

RR PE-1 is isolated, as follows:

```
*A:PE-1# configure port 1/1/[1..2] shutdown
```

When the hold timer expires, the BGP session goes down, and the BGP peers enter the helper mode, RR PE-1 as well as its clients. The following debug message occurs on PE-2 if debugging is enabled for graceful restart:

```
153 2018/04/05 12:50:26.757 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer VR 1: Group iBGP: Peer 192.0.2.1: entering helper mode due to reason hold_timer
_expiry
"
```

Log 99 logs the event as follows:

```
119 2018/04/05 12:50:26.757 UTC WARNING: BGP #2018 Base VR 1
"Peer 1: 192.0.2.1: graceful restart status changed to restarting"
```

The client PEs do not remove the routes they received from RR PE-1 immediately, but they mark these routes as stale and they keep using them. In the following list of IPv4 unicast routes, the x-status code indicates that the route is stale.

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>xi 172.16.225.0/30                                     100         None
      192.0.2.5                                           None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

When the BGP sessions are restored and an EOR is received for the AFI/SAFIs, the BGP routes are re-advertised and there are no longer any stale routes. However, if the stale routes timer expires before an EOR is received for the AFI/SAFIs, the stale routes are removed. The following command shows that there are no longer any BGP IPv4 routes in PE-2.

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

================================================================================
BGP IPv4 Routes
================================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
--------------------------------------------------------------------------------
No Matching Entries Found.
================================================================================
*A:PE-2#
```

When the stale routes timer expires before an EOR is received for the AFI/SAFIs, the GR phase is terminated and the PE is no longer a GR helper. The following debug messages are logged on PE-2 when debugging is enabled for GR:

```
154 2018/04/05 12:52:56.757 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
BGP trying to exit helper for peer Peer 1: 192.0.2.1 with reason stale-routes-
time expired for all address families
"

155 2018/04/05 12:52:56.757 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
BGP flushing stale routes for peer Peer 1: 192.0.2.1 AF All Address Families
"

156 2018/04/05 12:52:56.758 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer 1: 192.0.2.1: exit helper mode due to reason stale-routes-time expired
"
```

The following message is logged in log 99 on PE-2.

```
139 2018/04/05 12:52:56.758 UTC WARNING: BGP #2018 Base VR 1
"Peer 1: 192.0.2.1: graceful restart status changed to notHelping"
```

The situation on PE-1 is restored and the routes are re-advertised.

```
*A:PE-1# configure port 1/1/[1..2] no shutdown
```

In the example, the stale routes time is 150 seconds and the restart time 300 seconds. The helper mode stops when either of these timers expires. When the stale routes time is increased to 400 seconds and the restart time remains 300 seconds, the helper mode will stop when the restart time expires, as shown by the following debug message.

```
265 2018/04/05 13:40:44.758 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer 1: 192.0.2.1: exit helper mode due to reason restart-time expired
"
```

LLGR

Initially, LLGR will be configured with the same LLGR-stale time for all the configured AFI/SAFIs, but it is possible to configure LLGR with a different LLGR-stale time per AF. The LLGR-stale time is configured as **advertised-stale-time**—which is the value that is advertised to the BGP peer—but can be overridden locally without being advertised.

At first, LLGR will be enabled on the "iBGP" group on PE-1, PE-2, PE-3, and PE-5. Later, LLGR will also be enabled on the "eBGP" group on PE-2 and PE-4.

LLGR Enabled on iBGP Sessions

The following configuration enables LLGR as well as GR in the "iBGP" group on all PEs in AS 64496 for all the already configured AFI/SAFIs.

```
configure
    router
        bgp
            group "iBGP"
                graceful-restart
                    stale-routes-time 150
                    enable-notification
                    long-lived
                        advertised-stale-time 3600
                    exit
                exit
```

Neither GR nor LLGR is enabled in the "eBGP" group on PE-2 and PE-4. This makes no difference for the GR phase on PE-2; only for the LLGR phase.

When the RR PE-1 gets isolated and the hold timer for the BGP session expires, the GR phase starts for the "iBGP" group and the routes received from PE-1 are marked as stale, but remain in use. In the GR phase, the detailed information for the stale IPv4 route 172.16.225.0/30 on PE-2 shows the flags used, valid, best, IGP, and stale (not LLGR-stale), as follows. PE-2 will keep using the stale routes in the GR phase. PE-2 will not withdraw any stale routes and eBGP peer PE-4 remains unaware of the failure.

```
*A:PE-2# show router bgp routes detail
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496        Local AS:64496
===============================================================================
---snip---
===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes
```

```
Network        : 172.16.225.0/30
Nexthop        : 192.0.2.5
Path Id        : None
From           : 192.0.2.1
Res. Protocol  : ISIS                   Res. Metric    : 20
Res. Nexthop   : 192.168.23.2
Local Pref.    : 100                    Interface Name : int-PE-2-PE-3
---snip---
Community      : No Community Members
Cluster        : 192.0.2.1
Originator Id  : 192.0.2.5              Peer Router Id : 192.0.2.1
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Stale
Route Source   : Internal
---snip---
```

The routes keep the stale flag "x", as in the GR phase. The following IPv4 route is marked as stale on PE-2:

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>xi 172.16.225.0/30                                    100         None
      192.0.2.5                                          None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

The following detailed information for this route on PE-2 shows the LLGR-stale flag instead of the normal stale flag, as follows:

```
*A:PE-2# show router bgp routes detail
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
---snip---
===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 172.16.225.0/30
Nexthop        : 192.0.2.5
```

```
Path Id        : None
From           : 192.0.2.1
Res. Protocol  : ISIS                   Res. Metric    : 20
Res. Nexthop   : 192.168.23.2
Local Pref.    : 100                    Interface Name : int-PE-2-PE-3
---snip---
Community      : No Community Members
Cluster        : 192.0.2.1
Originator Id  : 192.0.2.5              Peer Router Id : 192.0.2.1
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  LlgrStale
---snip---
```

When debugging is enabled for GR, the following message on PE-2 is generated when the GR phase starts.

```
873 2018/04/06 10:44:40.453 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer VR 1: Group iBGP: Peer 192.0.2.1: entering helper mode due to reason hold_timer
_expiry
"
```

The following message on PE-2 is generated when the LLGR phase starts.

```
883 2018/04/06 10:47:10.453 UTC MINOR: DEBUG #2001 Base BGP
"BGP: RESTART
Peer VR 1: Group iBGP: Peer 192.0.2.1: Entering helper, LLGR Phase -
 reason llgr_Start_on_rtRtTm_pop
"
```

In the LLGR phase, the following command on PE-2 shows that, for the BGP session with RR PE-1, GR and LLGR are both enabled locally and on the BGP peer, and the GR and LLGR status of the peer PE-1 is "received restart request", so the LLGR phase is ongoing.

The advertised NLRIs of the RR PE-1 and its client PE-2 are similar, so the same AFI/SAFIs (and stale time) have been advertised by PE-2 and received from peer PE-1 for GR, GR notification, and LLGR. LLGR can only work if it is enabled on both BGP peers, which is the case.

```
*A:PE-2# show router bgp neighbor 192.0.2.1 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.1 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
NLRI(s) that peer supports restart for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) that peer saved forwarding for      : None
NLRI(s) that restart is negotiated for      : None
NLRI(s) of received end-of-rib markers      : None
NLRI(s) of all end-of-rib markers sent      : None
NLRI(s) peer supports NOTIFICATION GR for   : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
Restart time locally configured for peer    : 300 seconds
Restart time requested by the peer          : 300 seconds
Time until stale routes are deleted or
become long-lived stale                     : 150 seconds
Graceful restart status on the peer         : Rcvd restart request
Long-Lived GR status on the peer            : Rcvd restart request
Number of Restarts                          : 1
Last Restart at                             : 04/06/2018 09:19:52
-------------------------------------------------------------------------------
LLGR Configuration                          : Enabled
Peer's LLGR feature                         : Enabled
NLRI(s) peer signaled LLGR for & stale time
& F-bit                                     : ipv4 : 3600 seconds
                                              ipv6 : 3600 seconds
                                              vpn-ipv4 : 3600 seconds
                                              vpn-ipv6 : 3600 seconds
                                              l2-vpn : 3600 seconds
                                              flow-ipv4 : 3600 seconds
                                              flow-ipv6 : 3600 seconds
NLRI(s) LLGR negotiated for and stale time  : ipv4 : 3600 seconds
                                              ipv6 : 3600 seconds
                                              vpn-ipv4 : 3600 seconds
                                              vpn-ipv6 : 3600 seconds
                                              l2-vpn : 3600 seconds
                                              flow-ipv4 : 3600 seconds
                                              flow-ipv6 : 3600 seconds
LLGR Restart time overridden for the peer   : N/A
NLRI(s) LLGR advertised & stale time & F-bit: ipv4 : 3600 seconds
                                              ipv6 : 3600 seconds
                                              vpn-ipv4 : 3600 seconds
                                              vpn-ipv6 : 3600 seconds
                                              l2-vpn : 3600 seconds
                                              flow-ipv4 : 3600 seconds
                                              flow-ipv6 : 3600 seconds
===============================================================================
*A:PE-2#
```

On PE-2, the following command shows that GR and LLGR are disabled for the eBGP session with PE-4.

```
*A:PE-2# show router bgp neighbor 192.168.24.2 graceful-restart

===============================================================================
BGP Neighbor 192.168.24.2 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Disabled
GR Notification                             : Disabled
Peer's Graceful Restart feature             : Disabled
NLRI(s) that peer supports restart for      : None
NLRI(s) that peer saved forwarding for      : None
NLRI(s) that restart is negotiated for      : None
NLRI(s) of received end-of-rib markers      : None
NLRI(s) of all end-of-rib markers sent      : None
NLRI(s) peer supports NOTIFICATION GR for   : None
Restart time locally configured for peer    : 120 seconds
Restart time requested by the peer          : 0 seconds
```

```
Time until stale routes are deleted or
become long-lived stale                     : 360 seconds
Graceful restart status on the peer         : Not currently being helped
Long-Lived GR status on the peer            : Not currently being helped
Number of Restarts                          : 0
Last Restart at                             : Never
-------------------------------------------------------------------------------
LLGR Configuration                          : Disabled
Peer's LLGR feature                         : Disabled
NLRI(s) peer signaled LLGR for & stale time
& F-bit                                     : N/A
NLRI(s) LLGR negotiated for and stale time  : N/A
LLGR Restart time overridden for the peer   : N/A
NLRI(s) LLGR advertised & stale time & F-bit: N/A
===============================================================================
*A:PE-2#
```

In the LLGR phase, the stale routes remain stale, but are depreferenced. In this example, there are no alternative routes with a better preference, so the stale routes remain valid, best, and used. Traffic between PE-2, PE-3, and PE-5 is still uninterrupted.

However, the eBGP session between PE-2 and PE-4 does not have LLGR enabled. In the LLGR phase, the LLGR-stale routes are immediately withdrawn by PE-2; for example, the following BGP update withdraws the VPN-IPv4 routes toward PE-4. Therefore, VPN traffic can no longer be exchanged between VPRN 1 on PE-3 (or PE-5) and VPRN 1 on PE-4.

```
884 2018/04/06 10:47:10.458 UTC MINOR: DEBUG #2001 Base Peer 1: 192.168.24.2
"Peer 1: 192.168.24.2: UPDATE
Peer 1: 192.168.24.2 - Send BGP UPDATE:
    Withdrawn Length = 5
        172.16.225.0/30
    Total Path Attr Length = 39
    Flag: 0x90 Type: 15 Len: 35 Multiprotocol Unreachable NLRI:
        Address Family VPN_IPV4
        172.16.5.0/30 RD 64496:1 Label 0
        172.16.3.0/30 RD 64496:1 Label 0
"
```

Even though GR is also disabled for the eBGP session between PE-2 and PE-4, the routes are only withdrawn in the LLGR phase, not in the GR phase. GR is meant for short interruptions where the GR helper PE-2 pretends that the situation is normal and traffic can be forwarded based on stale routes, while LLGR is meant for longer failures and the neighbors need to be informed.

The ports on PE-1 are re-enabled and the routes are re-advertised followed by an EOR per AFI/SAFI, which terminates the LLGR phase.

```
*A:PE-1# configure port 1/1/[1..2] no shutdown
```

LLGR Enabled on eBGP Session

On PE-2 and PE-4, GR and LLGR are enabled for the "eBGP" group, as follows:

```
configure
    router
        bgp
            group "eBGP"
                graceful-restart
                    stale-routes-time 150
                    enable-notification
                    long-lived
                        advertised-stale-time 3600
                    exit
                exit
            exit
```

PE-2 will re-advertise the routes it sent to PE-4, but with well-known community **llgr-stale**. PE-4 was unaware of the GR phase; it only got involved in the LLGR phase. The following BGP update was sent by PE-2 to its eBGP peer PE-4 for the IPv4 address family:

```
339 2018/04/06 12:43:06.007 UTC MINOR: DEBUG #2001 Base Peer 1: 192.168.24.2
"Peer 1: 192.168.24.2: UPDATE
Peer 1: 192.168.24.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 27
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 64496 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.168.24.1
    Flag: 0xc0 Type: 8 Len: 4 Community:
        llgr-stale
    NLRI: Length = 5
        172.16.225.0/30
"
```

PE-4 does not mark the route as stale in the way that PE-2 does; the BGP route does not get the stale flag "x", as follows:

```
*A:PE-4# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
```

```
-------------------------------------------------------------------------------
u*>i  172.16.225.0/30                                    None        None
      192.168.24.1                                       None        -
      64496
-------------------------------------------------------------------------------
Routes : 1
```

The detailed information for this route on PE-4 shows the community **llgr-stale**, but no LLGR-stale flag, as follows:

```
*A:PE-4# show router bgp routes detail
===============================================================================
 BGP Router ID:192.0.2.4        AS:64500       Local AS:64500
===============================================================================
---snip---
===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 172.16.225.0/30
Nexthop        : 192.168.24.1
Path Id        : None
From           : 192.168.24.1
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 192.168.24.1
Local Pref.    : n/a                    Interface Name : int-PE-4-PE-2
---snip---
Community      : llgr-stale
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
Route Source   : External
AS-Path        : 64496
---snip---
```

Per-AF LLGR

The following configuration on PE-2 enables GR for the same address families as before, while LLGR will only be applied for IPv4 FlowSpec and IPv6 FlowSpec, with different LLGR-stale times. The default LLGR-stale time—**advertised-stale-time**—is 86400 seconds, but **helper-override-stale-time 0** in the iBGP group context overrides the LLGR-stale time to a zero value for the iBGP group. For FlowSpec routes, the **advertised-stale-time** is set to a value of 20000 seconds. For IPv4 FlowSpec, the **helper-override-stale-time** is set to 2000 seconds; for IPv6 FlowSpec, it is set to 3000 seconds. The forwarding bit is only set for non-forwarding AFs—**forwarding-bits-set non-fwd**—so it will be set for configuration routes, such as FlowSpec routes.

```
configure
```

```
    router
        bgp
            group "iBGP"
                family ipv4 ipv6 vpn-ipv4 vpn-ipv6 l2-vpn flow-ipv4 flow-ipv6
                graceful-restart
                    stale-routes-time 150
                    enable-notification
                    long-lived
                        advertised-stale-time 86400         # default
                        helper-override-stale-time 0
                        family flow-ipv4
                            advertised-stale-time 20000
                            helper-override-stale-time 2000
                        exit
                        family flow-ipv6
                            advertised-stale-time 20000
                            helper-override-stale-time 3000
                        exit
                        forwarding-bits-set non-fwd
                        no advertise-stale-to-all-neighbors     # default
                    exit
                exit
                peer-as 64496
                neighbor 192.0.2.1
                exit
```

With this configuration on PE-2, the LLGR phase will be reduced to zero seconds for all AFs except IPv4 FlowSpec and IPv6 FlowSpec, but the **helper-override-stale-time** is not advertised to the BGP peer; only the **advertised-stale-time** is advertised. The GR phase applies for all configured address families with the same timers. When the BGP configuration on PE-1 is preserved and LLGR is enabled for the same address families, the following command shows the GR information on PE-2 for peer PE-1.

```
*A:PE-2# show router bgp neighbor 192.0.2.1 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.1 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
NLRI(s) that peer supports restart for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) that peer saved forwarding for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) that restart is negotiated for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) of received end-of-rib markers      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) of all end-of-rib markers sent      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) peer supports NOTIFICATION GR for   : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
Restart time locally configured for peer    : 300 seconds
Restart time requested by the peer          : 300 seconds
```

```
Time until stale routes are deleted or
become long-lived stale                     : 150 seconds
Graceful restart status on the peer         : Not currently being helped
Long-Lived GR status on the peer            : Not currently being helped
Number of Restarts                          : 0
Last Restart at                             : Never
-------------------------------------------------------------------------------
LLGR Configuration                          : Enabled
Peer's LLGR feature                         : Enabled
NLRI(s) peer signaled LLGR for & stale time
& F-bit                                     : ipv4 : 3600 seconds (F)
                                              ipv6 : 3600 seconds (F)
                                              vpn-ipv4 : 3600 seconds (F)
                                              vpn-ipv6 : 3600 seconds (F)
                                              l2-vpn : 3600 seconds (F)
                                              flow-ipv4 : 3600 seconds (F)
                                              flow-ipv6 : 3600 seconds (F)
NLRI(s) LLGR negotiated for and stale time  : ipv4 : 0 seconds
                                              ipv6 : 0 seconds
                                              vpn-ipv4 : 0 seconds
                                              vpn-ipv6 : 0 seconds
                                              l2-vpn : 0 seconds
                                              flow-ipv4 : 2000 seconds
                                              flow-ipv6 : 3000 seconds
LLGR Restart time overridden for the peer   : N/A
NLRI(s) LLGR advertised & stale time & F-bit: ipv4 : 86400 seconds
                                              ipv6 : 86400 seconds
                                              vpn-ipv4 : 86400 seconds
                                              vpn-ipv6 : 86400 seconds
                                              l2-vpn : 86400 seconds(F)
                                              flow-ipv4 : 20000 seconds(F)
                                              flow-ipv6 : 20000 seconds(F)
===============================================================================
*A:PE-2#
```

LLGR is enabled on PE-1 and PE-2. BGP peer PE-1 has signaled LLGR-stale times of 3600 seconds for the supported AFI/SAFIs. PE-2 had advertised the default LLGR-stale time of 86400 seconds for all supported AFI/SAFIs except for the FlowSpec AFI/SAFIs, where the LLGR-stale time is 20000 seconds. On PE-2, the F-bit is only set for the non-forwarding routes; in this case, L2 VPN, IPv4 FlowSpec, and IPv6 FlowSpec.

The **helper-override-stale-time** is not advertised to the BGP peer, but considered for the local LLGR behavior (in bold). Only the FlowSpec AFs get a non-zero LLGR-stale time: 2000 seconds for IPv4 FlowSpec; 3000 seconds for IPv6 FlowSpec.

The following GR/LLGR information on peer PE-1 shows the advertised LLGR-stale time, not the **helper-override-stale-time** configured on PE-2.

```
*A:PE-1# show router bgp neighbor 192.0.2.2 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.2 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
```

```
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
NLRI(s) that peer supports restart for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) that peer saved forwarding for      : l2-vpn flow-ipv4 flow-ipv6
NLRI(s) that restart is negotiated for      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) of received end-of-rib markers      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) of all end-of-rib markers sent      : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
NLRI(s) peer supports NOTIFICATION GR for   : ipv4 ipv6 vpn-ipv4 vpn-ipv6
                                              l2-vpn flow-ipv4 flow-ipv6
Restart time locally configured for peer    : 300 seconds
Restart time requested by the peer          : 300 seconds
Time until stale routes are deleted or
become long-lived stale                     : 150 seconds
Graceful restart status on the peer         : Restart completed
Long-Lived GR status on the peer            : Restart completed
Number of Restarts                          : 4
Last Restart at                             : 04/06/2018 14:21:25
-------------------------------------------------------------------------------
LLGR Configuration                          : Enabled
Peer's LLGR feature                         : Enabled
NLRI(s) peer signaled LLGR for & stale time
& F-bit                                     : ipv4 : 86400 seconds
                                              ipv6 : 86400 seconds
                                              vpn-ipv4 : 86400 seconds
                                              vpn-ipv6 : 86400 seconds
                                              l2-vpn : 86400 seconds (F)
                                              flow-ipv4 : 20000 seconds (F)
                                              flow-ipv6 : 20000 seconds (F)
NLRI(s) LLGR negotiated for and stale time  : ipv4 : 86400 seconds
                                              ipv6 : 86400 seconds
                                              vpn-ipv4 : 86400 seconds
                                              vpn-ipv6 : 86400 seconds
                                              l2-vpn : 86400 seconds
                                              flow-ipv4 : 20000 seconds
                                              flow-ipv6 : 20000 seconds
LLGR Restart time overridden for the peer   : N/A
NLRI(s) LLGR advertised & stale time & F-bit: ipv4 : 3600 seconds(F)
                                              ipv6 : 3600 seconds(F)
                                              vpn-ipv4 : 3600 seconds(F)
                                              vpn-ipv6 : 3600 seconds(F)
                                              l2-vpn : 3600 seconds(F)
                                              flow-ipv4 : 3600 seconds(F)
                                              flow-ipv6 : 3600 seconds(F)
===============================================================================
*A:PE-1#
```

With this configuration, only the FlowSpec routes can get the LLGR-stale flag. No LLGR phase will start for the other AFs, so the stale routes of those AFs will be withdrawn when the GR phase ends.

It is possible to override the GR restart time to enter the LLGR phase immediately without going through the GR phase, as follows, on PE-2:

```
configure router bgp group iBGP graceful-restart long-lived helper-override-restart-
time 0
```

On PE-2, the locally overridden restart time is shown instead of the default restart time, while the restart time of the peer PE-1 remains 300 seconds, as follows:

```
*A:PE-2# show router bgp neighbor 192.0.2.1 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.1 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
---snip---
Restart time locally configured for peer    : 0 seconds
Restart time requested by the peer          : 300 seconds
---snip---
-------------------------------------------------------------------------------
LLGR Configuration                          : Enabled
Peer's LLGR feature                         : Enabled
---snip---
LLGR Restart time overridden for the peer   : 0 seconds
---snip---
```

PE-2 advertises the overridden restart time to the peer PE-1. The following restart times are shown for PE-1. Unlike peer PE-2, PE-1 did not override the restart time itself.

```
*A:PE-1# show router bgp neighbor 192.0.2.2 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.2 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
---snip---
Restart time locally configured for peer    : 300 seconds
Restart time requested by the peer          : 0 seconds
---snip---
-------------------------------------------------------------------------------
LLGR Configuration                          : Enabled
Peer's LLGR feature                         : Enabled
---snip---
LLGR Restart time overridden for the peer   : N/A
---snip---
```

When the BGP session goes down on PE-2, the GR phase is omitted because the restart time of zero seconds expires instantly, so the LLGR phase starts immediately, as follows.

```
*A:PE-2# show router bgp neighbor 192.0.2.1 graceful-restart

===============================================================================
BGP Neighbor 192.0.2.1 Graceful Restart
===============================================================================
Graceful Restart locally configured for peer: Enabled
GR Notification                             : Enabled
Peer's Graceful Restart feature             : Enabled
---snip---
Graceful restart status on the peer         : Restart completed
Long-Lived GR status on the peer            : Rcvd restart request
---snip---
```

When LLGR phase starts immediately, only the FlowSpec address families will be protected while all routes of the other AFs are withdrawn. The FlowSpec routes get the LLGR-stale flag and route updates to eBGP peer PE-4 will get the LLGR-stale community, as follows:

```
A:PE-2# show router bgp routes flow-ipv4 hunt
===============================================================================
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
---snip---
From           : 192.0.2.1
---snip---
Flags          : Used  Valid  Best  IGP  LlgrStale
---snip---

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
---snip---
To             : 192.168.24.2
---snip---
Community      : llgr-stale rate-limit: 0 kbps
---snip---
```

Conclusion

Graceful restart helpers avoid withdrawing BGP routes immediately when the BGP session goes down. Routes that were received from the failed router are marked as stale, but remain in use. When the BGP session is down for a longer time, such as hours or days, LLGR can take over when the GR ends, possibly only for a subset of AFI/SAFIs. In the LLGR phase, the LLGR-stale routes are depreferenced, but if they remain best and valid, they can be re-advertised to the BGP peers as LLGR-stale.

BGP Monitoring Protocol Basics

This chapter provides information about BGP Monitoring Protocol Basics.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS release 16.0.R2. BGP Monitoring Protocol (BMP) support was introduced in SR OS release 16.0.R1 for unicast IPv4/IPv6, VPN IPv4/IPv6, and labeled IPv4/IPv6. SR OS release 16.0.R4 provides an additional six address families: EVPN, L2VPN, multicast IPv4/IPv6, multicast VPN IPv4/IPv6.

Overview

The BGP Monitoring Protocol (BMP) is a unidirectional protocol for providers to monitor the behavior of BGP on their routers. A router communicates information about one or more BGP sessions to a BMP station, also known as a BMP collector. A router sends information in BMP messages to a BMP station. A BMP station never sends any messages to a router. BMP is described in detail in RFC 7854. Figure 110 shows an operational overview of BMP.

Figure 110 **BMP Operational Overview**

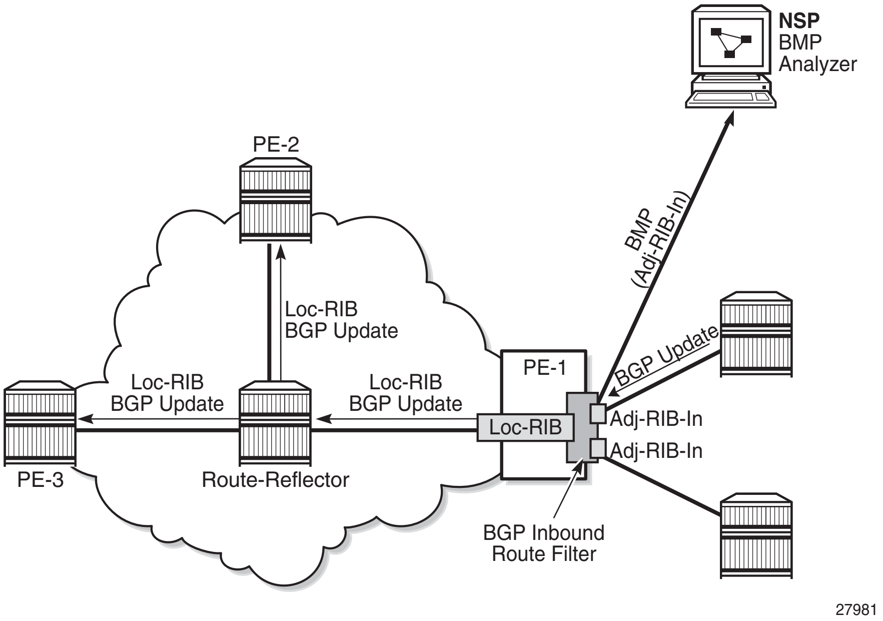


Table 11 lists the BMP message types that are defined in RFC 7854.

Table 11 **BMP Message Types**

BMP Message Type	Description
0	Route monitoring
1	Statistics report
2	Peer down notification
3	Peer up notification
4	Initiation message
5	Termination message
6	Route mirroring message

A BMP station (or BMP collector) typically is a dedicated server running network management or network controller software. Current examples of free and open-source BMP station software are OpenBMP and Open Daylight. Nokia has commercial BMP station support available through the Network Services Platform (NSP) controller. The simple operations and packet format of BMP resulted in many providers having created their own proprietary BMP-collector software.

BMP allows a router to report different types of information. A router can:

- send BMP messages with notifications when neighbors go into or out of Established state (for example, the peer goes "up" or "down"). These notifications are called BMP peer-up and peer-down messages.
- periodically send statistical information about one or more neighbors. This information consists of several counters; for example, how many routes are received from a specific neighbor, or how many of those routes were rejected or accepted because of an ingress policy. Other counters report how many errors were encountered; for example, AS-path loops, duplicate prefixes, withdrawals received, and so on.
- report the exact routes that were received from a neighbor. This action is called route monitoring. To do this, a router first re-encapsulates a BGP route into its original BGP update message, then encapsulates that BGP update message within a BMP route monitoring message to send it to the BMP station.

➡ **Note:** BMP on an SR OS router will only report information about routes that were received from a neighbor, which is the standard BMP behavior documented in RFC 7854. BMP will also report upon routes leaked or redistributed into the BGP RIB. A limitation of RFC 7854 is that BMP does not monitor routes sent toward a specific BGP neighbor. Nokia supports RFC 7854, so does not support monitoring of routes that were sent toward a BGP neighbor.

Configuration

Basic configuration of BMP

There are two main steps to enable BMP monitoring on an SR OS router:

Step 1. Configure a BMP station. This configuration identifies the target to which BMP information will be sent.

Step 2. Configure one or more BGP neighbors. These are the BGP peering sessions that will be monitored by BMP and the configured BMP station.

Configuring a BMP station

BMP stations and associated parameters are configured in global configuration mode. This allows the BMP station to reside either within the base router instance, or in a VPRN routing instance. The Nokia BMP implementation can monitor BGP peers in a base Internet service or in a VPRN service instance.

BMP will initiate a separate TCP session for each VPRN BGP instance monitored. The BMP router will use a different source TCP port number toward each configured TCP destination port number of the BMP station. For example, if there are four VPRN services configured in addition to the base router instance, the BMP router will instantiate five TCP sessions between the BMP router and the BMP station (one TCP session to monitor the base router instance, and four TCP sessions to monitor the VPRN services).

SR OS supports the configuration of up to eight BMP stations. To configure a BMP station, use the following command syntax:

```
*A:Dut-C# configure bmp station Antwerp create
```

This configuration example creates a BMP station with the name "Antwerp". This name must be used when configuring BGP peers to be monitored by this station. The name can also be used in **show router bmp** commands.

The next step is to configure (at a minimum) the IP address and the TCP destination port the BMP station is listening to. These parameters inform the BMP router where to reach the BMP station. BMP does not use a well-known port number; a provider can select any TCP port number. BMP sessions from an SR OS router can run over either TCP IPv4 or TCP IPv6.

The following configures the IP address 100.1.1.10 and port number 5000 of the BMP station:

```
configure
    bmp
        station "Antwerp" create
            connection
                station-address 100.1.1.10 port 5000
            exit
        exit
    exit
```

This configuration example creates a BMP station that can be used to monitor one or more BGP peers. Next, configure the BGP peers to be monitored by this station.

Assigning the BGP peers to be monitored

To configure one or more BGP neighbors to be monitored by the BMP station, first configure the **monitor** command in the BGP context or one of its subcontexts. This command can be configured at the BGP instance level, at the BGP group level, or at the neighbor level.

In the following example, monitoring is enabled (no shutdown) for all BGP peers defined in the **bgp** context, where the BMP reporting goes to BMP station *Antwerp*.

```
configure
    router
        bgp
            monitor
                station Antwerp
                no shutdown
            exit
            group internal-1
              --- snip ---
            Exit
            group internal-2
              --- snip ---
            exit
        exit
```

By default, BMP, including each individually configured station, is in the administrative shutdown state. To allow BMP to start the BMP sessions, administratively enable the BMP station:

```
configure
    bmp
        no shutdown
        station Antwerp
            no shutdown
        exit
    exit
```

All peers in the BGP instance of the base router are now monitored by station "Antwerp". At this stage, the router will only send BMP peer-up and peer-down messages to the BMP station. To send additional information (such as periodic statistics messages, or to report incoming BGP routes) requires explicit configuration.

Configuring periodic statistics messages

Enabling periodic statistics messages is done under the **configure bmp station** command. The command to enable periodic statistics is **stats-report-interval** <seconds>:

```
configure
    bmp
        station Antwerp
            stats-report-interval 600
        exit
    exit
```

This configuration example will cause the router to send statistics messages for each monitored peer to the BMP station every 10 minutes (600 seconds).

Verifying that the BMP session between router and BMP station works

To display the state of a BMP session to a BMP station, use the **show router bmp station** <station-name> command:

```
show router bmp station Antwerp
```

The output of the **show** command for BMP station Antwerp is as follows:

```
*A:Dut-C# show router bmp station "Antwerp"

===============================================================================
BMP Station "Antwerp" (monitoring router "Base")
===============================================================================
Admin State       : enabled              Global BMP State : enabled
Station Address   : 100.1.1.10           Station Port     : 5000
Via Router        : Base
Stats Report      : 30 seconds
Connect Interval  : 5 seconds            Local Routes     : not reporting
Reported families: ipv4

Session State     : ESTABLISHED          Last State Change: 08/29/2018 13:23:19
Reason Last Down  : admin shutdown       Last Msg Sent    : 08/29/2018 13:23:19
Local Address     : 100.1.1.3            Local Port       : 51446
Routes Timer      : 2 seconds left       Stats Timer      : 3 seconds left
Connect Timer     : not running          Monitored Peers  : 0 of 1
Initiation Msgs   : 1                    Goodbye Msgs     : 0
Peer Up Msgs      : 0                    Peer Down Msgs   : 0
Route Report Msgs: 0                     Stat Report Msgs : 0
Bytes Sent        : 276                  Output Queue     : 0/5
===============================================================================
*A:Dut-C#
```

The output consists of two blocks of information.

- The first block shows configuration information about this specific BMP station.
- The second block shows dynamic information about the current BMP session from the router instance to the BMP station.

Verify that the Session State is "ESTABLISHED".

Configuring BMP route monitoring

Configuring BMP route monitoring requires explicit configuration under the **monitor** command in the BGP instance context.

It is possible to configure BMP to report pre-policy routes, or post-policy routes, or both. Pre-policy routes are incoming routes as they were before applying any ingress policy. Post-policy routes are resulting routes in the Adj-RIB-In and reflect the routes after applying any BGP ingress policy.

Configuring BMP to report both pre- and post-policy routes will result in the doubling of BMP messages to the BMP station. This is because the router will send a route-monitor message for each pre-policy route, and for each post-policy route. This doubles the amount of resources consumed by BMP (such as bandwidth consumed on the link between the router and the BMP station, and CPU usage). The impact of enabling BMP route monitoring on the router CPU is similar to adding a BGP neighbor.

To configure route monitoring, use the **route-monitoring [pre-policy] [post-policy]** command in the monitor configuration mode in the BGP configuration context:

```
configure
    router
        bgp
            monitor
                station Antwerp
                route-monitoring pre-policy
                no shutdown
            exit
        exit
```

The BMP route monitoring is enabled within the context where the **monitor station** command is configured: in the general BGP context, the group context, or the neighbor context. With this configuration, the BMP router will start sending route monitoring messages for every route received from every neighbor in the base router BGP instance. This can be verified via the **show router bmp station** <station-name> command, which displays the counter for "Route Report Msgs:".

Advanced BMP configuration options

The BMP configuration can be fully customized. The following sections describe some additional configuration options.

Configuring route monitoring for different address families

When route monitoring is enabled, by default the BMP router will only report received IPv4 routes to the BMP station. This aligns with the default BGP behavior, where only unicast IPv4 is enabled when configuring a neighbor under BGP. To enable route monitoring for additional BGP address families, additional explicit configuration is required. The additional address families are available and can be configured under the **configure bmp station** command context, as follows:

```
configure
    bmp
        station Antwerp
            family
        exit
    exit
```

In SR OS release 16.0.R1, a Nokia BMP router supports route monitoring of six address families:

- unicast IPv4, unicast IPv6
- VPN-IPv4, VPN-IPv6
- label-IPv4, label-IPv6

SR OS release 16.0.R4 provides an additional six address families:

- EVPN
- L2VPN
- mcast-IPv4, mcast-IPv6
- mcast-VPN-IPv4, mcast-VPN-IPv6

Configuring monitoring of locally generated routes

RFC 7854 BMP reports only the routes in the Adj-RIB-In that were received from monitored neighbors. However, the BGP-RIB can hold more routes than those routes BGP has learned from neighbors. These locally generated routes are called imported or leaked routes.

Imported routes are learned via redistributing routes into BGP from external sources, like static, connected, IS-IS, or OSPF. Leaked routes are BGP routes from other BGP service instances that are leaked into the base router BGP.

To configure the Nokia BMP router to extend route reporting and report these imported and leaked routes to a configured BMP station, configure the **report-local-routes** command under the BMP station:

```
configure
    bmp
        station Antwerp
            report-local-routes
        exit
    exit
```

Configuring the frequency of router statistics reports

When periodic statistics are enabled, the router will send all the statistics as described in RFC 7854, section 4.8, except for statistic number 13 (number of duplicate update messages received).

The Nokia BMP router-supported statistics are:

- 0 - number of prefixes rejected by inbound policy
- 1 - number of duplicate prefix advertisements received
- 2 - number of duplicate withdraws received
- 3 - number of received updates invalidated due to cluster-list loop
- 4 - number of received updates invalidated due to AS-path loop
- 5 - number of received updates invalidated due to originator-id
- 6 - number of received updates invalidated due to as-confed loop
- 7 - total number of routes in Adj-RIB-In (all families)
- 8 - total number of routes in loc-RIB (all families)
- 9 - number of routes per address family in Adj-RIB-In (see Note)
- 10 - number of routes per address family in loc-RIB (see Note)
- 11 - number of updates subjected to treat-as-withdraw
- 12 - number of prefixes subjected to treat-as-withdraw
- 13 - not supported/reported by SR OS (number of duplicate update messages received)

➡ **Note:** These two statistics are per address family. The address family is specified as a BGP AFI/SAFI pair. Regardless of what families are configured or supported for route monitoring, a router will report the statistics of all address families that were negotiated with the neighbor.

The values shown in the preceding counters are the same values that are shown by the **show router <vrid> bgp neighbor <ip-addr> [detail]** command.

Customizing the TCP connection to the BMP station

BMP uses TCP sessions to send BMP messages to the BMP station. It is possible to customize the TCP-session settings using several configuration options. These options are under the **configure bmp station <name> connection** command context.

Setting the local address of the TCP session

For increased operational security, BMP collectors might restrict accepting BMP sessions from unknown routers. It is important to have a configuration option to force a BMP router to accept specific IP addresses. To enforce the source address of a BMP session, the provider can configure the "local-address <ip-address>".

A Nokia router BMP session can be over an IPv4 or IPv6 TCP session. The source IP address used by the BMP router can be configured using the **local-address** command. The local address can be an IPv4 or an IPv6 address. The address family (IPv4 or IPv6) must match the address family of the IP address configured in the **station-address <ip-address> port <portnr>** command:

```
configure
    bmp
        station "Antwerp"
            connection
                station-address 100:200:300::1 port 5000
                local-address 100:200:300::2
            exit
        exit
    exit
```

Setting the routing context of the BMP session

A Nokia router allows a provider to configure multiple virtual router instances.

The base router is such a virtual router. Each VRPN instance is also a virtual router.

A Nokia BMP router allows a provider to monitor a BGP VPRN session while the TCP connection of the BMP session is configured in another VPRN instance.

This functionality allows the provider to let a single BMP station connection, within a specific VPRN instance, monitor BGP sessions and instances resident in other virtual routers.

The TCP connection of a BMP session is by default active in the base router. This can be changed by adding additional VPRN context configuration when configuring a BMP station, as follows:

```
configure
    bmp
        station "Antwerp"
            connection
                router service-name vprn-22
            exit
        exit
```

Connect-retry command

When a router initiates a BMP session, it will try to establish the TCP connection to the BMP station. If this attempt fails, the router will wait a short while, then retry to bring up the connection. The time between two such attempts increases over time. The first attempt waits 3 seconds. After each failed attempt, the waiting time doubles (exponential increase). The maximum time to wait between two attempts is by default 2 minutes (120 seconds). This maximum waiting time is configurable, as follows:

```
configure
    bmp
        station "Antwerp"
            connection
                connect-retry 600
            exit
        exit
```

This configuration example will set the maximum waiting time between two connection attempts to 10 minutes (600 seconds).

TCP keepalives

BMP does not have any mechanism to detect the liveness of a BMP station. As the protocol is unidirectional, a router will not detect that a BMP station is down or unreachable, until it tries to send data to the station. During normal operation, the TCP layer will inform the BMP layer of an error when BMP tries to send a message to a BMP station that is down or unreachable. After discovering the TCP error, BMP will close the BMP session and try to re-establish a new session. However, when the BMP router has nothing to send to the unreachable BMP station, the station is not detected that easily.

Providers might need to detect a BMP failure even quicker. To do that, providers have the option to configure "TCP keepalives" on the BMP session. TCP keepalives are a feature of the TCP protocol. TCP keepalives are used to ensure the liveness of a TCP connection, even when no data is sent.

BMP on a Nokia router can use TCP keepalives. No special support is needed on the BMP station or host operating system because this functionality is a basic operation of the TCP session.

TCP keepalives are disabled by default. To enable a BMP session with TCP keepalives, configure:

```
configure
    bmp
        station "Antwerp"
            connection
                tcp-keepalive
                    no shutdown
                exit
            exit
        exit
    exit
```

The default operational values of TCP keepalives on a BMP session are:

- keep-idle (sometimes called keep-wait) 600 seconds
- keep-interval 15 seconds
- keep-count 4 times

A provider can change these values. Configuring more aggressive values-tuning values for faster convergence-will have a slight impact on CPU and bandwidth usage. Configuring less aggressive values lowers the risk of false positives. For normal BMP operation, the default values are a good starting point. The following is an example if a provider wants to use non-default TCP keepalive values.

```
configure
    bmp
        station "Antwerp"
            connection
                tcp-keepalive
                    keep-count 5
                    keep-idle 300
                    keep-interval 10
                    no shutdown
                exit
            exit
        exit
    exit
```

Conclusion

In this chapter, the basic operation of Nokia BMP technology is described. The BMP implementation on a Nokia router is fully dual-stack IPv4/IPv6 aware and supports the monitoring of active BGP neighbor state (up or down), the BGP pre- and post-policy routes received, and a set of associated statistics for the BGP Adj-RIB-In and RIB-IN.

Usually, the impact upon the router performance for each configured BMP station is similar to adding a BGP neighbor. The Nokia BMP implementation supports the monitoring of twelve address families (unicast IPv4/IPv6, VPN IPv4/IPv6, label IPv4/IPv6, EVPN, L2VPN, mcast-IPv4/IPv6, mcast-VPN-IPv4/IPv6) in SR OS Release 16.0.R4, and later.

The Nokia BMP implementation can use TCP timers to detect unreachable BMP collectors. There is support for monitoring BGP neighbors in the base router or in a VPRN instance and support for BMP collectors located in the GRT or in any other VPRN service instance.

BGP Multipath

This chapter provides information about BGP Multipath.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS release 14.0.R4.

In release 14.0.R1, BGP multipath has been improved by offering more control over the selection of the eligible multipaths. This chapter provides information about these enhancements.

Overview

When BGP multipath is enabled, traffic can be forwarded to an IP prefix destination over multiple BGP paths that are considered equal by the BGP decision process. BGP multipath is supported in base router and VPRNs. The **multipath** command specifies the maximum number of BGP paths that each BGP RIB can submit to the route table for an IP prefix. The equal cost multipath (ECMP) limit defines how many paths are selected for installation in the forwarding information base (FIB). Traffic in the data path that matches the IP prefix is load-balanced across the ECMP next hops on a per-packet hash calculation.

➡ **Note:** As described in chapter Separate BGP RIBs for Labeled Routes, labeled routes and unlabeled routes do not mix.

The BGP multipath enhancements in SR OS release 14.0.R1 are the following:

1. The **multipath** command in the base router and VPRN BGP contexts contains the following options:

   multipath *max-paths* [**ebgp** *ebgp-max-paths*] [**ibgp** *ibgp-max-paths*] [**restrict** {**same-neighbor-as | exact-as-path**}]

   – **max-paths** is the default maximum number of paths. It is overruled by **ebgp-max-paths** and **ibgp-max-paths**. However, if there is no maximum set for the number of eBGP paths or iBGP paths, then the maximum number of paths is set by **max-paths**.

   – **ebgp-max-paths** specifies the maximum number of paths that can be used when the best path is eBGP.

   – **ibgp-max-paths** specifies the maximum number of paths that can be used when the best path is iBGP.

   – **restrict same-neighbor-as** forces multipaths to have the same AS path length (unless **as-path-ignore** is configured) and the same neighbor AS.

   – **restrict exact-as-path** forces multipaths to have the exact same AS paths.

```
*A:PE-5# configure router bgp multipath
  - multipath <max-paths> [ebgp <ebgp-max-paths>] [ibgp <ibgp-max-paths>] [restrict
    {same-neighbor-as|exact-as-path}]
  - no multipath

 <max-paths>          : [1..16]
 <ebgp-max-paths>     : [1..16]
 <ibgp-max-paths>     : [1..16]
```

2. The **ebgp-ibgp-equal** command is added to the **best-path-selection** contexts in base router and VPRN BGP contexts. When this command is configured, as follows, the BGP decision process skips the step that prefers eBGP over iBGP. This enables load-balancing between eBGP and iBGP paths. The earlier **eibgp-loadbalance** command, from previous releases, can still be used, but it is only applicable in a VPRN, not in the base router.

```
*A:PE-5# configure router bgp best-path-selection
  - best-path-selection

 [no] always-compare* - Determine how the Multi-Exit Discriminator (MED) path
                        attribute is used in the BGP route selection process
 [no] as-path-ignore  - Determine whether the AS Path is used in determining the
                        best BGP route
 [no] compare-origin* - Enable/Disable compare validation state
 [no] deterministic-* - Enable/Disable deterministic Multi-Exit Discriminator
 [no] ebgp-ibgp-equal - Determine whether EBGP and IBGP learned paths are
                        considered equal
 [no] ignore-nh-metr* - Enable/Disable ignore next-hop metric
 [no] ignore-router-* - Enable/Disable ignore router-id
 [no] origin-invalid* - Enable/Disable origin invalid unusable routes.
```

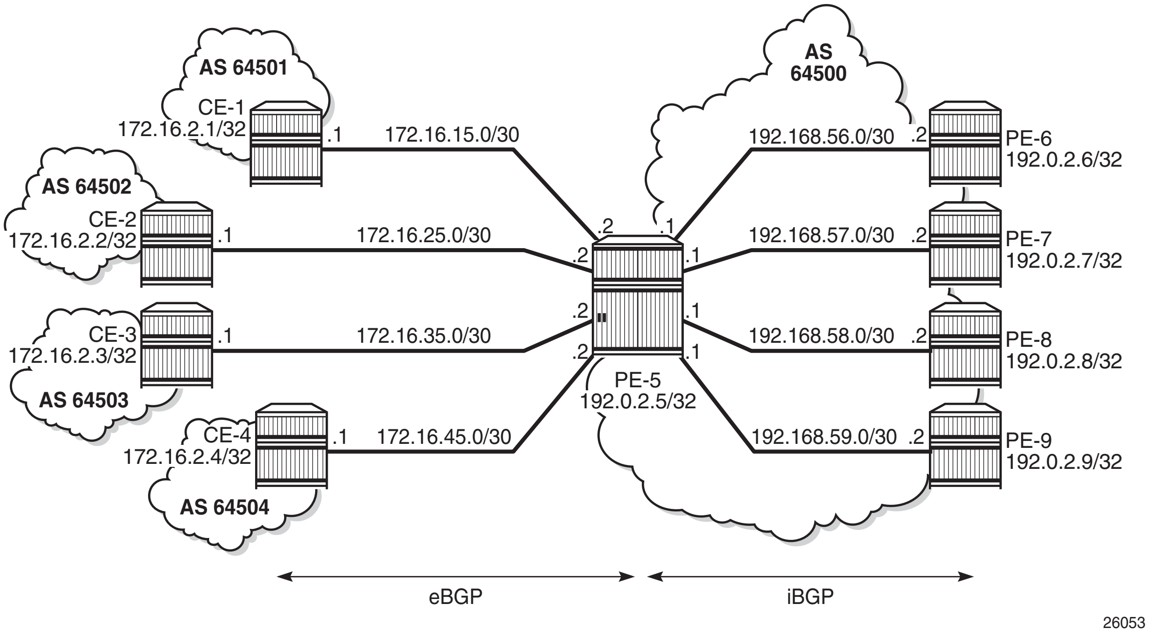
The **eibgp-loadbalance** command in a VPRN is used to provide ECMP over BGP-VPN (imported routes) and BGP routes. It is called eibgp-loadbalance because, in such scenarios, BGP-VPN is typically used between iBGP peers and BGP is used between eBGP peers. However, this is not always the case, so the name can be misleading.

Configuration

The examples in this section show the BGP configuration in the base router. For BGP multipath in a VPRN, the configuration is similar.

Figure 111 shows the example configuration with the used IP addresses. PE-5 has eBGP sessions with CEs in different autonomous systems (ASs) and iBGP sessions with PE-6, PE-7, PE-8, and PE-9.

Figure 111 **Example Topology**



The initial configuration includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS in AS 64500
- LDP in AS 64500

- BGP on all nodes (eBGP between CEs and PE-5; iBGP between PEs)
- Export policy "export-bgp" accepting routes from protocol direct on all nodes

The BGP configuration on CE-1 is as follows:

```
configure
    router
        autonomous-system 64501
        bgp
            min-route-advertisement 1
            group "eBGP"
                export "export-bgp"
                peer-as 64500
                neighbor 172.16.15.2
                exit
            exit
        exit
    exit
exit
```

The BGP configuration on the other nodes that advertise routes to PE-5 is similar.

The BGP configuration on PE-5 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            group "eBGP"
                neighbor 172.16.15.1
                    peer-as 64501
                exit
                neighbor 172.16.25.1
                    peer-as 64502
                exit
                neighbor 172.16.35.1
                    peer-as 64503
                exit
                neighbor 172.16.45.1
                    peer-as 64504
                exit
            exit
            group "iBGP"
                peer-as 64500
                neighbor 192.0.2.6
                exit
                neighbor 192.0.2.7
                exit
                neighbor 192.0.2.8
                exit
                neighbor 192.0.2.9
                exit
            exit
        exit
    exit
```

```
exit
```

The following will be configured and verified:

- BGP multipath with different limits for eBGP and iBGP paths
- BGP multipath with equal treatment for eBGP and iBGP paths
- BGP multipath restricted to the same neighbor AS
- BGP multipath restricted to the exact AS path
- EIBGP load-balancing in a VPRN

BGP Multipath with Different eBGP and iBGP Limits

On PE-5, BGP multipath is configured as follows:

```
*A:PE-5# configure router bgp multipath 8 ebgp 2 ibgp 3
```

It is mandatory to specify a maximum for BGP multipaths, as follows, but that is overruled by the individual limits for eBGP and iBGP. It is optional to configure limits for eBGP and iBGP.

```
*A:PE-5# configure router bgp multipath ebgp 2 ibgp 3
                                         ^
Error: Missing parameter
```

It is allowed to specify a lower value for multipath than for either eBGP or iBGP, as follows:

```
*A:PE-5# configure router bgp multipath 1 ebgp 2 ibgp 3
```

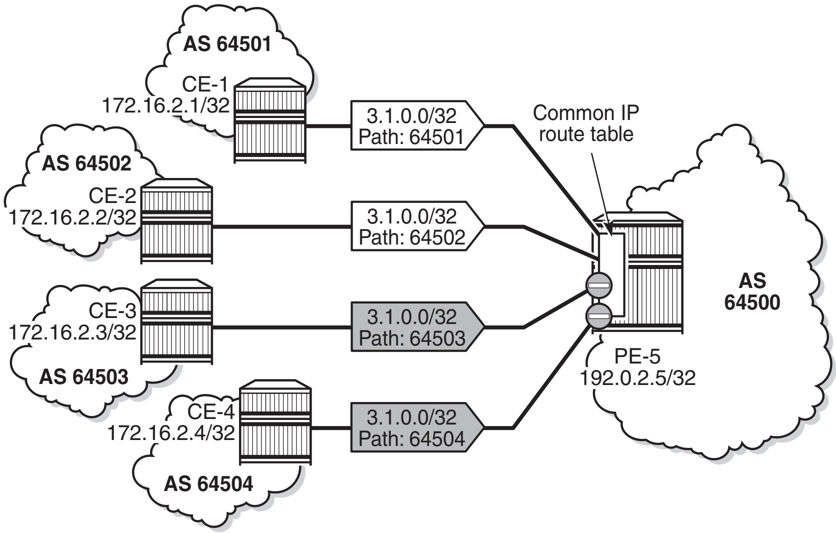
With this configuration, regardless of the value of multipath, there can be two eBGP routes for the same prefix and three iBGP routes for the same prefix. If the best route is eBGP, the multipath value is 2; if the best route is iBGP, the multipath value is 3. The value for multipath (1) is never used when limits for both eBGP and iBGP are configured.

```
*A:PE-5# configure router bgp multipath 3 ebgp 2
```

If the best route is eBGP, the multipath value is 2, and if the best route is iBGP, the multipath value is 3.

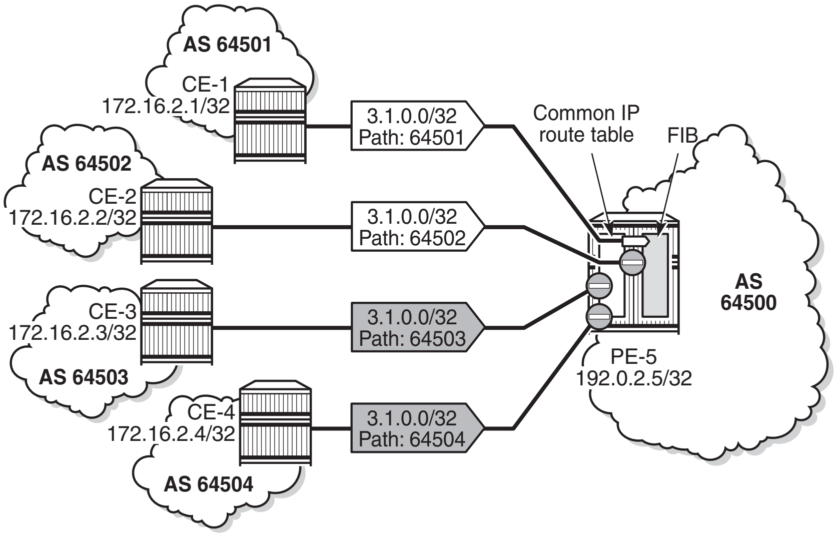
In this example, all four eBGP neighbors advertise prefix 3.1.0.0/32 to PE-5 and all four iBGP neighbors advertise prefix 3.2.0.0/32 to PE-5. PE-5 receives four eBGP routes for prefix 3.1.0.0/32, but only two will be added to the common IP route table, as shown in Figure 112.

Figure 112 **BGP Multipath with eBGP Limit 2**



These routes can only be added to the FIB if ECMP is configured to a value at least equal to the number of routes allowed in BGP multipath. By default, ECMP is disabled and only one route is added to the FIB, as shown in Figure 113.

Figure 113 **eBGP Multipath with Limit 2 and ECMP Disabled**



With ECMP disabled, only one of the four paths is active for prefix 3.1.0.0/32, as follows:

```
*A:PE-5# show router bgp routes 3.1.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.1.0.0/32                                         None        None
      172.16.15.1                                        None        -
      64501
*i    3.1.0.0/32                                         None        None
      172.16.25.1                                        None        -
      64502
*i    3.1.0.0/32                                         None        None
      172.16.35.1                                        None        -
      64503
*i    3.1.0.0/32                                         None        None
      172.16.45.1                                        None        -
      64504
-------------------------------------------------------------------------------
Routes : 4
```

In the remainder of the chapter, ECMP will be configured with a value of eight, implying that the routes added to the common IP route table will be added to the FIB as well. ECMP is configured on PE-5 as follows:

```
*A:PE-5# configure router ecmp 8
```

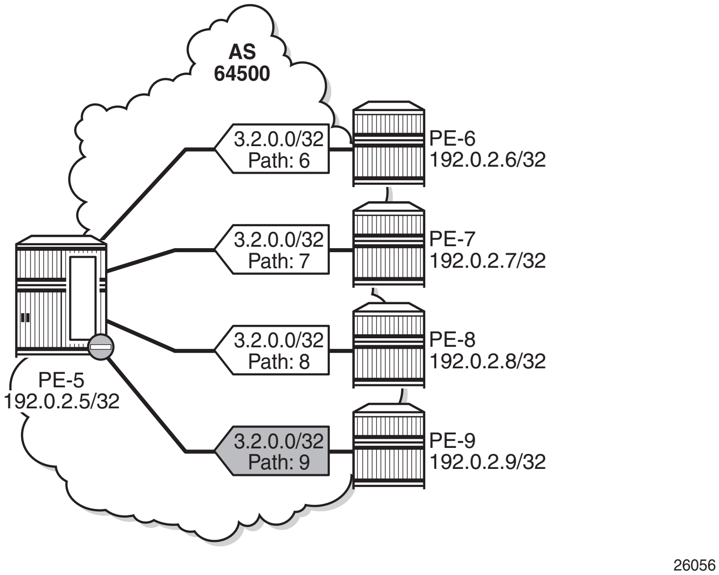
With ECMP configured with a limit of eight, two eBGP paths are active for prefix 3.1.0.0/32.

The first two of the following BGP routes for prefix 3.1.0.0/32 are used on PE-5:

```
*A:PE-5# show router bgp routes 3.1.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.1.0.0/32                                         None        None
      172.16.15.1                                        None        -
      64501
u*>i  3.1.0.0/32                                         None        None
      172.16.25.1                                        None        -
      64502
*>i   3.1.0.0/32                                         None        None
      172.16.35.1                                        None        -
      64503
*>i   3.1.0.0/32                                         None        None
      172.16.45.1                                        None        -
      64504
-------------------------------------------------------------------------------
Routes : 4
```

The four iBGP neighbors of PE-5 advertise prefix 3.2.0.0/32 to PE-5. BGP multipath has a limit of three for iBGP routes. Consequently, three BGP routes will be added to the common IP route table and to the FIB, as shown in Figure 114.

Figure 114 **BGP Multipath with iBGP Limit 3 and ECMP Limit 8**



Three iBGP paths are active for prefix 3.2.0.0/32, as follows:

```
*A:PE-5# show router bgp routes 3.2.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
```

```
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.2.0.0/32                                         100         None
      192.0.2.6                                          None        -
      6
u*>i  3.2.0.0/32                                         100         None
      192.0.2.7                                          None        -
      7
u*>i  3.2.0.0/32                                         100         None
      192.0.2.8                                          None        -
      8
*>i   3.2.0.0/32                                         100         None
      192.0.2.9                                          None        -
      9
-------------------------------------------------------------------------------
Routes : 4
```

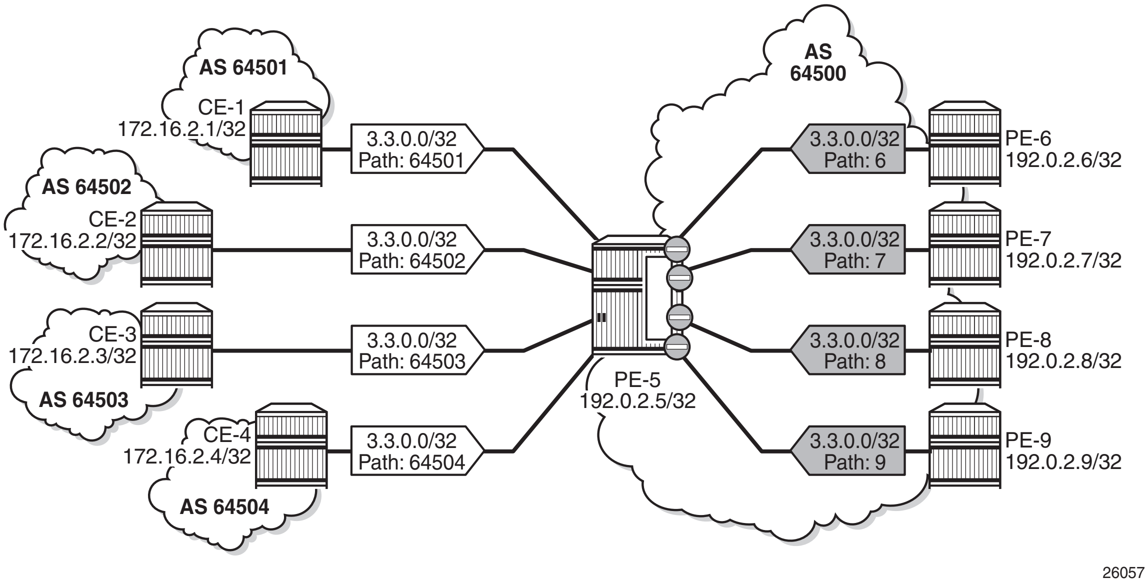
BGP Multipath with eBGP Equal to iBGP

It is optional to specify limits for eBGP and iBGP; an overall multipath limitation is sufficient, such as:

```
*A:PE-5# configure router bgp multipath 6
```

With this configuration, there can be six routes for the same prefix. These routes can be eBGP or iBGP routes. By default, eBGP routes are preferred and, therefore, only the four eBGP routes will be imported in the common IP route table, as shown in Figure 115.

Figure 115 **BGP Multipath with Limit 6 and eBGP Preferred**



Only the four eBGP paths are active, as follows:

```
*A:PE-5# show router bgp routes 3.3.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.0.0/32                                         None        None
      172.16.15.1                                        None        -
      64501
u*>i  3.3.0.0/32                                         None        None
      172.16.25.1                                        None        -
      64502
u*>i  3.3.0.0/32                                         None        None
      172.16.35.1                                        None        -
      64503
u*>i  3.3.0.0/32                                         None        None
      172.16.45.1                                        None        -
      64504
*i    3.3.0.0/32                                         100         None
      192.0.2.6                                          None        -
      6
```

```
*i    3.3.0.0/32                                            100         None
      192.0.2.7                                             None        -
      7
*i    3.3.0.0/32                                            100         None
      192.0.2.8                                             None        -
      8
*i    3.3.0.0/32                                            100         None
      192.0.2.9                                             None        -
      9
-------------------------------------------------------------------------------
Routes : 8
```

The BGP decision process prefers eBGP over iBGP, but this step can be skipped by configuring the following:

```
*A:PE-5# configure router bgp best-path-selection ebgp-ibgp-equal ipv4
```

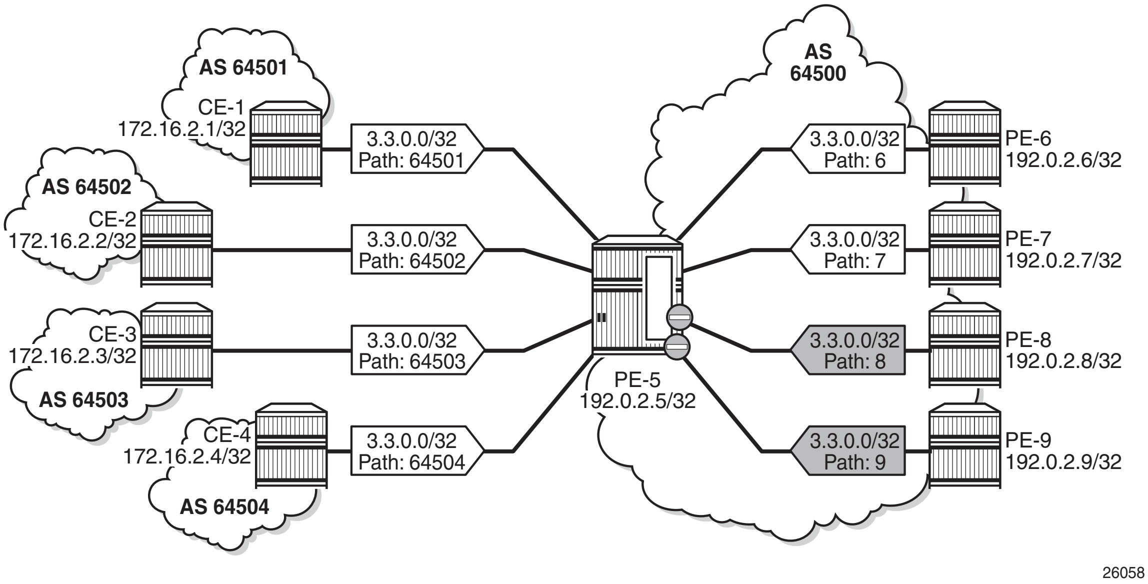
This configuration only skips one step in the BGP decision process. If the best route is still eBGP, the eBGP multipath limit applies; if the best route is iBGP, the iBGP multipath limit applies.

Optionally, other best path selection criteria can also be configured, such as ignore-nh-metric, as follows:

```
*A:PE-5# configure router bgp best-path-selection ignore-nh-metric
```

However, when all other path options are identical (such as local preference, MED, IGP cost, and other criteria from the BGP decision process), or when the best-path-selection is configured to ignore specific path options, and the only differentiator is an originator ID, the remaining steps in the BGP decision process do not exclude any routes. In that case, six of the eight eligible BGP paths will be included in the BGP multipath, as shown in Figure 116.

Figure 116   **BGP Multipath with Limit 6, eBGP Equal to iBGP, and Other Path Options Identical**



From the eight advertised BGP routes for prefix 3.3.0.0/32, six paths will be active, as follows:

```
*A:PE-5# show router bgp routes 3.3.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.0.0/32                                         None        None
      172.16.15.1                                        None        -
      64501
u*>i  3.3.0.0/32                                         None        None
      172.16.25.1                                        None        -
      64502
u*>i  3.3.0.0/32                                         None        None
      172.16.35.1                                        None        -
      64503
u*>i  3.3.0.0/32                                         None        None
      172.16.45.1                                        None        -
      64504
u*>i  3.3.0.0/32                                         100         None
```

```
      192.0.2.6                                               None        -
      6
u*>i  3.3.0.0/32                                              100         None
      192.0.2.7                                               None        -
      7
*>i   3.3.0.0/32                                              100         None
      192.0.2.8                                               None        -
      8
*>i   3.3.0.0/32                                              100         None
      192.0.2.9                                               None        -
      9
-------------------------------------------------------------------------------
Routes : 8
```

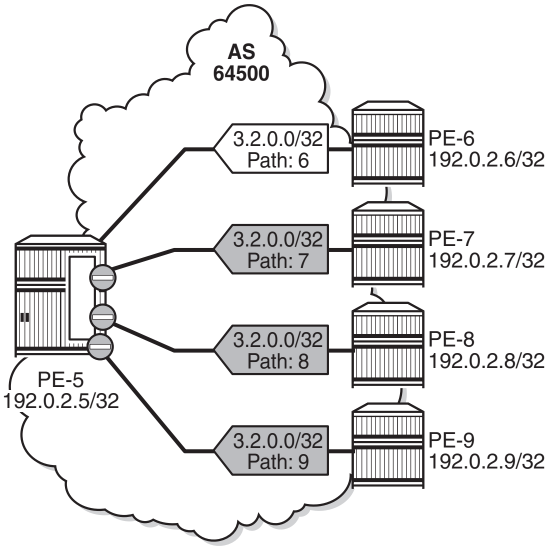
BGP Multipath Restricted to the Same Neighbor AS

BGP multipath can be configured with the restriction that the neighbor AS must be the same for all the active paths. When all routes have a different neighbor AS, only one path will be active. This can be shown for prefix 3.2.0.0/32 that is advertised by the iBGP neighbors. The BGP multipath configuration on PE-5 is as follows:

```
*A:PE-5# configure router bgp multipath 8 ebgp 2 ibgp 3 restrict same-neighbor-as
```

Figure 117 shows that with the restriction to the same neighbor AS, only one path is active because all BGP routes have a different neighbor AS.

Figure 117 **BGP Multipath Configured with Restriction to the Same Neighbor AS**



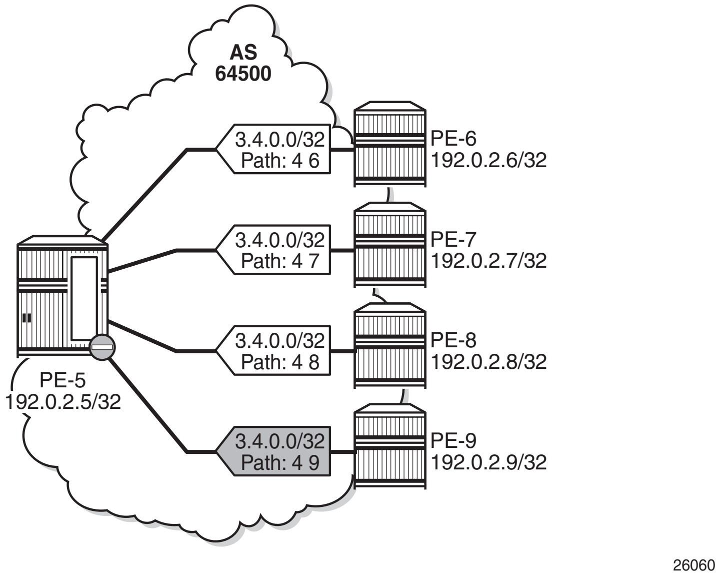
26059

Only one BGP path will be active, because all the other routes have a different neighbor AS, as follows:

```
*A:PE-5# show router bgp routes 3.2.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.2.0.0/32                                         100         None
      192.0.2.6                                          None        -
      6
*>i   3.2.0.0/32                                         100         None
      192.0.2.7                                          None        -
      7
*>i   3.2.0.0/32                                         100         None
      192.0.2.8                                          None        -
      8
*>i   3.2.0.0/32                                         100         None
      192.0.2.9                                          None        -
      9
-------------------------------------------------------------------------------
Routes : 4
```

Figure 118 shows that the iBGP neighbors also advertise prefix 3.4.0.0/32 with a different AS path, but the AS path is equally long and the neighbor AS is the same. Three of these BGP paths will be active.

Figure 118 BGP Multipath Restricted to the Same Neighbor AS: AS Paths with Same Length



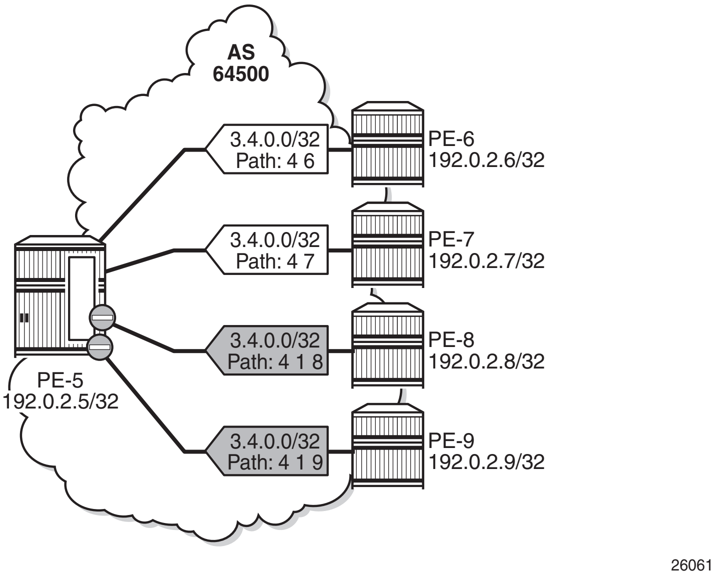
All iBGP neighbors have the same neighbor AS and an AS path of equal length. Three of the iBGP paths for prefix 3.4.0.0/32 are active, as follows:

```
*A:PE-5# show router bgp routes 3.4.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.4.0.0/32                                         100         None
      192.0.2.6                                          None        -
      4 6
u*>i  3.4.0.0/32                                         100         None
      192.0.2.7                                          None        -
      4 7
u*>i  3.4.0.0/32                                         100         None
      192.0.2.8                                          None        -
      4 8
*>i   3.4.0.0/32                                         100         None
      192.0.2.9                                          None        -
```

```
     4 9
-------------------------------------------------------------------------------
Routes : 4
```

The restriction that the neighbor AS must be the same does not overrule the BGP selection criterion that the shorter AS path is preferred. When the AS path is longer for the routes advertised by neighbors 192.0.2.8 and 192.0.2.9, only the BGP paths with the shorter AS path will be active, as shown in Figure 119.

Figure 119 **BGP Multipath Restricted to the Same Neighbor AS: AS Paths of Different Lengths**

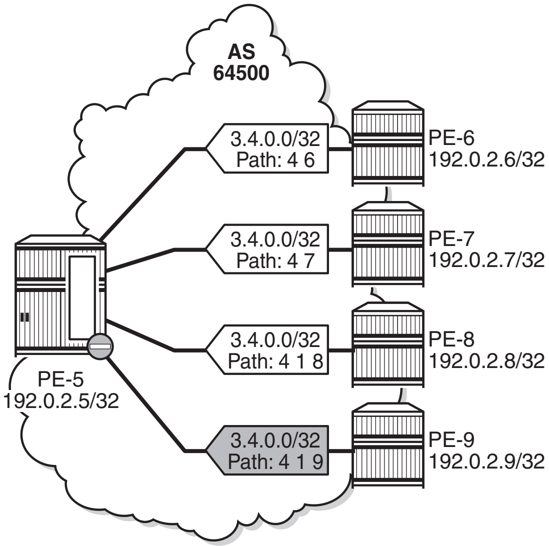


All BGP routes advertised by the iBGP neighbors have the same neighbor AS, but the AS path is longer for neighbors 192.0.2.8 and 192.0.2.9. The routes advertised by these neighbors will not be selected as best path and will not be added to the route table. Only the two BGP routes with the shorter AS path will be active, as follows:

```
*A:PE-5# show router bgp routes 3.4.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.4.0.0/32                                             100         None
      192.0.2.6                                              None        -
      4 6
u*>i  3.4.0.0/32                                             100         None
      192.0.2.7                                              None        -
      4 7
*i    3.4.0.0/32                                             100         None
      192.0.2.8                                              None        -
      4 1 8
*i    3.4.0.0/32                                             100         None
      192.0.2.9                                              None        -
      4 1 9
-------------------------------------------------------------------------------
Routes : 4
```

When the best path selection is configured to ignore the AS path, three paths will be active again, as shown in Figure 120.

Figure 120 **BGP Multipath Restricted to the Same Neighbor AS: AS Paths of Different Lengths, AS Path Ignored**



26062

The best path selection is reconfigured as follows:

```
*A:PE-5# configure router bgp best-path-selection as-path-ignore
```

Three of the four eligible BGP routes will be active, as follows:

```
*A:PE-5# show router bgp routes 3.4.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.4.0.0/32                                         100         None
      192.0.2.6                                          None        -
      4 6
u*>i  3.4.0.0/32                                         100         None
      192.0.2.7                                          None        -
      4 7
u*>i  3.4.0.0/32                                         100         None
      192.0.2.8                                          None        -
      4 1 8
*>i   3.4.0.0/32                                         100         None
      192.0.2.9                                          None        -
      4 1 9
-------------------------------------------------------------------------------
Routes : 4
```

The best selection path settings are restored as follows:

```
*A:PE-5# configure router bgp best-path-selection no as-path-ignore
```

BGP Multipath Restricted to the Exact AS Path

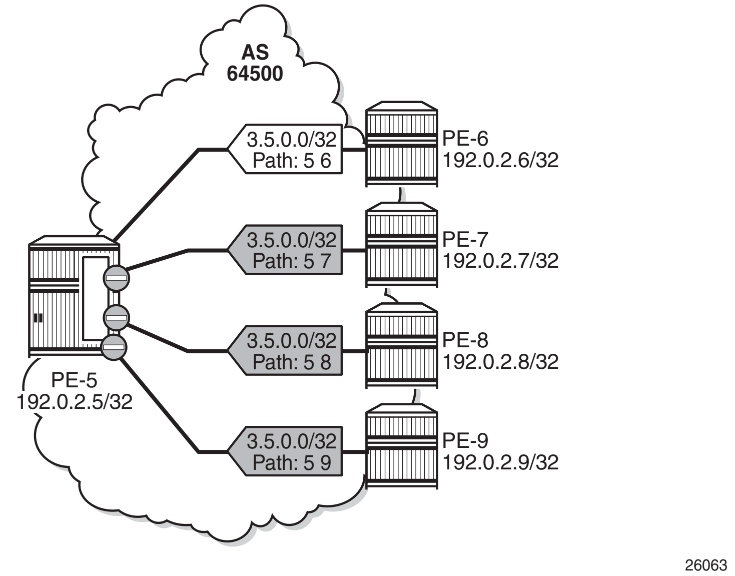
The BGP multipath configuration on PE-5 restricts BGP to only use identical AS paths, as follows:

```
*A:PE-5# configure router bgp multipath 8 ebgp 2 ibgp 3 restrict exact-as-path
```

The four iBGP neighbors advertise prefixes 3.5.0.0/32 and 3.6.0.0/32 to PE-5, see Figure 121 and Figure 122. The AS paths for prefix 3.5.0.0/32 are not identical, but the neighbor AS is the same, and the AS path is of equal length. The AS paths for prefix 3.6.0.0/32 are identical.

The BGP multipath configuration specifies that the AS paths must be identical, which is not the case for the received BGP routes for prefix 3.5.0.0/32. Only one BGP route will be imported in the route table, as shown in Figure 121.

Figure 121 **BGP Multipath Restricted to Exact Same AS. All AS Paths are Different.**

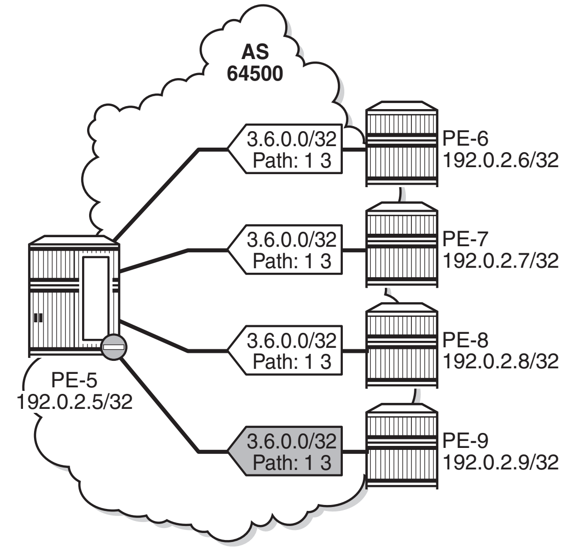


All the BGP routes for prefix 3.5.0.0/32 have a different AS path. Only the BGP route advertised by neighbor 192.0.2.6 will be active, as follows:

```
*A:PE-5# show router bgp routes 3.5.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.5.0.0/32                                         100         None
      192.0.2.6                                          None        -
      5 6
*>i   3.5.0.0/32                                         100         None
      192.0.2.7                                          None        -
      5 7
*>i   3.5.0.0/32                                         100         None
      192.0.2.8                                          None        -
      5 8
*>i   3.5.0.0/32                                         100         None
      192.0.2.9                                          None        -
      5 9
-------------------------------------------------------------------------------
Routes : 4
```

However, all the received BGP routes for prefix 3.6.0.0/32 have the same AS path. Three of these BGP paths will become active, as shown in Figure 122.

Figure 122 **BGP Multipath Restricted to Exact Same AS. All AS Paths are Identical**



26064

Three of the four received BGP routes for prefix 3.6.0.0/32 are used, as follows:

```
*A:PE-5# show router bgp routes 3.6.0.0/32
===============================================================================
 BGP Router ID:192.0.2.5        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.6.0.0/32                                         100         None
      192.0.2.7                                          None        -
      1 3
u*>i  3.6.0.0/32                                         100         None
```

```
      192.0.2.8                                               None         -
      1 3
u*>i  3.6.0.0/32                                              100          None
      192.0.2.9                                               None         -
      1 3
*i    3.6.0.0/32                                              100          None
      192.0.2.6                                               None         -
      1 3
-------------------------------------------------------------------------------
Routes : 4
```

EIBGP Load-Balancing in a VPRN

The **eibgp-loadbalance** command is used to perform ECMP over BGP-VPN (imported routes) and BGP routes, not to make eBGP routes equal to iBGP routes. Typically, BGP-VPN routes are used between iBGP peers and BGP routes are used between eBGP peers. This explains the naming of the command, but it can be misleading. Not all BGP-VPN routes are iBGP routes and not all BGP routes are eBGP routes. Four different types of routes are received by PE-5 in the following example:

- two BGP routes learned from eBGP peers CE-1 and CE-2
- two BGP-VPN routes learned from eBGP peers CE-3 and CE-4
- two BGP routes learned from iBGP peers PE-6 and PE-7
- two BGP-VPN routes learned from iBGP peers PE-8 and PE-9

VPRN 1 is configured on all nodes. On PE-5, ECMP must be configured in this VPRN to a value of 8, as follows:

```
*A:PE-5# configure service vprn 1 ecmp 8
```

In this example, the best path selection criteria will be such that iBGP routes are treated as equal to eBGP routes, as explained in section BGP Multipath with eBGP Equal to iBGP. This is configured for BGP routes and for VPN-BGP routes, as follows:

```
*A:PE-5# configure router bgp best-path-selection ebgp-ibgp-equal vpn-ipv4
*A:PE-5# configure service vprn 1 bgp best-path-selection ebgp-ibgp-equal ipv4
```

Optionally, other best path selection criteria can also be configured, such as ignore-nh-metric and as-path-ignore, as follows:

```
*A:PE-5# configure router bgp best-path-selection as-path-ignore vpn-ipv4
*A:PE-5# configure service vprn 1 bgp best-path-selection as-path-ignore ipv4

*A:PE-5# configure router bgp best-path-selection ignore-nh-metric
*A:PE-5# configure service vprn 1 bgp best-path-selection ignore-nh-metric
```

BGP multipath is configured with an eBGP multipath limit of 4 and an iBGP multipath limit of 4, as follows:

```
*A:PE-5# configure service vprn 1 bgp multipath 4 ebgp 4 ibgp 4
```

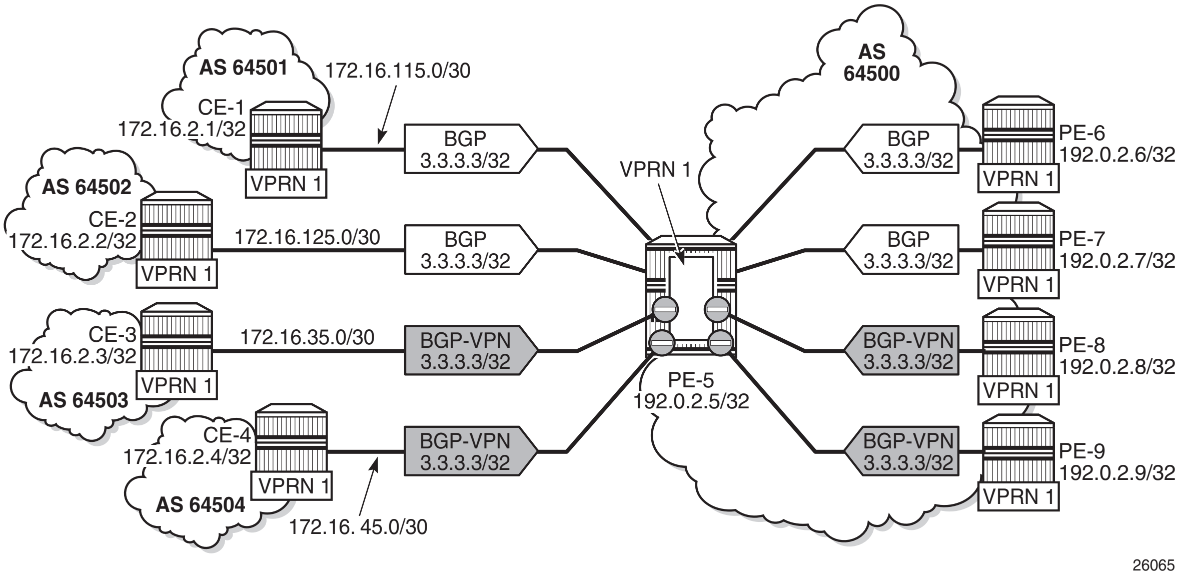
The **eibgp-loadbalance** command is not configured yet. By default, the BGP decision process prefers BGP routes over BGP-VPN routes, and therefore, the routing table only contains BGP routes and no BGP-VPN routes, as follows:

```
*A:PE-5# show router 1 route-table 3.3.3.3/32
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
3.3.3.3/32                                    Remote  BGP       00h01m11s  170
       172.16.115.1                                                 0
3.3.3.3/32                                    Remote  BGP       00h01m11s  170
       172.16.125.1                                                 0
3.3.3.3/32                                    Remote  BGP       00h01m11s  170
       192.168.156.2                                                0
3.3.3.3/32                                    Remote  BGP       00h01m11s  170
       192.168.157.2                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
```

The first two routes in this routing table are eBGP routes learned from CE-1 and CE-2. The latter two routes are iBGP routes learned from PE-6 and PE-7. EBGP routes and iBGP routes are treated as equal, but BGP routes are still preferred to BGP-VPN routes.

Figure 123 shows that the BGP routes are preferred. Even though it is allowed to have a maximum of four eBGP routes and four iBGP routes, only two eBGP routes are included and only two iBGP routes are included. No BGP-VPN routes are used, as follows.

Figure 123 **EBGP Equal to IBGP: No EIBGP Load-Balancing**



BGP-VPN routes will be treated equally to BGP routes after configuring the following command:

```
*A:PE-5# configure service vprn 1 bgp eibgp-loadbalance
```

When BGP-VPN routes are treated equal to BGP routes, the routing table for prefix 3.3.3.3/32 contains eight entries, as follows:

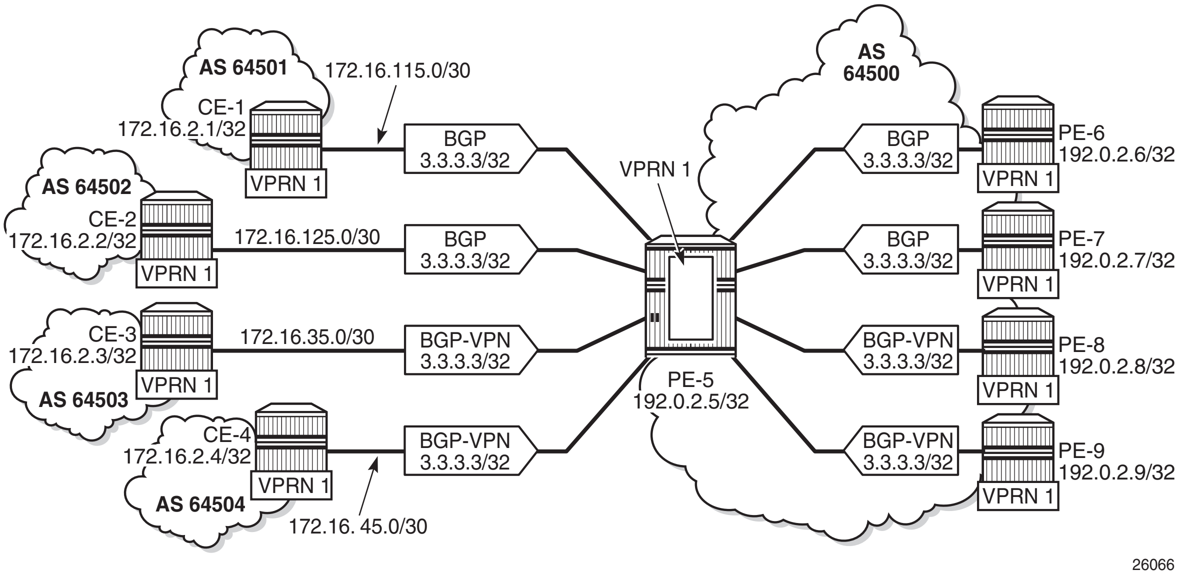
```
*A:PE-5# show router 1 route-table 3.3.3.3/32
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
3.3.3.3/32                                    Remote  BGP VPN   00h00m04s  170
       172.16.35.1 (tunneled)                                        0
3.3.3.3/32                                    Remote  BGP VPN   00h00m04s  170
       172.16.45.1 (tunneled)                                        0
3.3.3.3/32                                    Remote  BGP       00h00m04s  170
       172.16.115.1                                                  0
3.3.3.3/32                                    Remote  BGP       00h00m04s  170
       172.16.125.1                                                  0
3.3.3.3/32                                    Remote  BGP VPN   00h00m04s  170
       192.0.2.8 (tunneled)                                          0
3.3.3.3/32                                    Remote  BGP VPN   00h00m04s  170
       192.0.2.9 (tunneled)                                          0
3.3.3.3/32                                    Remote  BGP       00h00m04s  170
       192.168.156.2                                                 0
3.3.3.3/32                                    Remote  BGP       00h00m04s  170
       192.168.157.2                                                 0
```

```
-------------------------------------------------------------------------------
No. of Routes: 8
```

The first two BGP-VPN entries are eBGP routes learned from CE-3 and CE-4 and the other two BGP-VPN entries are iBGP routes learned from PE-8 and PE-9.

Figure 124 shows that when EIBGP load-balancing is configured, the BGP-VPN routes are equal to the BGP routes. All other selection criteria are also equal in this example and all eight routes are used in the routing table, as follows.

Figure 124 **EBGP Equal to IBGP: EIBGP Load-Balancing Enabled**



This example shows that the **eibgp-loadbalance** command can be used in combination with **best-path-selection ebgp-ibgp-equal**, because their scope is different.

Conclusion

BGP multipath allows the IP routing table to have multiple BGP paths to the same destination. Different path limits can be applied for eBGP and iBGP paths. It is possible to treat eBGP and iBGP routes as equal. Restrictions can be imposed related to AS path.

BGP Optimal Route Reflection for Hierarchical Networks

This chapter provides information about BGP Optimal Route Reflection for Hierarchical Networks.

Topics in this chapter include:

Applicability

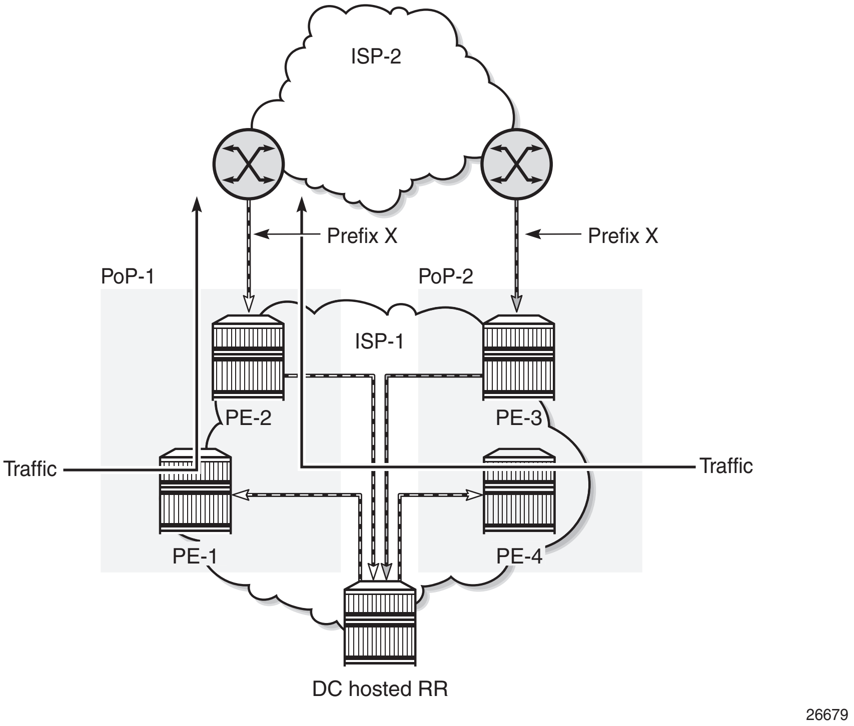
The information and configuration in this chapter are based on SR OS Release 15.0.R4.

Overview

BGP route reflectors are used in many networks. They improve network scalability by eliminating or reducing the need for a full-mesh of iBGP sessions.

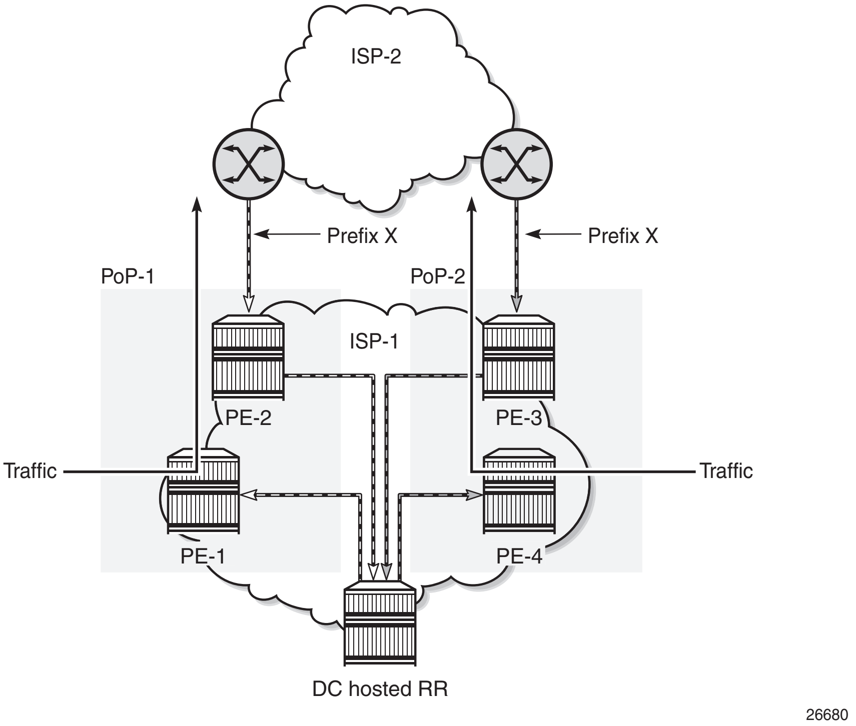
When a BGP route reflector receives multiple paths for the same IP destination, it normally selects and reflects a single best path in its routing domain to all clients in that domain, based on its own location in the domain. In Figure 125, the centralized route reflector RR for ISP-1 is located in the datacenter (DC), and receives prefix X from ISP-2 through PE-2 in Point of Presence (PoP) -1 and also through PE-3 in PoP-2. RR selects and reflects PE-2 as the best path to the remaining route reflector clients because RR is closer to PoP-1 than it is to PoP-2, so the traffic to destination X flows as indicated. Therefore, sending traffic to another autonomous system (AS) through the closest possible exit point from the local AS, known as hot-potato routing, cannot be achieved.

Figure 125 **Centralized Route Reflection**



Hot-potato routing can be achieved using a route reflector selecting and reflecting multiple best paths, for different subdomains and from point of view of a client in a subdomain, as outlined in *draft-ietf-idr-bgp-optimal-route-reflection* (ORR), and requires the route reflector to know the topology of each subdomain. In Figure 126, the route reflector calculates the best path for PoP-1 and reflects that to the clients in PoP-1 (PE-1), and it also calculates the best path for PoP-2 and reflects that to the clients in PoP-2 (PE-4).

Figure 126 **Centralized Route Reflection with ORR**



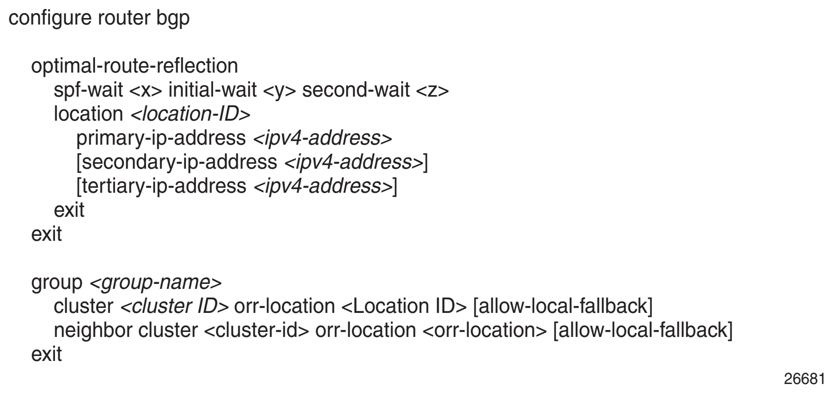
If the routing domain is flat, the route reflector is part of the routing domain and thus has a view on the entire topology through the Interior Gateway Protocol (IGP). See the BGP Optimal Route Reflection for Non-Hierarchical Networks chapter if the network topology is flat.

If the routing domain is hierarchical, the route reflector needs to extract the Link State Data Base (LSDB) from the subdomains it is not part of, which is achieved through BGP-Link-State (BGP-LS). The use of BGP-LS allows the route reflector to learn the IGP topology information for OSPF areas and IS-IS levels in which the route reflector is not a direct participant.

ORR CLI commands

The CLI command hierarchy used in the Nokia ORR implementation is shown in Figure 127. The BGP optimal-route-reflection context defines the Shortest Path First (SPF) parameters, and multiple locations. The locations are then referred to with the **cluster** command (residing in the BGP **group** context) through the **orr-location** argument.

Figure 127 **Configuring ORR in SR OS**



```
configure router bgp

  optimal-route-reflection
     spf-wait <x> initial-wait <y> second-wait <z>
     location <location-ID>
        primary-ip-address <ipv4-address>
        [secondary-ip-address <ipv4-address>]
        [tertiary-ip-address <ipv4-address>]
     exit
  exit

  group <group-name>
     cluster <cluster ID> orr-location <Location ID> [allow-local-fallback]
     neighbor cluster <cluster-id> orr-location <orr-location> [allow-local-fallback]
  exit
```

Up to sixteen locations can be created in the **optimal-route-reflection** context. Each location is identified through a Location-ID [1..16], and contains a **primary-ip-address** and, optionally, a **secondary-ip-address** and a **tertiary-ip-address**, for redundancy reasons. These addresses must correspond to loopback or system IP addresses of routers participating in the IGP protocols, and are used as the starting point (or seed) to start the SPF calculation. Because all clients in the same location receive the same optimal path for that location, these addresses should be close to the clients in that part of the network.

The SPF calculation is configurable with the **spf-wait** command. **Initial-wait** and **second-wait** are optional arguments. These timers define when to initiate the first, second, and subsequent SPF runs after a topology change occurs.

The Location-ID is referred to in the **orr-location** argument of the **cluster** command. Typically, **cluster** command applies to a BGP peer group; all neighbors in that group share the same Location-ID, unless the **cluster** command applies at a neighbor level. The **allow-local-fallback** option allows the RR to advertise the best reachable BGP path using its own location, but only when no BGP routes are reachable for some location. Otherwise, no path would be advertised to the clients in that location.

Properties

The following properties apply to ORR in SR OS:

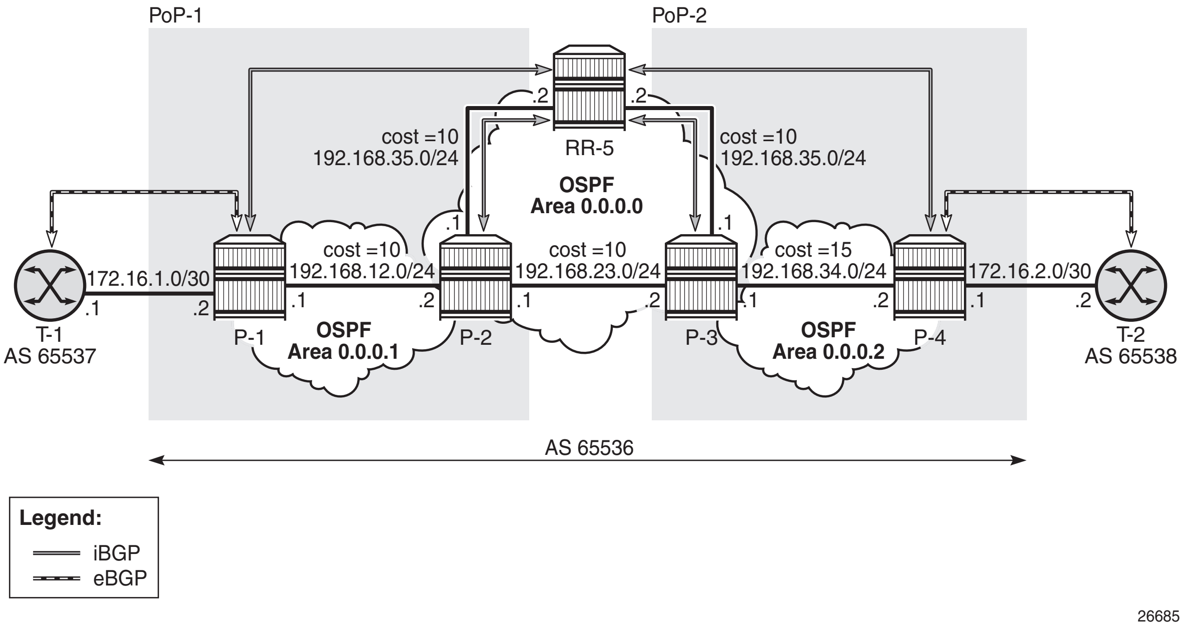
- ORR is supported in the Base router BGP instance.
- ORR is supported for the IPv4, label-IPv4, label-ipv6, VPN-IPv4, and VPN-IPv6 address families.

- ORR is supported with Add-Paths, meaning that Add-Paths advertised to ORR clients are also ORR location-based.

Configuration

Figure 128 shows the example topology. OSPF is used as the IGP for AS 65536, with RR-5 taking the role of the route reflector for clients P-1 to P-4. The OSPF backbone area is area 0.0.0.0, connecting routers P-2, P-3, and RR-5. Area 0.0.0.1 is a stub area interconnecting P-1 and P-2, and Area 0.0.0.2 is a stub area interconnecting P-3 and P-4. Both P-2 and P-3 are Area Border Routers (ABRs). Additionally, tester T-1 in AS 65537 peers with P-1, and tester T-2 in AS 65538 peers with P-4.

Figure 128 **Example Hierarchical Networking using OSPF**



The initial configuration on all nodes includes:

- Cards, MDAs and ports
- Router interfaces
- OSPF as IGP on all interfaces within AS 65536, with multiple non-backbone areas (alternatively, IS-IS can be used), and traffic engineering enabled

Route Reflection without ORR

RR-5 peers with clients P-1 to P-4, and because RR-5 is the route reflector, the **cluster** command is added, defining the cluster ID attribute value to use. The configuration for RR-5 is as follows:

```
# on RR-5
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                cluster 192.0.2.5
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-1 belongs to the cluster defined in the route reflector, so it does not need to be fully meshed with the other routers in the area; peering with the route reflectors in the area is sufficient for P-1 to receive updates. Typically, two route reflectors are provisioned for redundancy, but that does not apply in this example. P-1 also peers with T-1 in AS 65537 through eBGP, so the P-1 configuration is as follows:

```
# on P-1
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "eBGP"
                peer-as 65537
                neighbor 172.16.1.1
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-2 and P-3 also belong to the cluster defined in the route reflector, they only peer with the route reflector. Their configuration is as the same:

```
# on P2 and P3
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                next-hop-self
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-4 also belongs to the cluster defined in the route reflector, but peers with T-2 in AS 65538. The P-4 configuration is similar to the configuration of P-1.

T-1 announces prefix 10.0.11.0/24 to router P-1, and T-2 announces the same prefix to router P-4. As a result, traffic offered to P-1 for destination 10.0.11.0/24 is routed to T-1, as follows:

```
*A:P-1# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h14m08s  170
       172.16.1.1                                                   0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-1#
```

Traffic offered to P-2 for destination 10.0.11.0/24 is routed to P-1, as follows:

```
*A:P-2# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
```

```
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h10m41s   170
       192.168.12.1                                                 0
--------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:P-2#
```

Traffic offered to P-3 for destination 10.0.11.0/24 is routed to P-2, as follows:

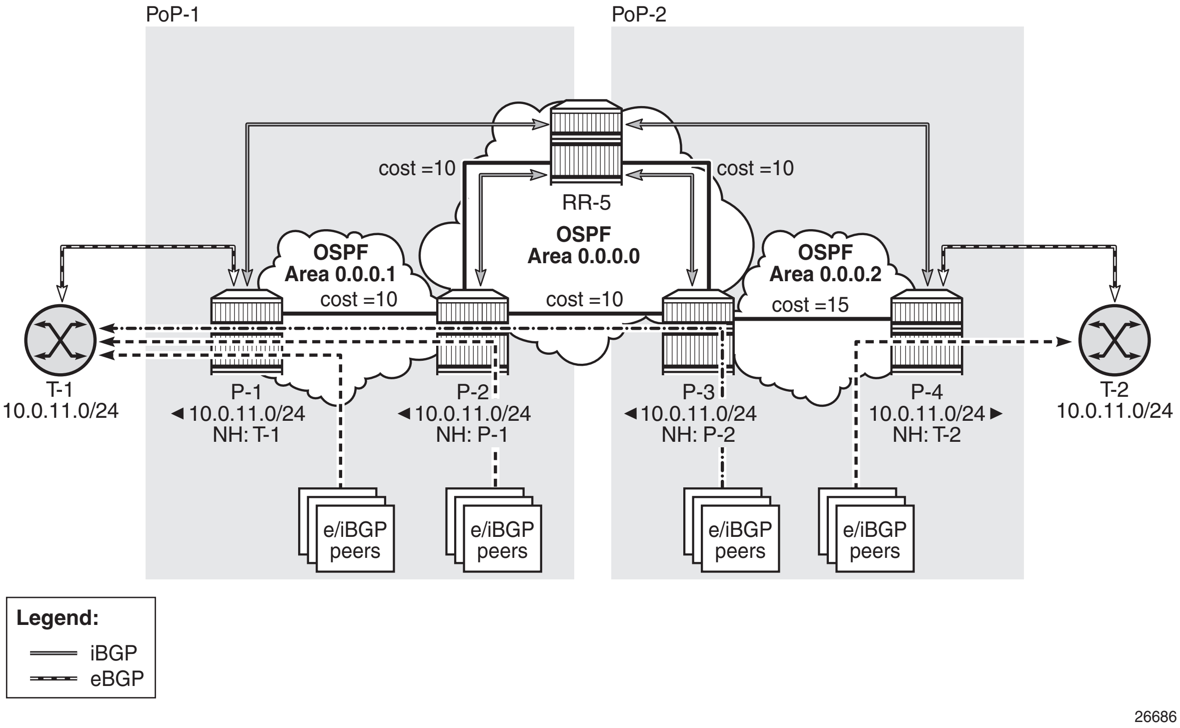
```
*A:P-3# show router route-table protocol bgp
================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h09m49s   170
       192.168.23.1                                                 0
--------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:P-3#
```

Traffic offered to P-4 for destination 10.0.11.0/24 is routed to T-2, as follows:

```
*A:P-4# show router route-table protocol bgp
================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h09m31s   170
       172.16.2.2                                                   0
--------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
================================================================================
*A:P-4#
```

This is summarized in Figure 129. RR-5 receives the updates from P-1 and P-4, and selects the best path based on its own position in the topology. The cost from RR-5 to P-1 is 20, and the cost from RR-5 to P-4 is 25, so RR-5 selects P-1 and reflects this path to all clients (except to P-1, because P-1 is the client where the path was learned from). Ultimately, P-1 only has one path, and so do P-2 and P-3. P-4 has two paths, but by default prefers the eBGP learned path over the iBGP learned path.

Figure 129 **Suboptimal Route Reflection**



Route Reflection with ORR

Implementing ORR using the hierarchical topology from Figure 5 requires changes in the non-backbone OSPF areas as well as changes to the route reflector.

Because the route reflector is part of the backbone area, and ABRs do not pass the LSAs describing the topology and the traffic engineering data for the non-backbone areas, that data must be extracted from the non-backbone areas and copied to the route reflector. This is achieved using BGP-LS, with additional support from OSPF.

In this example, BGP-LS is activated in P-1, P-4, and RR-5. P1 in Area 0.0.0.1 has **bgp-ls** activated with the **family** command. **Link-state-import-enable** is needed for P-1 to announce the LSDB and Traffic Engineering Database (TED) to the route reflector. On the same router P-1, OSPF is instructed to provide the **bgp-ls-identifier 1** using the **database-export** command. The configuration for P-1 is as follows:

```
# on P-1
configure
    router
        autonomous-system 65536
        ospf
            traffic-engineering
            database-export identifier 1 bgp-ls-identifier 1
            area 0.0.0.1
                --- snipped ---
            exit
            no shutdown
        exit
        bgp
            family ipv4 bgp-ls
            loop-detect discard-route
            split-horizon
            link-state-import-enable
            group "eBGP"
                peer-as 65537
                neighbor 172.16.1.1
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The configuration on P-4 is similar, and there the **bgp-ls-identifier** is set to 2. Routers P-2 and P-3 do not need to be reconfigured.

RR-5 in the backbone area also has **bgp-ls** activated with the **family** command, and **link-state-export-enable** is required for accepting and storing the LSDB and TED. No reconfiguration of OSPF is required in RR-5. The configuration for RR-5 is as follows:

```
# on RR-5
configure
    router
        autonomous-system 65536
        bgp
            family ipv4 bgp-ls
            loop-detect discard-route
            split-horizon
            link-state-export-enable
            group "iBGP"
                --- snipped ---
            exit
            no shutdown
        exit
    exit
exit
```

With these changes applied, the following command can be used for verification:

```
*A:RR-5# show router bgp summary all
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  65536       14    0 00h01m19s 1/0/0 (IPv4)
                           13    0           18/0/18 (LinkState)
192.0.2.2
Def. Instance  65536        5    0 00h01m19s 0/0/1 (IPv4)
                            6    0
192.0.2.3
Def. Instance  65536        5    0 00h01m19s 0/0/1 (IPv4)
                            7    0
192.0.2.4
Def. Instance  65536       14    0 00h01m13s 1/0/0 (IPv4)
                           15    0           18/0/18 (LinkState)
-------------------------------------------------------------------------------
*A:RR-5#
```

For implementing ORR using the hierarchical topology shown in Figure 5, the route reflector RR5 defines two locations in the **optimal-route-reflection** context. The primary IP address for location 1 is the P1 system IP address, and the primary IP address for location 2 is a loopback address for P-3. These addresses are used as the starting point for the SPF run. These **orr-locations** are then referred to from within the group definitions through the **cluster** command. Because RR-5 is not on the data path, there is no need for implementing the routes into the FIB, which is achieved through the **disable-route-table-install** command. The overall BGP configuration of RR-5 is as follows:

```
# on RR-5
configure
    router
        bgp
            --- snipped ---
            disable-route-table-install
            --- snipped ---
            optimal-route-reflection
```

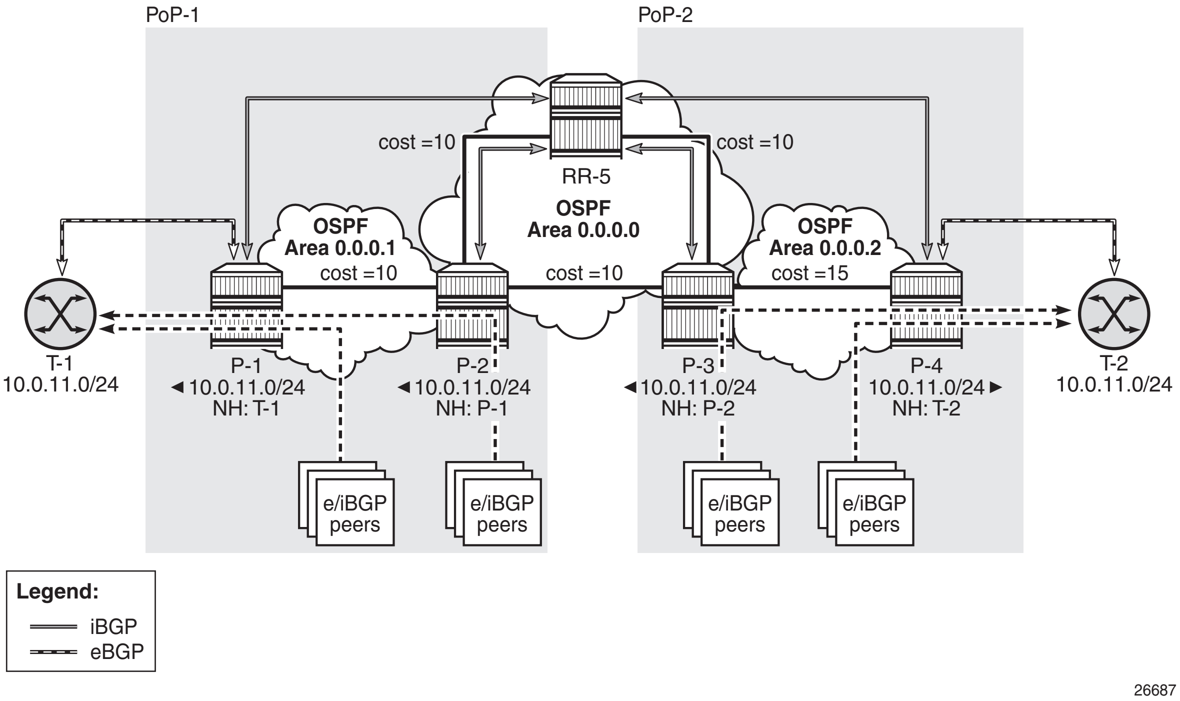
```
                spf-wait 1 initial-wait 1 second-wait 1
                location 1
                    primary-ip-address 192.0.2.1
                exit
                location 2
                    primary-ip-address 192.0.2.222
                exit
            exit
            group "iBGP-1"
                cluster 192.0.2.5 orr-location 1 allow-local-fallback
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
            group "iBGP-2"
                cluster 192.0.2.5 orr-location 2 allow-local-fallback
                peer-as 65536
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

In comparison with the previous scenario, there only is a change in the routing for prefix 10.0.11.0/24 on P-3, as follows.

```
*A:P-3#  show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h02m01s  170
       192.168.34.2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P3#
```

This is summarized in Figure 130. RR-5 receives the updates from P-1 and P-4, and now performs two SPF runs because two locations are used. The first SPF run uses the 192.0.2.1 address of P-1 as the starting point for the first location, selects the path via P-1 as the best path, and reflects that path to the remaining peers in the first location. The second SPF run uses the 192.0.2.222 loopback address of P-3 as the starting point for the second location, selects the path via P-4 as the best path, and reflects that path to the remaining peers in the second location.

Figure 130 **Optimal Route Reflection**



The following command provides the IGP distances for the configured reference points to all available BGP peers and all detected BGP next hops on the route reflector.

```
*A:RR-5# show router bgp optimal-route-reflection bgp-nh-info
===============================================================================
ORR BGP-NH Table (Router: Base)
===============================================================================
Location 1:
    Primary      : 192.0.2.1 [active]
    Secondary    : -
    Tertiary     : -
Location 2:
    Primary      : 192.0.2.222 [active]
    Secondary    : -
    Tertiary     : -
Age          : 00h07m43s
```

```
Spf wait     : 1
Initial wait : 1
Second wait  : 1
--------------------------------------------------------------------------------
Next Hop
   Loc    Dest-Prefix
                               DB-Source  Type          Proto     Metric    Pref
--------------------------------------------------------------------------------
192.0.2.1
    1     192.0.2.1/32
                               BGP-LS     Local         Local     0            0
    2     192.0.2.1/32
                               BGP-LS     Remote        OSPFv2    35          10
192.0.2.2
    1     192.0.2.2/32
                               BGP-LS     Remote        OSPFv2    10          10
    2     192.0.2.2/32
                               BGP-LS     Remote        OSPFv2    25          10
192.0.2.3
    1     192.0.2.3/32
                               BGP-LS     Remote        OSPFv2    20          10
    2     192.0.2.3/32
                               BGP-LS     Remote        OSPFv2    15          10
192.0.2.4
    1     192.0.2.4/32
                               BGP-LS     Remote        OSPFv2    35          10
    2     192.0.2.4/32
                               BGP-LS     Local         Local     0            0
--------------------------------------------------------------------------------
No. of BGP-NHs: 4
================================================================================
*A:RR-5#
```

Conclusion

BGP Optimal Route Reflection allows operators to optimize traffic streams through their network, even when the route reflector is placed out-of-path, for example in datacenters, thereby reducing the OPEX and CAPEX of route reflector deployment.

BGP Optimal Route Reflection for Non-Hierarchical Networks

This chapter provides information about BGP Optimal Route Reflection for Non-Hierarchical Networks.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

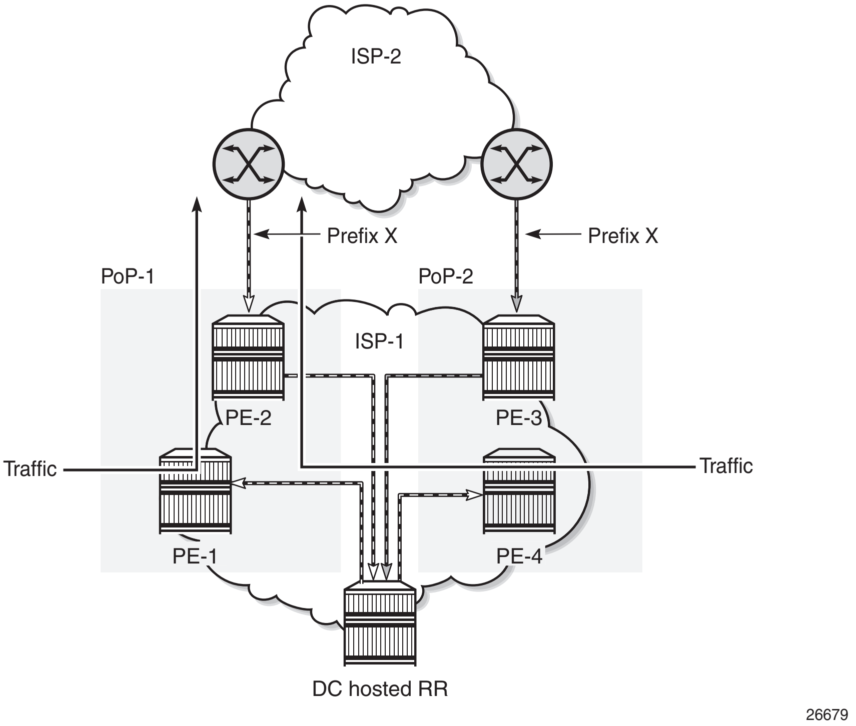
The information and configuration in this chapter are based on SR OS Release 15.0.R4.

Overview

BGP route reflectors are used in many networks. They improve network scalability by eliminating or reducing the need for a full-mesh of iBGP sessions.

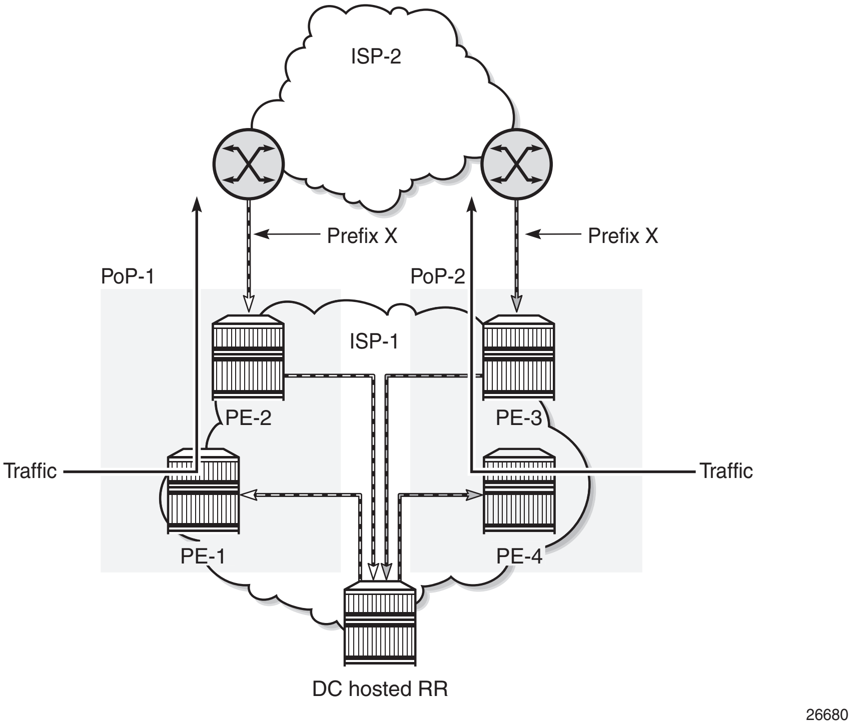
When a BGP route reflector receives multiple paths for the same IP destination, it normally selects and reflects a single best path in its routing domain to all clients in that domain, based on its own location in the domain. In Figure 131, the centralized route reflector RR for ISP-1 is located in the datacenter (DC), and receives prefix X from ISP-2 through PE-2 in Point of Presence (PoP) -1 and also through PE-3 in PoP-2. RR selects and reflects PE-2 as the best path to the remaining route reflector clients because RR is closer to PoP-1 than it is to PoP-2, so the traffic to destination X flows as indicated. Therefore, sending traffic to another autonomous system (AS) through the closest possible exit point from the local AS, known as hot-potato routing, cannot be achieved.

Figure 131 **Centralized Route Reflection**



Hot-potato routing can be achieved using a route reflector selecting and reflecting multiple best paths, for different subdomains and from the point of view of a client in a subdomain, as outlined in *draft-ietf-idr-bgp-optimal-route-reflection* (ORR), and requires the route reflector to know the topology of each subdomain. In Figure 132, the route reflector calculates the best path for PoP-1 and reflects that to the clients in PoP-1 (PE-1), and it also calculates the best path for PoP-2 and reflects that to the clients in PoP-2 (PE-4).

Figure 132 **Centralized Route Reflection with ORR**



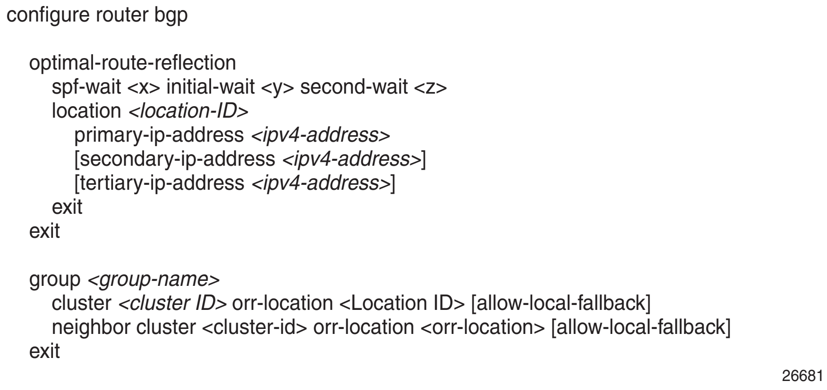
If the routing domain is flat, the route reflector is part of the routing domain and thus has a view on the entire topology through the Interior Gateway Protocol (IGP).

If the routing domain is hierarchical, the route reflector needs to extract the Link State Data Base (LSDB) from the subdomain it is not part of, which is achieved through BGP-Link-State (BGP-LS). The use of BGP-LS allows the route reflector to learn the IGP topology information for OSPF areas and IS-IS levels in which the route reflector is not a direct participant. See the BGP Optimal Route Reflection for Hierarchical Networks chapter if the network topology is hierarchical.

ORR CLI commands

The CLI command hierarchy used in the Nokia ORR implementation is shown in Figure 133. The BGP optimal-route-reflection context defines the Shortest Path First (SPF) parameters, and multiple locations. The locations are then referred to with the **cluster** command (residing in the BGP **group** context) through the **orr-location** argument.

Figure 133 **Configuring ORR in SR OS**



```
configure router bgp

 optimal-route-reflection
   spf-wait <x> initial-wait <y> second-wait <z>
   location <location-ID>
     primary-ip-address <ipv4-address>
     [secondary-ip-address <ipv4-address>]
     [tertiary-ip-address <ipv4-address>]
   exit
 exit

 group <group-name>
   cluster <cluster ID> orr-location <Location ID> [allow-local-fallback]
   neighbor cluster <cluster-id> orr-location <orr-location> [allow-local-fallback]
 exit
```

Up to sixteen locations can be created in the **optimal-route-reflection** context. Each location is identified through a Location-ID [1..16], and contains a **primary-ip-address** and, optionally, a **secondary-ip-address** and a **tertiary-ip-address**, for redundancy reasons. These addresses must correspond to loopback or system IP addresses of routers participating in the IGP protocols, and are used as the starting point (or seed) for the SPF calculation. Because all clients in the same location receive the same optimal path for that location, these addresses should be close to the clients in that part of the network.

The SPF calculation is configurable with the **spf-wait** command. **Initial-wait** and **second-wait** are optional arguments. These timers define when to initiate the first, second, and subsequent SPF runs after a topology change occurs.

The Location-ID is referred to in the **orr-location** argument of the **cluster** command. Typically, the **cluster** command applies to a BGP peer group; all neighbors in that group share the same Location-ID, unless the **cluster** command applies at a neighbor level. The **allow-local-fallback** option allows the RR to advertise the best reachable BGP path using its own location, but only when no BGP routes are reachable for some location. Otherwise, no path would be advertised to the clients in that location.

Properties

The following properties apply to ORR in SR OS:

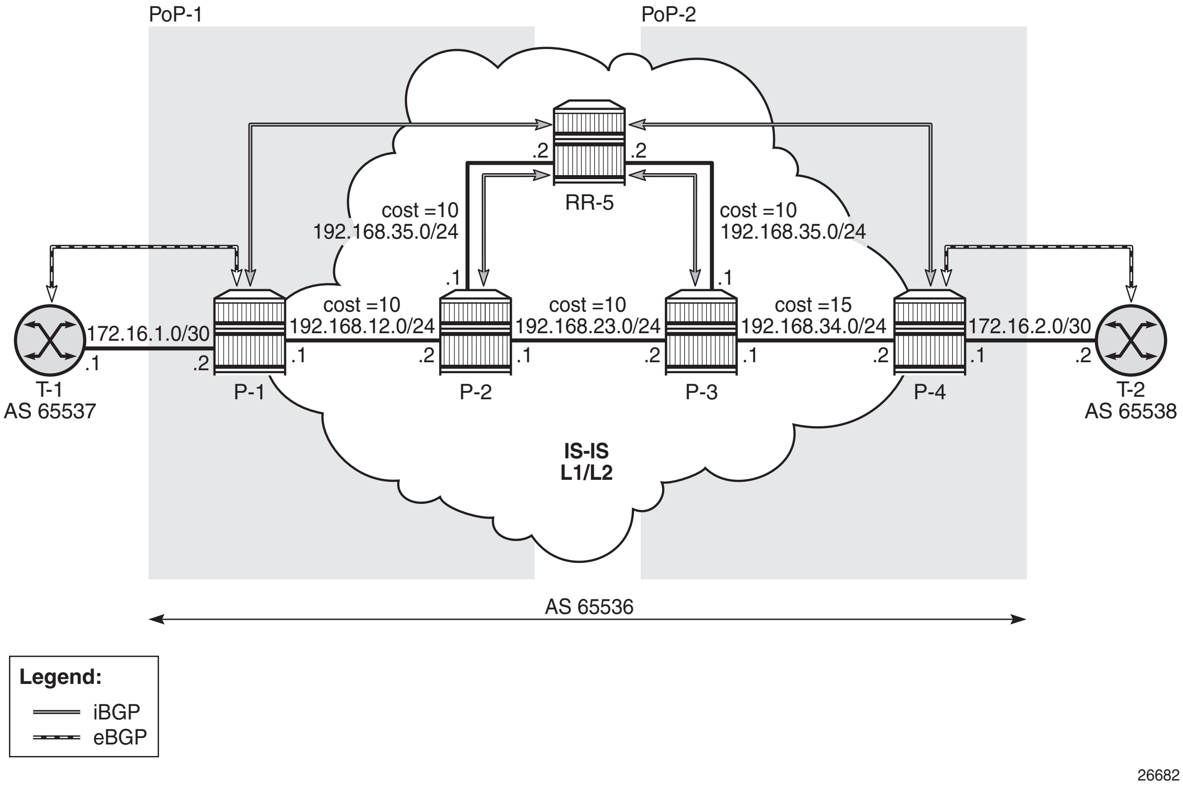
- ORR is supported in the Base router BGP instance.
- ORR is supported for the IPv4, label-IPv4, label-ipv6, VPN-IPv4, and VPN-IPv6 address families.

- ORR is supported with Add-Paths, meaning that Add-Paths advertised to ORR clients are also ORR location-based.?

Configuration

Figure 134 shows the example topology. IS-IS is used as the IGP for AS 65536, with RR-5 taking the role of the route reflector for clients P-1 to P-4. Additionally, tester T-1 in AS 65537 peers with P-1, and tester T-2 in AS 65538 peers with P-4.

Figure 134 **Example Non-Hierarchical Networking using IS-IS**



The initial configuration on all nodes includes:

- Cards, MDAs, and ports
- Router interfaces
- IS-IS as IGP on all interfaces within AS 65536, in a non-hierarchical way (alternatively, OSPF can be used), and traffic engineering enabled

The basic IS-IS configuration is very similar for all routers, including the route reflector. The RR-5 configuration is as follows:

```
# on RR-5
configure
    router
        isis 0
            area-id 49.0001
            traffic-engineering
            interface "system"
                no shutdown
            exit
            interface "int-RR5-P2"
                interface-type point-to-point
                no shutdown
            exit
            interface "int-RR5-P3"
                interface-type point-to-point
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

Route Reflection without ORR

RR-5 peers with clients P-1 to P-4, and because RR-5 is the route reflector, the **cluster** command is added, defining the cluster ID attribute value to use. The configuration for RR-5 is as follows:

```
# on RR-5
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                cluster 192.0.2.5
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-1 belongs to the cluster defined in the route reflector, so it does not need to be fully meshed with the other routers in the area; peering with the route reflectors in the area is sufficient for P-1 to receive updates. Typically, two route reflectors are provisioned for redundancy, but that does not apply in this example. P-1 also peers with T-1 in AS 65537 through eBGP, so the P-1 configuration is as follows:

```
# on P-1
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "eBGP"
                peer-as 65537
                neighbor 172.16.1.1
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-2 and P-3 also belong to the cluster defined in the route reflector; they only peer with the route reflector. Their configuration is the same:

```
# on P-2 and P-3
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                next-hop-self
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

P-4 also belongs to the cluster defined in the route reflector, but peers with T-2 in AS 65538. The P-4 configuration is similar to the configuration of P-1.

T-1 announces prefix 10.0.11.0/24 to router P-1, and T-2 announces the same prefix to router P-4. As a result, traffic offered to P-1 for destination 10.0.11.0/24 is routed to T-1, as follows:

```
*A:P-1# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h02m11s  170
       172.16.1.1                                                   0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-1#
```

Traffic offered to P-2 for destination 10.0.11.0/24 is routed to P-1, as follows:

```
*A:P-2# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.22.0/24                                  Remote  BGP       01d19h44m  170
       192.168.12.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-2#
```

Traffic offered to P-3 for destination 10.0.11.0/24 is routed to P-2, as follows:

```
*A:P-3# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h07m11s  170
       192.168.23.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
```

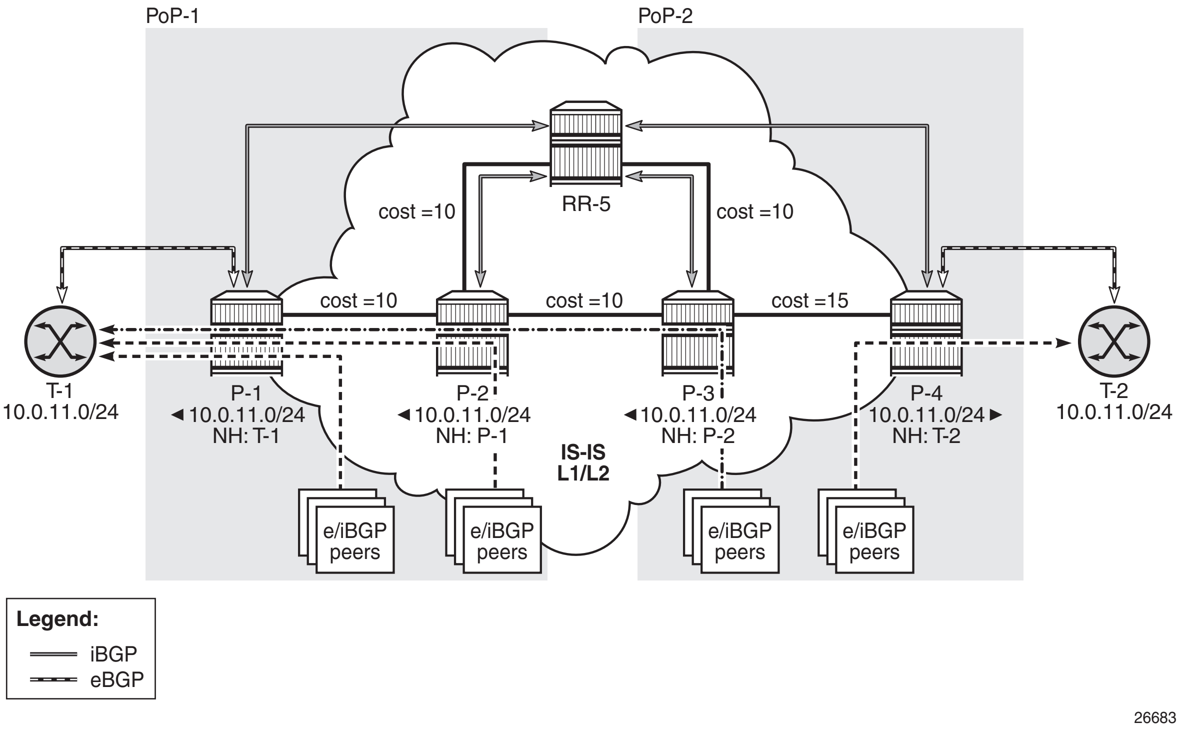
```
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

Traffic offered to P-4 for destination 10.0.11.0/24 is routed to T-2, as follows:

```
*A:P-4# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h17m59s  170
       172.16.2.2                                                   0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-4#
```

This is summarized in Figure 135. RR-5 receives the updates from P-1 and P-4, and selects the best path based on its own position in the topology. The cost from RR-5 to P-1 is 20, and the cost from RR-5 to P-4 is 25, so RR-5 selects P-1 and reflects this path to all clients (except to P-1, because P-1 is the client where the path was learned from). Ultimately, P-1 only has one path, and so do P-2 and P-3. P-4 has two paths, but by default prefers the eBGP learned path over the iBGP learned path.

Figure 135 **Suboptimal Route Reflection**



Route Reflection with ORR

For implementing ORR using the flat topology from Figure 5 the route reflector RR-5 defines two locations in the **optimal-route-reflection** context. The primary IP address for location 1 is the P-1 system IP address, and the primary IP address for location 2 is a loopback address for P-3. These addresses are used as the starting point for the SPF run. These **orr-locations** are then referred to from within the group definitions through the **cluster** command. The overall BGP configuration of RR-5 is as follows:

```
# on RR-5
configure
    router
        autonomous-system 65536
        bgp
            family ipv4
            loop-detect discard-route
            split-horizon
            optimal-route-reflection
                spf-wait 1 initial-wait 1 second-wait 1
                location 1
```

```
                    primary-ip-address 192.0.2.1
                exit
                location 2
                    primary-ip-address 192.0.2.222
                exit
            exit
            group "iBGP-1"
                cluster 192.0.2.5 orr-location 1 allow-local-fallback
                peer-as 65536
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
            group "iBGP-2"
                cluster 192.0.2.5 orr-location 2 allow-local-fallback
                peer-as 65536
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

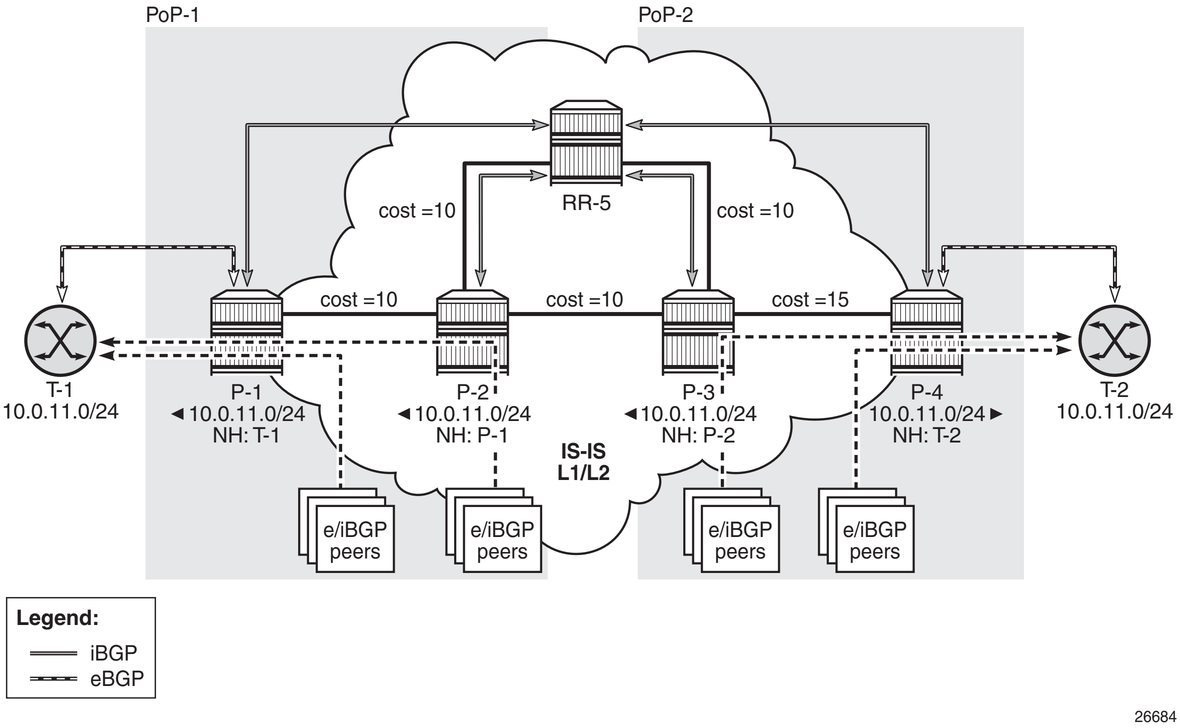
No changes are required in the clients.

T-1 announces prefix 10.0.11.0/24 to router P-1, and T-2 announces the same prefix to router P-4. In comparison with the previous scenario, there only is a change in the routing for this prefix on P-3, as follows:

```
*A:P-3# show router route-table protocol bgp
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.11.0/24                                  Remote  BGP       00h11m27s  170
       192.168.34.2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-3#
```

This is summarized in Figure 136. RR-5 receives the updates from P-1 and P-4, and now performs two SPF runs because two locations are used. The first SPF run uses the 192.0.2.1 address of P-1 as the starting point for the first location, selects the path via P-1 as the best path, and reflects that path to the remaining peers in the first location. The second SPF run uses the 192.0.2.222 loopback address of P-3 as the starting point for the second location, selects the path via P-4 as the best path, and reflects that path to the remaining peers in the second location.

Figure 136 **Optimal Route Reflection**



The following command provides the IGP distances for the configured reference points to all available BGP peers and all detected BGP next hops on the route reflector.

```
*A:RR-5# show router bgp optimal-route-reflection bgp-nh-info
===============================================================================
ORR BGP-NH Table (Router: Base)
===============================================================================
Location 1:
    Primary       : 192.0.2.1 [active]
    Secondary     : -
    Tertiary      : -
Location 2:
    Primary       : 192.0.2.222 [active]
    Secondary     : -
    Tertiary      : -
```

```
Age          : 00h08m26s
Spf wait     : 1
Initial wait : 1
Second wait  : 1
-------------------------------------------------------------------------------
Next Hop
   Loc    Dest-Prefix
                               DB-Source  Type         Proto     Metric    Pref
-------------------------------------------------------------------------------
192.0.2.1
    1     192.0.2.1/32
                               IGP        Local        Local     0            0
    2     192.0.2.1/32
                               IGP        Remote       ISIS      35          15
192.0.2.2
    1     192.0.2.2/32
                               IGP        Remote       ISIS      10          15
    2     192.0.2.2/32
                               IGP        Remote       ISIS      25          15
192.0.2.3
    1     192.0.2.3/32
                               IGP        Remote       ISIS      20          15
    2     192.0.2.3/32
                               IGP        Remote       ISIS      15          15
192.0.2.4
    1     192.0.2.4/32
                               IGP        Remote       ISIS      35          15
    2     192.0.2.4/32
                               IGP        Local        Local     0            0
-------------------------------------------------------------------------------
No. of BGP-NHs: 4
===============================================================================
*A:RR-5#
```

Conclusion

BGP Optimal Route Reflection allows operators to optimize traffic streams through their network, even when the route reflector is placed out-of-path, for example in datacenters, thereby reducing the OPEX and CAPEX of route reflector deployment.

BGP Prefix Limit per Address Family

This chapter provides information about BGP Prefix Limit per Address Family.

Topics in this chapter include:

Applicability

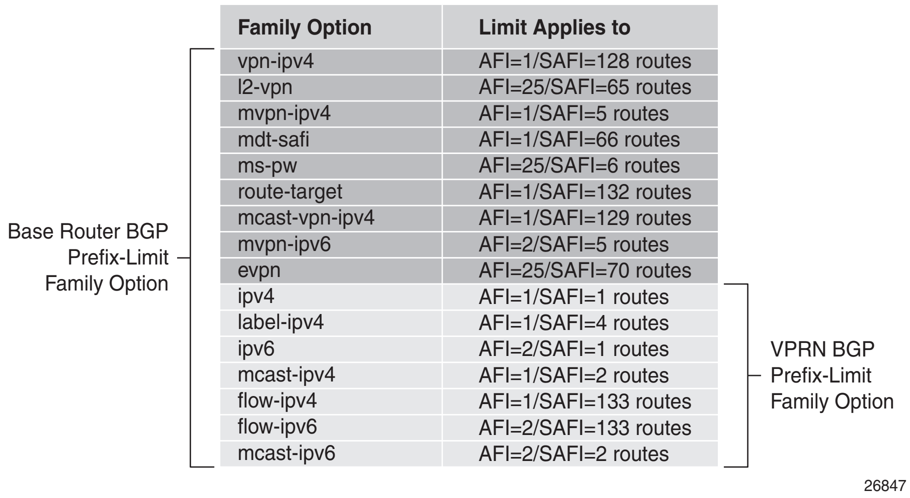
The information and configuration in this chapter are based on SR OS Release 15.0.R1.

Overview

A BGP per address family prefix limit can be defined to control the number of prefixes learned per neighbor or per group of neighbors in the base router or in a VPRN. This feature allows ISPs to secure their network from misbehaving or misconfigured peers. This feature can also be used to enforce the terms of a service contract.

Figure 137 lists the address families for which a prefix limit can be defined in the base router and in a VPRN.

Figure 137 **Supported Address Families**



Family Option	Limit Applies to
vpn-ipv4	AFI=1/SAFI=128 routes
l2-vpn	AFI=25/SAFI=65 routes
mvpn-ipv4	AFI=1/SAFI=5 routes
mdt-safi	AFI=1/SAFI=66 routes
ms-pw	AFI=25/SAFI=6 routes
route-target	AFI=1/SAFI=132 routes
mcast-vpn-ipv4	AFI=1/SAFI=129 routes
mvpn-ipv6	AFI=2/SAFI=5 routes
evpn	AFI=25/SAFI=70 routes
ipv4	AFI=1/SAFI=1 routes
label-ipv4	AFI=1/SAFI=4 routes
ipv6	AFI=2/SAFI=1 routes
mcast-ipv4	AFI=1/SAFI=2 routes
flow-ipv4	AFI=1/SAFI=133 routes
flow-ipv6	AFI=2/SAFI=133 routes
mcast-ipv6	AFI=2/SAFI=2 routes

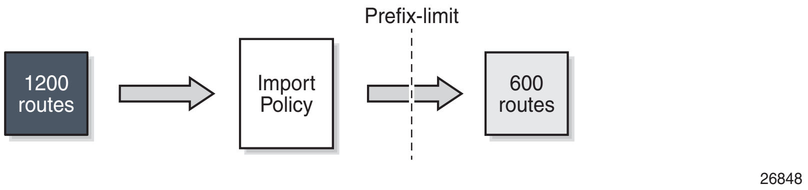
Base Router BGP Prefix-Limit Family Option
VPRN BGP Prefix-Limit Family Option
26847

If the number of received routes from a peer exceeds a defined per address family limit, the BGP session is torn down, the state is changed to disabled, the routes learned from that peer are deleted, and the RIB and FIB are recalculated. With the **log-only** option enabled, the BGP session is not torn down and no routes are deleted. An SNMP trap message is issued when exceeding the per address family threshold (default: 90%), and the per address family prefix limit.

Re-establishing the BGP session with the peer requires a manual intervention, or use of the **idle-timeout** option. The idle-timeout option defines the time in minutes after which the system attempts to re-establish the BGP session. The idle-timeout option can be given the value forever, which corresponds to the default behavior of requiring a manual intervention if the limit is exceeded.

The **post-import** option indicates that the limit should be applied only to the routes accepted by import policies; see Figure 138. A route rejected by an import policy will not be counted when checking against the prefix limit. Not specifying the post-import option results in routes being counted and verified against the prefix limit when they are received, before the import policy is executed, and might lead to BGP sessions being torn down unexpectedly.

Figure 138 **Post-Import Option**



BGP sessions will be torn down as soon as one of the address family prefix limits is exceeded, even when the limit for the other address family is not yet exceeded. In cases where this is important, consider defining two BGP sessions between two peers; the first using IPv4 for its transport, and the second using IPv6. In this way, an IPv4 limit being exceeded will not lead to IPv6 prefixes being affected.

Note that a VPN route carrying a route-target (VPN-IPv4, VPN-IPv6, L2-VPN, MVPN-IPV4, MVPN-IPv6) might not be retained in the RIB-IN if it is not imported by any service. If a VPN route is not stored in the RIB-IN, it is not counted and not checked against the prefix limit for its associated address family. If **mp-bgp-keep** is configured, or the router is a route reflector (using the cluster command) or an ASBR in an inter-AS VPRN model-B, then the VPN-IP route is always stored.

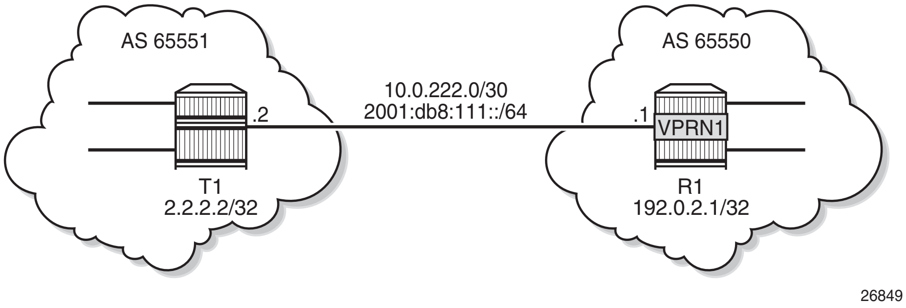
Configuration

Figure 139 shows the example topology. Tester T1 in AS 65551 peers with VPRN-1 hosted by R1 in AS 65550.

Two scenarios are considered:

- Prefix-limit without post-import
- Prefix-limit with post-import

Figure 139 **Example Setup**



Prefix Limit without Post-Import Option

T1 peers with VPRN-1 on R1, and has the IPv4 prefix limit set to 10, the threshold set to 50% and the idle-time set to 2 minutes, as follows:

```
# R1
configure
    service
        vprn 1 customer 1 create
            description "BGP prefix-limit on VPRN"
            autonomous-system 65550
            route-distinguisher 65550:1
            interface "int-R1.1-T1" create
                address 10.0.222.1/30
                ipv6
                    address 2001:db8:2222::1/64
                exit
                sap 1/1/4:11 create
                exit
            exit
            bgp
                family ipv4 ipv6
                loop-detect discard-route
                split-horizon
                group "int-R1.1-T1"
                    prefix-limit ipv4 10 threshold 50 idle-timeout 2
                    prefix-limit ipv6 8 threshold 80 idle-timeout 4
                    peer-as 65551
                    neighbor 10.0.222.2
                    exit
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

The debug configuration is as follows:

```
debug
    router "1"
        bgp
            packets neighbor 10.0.222.2
            events neighbor 10.0.222.2
        exit
    exit
exit
```

The debug output is sent to the log with log-id 1, as follows:

```
configure
    log
        log-id 1
            from debug-trace
            to memory
            no shutdown
        exit
    exit
exit
```

Initially the number of IPv4 routes received from T1 is below the threshold, and T1 gradually injects more IPv4 routes into VPRN-1. The following is a snapshot where three IPv4 routes and four IPv6 routes received and active in R1:

```
*A:R1# show router 1 bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:65550       Local AS:65550
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total BGP Paths         : 10          Total Path Memory           : 2016
Total IPv4 Remote Rts   : 3           Total IPv4 Rem. Active Rts  : 3
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total IPv6 Remote Rts   : 4           Total IPv6 Rem. Active Rts  : 4
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0

Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total FlowIpv4 Rem Rts  : 0           Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0           Total FlowIpv6 Rem Act Rts  : 0
Total LblIpv4 Rem Rts   : 0           Total LblIpv4 Rem. Act Rts  : 0
Total LblIpv6 Rem Rts   : 0           Total LblIpv6 Rem. Act Rts  : 0
Total LblIpv4 Bkp Rts   : 0           Total LblIpv6 Bkp Rts       : 0
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
```

```
                      PktSent OutQ
-------------------------------------------------------------------------------
10.0.222.2
               65551       21    0 00h03m24s 3/3/0 (IPv4)
                            9    0           4/4/0 (IPv6)
-------------------------------------------------------------------------------
*A:R1#
```

The IPv4 routes announced by T1 and received on R1, VPRN-1 are as follows:

```
*A:R1# show router 1 bgp routes
===============================================================================
 BGP Router ID:192.0.2.1        AS:65550       Local AS:65550
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.0.22.0/30                                       None        None
      10.0.222.2                                         None        -
      65551
u*>i  10.0.22.4/30                                       None        None
      10.0.222.2                                         None        -
      65551
u*>i  10.0.22.8/30                                       None        None
      10.0.222.2                                         None        -
      65551
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:R1#
```

When the sixth route is received, the threshold value (50% of 10 is 5) is exceeded and a message is generated and sent to log 99, as follows:

```
*A:R1# show log log-id 99

===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=104  (not wrapped)]

103 2017/04/12 12:20:21.59 CEST MINOR: BGP #2035 vprn1 Peer 2: 10.0.222.2
"VR 2: Group int-R1.1-CE: Peer 10.0.222.2: number of routes learned has exceeded 50
percentage of the configured maximum (10) for ipv4 family"
```

When the configured maximum number of BGP routes for IPv4 is exceeded, the peer is notified, as indicated in the following debug log:

```
*A:R1# show log log-id 1

===============================================================================
Event Log 1
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=100   next event=28  (not wrapped)]

27 2017/04/12 12:26:49.49 CEST MINOR: DEBUG #2001 vprn1 Peer 2: 10.0.222.2
"Peer 2: 10.0.222.2: NOTIFICATION
Peer 2: 10.0.222.2 - Send BGP NOTIFICATION: Code = 6 (CEASE) Subcode = 1
(Maximum prefixed reached)
  Data Length = 7  Data: 0x0 0x1 0x1 0x0 0x0 0x0 0xa
"

26 2017/04/12 12:26:49.49 CEST MINOR: DEBUG #2001 vprn1 BGP
"BGP: STATE
Peer 2:10.0.222.2 - Change State from ESTABLISHED to IDLE due to MAXPREFIX_EXCEEDED"

25 2017/04/12 12:26:49.49 CEST MINOR: DEBUG #2001 vprn1 Peer 2: 10.0.222.2
"Peer 2: 10.0.222.2: UPDATE
Peer 2: 10.0.222.2 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 20
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 6 AS Path:
        Type: 2 Len: 1 < 65551 >
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 10.0.222.2
    NLRI: Length = 5
        10.0.22.40/30
"
```

Therefore, the BGP session is torn down and the corresponding state is disabled, as follows:

```
*A:R1# show router 1 bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:65550       Local AS:65550
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total BGP Paths         : 8           Total Path Memory           : 1600
Total IPv4 Remote Rts   : 0           Total IPv4 Rem. Active Rts  : 0
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total IPv6 Remote Rts   : 0           Total IPv6 Rem. Active Rts  : 0
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0

Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total FlowIpv4 Rem Rts  : 0           Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0           Total FlowIpv6 Rem Act Rts  : 0
Total LblIpv4 Rem Rts   : 0           Total LblIpv4 Rem. Act Rts  : 0
Total LblIpv6 Rem Rts   : 0           Total LblIpv6 Rem. Act Rts  : 0
Total LblIpv4 Bkp Rts   : 0           Total LblIpv6 Bkp Rts       : 0
===============================================================================
BGP Summary
```

```
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
10.0.222.2
               65551       29    0 00h01m51s Disabled
                            4    0
-------------------------------------------------------------------------------
*A:R1#
```

Also this event is recorded in the system logs, as follows:

```
*A:R1# show log log-id 99

===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=114  (not wrapped)]

113 2017/04/12 12:28:58.51 CEST MINOR: BGP #2034 vprn1 Peer 2: 10.0.222.2
"VR 2: Group int-R1.1-CE: Peer 10.0.222.2: number of routes learned has exceeded
configured maximum (10) for ipv4 family"

112 2017/04/12 12:28:58.50 CEST WARNING: BGP #2005 vprn1 Peer 2: 10.0.222.2
"VR 2: Group int-R1.1-CE: Peer 10.0.222.2: sending notification: code CEASE
subcode MAX_PFX_RCHD"

111 2017/04/12 12:28:58.50 CEST WARNING: BGP #2039 vprn1 Peer 2: 10.0.222.2
"VR 2: Group int-R1.1-CE: Peer 10.0.222.2: moved from higher state ESTABLISHED
to lower state IDLE due to event MAXPREFIX_EXCEEDED"
```

When the idle-timeout expires, the system tries to re-establish the session. With the session re-established, the peer starts re-announcing its routes. As long as the number of routes announced is lower or equal to the limit, the session is maintained.

Prefix-Limit with Post-Import Option

Use caution when using the prefix-limit in combination with import policies. By default, the routes are counted when receiving them, that is before the import policy is enforced. To postpone the prefix-limit check, the post-import option must be used.

The BGP configuration for VPRN-1is then adapted as follows:

```
configure
    service
        vprn 1 customer 1 create
            bgp
```

```
                family ipv4 ipv6
                loop-detect discard-route
                import "import-10.0.22-ranges"
                split-horizon
                group "int-CE"
                    prefix-limit ipv4 10 threshold 50 idle-timeout 2 post-import
                    peer-as 65551
                    neighbor 10.0.222.2
                    exit
                exit
                no shutdown
            exit
        exit
    exit
exit
```

The *import-10.0.22-ranges* policy is defined as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "pfx-10.0.22-ranges"
                prefix 10.0.22.0/24 longer
            exit
            policy-statement "import-10.0.22-ranges"
                entry 10
                    from
                        prefix-list "pfx-10.0.22-ranges"
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
        exit
    exit
exit
```

When twelve IPv4 routes are received over this BGP session, six in the 10.0.22 range and six in the 10.0.23 range, then six routes are received and active in the routing table, as follows:

```
*A:R1# show router 1 route-table protocol bgp

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.0.22.0/30                                  Remote  BGP       00h14m21s  170
       10.0.222.2                                                   0
10.0.22.4/30                                  Remote  BGP       00h14m21s  170
       10.0.222.2                                                   0
10.0.22.8/30                                  Remote  BGP       00h14m21s  170
```

```
       10.0.222.2                                                   0
10.0.22.12/30                                 Remote  BGP       00h14m21s  170
       10.0.222.2                                                   0
10.0.22.16/30                                 Remote  BGP       00h14m21s  170
       10.0.222.2                                                   0
10.0.22.20/30                                 Remote  BGP       00h14m21s  170
       10.0.222.2                                                   0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:R1#
```

The BGP session remains established with twelve received routes and six of these being active, as follows:

```
*A:R1# show router 1 bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:65550       Local AS:65550
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 1           Total Peers                 : 1
Total BGP Paths         : 10          Total Path Memory           : 2016
Total IPv4 Remote Rts   : 12          Total IPv4 Rem. Active Rts  : 6
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total IPv6 Remote Rts   : 4           Total IPv6 Rem. Active Rts  : 0
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0

Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total FlowIpv4 Rem Rts  : 0           Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0           Total FlowIpv6 Rem Act Rts  : 0
Total LblIpv4 Rem Rts   : 0           Total LblIpv4 Rem. Act Rts  : 0
Total LblIpv6 Rem Rts   : 0           Total LblIpv6 Rem. Act Rts  : 0
Total LblIpv4 Bkp Rts   : 0           Total LblIpv6 Bkp Rts       : 0
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
10.0.222.2
               65551      237    0 00h16m05s 12/6/0 (IPv4)
                           67    0           4/0/0 (IPv6)
-------------------------------------------------------------------------------
*A:R1#
```

Without the post-import option the session would be torn down as soon as the number of received routes exceeded the configured prefix limit.

Conclusion

The BGP prefix limit per address family feature allows ISPs to protect their network from misbehaving or misconfigured peers, and can also be used to enforce the terms of a service contract.

BGP Route Leaking

This chapter provides information about BGP route leaking.

Topics in this chapter include:

Applicability

The information and configuration in this chapter wereoriginally written for SR OS release 14.0.R4. The CLI is updated to SR OS release 15.0.R4

Overview

Route leaking refers to the process of copying a route from one router context to another.

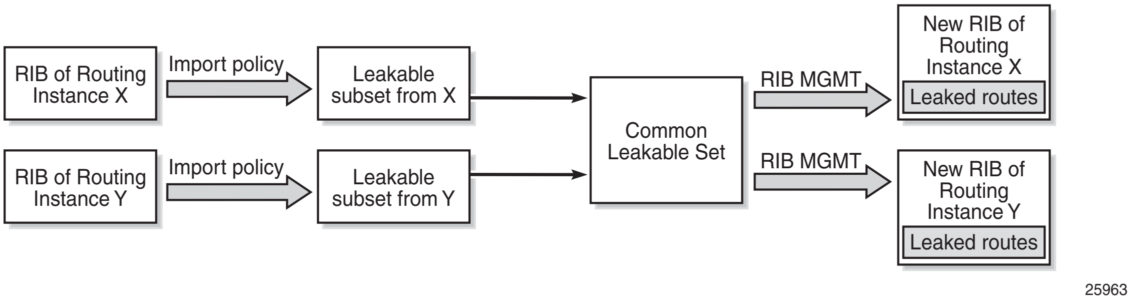
Network administrators may need to leak routes between routing instances in the same SR OS router. BGP route leaking is an alternative to using import/export policies based on communities to exchange routes between virtual router and forwarders (VRFs).

It is possible to leak a copy of a BGP route (including all its path attributes) from one routing instance to another in the same SR OS router. This BGP route leaking capability applies to IPv4, IPv6, and label-IPv4 routes. Leaking is supported from the GRT to a VPRN, from one VPRN to another VPRN, and from a VPRN to the GRT. Any valid BGP route for an IPv4 or IPv6 prefix can be leaked. A BGP route does not have to be the best path or used for forwarding in the source instance in order to be leaked, but it does have to be valid (that is, the next-hop must be resolved; the AS PATH must not exhibit a loop, for example).

An IPv4 or IPv6 BGP route becomes a candidate for leaking to another instance when it is specially marked by a BGP import policy. This special marking is achieved by accepting the route with a **bgp-leak** action in the route policy. Routes that are candidates for leaking to other instances show a **leakable** flag in the output of various show router bgp commands.

To copy a leakable BGP route from a source instance into the BGP RIB of a target instance, the target instance must be configured with a leak-import policy that matches and accepts the leakable route. There are separate leak-import policies for IPv4 and IPv6 routes. Up to 15 leak-import policies can be chained together for more complex examples. In the target instance, the **show router bgp routes** command displays leaked BGP RIB-IN routes in addition to direct RIB-IN routes learned from neighbors of the routing instance. A **leaked** flag is added to the leaked RIB-IN entries. Figure 140shows the process of BGP route leaking.

Figure 140 **BGP Route Leaking Process**



Leaked BGP routes can be advertised to BGP neighbors (peers) of the target routing instance. The BGP next hop of a leaked route is automatically reset to self whenever it is advertised to a peer of the target instance. Normal route advertisement rules apply: by default, the leaked route is advertised if it is the overall best path that is used as the active route to the destination and it is not blocked by the IBGP-to-IBGP split-horizon rule.

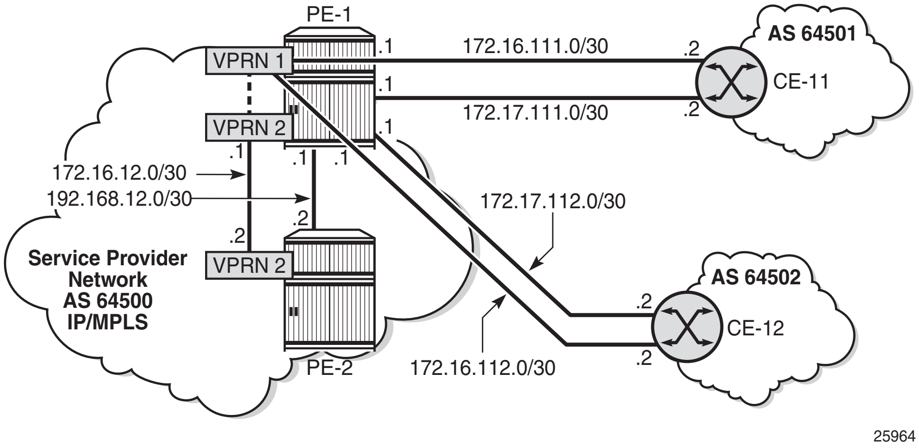
A BGP route leaked into a VPRN can be exported from the VPRN as a VPN-IPv4/v6 route if it matches the VRF export policy. Normal VPN export rules apply: by default, the leaked route is exported if it is the overall best path and it is used as the active route to the destination.

This chapter describes BGP route leaking only. For other routes, such as IS-IS, OSPF, RIP, and static routes, VPRN route leaking mechanisms apply that are protocol independent, see chapter Traffic Leaking from VPRN to GRT.

Configuration

Figure 141 shows the example topology used in this chapter, including the IPv4 addresses. For each of the examples, a dedicated figure will show the specific topology, which is a subset of the topology in Figure 2. The interfaces also have IPv6 addresses, which will be shown in Figure 6 and Figure 7. VPRN 2 also has CEs attached, but for simplicity, these are not shown on the figures and no CLI will be shown for any CE.

Figure 141 **Example Topology**



The following examples will be explained:

- Example 1 - BGP IPv4 Route Leaking between VPRNs. Global BGP Policy
- Example 2 - BGP IPv4 Route Leaking between VPRNs per Neighbor
- Example 3 - BGP IPv4 Route Leaking from VPRN to GRT per BGP Group
- Example 4 - BGP IPv4 Route Leaking from GRT to VPRN per Neighbor
- Example 5 - BGP IPv6 Route Leaking between VPRNs. Global VPRN BGP Configuration
- Example 6 - BGP IPv6 Route Leaking from GRT to VPRN and from VPRN to VPRN

Initial Configuration

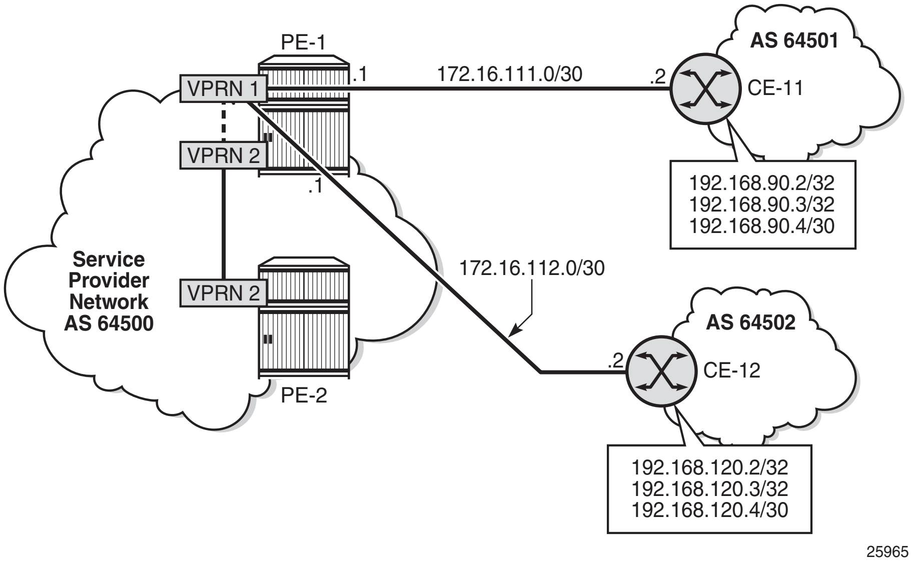
The nodes in the example topology have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces
- IGP (IS-IS or OSPF) between the PEs
- LDP between the PEs
- VPRN 1 on PE-1; VPRN 2 on PE-1 and PE-2
- BGP (IBGP between the PEs; EBGP between PE-1 and the CEs)
  - On the PEs, BGP is configured in the base router and in the VPRNs.
- Loopback addresses and black-hole static routes in the CEs. Different routes are exported to GRT and VPRN 1 on PE-1

Example 1 - BGP IPv4 Route Leaking between VPRNs. Global BGP Policy

Figure 142 shows the topology for this example. CE-11 exports routes such as 192.168.90.2/32 to VPRN 1 on PE-1, and CE-12 exports routes such as 192.168.120.2/32 to VPRN 1 on PE-1.

Figure 142 **BGP IPv4 Route Leaking Between VPRNs**



BGP leaking is disabled by default. The routing table for VPRN 1 on PE-1 includes routes that are learned from CE-11 and CE-12, as shown:

```
*A:PE-1# show router 1 route-table
===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.1/32                                 Local   Local     00h04m46s   0
       system                                                       0
172.16.111.0/30                               Local   Local     00h04m46s   0
       int-PE-1-CE-11                                               0
172.16.112.0/30                               Local   Local     00h04m46s   0
       int-PE-1-CE-12                                               0
192.168.90.2/32                               Remote  BGP       00h00m21s   170
       172.16.111.2                                                 0
192.168.90.3/32                               Remote  BGP       00h00m21s   170
       172.16.111.2                                                 0
192.168.90.4/30                               Remote  BGP       00h00m21s   170
       172.16.111.2                                                 0
192.168.120.2/32                              Remote  BGP       00h00m36s   170
       172.16.112.2                                                 0
192.168.120.3/32                              Remote  BGP       00h00m36s   170
       172.16.112.2                                                 0
192.168.120.4/32                              Remote  BGP       00h00m36s   170
       172.16.112.2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

These BGP routes are not leakable, by default, as follows:

```
*A:PE-1# show router 1 bgp routes ipv4 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
No Matching Entries Found.
===============================================================================
*A:PE-1
```

The routing table for VPRN 2 does not include any of these routes because BGP route leaking is disabled by default:

```
*A:PE-1# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.2.1/32                                 Local   Local     00h04m46s  0
       system                                                       0
172.16.2.2/32                                 Remote  BGP VPN   00h03m52s  170
       192.0.2.2 (tunneled)                                         0
172.16.12.0/30                                Local   Local     00h04m46s  0
       int-PE-1-PE-2_VPN2                                           0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

To configure BGP route leaking, an import policy is required in VPRN 1. The BGP route leaking policy is configured on PE-1, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-Leak-Policy"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
        exit
```

By adding the action accept **bgp-leak**, BGP routes are imported and marked as BGP leakable, meaning they are available to be copied - with their complete set of BGP path attributes - to the BGP RIB-IN of another routing instance.

The BGP route leaking policy can be applied in VPRN 1 in the general BGP configuration (as is the case here), in the group context, or per neighbor:

```
configure
    service
        vprn 1
            bgp
                import "BGP-Leak-Policy"
            exit
        exit
    exit
exit
```

With the preceding configuration, SR OS is marking all the BGP routes imported into the VPRN as leakable. The BGP routes originate from CE-11 or CE-12 in this example.

The following command shows which BGP routes in VPRN 1 are marked as leakable:

```
*A:PE-1# show router 1 bgp routes ipv4 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.168.90.2/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.90.3/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>?  192.168.90.4/30                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.120.2/32                                   None        None
      172.16.112.2                                       None        -
      64502
u*>i  192.168.120.3/32                                   None        None
      172.16.112.2                                       None        -
      64502
u*>?  192.168.120.4/32                                   None        None
      172.16.112.2                                       None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

The routes learned from CE-11 and CE-12 are leakable. The detailed output for any route in the preceding list shows the flag "leakable". The route source is external because the routes are imported (from CE-11 or CE-12):

```
*A:PE-1# show router 1 bgp routes 192.168.90.2/32 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 192.168.90.2/32
Nexthop        : 172.16.111.2
Path Id        : None
From           : 172.16.111.2
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 172.16.111.2
Local Pref.    : n/a                    Interface Name : int-PE-1-CE-11

--- snipped ---

Originator Id  : None                   Peer Router Id : 172.16.0.11
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leakable
Route Source   : External
AS-Path        : 64501

--- snipped ---

===============================================================================
*A:PE-1#
```

BGP leakable routes can be imported into another VPRN. Prefix lists can be used to filter specific routes for BGP leaking, but that is not configured in this example. The following import policy is configured on PE-1 to import BGP leakable routes:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

In each of the examples, the same import policy will be used. The import policy to import BGP leakable routes is applied in the VPRN 2 on PE-1 as follows:

```
configure
    service
        vprn 2
            bgp
                rib-management
                    ipv4
                        leak-import "Import-Leakable-Routes"
                    exit
                exit
            exit
        exit
    exit
exit
```

The following command shows that VPRN 2 imported leaked BGP routes from VPRN 1. The status code "l" indicates that the route is leaked.

```
*A:PE-1# show router 2 bgp routes ipv4 leaked
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 192.168.90.2/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>li 192.168.90.3/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>l? 192.168.90.4/30                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>li 192.168.120.2/32                                   100         None
      172.16.112.2 (VPRN 1)                              None        -
      64502
u*>li 192.168.120.3/32                                   100         None
      172.16.112.2 (VPRN 1)                              None        -
      64502
u*>l? 192.168.120.4/32                                   100         None
      172.16.112.2 (VPRN 1)                              None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

The flags in the detailed output for a particular leaked BGP route from the preceding list include the flag "leaked". The route source for this leaked route is VPRN 1 and all BGP attributes are preserved, as shown:

```
*A:PE-1# show router 2 bgp routes 192.168.90.2/32 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 192.168.90.2/32
Nexthop        : 172.16.111.2 (VPRN 1)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 172.16.111.2
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from VPRN 1
AS-Path        : 64501
Route Tag      : 0
Neighbor-AS    : 64501
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m36s

--- snipped ---

Routes : 1
===============================================================================
*A:PE-1#
```

The route-table for VPRN 2 in the neighbor PE-2 contains the leaked routes, as shown:

```
*A:PE-2# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.2.1/32                                 Remote  BGP VPN   00h12m14s  170
       192.0.2.1 (tunneled)                                         0
172.16.2.2/32                                 Local   Local     00h12m59s  0
       system                                                       0
172.16.12.0/30                                Local   Local     00h12m59s  0
       int-PE-2-PE-1_VPN2                                           0
192.168.90.2/32                               Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
192.168.90.3/32                               Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
192.168.90.4/30                               Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
192.168.120.2/32                              Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
192.168.120.3/32                              Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
192.168.120.4/32                              Remote  BGP       00h03m22s  170
       172.16.12.1                                                  0
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Example 2 - BGP IPv4 Route Leaking between VPRNs per Neighbor

The topology used for this example is the same as for Example 1; see Figure 142. Both CEs export the same routes as in the preceding example, and the BGP route leaking policy is identical:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-Leak-Policy"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

In the preceding example, the BGP route leaking policy was applied in the global BGP context in VPRN 1 and consequently, it applied to routes from all neighbors. In this example, the BGP route leaking policy is applied in VPRN 1 for neighbor CE-11 only, as follows:

```
configure
    service
        vprn 1
            bgp
                group "EBGP_64500to64501_IPv4"
                    neighbor 172.16.111.2
                        import "BGP-Leak-Policy"
                    exit
                exit
            exit
        exit
    exit
exit
```

This import policy implies that only routes learned from CE-11 will be leakable. The following command shows all the BGP routes learned in VPRN 1 on PE-1. Not all of these are leakable.

```
*A:PE-1# show router 1 bgp routes
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.168.90.2/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.90.3/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>?  192.168.90.4/30                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.120.2/32                                   None        None
      172.16.112.2                                       None        -
      64502
u*>i  192.168.120.3/32                                   None        None
      172.16.112.2                                       None        -
      64502
u*>?  192.168.120.4/32                                   None        None
      172.16.112.2                                       None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

Some routes are learned from CE-11 and other routes are learned from CE-12, and only the routes imported from CE-11 are leakable. The following command shows which routes are marked as leakable in the RT of the VPRN:

```
*A:PE-1# show router 1 bgp routes ipv4 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.168.90.2/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.90.3/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>?  192.168.90.4/30                                    None        None
      172.16.111.2                                       None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The BGP leakable routes can be imported into another VPRN instance. The import policy is the same as for Example 1:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This import policy is applied in VPRN 2 in the same way as in Example 1:

```
configure
    service
        vprn 2
            bgp
                rib-management
                    ipv4
                        leak-import "Import-Leakable-Routes"
                    exit
                exit
            exit
        exit
    exit
exit
```

The following command shows the leaked routes in VPRN 2. Each of these routes is leaked from VPRN 1, as indicated between brackets in the following output. Only routes learned from CE-11 in VPRN 1 are leaked to VPRN 2.

```
*A:PE-1# show router 2 bgp routes ipv4 leaked
===============================================================================
 BGP Router ID:192.0.2.1       AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 192.168.90.2/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>li 192.168.90.3/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>l? 192.168.90.4/30                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

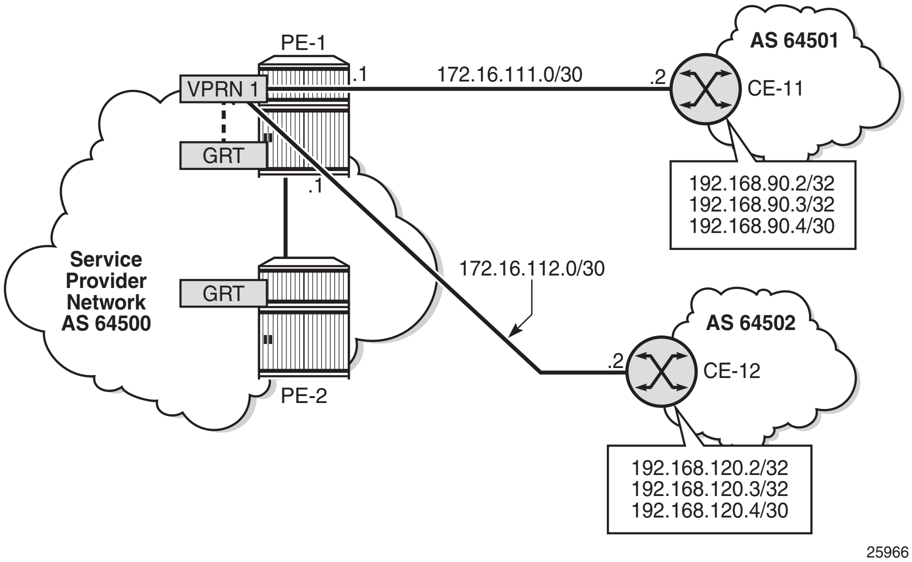
The detailed output for any of these BGP routes shows that the flag "leaked" is set and that the route source corresponds to VPRN 1, as shown for route 192.168.90.2/32:

```
*A:PE-1# show router 2 bgp routes 192.168.90.2/32 detail
================================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
================================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

================================================================================
BGP IPv4 Routes
================================================================================
Original Attributes

Network        : 192.168.90.2/32
Nexthop        : 172.16.111.2 (VPRN 1)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 172.16.111.2
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from VPRN 1
AS-Path        : 64501
--- snipped ---
Routes : 1
================================================================================
*A:PE-1#
```

Example 3 - BGP IPv4 Route Leaking from VPRN to GRT per BGP Group

Figure 143 shows the topology for this example. CE-11 and CE-12 export the same routes to VPRN 1. These routes will be marked as leakable and leaked to the GRT.

Figure 143 **BGP IPv4 Route Leaking from VPRN to GRT**



The routing table for VPRN 1 in PE-1 contains the BGP routes exported by CE-11 and CE-12, as follows:

```
*A:PE-1# show router 1 route-table

================================================================================
Route Table (Service: 1)
================================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
--------------------------------------------------------------------------------
172.16.1.1/32                                 Local   Local     00h14m51s   0
       system                                                       0
172.16.111.0/30                               Local   Local     00h14m51s   0
       int-PE-1-CE-11                                               0
172.16.112.0/30                               Local   Local     00h14m51s   0
       int-PE-1-CE-12                                               0
192.168.90.2/32                               Remote  BGP       00h00m26s   170
       172.16.111.2                                                 0
192.168.90.3/32                               Remote  BGP       00h00m26s   170
       172.16.111.2                                                 0
192.168.90.4/30                               Remote  BGP       00h00m26s   170
       172.16.111.2                                                 0
192.168.120.2/32                              Remote  BGP       00h03m15s   170
       172.16.112.2                                                 0
192.168.120.3/32                              Remote  BGP       00h03m15s   170
       172.16.112.2                                                 0
192.168.120.4/32                              Remote  BGP       00h03m15s   170
       172.16.112.2                                                 0
--------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
```

```
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The routing table of the base router does not include any of the BGP routes exported by the CEs, as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.17.111.0/30                               Local   Local     00h14m52s  0
       int-PE-1-CE-11                                               0
172.17.112.0/30                               Local   Local     00h14m52s  0
       int-PE-1-CE-12                                               0
192.0.2.1/32                                  Local   Local     00h14m52s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h14m32s  15
       192.168.12.2                                                 10
192.168.12.0/30                               Local   Local     00h14m52s  0
       int-PE-1-PE-2                                                0
-------------------------------------------------------------------------------
No. of Routes: 5
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The BGP routes are marked as leakable after applying the following configuration:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-Leak-Policy"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

This BGP route leaking policy can be applied in the general BGP configuration of VPRN 1, or per BGP group (as is the case here), or per BGP neighbor:

```
configure
    service
        vprn 1
            bgp
                group "EBGP_64500to64501_IPv4"
                    import "BGP-Leak-Policy"
                exit
            exit
        exit
    exit
exit
```

The following command shows the leakable BGP routes in VPRN 1:

```
*A:PE-1# show router 1 bgp routes ipv4 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.168.90.2/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>i  192.168.90.3/32                                    None        None
      172.16.111.2                                       None        -
      64501
u*>?  192.168.90.4/30                                    None        None
      172.16.111.2                                       None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The leakable BGP routes in VPRN 1 can be imported into the GRT. The import policy is identical to the import policy in the preceding examples, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This import policy is applied in the base router, as follows:

```
configure
    router
        bgp
            rib-management
                ipv4
                    leak-import "Import-Leakable-Routes"
                exit
            exit
        exit
    exit
exit
```

As a result, the leakable BGP routes in VPRN 1 are leaked to the GRT, as follows:

```
*A:PE-1# show router bgp routes ipv4 leaked
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500        Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 192.168.90.2/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>li 192.168.90.3/32                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
u*>l? 192.168.90.4/30                                    100         None
      172.16.111.2 (VPRN 1)                              None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The detailed information for any of these leaked routes shows that the flag "leaked" is present and that the route source is VPRN 1, as follows:

```
*A:PE-1# show router bgp routes 192.168.90.2/32 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 192.168.90.2/32
Nexthop        : 172.16.111.2 (VPRN 1)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 172.16.111.2
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from VPRN 1
AS-Path        : 64501
Route Tag      : 0
--- snipped ---
===============================================================================
*A:PE-1#
```

The GRT includes the leaked routes, as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.17.111.0/30                               Local   Local     00h18m08s  0
       int-PE-1-CE-11                                               0
172.17.112.0/30                               Local   Local     00h18m08s  0
       int-PE-1-CE-12                                               0
192.0.2.1/32                                  Local   Local     00h18m08s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h17m48s  15
       192.168.12.2                                                 10
192.168.12.0/30                               Local   Local     00h18m08s  0
       int-PE-1-PE-2                                                0
192.168.90.2/32                               Remote  BGP       00h01m22s  170
       172.16.111.2                                                 0
192.168.90.3/32                               Remote  BGP       00h01m22s  170
```

```
       172.16.111.2                                                 0
192.168.90.4/30                               Remote  BGP       00h01m22s  170
       172.16.111.2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 8
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

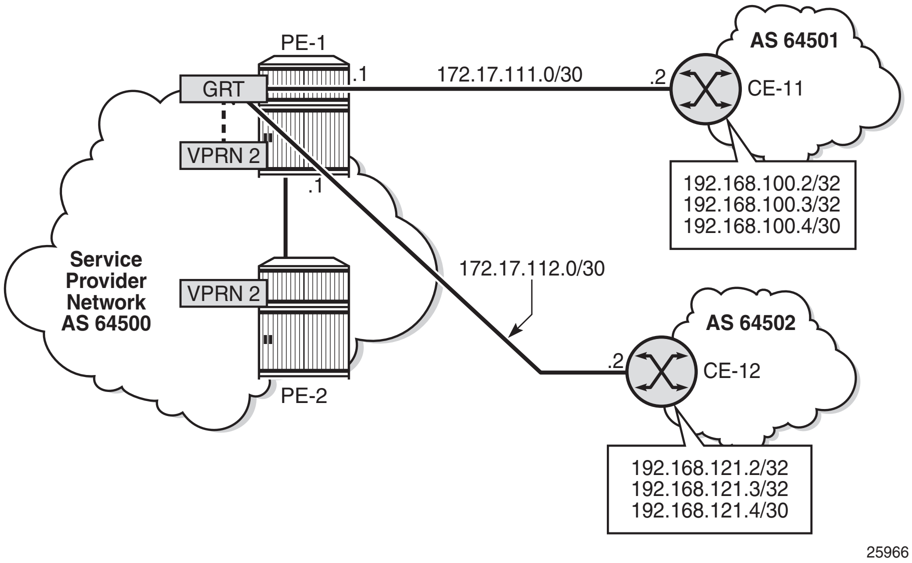
The GRT on neighbor PE-2 also includes the leaked routes, as follows:

```
*A:PE-2# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h17m50s  15
       192.168.12.1                                                 10
192.0.2.2/32                                  Local   Local     00h17m57s  0
       system                                                       0
192.168.12.0/30                               Local   Local     00h17m57s  0
       int-PE-2-PE-1                                                0
192.168.90.2/32                               Remote  BGP       00h00m58s  170
       192.168.12.1                                                 0
192.168.90.3/32                               Remote  BGP       00h00m58s  170
       192.168.12.1                                                 0
192.168.90.4/30                               Remote  BGP       00h00m58s  170
       192.168.12.1                                                 0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

Example 4 - BGP IPv4 Route Leaking from GRT to VPRN per Neighbor

Figure 144 shows the topology for this example, and the corresponding IP addresses. CE-11 exports routes such as 192.168.100.2/32 to the base router and CE-12 exports routes such as 192.168.121.2/32 to the base router. The routes will be leaked from the base router to VPRN 2.

Figure 144 **BGP IPv4 Route Leaking from GRT to VPRN**



The GRT in PE-1 includes BGP routes learned from CE-11 and CE-12, as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.17.111.0/30                               Local   Local     00h21m08s  0
       int-PE-1-CE-11                                               0
172.17.112.0/30                               Local   Local     00h21m08s  0
       int-PE-1-CE-12                                               0
192.0.2.1/32                                  Local   Local     00h21m08s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h20m48s  15
       192.168.12.2                                                 10
192.168.12.0/30                               Local   Local     00h21m08s  0
       int-PE-1-PE-2                                                0
192.168.100.2/32                              Remote  BGP       00h01m21s  170
       172.17.111.2                                                 0
192.168.100.3/32                              Remote  BGP       00h01m21s  170
       172.17.111.2                                                 0
192.168.100.4/30                              Remote  BGP       00h01m21s  170
       172.17.111.2                                                 0
192.168.121.2/32                              Remote  BGP       00h01m22s  170
       172.17.112.2                                                 0
192.168.121.3/32                              Remote  BGP       00h01m22s  170
       172.17.112.2                                                 0
192.168.121.4/30                              Remote  BGP       00h01m22s  170
       172.17.112.2                                                 0
```

```
-------------------------------------------------------------------------------
No. of Routes: 11
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The BGP leaking policy is the same as in the preceding examples:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-Leak-Policy"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

The BGP route leaking policy is applied on the base router for neighbor CE-11 only, as follows:

```
configure
    router
        bgp
            group "EBGP_64500to64501_IPv4"
                neighbor 172.17.111.2
                    import "BGP-Leak-Policy"
                exit
            exit
        exit
    exit
exit
```

The following command shows that only the routes imported from neighbor CE-11 are marked as leakable in the GRT:

```
*A:PE-1# show router bgp routes ipv4 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
```

```
================================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
--------------------------------------------------------------------------------
u*>i  192.168.100.2/32                                    None        None
      172.17.111.2                                        None        -
      64501
u*>i  192.168.100.3/32                                    None        None
      172.17.111.2                                        None        -
      64501
u*>?  192.168.100.4/30                                    None        None
      172.17.111.2                                        None        -
      64501
--------------------------------------------------------------------------------
Routes : 3
================================================================================
*A:PE-1#
```

The leakable BGP routes in the GRT can be imported into VPRN 2. The import policy is identical to the import policy in the preceding examples, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This import policy is applied in VPRN 2, as follows:

```
configure
    service
        vprn 2
            bgp
                rib-management
                    ipv4
                        leak-import "Import-Leakable-Routes"
                    exit
                exit
            exit
        exit
    exit
exit
```

The following command shows the imported leaked BGP routes in VPRN 2. The source of these leaked routes is the base router, not a VPRN.

```
*A:PE-1# show router 2 bgp routes ipv4 leaked
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 192.168.100.2/32                                   100         None
      172.17.111.2 (Base)                                None        -
      64501
u*>li 192.168.100.3/32                                   100         None
      172.17.111.2 (Base)                                None        -
      64501
u*>l? 192.168.100.4/30                                   100         None
      172.17.111.2 (Base)                                None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

Any of these leaked BGP routes has the flag "leaked", and the route source is the base router (leaked from base), as shown:

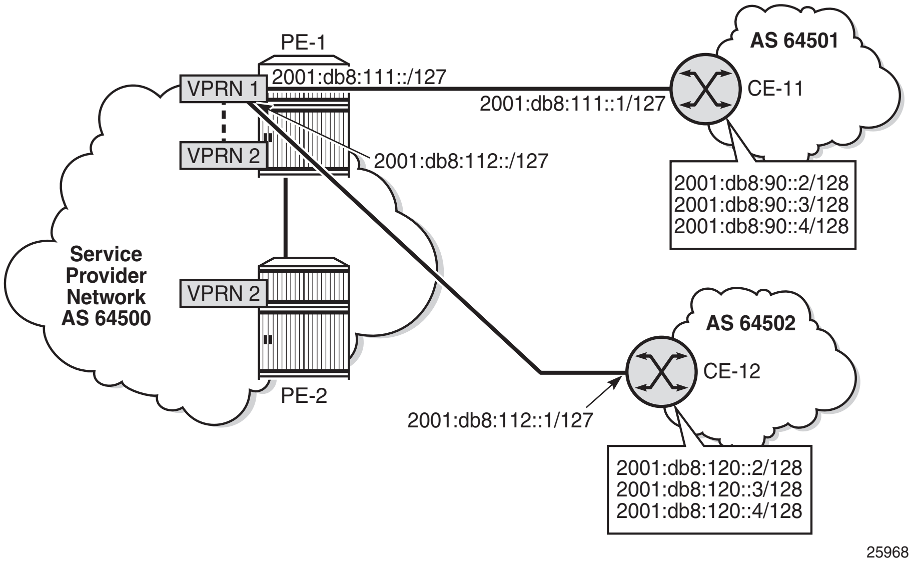
```
*A:PE-1# show router 2 bgp routes 192.168.100.2/32 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 192.168.100.2/32
Nexthop        : 172.17.111.2 (Base)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 172.17.111.2
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
```

```
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from Base
AS-Path        : 64501
--- snipped ---
===============================================================================
*A:PE-1#
```

Example 5 - BGP IPv6 Route Leaking between VPRNs. Global VPRN BGP Configuration

Figure 6 shows the topology and the IP addresses used for this example. CE-11 exports routes such as 2001:db8:90::2/128 to VPRN 1 on PE-1, and CE-12 exports routes such as 2001:db8:120::2/128 to VPRN 1 on PE-1.

Figure 145 **BGP IPv6 Route Leaking between VPRNs**



```
*A:PE-1# show router 1 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1:1/128                             Local   Local     00h25m40s  0
```

```
       system                                                      0
2001:db8:90::2/128                            Remote  BGP       00h01m22s  170
       2001:db8:111::1                                             0
2001:db8:90::3/128                            Remote  BGP       00h01m22s  170
       2001:db8:111::1                                             0
2001:db8:90::4/126                            Remote  BGP       00h01m22s  170
       2001:db8:111::1                                             0
2001:db8:111::/127                            Local   Local     00h25m40s  0
       int-PE-1-CE-11                                              0
2001:db8:112::/127                            Local   Local     00h25m40s  0
       int-PE-1-CE-12                                              0
2001:db8:120::2/128                           Remote  BGP       00h01m04s  170
       2001:db8:112::1                                             0
2001:db8:120::3/128                           Remote  BGP       00h01m04s  170
       2001:db8:112::1                                             0
2001:db8:120::4/126                           Remote  BGP       00h01m04s  170
       2001:db8:112::1                                             0
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The BGP route leaking policy is the same as for IPv4 routes:

```
configure
    router
        policy-options
            begin
            policy-statement "BGP-Leak-Policy"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

This import policy is applied in the BGP context of VPRN 1, as follows:

```
configure
    service
        vprn 1
            bgp
                import "BGP-Leak-Policy"
            exit
        exit
    exit
exit
```

With the preceding configuration, all the routes imported into the VPRN using BGP are marked as leakable.

The following command shows which BGP IPv6 routes are marked as leakable in VPRN 1:

```
*A:PE-1# show router 1 bgp routes ipv6 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2001:db8:90::2/128                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>i  2001:db8:90::3/128                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>?  2001:db8:90::4/126                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>i  2001:db8:120::2/128                                None        None
      2001:db8:112::1                                    None        -
      64502
u*>i  2001:db8:120::3/128                                None        None
      2001:db8:112::1                                    None        -
      64502
u*>?  2001:db8:120::4/126                                None        None
      2001:db8:112::1                                    None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

The BGP leakable routes can be imported into VPRN 2 when the following import policy is configured and applied in VPRN 2:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The only difference from IPv4 routes is that the policy is applied to the IPv6 context of the RIB management:

```
configure
    service
        vprn 2
            bgp
                rib-management
                    ipv6
                        leak-import "Import-Leakable-Routes"
                    exit
                exit
            exit
        exit
    exit
exit
```

The following command shows that the VPRN is importing the leaked BGP IPv6 routes from another VPRN instance:

```
*A:PE-1# show router 2 bgp routes ipv6 leaked
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 2001:db8:90::2/128                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>li 2001:db8:90::3/128                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>l? 2001:db8:90::4/126                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>li 2001:db8:120::2/128                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
u*>li 2001:db8:120::3/128                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
```

```
u*>l? 2001:db8:120::4/126                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

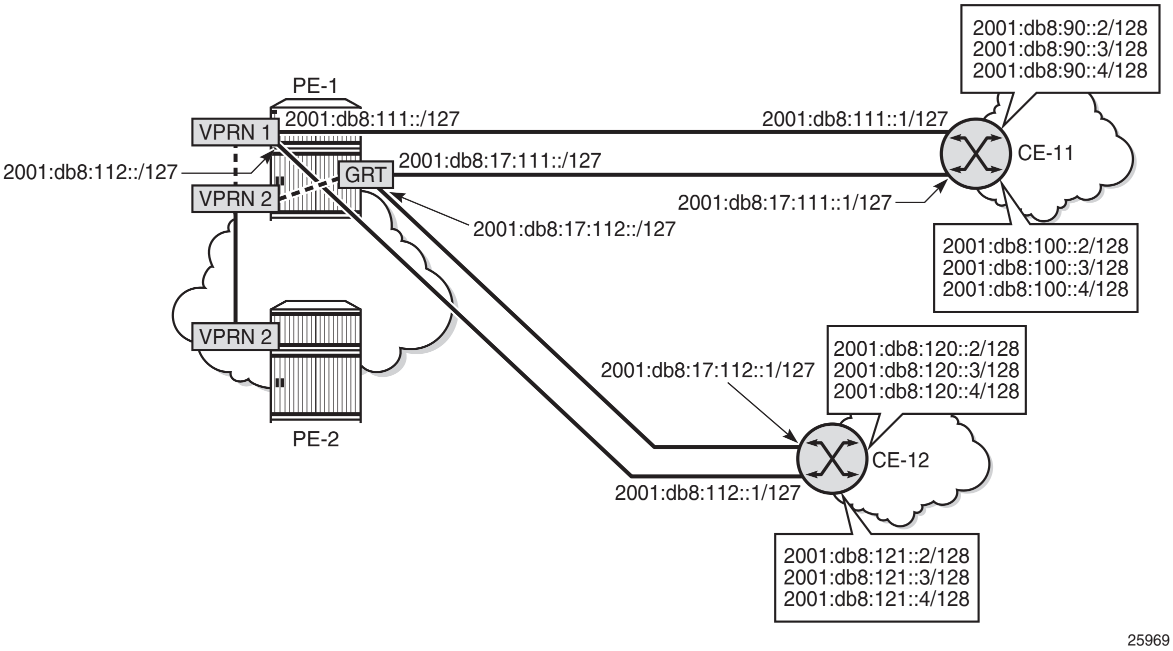
The BGP routes have the flag "leaked" and the route source is VPRN 1, as follows:

```
*A:PE-1# show router 2 bgp routes 2001:db8:90::2/128 detail
===============================================================================
 BGP Router ID:192.0.2.1       AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Original Attributes

Network        : 2001:db8:90::2/128
Nexthop        : 2001:db8:111::1 (VPRN 1)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 2001:db8:111::1
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from VPRN 1
AS-Path        : 64501
--- snipped ---
===============================================================================
*A:PE-1#
```

Example 6 - BGP IPv6 Route Leaking from GRT to VPRN and from VPRN to VPRN

Figure 146 shows the topology and the IPv6 addresses used in this example. CE-11 exports IPv6 routes such as 2001:db8:90::2/128 to VPRN 1 and IPv6 routes such as 2001:db8:100::2/128 to the GRT. CE-12 exports IPv6 routes such as 2001:db8:120::2/128 to VPRN 1 and IPv6 routes such as 2001:db8:121::2/128 to the GRT.

Figure 146 **BGP IPv6 Route Leaking from GRT and VPRN to VPRN**



The IPv6 routing table in the GRT contains routes exported by CE-11 and CE-12, as follows:

```
*A:PE-1# show router route-table ipv6

===============================================================================
IPv6 Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1/128                               Local   Local     00h31m24s  0
       system                                                      0
2001:db8::2/128                               Remote  ISIS      00h31m04s  15
       fe80::b:1ff:fe01:1-"int-PE-1-PE-2"                          10
2001:db8:12::/126                             Local   Local     00h31m23s  0
       int-PE-1-PE-2                                               0
```

```
2001:db8:17:111::/127                         Local   Local     00h31m22s  0
       int-PE-1-CE-11                                               0
2001:db8:17:112::/127                         Local   Local     00h31m23s  0
       int-PE-1-CE-12                                               0
2001:db8:100::2/128                           Remote  BGP       00h01m18s  170
       2001:db8:17:111::1                                           0
2001:db8:100::3/128                           Remote  BGP       00h01m18s  170
       2001:db8:17:111::1                                           0
2001:db8:100::4/126                           Remote  BGP       00h01m18s  170
       2001:db8:17:111::1                                           0
2001:db8:121::2/128                           Remote  BGP       00h01m22s  170
       2001:db8:17:112::1                                           0
2001:db8:121::3/128                           Remote  BGP       00h01m22s  170
       2001:db8:17:112::1                                           0
2001:db8:121::4/126                           Remote  BGP       00h01m22s  170
       2001:db8:17:112::1                                           0
-------------------------------------------------------------------------------
No. of Routes: 11
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The IPv6 routing table for VPRN 1 also contains routes exported by CE-11 and CE-12, as follows:

```
*A:PE-1# show router 1 route-table ipv6

===============================================================================
IPv6 Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
2001:db8::1:1/128                             Local   Local     00h31m22s  0
       system                                                       0
2001:db8:90::2/128                            Remote  BGP       00h02m03s  170
       2001:db8:111::1                                              0
2001:db8:90::3/128                            Remote  BGP       00h02m03s  170
       2001:db8:111::1                                              0
2001:db8:90::4/126                            Remote  BGP       00h02m03s  170
       2001:db8:111::1                                              0
2001:db8:111::/127                            Local   Local     00h31m22s  0
       int-PE-1-CE-11                                               0
2001:db8:112::/127                            Local   Local     00h31m21s  0
       int-PE-1-CE-12                                               0
2001:db8:120::2/128                           Remote  BGP       00h02m03s  170
       2001:db8:112::1                                              0
2001:db8:120::3/128                           Remote  BGP       00h02m03s  170
       2001:db8:112::1                                              0
2001:db8:120::4/126                           Remote  BGP       00h02m03s  170
       2001:db8:112::1                                              0
-------------------------------------------------------------------------------
No. of Routes: 9
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The policy to mark imported BGP routes as leakable can be identical to the policy used in the preceding examples. However, in this case, prefix-lists are added as a filter. The base router may accept routes such as 2001:db8:100::2/128 and 2001:db8:121::2/128.

```
configure
    router
        policy-options
            begin
            prefix-list "2001:db8:90::"
                prefix 2001:db8:90::/100 longer
            exit
            prefix-list "2001:db8:120::"
                prefix 2001:db8:120::/100 longer
            exit
            policy-statement "BGP-Leak-Policy_90_120"
                entry 10
                    from
                        protocol bgp
                        prefix-list "2001:db8:90::"
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        prefix-list "2001:db8:120::"
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

This import policy is applied in the general BGP settings for VPRN 1, as follows:

```
configure
    service
        vprn 1
            bgp
                import "BGP-Leak-Policy_90_120"
            exit
        exit
    exit
exit
```

In a similar way, the base router may accept routes such as 2001:8db:100::2/128 and 2001:8db:121::2/128:

```
configure
    router
        policy-options
            begin
            prefix-list "2001:db8:100::"
                prefix 2001:db8:100::/100 longer
            exit
            prefix-list "2001:db8:121::"
                prefix 2001:db8:121::/100 longer
            exit
            policy-statement "BGP-Leak-Policy_100_121"
                entry 10
                    from
                        protocol bgp
                        prefix-list "2001:db8:100::"
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
                entry 20
                    from
                        protocol bgp
                        prefix-list "2001:db8:121::"
                    exit
                    action accept
                        bgp-leak
                    exit
                exit
            exit
            commit
```

This BGP leaking policy is applied for neighbor CE-11 in the base router, as follows. The routes exported by CE-12 will not be marked as leakable.

```
configure
    router
        bgp
            group "EBGP_64500to64501_IPv6"
                neighbor 2001:db8:17:111::1
                    import "BGP-Leak-Policy_100_121"
                exit
            exit
        exit
    exit
exit
```

The following command shows which routes are marked as leakable in the GRT:

```
*A:PE-1# show router bgp routes ipv6 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2001:db8:100::2/128                                None        None
      2001:db8:17:111::1                                 None        -
      64501
u*>i  2001:db8:100::3/128                                None        None
      2001:db8:17:111::1                                 None        -
      64501
u*>?  2001:db8:100::4/126                                None        None
      2001:db8:17:111::1                                 None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-1#
```

The following command shows which routes are marked as leakable in VPRN 1:

```
*A:PE-1# show router 1 bgp routes ipv6 leakable
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2001:db8:90::2/128                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>i  2001:db8:90::3/128                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>?  2001:db8:90::4/126                                 None        None
      2001:db8:111::1                                    None        -
      64501
u*>i  2001:db8:120::2/128                                None        None
      2001:db8:112::1                                    None        -
      64502
u*>i  2001:db8:120::3/128                                None        None
      2001:db8:112::1                                    None        -
      64502
u*>?  2001:db8:120::4/126                                 None        None
      2001:db8:112::1                                     None        -
      64502
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

On PE-1, a policy is created to import the BGP leakable routes (the same as in the preceding examples), as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "Import-Leakable-Routes"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This import policy is configured for IPv6 routes in VPRN2, as follows:

```
configure
    service
        vprn 2
            bgp
                rib-management
                    ipv6
                        leak-import "Import-Leakable-Routes"
                    exit
                exit
            exit
        exit
    exit
exit
```

The following command shows the leaked IPv6 routes in VPRN 2:

```
*A:PE-1# show router 2 bgp routes ipv6 leaked
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
```

```
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>li 2001:db8:90::2/128                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>li 2001:db8:90::3/128                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>l? 2001:db8:90::4/126                                 100         None
      2001:db8:111::1 (VPRN 1)                           None        -
      64501
u*>li 2001:db8:100::2/128                                100         None
      2001:db8:17:111::1 (Base)                          None        -
      64501
u*>li 2001:db8:100::3/128                                100         None
      2001:db8:17:111::1 (Base)                          None        -
      64501
u*>l? 2001:db8:100::4/126                                100         None
      2001:db8:17:111::1 (Base)                          None        -
      64501
u*>li 2001:db8:120::2/128                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
u*>li 2001:db8:120::3/128                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
u*>l? 2001:db8:120::4/126                                100         None
      2001:db8:112::1 (VPRN 1)                           None        -
      64502
-------------------------------------------------------------------------------
Routes : 9
===============================================================================
*A:PE-1#
```

Some of these routes are leaked from the base router and some routes are leaked from VPRN 1. The detailed information for any of these leaked routes shows that the flag "leaked" is present. For route 2001:db8:100::2/128, the route source is the base router, as follows:

```
*A:PE-1# show router 2 bgp routes 2001:db8:100::2/128 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Original Attributes

Network        : 2001:db8:100::2/128
```

```
Nexthop        : 2001:db8:17:111::1 (Base)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 2001:db8:17:111::1
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from Base
AS-Path        : 64501
--- snipped ---
===============================================================================
*A:PE-1#
```

For route 2001:db8:90::2/128, the route source is VPRN 1, as follows:

```
*A:PE-1# show router 2 bgp routes 2001:db8:90::2/128 detail
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv6 Routes
===============================================================================
Original Attributes

Network        : 2001:db8:90::2/128
Nexthop        : 2001:db8:111::1 (VPRN 1)
Path Id        : None
From           : ::
Res. Protocol  : LOCAL                  Res. Metric    : 0
Res. Nexthop   : 2001:db8:111::1
Local Pref.    : 100                    Interface Name : int-PE-1-CE-11
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 0.0.0.0
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP  Leaked
Route Source   : Leaked from VPRN 1
AS-Path        : 64501
--- snipped ---
===============================================================================
*A:PE-1#
```

Conclusion

BGP provides many ways to manipulate routes. In this example, IPv4/IPv6 routes learned from BGP neighbors could be marked as "leakable" and imported into other routing instances (VPRN to VPRN, VPRN to GRT, GRT to VPRN) without the use of communities in the network policy.

BGP Weighted ECMP

This chapter provides information about BGP Weighted ECMP.

Topics in this chapter include:

Applicability

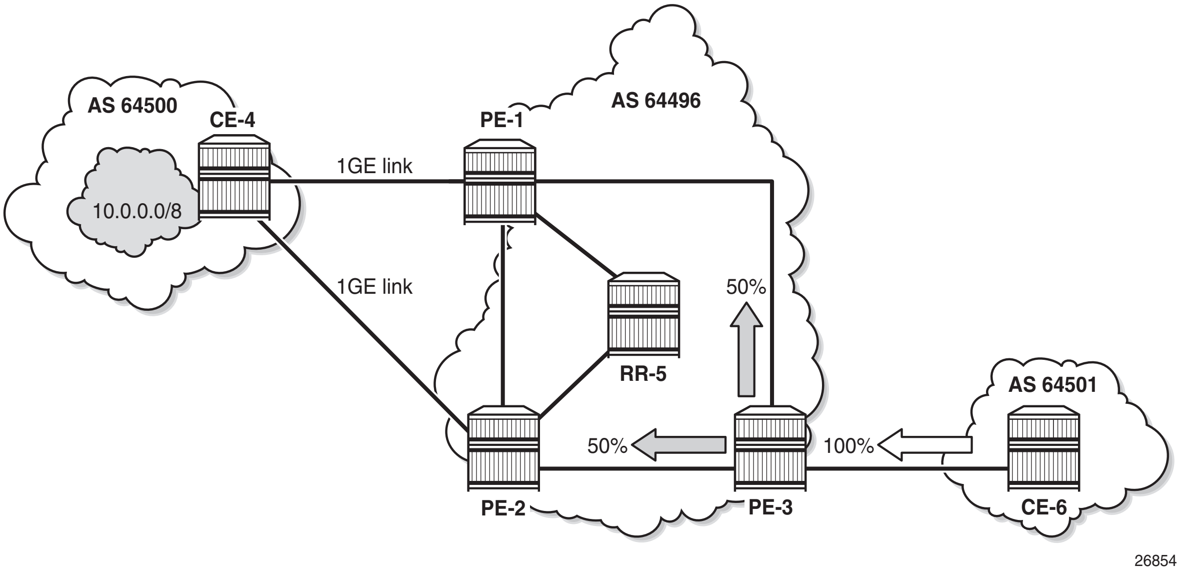
The information and configuration in this chapter was originally based on SR OS Release 15.0.R4. The CLI in the current edition is based on SR OS Release 16.0.R1.

Overview

Equal-cost multipath (ECMP) is a routing strategy that allows the installation of multiple next hops for an IP destination in the routing table. When used in conjunction with BGP multipath, the ingress router can forward traffic to an IP prefix destination in a load-balanced fashion across the available ECMP next hops. For more information about the implementation, see the BGP Multipath chapter.

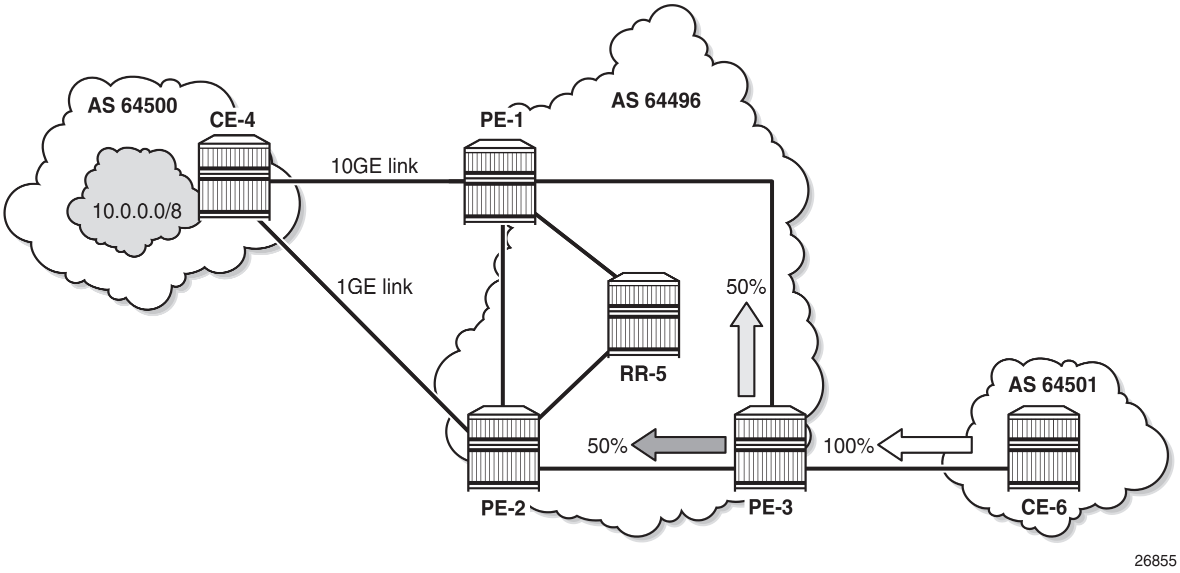
In the standard implementation, ECMP distributes traffic as evenly as possible across all the ECMP next hops. Figure 147 shows an example scenario where CE-4 is dual-homed to two PE routers and advertises the prefix 10.0.0.0/8. This prefix is then advertised within AS 64496 and received by PE-3, which in turn advertises it to CE-6 in AS 64501. PE-3 has BGP multipath and ECMP enabled, so the traffic toward destinations in 10.0.0.0/8 sent by CE-6 is load-balanced toward PE-1 and PE-2 as evenly as possible.

Figure 147 **Standard ECMP - Equal Bandwidth Links**



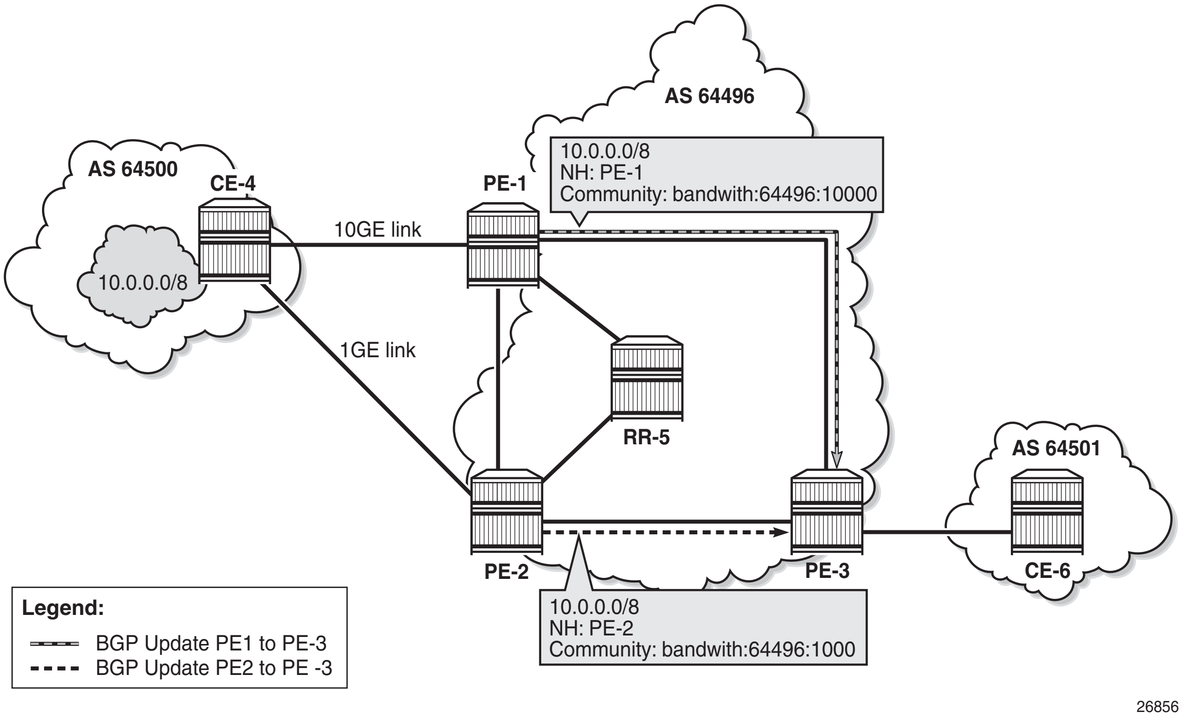
The behavior of equally distributing across the ECMP next hops may not be suitable under certain circumstances. Consider the same topology with the connection between CE-4 and PE-1 replaced with a 10GE link, while the CE-4 to PE-2 connection still is a 1GE link, as shown in Figure 148. In standard ECMP operation, when PE-3 sends 50% of traffic to PE-1 and 50% to PE-2, this may result in an under-utilization of the link between CE-4 and PE-1 or an over-utilization of the link between CE-4 and PE-2.

Figure 148 **Standard ECMP - Unequal Bandwidth Links**



BGP Weighted ECMP, also known as Unequal-Cost Multipath (UCMP), allows for the distribution of traffic in proportion to the relative bandwidth of each equal-cost path. This feature uses a BGP community called the Link Bandwidth Extended Community. Figure 149 shows that PE-1 and PE-2, with this functionality, can add a Link Bandwidth Extended Community to the BGP routes advertised toward other routers within AS 64496 that indicates the bandwidth of their PE-CE link.

Figure 149 Link Bandwidth Extended Community Advertisement



PE-3 can use the information in the Link Bandwidth Extended Community to distribute the traffic according to the relative bandwidth, or the "weight" of each path. Figure 150 shows that 91% of traffic is sent toward PE-1 with the 10GE link and 9% is sent toward PE-2 with the 1GE link.

Figure 150 **Weighted ECMP - Unequal Bandwidth Links**

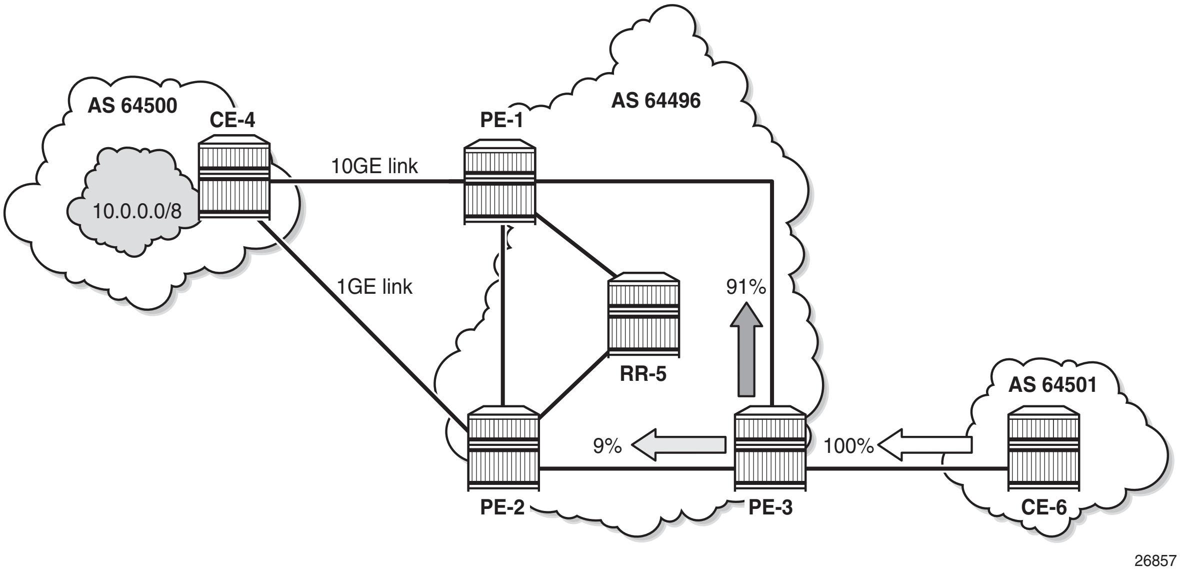


Figure 151 shows another example where the CE-4-to-PE-1 link is composed of four 1GE links that are part of a Link Aggregation Group (LAG) and the CE-4-to-PE-2 link is 1GE. Weighted ECMP can be used here to achieve an 80% to 20% distribution of traffic sent from PE-3 to PE-1 and PE-2, respectively.

Figure 151 **Weighted ECMP - Link Aggregation Group**

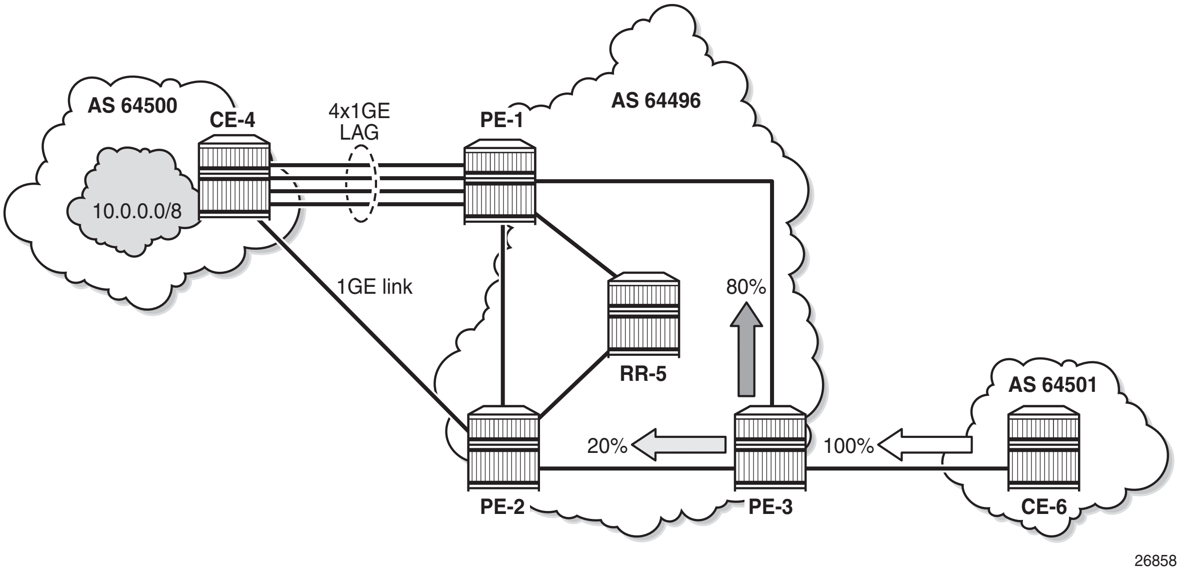


Figure 152 shows an example where PE-1 is connected to two eBGP routers in neighbor AS 64500. Using the weighted ECMP functionality, 91% of traffic is sent to CE-4 and 9% to CE-5, according to the relative bandwidth values.

Figure 152 **Standard ECMP - Unequal Bandwidth Links with eBGP**

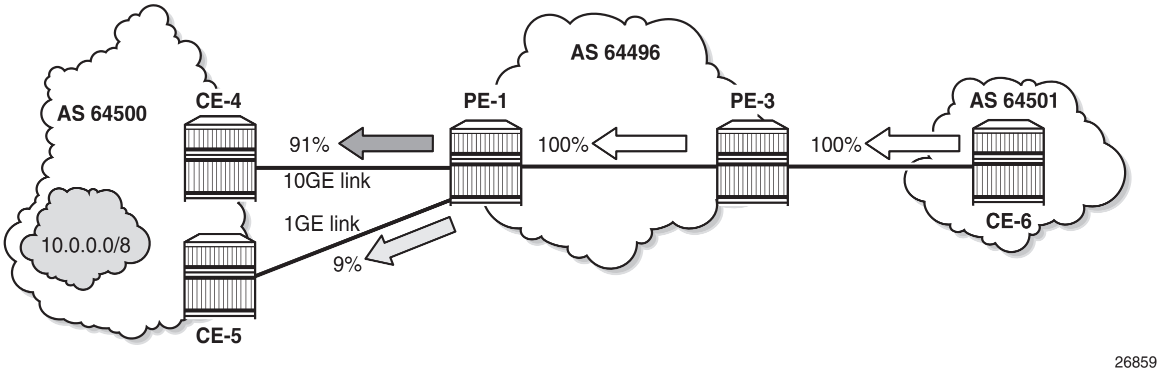
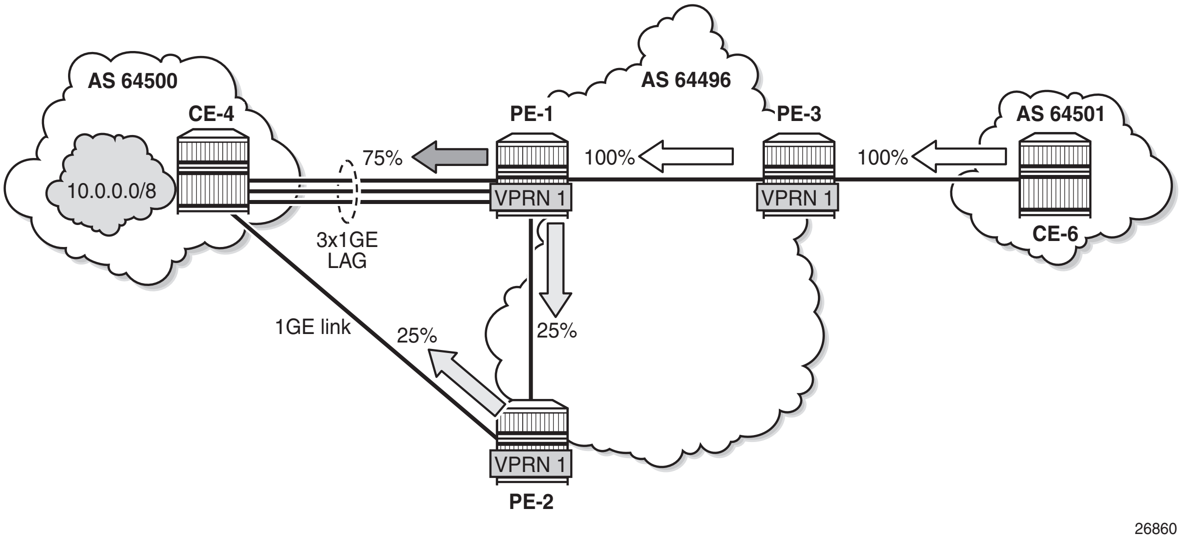


Figure 153 shows an example with a Layer 3 VPRN service. PE-1 receives prefix 10.0.0.0/8 from CE-4 via eBGP, and also from PE-2 via iBGP. PE-1 sets the Link Bandwidth Extended Community indicating 3GE on the route received from CE-4. PE-2 sets the community value indicating 1GE on the route it advertises to PE-1. With Exterior Interior Border Gateway Protocol (EIBGP) multipath (described in the BGP Multipath chapter) and ECMP within the VPRN, PE-1 can send 75% of traffic on the direct LAG link to CE-4 and 25% to PE-2, which then forwards that traffic to CE-4.

Figure 153 **Weighted ECMP - Unequal Bandwidth Links with VPRN**



Link Bandwidth Extended Community is defined in **draft-ietf-idr-link-bandwidth-06** and has the following characteristics:

- Signals the link bandwidth of a BGP path
- Has the following format: bandwidth:<as-number>:<value>
  - bandwidth is the community type
  - <as-number> is the local AS number
  - <value> is a fixed/static bandwidth in Mb/s (converted to IEEE floating point format in a BGP Update message)
- Optional and non-transitive attribute (not sent to other eBGP peers upon receipt)
- If a router changes the route next hop, it does not propagate the Link Bandwidth Extended Community
- A route can only have a single Link Bandwidth Extended Community
- SR OS routers automatically perform weighted load balancing if all the BGP updates received for a destination contain the Link Bandwidth Extended Community

Link Bandwidth Extended Community can be added to a BGP route with the following methods:

- **ebgp-link-bandwidth** command
- BGP import policy action
- VRF import policy action
- BGP export policy action

The **ebgp-link-bandwidth** command has the following characteristics:

- Configurable per BGP group or neighbor in base router or VPRN
- Adds a Link Bandwidth Extended Community to all (IPv4, IPv6, VPN-IPv4, VPN-IPv6, label-IPv4, label-IPv6) routes received from directly connected EBGP peers
- Bandwidth value is based on the speed of port or active LAG members
- Bandwidth is automatically adjusted for LAG interfaces based on the number of active LAG member ports

SR OS uses the following rules when BGP paths are received with Link Bandwidth Extended Communities:

a. If BGP multipath and ECMP are configured and all the eligible multipaths have a Link Bandwidth Extended Community, then weighted ECMP is performed on the relative bandwidth of each path.

b. If EIBGP multipath and ECMP are enabled in a VPRN and all the eligible next hops have a Link Bandwidth Extended Community, then weighted ECMP is performed based on the relative bandwidth of each path.

c. The Link Bandwidth Extended Community is not used as a criterion for two or more paths to be considered equal for BGP/EIBGP multipath purposes.

Configuration

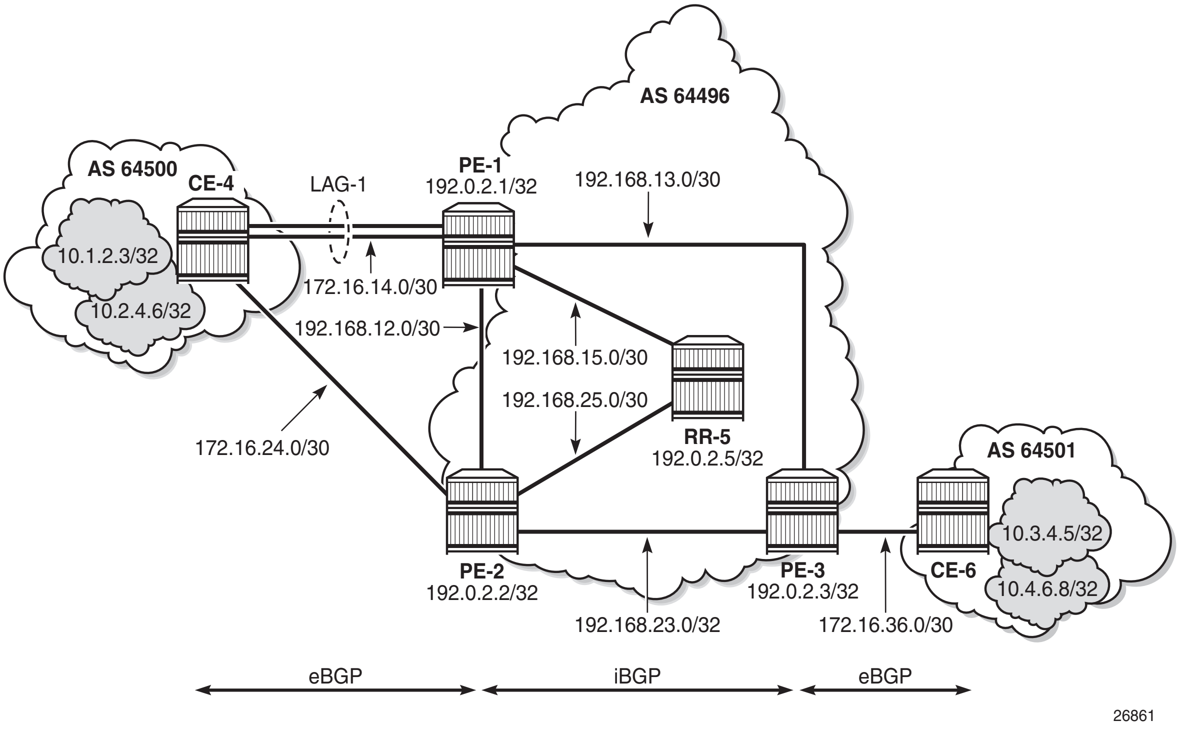
The following configuration examples for BGP weighted ECMP are covered in this chapter:

- BGP Weighted ECMP for IPv4 family using **ebgp-link-bandwidth** command
- BGP Weighted ECMP for IPv4 family using BGP import policy

Figure 154 shows the example topology for BGP Weighted ECMP for IPv4 family with the following characteristics:

- CE-4 in AS 64500 advertises both prefixes 10.1.2.3/32 and 10.2.4.6/32 to its eBGP peers PE-1 and PE-2 in AS 64496.
- RR-5 is route reflector for all PEs in AS 64496.
- Add-Path is configured on all PE routers and RR-5 with a send-limit of 2.
- CE-6 in AS 64501 advertises both prefixes 10.3.4.5/32 and 10.4.6.8/32 to its eBGP peer PE-3 in AS 64496.

Figure 154 **Example Topology - BGP Weighted ECMP for IPv4 Family**



Initial Configuration

The initial configuration on all nodes includes:

- Cards, MDAs, ports
- LAG configured for the link between CE-4 and PE-1 with two member links
- Router interfaces
- IS-IS as IGP on all interfaces within AS 64496 (alternatively, OSPF can be used)

BGP is configured on all the nodes. CE-4 peers with PE-1 and PE-2 and exports the 10.1.2.3/32 and 10.2.4.6/32 loopback prefixes to both eBGP peers, as follows:

```
# on CE-4
configure
    router
        interface "int-loopback-1"
            address 10.1.2.3/16
            loopback
            no shutdown
        exit
        interface "int-loopback-2"
            address 10.2.4.6/16
            loopback
            no shutdown
        exit
        autonomous-system 64500
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            policy-statement "policy-export-bgp"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        bgp
            multipath 2
            rapid-withdrawal
            split-horizon
            group "eBGP"
                export "policy-export-bgp"
                peer-as 64496
                neighbor 172.16.14.1
                exit
                neighbor 172.16.24.1
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on CE-6 is similar, except for the loopback interface addresses.

PE-1 peers with CE-4 in AS 65400 and RR-5 in AS 64496. Add-path is enabled on the iBGP group to advertise redundant BGP paths to the route reflector. The BGP configuration on PE-1 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            rapid-withdrawal
            split-horizon
            group "eBGP"
                peer-as 64500
                neighbor 172.16.14.2
                exit
            exit
            group "iBGP"
                next-hop-self
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
```

The BGP configuration on PE-2 and PE-3 is similar to that of PE-1.

RR-5 acts as a route reflector to all the PEs in AS 64496 with a cluster ID of 5.5.5.5. Add-path is enabled similarly to the PEs. The configuration on RR-5 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            rapid-withdrawal
            split-horizon
            group "iBGP"
                cluster 5.5.5.5
                peer-as 64496
                add-paths
                    ipv4 send 2 receive
                exit
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
            exit
            no shutdown
        exit
```

BGP Weighted ECMP for IPv4 Family using ebgp-link-bandwidth Command

PE-3 receives the prefixes 10.1.2.3/32 and 10.2.4.6/32 from PE-1 and PE-2 via the route reflector and indicates the ones received from PE-1 as the "used" or active route:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  10.1.2.3/32                                        100         None
      192.0.2.1                                          77          -
      64500
*i    10.1.2.3/32                                        100         None
      192.0.2.2                                          78          -
      64500
u*>i  10.2.4.6/32                                        100         None
      192.0.2.1                                          79          -
      64500
*i    10.2.4.6/32                                        100         None
      192.0.2.2                                          80          -
      64500
u*>i  10.3.4.5/32                                        None        None
      172.16.36.2                                        None        -
      64501
u*>i  10.4.6.8/32                                        None        None
      172.16.36.2                                        None        -
      64501
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-3#
```

ECMP and BGP multipath are enabled on PE-3 with the following commands:

```
*A:PE-3# configure router ecmp 2
*A:PE-3# configure router bgp multipath 2
```

As a result, PE-3 installs the routes from PE-2 as active, in addition to those from PE-1:

```
*A:PE-3# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
```

```
u*>i  10.1.2.3/32                                        100         None
      192.0.2.1                                          77          -
      64500
u*>i  10.1.2.3/32                                        100         None
      192.0.2.2                                          78          -
      64500
u*>i  10.2.4.6/32                                        100         None
      192.0.2.1                                          79          -
      64500
u*>i  10.2.4.6/32                                        100         None
      192.0.2.2                                          80          -
      64500
u*>i  10.3.4.5/32                                        None        None
      172.16.36.2                                        None        -
      64501
u*>i  10.4.6.8/32                                        None        None
      172.16.36.2                                        None        -
      64501
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-3#
```

The multiple next hops are also visible in the route table of PE-3:

```
*A:PE-3# show router route-table protocol bgp

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.2.3/32                                   Remote  BGP       00h03m36s  170
       192.168.13.1                                                 0
10.1.2.3/32                                   Remote  BGP       00h03m36s  170
       192.168.23.1                                                 0
10.2.4.6/32                                   Remote  BGP       00h03m36s  170
       192.168.13.1                                                 0
10.2.4.6/32                                   Remote  BGP       00h03m36s  170
       192.168.23.1                                                 0
10.3.4.5/32                                   Remote  BGP       00h08m54s  170
       172.16.36.2                                                  0
10.4.6.8/32                                   Remote  BGP       00h08m54s  170
       172.16.36.2                                                  0
-------------------------------------------------------------------------------
No. of Routes: 6
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The following command shows the routes received on PE-3 have no Community added (do not forget to add the keyword "expression" after the match statement, which appears in the second line below).

```
*A:PE-3# show router bgp routes 10.1.2.3/32 hunt brief |
                                         match "^Nexthop |Community" expression
Nexthop        : 192.0.2.1
Community      : No Community Members
Nexthop        : 192.0.2.2
Community      : No Community Members
*A:PE-3#
```

The following command output shows the ECMP-Weight outputs assigned to next hops 192.0.2.1 and 192.0.2.2. Both have a value of 1.

```
*A:PE-3# show router fib 1 10.1.2.3/32 extensive

===============================================================================
FIB Display (Router: Base)
===============================================================================
Dest Prefix             : 10.1.2.3/32
  Protocol              : BGP
  Installed             : Y
  Indirect Next-Hop     : 192.0.2.1
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.13.1
      Interface         : int-PE-3-PE-1
      ECMP-Weight       : 1
  Indirect Next-Hop     : 192.0.2.2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.23.1
      Interface         : int-PE-3-PE-2
      ECMP-Weight       : 1
===============================================================================
Total Entries : 1
===============================================================================
*A:PE-3#
```

The following command is executed on both PE-1 and PE-2 to automatically add a Link Bandwidth Extended Community on routes received from their eBGP neighbor CE-4:

```
*A:PE-1# configure router bgp group "eBGP" ebgp-link-bandwidth ipv4
```

```
*A:PE-2# configure router bgp group "eBGP" ebgp-link-bandwidth ipv4
```

PE-3 now receives the routes from PE-1 and PE-2 with Link Bandwidth Extended Communities corresponding to the interface bandwidth for each CE-PE link:

```
*A:PE-3# show router bgp routes 10.1.2.3/32 hunt brief |
                                         match "^Nexthop |Community" expression
Nexthop        : 192.0.2.1
```

```
Community      : bandwidth:64496:2000
Nexthop        : 192.0.2.2
Community      : bandwidth:64496:1000
*A:PE-3#
```

The following command output now shows the ECMP-Weight value assigned to next hop 192.0.2.1 is 2, relative to its two member interfaces in the LAG, whereas the ECMP-Weight value of 192.0.2.2 is still 1, because it has a single interface to CE-4:

```
*A:PE-3# show router fib 1 10.1.2.3/32 extensive

===============================================================================
FIB Display (Router: Base)
===============================================================================
Dest Prefix             : 10.1.2.3/32
  Protocol              : BGP
  Installed             : Y
  Indirect Next-Hop     : 192.0.2.1
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 2
    Resolving Next-Hop  : 192.168.13.1
      Interface         : int-PE-3-PE-1
      ECMP-Weight       : 1
  Indirect Next-Hop     : 192.0.2.2
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.23.1
      Interface         : int-PE-3-PE-2
      ECMP-Weight       : 1
===============================================================================
Total Entries : 1
===============================================================================
*A:PE-3#
```

If a tester tool is available, it can be used to test the traffic load-balancing behavior by using it to replace CE-4 and CE-6 in the topology. This would be the preferred option to get better results in observing the effect of weighted ECMP. Multiple flows (preferably a couple of hundred or thousands) should be created and sent between the tester ports. For a simple test, the SR OS rapid ping tool can be used to create traffic between the loopback interfaces of CE-6 and CE-4.

At least three flows need to be created in order to see traffic distributed over the two LAG links between CE-4 and PE-1 and the single link between CE-4 and PE-2. The loopback IP addresses on CE-4 and CE-6 have been specifically chosen to demonstrate the expected load balancing. The behavior might be different if different loopback IP addresses are used, because it will affect the load-balancing algorithm.

To facilitate the test, two more Telnet or SSH sessions are initiated to CE-6 (three in total) and the following commands are executed in each separate session:

First session:

```
*A:CE-6# ping 10.1.2.3 source 10.3.4.5 size 1200 count 100000 rapid
```

Second session:

```
*A:CE-6# ping 10.2.4.6 source 10.4.6.8 size 1200 count 100000 rapid
```

Third session:

```
*A:CE-6# ping 10.1.2.3 source 10.4.6.8 size 1200 count 100000 rapid
```

The **monitor** command outputs on PE-1 and PE-2 show the traffic from CE-6 to CE-4 is being distributed over the two LAG links on PE-1 and the single link on PE-2. In the ideal case, PE-1 would receive 66% and PE-2 would receive 33% of total traffic; however, it may not be possible to observe this effectively with only three ICMP flows.

On the PE-1 LAG link to CE-4, the following traffic is monitored:

```
*A:PE-1# monitor lag 1 interval 3 repeat 999 rate

===============================================================================
Monitor statistics for LAG ID 1
===============================================================================
Port-id      Input packets                   Output packets
             Input bytes                     Output bytes
             Input errors [Input util %]     Output errors [Output util %]
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
1/1/2        6418                            4936
             7977044                         6124556
             0                               0
1/1/5        3                               3
             204                             204
             0                               0
-------------------------------------------------------------------------------
Totals       6421                            4939
             7977248                         6124760
             0                               0

-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
1/1/2        300                             200
             375000                          250000
             0                         0.30  0                         0.20
1/1/5        0                               0
             0                               0
             0                         0.00  0                         0.00
-------------------------------------------------------------------------------
Totals       300                             200
```

```
             375000                           250000
             0                         0.15   0                         0.10

-------------------------------------------------------------------------------
At time t = 6 sec (Mode: Rate)
-------------------------------------------------------------------------------
1/1/2        300                              200
             375000                           250000
             0                         0.30   0                         0.20
1/1/5        0                                0
             0                                0
             0                         0.00   0                         0.00
-------------------------------------------------------------------------------
Totals       300                              200
             375000                           250000
             0                         0.15   0                         0.10

^C
*A:PE-1#
```

On the PE-2 to CE-4 link, the following traffic is monitored (output bytes and output packets):

```
*A:PE-2# monitor port 1/1/1 interval 3 repeat 999 rate

===============================================================================
Monitor statistics for Port 1/1/1
===============================================================================
                                                    Input                Output
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
At time t = 0 sec (Base Statistics)
-------------------------------------------------------------------------------
Octets                                               4227              15259178
Packets                                                50                 12254
Errors                                                  0                     0

-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Octets                                                  0                125000
Packets                                                 0                   100
Errors                                                  0                     0
Bits                                                    0               1000000
Utilization (% of port capacity)                     0.00                  0.10

-------------------------------------------------------------------------------
At time t = 6 sec (Mode: Rate)
-------------------------------------------------------------------------------
Octets                                                  0                125000
Packets                                                 0                   100
Errors                                                  0                     0
Bits                                                    0               1000000
Utilization (% of port capacity)                     0.00                  0.10

-------------------------------------------------------------------------------
At time t = 9 sec (Mode: Rate)
-------------------------------------------------------------------------------
```

```
Octets                                               0               125000
Packets                                              0                  100
Errors                                               0                    0
Bits                                                 0              1000000
Utilization (% of port capacity)                  0.00                 0.10

^C
*A:PE-2#
```

BGP Weighted ECMP for IPv4 Family using BGP Import Policy

The **ebgp-link-bandwidth** command, which was enabled in the previous step, is removed on PE-1 and PE-2:

```
*A:PE-1# configure router bgp group "eBGP" no ebgp-link-bandwidth
```

```
*A:PE-2# configure router bgp group "eBGP" no ebgp-link-bandwidth
```

The following policy is configured on PE-1 to manually add the Link Bandwidth Extended Community "bandwidth:64500:4000" to routes received from CE-4:

```
configure
    router
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            community "bandwidth-4G" members "bandwidth:64500:4000"
            policy-statement "policy-import-bandwidth-4G"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                        community add "bandwidth-4G"
                    exit
                exit
            exit
            commit
        exit
```

The policy is applied on PE-1 for the eBGP group in the import direction:

```
*A:PE-1# configure router bgp group "eBGP" import "policy-import-bandwidth-4G"
```

The following policy is configured on PE-2 to manually add the Link Bandwidth Extended Community "bandwidth:64500:2000" to routes received from CE-4:

```
configure
    router
        policy-options
            begin
            prefix-list "10.0.0.0/8"
                prefix 10.0.0.0/8 longer
            exit
            community "bandwidth-2G" members "bandwidth:64500:2000"
            policy-statement "policy-import-bandwidth-2G"
                entry 10
                    from
                        prefix-list "10.0.0.0/8"
                    exit
                    action accept
                        community add "bandwidth-2G"
                    exit
                exit
            exit
            commit
        exit
```

The policy is applied on PE-2 for the eBGP group in the import direction:

```
*A:PE-2# configure router bgp group "eBGP" import "policy-import-bandwidth-4G"
```

PE-3 receives the routes from PE-1 and PE-2 with Link Bandwidth Extended Communities as configured in the previous step:

```
*A:PE-3# show router bgp routes 10.1.2.3/32 hunt brief |
                                         match "^Nexthop |Community" expression
Nexthop        : 192.0.2.1
Community      : bandwidth:64500:4000
Nexthop        : 192.0.2.2
Community      : bandwidth:64500:2000
*A:PE-3#
```

Again, the following command output shows that the ECMP-Weight output assigned to next hop 192.0.2.1 has become 2:

```
*A:PE-3# show router fib 1 10.1.2.3/32 extensive

===============================================================================
FIB Display (Router: Base)
===============================================================================
Dest Prefix             : 10.1.2.3/32
  Protocol              : BGP
  Installed             : Y
  Indirect Next-Hop     : 192.0.2.1
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 2
    Resolving Next-Hop  : 192.168.13.1
      Interface         : int-PE-3-PE-1
      ECMP-Weight       : 1
  Indirect Next-Hop     : 192.0.2.2
```

```
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : 1
    Resolving Next-Hop  : 192.168.23.1
      Interface         : int-PE-3-PE-2
      ECMP-Weight       : 1
===============================================================================
Total Entries : 1
===============================================================================
*A:PE-3#
```

Note: Any dynamic changes to the Link Bandwidth Extended Community upon failure or bandwidth change of a LAG link are not possible with the policy functionality, as opposed to using the **ebgp-link-bandwidth** command.

Similar tests can be run using the rapid ping facility or an external tester tool as described in the previous section to check the packet forwarding behavior.

Conclusion

BGP Weighted ECMP allows modification of the standard load-balancing behavior to accommodate the relative link bandwidth values of different BGP next hops. This allows better utilization of the links in the network with different capacities. The bandwidth values are advertised by edge routers and carried within a BGP community called the Link Bandwidth Extended Community. SR OS routers automatically perform load balancing if all the BGP routes to a destination contain this community.

Dynamic BGP Peers

This chapter provides information about Dynamic BGP Peers.

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS routers and is based on SR OS Release 14.0.R7.

Overview

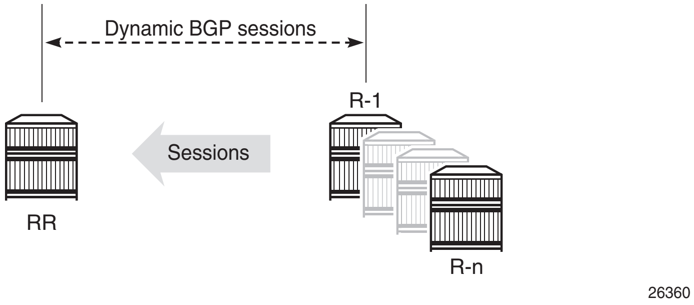
SR OS supports static and dynamic BGP sessions, where the static sessions are initiated toward explicitly configured non-passive neighbors, which are identified through an IPv4 or IPv6 address.

Neighbors must be part of a BGP peer group, and all neighbors in the same group share the same characteristics unless more specific characteristics are defined at the neighbor level.

SR OS will initiate TCP sessions toward explicitly configured non-passive neighbors, and listen for incoming TCP connections on port 179 for these configured neighbors. Sessions established with explicitly configured neighbors are considered static BGP sessions.

Dynamic BGP sessions can be established without explicitly configured neighbors; see Figure 155. The source address of a dynamic peer should match one of the configured IPv4 or IPv6 prefixes. SR OS will only listen for incoming TCP connections on port 179 for these prefixes (which defines passive mode). SR OS will never initiate connections toward dynamic peers. This is consistent with RFC 4271, which allows a BGP speaker to accept connections from unconfigured BGP peers.

Figure 155 **Establishing Dynamic BGP Sessions**



Dynamic BGP peering is also supported for ESM-routed subscriber hosts to improve deployment flexibility, but this is out of the scope of this chapter.

Characteristics

In SR OS, BGP groups and dynamic BGP peers have the following characteristics:

- A BGP group can support static and dynamic peers simultaneously.
- To support dynamic, unconfigured peers, multiple prefixes (IPv4/IPv6) can be associated with a group.
- A dynamic peer will be associated with a group, based on the source IP address of an incoming TCP connection. If multiple overlapping prefixes match, the prefix with the longest prefix length is used.
- A maximum number of dynamic peers can be configured per group and for the entire BGP instance. Whenever an incoming connection for a new dynamic session would cause either a group limit or the overall BGP limit to be exceeded, the connection attempt is rejected with a BGP Notification message.
- Dynamic peers are supported in the base router as well as in VPRN BGP instances.

Behavior

When a dynamic session is established, the following behavior will be observed when changes are made:

- If a new **prefix** entry is added to a group and this entry will become the longest prefix match for the IP address, then the session remains up, without interruption, if the new entry belongs to the same group as the one previously used to set up the dynamic session.
- If a new **prefix** entry is added to a group and this entry becomes the longest prefix match for the IP address, then the session is torn down immediately if the new entry belongs to a different group from the one previously used to set up the dynamic session. When the remote end attempts to reestablish the session, the parameters used locally are inherited from the new group.
- If a **neighbor** command is added to any group and its IP address matches the source IP address of an established dynamic session, then the dynamic session is torn down and the new session that is established inherits its local parameters from the **neighbor** configuration.

Using dynamic BGP peers can reduce the configuration file size of an SR OS router considerably, and is mainly used on route reflectors in large autonomous systems.

Configuration

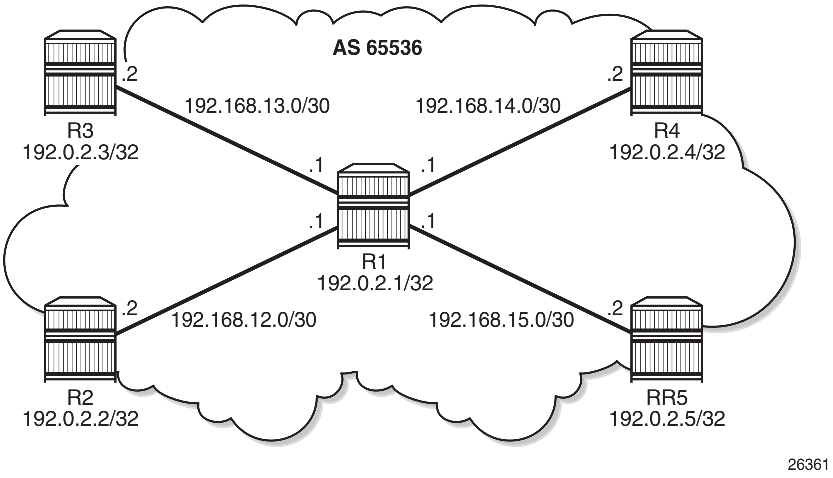
Figure 156 shows the example topology, and has the following characteristics:

- All nodes are part of AS 65536.
- BGP sessions are established between the routers of AS 65536, using RR5 as route reflector with R1, R2, R3, and R4 being the route reflector clients.

The initial configuration on the nodes includes:

- cards, MDAs, and ports
- router interfaces
- IS-IS between the PEs

***Figure 156* Dynamic BGP Peers**



BGP is configured between the route reflector clients and the route reflector for the IPv4 address family. The configuration on R1 is as follows:

```
# R1
configure
    router
        autonomous-system 65536
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                peer-as 65536
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The BGP configuration on the other route reflector clients is the same as on R1.

The initial route reflector RR5 BGP configuration is as follows:

```
# RR5
configure
    router
         autonomous-system 65536
         bgp
            loop-detect discard-route
            split-horizon
            dynamic-neighbor-limit 20
            group "iBGP"
                cluster 5.5.5.5
                peer-as 65536
                dynamic-peer-limit 10
                dynamic-neighbor
                    prefix 192.0.2.0/24
                exit
            exit
            no shutdown
        exit
    exit
exit
```

Dynamic neighbors are shown with the "D" flag, as follows:

```
*A:RR5# show router bgp summary all
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1(D)
Def. Instance  65536       64    0 00h30m53s 0/0/3 (IPv4)
                           67    0
192.0.2.2(D)
Def. Instance  65536       66    0 00h31m11s 1/1/2 (IPv4)
                           67    0
192.0.2.3(D)
Def. Instance  65536       67    0 00h31m49s 1/1/2 (IPv4)
                           68    0
192.0.2.4(D)
Def. Instance  65536       65    0 00h30m47s 1/1/2 (IPv4)
                           66    0
-------------------------------------------------------------------------------
*A:RR5#
```

The details for neighbor R2 show that the session is dynamic, as follows:

```
*A:RR5# show router bgp neighbor 192.0.2.2
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.2
Description          : (Not Specified)
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 50891
Peer Address         : 192.0.2.2
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.5
Peer Type            : Internal         Dynamic Peer         : Yes
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : IPv4
Remote Family        : IPv4
Hold Time            : 90             Keep Alive           : 30
Min Hold Time        : 0
Active Hold Time     : 90             Active Keep Alive    : 30
Cluster Id           : 5.5.5.5
--- snipped ---
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:RR5#
```

The route reflector RR5 is modified, as follows:

```
configure
    router
         bgp
            group "iBGP"
                cluster 5.5.5.5
                peer-as 65536
                dynamic-neighbor
                    prefix 192.0.2.0/24
                exit
                neighbor 192.0.2.1
                    keepalive 20
                    hold-time 60
                exit
            exit
            no shutdown
        exit
    exit
exit
```

Therefore, the properties of BGP group iBGP are as follows:

```
*A:RR5# show router bgp group "iBGP"
===============================================================================
BGP Group : iBGP
===============================================================================
-------------------------------------------------------------------------------
Group            : iBGP
-------------------------------------------------------------------------------
Description      : (Not Specified)
Group Type       : No Type              State            : Up
Peer AS          : 65536                Local AS         : 65536
Local Address    : n/a                  Loop Detect      : Discard
Import Policy    : None Specified / Inherited
Export Policy    : None Specified / Inherited
Hold Time        : 90                   Keep Alive       : 30
Min Hold Time    : 0
Cluster Id       : 5.5.5.5              Client Reflect   : Enabled
NLRI             : Unicast              Preference       : 170
TTL Security     : Disabled             Min TTL Value    : n/a
Graceful Restart : Disabled             Stale Routes Time: n/a
Restart Time     : n/a
```

```
Auth key chain   : n/a
Bfd Enabled      : Disabled             Disable Cap Nego : Disabled
Creation Origin  : manual
Flowspec Validate: Disabled             Default Route Tgt: Disabled
Aigp Metric      : Disabled
Split Horizon    : Enabled
Damp Peer Oscill*: Disabled
GR Notification  : Disabled             Fault Tolerance  : Disabled
Next-Hop Unchang*: None
Routes Resolve T*: Disabled

List of Static Peers
- 192.0.2.1 :

List of Dynamic Peers
- 192.0.2.2
- 192.0.2.3
- 192.0.2.4

Total Peers      : 4                    Established      : 4
-------------------------------------------------------------------------------
Peer Groups : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:RR5#
```

The BGP session toward R1 is static. The short session time is an indication that the BGP session toward R1 has been reestablished, as follows:

```
*A:RR5# show router bgp summary all
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  65536       95    0 00h01m33s 0/0/3 (IPv4)
                           16    0
192.0.2.2(D)
Def. Instance  65536        7    0 00h47m44s 1/1/2 (IPv4)
                            8    0
192.0.2.3(D)
Def. Instance  65536       94    0 00h45m04s 1/1/2 (IPv4)
                           99    0
192.0.2.4(D)
Def. Instance  65536       92    0 00h44m02s 1/1/2 (IPv4)
                           97    0
-------------------------------------------------------------------------------
*A:RR5# #
```

Reestablishment of the BGP session is also indicated in log 99, as follows:

```
62 2017/02/16 14:49:46.48 CET MINOR: BGP #2038 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: moved into established state"

61 2017/02/16 14:49:46.47 CET WARNING: BGP #2011 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: remote end closed connection"

60 2017/02/16 14:49:46.47 CET WARNING: BGP #2005 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: sending notification: code CEASE subcode CONN_COL
L_RES"

59 2017/02/16 14:49:46.45 CET WARNING: BGP #2039 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: moved from higher state ACTIVE to lower state IDL
E due to event CONFIG_CHG"

58 2017/02/16 14:49:46.43 CET WARNING: BGP #2011 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: remote end closed connection"
57 2017/02/16 14:49:46.43 CET WARNING: BGP #2005 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: sending notification: code CEASE subcode CONFIG_C
HG"

56 2017/02/16 14:49:46.43 CET WARNING: BGP #2039 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: moved from higher state CONNECT to lower state ID
LE due to event CONFIG_CHG"

55 2017/02/16 14:49:46.42 CET WARNING: BGP #2005 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: sending notification: code CEASE subcode CONFIG_C
HG"

54 2017/02/16 14:49:46.42 CET WARNING: BGP #2039 Base Peer 1: 192.0.2.1
"VR 1: Group iBGP: Peer 192.0.2.1: moved from higher state ESTABLISHED to lower stat
e IDLE due to event CONFIG_CHG"
```

New and more specific settings apply to R1, as follows:

```
*A:RR5# show router bgp neighbor 192.0.2.1
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.1
Description          : (Not Specified)
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 50907
Peer Address         : 192.0.2.1
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.5
Peer Type            : Internal         Dynamic Peer         : No
State                : Established      Last State           : Established
Last Event           : recvKeepAlive
Last Error           : Cease (Connection Collision Resolution)
Local Family         : IPv4
Remote Family        : IPv4
Hold Time            : 60               Keep Alive           : 20
Min Hold Time        : 0
Active Hold Time     : 60               Active Keep Alive    : 20
Cluster Id           : 5.5.5.5
--- snipped ---
```

```
-------------------------------------------------------------------------------
Neighbors : 1
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:RR5#
```

The properties of all dynamic peers can be displayed using a single command, as follows:

```
*A:RR5# show router bgp neighbor dynamic
===============================================================================
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.2
Description          : (Not Specified)
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 50894
Peer Address         : 192.0.2.2
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.5
Peer Type            : Internal         Dynamic Peer         : Yes
State                : Established      Last State           : Established
--- snipped ---
-------------------------------------------------------------------------------
Peer                 : 192.0.2.3
Description          : (Not Specified)
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 51020
Peer Address         : 192.0.2.3
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.5
Peer Type            : Internal         Dynamic Peer         : Yes
State                : Established      Last State           : Established
--- snipped ---
-------------------------------------------------------------------------------
Peer                 : 192.0.2.4
Description          : (Not Specified)
Group                : iBGP
-------------------------------------------------------------------------------
Peer AS              : 65536            Peer Port            : 50635
Peer Address         : 192.0.2.4
Local AS             : 65536            Local Port           : 179
Local Address        : 192.0.2.5
Peer Type            : Internal         Dynamic Peer         : Yes
State                : Established      Last State           : Established
--- snipped ---
-------------------------------------------------------------------------------
Neighbors : 3
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:RR5#
```

Lowering the dynamic peer limit will not tear down any existing BGP sessions, as follows:

```
*A:RR5# configure router bgp group "iBGP" dynamic-neighbor-limit 2
```

A hard reset of a running BGP session will result in that BGP session being torn down, as follows:

```
*A:RR5# clear router bgp neighbor 192.0.2.4 hard
```

The BGP peer fails to reconnect to the route reflector, because the peer limit has been reached, as follows:

```
66 2017/02/16 15:05:45.15 CET MINOR: BGP #2037 Base VR 1: Group iBGP
"192.0.2.4: Closing connection: reached dynamic peer limit (2) for BGP group iBGP"

65 2017/02/16 15:05:45.14 CET WARNING: BGP #2005 Base Peer 1: 192.0.2.4
"VR 1: Group iBGP: Peer 192.0.2.4: sending notification: code CEASE subcode HARD_RES
ET"

64 2017/02/16 15:05:45.14 CET WARNING: BGP #2039 Base Peer 1: 192.0.2.4
"VR 1: Group iBGP: Peer 192.0.2.4: moved from higher state ESTABLISHED to lower stat
e IDLE due to event ADMIN_RESET_HARD"

63 2017/02/16 15:05:45.12 CET INDETERMINATE: LOGGER #2010 Base Clear BGP
"Clear function clearRtrBgpNbr has been run with parameters: rtr-
name="Base" neighbor="192.0.2.4" type="hard".  The completion result is: success.  A
dditional error text, if any, is: "
```

Conclusion

The use of dynamic BGP peers provides ISPs the means to reduce the configuration file size for route reflectors. This reduces the number of configuration changes to be made to the network over time, which lowers the operational cost of running the network.

EBGP Route Resolution to a Static Route

This chapter provides information about EBGP Route Resolution to a Static Route.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R4. EBGP route resolution to a static route is supported in SR OS Release 14.0.R1, and later.

Overview

The configuration in this chapter resembles the configuration in chapter *VPRN Inter-AS VPN Model C (Layer 3 Services)*, but in this chapter the eBGP peering between the ASBRs is using loopback addresses instead of interface addresses.

Typically, service providers use interface IP addresses in eBGP sessions toward an Autonomous System Border Router (ASBR) of an untrusted ISP, but it is possible to use loopback addresses, such as system IP addresses. This requires the ASBRs to provide visibility on each other's loopback address; for example, by defining static routes. EBGP route resolution to a static route only works for ASBRs that are directly connected. As an alternative, RSVP-TE or LDP can be configured on the interfaces between the ASBRs, which is the only viable solution when the peering ASBRs are multiple hops away.

Configuring MPLS on the interface toward an ASBR of an untrusted ISP is considered insecure. For directly connected ASBRs, EBGP route resolution to a static route mitigates these security issues. On each ASBR, static routes are configured toward the loopback address of the peer ASBR. Additionally, the following command enables labeled routes to be resolved via a static route:

```
configure router bgp next-hop-resolution labeled-routes allow-static
```

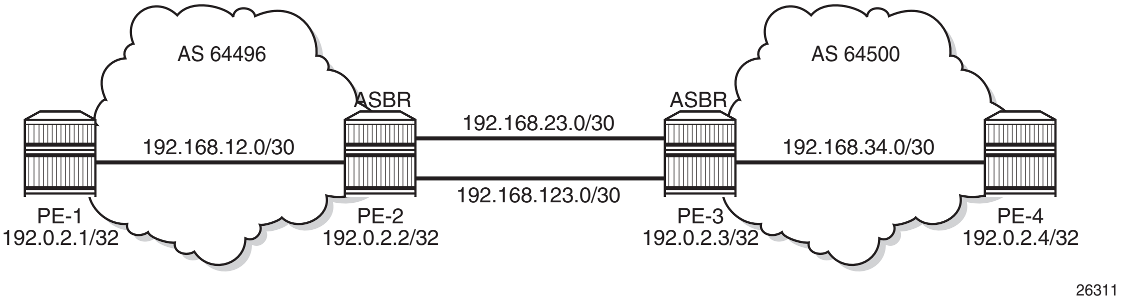
Even with this feature enabled, the system will first try to resolve the BGP next-hop to LDP or RSVP LSPs before the IP route table is attempted. The option is supported for the following address families:

- Labeled IPv4 routes
- VPN-IPv4 and VPN-IPv6 routes

Configuration

Figure 157 shows the example topology with four routers in two different ASs. PE-2 and PE-3 are ASBRs that are connected via two links, which implies that there will be multiple next-hops configured for the static route entry toward the loopback IP address of the eBGP peer. Also, Equal Cost Multi-Path (ECMP) and BGP multipath need to be enabled between these ASBRs.

Figure 157 **Example Topology**



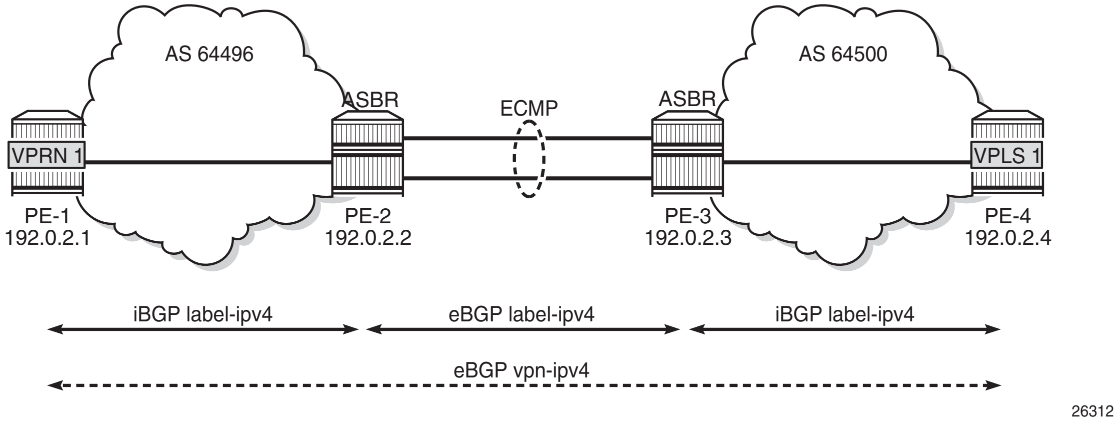
The initial configuration on the nodes includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS as IGP on the interfaces within an AS (alternatively, OSPF could be used)
- LDP on the interfaces within an AS

Figure 158 shows the BGP sessions to be configured:

- iBGP sessions for address family labeled IPv4 between the PEs within each AS
- eBGP sessions for address family labeled IPv4 between the ASBRs PE-2 and PE-3
- a multi-hop eBGP session for address family VPN-IPv4 between PE-1 and PE-4

Figure 158 **BGP Peering**



On PE-1, iBGP is configured for address family labeled IPv4, as follows. The configuration on PE-4 is similar.

```
configure
    router
        autonomous-system 64496
        bgp
            split-horizon
            group "iBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.2
                    family label-ipv4
                exit
            exit
```

The following export policy exports the loopback IP prefixes from PE-1 to ASBR PE-2 (and from PE-4 to ASBR PE-3):

```
configure
    router
        policy-options
            begin
            prefix-list "PE-sys"
                prefix 192.0.2.0/28 longer
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        protocol direct
                        prefix-list "PE-sys"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

On PE-2, iBGP and eBGP are configured for address family labeled IPv4, as follows. Two links are connecting PE-2 to PE-3 and, therefore, ECMP and BGP multipath are enabled. For more information about BGP multipath, see chapter BGP Multipath. The BGP configuration on PE-3 is similar.

```
configure
    router
        autonomous-system 64496
        ecmp 2
        bgp
            multipath 2 ebgp 2
            split-horizon
            group "eBGP"
                peer-as 64500
                neighbor 192.0.2.3
                    family label-ipv4
                    advertise-inactive
                exit
            exit
            group "iBGP"
                peer-as 64496
                neighbor 192.0.2.1
                    family label-ipv4
                exit
            exit
        exit
```

On the ASBRs, the BGP routes with the loopback IP addresses of the local AS PEs are not active because IGP routes are preferred. The **advertise-inactive** option ensures that the ASBRs additionally advertise these inactive routes to each other. PE-2 advertises prefix 192.0.2.1/32 to PE-3; PE-3 advertises prefix 192.0.2.4/32 to PE-2. This way, no export policy is required for the eBGP session between ASBRs. However, no prefixes can be exchanged between the ASBRs because the eBGP session is not in the established state yet; they still lack routing to each other's loopback IP address.

Eventually, the labeled IPv4 routes for prefixes PE-1 and PE-4 will be exchanged between ASBRs and forwarded to the PEs in the peer AS. PE-1 will have a route toward PE-4 in its routing table, and PE-4 will have a route toward PE-1. Both PEs can then set up a multi-hop eBGP session to each other for address family VPN-IPv4; for example, on PE-1, as follows:

```
configure
    router
        bgp
            group "eBGP_multihop"
                family vpn-ipv4
                peer-as 64500
                local-address 192.0.2.1
                neighbor 192.0.2.4
                    multihop 10
                    vpn-apply-export
                    export "EBGP-VPN-IPv4"
                exit
            exit
```

The export policy "EBGP-VPN-IPv4" is not defined and not required in this example, but usually some export policy would be used.

On PE-1, VPRN 1 is configured with loopback address 10.1.1.1/32, as follows:

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "loopback" create
                address 10.1.1.1/32
                loopback
            exit
            no shutdown
```

The configuration of PE-4 resembles the configuration of PE-1, whereas the configuration of PE-3 resembles that of PE-2.

This configuration for inter-AS VPN model C is almost identical to the configuration in chapter *VPRN Inter-AS VPN Model C*, with the difference that the eBGP session between the ASBRs does not use interface IP addresses, but loopback addresses. The problem is that the ASBRs cannot reach each other's loopback IP address, so the eBGP session between the ASBRs cannot be established, which can be verified in the BGP summary, as follows:

```
*A:PE-2# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496       14    0 00h04m49s 1/0/0 (Lbl-IPv4)
                           14    0
192.0.2.3
Def. Instance  64500        0    0 00h04m49s Connect
                            1    0
-------------------------------------------------------------------------------
```

```
*A:PE-2#
```

The state of the BGP session toggles between Active and Connect. The last event is an openFail, as follows:

```
*A:PE-2# show router bgp neighbor 192.0.2.3 detail | match "BGP Neighbor" post-lines 15
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.3
Description          : (Not Specified)
Group                : eBGP
-------------------------------------------------------------------------------
Peer AS              : 64500            Peer Port            : 0
Peer Address         : 192.0.2.3
Local AS             : 64496            Local Port           : 0
Local Address        : 0.0.0.0
Peer Type            : External         Dynamic Peer         : No
State                : Active           Last State           : Connect
Last Event           : openFail
Last Error           : Cease (Other Configuration Change)
Local Family         : LABEL-IPv4
*A:PE-2#
```

When the eBGP session between the ASBRs is not established, no IP prefixes will be learned from the peer AS. This implies that PE-1 will not have a route toward PE-4 in its routing table. Therefore, no multi-hop eBGP session can be established between PE-1 and PE-4, which can be shown as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h10m50s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h10m40s  15
       192.168.12.2                                                 10
192.168.12.0/30                               Local   Local     00h10m50s  0
       int-PE-1-PE-2                                                0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router bgp summary | match "BGP Summary" post-lines 15
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
```

```
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
               64496       29    0 00h12m53s 0/0/1 (Lbl-IPv4)
                           31    0
192.0.2.4
               64500        0    0 00h10m06s Connect
                            0    0
-------------------------------------------------------------------------------
*A:PE-1#
```

The state of the multi-hop eBGP session toggles between Active and Connect. The last event is openFail, as follows:

```
*A:PE-1# show router bgp neighbor 192.0.2.4 detail | match "BGP Neighbor" post-lines 15
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.4
Description          : (Not Specified)
Group                : eBGP_multihop
-------------------------------------------------------------------------------
Peer AS              : 64500            Peer Port            : 0
Peer Address         : 192.0.2.4
Local AS             : 64496            Local Port           : 0
Local Address        : 0.0.0.0
Peer Type            : External         Dynamic Peer         : No
State                : Connect          Last State           : Active
Last Event           : openFail
Last Error           : Unrecognized Error
Local Family         : VPN-IPv4
*A:PE-1#
```

The loopback IP addresses of the ASBRs can be made reachable by configuring static routes on each ASBR to the loopback IP address of the peer ASBR. This will be sufficient to establish the eBGP session between the ASBRs, but no BGP labeled IPv4 routes will be advertised to PE-1 and PE-4 yet. PE-2 and PE-3 are connected by two links and the static route entry contains two next-hops; for example, for PE-2, as follows. The configuration is similar for PE-3.

```
configure
    router
        static-route-entry 192.0.2.3/32
            next-hop 192.168.23.2
                no shutdown
            exit
            next-hop 192.168.123.2
                no shutdown
            exit
        exit
```

The routing table in ASBR PE-2 contains two routes toward PE-3, as follows:

```
*A:PE-2# show router route-table 192.0.2.3/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age         Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.3/32                                  Remote  Static    00h00m13s  5
       192.168.23.2                                                 1
192.0.2.3/32                                  Remote  Static    00h00m13s  5
       192.168.123.2                                                1
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

The eBGP session between the ASBRs is established; for example, on PE-2, as follows:

```
*A:PE-2# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496       40    0 00h17m38s 1/0/0 (Lbl-IPv4)
                           40    0
192.0.2.3
Def. Instance  64500        5    0 00h00m58s 1/0/1 (Lbl-IPv4)
                            6    0
-------------------------------------------------------------------------------
*A:PE-2#
```

However, the multi-hop eBGP session between PE-1 and PE-4 is not established yet. The state of the multi-hop eBGP session toggles between active and connect and the following output from PE-1 shows that the last event was openFail:

```
*A:PE-1# show router bgp neighbor 192.0.2.4 detail | match "BGP Neighbor" post-lines 15
BGP Neighbor
===============================================================================
-------------------------------------------------------------------------------
Peer                 : 192.0.2.4
```

```
Description          : (Not Specified)
Group                : eBGP_multihop
-------------------------------------------------------------------------------
Peer AS              : 64500           Peer Port            : 0
Peer Address         : 192.0.2.4
Local AS             : 64496           Local Port           : 0
Local Address        : 0.0.0.0
Peer Type            : External        Dynamic Peer         : No
State                : Connect         Last State           : Active
Last Event           : openFail
Last Error           : Unrecognized Error
Local Family         : VPN-IPv4
*A:PE-1#
```

PE-2 advertised an inactive route for prefix 192.0.2.1/32 to PE-3 and received from PE-3 an inactive route for prefix 192.0.2.4/32. The following output shows that the route for prefix 192.0.2.4/32 is not valid on PE-2:

```
*A:PE-2# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    192.0.2.1/32                                       100         None
      192.0.2.1                                          None        262141
      No As-Path
i     192.0.2.4/32                                       None        None
      192.0.2.3                                          None        262141
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

Consequently, PE-2 does not advertise this invalid route to its iBGP peer PE-1 and PE-1 will not have a route toward PE-4 in its routing table, as follows:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
```

```
192.0.2.1/32                                  Local   Local      00h23m51s  0
       system                                                        0
192.0.2.2/32                                  Remote  ISIS       00h23m41s  15
       192.168.12.2                                                  10
192.168.12.0/30                               Local   Local      00h23m51s  0
       int-PE-1-PE-2                                                 0
-------------------------------------------------------------------------------
No. of Routes: 3
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

PE-1 and PE-4 cannot set up a multi-hop eBGP session to one another to exchange routes for VPRN 1. This problem can be solved in two different ways:

1. Enable RSVP-TE or LDP on the interfaces between the ASBRs.
2. Enable the following option: **configure router bgp next-hop-resolution labeled-routes allow-static**.

It is risky to enable MPLS toward a peer ASBR belonging to an untrusted ISP, but it is required between distant ASBRs if loopback addresses are used in eBGP peering.

In the following section, the first solution is described (LDP is enabled on the interfaces between the ASBRs); the section after that describes how to enable eBGP route resolution to a static route.

Enable LDP toward Peer ASBR

LDP is configured on the interfaces between the ASBRs; for example, on PE-2, as follows. The configuration is similar on PE-3.

```
configure
    router
       ldp
            interface-parameters
                interface "int-PE-2-PE-3_1st"
                exit
                interface "int-PE-2-PE-3_2nd"
                exit
            exit
        exit
```

PE-2 now has a valid, best, and used route for prefix 192.0.2.4/32, as follows:

```
*A:PE-2# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    192.0.2.1/32                                       100         None
      192.0.2.1                                          None        262141
      No As-Path
u*>i  192.0.2.4/32                                       None        None
      192.0.2.3                                          None        262141
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

PE-1 has a valid route for prefix 192.0.2.4/32, as follows:

```
*A:PE-1# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.1       AS:64496        Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
      Nexthop (Router)                                   Path-Id     Label
Flag  Network                                            LocalPref   MED
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.4/32                                       100         None
      192.0.2.2                                          None        262138
      64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The following routing table shows that PE-1 has a BGP labeled route toward PE-4:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h43m23s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      00h43m13s  15
       192.168.12.2                                                 10
192.0.2.4/32                                  Remote  BGP_LABEL 00h17m57s  170
       192.0.2.2 (tunneled)                                         10
192.168.12.0/30                               Local   Local     00h43m23s  0
       int-PE-1-PE-2                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

A multi-hop eBGP session is established for address family VPN-IPv4 between PE-1 and PE-4, as follows:

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
Def. Instance  64496       93    0 00h44m01s 1/1/1 (Lbl-IPv4)
                           94    0
192.0.2.4
Def. Instance  64500       46    0 00h21m06s 1/1/1 (VpnIPv4)
                           47    0

-------------------------------------------------------------------------------
*A:PE-1#
```

The loopback address defined in VPRN 1 on PE-4 (10.2.2.2/32) is advertised as VPN-IPv4 route in this multi-hop eBGP session on PE-1, as follows:

```
*A:PE-1# show router bgp routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64500:1:10.2.2.2/32                                None        None
      192.0.2.4                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The routing table for VPRN 1 on PE-1 includes a BGP-VPN route to PE-4, as follows:

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.1.1/32                                   Local   Local     00h44m02s  0
       loopback                                                     0
10.2.2.2/32                                   Remote  BGP VPN   00h23m23s  170
       192.0.2.4 (tunneled:BGP)                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

To restore the configuration, LDP is disabled on the interfaces between the ASBRs, as follows for PE-2. The configuration is similar on PE-3.

```
configure
    router
        ldp
            interface-parameters
                interface "int-PE-2-PE-3_1st" shutdown
                no interface "int-PE-2-PE-3_1st"
                interface "int-PE-2-PE-3_2nd" shutdown
                no interface "int-PE-2-PE-3_2nd"
            exit
```

EBGP Route Resolution to a Static Route

The static routes are already configured on both ASBRs and the eBGP session between the ASBRs is established.

Multi-hop EBGP labeled IPv4 route resolution to a static route needs to be enabled on PE-2 and PE-3 using the following command:

```
configure router bgp next-hop-resolution labeled-routes allow-static
```

On PE-2, the labeled IPv4 route for prefix 192.0.2.4/32 is now valid, best, and used, as follows:

```
*A:PE-2# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    192.0.2.1/32                                       100         None
      192.0.2.1                                          None        262141
      No As-Path
u*>i  192.0.2.4/32                                       None        None
      192.0.2.3                                          None        262141
      64500
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-2#
```

PE-1 learns the following BGP labeled IPv4 route for prefix 192.0.2.4/32 from PE-2:

```
*A:PE-1# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.4/32                                        100         None
      192.0.2.2                                           None        262140
      64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The routing table on PE-1 contains a BGP labeled IPv4 route to 192.0.2.4/32:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     01h23m40s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      01h23m30s  15
       192.168.12.2                                                 10
192.0.2.4/32                                  Remote  BGP_LABEL 00h06m39s  170
       192.0.2.2 (tunneled)                                         10
192.168.12.0/30                               Local   Local     01h23m40s  0
       int-PE-1-PE-2                                                0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The multi-hop eBGP session between PE-1 in AS 64496 and PE-4 in AS 64500 is established, as follows:

```
*A:PE-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
Def. Instance  64496      164    0 01h18m54s 1/1/1 (Lbl-IPv4)
                          163    0
```

```
192.0.2.4
Def. Instance  64500       57    0 00h01m25s 1/1/1 (VpnIPv4)
                           11    0

-------------------------------------------------------------------------------
*A:PE-1#
```

The loopback address defined in VPRN 1 on PE-4 (10.2.2.2/32) is advertised as VPN-IPv4 route in this multi-hop eBGP session on PE-1, as follows:

```
*A:PE-1# show router bgp routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64500:1:10.2.2.2/32                                None        None
      192.0.2.4                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-1#
```

The routing table for VPRN 1 on PE-1 includes the following BGP-VPN route to 10.2.2.2/32:

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.1.1.1/32                                   Local   Local     01h20m41s  0
       loopback                                                     0
10.2.2.2/32                                   Remote  BGP VPN   00h05m26s  170
       192.0.2.4 (tunneled:BGP)                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The results are similar on PE-4 and PE-1, and on ASBRs PE-3 and PE-2.

For directly connected ASBRs, inter-AS VPN model C can be configured using loopback addresses on the ASBRs without the need to enable MPLS between the ASBRs.

Conclusion

Most service providers use interface IP addresses in eBGP sessions, in which case this feature is not needed. However, some providers build directly connected eBGP sessions based on loopback interfaces. The system interface of the peer ASBR must be reachable and the labeled IPv4 routes for the remote AS PEs must be advertised to the local AS PEs. This advertisement can be achieved by configuring static routes on the ASBRs to the loopback address of their eBGP peer and enabling the eBGP route resolution to a static route. Enabling eBGP route resolution to a static route is much more secure than enabling MPLS on the interface to the peer ASBR of an untrusted ISP. However, when the ASBRs are distant and loopback addresses are used for the eBGP peering, MPLS must be enabled between the ASBRs.

IS-IS Link Bundling

This chapter provides information about IS-IS link bundling.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

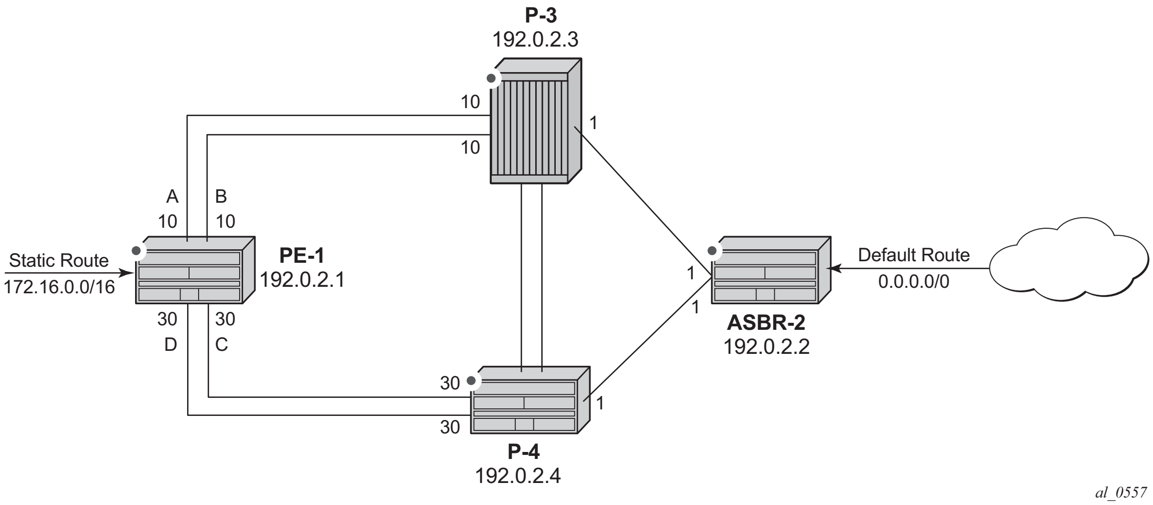
Applicability

The chapter was initially written for SR OS release 11.0.R6. However, the CLI in the current edition is based on release 14.0.R5.

Overview

Intermediate System to Intermediate System (IS-IS) Link Bundling allows for the grouping of a number of IS-IS interfaces into a single virtual link, called an IS-IS link group. It is used in conjunction with Equal Cost Multipath (ECMP) to dynamically change the metric of parallel IS-IS links if one or more links fail or suffer some sort of performance degradation.

Figure 159 **Link Bundle Schematic**



Consider the network in Figure 159, where a Provider Edge router PE-1 connects to a core network comprised of two Provider (P) routers and a single Autonomous System Border Router (ASBR). The links between PE-1 and P-3, and PE-1 and P-4 are 10 Gigabit Ethernet links. The links between ASBR-2 and P-3 and P-4 are both 100 Gigabit links. The link metrics are as shown in Figure 159.

In order to maximize the use of link bandwidth, ECMP is enabled on all routers and set to a value of 2, so that IP traffic flowing between PE1 and P-3, and PE-1 and P-4, is load balanced across the two links.

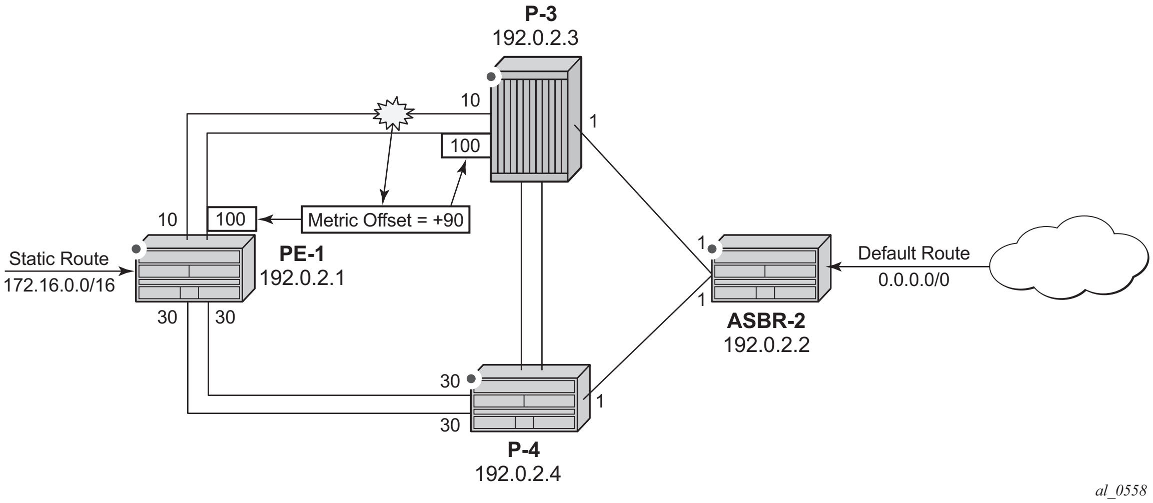
A default route is injected into the ASBR-2 router and re-distributed via a policy statement into IS-IS, so that traffic flowing from PE-1 to the ASBR is resolved by this route. Traffic flows between PE-1 and ASBR-2, using the path with the lowest IS-IS metric, via P-3 with a metric of 11. The second path PE-1 to ASBR-2 via P-4 has the same bandwidth, but a higher IS-IS metric of 31.

Traffic in the reverse direction flows toward a user subnet described by a static route configured on PE-1, which is redistributed into IS-IS using a policy statement. Once again, the shortest path between ASBR-2 and PE-1 is via P-3, so the bi-directional traffic flow is symmetric.

If one of the links between PE-1 and P-3 fails, traffic still flows via P-3, because the IS-IS metric is unchanged, but this now has less bandwidth than the second path via P-4. It is desirable to make use of the additional bandwidth of the second path, but this requires a change in metric. This can be achieved using IS-IS link bundling.

IS-IS link bundling allows for the creation of a group of IS-IS links, where the failure of a member link allows the metric of the remaining members of the link group to be increased by an offset value.

Figure 160 **Effect of Single Link Failure on Bundle Group**



Using Figure 160 as an example, the links between PE-1 and P-3 are included in a bundle group. To illustrate the change in metrics, a default static route is configured on ASBR-2 and re-distributed into IS-IS, and the path to this route is monitored at PE-1. Similarly, a static route to subnet 172.16.0.0/16 is configured on PE-1 and redistributed into IS-IS and viewed on ASBR-2.

Should one of the links between PE-1 and P-3 fail, the metric of the remaining members can be increased by an offset, for example 90, so that the metric of the remaining link becomes 10+90 = 100. The IS-IS metric between PE-1 and ASBR-2 via P-3 is now 101. The metric offset is applied to each remaining IS-IS interface individually and is advertised within the IS-IS database as the default cost in the TE-IS neighbors Type Length Variable (TLV).

The path between PE-1 and ASBR-2 via P-4 now has the lowest IS-IS metric, and any affected routers within the IS-IS area will try and re-route the traffic based on the new metric.

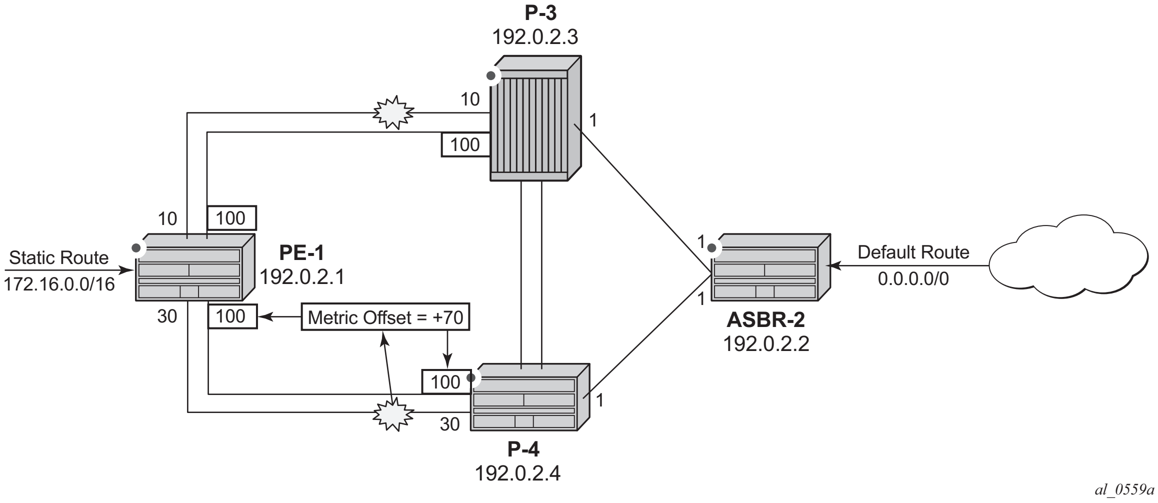
The fundamentals of this feature are:

- The treatment of all member links in a link group bundle as a single virtual interface.
- The increase in metric by a specific offset value of each remaining individual link within the group when a failure of one or more links occurs.

- The application of the offset occurs when the number of active links drops below a configured threshold.
- The offset is removed when the number of active links within the link group bundle reaches the configured reversion threshold.
- A link bundle is required on a router for the thresholds and offsets to apply.

Consider a second and subsequent failure where a link between PE-1 and P-4 also fails, so that there is only one active IS-IS interface between PE-1 and each of its neighboring P routers. This is shown in Figure 161.

Figure 161 **Double Link Failure**



In this case, the metric for the remaining link between PE-1 and P-4 can be increased by an offset value of +70 so that the IS-IS metric PE-1 to P-4 becomes 100, the same as that between PE-1 and P-3 when a link has failed.

PE-1 now sees two equal cost paths to the default route – one via P-3 and one via P-4, so there are still two 10Gigabit Ethernet links across which the traffic can be load shared.

This can be summarized using the following table, where ABCD are the 4 links as per Figure 159 and link status is Up (U) or Down (D).

Table 12 **Status of the links A, B, C, and D**

ABCD Status	A (metric,status)	B (metric,status)	C (metric,status)	D (metric,status)
UUUU	10 Transmit	10 Transmit	30 Idle	30 Idle
UDUU	100 Idle	Down	30 Transmit	30 Transmit

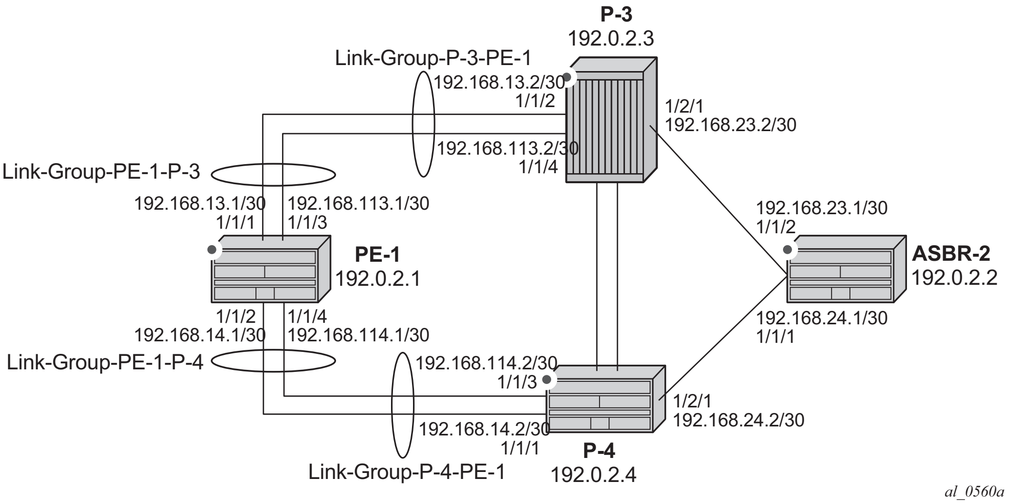
Table 12 **Status of the links A, B, C, and D (Continued)**

ABCD Status	A (metric,status)	B (metric,status)	C (metric,status)	D (metric,status)
UDUD	100 Transmit	Down	100 Transmit	Down
UUUD	10 Transmit	10 Transmit	100 Idle	Down

Configuration

The example topology is shown in Figure 162.

Figure 162 **Example Topology**



The PE-1 router configuration commands are as follows.

```
*A:PE-1# configure
    router
        interface "int-PE-1-P-3-1"
            address 192.168.13.1/30
            port 1/1/1
        exit
        interface "int-PE-1-P-3-2"
            address 192.168.113.1/30
            port 1/1/3
        exit
        interface "int-PE-1-P-4-1"
            address 192.168.14.1/30
            port 1/1/2
        exit
        interface "int-PE-1-P-4-2"
            address 192.168.114.1/30
            port 1/1/4
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        ecmp 2
```

The IP router configuration for the remaining routers can be derived from Figure 162.

The IS-IS network is a level 1 network.

The IS-IS configuration for PE-1, including the interface metrics is as follows:

```
*A:PE-1# configure
    router
        isis
            level-capability level-1
            area-id 49.0001
            advertise-passive-only
            level 1
                wide-metrics-only
            exit
            interface "system"
                passive
            exit
            interface "int-PE-1-P-3-1"
                interface-type point-to-point
                level 1
                    metric 10
                exit
            exit
            interface "int-PE-1-P-3-2"
                interface-type point-to-point
                level 1
                    metric 10
                exit
            exit
            interface "int-PE-1-P-4-1"
                interface-type point-to-point
                level 1
                    metric 30
                exit
            exit
            interface "int-PE-1-P-4-2"
                interface-type point-to-point
                level 1
                    metric 30
                exit
            exit
```

The IS-IS configuration for the remaining routers can be derived from Figure 162.

The following configuration is for the static route and export policy on ASBR-2. The configuration of the static route on PE-1 is similar.

```
*A:ASBR-2# configure router static-route-entry 0.0.0.0/0 black-hole no shutdown
```

```
*A:ASBR-2# configure router
        policy-options
            begin
            policy-statement "STATIC-ISIS"
                entry 10
                    from
                        protocol static
                    exit
                    to
                        level 1
                    exit
                    action accept
                        metric set igp
                    exit
                exit
            exit
            commit
        exit
```

```
*A:ASBR-2# configure router isis export "STATIC-ISIS"
```

Link Group Configuration

PE-1 contains 2 link groups. The first link group contains the IS-IS interfaces toward P-3. The second contains the interfaces toward P-4.

Each link-group is configured using a unique name, which is unique per router, and the IS-IS interface names are configured within the group as group members.

The metric offset value is the amount by which the IS-IS metric of active member links are increased when the number of links drops below a configured threshold.

The IS-IS link group configuration for PE-1 for the interfaces toward P-3 is as follows:

```
*A:PE-1# configure
    router
        isis
            link-group "Link-Group-PE-1-P-3"
                level 1
                    ipv4-unicast-metric-offset 90
                    member "int-PE-1-P-3-1"
                    member "int-PE-1-P-3-2"
                    revert-members 2
                    oper-members 2
                exit
            exit
        exit
```

Similarly, the IS-IS link group for PE-1 for the interfaces toward P-4 is:

```
*A:PE-1# configure
```

```
    router
        isis
            link-group "Link-Group-PE-1-P-4"
                level 1
                    ipv4-unicast-metric-offset 70
                    member "int-PE-1-P-4-1"
                    member "int-PE-1-P-4-2"
                    revert-members 2
                    oper-members 2
                exit
            exit
        exit
```

Within the link-group, two thresholds are configured:

- oper-members threshold
- revert-members threshold

If the number of operational links in the link-group drops below the oper-members value, then all interfaces associated with that IS-IS link group have their interface metric increased by the configured offset value. As a result, IS-IS then tries to reroute traffic over lower cost paths.

If the number of operational links in the link-group equals the revert-members threshold value, then all interfaces associated with that IS-IS link group have their interface metric decreased by the configured offset value.

In this configuration, there is a requirement to increase the metric of each interface within a link-group when a single interface fails. This means that the oper-members value is set to 2. In normal working circumstances, when both interfaces are active, the metric used is the configured interface metric. This means that the revert-members value must also be set to 2.

It is not possible to set the oper-members threshold to a value higher than that of the revert-members.

For completeness, the IS-IS configuration of each P-router is as follows.

P-3

```
*A:P-3# configure
    router
        isis
            level-capability level-1
            area-id 49.0001
            advertise-passive-only
            level 1
                wide-metrics-only
            exit
            interface "system"
                passive
            exit
            interface "int-P-3-PE-1-1"
                interface-type point-to-point
                level 1
                    metric 10
                exit
            exit
            interface "int-P-3-PE-1-2"
                interface-type point-to-point
                level 1
                    metric 10
                exit
            exit
            interface "int-P-3-ASBR-2"
                interface-type point-to-point
                level 1
                    metric 1
                exit
            exit
            link-group "Link-Group-P-3-PE-1"
                level 1
                    ipv4-unicast-metric-offset 90
                    member "int-P-3-PE-1-1"
                    member "int-P-3-PE-1-2"
                    revert-members 2
                    oper-members 2
                exit
            exit
        exit
```

P-4

```
*A:P-4# configure
    router
        isis
            level-capability level-1
            area-id 49.0001
            advertise-passive-only
            level 1
                wide-metrics-only
            exit
            interface "system"
                passive
            exit
            interface "int-P-4-PE-1-1"
                interface-type point-to-point
                level 1
                    metric 30
                exit
            exit
            interface "int-P-4-PE-1-2"
                interface-type point-to-point
                level 1
                    metric 30
                exit
            exit
            interface "int-P-4-ASBR-2"
                interface-type point-to-point
                level 1
```

```
                    metric 1
                exit
            exit
            link-group "Link-Group-P-4-PE-1"
                level 1
                    ipv4-unicast-metric-offset 70
                    member "int-P-4-PE-1-1"
                    member "int-P-4-PE-1-2"
                    revert-members 2
                    oper-members 2
                exit
            exit
        exit
```

An overview of all of the link groups can be shown using the following commands, in this case on node PE-1.

The link group status is as follows:

```
*A:PE-1# show router isis link-group-status
===============================================================================
Router Base ISIS Instance 0 Link-Group Status
===============================================================================
Link-group              Mbrs   Oper   Revert Active Level   State
                               Mbr     Mbr    Mbr
-------------------------------------------------------------------------------
Link-Group-PE-1-P-3     2      2      2      2      L1      normal
Link-Group-PE-1-P-4     2      2      2      2      L1      normal
===============================================================================
*A:PE-1#
```

The output for the individual link group members is as follows:

For "Link-Group-PE-1-P-3" at PE-1:

```
*A:PE-1# show router isis link-group-member-status level 1 "Link-Group-PE-1-P-3"
===============================================================================
Router Base ISIS Instance 0 Link-Group Member
===============================================================================
Link-group            I/F name                Level      State
-------------------------------------------------------------------------------
Link-Group-PE-1-P-3   int-PE-1-P-3-1          L1         Up
Link-Group-PE-1-P-3   int-PE-1-P-3-2          L1         Up
-------------------------------------------------------------------------------
Legend: BER = bitErrorRate
===============================================================================
*A:PE-1#
```

For "Link-Group-PE-1-P-4" at PE-1:

```
*A:PE-1# show router isis link-group-member-status level 1 "Link-Group-PE-1-P-4"
===============================================================================
Router Base ISIS Instance 0 Link-Group Member
===============================================================================
Link-group            I/F name                Level      State
```

```
-------------------------------------------------------------------------------
Link-Group-PE-1-P-4   int-PE-1-P-4-1          L1         Up
Link-Group-PE-1-P-4   int-PE-1-P-4-2          L1         Up
-------------------------------------------------------------------------------
Legend: BER = bitErrorRate
===============================================================================
*A:PE-1#
```

For P-3, the link group status is as follows:

```
*A:P-3# show router isis link-group-status
===============================================================================
Router Base ISIS Instance 0 Link-Group Status
===============================================================================
Link-group              Mbrs   Oper   Revert Active Level   State
                               Mbr     Mbr    Mbr
-------------------------------------------------------------------------------
Link-Group-P-3-PE-1     2      2      2      2      L1      normal
===============================================================================
*A:P-3#
```

For P-3, the link group member status is as follows:

```
*A:P-3# show router isis link-group-member-status level 1 "Link-Group-P-3-PE-1"
===============================================================================
Router Base ISIS Instance 0 Link-Group Member
===============================================================================
Link-group            I/F name                Level      State
-------------------------------------------------------------------------------
Link-Group-P-3-PE-1   int-P-3-PE-1-1          L1         Up
Link-Group-P-3-PE-1   int-P-3-PE-1-2          L1         Up
-------------------------------------------------------------------------------
Legend: BER = bitErrorRate
===============================================================================
*A:P-3#
```

Routing Table PE-1

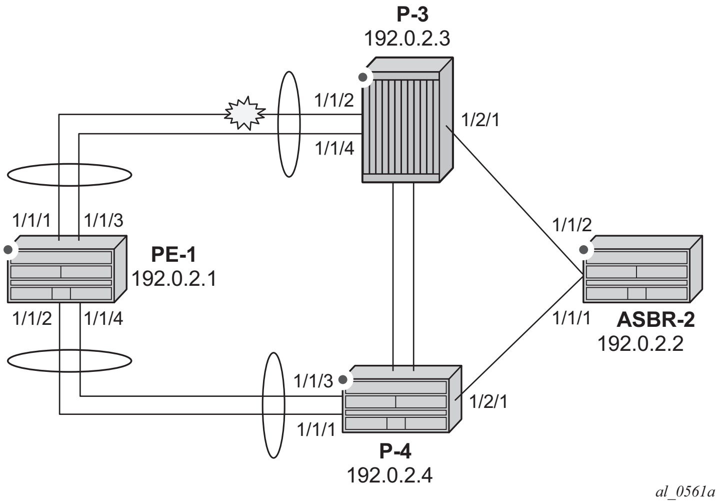
In a normal working state, the routing table for PE-1 contains the default route for forwarding traffic toward ASBR-2. Because ECMP is set to a value of 2, two entries are available with next-hops pointing toward P-3, as follows. The metric for each path is 11.

```
*A:PE-1# show router route-table 0.0.0.0/0

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
0.0.0.0/0                                     Remote  ISIS      00h02m27s  15
       192.168.13.2                                                 11
0.0.0.0/0                                     Remote  ISIS      00h02m27s  15
       192.168.113.2                                                11
```

```
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

Failure of link member PE-1 to P-3

Figure 163 **Link Failure**



One of the links between PE-1 and P-3 is put into a failed state by shutting down port 1/1/2 on P-3, as per Figure 163.

```
*A:P-3# configure port 1/1/2 shutdown
```

The route-table on PE-1 shows that the metric for the default route prefix, 0.0.0.0/0, has increased from 11 to 31, and the next-hops are now interface addresses on P-4, as follows:

```
A:PE-1# show router route-table 0.0.0.0/0
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
0.0.0.0/0                                     Remote  ISIS      00h01m14s  15
       192.168.14.2                                                 31
0.0.0.0/0                                     Remote  ISIS      00h01m14s  15
       192.168.114.2                                                31
-------------------------------------------------------------------------------
No. of Routes: 2
```

```
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The link-group status shows that the number of active members has fallen below the oper-members threshold and as a result, the metric offset has been applied.

```
*A:PE-1# show router isis link-group-status
===============================================================================
Router Base ISIS Instance 0 Link-Group Status
===============================================================================
Link-group              Mbrs   Oper   Revert Active Level   State
                               Mbr     Mbr    Mbr
-------------------------------------------------------------------------------
Link-Group-PE-1-P-3     2      2      2      1      L1      Offset-Applied
Link-Group-PE-1-P-4     2      2      2      2      L1      normal
===============================================================================
*A:PE-1#
```

Finally, the status of an individual link group member is as follows:

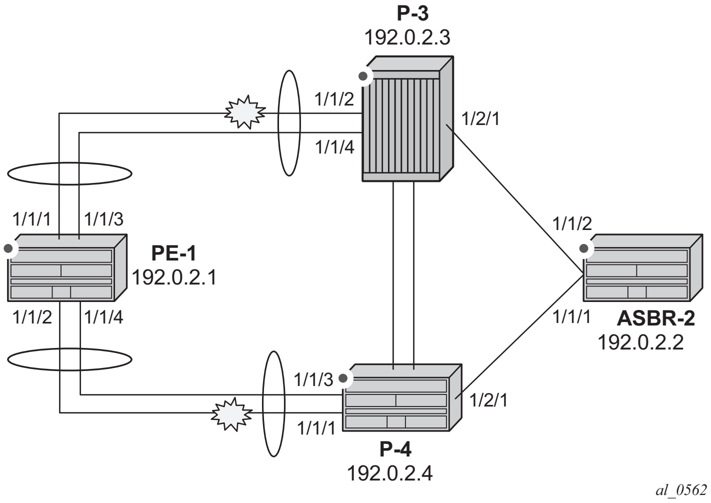
```
*A:PE-1# show router isis link-group-member-status "Link-Group-PE-1-P-3"
===============================================================================
Router Base ISIS Instance 0 Link-Group Member
===============================================================================
Link-group            I/F name                Level      State
-------------------------------------------------------------------------------
Link-Group-PE-1-P-3   int-PE-1-P-3-1          L1         If-Down
Link-Group-PE-1-P-3   int-PE-1-P-3-2          L1         Up
-------------------------------------------------------------------------------
Legend: BER = bitErrorRate
===============================================================================
*A:PE-1#
```

By examining the IS-IS database on PE-1, it can be seen that the link metric (TE-IS neighbor) toward P-3 has a metric of 100, comprised of the original metric of 10 plus the offset of 90.

```
*A:PE-1# show router isis database PE-1 detail
===============================================================================
Router Base ISIS Instance 0 Database
===============================================================================

Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : PE-1.00-00                                  Level     : L1
Sequence  : 0x9                    Checksum  : 0x9312   Lifetime  : 1016
Version   : 1                      Pkt Type  : 18       Pkt Ver   : 1
Attributes: L1                     Max Area  : 3
SysID Len : 6                      Used Len  : 156      Alloc Len : 1492

TLVs :
  Area Addresses:
    Area Address : (3) 49.0001
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-1
  Router ID   :
    Router ID   : 192.0.2.1
  I/F Addresses :
    I/F Address   : 192.0.2.1
    I/F Address   : 192.168.13.1
    I/F Address   : 192.168.14.1
    I/F Address   : 192.168.113.1
    I/F Address   : 192.168.114.1
  TE IS Nbrs   :
    Nbr   : P-3.00
    Default Metric  : 100
    Sub TLV Len     : 12
    IF Addr   : 192.168.113.1
    Nbr IP    : 192.168.113.2
  TE IS Nbrs   :
    Nbr   : P-4.00
    Default Metric  : 30
    Sub TLV Len     : 12
    IF Addr   : 192.168.14.1
    Nbr IP    : 192.168.14.2
  TE IS Nbrs   :
    Nbr   : P-4.00
    Default Metric  : 30
    Sub TLV Len     : 12
    IF Addr   : 192.168.114.1
    Nbr IP    : 192.168.114.2
  TE IP Reach   :
    Default Metric  : 0
    Control Info:    , prefLen 16
    Prefix   : 172.16.0.0
    Default Metric  : 0
    Control Info:    , prefLen 32
    Prefix   : 192.0.2.1

Level (1) LSP Count : 1

Displaying Level 2 database
-------------------------------------------------------------------------------
Level (2) LSP Count : 0
===============================================================================
*A:PE-1#
```

Failure of link member PE-1 to P-4:

Figure 164 **Second Link Failure**



If a link between PE-1 and P-4 now fails, simulated by shutting down port 1/1/1 on P-4, then the metric offset is applied to the link groups on PE-1 and P-4 as the number of active links has dropped below the oper-members threshold for the link groups Link-Group-PE-1-P-4 on PE-1 and Link-Group-P-4-PE-1on P-4.

```
*A:P-4# configure port 1/1/1 shutdown
```

The routing table for PE-1 now shows that there are still two equal cost paths for the default route prefix advertised by ASBR-2, as follows:

```
*A:PE-1# show router route-table 0.0.0.0/0

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
0.0.0.0/0                                     Remote  ISIS      00h01m16s  15
       192.168.113.2                                                101
0.0.0.0/0                                     Remote  ISIS      00h01m16s  15
       192.168.114.2                                                101
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

The metric for each routing table entry is 101, comprising of a cost of 100 for the PE-1 to P router link, where the link-group offset has been applied, and the cost of 1 for the P router to ASBR-2 router link.

By examining the IS-IS database on the PE-1 router, the updated metric for the link to neighbors P-3 and P-4 can be seen with the offset applied. These are seen in the "TE-IS Nbrs" TLV in the following output.

```
*A:PE-1# show router isis database PE-1 detail
===============================================================================
Router Base ISIS Instance 0 Database
===============================================================================
Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : PE-1.00-00                                Level     : L1
Sequence  : 0xa                    Checksum  : 0x5ebc   Lifetime  : 1105
Version   : 1                      Pkt Type  : 18       Pkt Ver   : 1
Attributes: L1                     Max Area  : 3
SysID Len : 6                      Used Len  : 131      Alloc Len : 1492

TLVs :
  Area Addresses:
    Area Address : (3) 49.0001
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-1
  Router ID   :
    Router ID   : 192.0.2.1
  I/F Addresses :
    I/F Address   : 192.0.2.1
    I/F Address   : 192.168.13.1
    I/F Address   : 192.168.14.1
    I/F Address   : 192.168.113.1
    I/F Address   : 192.168.114.1
  TE IS Nbrs   :
    Nbr   : P-3.00
    Default Metric  : 100
    Sub TLV Len     : 12
    IF Addr   : 192.168.113.1
    Nbr IP    : 192.168.113.2
  TE IS Nbrs   :
    Nbr   : P-4.00
    Default Metric  : 100
    Sub TLV Len     : 12
    IF Addr   : 192.168.114.1
    Nbr IP    : 192.168.114.2
  TE IP Reach   :
    Default Metric  : 0
    Control Info:    , prefLen 16
    Prefix   : 172.16.0.0
    Default Metric  : 0
    Control Info:    , prefLen 32
    Prefix   : 192.0.2.1

Level (1) LSP Count : 1

Displaying Level 2 database
```

```
-------------------------------------------------------------------------------
Level (2) LSP Count : 0
===============================================================================
*A:PE-1#
```

Conclusion

IS-IS link bundling allows service providers to configure multiple IS-IS interfaces as a single link group for ECMP purposes and allow link metric increases if an interface within the bundle group fails. This example provides the configuration for IS-IS link bundling, together with the associated commands and outputs which can be used for verifying and troubleshooting.

Next-Hop Resolution for Labeled BGP Routes

This chapter describes Next-Hop Resolution for Labeled BGP Routes. Topics include:

Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R7. In SR OS Release 15.0.R1, the next-hop (NH) resolution for labeled BGP routes was made consistent across different labeled route families—such as labeled IPv4, labeled IPv6, VPN-IPv4, VPN-IPv6—regardless of the type of the BGP peering (eBGP or iBGP).

Overview

BGP routes with the VPN-IPv4, VPN-IPv6, labeled IPv4, and labeled IPv6 address families are BGP routes whose Network Layer Reachability Information (NLRI) contains an MPLS label that is mapped to the route. BGP advertises labels that subsequently will be used in the data plane for MPLS forwarding. BGP labeled routes are fundamental to IP VPN services, 6PE services, inter-AS connectivity, and seamless MPLS network segmentation.

When a BGP speaker receives a BGP labeled route, it has the following options for resolving the NH of the route:

- It can resolve the NH to an MPLS tunnel, such as an LDP or RSVP tunnel. In this case, the router pushes a transport label on top of the BGP label and allows the BGP labeled packet to be transported to the NH router over a set of intermediate routers that lack context for forwarding using the BGP label.
- It can resolve the NH to a local interface if the NH is an address on a local subnet. No additional labels need to be pushed onto the top of the label stack.

- It can resolve the NH using a static route and no additional label needs to be pushed. BGP NH resolution using a static route is useful in the following cases:
    - The static route has a blackhole NH in an intentional Remotely Triggered Blackhole (RTBH) scenario. Blackholed static routes are used for BGP NH resolution even when the configuration does not allow BGP NH resolution using static routes.
    - The static route has a NH address of a loopback interface of a directly connected peer. By default, this option is disabled.
- It can resolve the NH using the Longest Prefix Match (LPM) in the route table with static routes, OSPF, IS-IS, and RIP routes. This is applicable for route reflectors (RRs) that are not in the data path, so they do not need to have tunnels. By default, this option is disabled.

NH resolution of BGP routes using tunnels is the same for eBGP and iBGP routes, and for VPN IP routes and label-unicast routes. The common NH resolution logic uses the following routes in order of preference:

1. Local or direct routes
2. Non-default static routes
    - Blackholed static routes
    - Non-blackholed non-default static routes, if allowed
3. RTM routes (including static, OSPF, IS-IS, and RIP), if allowed—only for RRs
    - When enabled, no routes are installed in the Forwarding Information Base (FIB) and no tunnels can be used.
4. Tunnels

Step 1. NH resolution using a local (interface) route

If possible, the BGP NH is resolved to a local interface route.

If the BGP NH is an IPv4-mapped IPv6 address in ::ffff:a.b.c.d format, the system first tries to find a local route matching the IPv6 address. When no match is found, the system tries to find a local route matching the extracted IPv4 address a.b.c.d.

Step 2. NH resolution using a non-default static route

If the BGP NH is an IPv4 address, the system looks for the non-default IPv4 static route that is the LPM of the address.

- If the LPM static route is blackholed, this static route is used, regardless of the **allow-static** command configuration.
- If the LPM static route is not blackholed, the static route is only used when the **allow-static** command is configured.

If the BGP NH is an IPv4-mapped IPv6 address in the ::ffff:a.b.c.d format, the system first tries to find the non-default static route that is the LPM of the full IPv6 address.

If no matching IPv6 static route is found, the system tries to find the non-default IPv4 route that is the LPM of the extracted IPv4 address a.b.c.d.

Step 3. NH resolution using any type of route in RTM—only on RR

This is only applicable for RRs that are not in the data path and configured with the **rr-use-route-table** and **disable-route-table-install** commands. The considered routes in the Route Table Manager (RTM) can be static, OSPF, IS-IS, or RIP.

If the BGP NH is an IPv4 address, the system searches the IPv4 RTM route that is the LPM of the address.

If the BGP NH is an IPv4-mapped IPv6 address in ::ffff:a.b.c.d format, the system first searches for the IPv6 route that is the LPM of the full IPv6 address. If no match is found, the system searches for an RTM route matching the extracted IPv4 address a.b.c.d.

Step 4. NH resolution using a tunnel in TTM

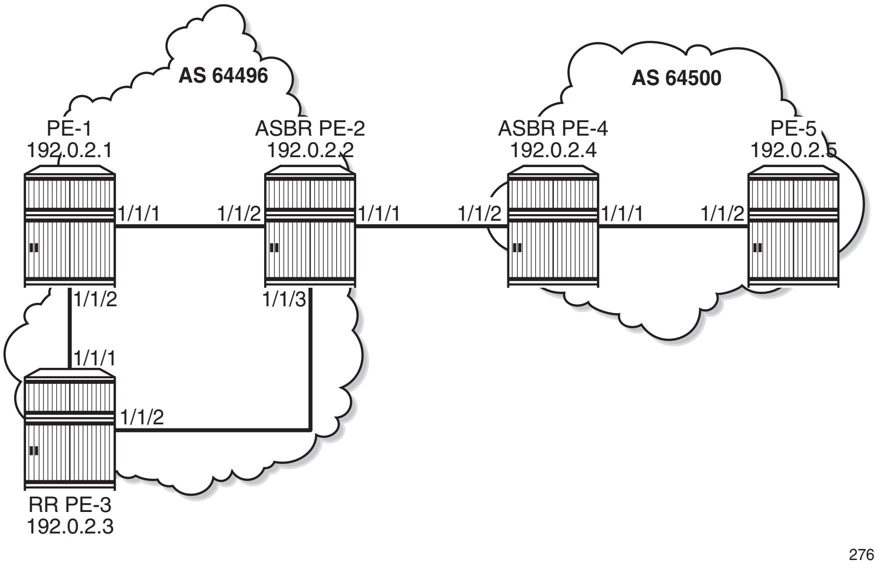
If the BGP NH is an IPv4 address, the Tunnel Table Manager (TTM) selects the tunnel table entry that matches the address prefix with the lowest preference and allowed by the applicable resolution filter. If the preference is the same, the tunnel table entry with the best metric is chosen, and so on.

If the BGP NH is an IPv4-mapped IPv6 address in ::ffff:a.b.c.d format, the system searches the most preferred TTM tunnel matching the extracted IPv4 address a.b.c.d that is allowed by the applicable resolution filter.

Configuration

Figure 165 shows the example topology with three routers in AS 64496 and two routers in AS 64500.

Figure 165 **Example Topology**



The initial configuration includes the following:

- Cards, MDAs, ports
- Router interfaces between the PEs
- IS-IS as IGP between the PEs within an AS, not between ASBRs PE-2 and PE-4
- LDP between the PE-1 and PE-2 in AS 64496 (not to the RR PE-3) and between PE-4 and PE-5 in AS 64500

The following scenarios will be configured in the following sections:

- NH resolution for labeled IPv4 routes
- NH resolution for iBGP VPN-IPv4/v6 routes
- NH resolution for inter-AS VPRN model B
- NH resolution for inter-AS VPRN model C

NH Resolution for Labeled IPv4 Routes

In the NH Resolution for Inter-AS VPRN Model C section, inter-AS VPRNs will be configured, as described in the *VPRN Inter-AS VPRN Model C* chapter. Within each AS, the PEs advertise their system addresses (192.0.2.x) as labeled IPv4 routes. The configuration of the export policy is as follows:

```
configure
    router
        policy-options
            begin
            prefix-list "PE-sys4"
                prefix 192.0.2.0/28 prefix-length-range 32-32
            exit
            policy-statement "export-bgp"
                entry 10
                    from
                        protocol direct
                        prefix-list "PE-sys4"
                    exit
                    to
                        protocol bgp-label
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

Within each AS, BGP group "iBGP" is configured for the VPN-IPv4, VPN-IPv6, and label-IPv4 address families. In AS 64496, PE-3 is configured as RR, as follows:

```
configure
    router
        bgp
            split-horizon
            group "iBGP"
                family vpn-ipv4 vpn-ipv6 label-ipv4
                cluster 192.0.2.3
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
        exit
```

Between the Autonomous System Border Routers (ASBRs) PE-2 and PE-4, BGP is configured for the label-IPv4 address family only. The initial configuration for the eBGP peering uses the interface address of the remote ASBR (such as 192.168.24.2), which is the standard way for eBGP peering between ASBRs. However, for demonstration purposes, loopback addresses will be configured later.

The BGP labeled routes for the system IP addresses are not used within an AS because IGP routes are preferred by the RTM, so they are inactive. However, BGP exports these inactive routes to the ASBR peer in the remote AS (**advertise-inactive**) where these routes will be used. The initial BGP configuration on PE-2 is as follows:

```
configure
```

```
    router
        bgp
            split-horizon
            group "eBGP4_local"
                neighbor 192.168.24.2
                    family label-ipv4
                    peer-as 64500
                    advertise-inactive
                exit
            exit
            group "iBGP"
                peer-as 64496
                neighbor 192.0.2.3
                    family label-ipv4
                exit
            exit
```

The default BGP NH resolution does not allow static routes and the only transport tunnel type that can be used for labeled IPv4 routes is LDP, as follows:

```
*A:PE-2# configure router bgp next-hop-resolution
*A:PE-2>config>router>bgp>next-hop-res# info detail
----------------------------------------------
            no use-bgp-routes
            no policy
            no weighted-ecmp
            shortcut-tunnel
                family ipv4
                    resolution-filter
                        no ldp
                        no rsvp
                        no bgp
                        no sr-isis
                        no sr-ospf
                        no sr-te
                    exit
                    no disallow-igp
                    resolution disabled
                exit
            exit
            labeled-routes
                no allow-static
                no rr-use-route-table
                transport-tunnel
                    family vpn
                        resolution-filter
                            ldp
                            no rsvp
                            no sr-isis
                            no sr-ospf
                            bgp
                            no sr-te
                            no udp
                        exit
                        resolution filter
                    exit
                    family label-ipv4
                        resolution-filter
```

```
                            ldp
                            no rsvp
                            no sr-isis
                            no sr-ospf
                            no bgp
                            no sr-te
                            no udp
                        exit
                        resolution filter
                    exit
                    family label-ipv6
                        resolution-filter
                            ldp
                            no rsvp
                            no sr-isis
                            no sr-ospf
                            no bgp
                            no sr-te
                            no udp
                        exit
                        resolution filter
                    exit
                exit
            exit
----------------------------------------------
```

Labeled IPv4 BGP NH Resolved to Local Route

The route table on PE-2 shows that the route to 192.0.2.5 on PE-5 is a BGP labeled IPv4 route with NH 192.168.24.2:

```
*A:PE-2# show router route-table
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
---snip---
192.0.2.5/32                                  Remote  BGP_LABEL 00h02m57s  170
       192.168.24.2                                                 0
---snip---
```

To verify that BGP NH resolution prefers local routes over static routes (if **allow-static- routes** is enabled), the following is configured on the ASBRs. For the following static routes between PE-2 and PE-4, additional loopback addresses are configured and a static route to the loopback address on the eBGP peer. The configuration on ASBR PE-2 is as follows:

```
configure
    router
        interface "loopback"
            address 10.0.0.2/32
            loopback
        exit
        static-route-entry 10.0.0.4/32
            next-hop 192.168.24.2
                no shutdown
            exit
        exit
```

On PE-2, the following additional eBGP group for the label IPv4 address family is configured and the BGP NH resolution for labeled routes is configured to allow static routes. The eBGP peer is only one hop away, so a **multihop** command is not required.

```
configure
    router
        bgp
            next-hop-resolution
                labeled-routes
                    allow-static
                exit
            exit
            group "eBGP4_static"
                neighbor 10.0.0.4
                    family label-ipv4
                    local-address 10.0.0.2
                    peer-as 64500
                    advertise-inactive
                exit
            exit
```

Another static route is configured to the system IP address of the eBGP peer with preference 25 to ensure that this static route is not preferred over the preceding static route with default preference 5. LDP is enabled on the interface between the ASBRs, such as "int-PE-2-PE-4" on PE-2. This makes it possible to resolve the BGP NH to an LDP tunnel. Also, an additional BGP group is configured for the labeled IPv4 address family to the system IP address of the eBGP peer, such as 192.0.2.4. The configuration on PE-2 is as follows:

```
configure
    router
        static-route-entry 192.0.2.4/32
            next-hop 192.168.24.2
                preference 25
                no shutdown
            exit
        exit
        ldp
            interface-parameters
                interface "int-PE-2-PE-4" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
        exit
        bgp
            group "eBGP4_tunnel"
                neighbor 192.0.2.4
                    family label-ipv4
                    peer-as 64500
                    advertise-inactive
                exit
            exit
```

This additional configuration does not result in a BGP NH resolution to an LDP tunnel, because the destination can also be reached via a static route, which is preferred. In the Labeled IPv4 BGP NH Resolved to Tunneled Route section, the configuration will be modified to exclude static routes from the NH resolution.

The following FIB on PE-2 shows that a labeled BGP route with resolved NH 192.168.24.2 is used for prefix 192.0.2.5/32. The BGP NH is not resolved to a tunnel.

```
*A:PE-2# show router fib 1 192.0.2.5/32
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
192.0.2.5/32                                                BGP_LABEL
  192.168.24.2 (int-PE-2-PE-4)
-------------------------------------------------------------------------------
Total Entries : 1
```

PE-2 has three labeled IPv4 BGP routes for prefix 192.0.2.5/32: the first route with local NH 192.168.24.2 (which is best and used), the second route with NH 10.0.0.4/32 (which can be reached via a static route), and the third route with NH 192.0.2.4 (which can be reached via a less preferred static route), as follows:

```
*A:PE-2# show router bgp routes 192.0.2.5/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       None        None
      192.168.24.2                                       None        262140
      64500
*i    192.0.2.5/32                                       None        None
```

```
      10.0.0.4                                           None        262140
      64500
*i    192.0.2.5/32                                       None        None
      192.0.2.4                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 3
```

Table 13 shows the default preferences in a route table. These preferences are configurable, except for the direct attached routes, which always have preference 0.

Table 13 **Default Preferences in Route Table**

Route Type	Preference
Direct Attached	0
Static	5
OSPF Internal	10
IS-IS Level 1 Internal	15
IS-IS Level 2 Internal	18
RIP	100
OSPF External	150
IS-IS Level 1 External	160
IS-IS Level 2 External	165
BGP	170

The following shows the BGP NHs with the resolving prefix and the resolved NH. On PE-2, all three NHs of the labeled IPv4 routes for prefix 192.0.2.5/32 have resolved NH 192.168.24.2. NH 192.168.24.2 has owner local and preference 0; NH 10.0.0.4 has owner static and default preference 5; NH 192.0.2.4 has owner static and preference 25 by configuration.

```
*A:PE-2# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
10.0.0.4                                                    5        STATIC
   10.0.0.4/32                                              Y        1
   192.168.24.2                                                          0
192.0.2.3                                                   15       ISIS
   192.0.2.3/32                                             Y        10
   192.168.23.2                                                      0
192.0.2.4                                                   25       STATIC
   192.0.2.4/32                                             Y        1
   192.168.24.2                                                      0
192.168.24.2                                                0        LOCAL
   192.168.24.0/30                                          Y        0
   192.168.24.2                                                      0
-------------------------------------------------------------------------------
Next Hops : 4
```

Labeled IPv4 BGP NH Resolved to Non-Default Static Route

When the BGP group "eBGP4_local" is disabled, the BGP NH can no longer be resolved to a local route. On the ASBRs PE-2 and PE-4, the following command disables the BGP group:

```
configure router bgp group "eBGP4_local" shutdown
```

The FIB on PE-2 shows that the route to prefix 192.0.2.5/32 is a labeled BGP route with resolved NH 192.168.24.2. This looks identical to the preceding output for the FIB when the BGP NH could be resolved to a local route, but in this case, the BGP NH is resolved to a non-default static route, as will be shown later.

```
*A:PE-2# show router fib 1 192.0.2.5/32

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
192.0.2.5/32                                                BGP_LABEL
  192.168.24.2 (int-PE-2-PE-4)
-------------------------------------------------------------------------------
Total Entries : 1
```

PE-2 now has only two valid labeled IPv4 BGP routes instead of three: the best and used route has NH 10.0.0.4 and the less preferred route has NH 192.0.2.4, as follows:

```
*A:PE-2# show router bgp routes 192.0.2.5/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       None        None
      10.0.0.4                                           None        262140
      64500
*i    192.0.2.5/32                                       None        None
      192.0.2.4                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

On PE-2, NH 10.0.0.4 and NH 192.0.2.4 are both resolved to NH 192.168.24.2, as follows. NH 10.0.0.4 has preference 5, which is better than the configured preference 25 for NH 192.0.2.4.

```
*A:PE-2# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
10.0.0.4                                                    5        STATIC
   10.0.0.4/32                                              Y        1
   192.168.24.2                                                      0
192.0.2.3                                                   15       ISIS
   192.0.2.3/32                                             Y        10
   192.168.23.2                                                      0
192.0.2.4                                                   25       STATIC
   192.0.2.4/32                                             Y        1
   192.168.24.2                                                      0
-------------------------------------------------------------------------------
Next Hops : 3
```

When the preferred static route with NH 10.0.0.4 becomes unavailable, the other static route takes over. The following command disables the static route with NH 10.0.0.4 on PE-2.

```
*A:PE-2# configure router static-route-entry 10.0.0.4/32 next-hop 192.168.24.2
shutdown
```

The FIB on PE-2 shows a labeled BGP route with resolved NH 192.168.24.2. Again, this FIB entry looks identical. The BGP NH is not resolved to a tunnel.

```
*A:PE-2# show router fib 1 192.0.2.5/32
```

```
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
192.0.2.5/32                                                BGP_LABEL
  192.168.24.2 (int-PE-2-PE-4)
-------------------------------------------------------------------------------
Total Entries : 1
```

On PE-2, the best and used labeled BGP route for prefix 192.0.2.5/32 has NH 192.0.2.4, as follows. The BGP route for prefix 192.0.2.5/32 with NH 10.0.0.4 is not valid, because the static route to 10.0.0.4/32 is disabled.

```
*A:PE-2# show router bgp routes 192.0.2.5/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       None        None
      192.0.2.4                                          None        262140
      64500
i     192.0.2.5/32                                       None        None
      10.0.0.4                                           None        262140
      64500
-------------------------------------------------------------------------------
Routes : 1
```

On PE-2, NH 10.0.0.4 is not resolved, because the static route is disabled. NH 192.0.2.4 has resolved NH 192.168.24.2, as follows:

```
*A:PE-2# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================

===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
10.0.0.4                                                    -        -
```

```
   Unresolved                                                        -
   --                                                                -
192.0.2.3                                                  15        ISIS
   192.0.2.3/32                                            Y         10
   192.168.23.2                                                      0
192.0.2.4                                                  25        STATIC
   192.0.2.4/32                                            Y         1
   192.168.24.2                                                      0
-------------------------------------------------------------------------------
Next Hops : 3
```

The configuration on ASBR PE-2 is restored as follows and the BGP NH will be resolved to the static route to 10.0.0.4 again:

```
*A:PE-2# configure router static-route-entry 10.0.0.4/32 next-hop 192.168.24.2 no
shutdown
```

Labeled IPv4 BGP NH Resolved to Tunneled Route

When the system does not allow BGH NH resolution to static routes, the tunneled route is selected. The following command configures BGP NH resolution for labeled routes to its default setting no **allow-static**:

```
*A:PE-2# configure router bgp next-hop-resolution labeled-routes no allow-static
```

On PE-2, the route table shows that the BGP labeled IPv4 route to 192.0.2.5/32 has NH 192.0.2.4, which is resolved to a tunnel, as follows:

```
*A:PE-2# show router route-table 192.0.2.5/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.5/32                                  Remote  BGP_LABEL 00h01m01s  170
       192.0.2.4 (tunneled)                                         1
-------------------------------------------------------------------------------
No. of Routes: 1
```

On PE-2, the following FIB shows that the BGP labeled route uses an LDP tunnel to the NH 192.0.2.4:

```
*A:PE-2# show router fib 1 192.0.2.5/32

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
```

```
-------------------------------------------------------------------------------
192.0.2.5/32                                                BGP_LABEL
  192.0.2.4 (Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 1
```

PE-2 has two labeled BGP routes to prefix 192.0.2.5/32: the route with NH 10.0.0.4 is not valid because it requires a static route, which is not allowed for BGP NH resolution; the best and used route has NH 192.0.2.4 (which is the NH that is reached by an LDP tunnel), as follows:

```
*A:PE-2# show router bgp routes 192.0.2.5/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                                  LocalPref   MED
      Nexthop (Router)                                         Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                             None        None
      192.0.2.4                                                None        262140
      64500
i     192.0.2.5/32                                             None        None
      10.0.0.4                                                 None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

On PE-2, the following BGP NH list shows that NH 192.0.2.4 is resolved using a static route with NH 192.168.24.2:

```
*A:PE-2# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================

===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
10.0.0.4                                                    5        STATIC
   10.0.0.4/32                                              Y        1
   192.168.24.2                                                      0
192.0.2.3                                                   15       ISIS
   192.0.2.3/32                                             Y        10
   192.168.23.2                                                         0
192.0.2.4                                                   25       STATIC
   192.0.2.4/32                                             Y        1
   192.168.24.2                                                      0
-------------------------------------------------------------------------------
Next Hops : 3
```

The configuration on the ASBRs is modified as follows and the BGP NH is resolved to the local route, to 192.168.24.2 again. Local routes prevail over tunneled routes.

```
configure router bgp group "eBGP4_local" no shutdown
```

Labeled IPv4 BGP NH Resolved to RTM Route on RR

RR PE-3 is not in the data path and **next-hop-self** is disabled, which is the default setting. PE-3 does not have LDP tunnels to PE-1 and PE-2, so BGP NH resolution to RTM routes needs to be allowed, by enabling **rr-use-route-table**. The following error is raised when attempting to configure **rr-use-route-table** without **disable-route-table-install**:

```
*A:PE-3# configure router bgp next-hop-resolution labeled-routes rr-use-route-table
INFO: BGP #1001 Configuration failed because of inconsistent values - BGP [VR 1] route-
table-for-label-routes cannot be set unless disable-route-table-install is set!
```

The option **disable-route-table-install** allows an RR to reflect routes without installing them in its FIB. This way, an RR can reflect more routes than it can install in its FIB.

The following configuration on RR PE-3 allows the use of the route table for labeled routes:

```
configure
    router
        bgp
            disable-route-table-install
            split-horizon
            next-hop-resolution
                labeled-routes
                    rr-use-route-table
                exit
            exit
            group "iBGP"
                family vpn-ipv4 vpn-ipv6 label-ipv4
                cluster 192.0.2.3
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
            exit
```

The following command shows that the labeled BGP route for 192.0.2.5/32 is not used in the RR. This is because the route is not installed in the FIB of the RR, which is allowed, because the RR is not in the data path and NHS is disabled.

```
*A:PE-3# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*i    192.0.2.1/32                                       100         None
      192.0.2.1                                          None        262141
      No As-Path
*>i   192.0.2.5/32                                       100         None
      192.0.2.2                                          None        262137
      64500
-------------------------------------------------------------------------------
Routes : 2
```

The following labeled BGP route has NH 192.0.2.2, which is resolved to an IS-IS route:

```
*A:PE-3# show router bgp next-hop 192.0.2.2
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================

===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
192.0.2.2                                                   15       ISIS
   192.0.2.2/32                                             N        10
   192.168.23.1                                                      0
-------------------------------------------------------------------------------
Next Hops : 1
```

RR PE-3 advertises this labeled BGP route to PE-1, which installs the route in its FIB, so it will be used, as follows:

```
*A:PE-1# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       100         None
      192.0.2.2                                          None        262137
      64500
-------------------------------------------------------------------------------
Routes : 1
```

The tunnel table on PE-1 has a BGP tunnel to 192.0.2.5 with NH 192.0.2.2 and an LDP tunnel to 192.0.2.2 with NH 192.168.12.2, as follows:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.5/32      bgp       MPLS  262145    12       192.0.2.2      1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

On PE-1, the BGP NH for route 192.0.2.5/32 is resolved to an LDP tunnel to PE-2, as follows:

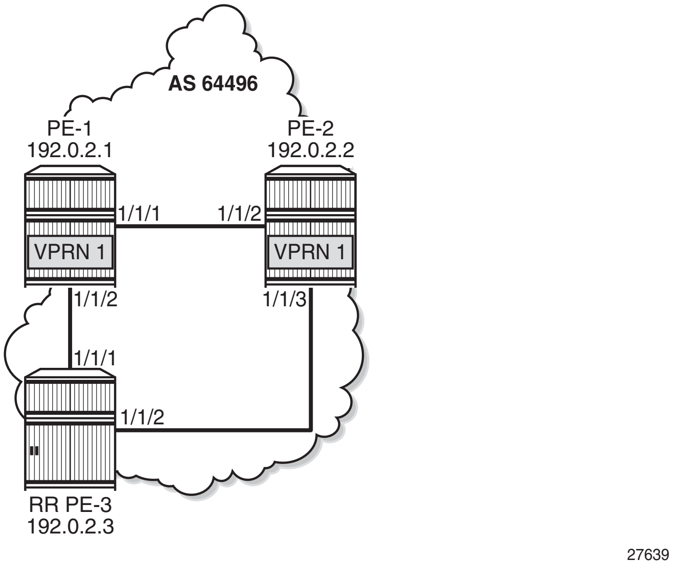
```
*A:PE-1# show router fp-tunnel-table 1

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol           Tunnel-ID
  Lbl
    NextHop                                                     Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.2/32                                 LDP                  -
  262143
  192.168.12.2                                                1/1/1:100
192.0.2.5/32                                 BGP                  -
  262135
  192.0.2.2                                                   LDP
```

```
-------------------------------------------------------------------------------
Total Entries : 2
```

NH Resolution for iBGP VPN-IPv4/v6 Routes

Figure 166 shows that VPRN 1 is configured on PE-1 and PE-2 in AS 64496.

Figure 166 **VPRN 1 in AS 64496**



On both PE-1 and PE-2, the VPN-IPv4 and VPN-IPv6 address families are configured in group "iBGP", as follows:

```
configure
    router
        bgp
            group "iBGP"
                neighbor 192.0.2.3
                    family vpn-ipv4 vpn-ipv6
                exit
            exit
```

On PE-1, VPRN 1 is configured as follows. The configuration on PE-2 is similar.

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
```

```
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "loopback1" create
                address 1.1.1.1/32
                ipv6
                    address 2001:db8::1:1:1:1/128
                exit
                loopback
            exit
            no shutdown
        exit
```

Even though only LDP is explicitly configured in the auto-bind tunnel resolution filter, the resolution filter allows LDP and BGP tunnels, as follows:

```
*A:PE-1# configure service vprn 1
*A:PE-1>config>service>vprn# info detail | match auto-bind-tunnel post-lines 14
            auto-bind-tunnel
                resolution-filter
                    no gre
                    ldp
                    no rsvp
                    no sr-isis
                    no sr-ospf
                    no sr-te
                    bgp
                    no udp
                exit
                resolution filter
                no ecmp
                no weighted-ecmp
            exit
```

VPRN 1 is only configured on nodes in AS 64496, so only LDP transport tunnels are used. The following tunnel table on PE-2 shows that an LDP tunnel toward PE-1 is available:

```
*A:PE-2# show router tunnel-table 192.0.2.1

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        192.168.12.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

PE-2 receives the following BGP VPN-IPv4 route with route distinguisher (RD) 64496:1 used in VPRN 1:

```
*A:PE-2# show router bgp routes vpn-ipv4 rd 64496:1
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:1.1.1.1/32                                 100         None
      192.0.2.1                                          None        262140
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

For iBGP VPN routes on a node that is not an RR, the NH can only be resolved using a tunnel in the TTM. If the BGP NH is an IPv4 address, the system uses the most preferred tunnel matching the address and allowed by the resolution filter. The resolution filter allows LDP and BGP, but within an AS, only LDP tunnels are used. The following FIB for VPRN 1 on PE-2 shows that the transport tunnel to NH 192.0.2.1 is an LDP tunnel:

```
*A:PE-2# show router 1 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
1.1.1.1/32                                                  BGP_VPN
  192.0.2.1 (VPRN Label:262140 Transport:LDP)
2.2.2.1/32                                                  LOCAL
  2.2.2.1 (loopback1)
-------------------------------------------------------------------------------
Total Entries : 2
```

The same is shown for BGP IPv6 routes:

```
*A:PE-2# show router bgp routes vpn-ipv6 rd 64496:1
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
```

```
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  64496:1:2001:db8::1:1:1:1/128                      100         None
      ::ffff:192.0.2.1                                   None        262140
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

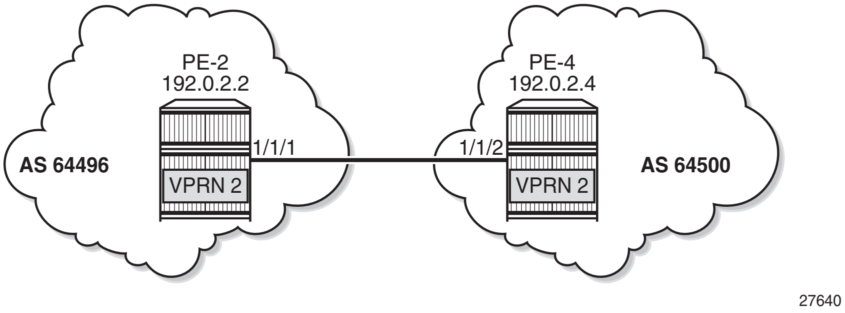
The following IPv6 FIB for VPRN 1 shows that a LDP tunnel is used to reach NH 192.0.2.1:

```
*A:PE-2# show router 1 fib 1 ipv6

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
2001:db8::1:1:1:1/128                                       BGP_VPN
  192.0.2.1 (VPRN Label:262140 Transport:LDP)
2001:db8::2:2:2:1/128                                       LOCAL
  2001:db8::2:2:2:1 (loopback1)
-------------------------------------------------------------------------------
Total Entries : 2
```

NH Resolution for Inter-AS VPRN Model B

Figure 167 shows that VPRN 2 is configured in AS 64496 and in AS 64500.

Figure 167 **VPRN 2 in AS 64496 and in AS 64500**



On PE-2, VPRN 2 is configured as follows. The service configuration on PE-4 is similar.

```
configure
    service
        vprn 2 customer 1 create
            route-distinguisher 2:2
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:2:2
            interface "loopback2" create
                address 2.2.2.2/32
                ipv6
                    address 2001:db8::2:2:2:2/128
                exit
                loopback
            exit
            no shutdown
        exit
```

BGP is configured for the VPN IP address families and BGP NH can be resolved to static routes. Multiple eBGP neighbors are defined, with NHs that can be resolved to a local, static, or tunneled route. The BGP configuration on PE-2 is as follows. The BGP configuration on PE-4 is similar.

```
*A:PE-2>config>router>bgp# info
----------------------------------------------
            enable-inter-as-vpn
            split-horizon
            next-hop-resolution
                labeled-routes
                    allow-static
                exit
            exit
            group "eBGP4_local"
                neighbor 192.168.24.2
                    family vpn-ipv4 vpn-ipv6
                    peer-as 64500
                exit
            exit
            group "eBGP4_static"
                neighbor 10.0.0.4
                    family vpn-ipv4 vpn-ipv6
                    local-address 10.0.0.2
                    peer-as 64500
                exit
            exit
            group "eBGP4_tunnel"
                neighbor 192.0.2.4
                    family vpn-ipv4 vpn-ipv6
                    peer-as 64500
                exit
            exit
            no shutdown
-----------------------------------------------
```

VPN IP NH Resolved to Local Route

PE-2 has three BGP VPN-IPv4 routes for prefix 4.4.4.2/32, as follows. The used route is NH 192.168.24.2, which is a local route.

```
*A:PE-2# show router bgp routes 4.4.4.2/32 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                             LocalPref   MED
      Nexthop (Router)                                    Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2:2:4.4.4.2/32                                      None        None
      192.168.24.2                                        None        262140
      64500
*i    2:2:4.4.4.2/32                                      None        None
      10.0.0.4                                            None        262140
      64500
*i    2:2:4.4.4.2/32                                      None        None
      192.0.2.4                                           None        262140
      64500
-------------------------------------------------------------------------------
Routes : 3
```

The IPv4 FIB on PE-2 shows prefix 4.4.4.2/32 with NH 192.168.24.2 on int-PE-2-PE-4, as follows. The NH is not resolved to a tunnel.

```
*A:PE-2# show router 2 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                               Protocol
  NextHop
-------------------------------------------------------------------------------
2.2.2.2/32                                                   LOCAL
  2.2.2.2 (loopback2)
4.4.4.2/32                                                   BGP_VPN
  192.168.24.2 (int-PE-2-PE-4)
-------------------------------------------------------------------------------
Total Entries : 2
```

In a similar way, the used VPN-IPv6 route on PE-2 has a NH resolved to a local route, as follows:

```
*A:PE-2# show router bgp routes 2001:db8::4:4:4:2/128 vpn-ipv6
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2:2:2001:db8::4:4:4:2/128                          None        None
      ::ffff:192.168.24.2                                None        262140
      64500
*i    2:2:2001:db8::4:4:4:2/128                          None        None
      ::ffff:10.0.0.4                                    None        262140
      64500
*i    2:2:2001:db8::4:4:4:2/128                          None        None
      ::ffff:192.0.2.4                                   None        262140
      64500
-------------------------------------------------------------------------------
Routes : 3
```

VPN IP NH Resolved to Static Route

When the eBGP session using the interface addresses is disabled, the next preferred NH resolution is static, which is allowed by configuration:

```
*A:PE-2# configure router bgp group "eBGP4_local" shutdown
```

On PE-2, the static route with the best preference is toward 10.0.0.4, as follows:

```
*A:PE-2# show router bgp routes 4.4.4.2/32 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
```

```
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2:2:4.4.4.2/32                                     None        None
      10.0.0.4                                           None        262140
      64500
*>i   2:2:4.4.4.2/32                                     None        None
      192.0.2.4                                          None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

On PE-2, NH 10.0.0.4 is resolved to 192.168.24.2, as follows:

```
*A:PE-2# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================

===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
10.0.0.4                                                    5        STATIC
   10.0.0.4/32                                              Y        1
   192.168.24.2                                                      0
192.0.2.4                                                   25       STATIC
   192.0.2.4/32                                             Y        1
   192.168.24.2                                                      0
-------------------------------------------------------------------------------
Next Hops : 2
```

This resolved NH 192.168.24.2 is the NH for prefix 4.4.4.2/32 in the FIB, as follows:

```
*A:PE-2# show router 2 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
2.2.2.2/32                                                  LOCAL
  2.2.2.2 (loopback2)
4.4.4.2/32                                                  BGP_VPN
  192.168.24.2 (int-PE-2-PE-4)
-------------------------------------------------------------------------------
Total Entries : 2
```

Also, for IPv6 routes on PE-2, the used route toward 2001:db8::4:4:4:2/128 has NH 10.0.0.4, as follows:

```
*A:PE-2# show router bgp routes 2001:db8::4:4:4:2/128 vpn-ipv6
```

```
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv6 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2:2:2001:db8::4:4:4:2/128                          None        None
      ::ffff:10.0.0.4                                    None        262140
      64500
*>i   2:2:2001:db8::4:4:4:2/128                          None        None
      ::ffff:192.0.2.4                                   None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

VPN IP NH Resolved to Tunneled Route

Note: This scenario is only for demonstration purposes. In an operational service provider network, no LDP sessions will be established to an untrusted AS (inter-AS VPRN model B is used for untrusted connections).

When the BGP configuration is changed to the default setting that static routes are not allowed for the NH resolution, the used BGP route toward 4.4.4.2/32 uses a tunnel toward the system address of the eBGP peer. The BGP configuration is modified as follows:

```
*A:PE-2# configure router bgp next-hop-resolution labeled-routes no allow-static
```

On PE-2, the used VPN-IPv4 route toward 4.4.4.2/32 has NH 192.0.2.4, as follows:

```
*A:PE-2# show router bgp routes 4.4.4.2/32 vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                                  LocalPref   MED
      Nexthop (Router)                                         Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  2:2:4.4.4.2/32                                           None        None
      192.0.2.4                                                None        262140
      64500
*i    2:2:4.4.4.2/32                                           None        None
      10.0.0.4                                                 None        262140
      64500
-------------------------------------------------------------------------------
Routes : 2
```

The tunnel table on PE-2 shows that an LDP tunnel is available toward 192.0.2.4/32, as follows:

```
*A:PE-2# show router tunnel-table 192.0.2.4/32

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.4/32      ldp       MPLS  65537     9        192.168.24.2   1
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

The following FIB on PE-2 shows that an LDP tunnel is used toward NH 192.0.2.4 to reach prefix 4.4.4.2/32:

```
*A:PE-2# show router 2 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
2.2.2.2/32                                                  LOCAL
  2.2.2.2 (loopback2)
4.4.4.2/32                                                  BGP_VPN
  192.0.2.4 (VPRN Label:262140 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 2
```

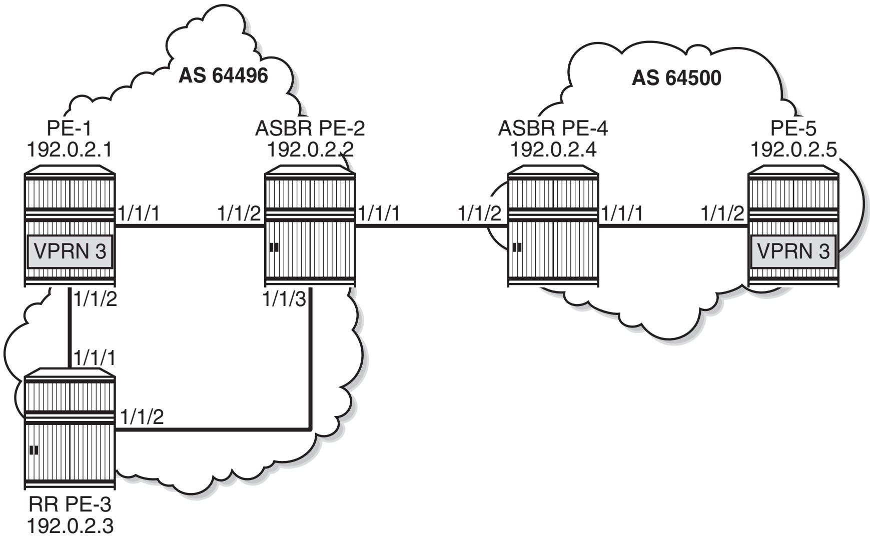
Similarly, the following IPv6 FIB on PE-2 shows that the same LDP tunnel is used toward NH 192.0.2.4 to reach prefix 2001:db8::4:4:4:2/128:

```
*A:PE-2# show router 2 fib 1 ipv6

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
  NextHop
-------------------------------------------------------------------------------
2001:db8::2:2:2:2/128                                       LOCAL
  2001:db8::2:2:2:2 (loopback2)
2001:db8::4:4:4:2/128                                       BGP_VPN
  192.0.2.4 (VPRN Label:262140 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 2
```

NH Resolution for Inter-AS VPRN Model C

Figure 168 shows the example topology with RR PE-3 in AS 64496. VPRN 3 is configured on PE-1 and PE-5.

Figure 168 **VPRN 3 - Inter-AS VPRN Model C**



27641

A labeled IPv4 eBGP session is established between ASBRs PE-2 and PE-4, and a multi-hop eBGP session is established between PE-1 and PE-5 for the VPN-IPv4 and VPN-IPv6 address families. The following BGP configuration is configured on PE-1. The configuration on PE-5 is similar.

```
*A:PE-1# configure router bgp
*A:PE-1>config>router>bgp# info
----------------------------------------------
            split-horizon
            group "iBGP"
                export "export-bgp"
                peer-as 64496
                neighbor 192.0.2.3
                    family vpn-ipv4 vpn-ipv6 label-ipv4
                exit
            exit
            group "eBGP_multihop"
                peer-as 64500
                neighbor 192.0.2.5
                    family vpn-ipv4 vpn-ipv6
                    local-address 192.0.2.1
                    multihop 10
                exit
            exit
            no shutdown
----------------------------------------------
```

The BGP configuration on RR PE-3 is as follows. The RR is configured with **disable-route-table-install**, so no routes are installed in the FIB; therefore, no eBGP multi-hop sessions can be established from the RR. The BGP NH will be resolved using the RTM. Local routes would be preferred, but there are no candidates. BGP NH resolution to static routes is not allowed in this configuration.

```
A:PE-3# configure router bgp
*A:PE-3>config>router>bgp# info
----------------------------------------------
            disable-route-table-install
            split-horizon
            next-hop-resolution
                labeled-routes
                    rr-use-route-table
                exit
            exit
            group "iBGP"
                cluster 192.0.2.3
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.1
                    family vpn-ipv4 vpn-ipv6 label-ipv4
                exit
                neighbor 192.0.2.2
                    family label-ipv4
                exit
            exit
            no shutdown
----------------------------------------------
```

On the ASBRs, BGP is only configured for the labeled IPv4 address family. The BGP configuration on PE-2 is as follows. The configuration on PE-4 is similar.

```
*A:PE-2# configure router bgp
*A:PE-2>config>router>bgp# info
----------------------------------------------
            split-horizon
            group "iBGP"
                family label-ipv4
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.3
                exit
            exit
            group "eBGP4_local"
                family label-ipv4
                advertise-inactive
                neighbor 192.168.24.2
                    peer-as 64500
                exit
            exit
            no shutdown
----------------------------------------------
```

On PE-1, VPRN 3 is configured as follows. The configuration is similar on PE-5.

```
configure
    service
        vprn 3 customer 1 create
            route-distinguisher 3:3
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:3:3
            interface "loopback3" create
                address 1.1.1.3/32
                ipv6
                    address 2001:db8::1:1:1:3/128
                exit
                loopback
            exit
            no shutdown
        exit
```

With the preceding configuration, the resolution filter in VPRN 3 allows the use of LDP and BGP tunnels, which can be verified as follows. BGP tunnels will be used for routes received from the peer AS.

```
*A:PE-1# configure service vprn 3
*A:PE-1>config>service>vprn# info detail | match auto-bind-tunnel post-lines 14
            auto-bind-tunnel
                resolution-filter
                    no gre
                    ldp
                    no rsvp
                    no sr-isis
                    no sr-ospf
                    no sr-te
                    bgp
                    no udp
                exit
                resolution filter
                no ecmp
                no weighted-ecmp
            exit
```

On PE-1, the VPN-IPv4 route for prefix 5.5.5.3/32 has NH 192.0.2.5 in the peer AS, as follows. Prefix 5.5.5.3/32 is the IP address of a loopback interface in VPRN 3 on PE-3.

```
*A:PE-1# show router bgp routes vpn-ipv4 rd 3:3
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3:3:5.5.5.3/32                                     None        None
      192.0.2.5                                          None        262141
      64500
-------------------------------------------------------------------------------
Routes : 1
```

On PE-1, the following tunnel table shows two tunnels: one LDP tunnel toward 192.0.2.2, and a BGP tunnel toward 192.0.2.5 in the remote AS.

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.5/32      bgp       MPLS  262146    12       192.0.2.2      1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The following FIB for VPRN 3 on PE-1 shows that the BGP tunnel is used for prefix 5.5.5.3/32 with NH 192.0.2.5:

```
*A:PE-1# show router 3 fib 1
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                            Protocol
  NextHop
-------------------------------------------------------------------------------
1.1.1.3/32                                                LOCAL
  1.1.1.3 (loopback3)
5.5.5.3/32                                                BGP_VPN
```

```
  192.0.2.5 (VPRN Label:262141 Transport:BGP)
-------------------------------------------------------------------------------
Total Entries : 2
```

On RR PE-3, the following VPN IP routes with NH 192.0.2.1 are reflected, but they are not installed in the FIB, so these are not used locally:

```
*A:PE-3# show router bgp routes vpn-ipv4 rd 3:3
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*>i   3:3:1.1.1.3/32                                     100         None
      192.0.2.1                                          None        262138
      No As-Path
*>i   3:3:5.5.5.3/32                                     100         None
      192.0.2.1                                          None        262137
      64500
-------------------------------------------------------------------------------
Routes : 2
```

On RR PE-3, NH 192.0.2.1 is resolved using the RTM, as follows:

```
*A:PE-3# show router bgp next-hop
===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
===============================================================================
BGP Next Hop
===============================================================================
Next Hop                                                    Pref     Owner
   Resolving Prefix                                         FibProg  Metric
   Resolved Next Hop                                                 Ref. Count
-------------------------------------------------------------------------------
192.0.2.1                                                   15       ISIS
   192.0.2.1/32                                             N        10
   192.168.13.1                                                      0
192.0.2.2                                                   15       ISIS
   192.0.2.2/32                                             N        10
   192.168.23.1                                                      0
-------------------------------------------------------------------------------
Next Hops : 2
```

Conclusion

The NH resolution of BGP routes using tunnels is consistent across different types of labeled route families (labeled IP and VPN-IP), both for eBGP and iBGP peering.

Policy Chaining and Logical Expressions

This chapter provides information about policy statement chaining and logical expressions.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter are based on SR OS release 14.0.R4. In SR OS releases earlier than 14.0.R1, only policy chaining was supported. SR OS release 14.0.R1 introduced support for route policy logical expressions using the logical operators AND, OR, and NOT. The **action drop** replaces the **action reject** in SR OS release 14.0.R4, and later.

Overview

Multiple policies can be chained together for sequential evaluation. For more complex evaluation logic, logical expressions (with operators AND, OR, and NOT) can be used. A logical expression can be included in a larger policy chain. Route policy logical expressions are supported in the following contexts:

- BGP export
- BGP import
- BGP leak-import (RIB leaking)
- VRF import
- VRF export
- GRT export (GRT leaking)

Table 14 shows a comparison between examples of policy chaining and policy logical expressions.

Table 14 **Policy Chaining Versus Policy Logical Expressions**

Policy chaining example	Policy logical expressions example
configure router bgp import "A" "B" "C"	configure router bgp import "[A] OR [B]"
For each route, policy A is evaluated first. • If policy A matches the route with **action next-policy**, then apply any route modifications and continue to evaluate policy B, and so on. • When the route is matched in a policy with **action accept** or **drop/reject**, the evaluation is completed.	Several logical operators can be used. This example shows an OR relationship between policy A and policy B. For each route, policy A is evaluated first. A true/false result is determined for policy A: • If true, the logical expression with operator OR is true already, and the evaluation is completed. • If false, then policy B is evaluated to determine the final true/false result. The final result is mapped back to a policy action (**accept**, **next-policy**, and so on).

Note: In SR OS release 14.0.R4, and later, the **action drop** replaces the **action reject**. The difference is that with **action drop**, it is possible to modify attributes in the same way as for **action accept**. This behavior is useful when a NOT operation makes a false expression true and the attributes are required. In a similar way, it is possible that an OR operation is true, even though the first policies that were evaluated were false. Routes are accepted when the final result is true and all policies that were evaluated will modify the route attributes. Examples of this behavior are included in the Configuration section.

To configure policy chaining - which may or may not include a policy logical expression - the syntax is the following:

```
*A:PE-1# configure router bgp export
  - export <plcy-or-long-expr> [<plcy-or-expr> [<plcy-or-expr>...(upto 14 max)]]
  - no export

 <plcy-or-long-expr>   : <policy-name> | <long-expr>
                         <policy-name>  - [64 chars max]
                         <long-expr>    - [255 chars max]
 <plcy-or-expr>        : <policy-name> | <expr>
                         <policy-name>  - [64 chars max]
                         <expr>         - [64 chars max]
```

A policy chain is 15 policies at maximum, one of which can be a logical expression. A logical expression in the policy chain can contain a maximum of 16 policies and can be anywhere in the command: the first operand can contain up to 255 characters, while the remaining operands can contain a maximum of 64 characters.

Policies in a logical expression need to be enclosed in square brackets. Like single policies, the logical expression can be enclosed in double quotes or not, as follows:

```
*A:PE-1# configure router bgp import [C]AND[A] "B"
*A:PE-1# configure router bgp import "[C]AND[A]" "B"
```

The logical expression can be anywhere in the policy chain, as follows:

```
*A:PE-1# configure router bgp import "B" [C]AND[A]
```

When quotes are used, spaces are allowed in the logical expression, as follows:

```
*A:PE-1# configure router bgp import "[A] AND [B]"
```

Without quotes, the logical operators are interpreted as policy names, as follows:

```
*A:PE-1# configure router bgp import [A] AND [B]
WARNING: CLI Policy "AND" does not exist.
```

The following message is raised when more than one logical expression is included in the policy chain:

```
*A:PE-1# configure router bgp import [C]AND[A] [C]AND[B]
INFO: BGP #1001 Configuration failed because of inconsistent values -
 BGP [VR 1] Policy stmts nbr logical expressions exceeded
```

The following message is raised when the logical expression is duplicated in the policy chain:

```
*A:PE-2# configure router bgp import [C]AND[A] "B" [C]AND[A]
INFO: BGP #1001 Configuration failed because of inconsistent values -
 BGP [VR 1] Policy stmts should be unique and set in order!
```

The following message is raised when the logical operators are not in uppercase:

```
*A:PE-1# configure router bgp import [A]and[B]
INFO: BGP #1001 Configuration failed because of inconsistent values -
 BGP [VR 1] Policy stmts expression format error
```

The following message is raised when more than 16 policies are in a policy expression:

```
*A:PE-
1# configure router bgp export [1]AND[2]AND[3]AND[4]AND[5]AND[6]AND[7]AND[8]AND[9]AN
D[10]AND[11]AND[12]AND[13]AND[14]AND[15]AND[16]


*A:PE-1# configure router bgp export [1]AND[2]AND[3]AND[4]AND[5]AND[6]AND[7]AND[8]
AND[9]AND[10]AND[11]AND[12]AND[13]AND[14]AND[15]AND[16]AND[17]
INFO: BGP #1001 Configuration failed because of inconsistent values -
 BGP [VR 1] Policy stmts expression format error
```

Route Policy Logical Expressions

Logical expressions are evaluated to be true or false. Table 15 shows the mapping of policy actions to Boolean values.

Table 15 **Boolean Values for the Policy Actions**

Policy action	Boolean value
Accept	True
Next-entry	True
Next-policy	True
Reject	False
Drop	False

➡ **Note:** : The policy **action drop** replaces **action reject** in release 14.0.R4, and later. The **action drop** supports route attribute modifications while **action reject** does not. SR OS automatically converts reject actions to drop actions.

Table 16 shows the evaluation actions for the logical operators NOT, OR, and AND.

Table 16 **Actions for the Logical Operators**

Logical Operator	Action
NOT <expr>	Swaps the true/false result of the expression.
<expr1> OR <expr2>	If expr1 is true, the result is true and expr2 need not be evaluated. If expr1 is false, expr2 must be evaluated. The final result is true if either expression is true; otherwise, it is false.
<expr1> AND <expr2>	If expr1 is false, the result is false and expr2 need not be evaluated. If expr1 is true, expr2 must be evaluated. The final result is true only if both expressions are true.

Table 17 shows the mapping of the final result of an expression to a policy action. Routes will be dropped when the entire expression is false.

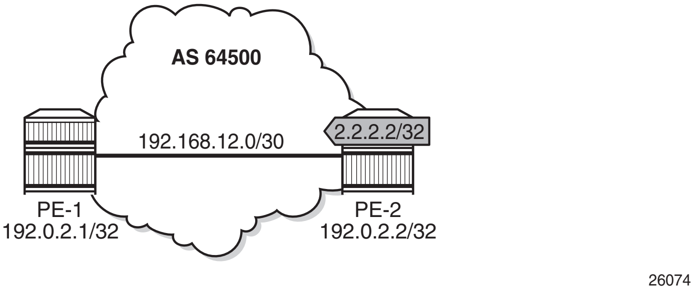
Table 17 **Mapping Final Result of an Expression to a Policy Action**

Final result	Action
True	**accept**, **next-entry**, or **next-policy** (depending on the last entry evaluated)
False	**drop/reject**

Configuration

Figure 1 shows the example topology including the advertised route.

Figure 169 **Example Topology**



The initial configuration of the routers includes the following:

- Cards, MDAs, ports
- Router interfaces
- IS-IS
- LDP
- BGP
- Export policy "export-bgp" accepting routes for prefix 2.2.2.2/32 on PE-2.

It is possible to configure VPRNs and assign policies to BGP in the VPRN, but in this chapter, all examples are for BGP in the base router.

Policy Chaining and Policy Logical Expressions

In this section, three route policies are configured that will add a community and set the local preference (LP): only policy C does not set LP. Policy C has **action next-policy**, and policies A and B have **action accept**. The configuration is as follows:

```
configure
    router
        policy-options
            begin
            community "A" members "1:1"
            community "B" members "2:2"
            community "C" members "3:3"
            policy-statement "A"
                entry 10
                    action accept
                        community add "A"
                        local-preference 110
                    exit
                exit
            exit
            policy-statement "B"
                entry 10
                    action accept
                        community add "B"
                        local-preference 200
                    exit
                exit
            exit
            policy-statement "C"
                entry 10
                    action next-policy
                        community add "C"
                    exit
                exit
            exit
            commit
```

Initially, policy chaining will be configured without a logical expression. Subsequently, policy chaining will be configured with only one policy logical expression and no other policies in the chain, as described in the following sections.

Policy Chaining without Logical Expression

Policy chaining may include one logical expression, but in this example, there is no policy logical expression in the chain.

Policy chaining is configured as follows on PE-1:

```
*A:PE-1# configure router bgp import "C" "A" "B"
```

PE-1 receives route 2.2.2.2/32 from PE-2. For each route, PE-1 evaluates policy C first. This policy adds community C (3:3) and has **action next-policy**, which implies that the next policy also needs to be evaluated. Policy A adds community A (1:1) and sets the LP to a value of 110 (by default, the local preference equals 100). Policy A has **action accept** and, therefore, the evaluation is completed. The local preference and the community are shown in the following output:

```
*A:PE-1# show router bgp routes hunt brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
Path Id        : None
From           : 192.0.2.2
Res. Nexthop   : 192.168.12.2
Local Pref.    : 110                    Interface Name : int-PE-1-PE-2
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : 1:1 3:3
---snip---
```

Policy Logical Expressions with Two Policies

In the following examples, the policy chain contains only a policy logical expression. When both policy A and policy B need to be executed, the logical operator to be used is AND. The sequence is important in this case, because both policies A and B set the LP and the last executed policy will set the final value for the LP. The following import policy expression is configured on PE-1:

```
*A:PE-1# configure router bgp import "[A]AND[B]"
```

Policy A is evaluated first and it adds community A (1:1) and sets LP 110. Then, policy B is evaluated, which adds community B (2:2) and sets LP 200.

```
*A:PE-1# show router bgp routes hunt brief | match "Local Pref."
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2
```

```
*A:PE-1# show router bgp routes hunt brief | match "Community"
Community      : 2:2 1:1
```

When the policy expression is [B]AND[A], the order is reversed. First, policy B sets LP 200, then policy A sets LP 110, as follows:

```
*A:PE-1# show router bgp routes hunt brief | match "Local Pref."
Local Pref.    : 110                    Interface Name : int-PE-1-PE-2

*A:PE-1# show router bgp routes hunt brief | match "Community"
Community      : 1:1 2:2
```

When the policy expression contains operator OR instead of AND, the first true expression results in a completed evaluation. Because both policy A and policy B result in a true expression, whichever policy is evaluated first is executed and the second one is skipped. When policy A is evaluated first and the result is true, policy B will be skipped. Therefore, the community will be A (1:1) and the LP 110, as follows:

```
*A:PE-1# configure router bgp import "[A]OR[B]"

*A:PE-1# show router bgp routes hunt brief | match "Local Pref."
Local Pref.    : 110                    Interface Name : int-PE-1-PE-2

*A:PE-1# show router bgp routes hunt brief | match "Community"
Community      : 1:1
```

Likewise, when policy B is evaluated first and the result is true, policy A will be skipped. The added community will be B (2:2) and the LP 200, as follows:

```
*A:PE-1# configure router bgp import "[B]OR[A]"

*A:PE-1# show router bgp routes hunt brief | match "Local Pref."
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2

*A:PE-1# show router bgp routes hunt brief | match "Community"
Community      : 2:2
```

The logical operator NOT swaps the result from true to false, and vice versa. When policy A is evaluated as true, NOT[A] is false. A false expression in an AND relationship leads to a false result. The next policy in the logical expression need not be evaluated. No communities will be added and no LP will be set (the default value for LP is 100). The route will be rejected as invalid, as follows:

```
*A:PE-1# configure router bgp import "NOT[A]AND[B]"

*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
```

```
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 100                    Interface Name : int-PE-1-PE-2
---snip---
Community      : No Community Members
---snip---
Flags          : Invalid  IGP  Rejected
---snip---
```

However, a false NOT[A] expression in an OR relation may still lead to the expression being evaluated to true, as follows:

```
*A:PE-1# configure router bgp import "NOT[A]OR[B]"
```

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 2:2 1:1
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

Policy B is evaluated as true for the route and, therefore, the entire logical expression "NOT[A]OR[B]" is true, and the route is accepted. Every policy in the expression that was evaluated, before the entire logical expression was recognized to be true, is executed, including policy A. This implies that policy A adds community A (1:1) to the route and sets LP to a value of 110. Then, policy B adds community B (2:2) to the route and overwrites the LP to a value of 200.

The import policy "[B] OR NOT[A]" is true after the first policy is evaluated as true. Only policy B is executed and the assigned community is B (2:2) and the LP is 200, as follows:

```
*A:PE-1# configure router bgp import "[B] OR NOT[A]"
```

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2
```

```
---snip---
Community      : 2:2
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

Table 18 summarizes the results for these different scenarios.

Table 18 **Assigned LP and Communities for the Import Logical Expressions**

Import Logical Expression	Assigned LP	Assigned Community
import "[A] AND [B]"	200	2:2 1:1
import "[B] AND [A]"	110	1:1 2:2
import "[A] OR [B]"	110	1:1
import "[B] OR [A]"	200	2:2
import "NOT[A] AND [B]"	None	None (Route rejected)
import "NOT[A] OR [B]"	200	2:2 1:1
Import "[B] OR NOT[A]"	200	2:2

Policy Logical Expressions with Three Policies

In policy chaining, the next policy in the chain will be evaluated when the action is **next-policy**. In policy logical expressions, the next policy will be evaluated depending on the logical operator and the Boolean value for the previous policies in the expression.

Policy C has **action next-policy** instead of **accept** and adds community C (3:3), but does not set the LP.

Several logical expressions can be made with policies A, B, and C. The following import policy has all three policies in an AND relationship. The expression is evaluated as true and all policies are executed: three communities are added and the LP is set.

```
*A:PE-1# configure router bgp import "[C]AND[A]AND[B]"
```

The first policy adds community C (3:3), the second policy adds community A (1:1) and sets LP 110, and the third policy adds community B (2:2) and sets LP 200, as follows:

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 2:2 1:1 3:3
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

The import policy "[C]AND[A]OR[B]" results in the first two being executed. Policy C is evaluated as true, and the logical operation is AND. Therefore, the next policy needs to be evaluated too. Policy A is also evaluated as true and the next operation is OR. The final result is evaluated as true without evaluating policy B. The communities added are C and A (3:3 and later 1:1) and the LP is 110.

```
*A:PE-1# configure router bgp import "[C]AND[A]OR[B]"
```

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 110                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 1:1 3:3
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

The import policy "[C]OR[A]OR[B]" is evaluated as true after the first policy is evaluated as true. Even though the action in policy C is **next-policy**, the next policy in this expression does not need to be evaluated, because the expression is true. Only policy C is executed and it adds the community C (3:3), but does not configure the LP, as follows:

```
*A:PE-1# configure router bgp import "[C]OR[A]OR[B]"
```

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
```

```
Local Pref.    : 100                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 3:3
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

However, if the policy chain contains not only a logical expression, but also single policies, the action next-policy ensures that a following policy in the chain is executed; for example, policy D in the following policy chain:

```
*A:PE-1# configure router bgp import "[C]OR[A]OR[B]" "D"
```

The expression "[C]OR[A]OR[B]" is true after policy C has been evaluated, but policy C has as **action next-hop** and policy D is the next policy to be evaluated.

The import policy "[C]OR[A]AND[B]" evaluates policy C as true. Policy C has an OR relation with policy A in the logical expression {[C]OR[A]}, and therefore, policy A need not be evaluated. There is an AND relation with policy B and policy B is evaluated as true. Therefore, the entire logical expression "[C]OR[A]AND[B]" is true and the route will be accepted. Both policy C and B are executed. First, policy C adds community C (3:3), then policy B adds community B (2:2) and sets LP 200, as follows:

```
*A:PE-1# configure router bgp import "[C]OR[A]AND[B]"
```

```
*A:PE-1# show router bgp routes hunt brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 200                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 2:2 3:3
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

Table 19 summarizes the results for these different scenarios.

Table 19 Assigned LP and Communities for the Import Logical Expressions

Import Logical Expression	Assigned LP	Assigned Community
import "[C] AND [A] AND [B]"	200	2:2 1:1 3:3
import "[C] AND [A] OR [B]"	110	1:1 3:3
import "[C] OR [A] OR [B]"	None	3:3
import "[C] OR [A] AND [B]"	200	2:2 3:3

Combinations of Policy Logical Operations Using Brackets

For this section, the following communities and policies are configured on PE-1. All these policies have a **from** condition that matches a community (D, E, F, G). Besides these policies, there are also export policies on PE-2 that add one or more communities (D, E, F, G) to the advertised routes. On PE-1, incoming route 2.2.2.2/32 will have one or more communities that may or may not match the **from** condition in the following route policies.

```
configure
    router
        policy-options
            begin
            community "D" members "4:4"
            community "E" members "5:5"
            community "F" members "6:6"
            community "G" members "7:7"
            policy-statement "D"
                entry 10
                    from
                        community "D"
                    exit
                    action accept
                        local-preference 4
                    exit
                exit
                default-action drop
                exit
            exit
            policy-statement "E"
                entry 10
                    from
                        community "E"
                    exit
                    action accept
                        local-preference 5
                    exit
                exit
                default-action drop
                exit
            exit
            policy-statement "F"
                entry 10
                    from
                        community "F"
                    exit
                    action accept
                        local-preference 6
                    exit
                exit
                default-action drop
                exit
            exit
            policy-statement "G"
                entry 10
                    from
                        community "G"
                    exit
                    action accept
                        local-preference 7
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

The received routes have community E (5:5) present. The following import policy is configured on PE-1:

```
*A:PE-1# configure router bgp import "([D]AND[E])OR([F]AND[G])"
```

The first policy that is evaluated requires community D (4:4) to be present. This is not the case and the expression between brackets, ([D]AND[E]), is false. Policy E need not be evaluated. The next policy to be evaluated is F and it requires community F (6:6), which is not present. The second expression between brackets, ([F]AND[G]), is therefore also false and policy G need not be evaluated. The entire policy logical expression is false and the route will be rejected.

The following commands show what policy evaluation caused the route to be rejected. For the entire logical expression "([D]AND[E])OR([F]AND[G])", the last policy that was evaluated, and that caused the route to be rejected, was policy F, as follows:

```
*A:PE-1# show router bgp policy-test "([D]AND[E])OR([F]AND[G])" family ipv4
prefix 0.0.0.0/0 longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

```
===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Rejected by Logical expression last policy F Default action
      2.2.2.2/32
-------------------------------------------------------------------------------
 Total Routes : 1 Routes rejected : 1
```

For the logical expression "[D]AND[E]", the last policy that was evaluated, and that led to the conclusion that the expression was false, was policy D, as follows:

```
*A:PE-1# show router bgp policy-test "[D]AND[E]" family ipv4 prefix 0.0.0.0/0
longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Rejected by Logical expression last policy D Default action
      2.2.2.2/32
-------------------------------------------------------------------------------
 Total Routes : 1 Routes rejected : 1
```

For the logical expression "[F]AND[G]", the last policy that was evaluated, and that led to the conclusion that the expression was false, was policy F, as follows:

```
*A:PE-1# show router bgp policy-test "[F]AND[G]" family ipv4 prefix 0.0.0.0/0
longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Rejected by Logical expression last policy F Default action
      2.2.2.2/32
-------------------------------------------------------------------------------
 Total Routes : 1 Routes rejected : 1
```

The logical expression "([D]AND[E])OR([F]AND[G])" is false and, therefore, the route is rejected, as follows. No LP will be set. Community E (5:5) was already present in the incoming route.

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network         : 2.2.2.2/32
Nexthop         : 192.0.2.2
---snip---
Local Pref.     : 100                    Interface Name : int-PE-1-PE-2
---snip---
Community       : 5:5
---snip---
Flags           : Invalid  IGP  Rejected
---snip---
```

In the second example, the incoming route contains communities D (4:4) and E (5:5). The same policy logical expression "([D]AND[E])OR([F]AND[G])" is evaluated as true because both policy D and policy E are true. There is an OR relationship with the rest of the expression and, therefore, the entire logical expression is true. Policy E is the last policy to be evaluated, as follows:

```
*A:PE-1# show router bgp policy-test "([D]AND[E])OR([F]AND[G])" family ipv4
prefix 0.0.0.0/0 longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Accepted by Logical expression last policy E Entry 10
      2.2.2.2/32
-------------------------------------------------------------------------------
Routes : 1
```

The route is accepted as valid and gets LP 5. The communities D (4:4) and E (5:5) were already present for the incoming route. The first policy that was executed, was policy D and it set the LP to a value of 4. Policy E was the second and last policy that was executed and it set the LP to a value of 5, as follows:

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
```

```
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 5                      Interface Name : int-PE-1-PE-2
---snip---
Community      : 5:5 4:4
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.2
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  IGP
---snip---
```

For the third example, the incoming route contains communities D (4:4) and E (5:5). The logical expression "([D]OR[E])AND([F]OR[G])" will be evaluated as false and the route will be rejected, follows:

```
*A:PE-1# configure router bgp import "([D]OR[E])AND([F]OR[G])"
```

First, policy D is evaluated as true because community D (4:4) is present. Policy D has an OR relationship with policy E, which will be true without the need to evaluate policy E. The next policy to be evaluated is F. Policy F requires the community F (6:6) to be present, which is not the case. The logical expression [F]OR[G] can only be true if policy G is true. Policy G requires community G (7:7) to be present, which is false. The last policy that was evaluated before the route was rejected was policy G, as follows:

```
*A:PE-1# show router bgp policy-test "([D]OR[E])AND([F]OR[G])" family ipv4
prefix 0.0.0.0/0 longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500        Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Rejected by Logical expression last policy G Default action
      2.2.2.2/32
-------------------------------------------------------------------------------
 Total Routes : 1 Routes rejected : 1
```

The route was rejected and, therefore, no policy was executed. The LP kept its default value of 100, as follows:

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
```

```
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 100                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 5:5 4:4
---snip---
Flags          : Invalid  IGP  Rejected
---snip---
```

For the fourth example, the incoming route has communities E (5:5) and G (7:7). The logical expression "([D]OR[E])AND([F]OR[G])" will be evaluated as true and the route will be accepted. First, policy D is evaluated as false. Policy D has an OR relationship with policy E, which will be evaluated as true. Consequently, the expression [D]OR[E] is true. This expression has an AND relationship with the expression [F]OR[G].

The next policy to be evaluated is F. Policy F requires the community F (6:6) to be present, which is false. The logical expression [F]OR[G] can only be true if policy G is true. Policy G requires community G (7:7) to be present, which is true. This makes [F]OR[G] true and also the entire expression "([D]OR[E])AND([F]OR[G])".

The last policy that was evaluated before the route was accepted was policy G, as follows:

```
*A:PE-1# show router bgp policy-test "([D]OR[E])AND([F]OR[G])" family ipv4
prefix 0.0.0.0/0 longer display-rejects brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
      Network
-------------------------------------------------------------------------------
Accepted by Logical expression last policy G Entry 10
      2.2.2.2/32
-------------------------------------------------------------------------------
Routes : 1
```

The route was accepted and has the changes of all policies that were evaluated: initially, policy D set the LP to 4. This value was overwritten by policy E to 5, by policy F to 6, and finally by policy G to a value of 7, as follows:

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
```

```
Network         : 2.2.2.2/32
Nexthop         : 192.0.2.2
---snip---
Local Pref.     : 7                     Interface Name : int-PE-1-PE-2
---snip---
Community       : 5:5 7:7
---snip---
Flags           : Used  Valid  Best  IGP
---snip---
```

Table 20 summarizes the results for these different scenarios.

Table 20 **Assigned LP for the Import Logical Expressions**

Ingress Community	Import Logical Expression	Assigned LP
5:5	import "([D]AND[E])OR([F]AND[G])"	Prefix rejected
5:5 4:4	import "([D]AND[E])OR([F]AND[G])"	5
5:5 4:4	import "([D]OR[E])AND([F]OR[G])"	Prefix rejected
5:5 7:7	import "([D]OR[E])AND([F]OR[G])"	7

Modification of Attributes While Processing

During the policy evaluation process, some prefix attributes can be modified while processing, and these modified attributes can be used as criteria for other policies in the logical expression.

In the following example, two route policies are configured:

- Policy X adds a new community Y (11:11) to the incoming route update.
- Policy Y uses community Y (11:11) as the only match criterion and removes communities X and Y. Policy Y also sets the LP to a value of 9, which is used here as an indication that policy Y was executed.

An export policy on PE-2 adds community X (10:10) to prefix 2.2.2.2/32 (not shown here).

Route policies X and Y are configured as follows on PE-1:

```
configure
    router
        policy-options
            begin
            community "X" members "10:10"
            community "Y" members "11:11"
            policy-statement "X"
    entry 10
        from
            community "X"
        exit
        action accept
            community add "Y"
        exit
    exit
exit
policy-statement "Y"
    entry 10
        from
            community "Y"
        exit
        action accept
            community remove "X" "Y"
            local-preference 9
        exit
    exit
exit
commit
```

When no import policy is applied on PE-1, the received route 2.2.2.2/32 has community 10:10 and the default LP, as follows:

```
*A:PE-1# show router bgp routes hunt brief
---snip---
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 100                    Interface Name : int-PE-1-PE-2
---snip---
Community      : 10:10
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

The import policy "[X]AND[Y]" is configured on PE-1, as follows:

```
*A:PE-1# configure router bgp import "[X]AND[Y]"
```

The route update contains community X (10:10) and policy X is evaluated as true. Policy X adds community Y (11:11) to the route. Policy Y requires this community and is evaluated as true. Therefore, the entire logical expression "[X]AND[Y]" is true and the route is accepted. Policy Y removes communities X (10:10) and Y (11:11), and sets the LP to a value of 9, as follows:

```
*A:PE-1# show router bgp routes hunt brief
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
Network        : 2.2.2.2/32
Nexthop        : 192.0.2.2
---snip---
Local Pref.    : 9                     Interface Name : int-PE-1-PE-2
---snip---
Community      : No Community Members
---snip---
Flags          : Used  Valid  Best  IGP
---snip---
```

Conclusion

Route policy chaining and logical expressions allow complex route processing logic to be broken into smaller components. These policy components are reusable and facilitate the process of updating route control logic. Logical expressions offer more flexible combinations of policy statements.

Pop-Label for /32 Label-IPv4 BGP Routes

This chapter describes the Pop-Label for /32 Label-IPv4 BGP routes. Topics include:

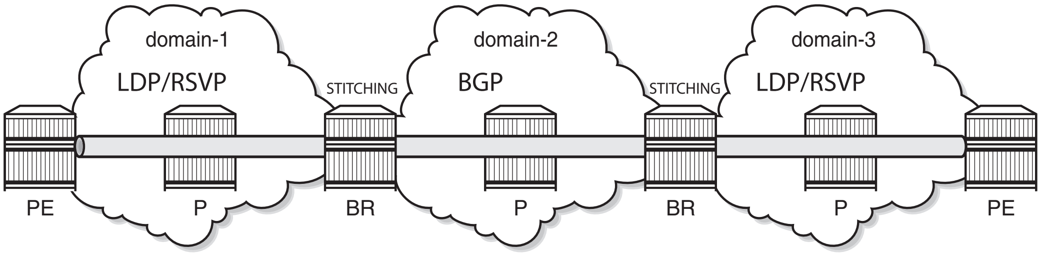
Applicability

The information and configuration in this chapter are based on SR OS Release 15.0.R5.

Overview

Labeled IPv4 routes are used in seamless MPLS and in VPRN inter-AS model C scenarios. In these scenarios, transport tunnels run through multiple domains, where the Area Border Routers (ABRs) or Autonomous System Border Routers (ASBRs) effectively stitch LDP/RSVP tunnels to BGP tunnels. For inter-AS model C, the domain is an autonomous system; for seamless MPLS, the domain is a part of an autonomous system. In either case, an end-to-end transport tunnel can be considered as a concatenation of multiple transport tunnels; see Figure 170.

Figure 170 **Stitching RSVP/LDP Tunnels to BGP Tunnels**



27611

Release 15.0.R1 enhanced the BGP support at the border router (ABR or ASBR) for /32 label-IPv4 BGP routes that are originated by exporting static, OSPF, or IS-IS routes from the route table into BGP. Before Release 15.0.R1, the advertisement of this type of BGP route always created a swap Ingress Label Mapping (ILM) entry in the data path, thereby stitching the BGP tunnel to the RSVP/LDP tunnel going to the destination as indicated by the /32 route.

Release 15.0.R1 provided a tighter coupling between the LDP/RSVP-TE and the BGP tunnels stitched at the ABR or ASBR, as follows:

1. By implementing an **accept** policy action (without the **advertise-label pop** modifier) for the /32 addresses in a **route-table-import** policy. The router advertises a /32 label-IPv4 route with a label that is swapped when an LDP/RSVP-TE is available, and withdrawn when the last LDP/RSVP-TE tunnel to that /32 prefix goes down. This applies to PEs with services, but should not be applied for route reflectors (RRs) when VPN addresses will be exchanged across eBGP sessions, because withdrawing labels for RRs would break the exchange of VPN routes. For the use of the **route-table-import** command, see the Separate BGP RIBs for Labeled Routes chapter.
2. By implementing the **accept** policy action with the a**dvertise-label pop** modifier for some system addresses in a **route-table-import** policy. The router advertises a /32 label-IPv4 route with a label that is popped rather than swapped, in case no LDP/RSVP-TE tunnel is available to that /32 prefix. This particularly applies to infrastructure nodes, for example off data path RRs, which do not participate in MPLS. RRs in different ASs, for example, still must be able to peer with each other through a multi-hop eBGP session, for the exchange of VPN routes belonging to the different services.

The **advertise-label pop** modifier can be used for the label-IPv4 redistribution of /32 prefixes of:

- OSPF and IS-IS routes
- Static routes:
  - Direct next-hop
  - Indirect next-hop
  - Blackhole

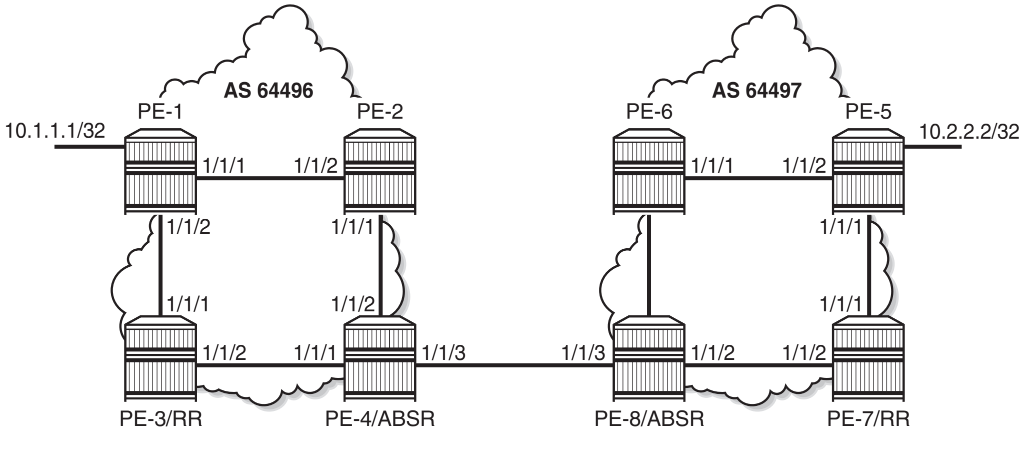
Redistributing /32 blackhole static routes does not require the **advertise-label pop** modifier; the label-IPv4 route is always advertised to the peer AS, and popped by the data plane.

The configuration in this chapter describes the redistribution of /32 prefixes for IS-IS routes. The redistribution of /32 routes for OSPF and the different static route types is similar.

Configuration

Figure 171 shows the example topology, depicting the inter-AS scenario also used in the Inter-AS VPRN Model C chapter. PE-1 and PE-5 host VPRN service 1, with 10.1.1.1/32 and 10.5.5.5/32 being the loopback addresses for this service on PE-1 and PE-5, respectively. In AS 64496, PE-3 is the IPv4 VPN RR, and PE-4 is the label-IPv4 RR toward clients PE-1 and PE-2. In AS 64497, PE-7 is the IPv4 VPN RR, and PE-8 is the label-IPv4 RR toward clients PE-5 and PE-6. IS-IS is the IGP for AS 64496 and 64497, and PE-4 and PE-8 are their respective ASBRs. Additionally, and in support for model C, the PE-3 and PE-7 RRs require a multi-hop IPv4 VPN eBGP connection.

Figure 171 **Example Topology**



The initial configuration includes:

1. Cards, MDAs, and ports.
2. Router interfaces.
3. IS-IS as IGP on all interfaces within AS 64496 and AS 64497 (alternatively, OSPF can be used).
4. LDP configured between PE-1, PE-2, and PE-4 in AS 64496, and between PE-5, PE-6, and PE-8 in AS 64497. The PE-3 and PE-7 RRs are off data path and do not have LDP enabled.

Base Configuration

In this example topology, the ASBRs generate the labeled routes, so that no BGP policies are required for generating labeled routes on the other PEs. The transport tunnels available in ASs 64496 and 64497 are LDP tunnels.

PE-1 and PE-2 peer with RR PE-3 for exchanging IPv4 VPN routes, and with RR PE-4 for receiving label-IPv4 routes. This enables PE-1 and PE-2 to exchange service traffic with the PEs in the peer AS. Their internal BGP configuration is as follows:

```
# on PE-1 and PE-2
configure
    router
        autonomous-system 64496
        bgp
            loop-detect discard-route
            split-horizon
            group "iBGP"
                peer-as 64496
                neighbor 192.0.2.3
                    family vpn-ipv4
                exit
                neighbor 192.0.2.4
                    family label-ipv4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

PE-3 is the IPv4 VPN RR for internal clients, using cluster ID 192.0.2.3, so it maintains iBGP sessions with PE-1 and PE-2. PE-3 also maintains an eBGP session with PE-7, which is the RR for clients PE-5 and PE-6 in AS 64497. The **vpn-apply-import**, **vpn-apply-export**, and **import** and **export** commands can be used at bgp, group, or neighbor level for selectively exchanging dedicated VPN routes. The BGP configuration for RR PE-3 is as follows:

```
# on PE-3, RR
configure
    router
        autonomous-system 64496
        bgp
            loop-detect discard-route
            split-horizon
            group "eBGP-vpn"
                peer-as 64497
                local-address 192.0.2.3
                neighbor 192.0.2.7
                    family vpn-ipv4
                    multihop 10
                    vpn-apply-import
                    vpn-apply-export
                exit
            exit
            group "iBGP-vpn"
                cluster 192.0.2.3
                peer-as 64496
                neighbor 192.0.2.1
                    family vpn-ipv4
                exit
                neighbor 192.0.2.2
                    family vpn-ipv4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

PE-4 is the label-IPv4 RR for internal clients, using cluster ID 192.0.2.4, so it maintains iBGP sessions with PE-1 and PE-2. PE-4 imposes **next-hop-self** on the iBGP advertised label-IPv4 routes. PE-4 also maintains an eBGP session with PE-8, and requires the **advertise-inactive** command for stitching to apply. The reason for the **advertise-inactive** command is that the system IP addresses for PEs are advertised in IGP and in BGP. Because the IGP has a lower preference value than BGP, the BGP routes are rendered inactive. By default, inactive BGP routes are not advertised to the peer AS, and the **advertise-inactive** command bypasses this issue. The BGP configuration for PE-4 is as follows:

```
# on PE-4, ASBR
configure
    router
        autonomous-system 64496
        bgp
            loop-detect discard-route
            enable-inter-as-vpn
            split-horizon
            rib-management
                label-ipv4
                    route-table-import "to-AS64497"
                exit
            exit
            group "eBGP-label"
                export "exp-ALL"
                advertise-inactive
                neighbor 192.168.48.2
                    family label-ipv4
                    peer-as 64497
                exit
            exit
            group "iBGP-label"
                next-hop-self
                cluster 192.0.2.4
                peer-as 64496
                neighbor 192.0.2.1
                    family label-ipv4
                exit
                neighbor 192.0.2.2
```

```
                    family label-ipv4
                exit
            exit
            no shutdown
        exit
    exit
exit
```

The *PE-pfxs* prefix list is the set of exact /32 addresses of the PEs in AS 64496, excluding the RR. The *RR-pfxs* prefix list is the exact /32 address of RR PE-3. The *to-AS64497* policy in ASBR PE-4 matches the *PE-pfxs* prefix list in entry 10 with action accept (without modifier), and the *RR-pfxs* prefix list in entry 20 with action accept and the advertise-label pop modifier. The *exp-ALL* policy is used to advertise the combined set of prefixes to the peer AS. These policies are defined on ASBR PE-4 as follows:

```
# on PE-4, ASBR
configure
    router
        policy-options
            begin
            prefix-list "PE-pfxs"
                prefix 192.0.2.1/32 exact
                prefix 192.0.2.2/32 exact
            exit
            prefix-list "RR-pfxs"
                prefix 192.0.2.3/32 exact
            exit
            policy-statement "exp-ALL"
                entry 10
                    from
                        prefix-list "PE-pfxs" "RR-pfxs"
                    exit
                    action accept
                    exit
                exit
            exit
            policy-statement "to-AS64497"
                entry 10
                    from
                        prefix-list "PE-pfxs"
                    exit
                    action accept
                    exit
                exit
                entry 20
                    from
                        prefix-list "RR-pfxs"
                    exit
                    action accept
                        advertise-label pop
                    exit
                exit
            exit
            commit
        exit
    exit
exit
```

Because PE-3 is deliberately placed off the data path, not participating in MPLS, an indirect static route is added to its configuration so that it can establish an eBGP session with PE-7, as follows:

```
configure
    router
        static-route-entry 192.0.2.7/32
            indirect 192.0.2.4
                tunnel-next-hop
                    resolution disabled
                exit
                no shutdown
            exit
        exit
    exit
exit
```

The configuration of the PEs in AS 64497 is similar to the PEs in AS 64496; see Figure 171 for the addresses required.

Redistributing IGP /32 Routes to Label-IPv4 Routes

With the configuration as indicated in the previous section, PE-4 advertises the system addresses used in AS 64496 to PE-8 in the peer AS as label-IPv4 routes. The label-IPv4 routes advertised are as follows:

```
*A:PE-4# show router bgp neighbor 192.168.48.2 advertised-routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     192.0.2.1/32                                       n/a         20
      192.168.48.1                                       None        262138
      64496
i     192.0.2.2/32                                       n/a         10
      192.168.48.1                                       None        262139
      64496
i     192.0.2.3/32                                       n/a         10
```

```
      192.168.48.1                                       None        262140
      64496
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-4#
```

The label-IPv4 routes are accepted and put in the routing table of PE-8. The next-hop for all the label-IPv4 routes is 192.168.48.1, as follows:

```
*A:PE-8# show router route-table 192.0.2.0/24 longer

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  BGP_LABEL 00h13m54s  170
       192.168.48.1                                                 0
192.0.2.2/32                                  Remote  BGP_LABEL 00h13m54s  170
       192.168.48.1                                                 0
192.0.2.3/32                                  Remote  BGP_LABEL 00h02m46s  170
       192.168.48.1                                                 0
192.0.2.5/32                                  Remote  ISIS      03d06h12m  18
       192.168.68.1                                                 20
192.0.2.6/32                                  Remote  ISIS      03d06h12m  18
       192.168.68.1                                                 10
192.0.2.7/32                                  Remote  ISIS      03d06h12m  18
       192.168.78.1                                                 10
192.0.2.8/32                                  Local   Local     03d06h12m  0
       system                                                       0
-------------------------------------------------------------------------------
No. of Routes: 7
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-8#
```

Also, PE-8 is advertising label-IPv4 routes to PE-4, so that PE-4 ultimately has LDP and BGP tunnels available to destinations in its own and its peer AS, respectively, as follows:

```
*A:PE-4# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65547     9        192.168.24.1    20
192.0.2.2/32      ldp       MPLS  65537     9        192.168.24.1    10
192.0.2.5/32      bgp       MPLS  262145    12       192.168.48.2   1000
192.0.2.6/32      bgp       MPLS  262147    12       192.168.48.2   1000
```

```
192.0.2.7/32      bgp       MPLS  262146    12       192.168.48.2   1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-4#
```

PE-4 effectively stitches the BGP tunnels to the LDP tunnels, as follows:

```
*A:PE-4# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                      Received       Advertised     Label
                             Label          Label          Origin
-------------------------------------------------------------------------------
0.0.0.0                      0              262140         Edge
192.0.2.1                    262142         262138         InternalLdp
192.0.2.2                    262143         262139         InternalLdp
192.168.48.2                 262138         262137         External
192.168.48.2                 262139         262135         External
192.168.48.2                 262140         262136         External
-------------------------------------------------------------------------------
Total Labels allocated:   6
===============================================================================
*A:PE-4#
```

The first entry in this table, with advertised label 262140, is used for tunnels for which PE-4 is the end-point, so that no stitching is required. This is indicated by setting the next-hop, the received label, and the label origin to 0.0.0.0, 0, and Edge, respectively.

The second and third entries, with advertised labels 262138 and 262139, are used for tunnels to PE-1 and PE-2, respectively. Taking PE-1 as an example, label 262138 is swapped to label 262142, where 262142 is assigned through LDP (label origin is InternalLdp), thereby stitching the BGP tunnel to the LDP tunnel, and vice versa.

The last three entries, with advertised labels 262137, 262135, and 262136, and received labels 262138, 262139, and 262140, respectively, are used for tunnels to the PEs in the peer AS, which can be verified by displaying the label-IPv4 routes received by PE-4, as follows:

```
*A:PE-4# show router bgp neighbor 192.168.48.2 received-routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       n/a         20
      192.168.48.2                                       None        262138
      64497
u*>i  192.0.2.6/32                                       n/a         10
      192.168.48.2                                       None        262139
      64497
u*>i  192.0.2.7/32                                       n/a         10
      192.168.48.2                                       None        262140
      64497
-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-4#
```

Verifying the content of the RIB provides an alternative to check whether tunnels are stitched. A check is performed for PE-1, which can have services defined, and for PE-3, which does not have any.

Checking for the 192.0.2.1/32 prefix in the PE-4 RIB shows that label 262138 is advertised to 192.168.48.2, and the label type is swap, as follows. This is consistent with the output from the previous commands.

```
*A:PE-4# show router bgp routes 192.0.2.1/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.168.48.1
Path Id        : None
To             : 192.168.48.2
Res. Nexthop   : n/a
Local Pref.    : n/a                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 20
AIGP Metric    : None
Connector      : None
```

```
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                    Peer Router Id : 192.0.2.8
IPv4 Label     : 262138                  Label Type     : SWAP
Lbl Allocation : NEXT-HOP
Origin         : IGP
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: NotFound
Source Class   : 0                       Dest Class     : 0
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

Checking for the 192.0.2.3/32 prefix in the PE-4 RIB shows that label 262140 is advertised to 192.168.48.2, and the label type is pop, as follows:

```
*A:PE-4# show router bgp routes 192.0.2.3/32 label-ipv4 hunt
===============================================================================
 BGP Router ID:192.0.2.4        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------


-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.3/32
Nexthop        : 192.168.48.1
Path Id        : None
To             : 192.168.48.2
Res. Nexthop   : n/a
Local Pref.    : n/a                     Interface Name : NotAvailable
Aggregator AS  : None                    Aggregator     : None
Atomic Aggr.   : Not Atomic              MED            : 10
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                    Peer Router Id : 192.0.2.8
IPv4 Label     : 262140                  Label Type     : POP
Lbl Allocation : NEXT-HOP
Origin         : IGP
AS-Path        : 64496
Route Tag      : 0
Neighbor-AS    : 64496
Orig Validation: NotFound
```

```
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-4#
```

RR/PE-3 and RR/PE-7 have a multi-hop eBGP session established and are exchanging VPN routes, as follows:

```
 *A:PE-3# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
ServiceId          AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.1
Def. Instance  64496     9437    0 03d04h27m 1/0/1 (VpnIPv4)
                         9230    0
192.0.2.2
Def. Instance  64496     9434    0 03d06h34m 0/0/3 (VpnIPv4)
                         9476    0
192.0.2.7
Def. Instance  64497     8874    0 00h15m02s 1/0/1 (VpnIPv4)
                           75    0

-------------------------------------------------------------------------------
*A:PE-3#
```

Communication between PE-1 and PE-5 is verified with a ping:

```
*A:PE-1# ping router 1 10.5.5.5
PING 10.5.5.5 56 data bytes
64 bytes from 10.5.5.5: icmp_seq=1 ttl=64 time=4.99ms.
64 bytes from 10.5.5.5: icmp_seq=2 ttl=64 time=4.72ms.
64 bytes from 10.5.5.5: icmp_seq=3 ttl=64 time=4.94ms.
64 bytes from 10.5.5.5: icmp_seq=4 ttl=64 time=5.24ms.
64 bytes from 10.5.5.5: icmp_seq=5 ttl=64 time=4.70ms.

---- 10.5.5.5 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 4.70ms, avg = 4.92ms, max = 5.24ms, stddev = 0.198ms
*A:PE-1#
```

Shutting down LDP on PE-1 results in PE-4 withdrawing the label-IPv4 route to 192.0.2.1, as follows:

```
156 2018/04/27 16:01:56.671 CEST MINOR: DEBUG #2001 Base Peer 1: 192.168.48.2
"Peer 1: 192.168.48.2: UPDATE
Peer 1: 192.168.48.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 15
    Flag: 0x90 Type: 15 Len: 11 Multiprotocol Unreachable NLRI:
        Address Family LBL-IPV4-Labeled
        192.0.2.1/32 Label 0
"
```

Conclusion

Implementing the **advertise-label pop** policy action in a **route-table-import** policy provides operators the means to save on resources used in the network.

Separate BGP RIBs for Labeled Routes

This chapter provides information about separate border gateway protocol (BGP) route information bases (RIBs) for labeled-unicast routes.

Topics in this chapter include:

Applicability

The information and configuration in this chapter are based on SR OS release 14.0.R4.

Release 14.0.R4 introduced separate BGP RIBs for labeled-unicast routes.

Overview

In SR OS releases earlier than 14.0.R4, BGP maintained one RIB for both unlabeled IPv4 and labeled-unicast IPv4 routes and another RIB for both unlabeled IPv6 and labeled-unicast IPv6 routes. In SR OS release 14.0.R4, and later, the labeled-unicast and unlabeled routes are held in separate RIBs: IPv4, label-IPv4, IPv6, and label-IPv6.

Table 21 compares the previous BGP RIB architecture to the current one.

Table 21 **Comparison of the BGP RIB Architectures**

	SR OS earlier than 14.0.R4	SR OS 14.0.R4 and later
BGP peer with both labeled-unicast (SAFI 4) and unlabeled IP (SAFI 1) routes	SAFI 1 and SAFI 4 are mutually exclusive. Received routes with the non-negotiated SAFI are treated as route withdrawals.	Both labeled and unlabeled routes can be advertised to the same neighbor.

Table 21 **Comparison of the BGP RIB Architectures (Continued)**

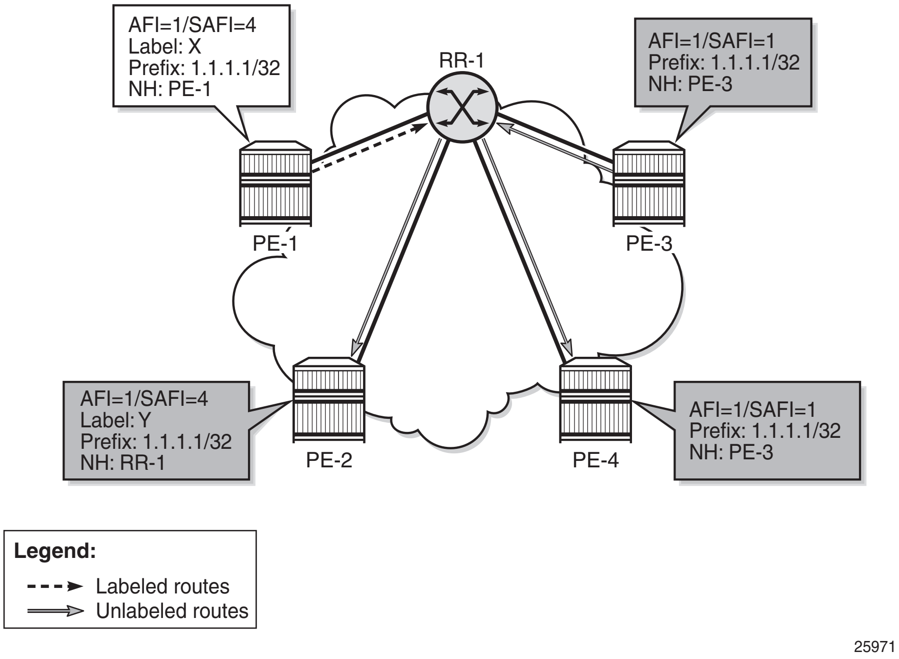
	SR OS earlier than 14.0.R4	SR OS 14.0.R4 and later
RR default behavior	BGP RR re-advertises all types of routes to all its clients, including labeled routes to unlabeled peers and vice versa.	BGP RR only re-advertises labeled-unicast routes to labeled-unicast peers and unlabeled routes to unlabeled IP peers, unless route-table-import policies are used to allow route leaking.
Route-table import policies	No route-table import policies supported. All active routes are automatically imported to BGP RIB.	Route-table import policies can save BGP memory by blocking the import of unnecessary IP routes. All unlabeled and all labeled BGP-owned routes are by default imported into both BGP RIBs.

BGP Common IPv4 RIB Implementation

In SR OS releases earlier than 14.0.R4, unlabeled IP routes and labeled IP routes for the same prefix were considered different paths to the same destination. SR OS could not advertise both labeled and unlabeled routes to the same neighbor. If the advertise-label ipv4 command was configured on a session, labeled routes were advertised, and unlabeled routes were not expected and not sent. For labeled routes, scarce MPLS datapath resources were consumed. In many networks, advertising a labeled route was only required for certain prefixes and not the entire IPv4 or IPv6 routing table.

Another issue with the common IPv4 RIB architecture is shown in Figure 172.

Figure 172 **RR-1 with Common IPv4 RIB**



Route reflector RR-1 has two clients sending labeled-IPv4 routes (PE-1, PE-2) and two other clients sending unlabeled IPv4 routes (PE-3, PE-4). The RR is intended to be a control plane RR and not configured to perform next-hop-self toward any of the clients. When both PE-1 (labeled) and PE-3 (unlabeled) send a route for IPv4 prefix 1.1.1.1/32, the RR uses the BGP decision process to select the best path for 1.1.1.1/32. Assume that PE-3 sends the best path. This path will be advertised to clients PE-2 (labeled) and PE-4 (unlabeled).

The route received from PE-3 is unlabeled, and can be advertised as an unlabeled route to peer PE-4. However, toward PE-2, the route must get a label Y, according to the configured label in the **advertise-label ipv4** command, but the next hop cannot remain PE-3, because PE-3 does not expect any labels. Therefore, RR-1 must set next-hop-self and data traffic between PE-2 and PE-3 will flow via RR-1 where label Y will be pushed or popped. This is not a desirable situation.

BGP Separate Labeled-IPv4 RIB Implementation

In SR OS release 14.0.R4, and later, a separate RIB is used for labeled-IPv4 routes. With this implementation, client PE-2 learns the best labeled-IPv4 route and client PE-4 learns the best unlabeled IPv4 route. RR-1 does not need to set next-hop-self and traffic can be sent directly from PE-2 to PE-1 and from PE-4 to PE-3. The RR is used only for control traffic, as intended. Figure 173 shows how RR-1 sends a labeled-IPv4 route to PE-2 with label X and next hop PE-1.

Figure 173 **RR-1 with Separate Labeled-IPv4 RIB Implementation**

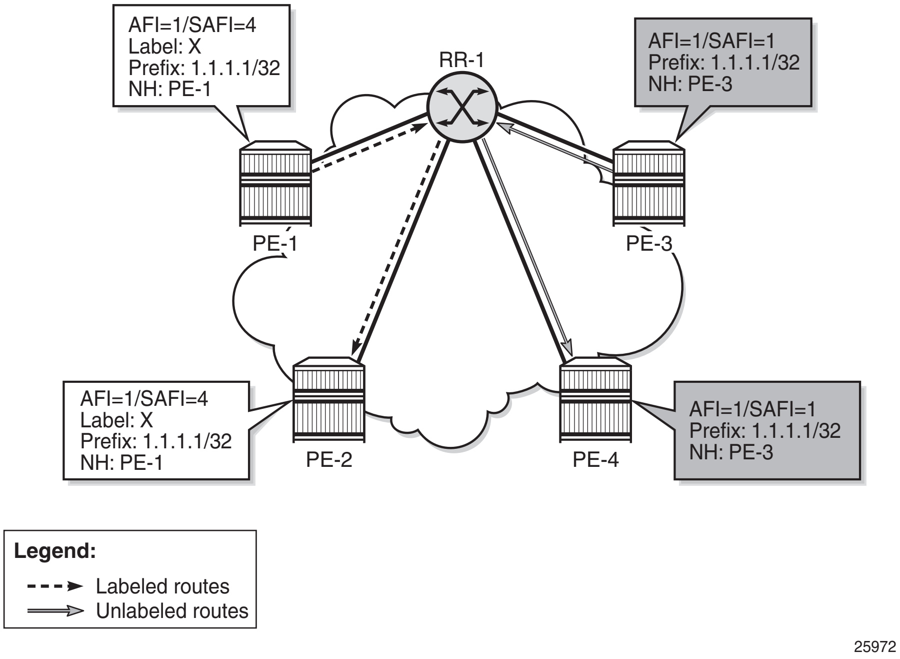
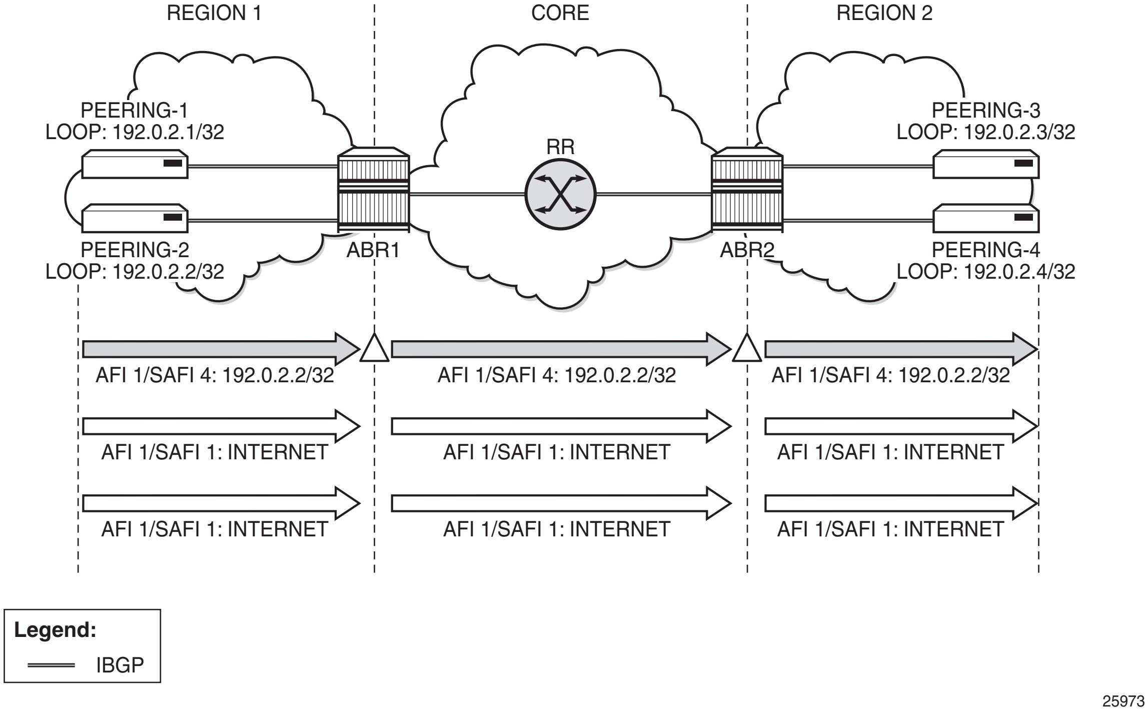


Figure 174 shows a seamless MPLS use case, which is a good example of the coexistence of labeled (AFI 1/SAFI 4) and unlabeled (AFI 1/SAFI 1) BGP sessions.

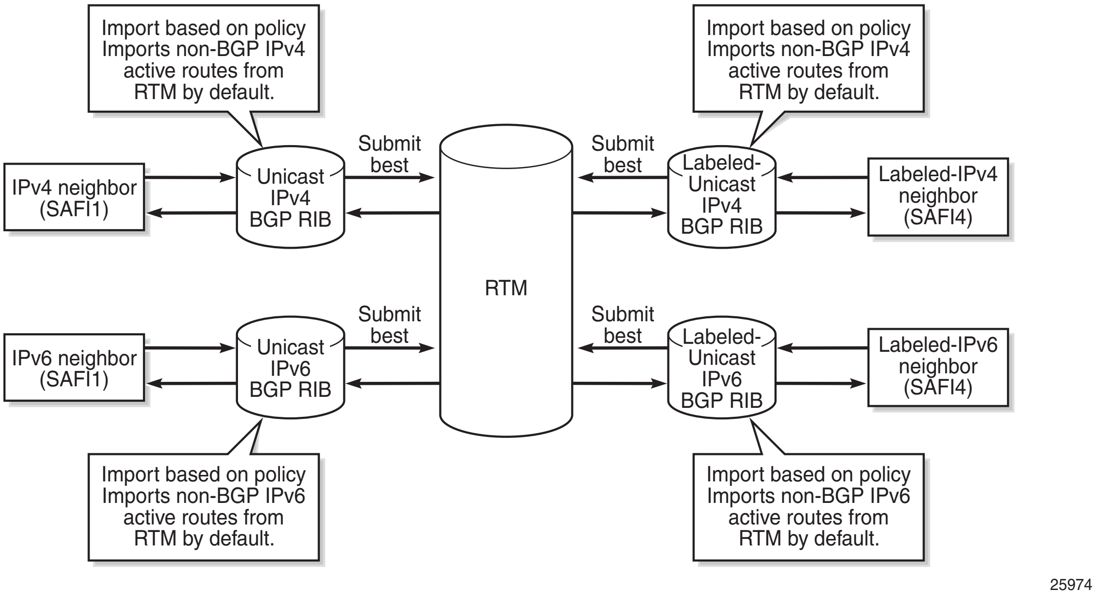
Figure 174 **Seamless MPLS - Separate Labeled-IPv4 Implementation**



RIB Architecture

Figure 175 shows the system architecture with four separate RIBs for IPv4 and IPv6 routes.

Figure 175 **System Architecture with Separate RIBs for Labeled-Unicast and Unlabeled Routes**



Labeled-unicast routes from peers are stored in a labeled RIB and unlabeled routes from the same or different peers are stored in a non-labeled RIB. Both labeled and unlabeled routes can be sent and received to and from the same peer. Different sets of routes can be advertised to labeled/unlabeled peers. Labeled and unlabeled BGP sessions are using the common equal cost multipath (ECMP) and multipath limit.

More user control is provided over the RTM route import process. By default, a RIB imports all non-BGP active routes from RTM, but a user-defined route policy can be applied. Route policies can be used to reduce BGP memory usage.

The configuration for labeled IP routes has changed in SR OS release 14.0.R4:

- The **advertise-label** command is deprecated.
- Address families mapped to the RIBs are: **ipv4, label-ipv4, ipv6, label-ipv6**.
- In route policies, protocol types **bgp** and **bgp-label** can be used.
- The default RTM preference for labeled IP routes is configurable (**label-preference**) in the BGP context of the base router or a VPRN. The default preference is 170.

The consequences of having four separate RIBs instead of two are the following:

- Increased BGP memory usage is possible because the same route may be present in two RIBs. This can be avoided by a good route-table-import policy.

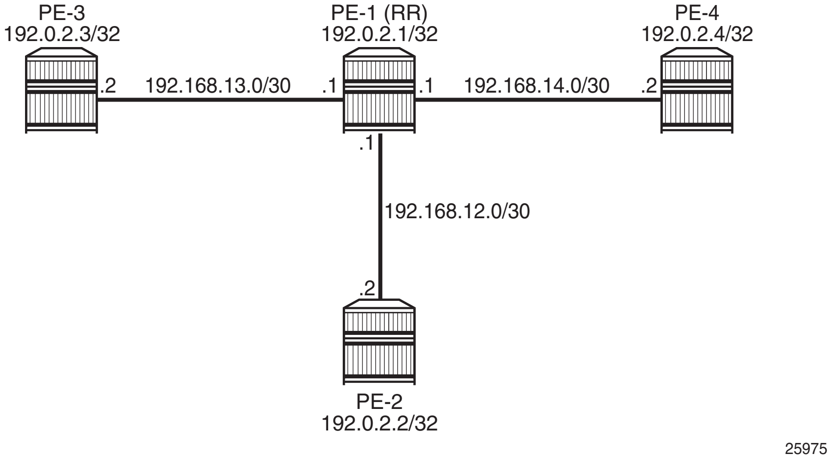
- By default, BGP unlabeled routes are no longer advertised to BGP label peers and vice versa. However, non-default route-table-import policies can accept BGP routes of the other type.
- The paths that are advertised to a peer for a prefix may be different, even if full routes are leaked between labeled and unlabeled RIBs. Leaking is done through the RTM and each RIB submits its best path. However, by default, the same route will be sent to both the labeled and unlabeled peer if the route-table-import policy allows the import to the BGP/BGP-Label owned routes.
- Care must be taken during upgrade.

Configuration

All the examples are based on labeled and unlabeled IPv4 addresses. For IPv6, the configuration is similar.

Figure 176 shows the example topology using IPv4 addresses.

Figure 176 **Example IPv4 Topology**



The initial configuration includes:

- Cards, MDAs, ports
- Router interfaces
- IS-IS in AS 64500 (PE-1, PE-2, PE-4)
- LDP in AS 64500
- Loopback addresses 3.3.3.3/32 in PE-3 and 4.4.4.4/32 in PE-4

- Export policy "export-bgp" accepting routes from protocol direct on all nodes

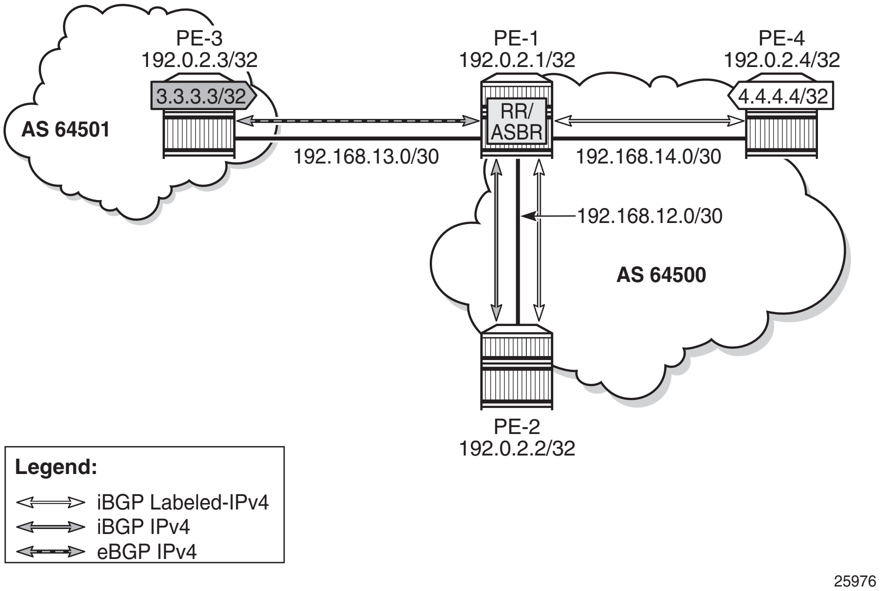
The following will be configured and verified:

1. Coexistence of labeled and unlabeled address families for BGP
2. Applying next-hop-self
3. Export Policy to Advertise Route as Labeled/Unlabeled
4. Behavior of RR with a mix of labeled and unlabeled iBGP sessions

Coexistence of Labeled and Unlabeled Address Families for BGP

Figure 177 shows the eBGP and iBGP sessions that are established between the nodes and the routes advertised for the loopback addresses.

Figure 177 **BGP Sessions**



PE-1 acts as RR for PE-2 and PE-4, and it is an autonomous system border router (ASBR) toward PE-3. PE-1 has two single-family connections: unlabeled IPv4 to PE-3 and labeled IPv4 to PE-4. PE-1 also has one dual-family connection to PE-2. The BGP configuration on PE-1 is as follows:

```
configure
```

```
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            split-horizon
            group "eBGP"
                peer-as 64501
                neighbor 192.168.13.2
                    family ipv4
                exit
            exit
            group "iBGP"
                cluster 192.0.2.1
                export "export-bgp"
                peer-as 64500
                neighbor 192.0.2.2
                    family ipv4 label-ipv4
                exit
                neighbor 192.0.2.4
                    family label-ipv4
                exit
            exit
        exit
    exit
exit
```

The BGP configuration on PE-2 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            split-horizon
            group "iBGP"
                export "export-bgp"
                peer-as 64500
                neighbor 192.0.2.1
                    family ipv4 label-ipv4
                exit
            exit
        exit
    exit
exit
```

The BGP configuration on PE-3 in AS 64501 is as follows:

```
configure
    router
        autonomous-system 64501
        bgp
            min-route-advertisement 1
            split-horizon
            group "eBGP"
                export "export-bgp"
                peer-as 64500
                neighbor 192.168.13.1
```

```
                    family ipv4
                exit
            exit
        exit
    exit
exit
```

The BGP configuration on PE-4 is as follows:

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            split-horizon
            group "iBGP"
                export "export-bgp"
                peer-as 64500
                neighbor 192.0.2.1
                    family label-ipv4
                exit
            exit
        exit
    exit
exit
```

The BGP summary on PE-1 shows that there is a dual-family connection with PE-2: IPv4 and Lbl-IPv4. PE-1 has an Lbl-IPv4 connection with PE-4 and an IPv4 connection with PE-3.

```
*A:PE-1# show router bgp summary
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
BGP Admin State         : Up          BGP Oper State              : Up
Total Peer Groups       : 2           Total Peers                 : 3
Total BGP Paths         : 14          Total Path Memory           : 2808
Total IPv4 Remote Rts   : 5           Total IPv4 Rem. Active Rts  : 2
Total McIPv4 Remote Rts : 0           Total McIPv4 Rem. Active Rts: 0
Total McIPv6 Remote Rts : 0           Total McIPv6 Rem. Active Rts: 0
Total IPv6 Remote Rts   : 0           Total IPv6 Rem. Active Rts  : 0
Total IPv4 Backup Rts   : 0           Total IPv6 Backup Rts       : 0

Total Supressed Rts     : 0           Total Hist. Rts             : 0
Total Decay Rts         : 0

Total VPN Peer Groups   : 0           Total VPN Peers             : 0
Total VPN Local Rts     : 0
Total VPN-IPv4 Rem. Rts : 0           Total VPN-IPv4 Rem. Act. Rts: 0
Total VPN-IPv6 Rem. Rts : 0           Total VPN-IPv6 Rem. Act. Rts: 0
Total VPN-IPv4 Bkup Rts : 0           Total VPN-IPv6 Bkup Rts     : 0

Total VPN Supp. Rts     : 0           Total VPN Hist. Rts         : 0
Total VPN Decay Rts     : 0

Total L2-VPN Rem. Rts   : 0           Total L2VPN Rem. Act. Rts   : 0
Total MVPN-IPv4 Rem Rts : 0           Total MVPN-IPv4 Rem Act Rts : 0
Total MDT-SAFI Rem Rts  : 0          Total MDT-SAFI Rem Act Rts  : 0
Total MSPW Rem Rts      : 0          Total MSPW Rem Act Rts      : 0
Total RouteTgt Rem Rts  : 0          Total RouteTgt Rem Act Rts  : 0
Total McVpnIPv4 Rem Rts : 0          Total McVpnIPv4 Rem Act Rts : 0
Total MVPN-IPv6 Rem Rts : 0          Total MVPN-IPv6 Rem Act Rts : 0
Total EVPN Rem Rts      : 0          Total EVPN Rem Act Rts      : 0
Total FlowIpv4 Rem Rts  : 0          Total FlowIpv4 Rem Act Rts  : 0
Total FlowIpv6 Rem Rts  : 0          Total FlowIpv6 Rem Act Rts  : 0
Total LblIpv4 Rem Rts   : 5          Total LblIpv4 Rem. Act Rts  : 1
Total LblIpv6 Rem Rts   : 0          Total LblIpv6 Rem. Act Rts  : 0
Total LblIpv4 Bkp Rts   : 0          Total LblIpv6 Bkp Rts       : 0
===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
               64500        5    0 00h00m25s 2/0/6 (IPv4)
                            8    0           2/0/5 (Lbl-IPv4)
192.0.2.4
               64500        5    0 00h00m32s 3/1/4 (Lbl-IPv4)
                            7    0
192.168.13.2
               64501        5    0 00h00m43s 3/2/0 (IPv4)
                            5    0
-------------------------------------------------------------------------------
*A:PE-1#
```

The unlabeled IPv4 routes on PE-1 include unlabeled routes imported from PE-2 and PE-3, including the loopback address 3.3.3.3/32 advertised by PE-3, as follows:

```
*A:PE-1# show router bgp routes ipv4
===============================================================================
 BGP Router ID:192.0.2.1       AS:64500        Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.3.3/32                                         None        None
      192.168.13.2                                       None        -
      64501
*i    192.0.2.2/32                                       100         None
      192.0.2.2                                          None        -
      No As-Path
u*>i  192.0.2.3/32                                       None        None
```

```
      192.168.13.2                                       None        -
      64501
*i    192.168.12.0/30                                    100         None
      192.0.2.2                                          None        -
      No As-Path
*i    192.168.13.0/30                                    None        None
      192.168.13.2                                       None        -
      64501
-------------------------------------------------------------------------------
Routes : 5
===============================================================================
*A:PE-1#
```

The labeled-unicast IPv4 routes on PE-1 include labeled routes imported from PE-2 and PE-4, including the loopback address 4.4.4.4/32 advertised by PE-4, as follows:

```
*A:PE-1# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  4.4.4.4/32                                         100         None
      192.0.2.4                                          None        262139
      No As-Path
*i    192.0.2.2/32                                       100         None
      192.0.2.2                                          None        262139
      No As-Path
*i    192.0.2.4/32                                       100         None
      192.0.2.4                                          None        262140
      No As-Path
*i    192.168.12.0/30                                    100         None
      192.0.2.2                                          None        262140
      No As-Path
*i    192.168.14.0/30                                    100         None
      192.0.2.4                                          None        262138
      No As-Path
-------------------------------------------------------------------------------
Routes : 5
===============================================================================
*A:PE-1#
```

PE-2 imports the prefix 3.3.3.3/32 in its unlabeled RIB, as follows:

```
*A:PE-2# show router bgp routes 3.3.3.3/32
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     3.3.3.3/32                                         100         None
      192.168.13.2                                       None        -
      64501
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

PE-2 imports the prefix 4.4.4.4/32 in its labeled RIB, as follows:

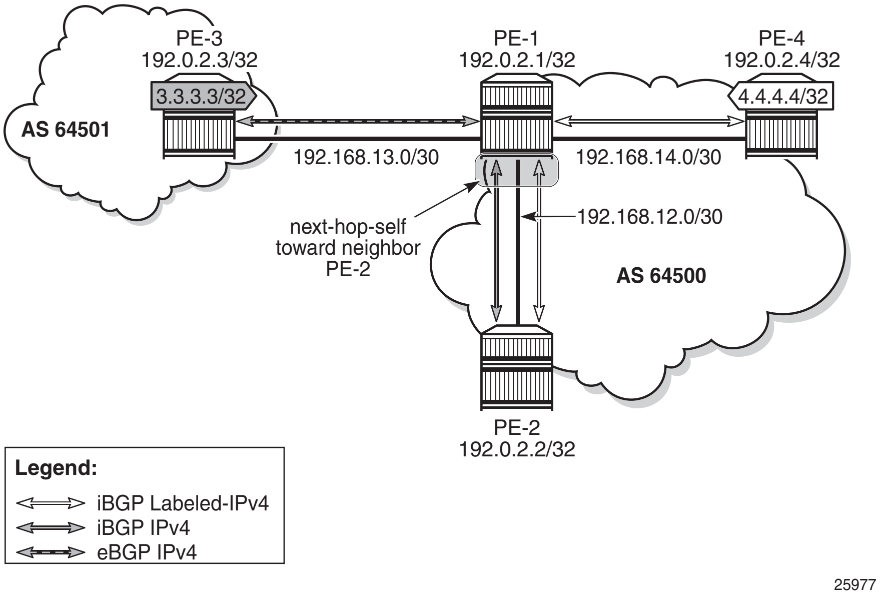
```
*A:PE-2# show router bgp routes 4.4.4.4/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  4.4.4.4/32                                         100         None
      192.0.2.4                                          None        262139
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

As expected, the prefixes from address family label-ipv4 are advertised independently from the prefixes from address family ipv4.

Applying Next-Hop-Self

Figure 178 shows that PE-1 applies next-hop-self for BGP updates toward PE-2.

Figure 178 **PE-1 Applies Next-Hop-Self toward Neighbor PE-2**



On PE-1, next-hop-self is enabled for neighbor PE-2 only, as follows:

```
*A:PE-1# configure router bgp group "iBGP" neighbor 192.0.2.2 next-hop-self
```

This applies to both address families. The next hop for unlabeled route 3.3.3.3/32 will be 192.0.2.1, as follows:

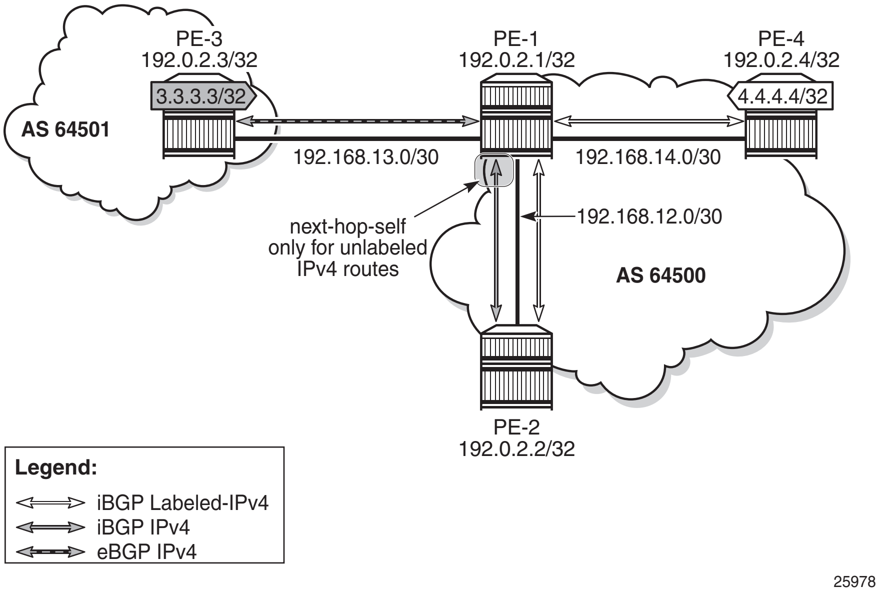
```
*A:PE-2# show router bgp routes 3.3.3.3/32
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.3.3/32                                         100         None
      192.0.2.1                                          None        -
      64501
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The labeled-unicast route 4.4.4.4/32 also has next hop 192.0.2.1, as follows:

```
*A:PE-2# show router bgp routes 4.4.4.4/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  4.4.4.4/32                                         100         None
      192.0.2.1                                          None        262136
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

Figure 179 shows that next-hop-self is applied to unlabeled IPv4 routes only.

Figure 179 **Applying Next-Hop-Self to Unlabeled IP-4 Routes to Neighbor PE-2**



An export policy is configured to ensure that next-hop-self is only applied for address family ipv4. The route policy is configured as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "export-nhs"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                        next-hop-self
                    exit
                exit
                entry 20
                    from
                        protocol bgp-label
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The next-hop-self configuration for neighbor PE-2 is replaced by export policy "export-nhs", as follows:

```
*A:PE-1# configure router bgp group "iBGP" neighbor 192.0.2.2 no next-hop-self
*A:PE-1# configure router bgp group "iBGP" neighbor 192.0.2.2 export "export-nhs"
```

With this export policy, only the unlabeled route 3.3.3.3/32 will have next hop 192.0.2.1, while the labeled-unicast route 4.4.4.4/32 will have next hop 192.0.2.4, as follows:

```
*A:PE-2# show router bgp routes 3.3.3.3/32
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.3.3/32                                         100         None
      192.0.2.1                                          None        -
      64501
-------------------------------------------------------------------------------
Routes : 1
```

```
===============================================================================
*A:PE-2# show router bgp routes 4.4.4.4/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  4.4.4.4/32                                         100         None
      192.0.2.4                                          None        262139
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

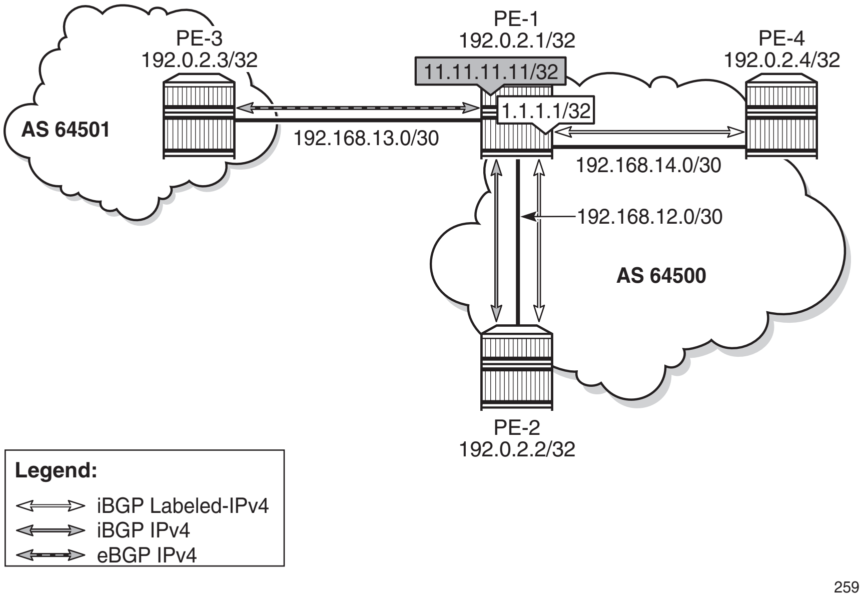
The export policy "export-nhs" toward neighbor PE-2 is removed as follows:

```
*A:PE-1# configure router bgp group "iBGP" neighbor 192.0.2.2 no export
```

Export Policy to Advertise Route as Labeled/Unlabeled

Figure 180 shows that two loopback addresses are configured in PE-1 to be advertised: prefix 1.1.1.1/32 and 11.11.11.11/32. Initially, there is no route policy applied for a selective export as labeled or unlabeled route.

Figure 180 **PE-1 Advertises Prefixes 1.1.1.1/32 and 11.11.11.11/32**



By default, these prefixes will be advertised as both labeled and unlabeled routes toward dual-family neighbor PE-2. On PE-2, the unlabeled IPv4 RIB contains prefixes 1.1.1.1/32 and 11.11.11.11/32, as follows:

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  1.1.1.1/32                                         100         None
      192.0.2.1                                          None        -
      No As-Path
---snip---
u*>i  11.11.11.11/32                                     100         None
      192.0.2.1                                          None        -
      No As-Path
---snip---
```

The labeled-IPv4 RIB on PE-2 also contains prefixes 1.1.1.1/32 and 11.11.11.11/32, as follows:

```
*A:PE-2# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
*>i   1.1.1.1/32                                         100         None
      192.0.2.1                                          None        262136
      No As-Path
---snip---
*>i   11.11.11.11/32                                     100         None
      192.0.2.1                                          None        262135
      No As-Path
---snip---
```

In many cases, it is not required to advertise both a labeled route and an unlabeled route. The following policy is configured to advertise prefix 1.1.1.1/32 as a labeled-IPv4 route and prefix 11.11.11.11/32 as an unlabeled IPv4 route:

```
configure
    router
        policy-options
            begin
            prefix-list "1.1.1.1/32"
                prefix 1.1.1.1/32 exact
            exit
            prefix-list "11.11.11.11/32"
                prefix 11.11.11.11/32 exact
            exit
            policy-statement "export-bgp1"
                entry 10
                    from
                        prefix-list "1.1.1.1/32"
                    exit
                    to
                        protocol bgp-label
                    exit
                    action accept
                    exit
                exit
                entry 30
                    from
                        prefix-list "11.11.11.11/32"
                    exit
                    to
```

```
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
```

This policy is applied on PE-1 as an export policy for neighbor PE-2, as follows:

```
*A:PE-1# configure router bgp group "iBGP" neighbor 192.0.2.2 export "export-bgp1"
```

Prefix 11.11.11.11/32 is received as an unlabeled route on PE-2 and stored in the unlabeled RIB, but prefix 1.1.1.1/32 is not, as follows:

```
*A:PE-2# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  3.3.3.3/32                                         100         None
      192.168.13.2                                       None        -
      64501
u*>i  11.11.11.11/32                                     100         None
      192.0.2.1                                          None        -
      No As-Path
u*>i  192.0.2.3/32                                       100         None
      192.168.13.2                                       None        -
      64501
-------------------------------------------------------------------------------
Routes : 3
```

On PE-2, prefix 1.1.1.1/32 is received as a labeled route and stored in the labeled-IPv4 RIB, but prefix 11.11.11.11/32 is not, as follows:

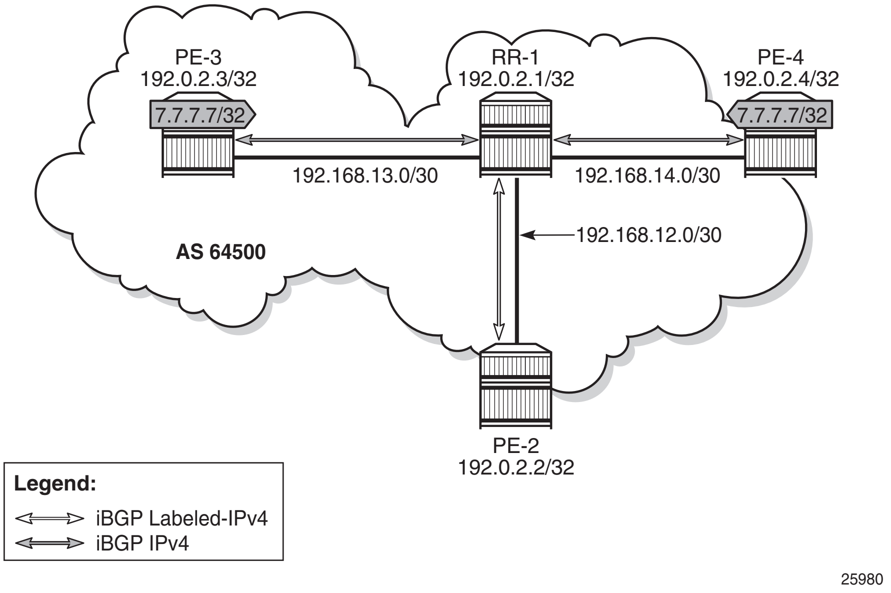
```
*A:PE-2# show router bgp routes label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  1.1.1.1/32                                         100         None
      192.0.2.1                                          None        262136
      No As-Path
u*>i  4.4.4.4/32                                         100         None
      192.0.2.4                                          None        262138
      No As-Path
-------------------------------------------------------------------------------
Routes : 2
```

This selective route advertisement from PE-1 reduces the memory usage for the RIBs on PE-2.

RR Behavior with a Mix of Labeled and Unlabeled BGP Sessions

Figure 181 shows a slightly different setup, with all PEs in the AS 64500 and RR-1 acting as the RR for all PEs. There are no dual-family connections. PE-3 and PE-4 have an unlabeled BGP session with RR-1 and PE-2 has a labeled BGP connection with RR-1. RR-1 has add-path=2 capability configured for neighbor PE-2. RR-1 receives the same prefix 7.7.7.7/32 from two neighbors: PE-3 and PE-4.

Figure 181 **RR with Labeled and Unlabeled BGP Sessions**



On RR-1, BGP is configured as follows:

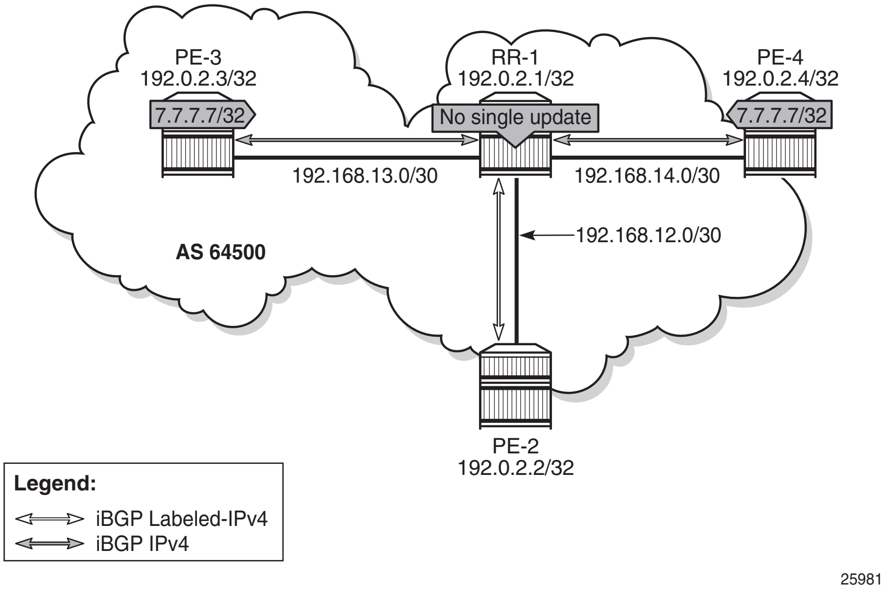
```
configure
    router
        bgp
            min-route-advertisement 1
            split-horizon
            group "iBGP"
                cluster 192.0.2.1
                export "export-bgp"
                peer-as 64500
                neighbor 192.0.2.2
                    family label-ipv4
                    add-paths
                        label-ipv4 send 2 receive
                    exit
                exit
                neighbor 192.0.2.3
                    family ipv4
                exit
                neighbor 192.0.2.4
                    family ipv4
                exit
            exit
        exit
    exit
exit
```

RR-1 receives the prefix 7.7.7.7/32 from neighbors PE-3 and PE-4, as follows:

```
*A:RR-1# show router bgp routes 7.7.7.7/32
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  7.7.7.7/32                                         100         None
      192.0.2.3                                          None        -
      No As-Path
*i    7.7.7.7/32                                         100         None
      192.0.2.4                                          None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 2
```

Both routes are unlabeled and BGP updates from unlabeled sessions are by default not exported to a labeled-IPv4 session, as shown in Figure 182.

Figure 182 **Updates from Unlabeled Sessions Not Propagated to Labeled Sessions (Default)**



PE-2 will not receive prefix 7.7.7.7/32, neither as unlabeled route, nor as labeled route, as follows:

```
*A:PE-2# show router bgp routes 7.7.7.7/32 ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                          LocalPref   MED
      Nexthop (Router)                                 Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-2#
*A:PE-2# show router bgp routes 7.7.7.7/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                                  LocalPref   MED
      Nexthop (Router)                                         Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-2#
```

A route policy is created on RR-1 to accept both labeled and unlabeled routes, as follows:

```
configure
    router
        policy-options
            begin
            policy-statement "import-all"
                entry 10
                    from
                        protocol bgp
                    exit
                    action accept
                    exit
                exit
                entry 20
                    from
                        protocol bgp-label
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This policy accepts all routes, labeled and unlabeled. For route 7.7.7.7/32 to be advertised to the labeled peer PE-2, it is sufficient to have a policy with only entry 10 that says from protocol bgp action accept. However, the preceding policy can also be used to import labeled routes to be advertised to unlabeled peers.

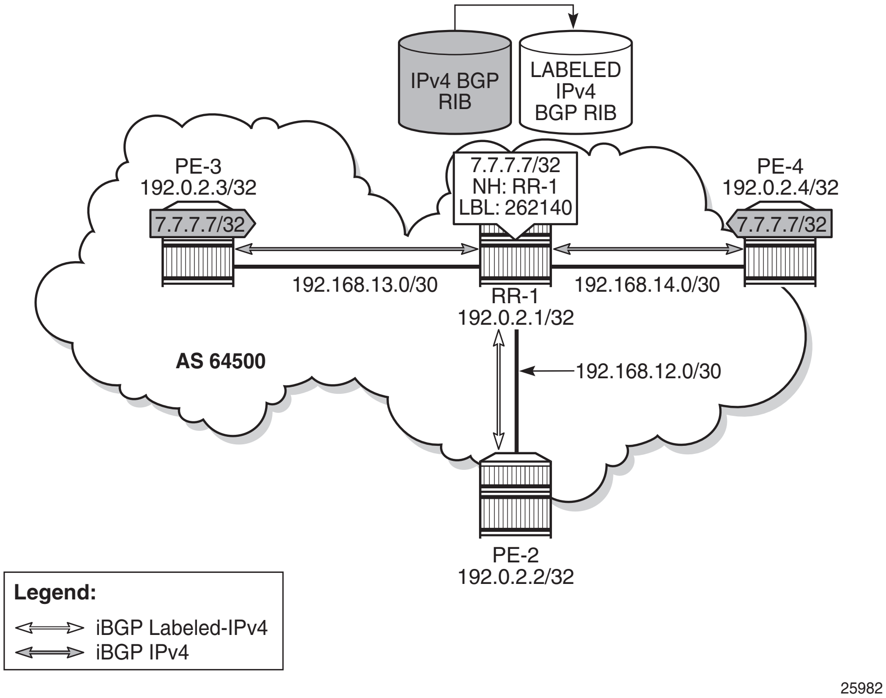
The following policy is applied as route-table-import policy in BGP RIB management, both for unlabeled IPv4 routes and labeled-IPv4 routes on RR-1:

```
configure
    router
        bgp
            rib-management
                ipv4
                    route-table-import "import-all"
                exit
                label-ipv4
                    route-table-import "import-all"
                exit
            exit
        exit
    exit
exit
```

For allowing unlabeled route 7.7.7.7/32 to be advertised on a labeled session, it is sufficient to have a route-table-import for labeled-IPv4 only. However, the configuration allows for RIB leaking in both ways: from unlabeled IPv4 BGP RIB to labeled-IPv4 BGP RIB and vice versa.

Figure 183 shows this RIB leaking process.

Figure 183 **RIB Leaking from IPv4 BGP RIB to Labeled-IPv4 BGP RIB**



After applying this RIB leaking, RR-1 will advertise prefix 7.7.7.7/32 to PE-2. Therefore, RR-1 needs to add a label to the route and RR-1 needs to set next-hop-self. RR-1 advertises only one labeled route for prefix 7.7.7.7/32, with next hop 192.0.2.1, as follows:

```
*A:RR-1# show router bgp neighbor 192.0.2.2 label-ipv4 advertised-routes
===============================================================================
 BGP Router ID:192.0.2.1        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
```

```
===============================================================================
Flag  Network                                                LocalPref   MED
      Nexthop (Router)                                       Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     7.7.7.7/32                                             100         None
      192.0.2.1                                              None        262140
      No As-Path
---snip---
```

The BGP update message is as follows:

```
9 2016/09/07 11:44:01.86 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.2
"Peer 1: 192.0.2.2: UPDATE
Peer 1: 192.0.2.2 - Send BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 49
    Flag: 0x90 Type: 14 Len: 17 Multiprotocol Reachable NLRI:
        Address Family LBL-IPV4-Labeled
        NextHop len 4 NextHop 192.0.2.1
        7.7.7.7/32 Label 262140
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.3
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        192.0.2.1
"
```

On PE-2, the following labeled BGP route is imported:

```
*A:PE-2# show router bgp routes 7.7.7.7/32 label-ipv4
===============================================================================
 BGP Router ID:192.0.2.2        AS:64500       Local AS:64500
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
===============================================================================
BGP Routes
===============================================================================
Flag  Network                                                LocalPref   MED
      Nexthop (Router)                                       Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  7.7.7.7/32                                             100         None
      192.0.2.1                                              None        262140
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
```

Conclusion

The BGP RIB architecture in SR OS release 14.0.R4, and later, with separate RIBs for unlabeled and labeled-unicast routes supports unlabeled sessions and labeled sessions in parallel. By default, labeled routes are not advertised to unlabeled sessions and vice versa. Route-table import policies for RIB management allow route leaking between separate RIBs: unlabeled BGP RIB and labeled-unicast BGP RIB.

MPLS

In This Section

This section provides MPLS configuration information for the following topics:

- Tunneling of ICMP Reply Packets over MPLS LSPs
- Unnumbered Interfaces in RSVP-TE and LDP

Automatic Bandwidth Adjustment in P2P LSPs

This chapter provides information about automatic bandwidth adjustment in P2P LSPs.

Topics in this chapter include:

Applicability

Automatic bandwidth adjustment was first introduced in SR OS release 8.0.R4. Overflow triggers are supported from 8.0.R4 onward; underflow triggers from 12.0.R1 onward. From 12.0.R4 onward, auto-bandwidth adjustment is also supported on LSPs that have secondary paths.

Initially, this chapter was written for SR OS release 13.0.R2, but the CLI in the current edition corresponds to SR OS release 16.0.R3.

Overview

Automatic bandwidth adjustment refers to the capability of an ingress Label Edge Router (iLER) to dynamically adjust the bandwidth of a Resource Reservation Protocol (RSVP) Label Switched Path (LSP) tunnel based on active measurement of the traffic rate into the tunnel. The bandwidth assigned to an RSVP LSP tunnel is taken into account by the control plane, to verify that sufficient bandwidth is available for a new LSP or for an increase or decrease in bandwidth for an existing LSP. The actual bandwidth in the data plane is not capped by this setting. QoS mechanisms can be set up to filter and police the traffic in the data plane, but that is beyond the scope of this chapter.

Auto-bandwidth adjustment uses the existing LSP egress statistics feature to track the bandwidth on a specific LSP. When egress statistics are enabled, the Control Processing Module (CPM) collects statistics from all IOMs forwarding traffic belonging to the LSP (whether the traffic is currently leaving the ingress LER via the primary path, a secondary path, or an FRR detour/bypass path). The egress statistics have counts for the number of packets and bytes forwarded per LSP on a per-forwarding class, per priority (in-profile versus out-of-profile) basis.

For the actual bandwidth adjustment, Make-Before-Break (MBB) is used. No traffic interruption is noticed. If an auto-bandwidth attempt fails, there will be 5 retries and, if they all fail, the bandwidth remains unchanged. The next attempt may occur with the next trigger.

Retries follow the retry-limit (5 in this case), retry-timer (by default 30s), and exponential back-off timer, if enabled in MPLS.

Auto-bandwidth adjustment can be triggered in four different ways:

1. Periodic trigger

   The iLER determines at the end of each adjust-interval whether to attempt an auto-bandwidth adjustment.
2. Overflow or underflow trigger

   The measured bandwidth of an LSP has increased or decreased significantly since the start of the current adjust-interval. It may be preferable to adjust the bandwidth of the LSP after a number of overflow/underflow samples, rather than wait for the adjust-interval to end (default: 24 h).
3. Manual trigger

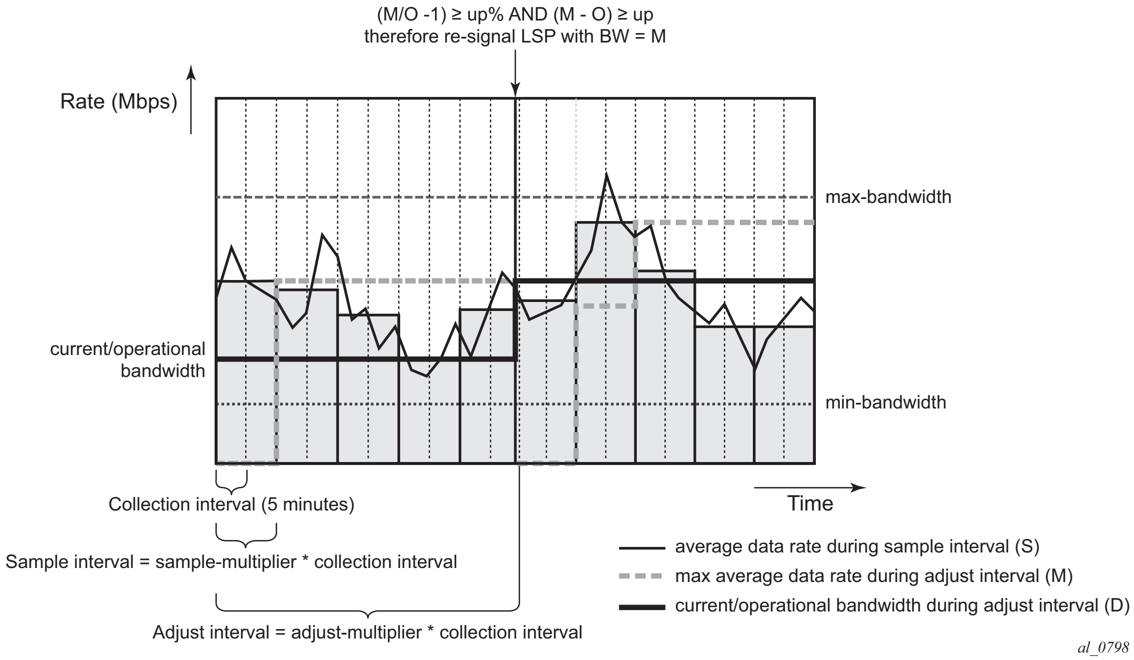
   An operator launches a **tools** command to trigger an auto-bandwidth adjustment.
4. Active path change

   The LSP has a primary and one or more secondary paths. When there is a change from the primary path to a secondary path without the LSP going down, an auto-bandwidth MBB is triggered. When the primary path becomes active again, another auto-bandwidth MBB is triggered.

Periodic Trigger

Figure 184 shows the different time intervals and bandwidths defined in the auto-bandwidth adjustment implementation. In this example, there will be an auto-bandwidth attempt when the adjust-interval elapses (periodic trigger). If the auto-bandwidth algorithm is met, the current bandwidth is increased. The parameters are explained after the figure.

Figure 184 **Auto-Bandwidth Adjustment Implementation**



The time intervals are:

- Collection interval in minutes. This is a global parameter to be set in an accounting policy. Range: 5 to 120 minutes. Default: 5 minutes.

  For this kind of record type, the minimum interval is 5 minutes. For policies using a record type of SAA or PM, the minimum is 1 minute.

```
*A:PE-1# configure log accounting-policy 10 collection-interval
  - collection-interval <minutes>
  - no collection-interval

 <minutes>            : [1..120]

*A:PE-1# configure log accounting-policy 10 collection-interval 1
MAJOR: LOG #1076 Except for policies using a record type of SAA or PM the minimum
interval is 5 mins
```

  – Sample interval: sample-multiplier * collection interval

  – Sample-multiplier is configurable globally in the MPLS context or per LSP. Default value: 1. In Figure 184, the sample multiplier equals 2 for a sample interval of 2 * 5 minutes = 10 minutes.

- Adjust-interval: adjust-multiplier * collection interval

  – Nokia recommends that the adjust-multiplier is an integer multiple of the sample-multiplier.

  – Adjust-multiplier is configurable globally in the MPLS context or per LSP. Default value: 288 (288 * 5 minutes = 1440 minutes = 24 h). In Figure 184, the adjust multiplier equals 10 for an adjust-interval of 10 * 5 minutes = 50 minutes.

```
*A:PE-1# configure router mpls auto-bandwidth-multipliers
  - auto-bandwidth-multipliers sample-multiplier <number1> adjust-multiplier <number2>
  - no auto-bandwidth-multipliers

 <number1>            : [1..511]
 <number2>            : [1..16383]


*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth multipliers
  - multipliers sample-multiplier <num1> adjust-multiplier <num2>
  - no multipliers

 <num1>               : [1..511]
 <num2>               : [1..16383]
```

The different bandwidths are:

- Minimum bandwidth: configured minimum bandwidth in Mbps that the auto-bandwidth adjustment can signal for the LSP.   Granularity: 1 Mbps. Default: 0 Mbps.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth min-bandwidth
  - min-bandwidth <mbps>
  - no min-bandwidth

 <mbps>               : [0..100000]
```

- Maximum bandwidth: configured maximum bandwidth in Mbps that the auto-bandwidth adjustment can signal for the LSP.   Granularity: 1 Mbps. Default: 100 Mbps.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth max-bandwidth
  - max-bandwidth <mbps>
  - no max-bandwidth

 <mbps>               : [0..100000]
```

- Current bandwidth or operational bandwidth (O): currently reserved bandwidth in Mbps for the LSP in the control plane. This is the operational bandwidth that is maintained in the Management Information Base (MIB) and is the bandwidth that will be auto-adjusted. Granularity: 1 Mbps.
- Sampled bandwidth (S): average data rate for the last sample interval.
- Measured bandwidth (M): maximum averaged (per sample interval) data rate in the current adjust-interval. The SR OS keeps track of the maximum average data rate of each LSP since the last reset of the adjust-count.

- Signaled bandwidth: bandwidth in Mbps that is provided to the CSPF algorithm and signaled in the RSVP SENDER_TSPEC and FLOWSPEC objects, when an auto-bandwidth adjustment is attempted. Granularity: 1 Mbps.

The other auto-bandwidth parameters for periodically triggered auto-bandwidth adjustment are:

- Up% (adjust-up in percent): minimum increase in bandwidth from current to measured bandwidth, expressed as a percentage of the current bandwidth. Default: 5%.
- Up (adjust-up bw): minimum increase in bandwidth as absolute bandwidth in Mbps. Up = measuredBW – currentBW. Granularity: 1 Mbps. Default: 0 Mbps.

```
**A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth adjust-up
  - adjust-up <percent> [bw <mbps>]
  - no adjust-up

 <percent>            : [0..100]
 <mbps>               : [0..100000]
```

- Down% (adjust-down in percent): minimum decrease in bandwidth from current to measured bandwidth, expressed as a percentage of the current bandwidth. Default: 5%.
- Down (adjust-down bw): minimum decrease in bandwidth as absolute bandwidth in Mbps. Down = currentBW – measuredBW. Granularity: 1 Mbps. Default: 0 Mbps.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth adjust-down
  - adjust-down <percent> [bw <mbps>]
  - no adjust-down

 <percent>            : [0..100]
 <mbps>               : [0..100000]
```

In Figure 184, the minimum and maximum bandwidths mark the bandwidth range where auto-bandwidth adjustments are allowed. The sample interval is two collection intervals long (2 * 5 minutes = 10 minutes). The adjust-interval is 10 collection intervals long (10 * 5 minutes = 50 minutes). Initially, the current bandwidth (O) equals the configured bandwidth for the primary path. It is good practice to give that same value to the minimum bandwidth for auto-bandwidth. The system doesn’t confirm this and these bandwidths are independent from each other.

In this example, the sampled bandwidth exceeds the current bit rate in most of the sample intervals. The maximum sampled bandwidth in the current adjust-interval corresponds to the measured bandwidth (M). When auto-bandwidth adjustment is triggered at the end of the adjust-interval, this measured bandwidth will be signaled and, after a successful adjustment, will be the new current bandwidth. After the auto-bandwidth adjustment, a new adjust-interval starts and the measured bandwidth is reset to 0. As long as the first sample interval of the new adjust-interval is not finished, the measured bandwidth equals 0 and auto-adjustment would be impossible even when triggered manually.

The auto-bandwidth attempt follows these rules:

- When measuredBW ≥ currentBW
  - if {(measuredBW / currentBW – 1) ≥ up%} &&{(measuredBW – currentBW) ≥ up
    then signaledBW = max{(min(measuredBW, maxBW)), minBW}
- When measuredBW ≤ currentBW
  - if {(1 – measuredBW/currentBW) ≥ down%} && {(currentBW – measuredBW) ≥ down}
    then signaledBW = min{(max(measuredBW, minBW)), maxBW}

CLI configured bandwidths have a granularity of 1 Mbps, while the threshold calculations with measured bandwidth are performed at full precision. This means that the signaled bandwidth in the RSVP message is rounded up to the nearest integer multiple of 1 Mbps.

Overflow/Underflow Trigger

Auto-bandwidth adjustment can also be triggered by overflow or underflow. When the bandwidth changes drastically, the bandwidth can be auto-adjusted after a number of consecutive overflow/underflow samples. In this case, there is no need to wait for the adjust-interval to end (by default: 24 h).

The parameters used in case of overflow are:

- Overflow sample: a sample interval counts as an overflow sample if the sampled bandwidth is higher than the current bandwidth by at least the configured overflow thresholds.
- Overflow-limit/overflow-count: an auto-bandwidth adjustment occurs after this number of consecutive overflow samples.
- Threshold%: minimum difference between sampled bandwidth and current bandwidth, expressed as a percentage of the current bandwidth.

- Threshold bw: minimum difference between sampled bandwidth and current bandwidth in Mbps. Default value: 0.

```
A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth overflow-limit
  - overflow-limit <number> threshold <percent> [bw <mbps>]
  - no overflow-limit

 <number>             : [1..10]
 <percent>            : [0..100]
 <mbps>               : [0..100000]
```

The rules for overflow-triggered auto-bandwidth adjustment are as follows:

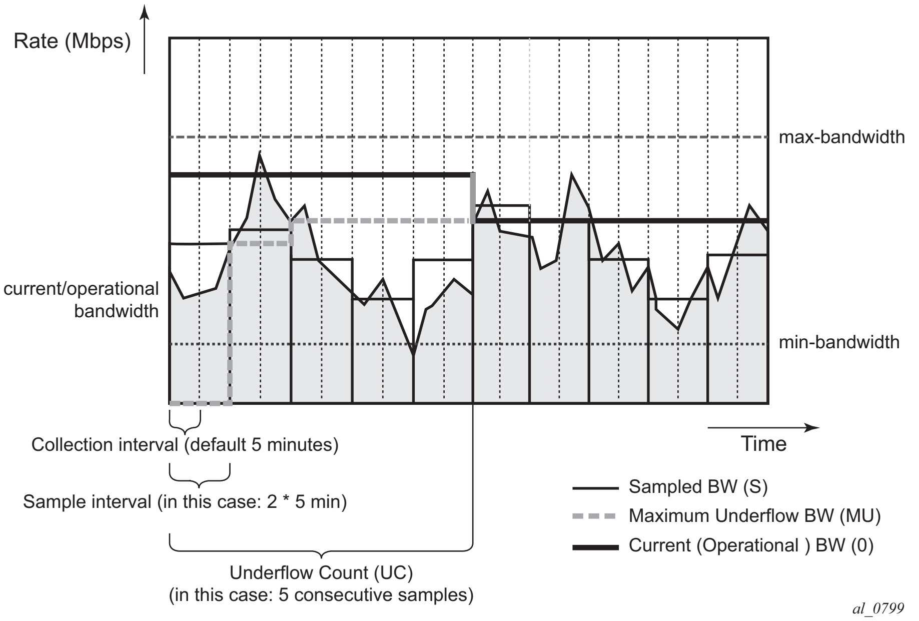
- Overflow sample: {(sampledBW / currentBW – 1) ≥ threshold%} && {(sampledBW – currentBW) ≥ thresholdBW}
- The signaled bandwidth will be:
  - if (measuredBW ≥ maxBW) then signaledBW = maxBW
  - if (measuredBW ≤ minBW) then signaledBW = minBW
  - else signaledBW = measuredBW

The parameters used in case of underflow are:

- Underflow sample: a sample interval counts as an underflow sample if the sampled bandwidth is lower than the current bandwidth by at least the configured underflow thresholds.
- Underflow-limit/underflow-count: an auto-bandwidth adjustment occurs after this number of consecutive underflow samples.
- Threshold%: minimum difference between current bandwidth and sampled bandwidth, expressed as a percentage of the current bandwidth.
- Threshold BW: minimum difference between current bandwidth and sampled bandwidth in Mbps. Default value: 0.
- Maximum underflow bandwidth (MU): maximum sampled bandwidth in the consecutive underflow samples.

```
A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth underflow-limit
  - underflow-limit <number> threshold <percent> [bw <mbps>]
  - no underflow-limit

 <number>             : [1..10]
 <percent>            : [0..100]
 <mbps>               : [0..100000]
```

Figure 185 **Underflow-Triggered Auto-Bandwidth Implementation**



In Figure 185, the adjust-interval is not displayed. It is assumed to be the default of 288 collection intervals (24 h). The figure only shows five consecutive underflow samples. The underflow-limit equals 5. In each of the samples, the sample bandwidth is below the underflow threshold. The maximum sampled bandwidth of these five samples corresponds to the maximum underflow bandwidth. This bandwidth will be signaled when auto-bandwidth adjustment is triggered because the underflow count is reached.

The rules for underflow-triggered auto-bandwidth adjustment are as follows:

- Underflow sample:
  - {(1 - sampledBW / currentBW) ≥ threshold%} && {(currentBW – sampledBW) ≥ thresholdBW}
- Underflow count/underflow limit: after that many consecutive underflow samples, an auto-bandwidth adjustment is triggered.
- The signaled bandwidth will be:
  - if (maxUnderflowBW ≥ maxBW) then signaledBW = maxBW
  - if (maxUnderflowBW ≤ minBW) then signaledBW = minBW else signaledBW = maxUnderflowBW

If the adjustment is successful, the sample counter within the adjust-interval is reset, along with other parameters, such as the maximum underflow bandwidth, the measured bandwidth, and the underflow count. The next adjust-interval will elapse in 24 h.

If the adjustment fails, there will be 5 retries. If they all fail, only the underflow count and the maximum underflow bandwidth are reset. The current adjust-interval continues.

Manual Trigger

Besides the periodic trigger and the overflow/underflow trigger, an operator can launch a **tools** command to trigger an auto-bandwidth adjustment.

```
A:PE-1# tools perform router mpls adjust-autobandwidth
```

This **tools** command can be launched with or without explicit LSP name. In the latter case, all active LSPs are attempted for auto-bandwidth.

```
A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
  - adjust-autobandwidth [lsp <lsp-name> [force [bandwidth <mbps>]]]

 <lsp-name>           : [64 chars max]
 <force>              : keyword
 <mbps>               : [0..100000]

A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
```

This command (without the keyword **force**) triggers a new auto-bandwidth calculation according to the rules of periodic triggered type. If the LSP already has the correct reserved bandwidth, the following message is returned.

```
A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
MINOR: CLI lsp LSP-PE-1-PE-2 active path is already at the requested value 12 Mbps.
```

If the keyword **force** is added without a specific value for the bandwidth, there is no threshold checking. The bandwidth can also be adjusted if the difference in bandwidth is below the thresholds. The granularity remains 1 Mbps.

```
A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2" force
```

The rules for the signaled bandwidth are unchanged:

- if (measuredBW ≥ maxBW) then signaledBW = maxBW
- if (measurealedBdBW ≤ minBW) then signaledBW = minBW
  else signW = measuredBW

If the keyword **force** with **bandwidth** (in Mbps) option is given, the signaled bandwidth is set to this configured bandwidth, even if it is a value below the minimum or higher than the maximum bandwidth.

```
A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2" force
bandwidth 30
```

After a manually triggered auto-bandwidth MBB, no counters are reset. The ongoing adjust-interval is not aborted.

A clear command resets all counters and timers associated with auto-bandwidth adjustment on a specified LSP.

```
A:PE-1# clear router mpls lsp-autobandwidth "LSP-PE-1-PE-2"
```

Passive Monitoring

The system offers the option to measure the bandwidth of an LSP without taking any action to adjust the bandwidth reservation.

```
A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth monitor-bandwidth
```

Auto-Bandwidth Based on Forwarding Class

From 11.0.R4 onward, the bandwidth can be calculated as a weighted sum of all the traffic in the eight forwarding classes. By default, all forwarding classes have the same weight: 100%, but that sampling weight is configurable.

```
A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth fc
  - fc <fc-name> sampling-weight <sampling-weight>
  - no fc <fc-name>

 <fc-name>            : be|l2|af|l1|h2|ef|h1|nc
 <sampling-weight>    : [0..100]
```

Active Path Change

Auto-bandwidth adjustment is also supported on LSPs that have secondary paths. If the secondary path is standby, an auto-bandwidth MBB can be triggered when the active path changes from primary to secondary. The secondary/standby path is only initialized at its configured bandwidth when it is established, and the bandwidth is adjusted only when it becomes active. This happens when the primary path goes down or becomes degraded. When another path becomes active, the bandwidth used to signal the auto-bandwidth MBB will be the operational bandwidth of the previous path.

The definition for current bandwidth is modified for this feature:

- Current bandwidth: last known reserved bandwidth for the LSP. This may be for a different path than the active one.

  Auto-bandwidth adjustment will only take place on the active path. When the active path changes, the current bandwidth is updated to the operational bandwidth of the new active path.
- For a secondary path that is signaled as standby, if the active path for an LSP changes without the LSP going down, an auto-bandwidth MBB is triggered on the new active path. The signaled bandwidth is the operational bandwidth of the previous path. The reserved bandwidth of the new active path will be its configured bandwidth until the MBB succeeds.
- For a secondary path where the active path goes down, the LSP will go down temporarily until the secondary path is set up. When the LSP goes down, all statistics and counters are cleared, so the previous path operational bandwidth is lost. There will be no immediate bandwidth adjustment on the secondary path.

The following rules apply to determine the signaled bandwidth of the new active path.

- For a path that is operationally down, signaledBW = configuredBW.
- For the first 5 MBB attempts on the path that just became active,

  signaledBW = currentBW (operational bandwidth of the previous path).

For the remaining MBB attempts, signaledBW = operationalBW.

- For all MBBs other than auto-bandwidth MBB on the active path,

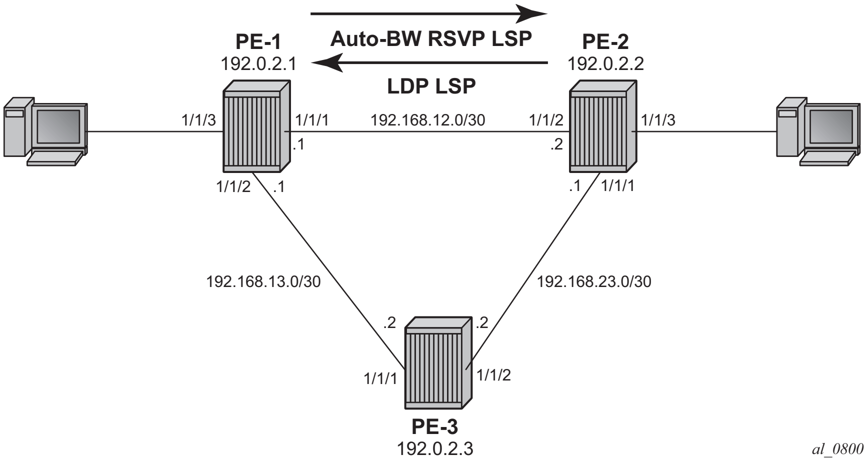
  MBB signaledBW = operationalBW.
- For an MBB on the inactive (standby) path, MBB signaledBW = configuredBW.

When the system reverts from a secondary standby path to the primary path, a Delayed Retry MBB is attempted to bring the bandwidth on the standby path back to the configured bandwidth. MBB is attempted once, and if it fails, the standby is torn down. A Delayed Retry MBB has the highest priority among MBBs, so it will take precedence over any other MBB in progress on the standby path, such as configuration change or pre-emption.

Configuration

Figure 186 shows the example setup. The focus will be on the RSVP LSP from PE-1 to PE-2.

Figure 186 **Example Setup for Auto-Bandwidth Point-to-Point LSPs**



Base Configuration

The cards, MDAs and ports need to be configured.

Configure the interfaces on all nodes. For PE-1:

```
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-3"
            address 192.168.13.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
```

As an IGP, OSPF or IS-IS can be used. In this example, OSPF is configured. Traffic engineering should be enabled. For PE-1:

```
configure
    router
        ospf
            traffic-engineering
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-1-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-1-PE-3"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
```

Optionally, enable LDP on all interfaces. Link-layer LDP is not a prerequisite for using auto-bandwidth RSVP LSPs. In this example, the SDP from PE-2 to PE-1 uses an LDP LSP, but it could have been an RSVP-TE LSP instead.   For PE-2:

```
configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-PE-1-PE-3" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

Enable MPLS and RSVP on all nodes and add all interfaces to the MPLS context. They will automatically be added to the RSVP context. For PE-1:

```
configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
```

```
                interface "int-PE-1-PE-3"
                exit
                no shutdown
            exit
            rsvp no shutdown
```

Configure a path with no explicitly defined hops and LSP LSP-PE-1-PE-2 on PE-1:

```
configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "loose"
                exit
                no shutdown
            exit
```

In the example, traffic needs to be injected into the LSP tunnel. For that, a VPLS service is created. For PE-1, an SDP using the RSVP LSP to PE-2 is created.

```
configure
    service
        sdp 212 mpls create
            description "SDP-PE-1-PE-2-overRSVP-TE"
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
```

On PE-2, an SDP using LDP is created toward PE-1.

```
configure
    service
        sdp 121 mpls create
            description "SDP-PE-2-PE-1-overLDP"
            far-end 192.0.2.1
            ldp
            no shutdown
        exit
```

On PE-1 and PE-2, a VPLS is created. For PE-1:

```
configure
    service
        vpls 100 name "VPLS 100" customer 1 create
            sap 1/1/3 create
            exit
            spoke-sdp 212:100 create
            exit
            no shutdown
            exit
```

The configuration on PE-2 is similar.

Pre-requisites for Auto-Bandwidth LSP Configuration

Enable Constrained Shortest Path First (CSPF) on the LSP by adding the keyword **cspf**.

```
configure router mpls lsp “LSP-PE-1-PE-2” cspf
```

The bandwidth of the LSP will be adjusted in a Make-Before-Break (MBB) manner. Enable MBB on the LSP by adding the keyword **adaptive** to the primary path.

```
configure router mpls lsp “LSP-PE-1-PE-2” primary “loose” adaptive
```

Enter a value for the bandwidth in Mbps for the primary path. It is good practice to configure the same value as for the minimum bandwidth in the auto-bandwidth settings.

```
configure router mpls lsp “LSP-PE-1-PE-2” primary “loose” bandwidth 2
```

Auto-Bandwidth LSP Configuration

MPLS auto-bandwidth adjustment allows the ingress LER to dynamically adjust the bandwidth of an RSVP tunnel based on active measurements of the traffic rate into the tunnel. Therefore, LSP egress statistics need to be enabled on the iLER.

Auto-bandwidth adjustment requires an accounting policy to be defined and operational. The accounting policy specifies the collection interval for LSP statistics collection, which is fundamental to the auto-bandwidth algorithm. The minimum interval for this type of collection is 5 minutes, which is the default value.

```
configure
    log
        accounting-policy 10
            record combined-mpls-lsp-egress
            to no-file
            no shutdown
       exit
```

An accounting policy of record type **combined-mpls-lsp-egress** doesn’t need a reference to a specific file ID.

From the moment auto-bandwidth is enabled with an LSP context, the record combined-mpls-lsp-egress inside the accounting policy will also take bandwidth measurements.

```
configure log accounting-policy 10 to no-file
```

When the **no-file** is configured, no LSP statistics are stored anymore. The MPLS auto-bandwidth feature retrieves it LSP statistics information directly from the statistics module.

However, the accounting policy can reference a file and, therefore, a CF card. An additional CF card may be required in each node as a storage location.

```
configure
    log
        file-id 66
            location cf1:
            rollover 15 retention 1
        exit
        accounting-policy 66
            record combined-mpls-lsp-egress
            to file 66
            no shutdown
        exit
```

In the remainder of the example, the accounting policy will reference to no-file.

After the accounting policy has been created, egress statistics can be enabled on the LSP.

```
configure
    router
        mpls
            lsp "LSP-PE-1-PE-2"
                egress-statistics
                    no shutdown
                    collect-stats
                    accounting-policy 10
                exit
```

The system does not verify whether egress statistics have been enabled on the LSP. When a user configures auto-bandwidth adjustment, but without enabling egress statistics, no auto-bandwidth measurements and adjustments are performed. The operational state of auto-bandwidth (AB OpState) is down.

Enable auto-bandwidth with default settings by adding the keyword auto-bandwidth to the LSP.

```
configure router mpls lsp LSP-PE-1-PE-2 auto-bandwidth
```

The actual values are shown in the following output. They are explained after the output.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 0 Mbps                    Auto BW Max          : 100000 Mbps
AB Up Thresh    : 5 percent                 AB Down Thresh       : 5 percent
AB Up BW        : 0 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 2 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288+                      AB Samp Mul          : 1+
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : n/a                       AB Next Adj          : 1440 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 0 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : none                      AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

The plus sign (+) indicates that the value is inherited from the global MPLS settings (AB Adj Mul: 288+ and AB Samp Mul: 1+). The sample-multiplier and the adjust-multiplier can both be configured globally in the MPLS context or overruled by the settings per LSP. In this example, nothing has been configured in the MPLS context or in the LSP. Therefore, the default values as defined in the MPLS context are applicable.

Auto-Bandwidth – Periodic Trigger (Normal)

The default collection interval is 5 minutes. The sample-multiplier is 1, by default. The sample interval equals 1 * 5 minutes = 5 minutes. The adjust-multiplier is 288, by default 288. The adjustment interval equals 288 * 5 minutes = 1440 minutes (24 hours).

The auto-bandwidth settings for the LSP are modified as follows:

```
configure
    router
        mpls
            lsp LSP-PE-1-PE-2
                auto-bandwidth
                    multipliers sample-multiplier 1 adjust-multiplier 3
                    adjust-up 10 bw 1
                    adjust-down 5 bw 0       ## default
                    max-bandwidth 20
                    min-bandwidth 2
                exit
```

In the example, the bandwidth of the LSP can be auto-adjusted every 15 minutes (after 3 intervals of 5 minutes). For a decrease in bandwidth, the default settings apply and no explicit command is required in this example. That means that the current bandwidth will be reduced when the difference in bandwidth is at least 5%. There is no absolute decrease (in Mbps) defined. For an increase in bandwidth, there will only be an adjustment when the increase is at least 10% and at least 1 Mbps. The minimum bandwidth is 2 Mbps. This equals the bandwidth set in the path in the LSP (recommended). The maximum bandwidth equals 20 Mbps. The system will not compare the minimum or maximum bandwidth to the configured bandwidth for the path.

Display the actual auto-bandwidth data after 5, 10, and 15 minutes.

There are different bandwidths displayed:

- The AB Curr BW is the operational bandwidth during the adjustment interval. It is initially the configured bandwidth of the path in the LSP, but it can be auto-adjusted. This bandwidth is taken into account in the control plane when an LSP is set up or modified in case of MBB. The real data rate in the data plane may exceed this operational bandwidth.
- The ABMaxAvgRt is the measured bandwidth, meaning the maximum averaged bandwidth (calculated every sample interval of 5 minutes) in the adjustment interval of 15 minutes (AB Adj Time: 15 Min).
- The AB Lst AvgRt is the sampled bandwidth, averaged over the latest sample interval of 5 minutes (AB Samp Intv: 5 Mins).

After 5 minutes, one collection interval has elapsed within the adjust-interval (AB Adj Cnt = 1) and the next adjustment time is in 10 minutes (AB Next Adj = 10 Min). The current bandwidth equals 2 Mbps, while the measured and the sampled bandwidths are much higher: 12 Mbps.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
```

```
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 2 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 3                         AB Samp Mul          : 1
AB Adj Time     : 15 Mins                   AB Samp Time         : 5 Mins
AB Adj Cnt      : 1                         AB Samp Cnt          : 0
AB Last Adj     : n/a                       AB Next Adj          : 10 Mins
ABMaxAvgRt      : 12 Mbps                   AB Lst AvgRt         : 12 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : none                      AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

After 10 minutes, another collection interval has elapsed in the adjust-interval (**AB Adj Cnt** = 2) and the next adjustment time is in 5 minutes (**AB Next Adj** = 5 Min).

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 2 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 3                         AB Samp Mul          : 1
AB Adj Time     : 15 Mins                   AB Samp Time         : 5 Mins
AB Adj Cnt      : 2                         AB Samp Cnt          : 0
AB Last Adj     : n/a                       AB Next Adj          : 5 Mins
ABMaxAvgRt      : 12 Mbps                   AB Lst AvgRt         : 12 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
```

```
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : none                      AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

After 15 minutes, auto-bandwidth adjustment occurs. **AB Adj Cause** is normal for periodically triggered adjustments. The next adjustment interval will elapse in 15 minutes. The measured bandwidth **ABMaxAvgRt** is reset to 0 after a successful adjustment.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 12 Mbps                   AB Samp Intv         : 5 Mins
AB Adj Mul      : 3                         AB Samp Mul          : 1
AB Adj Time     : 15 Mins                   AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 09:26:45       AB Next Adj          : 15 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 12 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : normal                    AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

The periodic trigger type rules for auto-bandwidth are:

- When measuredBW ≥ currentBW

- if {(measuredBW / currentBW – 1) ≥ up%} &&{(measuredBW – currentBW) ≥ up

  then signaledBW = max{(min(measuredBW, maxBW)), minBW}

In this case, the measuredBW (13 Mbps) is greater than the currentBW (2 Mbps). The increase is at least 10% (up%) and at least 1 Mbps (up). The bandwidth will be adjusted. The new bandwidth that will be signaled is calculated as follows:

```
signaledBW= max{(min(measuredBW, maxBW)), minBW}
signaledBW= max {(min (12 Mbps, 20 Mbps)), 2 Mbps}
signaledBW= max {12 Mbps, 2 Mbps}
signaledBW= 12 Mbps
```

Whenever an auto-bandwidth adjustment is performed, a message is stored in log 99.

```
*A:PE-1# show log log-id 99 application "mpls"

===============================================================================
Event Log 99
===============================================================================
Description : Default System Log
Memory Log contents  [size=500   next event=62  (not wrapped)]

91 2018/09/21 09:26:46.388 UTC WARNING: MPLS #2014 Base VR 1:
"LSP path LSP-PE-1-PE-2::loose resignaled as result of autoBandwidth MBB"
```

When the maximum bandwidth is modified to a value that is lower than the current bandwidth, an adjustment occurs at the end of the adjustment interval.

```
*A:PE-1# configure router mpls lsp LSP-PE-1-PE-2 auto-bandwidth max-bandwidth 10
```

The current bandwidth will be reduced to 10 Mbps (for a measured bandwidth of 12 Mbps).

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                    AB OpState          : Up
Auto BW Min     : 2 Mbps                     Auto BW Max         : 10 Mbps
AB Up Thresh    : 10 percent                 AB Down Thresh      : 5 percent
AB Up BW        : 1 Mbps                     AB Down BW          : 0 Mbps
AB Curr BW      : 10 Mbps                    AB Samp Intv        : 5 Mins
AB Adj Mul      : 3                          AB Samp Mul         : 1
AB Adj Time     : 15 Mins                   AB Samp Time         : 5 Mins
AB Adj Cnt      : 1                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 09:41:45       AB Next Adj          : 10 Mins
ABMaxAvgRt      : 13 Mbps                   AB Lst AvgRt         : 13 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : normal                    AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================*
```

Auto-Bandwidth - Passive Monitoring

When passive monitoring is enabled, no automatic bandwidth adjustments occurs. When the maximum bandwidth is again raised to 20 Mbps, the bandwidth will not be auto-adjusted even if the measured bandwidth is high enough.

```
configure router mpls lsp LSP-PE-1-PE-2 auto-bandwidth max-bandwidth 20
configure router mpls lsp LSP-PE-1-PE-2 auto-bandwidth monitor-bandwidth
```

The system monitors the bandwidth, but without taking action at the end of the adjust-interval.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 10 Mbps                   AB Samp Intv         : 5 Mins
AB Adj Mul      : 3                         AB Samp Mul          : 1
AB Adj Time     : 15 Mins                   AB Samp Time         : 5 Mins
AB Adj Cnt      : 2                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 09:41:45       AB Next Adj          : 5 Mins
ABMaxAvgRt      : 13 Mbps                   AB Lst AvgRt         : 13 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt        : 0
```

```
ABUndflThrs     : 0 percent                 AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : normal                    AB Monitor BW        : True
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

The value for the parameter **AB Monitor BW** is True

For the remainder of the chapter, there is no passive monitoring. The settings are restored to normal:

```
configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth no monitor-bandwidth
```

Auto-Bandwidth – Overflow and Underflow Trigger Type

With default settings, the adjustment interval is 24 hours. If the bandwidth changes significantly since the start of the current adjust-interval, overflow and underflow triggers can be used. This will speed up the auto-bandwidth adjustment.

Stop auto-bandwidth in order to force an MBB attempt toward the configured primary path bandwidth (2Mbps in this example).

```
configure router mpls lsp "LSP-PE-1-PE-2" no auto-bandwidth
```

Check the operational bandwidth of the LSP.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
LSP Type        : RegularLsp                LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.1                 To                   : 192.0.2.2
Adm State       : Up                        Oper State           : Up

---snip---
```

```
Primary(a)      : loose
                                            Up Time              : 0d 01:53:02
Bandwidth       : 2 Mbps
===============================================================================
*A:PE-1#
```

Enable auto-bandwidth with similar settings as before and add overflow and underflow triggers. The multipliers are default. Therefore, a periodically triggered auto-adjustment will only take place once every 24 hours.

```
configure
    router
        mpls
            lsp LSP-PE-1-PE-2
                auto-bandwidth
                    multipliers sample-multiplier 1 adjust-multiplier 288
                    adjust-up 10 bw 1
                    max-bandwidth 20
                    min-bandwidth 2
                    overflow-limit 1 threshold 10 bw 2
                    underflow-limit 3 threshold 10 bw 2
                exit
```

The overflow count indicates the number of consecutive times that the overflow condition is detected at the end of a sample interval. Auto-bandwidth adjustment occurs after that number of overflow samples is reached, in this case, after the first overflow sample (overflow-limit = 1). The conditions for an overflow sample are:

{(sampledBW / currentBW – 1) ≥ threshold%} && {(sampledBW – currentBW) ≥ thresholdBW}

{(11 Mbps/2Mbps – 1) ≥ 0,1} && {(11Mbps – 2Mbps) ≥ 2Mbps}

The signaled bandwidth will be:

- if (measuredBW ≥ maxBW) then signaledBW = maxBW
- if (measuredBW ≤ minBW) then signaledBW = minBW
  else signaledBW = measuredBW
- if (11 Mbps ≥ 20 Mbps) then signaledBW = 20 Mbps;
- if (11 Mbps ≤ 2 Mbps) then signaledBW = 2 Mbps;
  else signaledBW = 11 Mbps

Display the auto-bandwidth data. The **AB Adj Cause** is now **overflow**. The overflow limit is the configured value of 1 (**AB Ovfl Lmt**). The overflow count has been reset (**AB Ovfl Cnt = 0**) after the auto-bandwidth was adjusted, along with the **ABMaxAvgRt**. This is the start of a new adjust-interval of 24 hours.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth
```

```
===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW          : Enabled                 AB OpState          : Up
Auto BW Min      : 2 Mbps                  Auto BW Max         : 20 Mbps
AB Up Thresh     : 10 percent              AB Down Thresh      : 5 percent
AB Up BW         : 1 Mbps                  AB Down BW          : 0 Mbps
AB Curr BW       : 12 Mbps                 AB Samp Intv        : 5 Mins
AB Adj Mul       : 288                     AB Samp Mul         : 1
AB Adj Time      : 1440 Mins               AB Samp Time        : 5 Mins
AB Adj Cnt       : 0                       AB Samp Cnt         : 0
AB Last Adj      : 09/21/2018 11:16:45     AB Next Adj         : 1440 Mins
ABMaxAvgRt       : 0 Mbps                  AB Lst AvgRt        : 12 Mbps
AB Ovfl Lmt      : 1                       AB Ovfl Cnt         : 0
ABOvflThres      : 10 percent              AB Ovfl BW          : 2 Mbps
AB UndflLmt      : 3                       AB Undrfl Cnt       : 0
ABUndflThrs      : 10 percent              AB Undrfl BW        : 2 Mbps
ABMaxUndflBW     : 0 Mbps
AB Adj Cause     : overflow                AB Monitor BW       : False
Be Weight        : 100 percent             Af Weight           : 100 percent
L1 Weight        : 100 percent             L2 Weight           : 100 percent
Nc Weight        : 100 percent             Ef Weight           : 100 percent
H1 Weight        : 100 percent             H2 Weight           : 100 percent
===============================================================================
*A:PE-1#
```

In the following example, the current bandwidth is 12 Mbps, but the bandwidth dropped to 7 Mbps and the conditions for underflow are met. At the end of a sample interval, the sampled bandwidth is reduced by at least 10% and at least 2 Mbps, and this becomes an underflow sample. The conditions for an underflow sample are:

{(1 - sampledBW / currentBW) ≥ threshold%} && {(currentBW – sampledBW) ≥ thresholdBW}

{(1 - 7 Mbps / 12 Mbps) ≥ 0,1} && {(12Mbps – 7Mbps) ≥ 2Mbps}

The underflow limit equals 3, so an auto-bandwidth adjustment can only take place after the third consecutive underflow sample. The new bandwidth will equal the **maximum sampled underflow bandwidth (ABMaxUndflBW)**. This is the maximum sampled bandwidth in the three consecutive underflow samples.

The signaled bandwidth will be:

- if (maxUnderflowBW ≥ maxBW) then signaledBW = maxBW
- if (maxUnderflowBW≤ minBW) then signaledBW = minBW
  else signaledBW = maxUnderflowBW

- if (7 Mbps≥ 20 Mbps) then signaledBW = 20 Mbps;
- if (7 Mbps ≤ 2 Mbps) then signaledBW = 2 Mbps;

  else signaledBW = 7 Mbps

The following output shows the auto-bandwidth data after two consecutive underflow samples (**AB Underfl Cnt: 2**). The maximum sampled underflow bandwidth equals 7 Mbps. No bandwidth adaptation can take place until there are three consecutive underflow samples (**AB UndflLmt: 3**).

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                    AB OpState           : Up
Auto BW Min     : 2 Mbps                     Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                 AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                     AB Down BW           : 0 Mbps
AB Curr BW      : 12 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                        AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                  AB Samp Time         : 5 Mins
AB Adj Cnt      : 2                          AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 11:16:45        AB Next Adj          : 1430 Mins
ABMaxAvgRt      : 7 Mbps                     AB Lst AvgRt         : 5 Mbps
AB Ovfl Lmt     : 1                          AB Ovfl Cnt          : 0
ABOvflThres     : 10 percent                 AB Ovfl BW           : 2 Mbps
AB UndflLmt     : 3                          AB Undrfl Cnt        : 2
ABUndflThrs     : 10 percent                 AB Undrfl BW         : 2 Mbps
ABMaxUndflBW    : 7 Mbps
AB Adj Cause    : overflow                   AB Monitor BW        : False
Be Weight       : 100 percent                Af Weight            : 100 percent
L1 Weight       : 100 percent                L2 Weight            : 100 percent
Nc Weight       : 100 percent                Ef Weight            : 100 percent
H1 Weight       : 100 percent                H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

After a successful auto-bandwidth adjustment, the **ABMaxUndflBW** is reset, along with the **AB Adj Cnt**, **AB Underfl Cnt**, and **ABMaxAvgRt**.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
==============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 7 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 11:31:45       AB Next Adj          : 1440 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 5 Mbps
AB Ovfl Lmt     : 1                         AB Ovfl Cnt          : 0
ABOvflThres     : 10 percent                AB Ovfl BW           : 2 Mbps
AB UndflLmt     : 3                         AB Undrfl Cnt        : 0
ABUndflThrs     : 10 percent                AB Undrfl BW         : 2 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : underflow                 AB Monitor BW        : False
Be Weight       : 100 percent               Af Weight            : 100 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

If the overload or underload trigger condition is met at the end of an adjust-interval, the auto-bandwidth adjustment is normal, based on the periodic trigger. Overflow and underflow auto-bandwidth adjustments only take place when the adjust-interval is not yet completed.

Auto-Bandwidth – Manual Trigger Type

As before, auto-bandwidth adjustment is disabled to revert to a bandwidth of 2 Mbps, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" no auto-bandwidth
```

Afterward, auto-bandwidth adjustment is configured on PE-1, as follows:

```
*A:PE-1#
configure
    router
        mpls
            lsp LSP-PE-1-PE-2
                auto-bandwidth
                    multipliers sample-multiplier 1 adjust-multiplier 288
                    adjust-up 10 bw 1
                    max-bandwidth 20
                    min-bandwidth 2
                exit
            exit
```

The auto-bandwidth adjustment can be triggered manually at all times by the following command (with or without the keyword force).

```
tools perform router mpls adjust-autobandwidth
tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2" force
```

When no specific LSP is referred to, auto-bandwidth will be attempted on all LSPs. If the LSP already has the requested bandwidth, the following output is returned.

```
*A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
MINOR: CLI No Thresholds crossed for lsp LSP-PE-1-PE-2.
```

By adding the keyword **force**, there is no check whether the thresholds are crossed. However, the granularity is 1 Mbps. In this case, it is not possible to signal a bandwidth that is at least 1 Mbps different, so the following error message is returned.

```
*A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2" force
MINOR: CLI lsp LSP-PE-1-PE-2 active path is already at the requested value 13 Mbps.
```

If the first sample interval has not yet expired, the following error message is returned.

```
*A:PE-1# tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2"
MINOR: CLI No Autobandwidth Averages computed for lsp LSP-PE-1-PE-2.
```

If the tools command is launched after the first sample interval has expired (**ABMaxAvgRt** is filled in), the bandwidth can be adjusted. The **AB Adj Cause** is manual.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 13 Mbps                   AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 1                         AB Samp Cnt           : 0
AB Last Adj     : 09/21/2018 12:25:37       AB Next Adj           : 1435 Mins
ABMaxAvgRt      : 13 Mbps                   AB Lst AvgRt          : 13 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt           : 0
ABOvflThres     : 0 percent                 AB Ovfl BW            : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt         : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW          : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : manual                    AB Monitor BW         : False
Be Weight       : 100 percent               Af Weight             : 100 percent
L1 Weight       : 100 percent               L2 Weight             : 100 percent
Nc Weight       : 100 percent               Ef Weight             : 100 percent
H1 Weight       : 100 percent               H2 Weight             : 100 percent
===============================================================================
*A:PE-1#
```

The counters are not reset after a manually triggered auto-bandwidth adjustment. The adjust-interval is not interrupted, the measured bandwidth and the maximum underflow bandwidth are not reset, and the overflow and underflow count are not reset.

Launch the tools command with the keyword **force** and a bandwidth value. This will set the current bandwidth to this value, even if the value is not within the allowed range between the minimum and maximum bandwidth.

```
tools perform router mpls adjust-autobandwidth lsp "LSP-PE-1-PE-2" force bandwidth 30
```

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth
===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState            : Up
Auto BW Min     : 2 Mbps                    Auto BW Max           : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh        : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW            : 0 Mbps
AB Curr BW      : 30 Mbps                   AB Samp Intv          : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul           : 1
AB Adj Time     : 1440 Mins                 AB Samp Time          : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt           : 0
AB Last Adj     : 09/21/2018 12:31:29       AB Next Adj           : 1440 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt          : 0 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt           : 0
ABOvflThres     : 0 percent                 AB Ovfl BW            : 0 Mbps
AB UndflLmt     : 0                         AB Undrfl Cnt         : 0
ABUndflThrs     : 0 percent                 AB Undrfl BW          : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : manual                    AB Monitor BW         : False
Be Weight       : 100 percent               Af Weight             : 100 percent
L1 Weight       : 100 percent               L2 Weight             : 100 percent
```

```
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

Manually triggered auto-bandwidth adjustments are also performed using MBB procedures.

Auto-Bandwidth Adjustment Based on Forwarding Class Subset

With the configuration applied so far, there is no distinction between traffic from different forwarding classes (FCs). The average data rate is the sum of the traffic from all eight FCs. From 11.0.R4 onward, it is possible to provide a sampling weight for each Forwarding Class (FC) for each auto-bandwidth LSP. The average data rate is now the weighted sum of the traffic from all FCs.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth fc
  - fc <fc-name> sampling-weight <sampling-weight>
  - no fc <fc-name>

 <fc-name>            : be|l2|af|l1|h2|ef|h1|nc
 <sampling-weight>    : [0..100]


configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth fc be sampling-weight 50
configure router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth fc af sampling-weight 80


*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 30 Mbps                   AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 12:31:29       AB Next Adj          : 1435 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 0 Mbps
AB Ovfl Lmt     : 0                         AB Ovfl Cnt          : 0
ABOvflThres     : 0 percent                 AB Ovfl BW           : 0 Mbps
```

```
AB UndflLmt     : 0                          AB Undrfl Cnt        : 0
ABUndflThrs     : 0 percent                  AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : manual                     AB Monitor BW        : False
Be Weight       : 50 percent                 Af Weight            : 80 percent
L1 Weight       : 100 percent                L2 Weight            : 100 percent
Nc Weight       : 100 percent                Ef Weight            : 100 percent
H1 Weight       : 100 percent                H2 Weight            : 100 percent
==============================================================================
*A:PE-1#
```

The sampling-weight values can be changed while auto-bandwidth is enabled. The auto-bandwidth algorithm will be reset on the LSP at the end of the current collection interval. At that time, the current bandwidth will not be adjusted and the following counters will be reset to 0: sample count, adjust count, overflow count, underflow count, max average data rate, and max average underflow data rate.

Auto-Bandwidth on LSPs with Secondary Paths

When the active path goes down or becomes degraded, the bandwidth used to signal the auto-bandwidth MBB will be the operational bandwidth of the previous active path. The parameter current-bandwidth requires a modified definition:

Current-bandwidth — The last known reserved bandwidth for the LSP (this may be for a different path than the active one).

When the active path changes, the current bandwidth is updated to the operational bandwidth of the new active path. While the auto-bandwidth MBB on the active path is in progress, a statistics sample might be triggered because the intervals aren’t reset when the active path changes. It is possible that an auto-adjustment is needed. The in-progress auto-bandwidth MBB will be restarted with retry attempts to 0 and signaled bandwidth equal to the new measured bandwidth. If after five attempts, auto-bandwidth MBB fails, the current bandwidth and secondary **oper-bw** remain unchanged.

For a secondary/standby path, if the active path changes without the LSP going down, an auto-bandwidth MBB is triggered for the new active path. The bandwidth used to signal the MBB is the operational bandwidth of the previous active path (current bandwidth).

If the primary path is not currently active, but it has not gone down, then any MBB should use the configured bandwidth for the primary path.

Create two new strict paths and assign them to the LSP. The primary path is the direct strict path from PE-1 to PE-2. There are two secondary paths: **path-PE-1-PE-3-PE-2** and **loose**. The first one is standby, the latter is not.

```
configure
    router
        mpls
            path path-PE-1-PE-2
                hop 10 192.0.2.2 strict
                no shutdown
            exit
            path path-PE-1-PE-3-PE-2
                hop 10 192.0.2.3 strict
                hop 20 192.0.2.2 strict
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                cspf
                fast-reroute facility
                    no node-protect
                exit
                primary loose shutdown
                no primary loose
                primary path-PE-1-PE-2
                    adaptive
                    bandwidth 2
                exit
                secondary path-PE-1-PE-3-PE-2
                    adaptive
                    bandwidth 2
                    standby
                exit
                secondary loose
                    adaptive
                    bandwidth 2
                exit
                no shutdown
            exit
```

Stop the auto-bandwidth MBB to have the current bandwidth equal to the bandwidth configured for the primary path (2 Mbps).

```
configure router mpls lsp "LSP-PE-1-PE-2" no auto-bandwidth
```

Configure auto-bandwidth with the following settings:

```
configure
    router
        mpls
            lsp "LSP-PE-1-PE-2"
                auto-bandwidth
                    multipliers sample-multiplier 1 adjust-multiplier 288
                    adjust-up 10 bw 1
                    max-bandwidth 20
                    min-bandwidth 2
                    overflow-limit 2 threshold 10
                    underflow-limit 3 threshold 10
                    fc be sampling-weight 50
                    fc af sampling-weight 80
                exit
```

Initially, the current bandwidth is the configured bandwidth of the primary path: 2 Mbps, but in case of overflow, it will be increased after two overflow samples (10 minutes).

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

================================================================================
MPLS LSP (Auto Bandwidth)
================================================================================
Legend :
    + - Inherited
================================================================================
--------------------------------------------------------------------------------
Type : Originating
--------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 6 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 12:51:45       AB Next Adj          : 1440 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 6 Mbps
AB Ovfl Lmt     : 2                         AB Ovfl Cnt          : 0
ABOvflThres     : 10 percent                AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 3                         AB Undrfl Cnt        : 0
ABUndflThrs     : 10 percent                AB Undrfl BW         : 0 Mbps
ABMaxUndflBW: 0 Mbps
AB Adj Cause: overflow                      AB Monitor BW  : False
Be Weight       : 50 percent                Af Weight            : 80 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
================================================================================
*A:PE-1#
```

Shut down port 1/1/1 on PE-1 to initiate a failure on the primary path.

```
*A:PE-1# configure port 1/1/1 shutdown
```

Verify that the secondary/standby path is now active.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" activepath

================================================================================
MPLS LSP: LSP-PE-1-PE-2 (active paths)
================================================================================
Legend :
    #  - Manually switched path
    #F - Manually forced switched path
================================================================================
LSP Name     : LSP-PE-1-PE-2
LSP Id       : 53308
```

```
Path Name    : path-PE-1-PE-3-PE-2
Active Path  : Standby
To           : 192.0.2.2                         LSP Type     : dynamic

===============================================================================
*A:PE-1#
```

Check the auto-bandwidth data on the LSP. The current bandwidth for the LSP is the same as it used to be for the primary path. **AB Adj Cause** = activePathChange.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 6 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 0                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 12:52:50       AB Next Adj          : 1440 Mins
ABMaxAvgRt      : 0 Mbps                    AB Lst AvgRt         : 6 Mbps
AB Ovfl Lmt     : 2                         AB Ovfl Cnt          : 0
ABOvflThres     : 10 percent                AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 3                         AB Undrfl Cnt        : 0
ABUndflThrs     : 10 percent                AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : activePathChange          AB Monitor BW        : False
Be Weight       : 50 percent                Af Weight            : 80 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

The original situation is restored.

```
*A:PE-1# configure port 1/1/1 no shutdown
```

The primary path comes up again.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" activepath

===============================================================================
MPLS LSP: LSP-PE-1-PE-2 (active paths)
===============================================================================
Legend :
    #  - Manually switched path
    #F - Manually forced switched path
===============================================================================
LSP Name     : LSP-PE-1-PE-2
LSP Id       : 53312
Path Name    : path-PE-1-PE-2
Active Path  : Primary
To           : 192.0.2.2                         LSP Type     : dynamic

===============================================================================
*A:PE-1#
```

The auto-bandwidth data again shows **AB Adj Cause**: activePathChange, but with a different timestamp.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" auto-bandwidth

===============================================================================
MPLS LSP (Auto Bandwidth)
===============================================================================
Legend :
    + - Inherited
===============================================================================
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-PE-2
Auto BW         : Enabled                   AB OpState           : Up
Auto BW Min     : 2 Mbps                    Auto BW Max          : 20 Mbps
AB Up Thresh    : 10 percent                AB Down Thresh       : 5 percent
AB Up BW        : 1 Mbps                    AB Down BW           : 0 Mbps
AB Curr BW      : 6 Mbps                    AB Samp Intv         : 5 Mins
AB Adj Mul      : 288                       AB Samp Mul          : 1
AB Adj Time     : 1440 Mins                 AB Samp Time         : 5 Mins
AB Adj Cnt      : 1                         AB Samp Cnt          : 0
AB Last Adj     : 09/21/2018 12:55:31       AB Next Adj          : 1435 Mins
ABMaxAvgRt      : 6 Mbps                    AB Lst AvgRt         : 6 Mbps
AB Ovfl Lmt     : 2                         AB Ovfl Cnt          : 0
ABOvflThres     : 10 percent                AB Ovfl BW           : 0 Mbps
AB UndflLmt     : 3                         AB Undrfl Cnt        : 0
ABUndflThrs     : 10 percent                AB Undrfl BW         : 0 Mbps
ABMaxUndflBW    : 0 Mbps
AB Adj Cause    : activePathChange          AB Monitor BW        : False
Be Weight       : 50 percent                Af Weight            : 80 percent
L1 Weight       : 100 percent               L2 Weight            : 100 percent
Nc Weight       : 100 percent               Ef Weight            : 100 percent
H1 Weight       : 100 percent               H2 Weight            : 100 percent
===============================================================================
*A:PE-1#
```

Conclusion

Auto-bandwidth adjustment can be enabled on point-to-point LSPs in order to make a realistic bandwidth reservation, based on active iLER traffic monitoring. A user has control over how the bytes count for the different FCs by providing a sampling-weight factor. This can influence the average data rate over the sample interval.

The bandwidth is taken into account in the control plane when LSPs are established or when they change their bandwidth using MBB. The bandwidth in the data plane is not restricted by this setting.

Auto-bandwidth adjustment can be triggered in different ways: periodically, in case of overflow/underflow, manually, and in case of an active path change. It is also possible to have passive monitoring where no adjustment is done.

Automatic Creation of RSVP-TE LSPs

This chapter provides information about automatic creation of Resource Reservation Protocol with Traffic Engineering (RSVP-TE) Label Switched Paths (LSPs).

Topics in this chapter include:

Applicability

This feature is applicable to SR OS with no hardware constraints because this is a control-plane feature only.

This chapter was originally written for SR OS release 11.0.R6, but the CLI in this edition corresponds to SR OS release 16.0.R3.

Overview

Automatic creation of RSVP-TE LSPs enables the automated creation of point-to-point RSVP-TE LSPs within a single Interior Gateway Protocol (IGP) Intermediate System to Intermediate System (IS-IS) level or Open Shortest Path First (OSPF) area that can subsequently be used by services and/or IGP shortcuts. The feature is divided into two components: creation of an RSVP-TE LSP mesh, and creation of single-hop RSVP-TE LSPs. Although both can be used simultaneously, it is likely that one or the other is used.

When creating an RSVP-TE LSP mesh, the mesh can be full or partial, the extent of which is governed by a prefix list containing the system addresses of all nodes that should form part of the mesh. When using single-hop RSVP-TE LSPs, point-to-point LSPs are established to all directly connected neighbors. The purpose of these single-hop LSPs is to allow for Equal Cost Multi-Path (ECMP) load-balancing of traffic using LDP-over-RSVP, which is not possible using native RSVP LSPs.

The use of automatically created RSVP-TE LSPs avoids manual configuration of RSVP-TE LSP meshes. Even when provisioning tools—such as 5620 SAM—are used to automatically provision these LSPs, auto-mesh still provides a benefit by avoiding increased configuration file sizes.

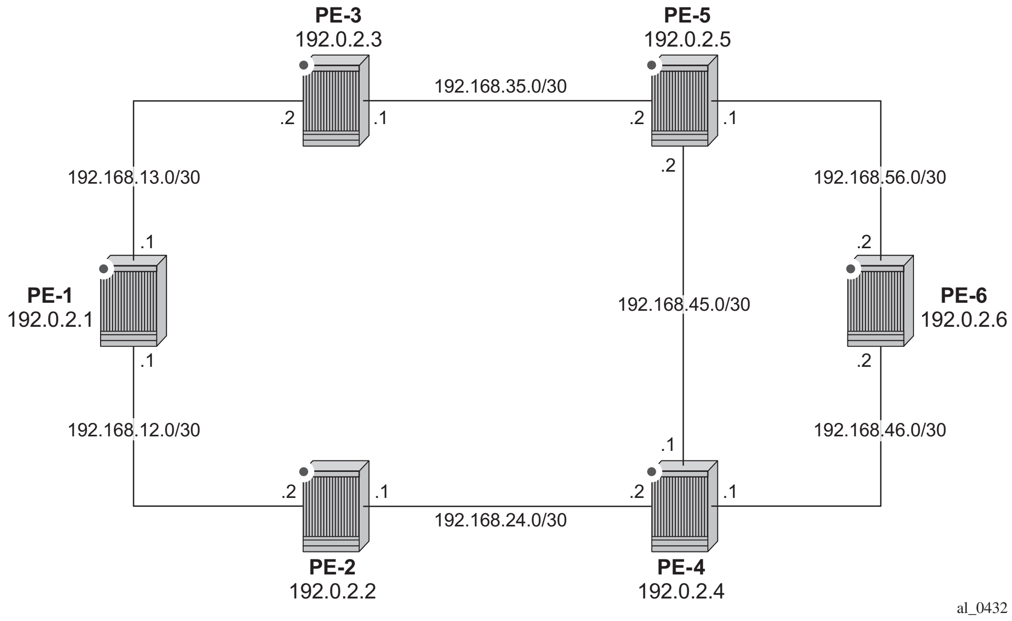
The use of automatically created Targeted Label Distribution Protocol (T-LDP) sessions is also described when using the automatically created RSVP LSPs for Layer 2 services.

Configuration

Example Topology

The example topology is shown in Figure 187. All routers participate in a single IS-IS Level 2 area that has traffic engineering enabled. Multi-Protocol Label Switching (MPLS) and RSVP are enabled on every interface, but no LSPs are initially provisioned. All routers are Border Gateway Protocol (BGP) speakers and form part of Autonomous System (AS) 64496. PE-5 is a Route Reflector and the remaining routers are IBGP clients for the VPN-IPv4 and L2-VPN address families. The objective of this example is to demonstrate how to automatically create transport LSPs using RSVP or LDP-over-RSVP, and then create services that utilize those LSPs. The exchange of BGP routes is needed for those services.

Figure 187 **Example Topology**



Automatic Creation of an RSVP-TE LSP Mesh

To start the process of automatically creating an RSVP-TE LSP mesh, the user must create a route policy referencing a prefix-list. This prefix-list contains the system addresses of all nodes that are required to be in the mesh, and can be entered as a series of /32 addresses, or simply as a range as follows. This range encompasses all of the system addresses of the nodes in the example topology as the requirement is to make a full mesh.

```
configure
    router
        policy-options
            begin
            prefix-list "System-Addresses"
                prefix 192.0.2.0/24 prefix-length-range 32-32
            exit
            policy-statement "Remote-PEs"
                entry 10
                    from
                        prefix-list "System-Addresses"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

After the route policy is created, the user must create an LSP template containing the common parameters which are used to establish all point-to-point LSPs within the mesh. For an RSVP-TE LSP mesh, the **lsp-template** must be configured with the creation-time attribute **mesh-p2p**. Upon creation of the template, CSPF is automatically enabled (and cannot be disabled), and the template must reference a **default-path** before it can be placed in a **no shutdown** state. In the example contained in the following output, the template refers to a path named “loose” that has no strict or loose hops defined, meaning the system will dynamically calculate the path while considering other specified constraints. The LSP template in this output also stipulates **fast-reroute facility** bypass protection. The default behavior is no node-protect, so this configuration requests link protection only. FRR one-to-one protection is not supported for automatically created RSVP-TE LSPs; therefore facility bypass is the only form of protection supported. Finally, the template is placed in a no shutdown state.

Next, the user must associate the LSP template with the previously defined route-policy, and this is accomplished using the **auto-lsp lsp-template** command. In this example, the LSP template “Full-Mesh” is associated with the policy-statement “Remote-PEs” that in turn references a prefix-list containing all system addresses in the example topology. Up to five policies can be associated with an LSP template at the same time. If a policy associated with an LSP template is modified in order to add or remove prefixes, the system immediately re-evaluates the policy and the prefix-list to determine if one or more LSPs need to be established, or one or more LSPs need to be torn down.

```
configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp-template "Full-Mesh" mesh-p2p
                default-path "loose"
                cspf
                fast-reroute facility
                exit
                no shutdown
            exit
            auto-lsp lsp-template "Full-Mesh" policy "Remote-PEs"
            no shutdown
        exit
```

When the **auto-lsp lsp-template** command is entered, the system commences the process of establishing the point-to-point LSPs. The prefixes defined in the prefix list are checked, and if a prefix corresponds to a router ID that is present in the Traffic Engineering Database (TED), the system instantiates a CSPF computed primary path to that prefix using the parameters specified in the LSP template. With the previously defined configuration applied at PE-6, the existence of point-to-point

RSVP LSPs to every node in the example topology can be verified as shown in the following output. The LSP name is automatically constructed as TemplateName-DestIPv4Address-TunnelId. The LSP name signaled in the Session Attribute object concatenates the LSP name with the path name (for example Full-Mesh-192.0.2.1-61441::loose).

```
*A:PE-6# show router mpls lsp

================================================================================
MPLS LSPs (Originating)
================================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
--------------------------------------------------------------------------------
Full-Mesh-192.0.2.1-61441          192.0.2.1        61441   Yes       Up   Up
Full-Mesh-192.0.2.2-61442          192.0.2.2        61442   Yes       Up   Up
Full-Mesh-192.0.2.3-61443          192.0.2.3        61443   Yes       Up   Up
Full-Mesh-192.0.2.4-61444          192.0.2.4        61444   Yes       Up   Up
Full-Mesh-192.0.2.5-61445          192.0.2.5        61445   Yes       Up   Up
--------------------------------------------------------------------------------
LSPs : 5
================================================================================
*A:PE-6#
```

The LSP template requests FRR link protection. At PE-6, this protection can be verified by querying each primary LSP. In the following output, the primary LSP to PE-1 (Full-Mesh-192.0.2.1-61441) is signaled through PE-5 (192.0.2.5) and PE-3 (192.0.2.3), and the presence of the @ indicator after each hop denotes that link protection is available to the primary path.

```
*A:PE-6# show router mpls lsp path "Full-Mesh-192.0.2.1-61441" detail
| match expression "LSP Name|Actual Hops" post-lines 4
LSP Name         : Full-Mesh-192.0.2.1-61441
From             : 192.0.2.6            To                   : 192.0.2.1
Admin State      : Up                   Oper State           : Up
Path Name        : loose
Path LSP ID      : 44544                Path Type            : Primary
Actual Hops      :
    192.168.56.2 (192.0.2.6) @                   Record Label        : N/A
 -> 192.168.56.1 (192.0.2.5) @                   Record Label        : 524275
 -> 192.168.35.1 (192.0.2.3) @                   Record Label        : 524280
 -> 192.168.13.1 (192.0.2.1)                     Record Label        : 524278
*A:PE-6#
```

Finally, it can be verified that the signaled LSPs are placed in the tunnel table and made available to the tunnel table manager so they can be used by applications and services.

```
*A:PE-6# show router tunnel-table

================================================================================
IPv4 Tunnel Table (Router: Base)
================================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
```

```
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  61441     7        192.168.56.1   300
192.0.2.2/32      rsvp      MPLS  61442     7        192.168.46.1   200
192.0.2.3/32      rsvp      MPLS  61443     7        192.168.56.1   200
192.0.2.4/32      rsvp      MPLS  61444     7        192.168.46.1   100
192.0.2.5/32      rsvp      MPLS  61445     7        192.168.56.1   100
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-6#
```

When the LSP template is in use and LSPs are instantiated, it is necessary to place the template into a shutdown state to change any parameters that cannot be handled as a Make-Before-Break (MBB). This essentially includes all LSP parameters with the exception of bandwidth and FRR without node-protection. Modification of any other parameters requires a shutdown of the LSP template and a re-signal of the LSP once the LSP template is placed in the no shutdown state again. MBB is supported for timer-based and manual re-signaling of the automatically created LSPs.

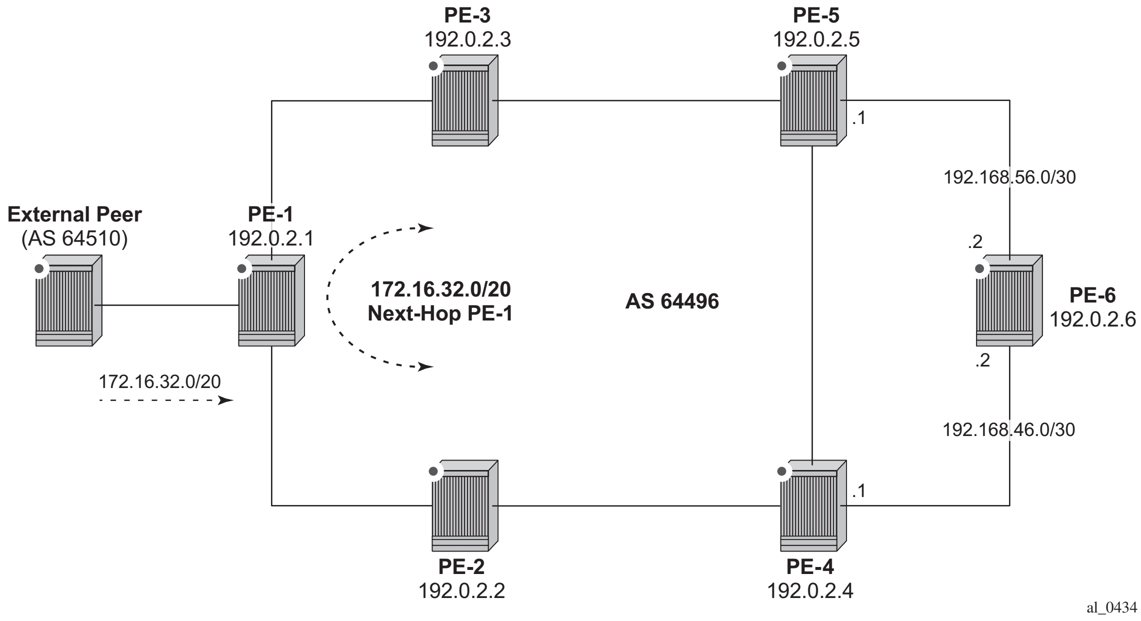
Service and Application Verification

With the RSVP-TE LSP mesh in place, it is now possible to create services and applications to utilize those LSPs. These applications and services include Layer 2 and Layer 3 VPNs, resolution of BGP labeled routes and resolution of BGP, IGP, and static routes. However, the automatically created LSPs are not available for explicit binding in a statically provisioned Service Distribution Point (SDP).

IGP Shortcuts

Figure 188 demonstrates the use of IGP shortcuts. Prefix 172.16.32.0/20 is advertised to PE-1 from an external peer in AS 64510, which PE-1 subsequently advertises into IBGP, imposing Next-Hop-Self in the process. For more details on IGP shortcuts, see the IGP Shortcuts chapter.

Figure 188 **IGP Shortcuts with RSVP-TE Auto-Mesh**



The objective is for PE-6 to use the automatically created LSP to PE-1 as an IGP shortcut (typically implemented in order to maintain a “BGP-free” core). IGP shortcuts for BGP are enabled under the main BGP context using the command **next-hop-resolution shortcut-tunnel** with options for **rsvp**, **ldp** or **bgp**. Because the example topology only has (automatically created) RSVP-TE LSPs, this option is selected. In fact, there are four more options: **sr-isis**, **sr-ospf, sr-policy**, and **sr-te**. These are related to segment routing, which is beyond the scope of this chapter.

```
configure router bgp next-hop-resolution shortcut-tunnel family
  - family <family>

 <family>             : ipv4


 [no] disallow-igp    - Allow/Disallow IGP shortcuts
 [no] enforce-strict* - Enable/Disable Use of admin-tags for resolving routes for the
                        Next-hop families
      resolution      - Configure resolution state of BGP unlabelled routes to tunnels
      resolution-fil* + Configure specific tunnels to be used for resolving BGP
                        unlabelled routes

configure router bgp next-hop-resolution shortcut-tunnel family ipv4 resolution-filter
  - resolution-filter

 [no] bgp             - Use BGP tunnelling for next hop resolution
 [no] ldp             - Use LDP tunnelling for next hop resolution
 [no] rsvp            - Use RSVP tunnelling for next hop resolution
 [no] sr-isis         - Use sr-isis for next hop resolution
 [no] sr-ospf         - Use sr-ospf for next hop resolution
 [no] sr-policy       - Use sr-policy for next hop resolution
 [no] sr-te           - Use sr-te for next hop resolution
```

```
configure
    router
        bgp
            next-hop-resolution
                shortcut-tunnel
                    family ipv4
                        resolution-filter
                            rsvp
                        exit
                        resolution filter
                    exit
                exit
            exit
```

When the shortcuts are enabled, the route-table (and FIB) can be validated to ensure that the programmed next hop is the advertising BGP speaker (as opposed to the IGP next hop), and that traffic is tunneled to that next hop through an RSVP LSP. In this case, the RSVP LSP is the LSP with tunnel ID 61441, which is the LSP to PE-1.

```
*A:PE-6# show router route-table 172.16.32.0/20

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.32.0/20                                Remote  BGP       00h08m16s  170
       192.0.2.1 (tunneled:RSVP:61441)                              0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

Layer 3 VPN

Layer 3 VPNs can utilize the automatically created LSPs by using the **auto-bind-tunnel** feature configured with the **rsvp** option (possibly in combination with LDP). The option to include both RSVP and LDP allows the system to use an RSVP LSP if one exists, and if not, to revert to an LDP-based LSP. A Virtual Private Routed Network (VPRN) is configured in this manner at PE-1 and PE-6. PE-1 is configured with a loopback address of 172.16.1.1/24 and advertises the VPN-IPv4 prefix 172.16.1.0/24 into IBGP, while PE-6 is configured with a loopback address of 172.16.6.1/24 and advertises the VPN-IPv4 prefix 172.16.6.0/24 into IBGP. The following output illustrates the configuration at PE-6. The only difference at PE-1 is the IP address assigned to the loopback interface.

```
*A:PE-6#
configure
    service
        vprn 1 name "VPRN 1" customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "loopback" create
                address 172.16.6.1/24
                loopback
            exit
            no shutdown
        exit
```

Before a VPN-IPv4 prefix is considered valid, the receiving SR OS PE router must be able to resolve the BGP next hop to an LSP in the tunnel table (if not, the prefix is held in Routing Information Base RIB-IN and flagged as invalid). At PE-6, it is possible to verify that the VPN-IPv4 prefix 172.16.1.0/24 received from PE-1 is correctly resolved by looking at the VPRN-specific route table. In the following output, the VPN-IPv4 prefix 172.16.1.0/24 with a next hop of PE-1 (192.0.2.1) is correctly resolved to an RSVP LSP with a tunnel ID of 61441.

```
*A:PE-6# show router 1 route-table 172.16.1.0/24

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.0/24                                 Remote  BGP VPN   01h51m58s  170
       192.0.2.1 (tunneled:RSVP:61441)                              0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

Layer 2 VPN

As previously described, automatically created RSVP LSPs cannot be referenced by statically provisioned SDPs. Without the ability for SDPs to explicitly reference automatically created RSVP LSPs, there is little value in manually defining SDPs within Layer 2 service constructs (there is little point in referring to an SDP that cannot bind to the underlying RSVP mesh). Therefore, in order to deliver Layer 2 services, there is a requirement to adopt a model within the service construct that permits automatic creation of SDP bindings, and this is achieved using a pseudowire-template dictating the characteristics of the SDP. The secondary effect of using pseudowire-templates to dynamically create SDPs is that these automatically created SDPs can currently only use LDP or BGP as a transport tunnel, not RSVP. The solution is to enable LDP-over-RSVP.

This can be implemented using static provisioning of peers as shown in the next output, or it can be done using automatic creation of T-LDP sessions. Regardless of the method, a reciprocal configuration must exist at both peer endpoints. The static per-peer configuration is applied in the **targeted-session** context specifying the remote peer system IP address, and the keyword **tunneling,** which enables tunneling of LDP FECs over RSVP LSPs with a far-end address matching that of the T-LDP peer. At a global level, the **prefer-tunnel-in-tunnel** command is shown, but is only required when a next hop router advertises a FEC over link-level LDP and T-LDP. In this case, by default, the system would prefer the link-level LDP tunnel, so the **prefer-tunnel-in-tunnel** instructs the system to prefer an LDP-over-RSVP tunnel if it is available. Although link-layer LDP is not present in the example topology, the command is included because the presence of link-layer LDP is common.

```
configure
    router
        ldp
            prefer-tunnel-in-tunnel
            interface-parameters
            exit
            targeted-session
                peer 192.0.2.1
                    tunneling
                    exit
                exit
            exit
            no shutdown
        exit
```

The following output provides an example demonstrating the automatic creation of T-LDP sessions. No explicit reference is made to specific peers, but rather a **peer-template** is configured containing the parameters which apply to all T-LDP sessions spawned by this template. In this example, only the **tunneling** command is required. A **peer-template-map** is then used to create a mapping between the **peer-template** (TLDP-Mesh) and a **policy** defining the IP addresses of remote nodes to which T-LDP sessions should be established. In this example, the policy “Remote-PEs” is the same policy previously used by the auto-created RSVP LSP mesh.

```
configure
    router
        ldp
            prefer-tunnel-in-tunnel
            interface-parameters
            exit
            targeted-session
                peer-template "TLDP-Mesh"
                    tunneling
                exit
                peer-template-map peer-template "TLDP-Mesh" policy "Remote-PEs"
            exit
            no shutdown
        exit
```

Regardless of whether T-LDP sessions are explicitly provisioned, or dynamically created using a peer-template, the result is that a targeted LDP session is established which can be used for advertising address and service FECs, and which is capable of tunneling LDP over RSVP.

```
*A:PE-6# show router ldp targ-peer 192.0.2.1 detail

===============================================================================
LDP IPv4 Targeted Peers
===============================================================================
-------------------------------------------------------------------------------
192.0.2.1
-------------------------------------------------------------------------------
Admin State        : Up               Oper State           : Up
Last Oper Chg      : 0d 00:05:15
Hold Time          : 45               Hello Factor         : 3
Oper Hold Time     : 45
Hello Reduction    : Disabled         Hello Reduction Fctr : 3
Keepalive Timeout  : 40               Keepalive Factor     : 4
Active Adjacencies : 1                Last Modified        : Never
Auto Created       : Yes
Creator            : template         Template Name        : TLDP-Mesh
Tunneling          : Enabled
Lsp Name           : None
Local LSR          : None             32-BitLocalLsr       : Disabled
Local-LSR ID adv.  : Disabled
Community          :
BFD Status         : Disabled
===============================================================================
No. of IPv4 Targeted Peers: 1
===============================================================================
*A:PE-6#
```

To create VPLS services using dynamically-created SDPs, BGP Auto-Discovery (BGP-AD) must be used together with LDP (or BGP) pseudowire signaling, for more details see the LDP VPLS Using BGP-Auto Discovery chapter.

In the following output, PE-6 uses BGP-AD and LDP signaling. The same configuration is applied at PE-1. The **vpls-id** is configured in the **bgp-ad** context. The VPLS ID is a network-wide identifier assigned to all VPLS Switch Instances (VSIs) belonging to the same VPLS, and is carried in VPLS Network Layer Reachability Information (NLRI) as an extended community attribute. A second parameter used for BGP-AD and carried in the VPLS NLRI is the VSI-ID, which uniquely identifies each VSI. The VSI-ID is automatically derived from the global ASN, the VPLS service ID, and the system IP address. To automatically create SDPs, the **bgp** context of the VPLS service refers to a **pw-template** defining the parameters of the pseudowire. In this example, the use of the hash (entropy) label is enabled in the pseudowire template, and a **split-horizon-group,** SHG, is applied.

```
configure
    service
        pw-template 2 name "PW2" create
            hash-label
            split-horizon-group "SHG"
            exit
        exit
        vpls 2 name "VPLS 2" customer 1 create
            bgp
                pw-template-binding 2
                exit
            exit
            bgp-ad
                vpls-id 64496:2
                no shutdown
            exit
            sap 1/1/4:2 create
            exit
            no shutdown
        exit
```

The following service information provides the BGP and BGP-AD operational parameters, and shows that an SDP of type **BgpAd** (32767:4294967295) has been automatically created. Both the SDP and the SAP are operationally up.

```
*A:PE-6# show service id 2 bgp

===============================================================================
BGP Information
===============================================================================
Bgp Instance         : 1
Vsi-Import           : None
Vsi-Export           : None
Route Dist           : None
Oper Route Dist      : 64496:2
Oper RD Type         : derivedVpls
Rte-Target Import    : None                 Rte-Target Export: None
Oper RT Imp Origin   : derivedVpls          Oper RT Import   : 64496:2
Oper RT Exp Origin   : derivedVpls          Oper RT Export   : 64496:2

PW-Template Id       : 2                    PW-Template SHG  : None
Oper Group           : None
Mon Oper Group       : None
BFD Template         : None
BFD-Enabled          : no                   BFD-Encap        : ipv4
Import Rte-Tgt       : None
-------------------------------------------------------------------------------
===============================================================================
*A:PE-6#

*A:PE-6# show service id 2 bgp-ad

-------------------------------------------------------------------------------
BGP Auto-discovery Information
-------------------------------------------------------------------------------
Admin State       : Up
Vpls Id           : 64496:2
Prefix            : 192.0.2.6
-------------------------------------------------------------------------------
*A:PE-6#

*A:PE-6# show service id 2 base | match "Service Access" post-lines 10
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:2                              q-tag        1518    1518    Up   Up
sdp:32767:4294967295 SB(192.0.2.1)       BgpAd        0       1548    Up   Up
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-6#
```

To create Epipe services using dynamically created SDPs, two options exist. Either LDP FEC 129 signaling can be used, which in turn dictates the presence of pseudowire routing information, or BGP-VPWS based signaling can be used, for more details, see the *BGP Virtual Private Wire Services* chapter. This example illustrates the use of BGP VPWS, but in either case, only single-segment pseudowires are supported. The following output shows the configuration requirements for a basic BGP-based Epipe service at PE-6. Once again a **pw-template** is used to define the characteristics of the pseudowire, and this template is referenced in the **bgp** context of the Epipe service. The **bgp** context is also where the **route-distinguisher** and **route-target** values are configured, which are carried in the VPWS NLRI and extended communities respectively. The **ve-name**, **ve-id**, and **remote-ve-name** are all configured in the **bgp-vpws** context. The **ve-id** is carried in the VPWS NLRI, and when a PE router receives a VPWS NLRI to try to establish an Epipe service, the **ve-id** from the NLRI is validated against the **ve-id** configured in the **remote-ve-name**. These must match before the Epipe becomes operational.

```
configure
    service
        pw-template 3 name "PW3" create
            hash-label
        exit
        epipe 3 name "Epipe 3" customer 1 create
            bgp
                route-distinguisher 64496:3
                route-target export target:64496:3 import target:64496:3
                pw-template-binding 3
                exit
            exit
            bgp-vpws
                ve-name "PE-6"
                    ve-id 6
                exit
                remote-ve-name "PE-1"
                    ve-id 1
                exit
                no shutdown
            exit
            sap 1/1/4:3 create
            exit
            no shutdown
        exit
```

The basic service information is truncated to show only the relevant information in order to verify that the service is operational. SDP (32766:4294967294) has been automatically created and is of type **BgpVpws**. Both the SDP and the SAP are operationally up.

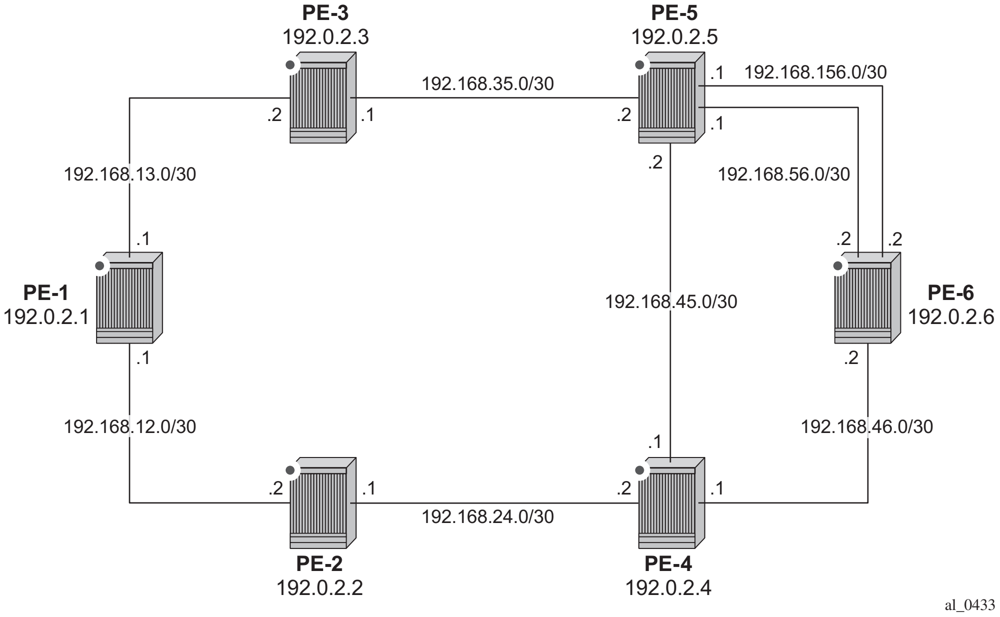
```
*A:PE-6# show service id 3 base | match "Service Access" post-lines 10
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/4:3                              q-tag        1518    1518    Up   Up
sdp:32766:4294967294 SB(192.0.2.1)       BgpVpws      0       1548    Up   Up
===============================================================================
*A:PE-6#
```

Automatic Creation of RSVP Single-Hop LSPs

As previously discussed, the purpose of a single-hop LSP mesh is to allow for ECMP load-balancing of traffic using LDP-over-RSVP. ECMP load-balancing could be implemented using LDP over a partial or full mesh of RSVP-TE LSPs, but the use of single-hop LSPs additionally allows for load-balancing across a number of parallel RSVP LSPs between nodes. To illustrate ECMP load-balancing over multiple parallel RSVP LSPs, the example topology of Figure 187 is modified to include a parallel link between PE-6 and PE-5 as shown in Figure 189. In addition, all routers are enabled for ECMP=2, as follows.

```
configure router ecmp 2
```

Figure 189 **Example Topology for Single-Hop LDP-over-RSVP with ECMP**



Unlike the automatically created RSVP-TE LSP mesh previously described, the automatically created single-hop RSVP-TE LSPs have no requirement for a prefix-list to be referenced containing the prefixes of the remote nodes that form part of the mesh. In the case of automatically created single-hop LSPs, the TE database keeps track of each TE link which comes up to a directly connected IGP neighbor. The system then establishes a single-hop LSP with a destination address matching the router ID of the neighbor and with a strict hop consisting of the address of the interface used by the TE link.

The first requirement is to create an LSP template containing the common parameters used to establish each single-hop LSP. For a single-hop LSP mesh, the **lsp-template** must be configured with the creation-time attribute **one-hop-p2p**. Upon creation of the template, **cspf** is automatically enabled (and cannot be disabled), and the **hop-limit** is set to a value of **2**. The hop limit defines the number of nodes the LSP may traverse, and, because these are single-hop LSPs to adjacent neighbors, a limit of 2 is sufficient. The template must also reference a **default-path** before it can be placed in the no shutdown state. The following example references a path named “loose” that has no strict or loose hops defined. When the RSVP PATH message is actually generated to create the one-hop LSP, it contains one strict-hop to the interface address of the neighbor; and as destination the system address of the adjacent node.

The next requirement is to trigger the creation of single-hop LSPs, and this is achieved using the **auto-lsp lsp-template** command. In this example, the LSP template “Single-Hop” is referenced, and the command is completed with the keyword **one-hop** to indicate the creation of single-hop LSPs. Unlike an RSVP-TE mesh, there is no requirement to reference a route policy. In the example, the auto-lsp with LSP template “Full-Mesh” is removed on all PEs.

```
configure router mpls no auto-lsp lsp-template "Full-Mesh"
```

The following one-hop LSP template is created on all nodes:

```
configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp-template "Single-Hop" one-hop-p2p
                default-path "loose"
                cspf
                hop-limit 2
                no shutdown
            exit
            auto-lsp lsp-template "Single-Hop" one-hop
            no shutdown
        exit
    exit
exit
```

Once the **auto-lsp lsp-template** command is entered, the system starts the process of establishing the single-hop LSPs. A check is made of the TE database for every TE link to a directly connected IGP neighbor, and a single-hop LSP is established across each TE link. The following output is taken from PE-6 and shows the automatically created single-hop LSPs. The LSP names are automatically constructed as TemplateName-DestIPv4Address-TunnelId. The LSP name signaled in the session attribute object concatenates the LSP name with the path name (for example Single-Hop-192.0.2.4-61449::loose). Recall from Figure 189 that PE-6 has a single TE-enabled link to PE-4, and two TE-enabled links to PE-5, therefore with ECMP=2, there is one LSP to PE-4 (192.0.2.4) and two LSPs to PE-5 (192.0.2.5). However, if ECMP=1, only one single-hop LSP would be signaled to PE-5.

```
*A:PE-6# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
Single-Hop-192.0.2.4-61448         192.0.2.4        61448   No        Up   Up
Single-Hop-192.0.2.5-61449         192.0.2.5        61449   No        Up   Up
Single-Hop-192.0.2.5-61450         192.0.2.5        61450   No        Up   Up
```

```
-------------------------------------------------------------------------------
LSPs : 3
===============================================================================
*A:PE-6#
```

The purpose of single-hop LSPs is to enable ECMP load-balancing using LDP-over-RSVP, so there is a requirement to configure T-LDP sessions between RSVP LSP endpoints. This can be implemented using static peer provisioning, or it can be done using automatic creation of T-LDP sessions, both of which have been previously described and they are therefore not repeated. In this example, the automatic creation of T-LDP sessions approach is used, and T-LDP sessions are created to adjacent neighbors that are capable of tunneling inside RSVP.

```
*A:PE-6# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.4:0         Targeted  Established   17        18        0d 00:00:45
192.0.2.5:0         Targeted  Established   18        19        0d 00:00:51
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-6#
```

To validate the ECMP load-balancing capability, PE-5 is configured to advertise prefix 172.16.5.0/24 to PE-6. In turn, PE-6 is configured for **ibgp-multipath** to enable load-balancing over IGP links to the BGP next hop address, **next-hop-resolution shortcut-tunnel resolution-filter ldp** to enable tunneling of traffic destined toward the BGP next hop in MPLS, and **ecmp 2**. For additional information on BGP multipath, see the *BGP Multipath* chapter.

```
configure
    router
        bgp
            ibgp-multipath
            next-hop-resolution
                shortcut-tunnel
                    family ipv4
                        resolution-filter
                            ldp
                        exit
                        resolution filter
                    exit
                exit
            exit
            no shutdown
```

The prefix 172.16.5.0/24 advertised by PE-5 is learned at PE-6 and installed in the RIB/FIB with PE-5’s system address (192.0.2.5) as next hop.

```
*A:PE-6# show router bgp routes 172.16.5.0/24
===============================================================================
 BGP Router ID:192.0.2.6        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  172.16.5.0/24                                      100         None
      192.0.2.5                                          None        -
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-6#
```

Checking the FIB for the next hop address 192.0.2.5, it can be verified that both links are installed as next hop addresses, meaning that ECMP load-balancing is active.

```
*A:PE-6# show router fib 1 192.0.2.5/32

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                            Protocol
  NextHop
-------------------------------------------------------------------------------
192.0.2.5/32                                              ISIS
  192.168.56.1 (int-PE-6-PE-5)
  192.168.156.1 (int-PE-6-PE-5-2nd)
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-6#
```

Conclusion

Automatic creation of RSVP-TE LSPs provides a good solution for reducing the amount of provisioning activity required when configuring RSVP LSPs. However, there are some constraints with regard to the way that services are deployed on top of those LSPs. SDPs cannot explicitly reference automatically-created RSVP LSPs, which means that automatically created SDPs need to be used for Layer 2 services. In turn, automatically-created SDPs can only use LDP or BGP as a transport tunnel (not RSVP), therefore, in order to use the automatically created RSVP mesh, LDP over RSVP must be used. These caveats need to be fully understood before considering deployment of automatically created RSVP-TE LSPs.

BFD for RSVP-TE and LDP LSPs

This chapter provides information about BFD for RSVP-TE and LDP LSPs.

Topics in this chapter include:

Applicability

The information and configuration in this chapter were based on SR OS Release 15.0.R7. In the current edition these are based on 16.0.R1.

Overview

SR OS supports RFC 5884, *Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)*, and enables LSPs to be monitored between the ingress and egress LERs, regardless of the number of LSRs that the LSP traverses. When enabled, faults to individual LSPs can be detected quickly, so BFD for MPLS LSPs is ideal for monitoring LSPs carrying high-value services, where detecting forwarding failures in a minimum amount of time is critical. The LSPs can be established through RSVP-TE or LDP.

Enabling BFD for LSPs avoids manual hop-by-hop troubleshooting of each element along the LSP. BFD sessions are created and run end-to-end, from ingress to egress, so BFD session state is maintained in the ingress LER (iLER) and egress LER (eLER), but not in intermediate LSRs. If an LSP BFD session changes state, an SNMP trap is generated. Because LSPs are unidirectional, a routed return path is used for the BFD control packets from the eLER toward the iLER.

BFD is only used for fault detection, and will not redirect traffic to an alternate path. On detection of a failure, BFD informs other software components, which then can redirect traffic to avoid faulty links.

BFD is used in conjunction with LSP ping for MPLS LSP fault detection:

- LSP ping is used for bootstrapping the BFD session, at which time local and remote discriminator values are exchanged.
- BFD is used to exchange fault detection packets at the required detection interval.
- LSP ping is used to periodically verify the control plane against the data plane by ensuring that the LSP is mapped to the same FEC at the egress as at the ingress.

BFD can be used for RSVP-TE and LDP LSPs. If BFD is applied to RSVP-TE LSPs, it only runs on the currently active path. It cannot determine if any non-active paths (for example, a secondary path or primary path during reversion) that the system might switch to are up and forwarding. If BFD is applied to LDP LSPs, the session runs on the path defined by the underlying IGP.

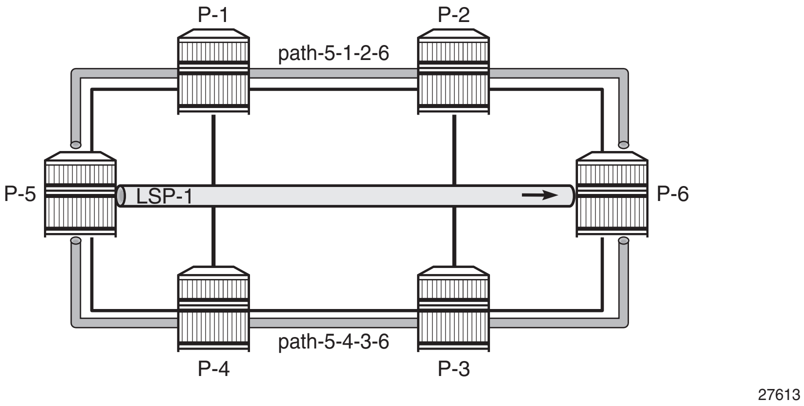
BFD for LSPs can be combined with a failure action. For RSVP-TE LSPs, the failure action can be down or failover; see the BFD for RSVP-TE LSPs with Failure-Action for more information. For LDP LSPs, the failure action can only be down.

Configuration

BFD for RSVP-TE LSPs

Figure 190 shows the example topology for BFD for RSVP-TE LSPs.

Figure 190 **BFD for RSVP-TE LSPs - Topology**



The initial configuration includes:

- Cards, MDAs, and ports
- Router interfaces
- IS-IS as IGP on all interfaces (alternatively, OSPF can be used), with traffic engineering enabled
- MPLS and RSVP-TE enabled on all interfaces

Base Configuration

The example topology from Figure 190 has an LSP defined on P-5 using two strict paths, where *path-5-1-2-6* is taking the upper path and used as the primary path, and *path-5-4-3-6* is the lower path and used as the secondary, as follows:

```
configure
    router
        mpls
            path "path-5-1-2-6"
                hop 10 192.168.15.1 strict
                hop 20 192.168.12.2 strict
                hop 30 192.168.26.2 strict
                no shutdown
            exit
            path "path-5-4-3-6"
                hop 10 192.168.45.1 strict
                hop 20 192.168.34.1 strict
                hop 30 192.168.36.2 strict
                no shutdown
            exit
            lsp "lsp-1"
                to 192.0.2.6
                cspf
                primary "path-5-1-2-6"
                exit
                secondary "path-5-4-3-6"
                exit
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

BFD for RSVP-TE LSPs Configuration

There are four steps to configure BFD for RSVP-TE LSPs:

1. Configure a BFD template.
2. Enable LSP BFD on the tail node.

3. Apply the BFD template to the LSP or LSP path.
4. Enable BFD on the LSP or LSP path.

Step 1

The **bfd-template** provides the control packet timer values for the BFD session to use at the LSP head end. The BFD state machine at the tail end initially uses system-wide default parameters (minimum transmit and receive intervals are both 1 second), because an LSP is unidirectional so no configuration for the LSP exists at the tail end. The head end then attempts to adjust the control packet timer values when it transitions to the INIT state.

The general command to define a BFD template is as follows:

```
config
    router
        bfd
            bfd-template name
                transmit-interval transmit-interval
                receive-interval receive-interval
                echo-receive echo-interval
                multiplier multiplier
                type {cpm-np}
            exit
```

However, network processor BFD is not supported for LSPs, and the minimum supported receive or transmit timer interval is 1 second. Therefore, an error is generated if a user tries to apply a BFD template of **type cpm-np** or any unsupported transmit or receive interval value to an LSP. An error is also generated when the user attempts to commit changes to a BFD template that is already applied to an LSP where the new values are invalid for lsp-bfd.

BFD templates may be used by different BFD applications (for example, LSPs or pseudowires). If the BFD timer values are changed in a template, the BFD sessions on LSPs or spoke-SDPs to which that template applies try to renegotiate their timers to the new values.

In this example, the BFD template used is configured as follows:

```
configure
    router
        bfd
            bfd-template "bfdt-1"
                no type
                transmit-interval 2000
                receive-interval 2000
                multiplier 5
                echo-receive 100
            exit
        exit
    exit
```

Step 2

LSP BFD is enabled or disabled on a node-wide basis with the **[no] bfd-sessions** *maxlimit* command in the **config router lsp-bfd** context. The *maxlimit* parameter configures the maximum number of LSP BFD sessions that can be established. This is required at the tail end of the LSP.

In this example, the tail node is configured as follows:

```
configure router lsp-bfd bfd-sessions 10
```

Because BFD resources are shared by different BFD applications, the limit defined here must provide sufficient resources for other applications.

Steps 3 and 4

LSP BFD is applicable to configured RSVP LSPs as well as to mesh P2P and one-hop P2P auto-LSPs. It is configured on an RSVP-TE LSP, or on the primary path of an RSVP-TE LSP, under the **bfd** context at the LSP head end.

A BFD template must always be configured first. BFD is then enabled using the bfd-enable command.

To apply and enable the BFD template at LSP level, the command is as follows:

```
configure
    router
        mpls
            lsp name
                bfd
                    [no] bfd-template name
                    [no] bfd-enable
                exit
```

When BFD is configured at the LSP level, BFD packets follow the currently active path of the LSP.

To apply and enable the BFD template at primary path level, the command is as follows:

```
config
    router
        mpls
            lsp name
                primary path-name
                    bfd
                        [no] bfd-template name
                        [no] bfd-enable
                    exit
```

It is not possible to configure LSP BFD on a secondary path or on P2MP LSPs.

LSP BFD at the LSP level and the path level are mutually exclusive. That is, if LSP BFD is already configured for the LSP, its configuration for the path is blocked. Likewise, it cannot be configured on the LSP if it is already configured at the path level.

LSP BFD is supported on auto-LSPs. In that case, LSP BFD is configured on mesh P2P and one-hop P2P auto-LSPs using the LSP template.

In this example, on the head-end node, the BFD template is applied to the LSP, as follows:

```
configure
    router
        mpls
            lsp "lsp-1"
                bfd
                    bfd-template "bfdt-1"
                    bfd-enable
                exit
```

BFD Verification

The details of the MPLS LSP path show that BFD is enabled and using BFD template *bfdt-1*, as follows:

```
*A:P-5# show router mpls lsp detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : lsp-1
LSP Type        : RegularLsp                LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.5                 To                   : 192.0.2.6
Adm State       : Up                        Oper State           : Up
LSP Up Time     : 0d 00:03:28               LSP Down Time        : 0d 00:00:00
Transitions     : 1                         Path Changes         : 1
Retry Limit     : 0                         Retry Timer          : 30 sec
Signaling       : RSVP                      Resv. Style          : SE
Hop Limit       : 255                       Negotiated MTU       : 1564
Adaptive        : Enabled                   ClassType            : 0
FastReroute     : Disabled                  Oper FR              : Disabled
CSPF            : Enabled                   ADSPEC               : Disabled
Metric          : N/A                       Use TE metric        : Disabled
Load Bal Wt     : N/A                       ClassForwarding      : Disabled
Include Grps    :                           Exclude Grps         :
```

```
None                                           None
Least Fill      : Disabled
BFD Template    : bfdt-1                    BFD Ping Intvl       : 60
BFD Enable      : True                      BFD Failure-action   : None

--- snipped ---

Primary(a)      : path-5-1-2-6              Up Time              : 0d 00:03:28
Bandwidth       : 0 Mbps
Secondary       : path-5-4-3-6              Down Time            : 0d 00:03:28
Bandwidth       : 0 Mbps
===============================================================================
*A:P-5#
```

Initially, the BFD session is running over *path-5-1-2-6*, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-1-2-6                                  Up           45         49
  192.0.2.6                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

At the head end, the BFD session details are as follows:

```
*A:P-5# show router bfd session detail lsp-rsvp head

===============================================================================
BFD On LSP Session
===============================================================================
Rsvp Session Na: lsp-1::path-5-1-2-6
Remote Address : 192.0.2.6
Lsp Id         : 15360                    Tunnel Id        : 1
Oper State     : Up                       Protocols        : rsvpLsp
Recd Msgs      : 62                       Sent Msgs        : 59
Up Time        : 0d 00:01:54              Up Transitions   : 1
Down Time      : None                     Down Transitions : 0
                                          Version Mismatch : 0

Forwarding Information

Local Discr    : 2                        Local State      : Up
Local Diag     : 0 (None)
Local Mode     : Async
Local Min Tx   : 2000                     Local Mult       : 5
Last Sent      : 07/12/2018 09:28:12      Local Min Rx     : 2000
Type           : central
Remote Discr   : 1                        Remote State     : Up
Remote Diag    : 0 (None)                 Remote Mode      : Async
Remote Min Tx  : 1000                     Remote Mult      : 3
Last Recv      : 07/12/2018 09:28:11      Remote Min Rx    : 1000
===============================================================================
===============================================================================
*A:P-5#
```

At the tail end, the BFD session details are as follows:

```
*A:P6# show router bfd session detail lsp-rsvp tail

===============================================================================
BFD On LSP Session
===============================================================================
Rsvp Session Na: lsp-1::path-5-1-2-6
Remote Address : 192.0.2.5
Lsp Id         : 15360                    Tunnel Id        : 1
Oper State     : Up                       Protocols        : rsvpLsp
Recd Msgs      : 118                      Sent Msgs        : 122
Up Time        : 0d 00:03:53              Up Transitions   : 1
Down Time      : None                     Down Transitions : 0
                                          Version Mismatch : 0

Forwarding Information

Local Discr    : 1                        Local State      : Up
Local Diag     : 0 (None)
Local Mode     : Async
Local Min Tx   : 1000                     Local Mult       : 3
Last Sent      : 07/12/2018 09:30:11      Local Min Rx     : 1000
Type           : central
Remote Discr   : 2                        Remote State     : Up
Remote Diag    : 0 (None)                 Remote Mode      : Async
Remote Min Tx  : 2000                     Remote Mult      : 5
Last Recv      : 07/12/2018 09:30:10      Remote Min Rx    : 2000
===============================================================================
===============================================================================
*A:P6#
```

A failure on the upper path is detected by BFD quickly, which is emulated by bringing down the link between P-1 and P-2. This results in the BFD session being re-established on *path-5-4-3-6*, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-4-3-6                                  Up            9         12
  192.0.2.6                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

Bringing up the link between P-1 and P-2 will result in the primary path becoming active again.

The ping bootstrap and periodic verification information for BFD on LSPs can be displayed at the head end, as follows:

```
*A:P-5# show test-oam lsp-bfd lsp-name "lsp-1"

-------------------------------------------------------------------------------
LSP Ping Bootstrap and Periodic Verification Information for BFD on LSPs
-------------------------------------------------------------------------------
OAM Operational State    : Bootstrapped - Sending Periodic Verification
FEC Type                 : RSVP
LSP Name : lsp-1
LSP Path Status          : active
Source Address           : 192.0.2.5
Replying Node            : 192.0.2.6
Latest Return Code       : EgressRtr (3)
Latest Return Subcode    : 1
Local BFD Discriminator  : 4            Remote BFD Discriminator : 3
LSP Ping Tx Interval (s) : 60           Bootstrap Retry Count    : 0
Tx LSP Ping Requests     : 1            Rx LSP Ping Replies      : 1
-------------------------------------------------------------------------------
No. of matching BFD on LSP sessions: 1
-------------------------------------------------------------------------------
*A:P-5#
```

BFD sessions changing state are trapped, so these are logged to log-id 99, as follows:

```
107 2018/07/12 09:31:20.067 CEST MINOR: BFD #2004 Base 192.0.2.6
"The protocol (RSVP LSP) using BFD session on node 192.0.2.6 has been added"

108 2018/07/12 09:31:21.066 CEST MINOR: BFD #2004 Base 192.0.2.6
"The protocol (RSVP LSP) using BFD session on node 192.0.2.6 has been cleared"

109 2018/07/12 09:31:21.066 CEST MINOR: BFD #2003 Base 192.0.2.6
"The lspHead BFD Session with Local Discriminator 3 on 192.0.2.6 has been deleted"

110 2018/07/12 09:31:21.068 CEST MAJOR: SVCMGR #2316 Base
"Processing of a SDP state change event is finished and the status of all
                               affected SDP Bindings on SDP 56 has been updated."
```

```
111 2018/07/12 09:31:24.066 CEST WARNING: MPLS #2012 Base VR 1:
"LSP path lsp-1::path-5-4-3-6 is operationally disabled
                                                  ('shutdown') because noError"

112 2018/07/12 09:31:25.675 CEST MINOR: BFD #2002 Base 192.0.2.6
"The lspHead BFD session with Local Discriminator 4 on 192.0.2.6 is up"
```

The **tools** command for displaying LSP details at the head end also includes BFD related information, if applicable, as follows:

```
*A:P-5# tools dump router mpls lspinfo "lsp-1" detail
LSP "lsp-1"  LspIdx 1  LspType Dynamic  State LSPS_UP  Flags 0x2000
NumPaths 2  NumSdps 1  NumCBFSdps 0  NumFltrEntries 0
HoldTimeRemaining 0secs  ClassType 0  Metric 0  OperMetric 30
LDPoRsvp Include  VprnAutoBind Include IgpShortCut Include BgpShortCut Include
BgpTransTunnel Include IpShCutTtlPropLocal TRUE IpShCutTtlPropTans TRUE
RelativeMetricOffset 2147483647 EntropyLbl inherit MTU 1564 InUseByLdp FALSE
LspAdminState : 2 LspOperState : 2 lspRowStatus : 1
ClassForwarding: Disabled
BFD Enabled  Template bfdt-1  PingInterval 60 Failure-Action None
Path Profile:
  None
Admin Tags:
  None
  Path "path-5-1-2-6"  LspId 15364  PathType Primary  ActivePath Yes
  RowStatus 1  LastChange 000 00:11:03.310  AdminState :2  OperState :2
                                                OperStateChange 000 00:01:16.980
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.15.2 EgrLnkId 0
                                                        RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.15.1 IngLnkId 0 EgrIp 192.168.12.1 EgrLnkId 0
                                                        RtrId 192.0.2.1 Flag 0x0
      Hop[3] IngIp 192.168.12.2 IngLnkId 0 EgrIp 192.168.26.1 EgrLnkId 0
                                                        RtrId 192.0.2.2 Flag 0x0
      Hop[4] IngIp 192.168.26.2 IngLnkId 0 EgrIp 192.0.2.6 EgrLnkId 0
                                                        RtrId 192.0.2.6 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x40000000
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 0 HopLimit 255 BW 0Mbps
    TotIgpCost 30 OperMetric 30 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 30  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No
  Path "path-5-4-3-6"  LspId 15362  PathType Secondary  ActivePath No
  RowStatus 1  LastChange 000 00:11:03.310  AdminState :2  OperState :3
                                                OperStateChange 000 00:01:11.980
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_DOWN  Flags 0x40000
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 255 HopLimit 255 BW 0Mbps
    TotIgpCost 0 OperMetric 16777215 MTU 0
    BFD Disabled  Template n/a  PingInterval 60
```

```
Total Ingress LSP Count         : 1
*A:P-5#
```

The current BFD session information for RSVP LSPs can be displayed using a **tools** command at the head end, as follows:

```
*A:P-5# tools dump router bfd lsp-rsvp head
--------------------------------------------------------------------------------
FEC: (PTR 0x27985ec8)
 RSVP : vrId: 1 (To: 192.0.2.6 - 1 - 192.0.2.5), Sender (192.0.2.5 - 15364)

  Session: lsp-1::path-5-1-2-6    refCnt = 1
  PingIntvl: 60 Flags: 0x6  ProtNhidx: 13  NumNextHop: 1
  TemplName: bfdt-1   LspName: lsp-1 TunnelId: 1  NumLspUser: 0
  NextHop: 192.168.15.1  IfIndex: 1  isBackup: N
    PGId: 0 [State: N/A]   NhIdx: 13
    Label:-  [0]524287
       BFD Handle: 4   State: UP   LastEvent: UP
       BFD UserId: 24   TmrActive: N [0]    NumRetry: 0
       DstAddr: 127.0.0.4   LocalDiscr: 4   RemoteDiscr: 0
--------------------------------------------------------------------------------
Total FEC Count in Head: 1

*A:P-5#
```

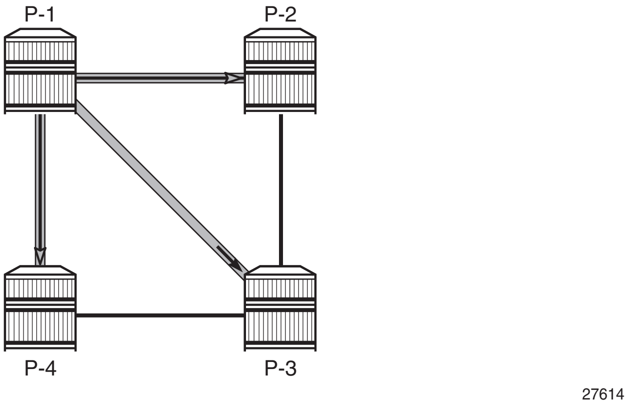
Other **tools** commands can display BFD LSP information at the tail end, as follows:

```
*A:P-6# tools dump test-oam lsp-bfd tail

 -------------------------------------------------------------------------------
 Total Number of Active Tail Cache Sessions : 1
 -------------------------------------------------------------------------------

 VrId              : 1
 RemoteBfdDisc     : 5
 LocalBfdDisc      : 6
 FecType           : rsvp_ipv4(3)
 LspId             : 37898
 TunnelId          : 1
 SenderIp          : 192.0.2.5
 TunnEndIp         : 192.0.2.6
 ExtTunnId         : 192.0.2.5
 Bootstrap Echo Rx : rcvd 2018/03/02 14:31:02.00 UTC
                     handle 5 seqNum 2 rc 3 rsc 1
 Last Echo Req Rx  : rcvd 2018/03/02 14:43:07.00 UTC
                     handle 5 seqNum 14 rc 3 rsc 1
 -------------------------------------------------------------------------------
 Number of Matched Tail Cache Sessions : 1
 -------------------------------------------------------------------------------
*A:P-6#
```

BFD for LDP LSPs

Figure 191 shows the example topology for BFD for LDP LSPs.

Figure 191 **BFD for LDP LSPs - Topology**



The initial configuration includes:

- Cards, MDAs, and ports
- Router interfaces
- IS-IS as IGP on all interfaces (alternatively, OSPF can be used), with traffic engineering enabled

Base Configuration

The example topology from Figure 2 has LDP configured on all interfaces. LDP automatically generates and distributes labels across the network, so for the topology in Figure 2, twelve tunnels are created so that every node can reach any other node; only the tunnels originating in P-1 are shown. The LDP configuration for P-1 is as follows; the LDP configuration for P-2, P-3, and P-4 is similar.

```
# on P-1
configure
    router
        ldp
             interface-parameters
                interface "int-P-1-P-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-P-1-P-4" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-P-1-P-5" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
        exit
    exit
exit
```

BFD for LDP LSPs Configuration

There are six steps to configure BFD for LDP LSPs:

1. Create a BFD template.
2. Enable LSP BFD on the tail node.
3. Create a prefix list.
4. Configure LSP BFD for LDP.
5. Apply the BFD template to the LSP.
6. Enable BFD on the LSP.

Step 1

The general command to define a BFD template is the same as before, so is not repeated.

In this example, the BFD template used is configured as follows:

```
# on P-1
configure
    router
        bfd
            bfd-template "bfdt-2"
                no type
                transmit-interval 1000
                receive-interval 1000
                multiplier 3
                echo-receive 100
            exit
        exit
    exit
exit
```

Step 2

The command to enable or disable LSP BFD on a node-wide basis at the tail end of the tunnels is the same as before, and is not repeated in this section.

In this example, the tail node is configured as follows:

```
# on P-2 and P-3
configure router lsp-bfd bfd-sessions 5
```

Step 3

When high-value services are relying on the LDP tunnels between P-1, P-2, and P-3, a prefix list with the system IP addresses (or other routable loopback addresses) of P-2 and P-3 can be used in P-1 for monitoring these tunnels. In this example, the *pfx-lst-1* prefix list is defined as follows:

```
# on P-1
configure
    router
        policy-options
            begin
            prefix-list "pfx-lst-1"
                prefix 192.0.2.2/32 exact
                prefix 192.0.2.3/32 exact
            exit
            commit
        exit
    exit
exit
```

Step 4, 5, and 6

LSP BFD is configured for LDP using the following commands:

```
config
    router
        ldp
            lsp-bfd prefix-list-name
                priority priority-level
                bfd-template bfd-template-name
                source-address ip-address
                bfd-enable
                lsp-ping-interval seconds
            exit
```

A BFD template must always be applied first. BFD is then enabled using the **bfd-enable** command.

The priority level is set to one, by default, and is used in case a prefix appears in the multiple prefix list; see the *MPLS User Guide* for more information. The source address can be any local address routable by the other nodes in the network; by default, the system IP address is used. The LSP ping interval defines how frequently ping messages must be sent on the LSP.

The BFD template is applied to the head-end node in the lsp-bfd context, as follows. The **lsp-bfd** command takes the prefix-list name defined in step 3 as its argument.

```
# on P-1
configure
    router
        ldp
            lsp-bfd "pfx-lst-1"
                bfd-template "bfdt-2"
                source-address 192.0.2.1
                bfd-enable
            exit
        exit
    exit
exit
```

BFD Verification

The prefix lists applied to LDP BFD are as follows:

```
A:P-1# show router ldp lsp-bfd
===========================================================
BFD on LDP LSP Configuration Summary
===========================================================
Prio   Prefix List Name                 Enabled   Prefixes
-----------------------------------------------------------
1      pfx-lst-1                            Yes          2
-----------------------------------------------------------
No. of prefix lists: 1
===========================================================
A:P-1#
```

The LDP BFD information for prefix list *pfx-lst-1* is as follows:

```
A:P-1# show router ldp lsp-bfd "pfx-lst-1"
==============================================================================
BFD on LDP LSP Configuration Detail
==============================================================================
Prefix List       : pfx-lst-1
Prefix Count      : 2
BFD Template      : bfdt-2
Source Address    : 192.0.2.1
BFD Enable        : Yes                Failure Action    : none
LSP Ping Interval : 60 seconds         Priority          : 1
==============================================================================
A:P-1#
```

The prefixes of prefix list *pfx-lst-1* to which the system tries to establish BFD sessions are as follows:

```
A:P-1# show router ldp lsp-bfd "pfx-lst-1" prefixes

=========================================================================
BFD on LDP LSP Prefix List "pfx-lst-1" (Enabled)
=========================================================================
Prefix                                       Operational State
-------------------------------------------------------------------------
192.0.2.2/32                                 Up
192.0.2.3/32                                 Up
-------------------------------------------------------------------------
No. of prefixes: 2
=========================================================================
A:P-1#
```

The LDP BFD session data created and maintained at the head end P-1 is as follows:

```
A:P-1# show router bfd session lsp-ldp head

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
192.0.2.2/32                                         Up         4288       4289
  N/A                                                 3         1000       1000
  ldpLsp                                        central          N/A        N/A
192.0.2.3/32                                         Up         4285       4287
  N/A                                                 3         1000       1000
  ldpLsp                                        central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 2
===============================================================================
A:P-1#
```

The LDP BFD session data created and maintained at the tail end P-3 is as follows:

```
A:P-3# show router bfd session lsp-ldp tail

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
```

```
-------------------------------------------------------------------------------
192.0.2.3/32                                         Up         4456       4454
  192.0.2.1                                           3         1000       1000
  ldpLsp                                        central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
A:P-3#
```

The ping bootstrap and periodic verification information for BFD on LSPs can be displayed at the head end, as follows:

```
*A:P-1# show test-oam lsp-bfd

-------------------------------------------------------------------------------
LSP Ping Bootstrap and Periodic Verification Information for BFD on LSPs
-------------------------------------------------------------------------------
OAM Operational State    : Bootstrapped - Sending Periodic Verification
FEC Type                 : LDP
Prefix                   : 192.0.2.2/32
Source Address           : 192.0.2.1
Replying Node            : 192.0.2.2
Latest Return Code       : EgressRtr (3)
Latest Return Subcode    : 1
Local BFD Discriminator  : 3           Remote BFD Discriminator : 1
LSP Ping Tx Interval (s) : 60          Bootstrap Retry Count    : 0
Tx LSP Ping Requests     : 2           Rx LSP Ping Replies      : 2
-------------------------------------------------------------------------------
OAM Operational State    : Bootstrapped - Sending Periodic Verification
FEC Type                 : LDP
Prefix                   : 192.0.2.3/32
Source Address           : 192.0.2.1
Replying Node            : 192.0.2.3
Latest Return Code       : EgressRtr (3)
Latest Return Subcode    : 1
Local BFD Discriminator  : 4           Remote BFD Discriminator : 1
LSP Ping Tx Interval (s) : 60          Bootstrap Retry Count    : 0
Tx LSP Ping Requests     : 2           Rx LSP Ping Replies      : 2
-------------------------------------------------------------------------------
No. of matching BFD on LSP sessions: 2
-------------------------------------------------------------------------------
*A:P-1#
```

BFD sessions changing state are trapped, so these are logged to log-id 99, as follows:

```
75 2018/07/12 09:40:37.808 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (LDP LSP) using BFD session on node 192.0.2.2 has been cleared"

76 2018/07/12 09:40:37.808 CEST MINOR: BFD #2003 Base 192.0.2.2
"The lspHead BFD Session with Local Discriminator 1 on 192.0.2.2 has been deleted"

77 2018/07/12 09:40:39.808 CEST MINOR: BFD #2004 Base 192.0.2.3
"The protocol (LDP LSP) using BFD session on node 192.0.2.3 has been cleared"

78 2018/07/12 09:40:39.808 CEST MINOR: BFD #2003 Base 192.0.2.3
"The lspHead BFD Session with Local Discriminator 2 on 192.0.2.3 has been deleted"
```

```
79 2018/07/12 09:41:07.808 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (LDP LSP) using BFD session on node 192.0.2.2 has been added"

80 2018/07/12 09:41:10.808 CEST MINOR: BFD #2004 Base 192.0.2.3
"The protocol (LDP LSP) using BFD session on node 192.0.2.3 has been added"

81 2018/07/12 09:41:11.680 CEST MINOR: BFD #2002 Base 192.0.2.2
"The lspHead BFD session with Local Discriminator 3 on 192.0.2.2 is up"

82 2018/07/12 09:41:13.977 CEST MINOR: BFD #2002 Base 192.0.2.3
"The lspHead BFD session with Local Discriminator 4 on 192.0.2.3 is up"
```

The current BFD session information for LDP LSPs can be displayed using a **tools** command, as follows:

```
A:P-1# tools dump router bfd lsp-ldp prefix 192.0.2.3/32
--------------------------------------------------------------------------------
FEC: (PTR 0x27985f68)
 LDP HEAD: vrId: 1 (To: 192.0.2.3/32), Sender (192.0.2.1)
  PingIntvl: 60 Flags: 0x6  ProtNhidx: 16  NumNextHop: 1
  TemplName: bfdt-2   LspName:  TunnelId: 65538  NumLspUser: 0
  NextHop: 192.168.12.2  IfIndex: 1  isBackup: N
    PGId: 3 [State: UP]   NhIdx: 16
    Label:-  [0]524283
       BFD Handle: 4   State: UP   LastEvent: UP
       BFD UserId: 25   TmrActive: N [0]    NumRetry: 1
       DstAddr: 127.0.0.4   LocalDiscr: 4   RemoteDiscr: 0
--------------------------------------------------------------------------------
Total FEC Count in Head: 1

Total FEC Count in Tail: 0

A:P-1#
```

The BFD templates used by LDP can also be listed using a **tools** command, as follows:

```
*A:P-1# tools dump router ldp lsp-bfd bfd-templates-in-use

===================================================================
BFD on LDP LSP BFD Template Summary
===================================================================
Prefix List Name                 BFD Template Name
-------------------------------------------------------------------
pfx-lst-1                        bfdt-2
-------------------------------------------------------------------
No. of prefix lists: 1
===================================================================
*A:P-1#
```

Conclusion

BFD is supported for RSVP-TE and LDP LSPs and is ideal for monitoring LSPs carrying high-value services, where detecting failures in a minimum amount of time is critical.

BFD for RSVP-TE LSPs with Failure-Action

This chapter describes BFD for RSVP-TE LSPs with Failure-Action. Topics include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

The information and configuration in this chapter were based on SR OS Release 15.0.R7. In the current edition these are based on 16.0.R1.

Overview

Using the **failure-action** command, the operator can configure the action taken by the system if a BFD session fails for an RSVP LSP or LDP prefix list. See the BFD for RSVP-TE and LDP LSPs chapter for a general description on configuring BFD for LSPs.

When failure-action failover is configured, and the LSP BFD session goes down on the currently active path, the LSP switches from the primary path to a secondary path, or from the currently active secondary path to the next best preference secondary path if the currently active path was a secondary.

When failure-action down is configured, the LSP is registered as unusable in the Tunnel Table Manager (TTM) when BFD on the LSP goes down. A tunnel being registered as unusable in TTM is not available to RTM and all routes using that tunnel are withdrawn. SDP auto-bind will not use an LSP until it is registered as usable. Traffic cannot pass through that LSP, even when secondary paths are available for that LSP.

In either case, SNMP traps are raised when the BFD state machine for the LSP transitions.

Nokia recommends configuring the BFD control packet timer intervals long enough to deal with transient data path disruptions that may occur when the underlying transport network recovers following a failure.

LSP BFD only runs on the currently active path. It cannot determine if any non-active paths (for example, a secondary path or primary path during reversion) that the system might switch to are up and forwarding.

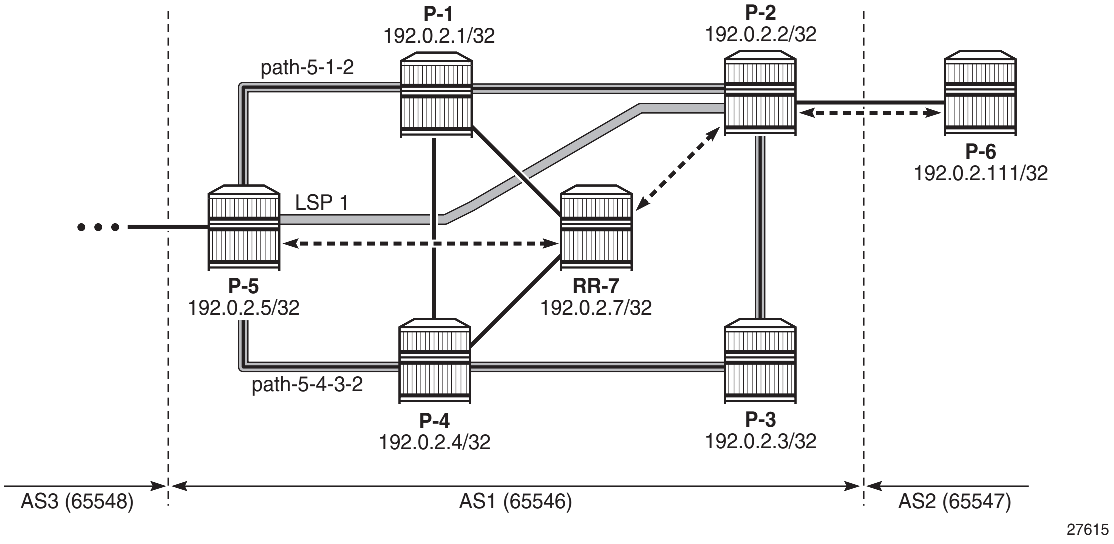
When BFD failure-action is configured on an RSVP-TE LSP directly, the action can be failover or down. When BFD failure-action is configured on an RSVP-TE LSP indirectly, through an LSP template, the only action available is down. This chapter only covers the direct configuration of a failure-action.

Configuration

Failure-Action Failover

Figure 192 shows the topology used for failure-action failover. A BGP shortcut is defined in AS 65545 running between the AS Border Routers (ASBRs) P-5 and P-2. That shortcut is an RSVP-TE LSP composed of two paths, where the first path is the upper path from P-5 via P-1 to P-2, and the second path is the lower path from P-5 via P-4 and P-3 to P-2.

Figure 192 **Failure-Action Failover Topology**



The initial configuration includes:

- Cards, MDAs, and ports
- Router interfaces

- IS-IS as IGP on all interfaces (alternatively, OSPF can be used), with traffic engineering enabled
- MPLS and RSVP-TE enabled on all interfaces
- BGP configured, with RR-7 being the route reflector for clients P-5 and P-2 of AS 65545, and P-6 located in AS 65546 and connected to P-2. P-6 advertises its prefix 192.0.2.111/32 to AS 65545.

The *lsp-1* LSP from Figure 1 is defined as follows. The primary path is *path-5-1-2*, and *path-5-4-3-2* is the secondary standby path; the two paths are established when the LSP is brought up, to minimize traffic loss in case of a failure. BFD template *bfdt-1* with a failure-action failover is applied to *lsp-1* at the LSP level.

```
# on P-5
configure
    router
        mpls
            lsp "lsp-1"
                to 192.0.2.2
                cspf
                bfd
                    bfd-template "bfdt-1"
                    bfd-enable
                    failure-action failover
                exit
                primary "path-5-1-2"
                exit
                secondary "path-5-4-3-2"
                    standby
                exit
                no shutdown
            exit
        exit
    exit
exit
```

The details of the LSP show the configured failure action, as follows:

```
*A:P-5# show router mpls lsp "lsp-1" detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : lsp-1
LSP Type        : RegularLsp                LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.5                 To                   : 192.0.2.2
Adm State       : Up                        Oper State           : Up
LSP Up Time     : 0d 00:15:29               LSP Down Time        : 0d 00:00:00
```

```
Transitions     : 1                         Path Changes        : 1
Retry Limit     : 0                         Retry Timer         : 30 sec
--- snipped ---
Least Fill      : Disabled
BFD Template    : bfdt-1                    BFD Ping Intvl      : 60
BFD Enable      : True                      BFD Failure-action  : Failover

--- snipped ---

Primary(a)      : path-5-1-2                Up Time             : 0d 00:15:29
Bandwidth       : 0 Mbps
Standby         : path-5-4-3-2              Up Time             : 0d 00:15:29
Bandwidth       : 0 Mbps
===============================================================================
*A:P-5#
```

With this configuration, the BFD session is running over the upper path, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-1-2                                    Up          491        494
  192.0.2.2                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

BGP route 192.0.2.111/32 is advertised by P-6 out of AS 65546, as follows. This route is learned by P-5 via RR7.

```
*A:P-5# show router bgp routes
===============================================================================
 BGP Router ID:192.0.2.5       AS:65545        Local AS:65545
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
```

```
-------------------------------------------------------------------------------
u*>i  192.0.2.111/32                                     100         None
      192.0.2.2                                          1           -
      65546
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:P-5#
```

To keep the core of AS 65545 BGP free, traffic is tunneled through the *lsp-1* LSP, as follows:

```
*A:P-5# show router route-table 192.0.2.111/32
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.111/32                                Remote  BGP       00h41m40s  170
       192.0.2.2 (tunneled:RSVP:1)                                  0
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-5#
```

The LSP and path details can also be shown using a **tools dump** command, as follows. LSP lsp-1 is up, *path-5-1-2* is the active path taking three hops, so the operational metric is 20, and *path-5-4-3-2* is the standby path, but not active.

```
*A:P-5# tools dump router mpls lspinfo "lsp-1" detail
LSP "lsp-1"  LspIdx 1  LspType Dynamic  State LSPS_UP  Flags 0x2000
NumPaths 2  NumSdps 0  NumCBFSdps 0  NumFltrEntries 0
HoldTimeRemaining 0secs  ClassType 0  Metric 0  OperMetric 20
LDPoRsvp Include  VprnAutoBind Include IgpShortCut Include BgpShortCut Include
BgpTransTunnel Include IpShCutTtlPropLocal TRUE IpShCutTtlPropTans TRUE
RelativeMetricOffset 2147483647 EntropyLbl inherit MTU 1564 InUseByLdp FALSE
LspAdminState :2 LspOperState : 2 lspRowStatus : 1
ClassForwarding: Disabled
BFD Enabled  Template bfdt-1  PingInterval 60 Failure-Action Failover
Path Profile:
  None
  Path "path-5-1-2"  LspId 47628  PathType Primary  ActivePath Yes
  RowStatus 1  LastChange 000 01:49:00.620 AdminState :2 OperState :2
                                                OperStateChange 000 01:20:57.810
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.15.2 EgrLnkId 0
                                                        RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.15.1 IngLnkId 0 EgrIp 192.168.12.1 EgrLnkId 0
                                                        RtrId 192.0.2.1 Flag 0x0
      Hop[3] IngIp 192.168.12.2 IngLnkId 0 EgrIp 192.0.2.2 EgrLnkId 0
                                                        RtrId 192.0.2.2 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x40000000
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 0 HopLimit 255 BW 0Mbps
    TotIgpCost 20 OperMetric 20 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 20  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No
  Path "path-5-4-3-2"  LspId 47626  PathType Standby  ActivePath No
  RowStatus 1  LastChange 000 01:49:00.620 AdminState :2 OperState :2
                                               OperStateChange 000 01:26:55.810
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.45.2 EgrLnkId 0
                                                       RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.45.1 IngLnkId 0 EgrIp 192.168.34.2 EgrLnkId 0
                                                       RtrId 192.0.2.4 Flag 0x0
      Hop[3] IngIp 192.168.34.1 IngLnkId 0 EgrIp 192.168.23.2 EgrLnkId 0
                                                       RtrId 192.0.2.3 Flag 0x0
      Hop[4] IngIp 192.168.23.1 IngLnkId 0 EgrIp 192.0.2.2 EgrLnkId 0
                                                       RtrId 192.0.2.2 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x0
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 255 HopLimit 255 BW 0Mbps
    TotIgpCost 30 OperMetric 30 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 30  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No

Total Ingress LSP Count         : 1
*A:P-5#
```

Bringing down the link between P-1 and P-2 results in the secondary path, *path-5-4-3-2* of LSP *lsp-1*, becoming active, and the BFD session is re-established on that path, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
```

```
-------------------------------------------------------------------------------
lsp-1::path-5-4-3-2                                  Up            43         46
  192.0.2.2                                           5          2000       2000
  rsvpLsp                                       central           N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

BFD sessions changing state are logged in the trap log, as follows:

```
137 2018/07/12 14:09:03.069 CEST WARNING: MPLS #2010 Base VR 1:
"LSP lsp-2 is operationally disabled ('shutdown') because noPathIsOperational"

138 2018/07/12 14:09:03.070 CEST MINOR: BFD #2004 Base 192.0.2.3
"The protocol (RSVP LSP) using BFD session on node 192.0.2.3 has been cleared"

139 2018/07/12 14:09:03.070 CEST MINOR: BFD #2003 Base 192.0.2.3
"The lspHead BFD Session with Local Discriminator 10 on 192.0.2.3 has been deleted"

140 2018/07/12 14:13:21.406 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (RSVP LSP) using BFD session on node 192.0.2.2 has been cleared"

141 2018/07/12 14:13:21.406 CEST MINOR: BFD #2003 Base 192.0.2.2
"The lspHead BFD Session with Local Discriminator 9 on 192.0.2.2 has been deleted"

142 2018/07/12 14:13:31.510 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (RSVP LSP) using BFD session on node 192.0.2.2 has been added"

143 2018/07/12 14:13:35.895 CEST MINOR: BFD #2002 Base 192.0.2.2
"The lspHead BFD session with Local Discriminator 11 on 192.0.2.2 is up"

144 2018/07/12 14:49:31.106 CEST WARNING: MPLS #2012 Base VR 1:
"LSP path lsp-1::path-5-1-2 is operationally disabled ('shutdown') because noError"

145 2018/07/12 14:49:31.107 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (RSVP LSP) using BFD session on node 192.0.2.2 has been cleared"

146 2018/07/12 14:49:31.107 CEST MINOR: BFD #2003 Base 192.0.2.2
"The lspHead BFD Session with Local Discriminator 11 on 192.0.2.2 has been deleted"

147 2018/07/12 14:49:31.107 CEST MINOR: BFD #2004 Base 192.0.2.2
"The protocol (RSVP LSP) using BFD session on node 192.0.2.2 has been added"

148 2018/07/12 14:49:35.496 CEST MINOR: BFD #2002 Base 192.0.2.2
"The lspHead BFD session with Local Discriminator 12 on 192.0.2.2 is up"
```

The **tools dump** command shows that *lsp-1* is still up, and *path-5-4-3-2* is active with four hops, so the LSP operational metric is 30, as follows:

```
*A:P-5# tools dump router mpls  lspinfo "lsp-1" detail
LSP "lsp-1"  LspIdx 1  LspType Dynamic  State LSPS_UP  Flags 0x2000
NumPaths 2  NumSdps 0  NumCBFSdps 0  NumFltrEntries 0
HoldTimeRemaining 0secs  ClassType 0  Metric 0  OperMetric 30
LDPoRsvp Include  VprnAutoBind Include IgpShortCut Include BgpShortCut Include
BgpTransTunnel Include IpShCutTtlPropLocal TRUE IpShCutTtlPropTans TRUE
RelativeMetricOffset 2147483647 EntropyLbl inherit MTU 1564 InUseByLdp FALSE
```

```
LspAdminState : 2 LspOperState : 2 lspRowStatus : 1
ClassForwarding: Disabled
BFD Enabled  Template bfdt-1  PingInterval 60 Failure-Action Failover
Path Profile:
  None
Admin Tags:
  None
  Path "path-5-1-2"  LspId 47630  PathType Primary  ActivePath No
  RowStatus 1  LastChange 000 02:08:08.210  AdminState :2  OperState :3
                                                  OperStateChange 000 00:09:05.190
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_DOWN  Flags 0x0
    RetryAttempts 18  RetryInterval 30  NextRetryIn 1secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 0 HopLimit 255 BW 0Mbps
    TotIgpCost 0 OperMetric 16777215 MTU 0
    BFD Disabled  Template n/a  PingInterval 60
  Path "path-5-4-3-2"  LspId 47626  PathType Standby  ActivePath Yes
  RowStatus 1  LastChange 000 02:08:08.210  AdminState :2  OperState :2
                                                  OperStateChange 000 01:46:03.400
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.45.2 EgrLnkId 0
                                                          RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.45.1 IngLnkId 0 EgrIp 192.168.34.2 EgrLnkId 0
                                                          RtrId 192.0.2.4 Flag 0x0
      Hop[3] IngIp 192.168.34.1 IngLnkId 0 EgrIp 192.168.23.2 EgrLnkId 0
                                                          RtrId 192.0.2.3 Flag 0x0
      Hop[4] IngIp 192.168.23.1 IngLnkId 0 EgrIp 192.0.2.2 EgrLnkId 0
                                                          RtrId 192.0.2.2 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x40000000
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 255 HopLimit 255 BW 0Mbps
    TotIgpCost 30 OperMetric 30 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 30  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No

Total Ingress LSP Count         : 1
*A:P-5#
```

While the system is busy establishing the secondary path because that path was not established yet, and before the BFD session is re-established on that secondary path, the LSP is in the degraded state, as follows. This is particularly apparent when the secondary path is defined as a standby path:

```
*A:P-5# tools dump router mpls lspinfo "lsp-1"
LSP "lsp-1"  LspIdx 1  LspType Dynamic  State LSPS_DEGRADED  Flags 0x2000
NumPaths 2  NumSdps 0  NumCBFSdps 0  NumFltrEntries 0
HoldTimeRemaining 0secs  ClassType 0  Metric 0  OperMetric 20
LDPoRsvp Include  VprnAutoBind Include IgpShortCut Include BgpShortCut Include
BgpTransTunnel Include IpShCutTtlPropLocal TRUE IpShCutTtlPropTans TRUE
RelativeMetricOffset 2147483647 EntropyLbl inherit MTU 1564 LspAdminState :2
```

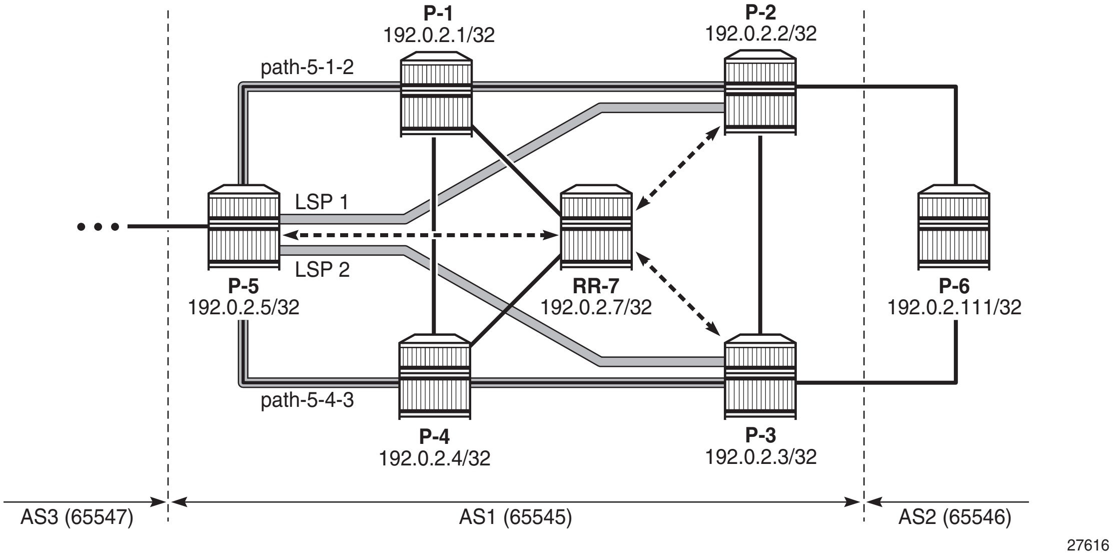
```
LspOperState : 2 lspRowStatus : 1
ClassForwarding: Disabled
BFD Enabled  Template bfdt-1  PingInterval 60 Failure-Action Failover

Total Ingress LSP Count         : 1
*A:P-5#
```

In summary, the secondary path, *path-5-4-3-2*, becoming active does not result in a change of the BGP next-hop. Traffic continues to flow from P-5 to P-6 via LSP *lsp-1*, but now via the lower path. The BFD failure-action failover combined with standby secondary paths can help detect failures faster, with minimal traffic loss, which is especially useful in larger domains, or when the LSP passes through multiple domains.

Failure-Action Down

Figure 193 shows the topology used for failure-action down. BGP shortcuts are defined in AS 65545 running between the ASBRs P-5, P-2, and P-3. A first shortcut is offered through an RSVP-TE LSP called *lsp-1*, with a single path from P-5 via P-1 to P-2; the second shortcut is offered through another RSVP-TE LSP called lsp-2, with a single path from P-5 via P-4 to P-3. When *lsp-1* fails and the failure gets detected by BFD, traffic starts using *lsp-2*, implying a change of the BGP next-hop; this scenario being an edge Prefix-Independent Convergence (PIC) scenario. See the BGP Fast Reroute chapter for more information about edge PIC.

Figure 193 **Failure-Action Down Topology**



The initial configuration includes:

- Cards, MDAs, and ports
- Router interfaces
- IS-IS as IGP on all interfaces (alternatively, OSPF can be used), with traffic engineering enabled
- MPLS and RSVP-TE enabled on all interfaces
- BGP configured, with RR-7 being the route reflector for clients P-5 and P-2 in AS 65545, and P-6 located in AS 65546 and connected to P-2 and P-3. P-6 reports prefix 192.0.2.111/32 to P-2 and P-3.

The LSPs from Figure 2 are configured as follows. The paths referred to from these LSPs are fully strict paths, using interface IP addresses. Only *lsp-1* has BFD enabled, and failure-action down configured.

```
# on P-5
configure
    router
        mpls
            lsp "lsp-1"
                to 192.0.2.2
                cspf
                bfd
                    bfd-template "bfdt-1"
                    bfd-enable
                    failure-action down
                exit
                primary "path-5-1-2"
                exit
                secondary "path-5-4-3-2"
                    standby
                exit
                no shutdown
            exit
            lsp "lsp-2"
                to 192.0.2.3
                cspf
                primary "path-5-4-3"
                exit
                no shutdown
            exit
        exit
    exit
exit
```

The details of the LSP show the configured failure-action, as follows:

```
*A:P-5# show router mpls lsp "lsp-1" detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : lsp-1
LSP Type        : RegularLsp                LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.5                 To                   : 192.0.2.2
Adm State       : Up                        Oper State           : Up
LSP Up Time     : 0d 02:07:58               LSP Down Time        : 0d 00:00:00
Transitions     : 5                         Path Changes         : 12
Retry Limit     : 0                         Retry Timer          : 30 sec
Signaling       : RSVP                      Resv. Style          : SE
Hop Limit       : 255                       Negotiated MTU       : 1564
Adaptive        : Enabled                   ClassType            : 0
FastReroute     : Disabled                  Oper FR              : Disabled
CSPF            : Enabled                   ADSPEC               : Disabled
Metric          : N/A                       Use TE metric        : Disabled
Load Bal Wt     : N/A                       ClassForwarding      : Disabled
Include Grps    :                           Exclude Grps         :
None                                           None
Least Fill      : Disabled
BFD Template    : bfdt-1                    BFD Ping Intvl       : 60
BFD Enable      : True                      BFD Failure-action   : Down

Revert Timer    : Disabled                  Next Revert In       : N/A
Entropy Label   : Enabled+                  Oper Entropy Label   : Enabled
Negotiated EL   : Disabled
Auto BW         : Disabled
LdpOverRsvp     : Enabled
VprnAutoBind    : Enabled
IGP Shortcut    : Enabled                   BGP Shortcut         : Enabled
IGP LFA         : Disabled                  IGP Rel Metric       : Disabled
BGPTransTun     : Enabled
Oper Metric     : 20
Prop Adm Grp    : Disabled
PCE Report      : Disabled+
PCE Compute     : Disabled                  PCE Control          : Disabled
Path Profile    : None
Admin Tags      : None

Primary(a)      : path-5-1-2                Up Time              : 0d 00:14:16
Bandwidth       : 0 Mbps
Standby         : path-5-4-3-2              Up Time              : 0d 02:07:45
Bandwidth       : 0 Mbps
===============================================================================
*A:P-5#
```

Multiple BGP paths are available out of P-5 to reach P-6, as follows. The path via P-2 is the currently active path, the path via P-3 is the standby path.

```
*A:P-5# show router route-table protocol bgp alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
```

```
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.111/32                                Remote  BGP       00h18m21s  170
       192.0.2.2 (tunneled:RSVP:1)                                  0
192.0.2.111/32 (Backup)                       Remote  BGP       00h18m21s  170
       192.0.2.3 (tunneled:RSVP:2)                                  0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:P-5#
```

The BFD session is running over the active path, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts   Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl  Rx Intvl
  Protocols                                        Type     LAG Port    LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-1-2                                    Up           87         90
  192.0.2.2                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

The BFD session running over the active path is also indicated in the output of the **tools** command, as follows:

```
*A:P-5# tools dump router mpls lspinfo "lsp-1" detail
LSP "lsp-1"  LspIdx 1  LspType Dynamic  State LSPS_UP  Flags 0x2000
NumPaths 2  NumSdps 0  NumCBFSdps 0  NumFltrEntries 0
HoldTimeRemaining 0secs  ClassType 0  Metric 0  OperMetric 20
LDPoRsvp Include  VprnAutoBind Include IgpShortCut Include BgpShortCut Include
BgpTransTunnel Include IpShCutTtlPropLocal TRUE IpShCutTtlPropTans TRUE
RelativeMetricOffset 2147483647 EntropyLbl inherit MTU 1564 InUseByLdp FALSE
LspAdminState : 2 LspOperState : 2 lspRowStatus : 1
ClassForwarding: Disabled
BFD Enabled  Template bfdt-1  PingInterval 60 Failure-Action Down
Path Profile:
  None
Admin Tags:
  None
  Path "path-5-1-2"  LspId 47630  PathType Primary  ActivePath Yes
  RowStatus 1  LastChange 000 03:09:11.210  AdminState :2  OperState :2
                                              OperStateChange 000 00:53:37.400
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.15.2 EgrLnkId 0
                                                      RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.15.1 IngLnkId 0 EgrIp 192.168.12.1 EgrLnkId 0
                                                      RtrId 192.0.2.1 Flag 0x0
      Hop[3] IngIp 192.168.12.2 IngLnkId 0 EgrIp 192.0.2.2 EgrLnkId 0
                                                      RtrId 192.0.2.2 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x40000000
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 0 HopLimit 255 BW 0Mbps
    TotIgpCost 20 OperMetric 20 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 20  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No
  Path "path-5-4-3-2"  LspId 47632  PathType Standby  ActivePath No
  RowStatus 1  LastChange 000 00:04:41.000  AdminState :2  OperState :2
                                              OperStateChange 000 00:04:40.990
    TE Computed Hop List:
      Hop[1] IngIp 192.0.2.5 IngLnkId 0 EgrIp 192.168.45.2 EgrLnkId 0
                                                      RtrId 192.0.2.5 Flag 0x0
      Hop[2] IngIp 192.168.45.1 IngLnkId 0 EgrIp 192.168.34.2 EgrLnkId 0
                                                      RtrId 192.0.2.4 Flag 0x0
      Hop[3] IngIp 192.168.34.1 IngLnkId 0 EgrIp 192.168.23.2 EgrLnkId 0
                                                      RtrId 192.0.2.3 Flag 0x0
      Hop[4] IngIp 192.168.23.1 IngLnkId 0 EgrIp 192.0.2.2 EgrLnkId 0
                                                      RtrId 192.0.2.2 Flag 0x0
    Reported to PCE: No, Delegated to PCE: No
    LspPath FsmState LSP_PATH_S_UP  Flags 0x0
    RetryAttempts 0  RetryInterval 30  NextRetryIn 0secs
    Class Type 0 SetupPri 7 HoldPri 0 Pref 255 HopLimit 255 BW 0Mbps
    TotIgpCost 30 OperMetric 30 MTU 1564
    BFD Disabled  Template n/a  PingInterval 60
    Oper Values:
        Class Type 0 SetupPri 7 HoldPri 0 HopLimit 255 BW 0Mbps
        RecordRoute RecordLabel No Adspec
        No PropagateAdminGroup Exclude 0x00000000 Include 0x00000000
        No FRR
        Metric 30  CSPF No Least Fill Intra-area
        PCE-Computed No PCE-Reported No PCE-Controlled No

Total Ingress LSP Count         : 1
*A:P-5#
```

Emulating a path failure by bringing down port 1/1/2 on P-2 leads to the secondary path becoming active, as follows:

```
*A:P-5# show router mpls lsp "lsp-1" path

===============================================================================
MPLS LSP lsp-1 Path
===============================================================================
-------------------------------------------------------------------------------
LSP Name         : lsp-1
To               : 192.0.2.2
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
path-5-1-2                       n/a             Primary    n/a       Up   Dwn
path-5-4-3-2                     192.168.45.1    Standby    1/1/2     Up   Up
===============================================================================
*A:P-5#
```

The BFD session on the now unused *lsp-1* changes from *path-5-1-2* to *path-5-4-3-2*, as follows.

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-4-3-2                                  Up           58         62
  192.0.2.2                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 1
===============================================================================
*A:P-5#
```

Even though the secondary path is now active, because of the failure-action down declaration on *lsp-1*, *lsp-1* is registered as unusable in the TTM, so BGP traffic is diverted into *lsp-2*, and the BGP next-hop changes, as follows:

```
*A:P-5# show router route-table protocol bgp alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.111/32 (Backup)                       Remote  BGP       00h03m41s  170
       192.0.2.2 (tunneled:RSVP:1)                                  0
192.0.2.111/32                                Remote  BGP       00h03m41s  170
       192.0.2.3 (tunneled:RSVP:2)                                  0
```

```
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:P-5#
```

Because a BFD session running on a secondary unused path can be confusing to operators and is taking up resources, Nokia recommends redefining the LSPs to only use a single path, as follows:

```
# on P-5
configure
    router
        mpls
            lsp "lsp-1"
                to 192.0.2.2
                cspf
                bfd
                    bfd-template "bfdt-1"
                    bfd-enable
                    failure-action down
                exit
                primary "path-5-1-2"
                exit
                no shutdown
            exit
            lsp "lsp-2"
                to 192.0.2.3
                cspf
                bfd
                    bfd-template "bfdt-1"
                    bfd-enable
                    failure-action down
                exit
                primary "path-5-4-3"
                exit
                no shutdown
            exit
        exit
    exit
exit
```

Monitoring both LSPs with BFD provides an even higher level of security.

With this configuration applied, two BFD sessions are active in a non-failure condition of the network, as follows:

```
*A:P-5# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
```

```
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
lsp-1::path-5-1-2                                    Up          157        159
  192.0.2.2                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
lsp-2::path-5-4-3                                    Up          148        152
  192.0.2.3                                           5         2000       2000
  rsvpLsp                                       central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 2
===============================================================================
*A:P-5#
```

Conclusion

The BFD failure-action failover and down, optionally combined with standby secondary paths, can help detect failures faster with minimal traffic loss on switchover, which is especially useful in larger domains or when the LSP passes through multiple domains.

Class-Based Forwarding

This chapter provides information about class-based forwarding

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS routers and was initially written for release 13.0.R7. The CLI in this edition corresponds to release 14.0.R1.

Before SR OS 13.0.R6, class-based forwarding (CBF) was only applicable to resource reservation protocol (RSVP) label switched paths (LSPs) in multi-LSP service distribution points (SDPs) used by services. In release 13.0.R6, CBF of label distribution protocol (LDP) prefix packets over interior gateway protocol (IGP) shortcuts was introduced. Both implementations are described in this chapter.

Overview

In large networks, services are typically required from any PE to any other PE, and can traverse multiple domains. Within a service, different traffic classes can coexist, each with different requirements for latency and jitter.

With CBF, packets with different forwarding classes (FCs) can be forwarded on different LSPs. When equal-cost multipath (ECMP) routing without CBF is used, packets are distributed over the whole set of LSPs, without any distinction for the FC. CBF is based on FC, not on the traffic being in-profile or out-of-profile, and is a local decision, which makes it easier for interoperability.

This feature may be useful when certain links have less bandwidth and should only be used for high priority traffic. A service provider might decide to send real-time traffic over a shorter path with less bandwidth, while the bulk of the traffic takes a longer path with more bandwidth.

Configuration

The initial implementation for CBF was based on RSVP LSPs. This is described first, followed by CBF of LDP prefix packets over IGP shortcuts.

CBF over RSVP LSPs

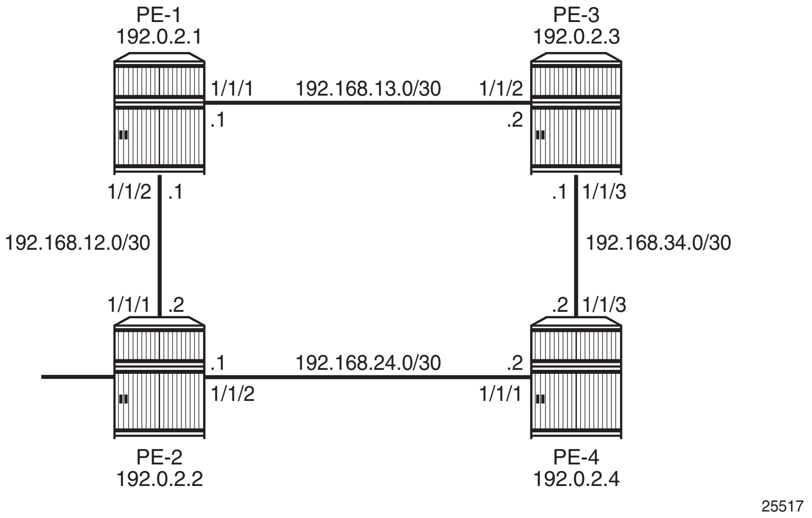
CBF over RSVP LSPs allows a service packet to be forwarded over a specific RSVP LSP, part of an SDP, based on its service ingress determined FC, typically controlled by a default or operator-defined **sap-ingress** policy. A default LSP is configured for all traffic that does not match the FCs that are explicitly configured in the SDP. CBF over RSVP LSPs is also used to forward a received packet that has been classified at the ingress service access point (SAP) into an FC, if the LSP that supports its FC is not available.

Note: CBF can also be enabled on static LSPs in the same way, but that scenario is not common.

A multicast LSP is configured for broadcast, unknown, and multicast (BUM) traffic. When no multicast LSP is defined, BUM traffic uses the default LSP. Because there are multiple LSPs per SDP, CBF over RSVP LSPs increases the number of RSVP sessions.

The test topology is shown in Figure 194. This topology will be extended in a subsequent use case.

Figure 194 **Test Topology for CBF on RSVP LSPs**



Initial Configuration

All nodes have the following initial configuration:

- Cards, media dependent adapters (MDAs), ports
- Router interfaces
- IGP open shortest path first (OSPF) or intermediate system to intermediate system (IS-IS)
- Multiprotocol label switching (MPLS) enabled on all router interfaces
- RSVP enabled
- RSVP LSPs

As an example, the configuration on PE-1 is shown. The configuration for PE-2 is similar.

```
*A:PE-1# configure router
        interface "int-PE-1-P-3"
            address 192.168.13.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
```

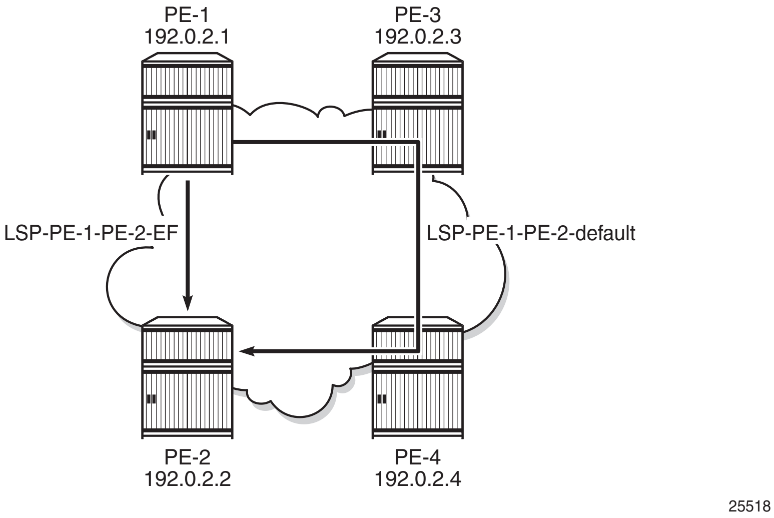
```
*A:PE-1# configure router
        ospf
            traffic-engineering
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-1-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-1-P-3"
                    interface-type point-to-point
                exit
            exit
        exit
*A:PE-1# configure router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-P-3"
            exit
            no shutdown
        exit
*A:PE-1# configure router rsvp no shutdown
```

The RSVP LSPs use paths with strict hops:

```
*A:PE-1# configure router
        mpls
            path "direct-PE-1-PE-2"
                hop 10 192.168.12.2 strict
                no shutdown
            exit
            path "indirect-PE-1-PE-2"
                hop 10 192.168.13.2 strict
                hop 20 192.168.34.2 strict
                hop 30 192.168.24.1 strict
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2-EF"
                to 192.0.2.2
                primary "direct-PE-1-PE-2"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2-default"
                to 192.0.2.2
                primary "indirect-PE-1-PE-2"
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The configured RSVP LSPs are shown in Figure 195.

Figure 195 **LSPs with Direct and Indirect Path toward PE-2**



Configure SDPs

In the initial configuration, the traffic is by default sent on LSP "LSP-PE-1-PE-2-default", except when the FC is expedited forwarding (EF). The traffic with FC EF is sent over LSP "LSP-PE-1-PE-2-EF". Both LSPs are assigned to SDP 122 on PE-1:

```
*A:PE-1# configure service
        sdp 122 mpls create
            description "SDP-PE-1-PE-2"
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2-EF"
            lsp "LSP-PE-1-PE-2-default"
            path-mtu 1514
            class-forwarding default-lsp "LSP-PE-1-PE-2-default"
                fc "ef" lsp "LSP-PE-1-PE-2-EF"
                no shutdown
            exit
            no shutdown
        exit
```

Note: A change in the CBF configuration may result in a change of forwarding behavior.

The configuration of SDP 212 from PE-2 to PE-1 is similar.

For the SDPs to get operational up, targeted-LDP (T-LDP) must be enabled on the PE nodes:

```
*A:PE-1# configure router ldp no shutdown
```

The state of the SDP can be verified as follows:

```
*A:PE-1# show service sdp
============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
122    1514    1514    192.0.2.2        Up   Up          MPLS    R     TLDP
----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE
============================================================================
*A:PE-1#
```

Some Considerations

- When CBF is enabled, a default LSP must be configured. An error is raised when class-forwarding is enabled without a default LSP:

```
*A:PE-1>config>service>sdp# class-forwarding
MINOR: CLI Default LSP must be specified.
```

- Only one LSP can be configured as the default LSP in the SDP context. Configuring another LSP as the default LSP overwrites the originally configured LSP.
- Only one LSP can be assigned to an FC in the SDP context. Configuring another LSP for an FC overwrites the originally configured LSP.
- An LSP can be assigned to multiple FCs in the SDP context.
- The default LSP can also be assigned to one or more FCs in the SDP context.

```
*A:PE-1>config>service# info
----------------------------------------------
        sdp 122 mpls create
            description "RSVP-SDP-PE-1-PE-2"
            far-end 192.0.2.2
            lsp "LSP-PE-1-PE-2-EF"
            lsp "LSP-PE-1-PE-2-default"
            path-mtu 1514
            keep-alive
                shutdown
            exit
            class-forwarding default-lsp "LSP-PE-1-PE-2-default"
                fc "af" lsp "LSP-PE-1-PE-2-default"
                fc "be" lsp "LSP-PE-1-PE-2-default"
                fc "ef" lsp "LSP-PE-1-PE-2-EF"
                multicast-lsp "LSP-PE-1-PE-2-default"
                no shutdown
            exit
            no shutdown
        exit
```

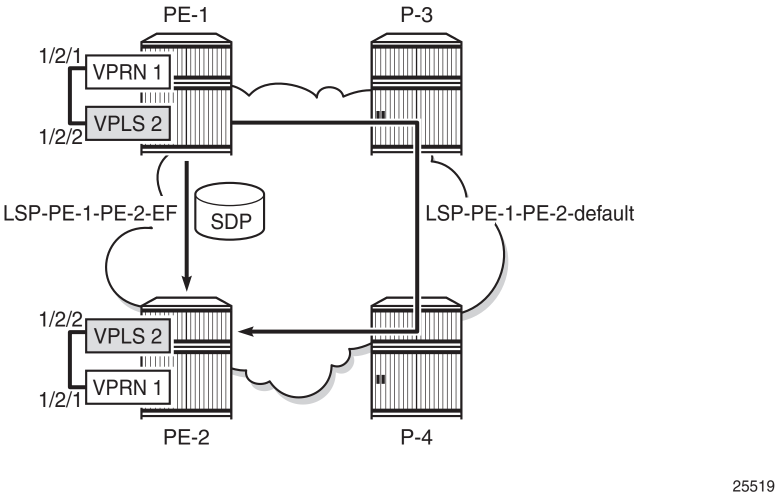
- The SDP goes down when the default LSP goes down.
- When CBF is configured on the LSP/LSP template as well as on the SDP, the configuration on the LSP/LSP template is ignored. Configuring CBF in the LSP/LSP template context is required in another use case: CBF of LDP Prefix Packets over IGP Shortcuts. This is described later in this chapter.

Configure Services for Traffic Verification

To verify that EF traffic is sent over the direct link while traffic with a different FC is sent over the longer path, two services are configured on PE-1 and PE-2.

To generate traffic, ping messages will be sent in virtual private routed network 1 (VPRN 1). The outgoing traffic is sent to virtual private LAN service 2 (VPLS 2) where the FC can be modified to EF if needed. VPLS 2 has the spoke SDP with the different LSPs and CBF. This is shown in Figure 196.

Figure 196 **CBF on RSVP LSPs - Services**



```
*A:PE-1# configure service
        vprn 1 customer 1 create
            route-distinguisher 64496:11
            vrf-target target:64496:1
            interface "loopback1" create
                address 172.31.1.1/32
                loopback
            exit
            interface "int-PE-1-PE-2-VPRN1" create
                address 192.168.112.1/30
                sap 1/2/1 create
                exit
            exit
            static-route-entry 172.31.2.1/32
                next-hop 192.168.112.2
                    no shutdown
                exit
            exit
            no shutdown
        exit
*A:PE-1# configure service
        vpls 2 customer 1 create
            description "VPLS to modify FC for traffic from VPRN 1"
            sap 1/2/2 create
            exit
            spoke-sdp 122:2 create
            exit
            no shutdown
        exit
```

The configuration of the services on PE-2 is similar.

The operational state of the spoke SDP can be verified as follows:

```
*A:PE-1# show service sdp-using
===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End              Opr   I.Label E.Label
                                                          State
-------------------------------------------------------------------------------
2          122:2              Spok   192.0.2.2            Up    262139  262139
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

Verify CBF on LSP for Specific FC

To verify that EF traffic is sent out on "LSP-PE-1-PE-2-EF", via port 1/1/2 on PE-1, the configuration of VPLS 2 needs to be modified. The default FC is set to EF.

The SAP ingress policy to set the FC to EF is applied in VPLS 2 on PE-1 as follows:

```
*A:PE-1# configure qos sap-ingress 2 create
```

```
            default-fc ef
            exit
*A:PE-1# configure service
        vpls 2
            sap 1/2/2 create
                ingress qos 2
            exit
        exit
```

The configuration is identical for PE-2.

First, the port statistics are cleared.

```
*A:PE-1# clear port 1/1/[1..2] statistics
```

One thousand ping messages will be sent from VPRN 1 on PE-1 to the loopback address in VPRN 1 on PE-2:

```
*A:PE-1# ping router 1 172.31.2.1 rapid count 1000
```

The port statistics are verified. The FC is equal to EF, so the LSP "LSP-PE-1-PE-2-EF" is used and the traffic is sent via port 1/1/2.

```
*A:PE-1# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                        21            1740              21            1708
===============================================================================
*A:PE-1# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      1033          126954            1033          126898
===============================================================================
```

Note: The traffic on the unused port (1/1/1) is not strictly equal to zero and the traffic on the used port (1/1/2) is not strictly equal to 1000, because of other traffic, such as OSPF messages.

Verify CBF on Default LSP

The SAP ingress policy to define the FC is removed from SAP 1/2/2 of VPLS 2. Traffic that enters a SAP where no SAP ingress policy with FC is defined always gets FC best effort (BE). Traffic with FC BE will be sent on the default LSP "LSP-PE-1-PE-2-default" on port 1/1/1 on PE-1, not on the direct link to PE-2.

```
*A:PE-1# configure service vpls 2 sap 1/2/2 ingress no qos 2
*A:PE-1# clear port 1/1/[1..2] statistics
*A:PE-1# ping router 1 172.31.2.1 rapid count 1000
```

The port statistics are verified. The FC is not equal to EF, so the default LSP is used and the traffic is sent via port 1/1/1.

```
*A:PE-1# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1015          125284            1015          125284
===============================================================================
*A:PE-1# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        12            1068              12            1068
===============================================================================
```

Because no LSP has been configured to transport packets of classes other than EF, the default LSP will be used for all traffic classes different from EF. A similar outcome occurs when the FC of the ping messages is set to assured forwarding (AF), or any other FC different from EF.

The SAP ingress policy to set the FC to AF is applied in VPLS 2 on PE-1 as follows:

```
*A:PE-1# configure qos sap-ingress 2 create
            default-fc af
            exit
*A:PE-1# configure service
        vpls 2
            sap 1/2/2 create
                ingress qos 2
            exit
        exit
```

The configuration is identical for PE-2.

The port statistics are cleared and ping messages are sent. The result shows that the same port (that is, 1/1/1) is used for the generated traffic:

```
*A:PE-1# clear port 1/1/[1..2] statistics
*A:PE-1# ping router 1 172.31.2.1 rapid count 1000
*A:PE-1# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1005          124342            1007          124750
===============================================================================
*A:PE-1# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        10             846              10             864
===============================================================================
*A:PE-1#
```

As indicated, the default LSP is used for all traffic with an FC that is not mapped to a specific LSP, which includes the case where a mapping FC-to-LSP has been defined but the LSP is not available. In the current example, all traffic with FC EF will be sent on LSP "LSP-PE-1-PE-2-EF", unless that LSP is unavailable. The default LSP will carry traffic with FC EF in that latter case.

The SAP ingress policy to assign the FC EF is applied in VPLS 2 on PE-1 as follows:

```
*A:PE-1# configure qos sap-ingress 2 create
            default-fc ef
            exit
*A:PE-1# configure service
        vpls 2
            sap 1/2/2 create
                ingress qos 2
            exit
        exit
```

The configuration is identical for PE-2.

To make the LSP "LSP-PE-1-PE-2-EF" unavailable, it is put in the shutdown state on PE-1. LSP "LSP-PE-2-PE-1-EF" on PE-2 is also put in the shutdown state.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2-EF" shutdown
*A:PE-2# configure router mpls lsp "LSP-PE-2-PE-1-EF" shutdown
```

The port statistics are cleared and ping messages are sent. The packets are classified as FC EF in VPLS 2:

```
*A:PE-1# clear port 1/1/[1..2] statistics
*A:PE-1# ping router 1 172.31.2.1 rapid count 1000
```

Because LSP "LSP-PE-1-PE-2-EF" (which uses port 1/1/2) is not operational, the traffic is sent on the default LSP on port 1/1/1 instead:

```
*A:PE-1# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1010          124858            1009          124620
===============================================================================
*A:PE-1# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        14            1046              14            1046
===============================================================================
```

The LSP "LSP-PE-1-PE-2-EF" is re-enabled and the default LSP is now disabled. In this case, the SDP goes down. Ping messages will time out. The SAP ingress quality of service (QoS) policy in VPLS 2 is removed and the FC of the ping messages will no longer be changed to EF.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2-EF" no shutdown
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2-default" shutdown
*A:PE-1# configure service vpls 2 sap 1/2/2 ingress no qos 2
```

The SDP is operational down when the default LSP is down, as can be verified:

```
*A:PE-1# show service sdp-using
===============================================================================
SDP Using
===============================================================================
SvcId      SdpId              Type   Far End              Opr   I.Label E.Label
                                                          State
-------------------------------------------------------------------------------
2          122:2              Spok   192.0.2.2            Down  262139  262139
-------------------------------------------------------------------------------
Number of SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The default LSP is re-enabled.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2-default" no shutdown
```

Define Multicast LSP for BUM Traffic

The multicast LSP specifies the LSP in the SDP to use to forward BUM traffic. The LSP name must exist and must have been associated with this SDP. In this example, the default LSP is configured as the multicast LSP.

```
*A:PE-1# configure service sdp 122 class-forwarding multicast-lsp "LSP-PE-1-PE-2-
default"
*A:PE-2# configure service sdp 212 class-forwarding multicast-lsp "LSP-PE-2-PE-1-
default"
```

The SAP ingress policy to set the default FC to EF is enabled on SAP 1/2/2 in VPLS 2 in PE-1 and PE-2.

```
*A:PE-1# configure service vpls 2 sap 1/2/2 ingress qos 2
```

Ping messages are classified as FC EF in VPLS 2. This traffic is sent over "LSP-PE-1-PE-2-EF". However, if the traffic is unknown, it will be sent over the multicast LSP, which is "LSP-PE-1-PE-2-default". To get unknown traffic, the forwarding database (FDB) is cleared and MAC learning is disabled in VPLS 2 on PE-1.

```
*A:PE-1# clear service id 2 fdb all
*A:PE-1# configure service vpls 2 disable-learning
```

Clear port statistics and launch one thousand rapid ping messages in VPRN 1 on PE-1.

```
*A:PE-1# clear port 1/1/[1..2] statistics
*A:PE-1# ping router 1 172.31.2.1 rapid count 1000
```

Verify the port statistics. BUM traffic is sent on the multicast LSP "LSP-PE-1-PE-2-default" and the egress port on PE-1 is 1/1/1:

```
*A:PE-1# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1010          124684            1011          124850
===============================================================================
*A:PE-1# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                        14            1200              15            1310
===============================================================================
```

Re-enable MAC learning in VPLS 2.

```
*A:PE-1# configure service vpls 2 no disable-learning
```

Verify CBF in SDP

The CBF related information for SDP 122 in VPLS 2 on PE-1can be verified as follows:

```
*A:PE-1# show service id 2 sdp 122 detail
---snip---
-------------------------------------------------------------------------------
RSVP/Static LSPs
-------------------------------------------------------------------------------
Associated LSP List :
Lsp Name           : LSP-PE-1-PE-2-EF
Admin State        : Up                       Oper State        : Up
Time Since Last Tr*: 00h26m41s

Lsp Name           : LSP-PE-1-PE-2-default
Admin State        : Up                       Oper State        : Up
Time Since Last Tr*: 00h21m35s

-------------------------------------------------------------------------------
Class-based forwarding :
-------------------------------------------------------------------------------
Class forwarding   : Enabled                  EnforceDSTELspFc  : Disabled
Default LSP        : LSP-PE-1-PE-2-default    Multicast LSP     : LSP-PE-1-PE-*
===============================================================================
FC Mapping Table
===============================================================================
FC Name             LSP Name
-------------------------------------------------------------------------------
ef                  LSP-PE-1-PE-2-EF
===============================================================================
---snip---
```

The detailed information for the SDP contains all the LSPs configured in the SDP. CBF is enabled with default LSP "LSP-PE-1-PE-2-default". The multicast LSP is "LSP-PE-PE-2-default", but the name is truncated. The FC mapping table lists all FCs that have an LSP assigned. Each FC can only have one LSP assigned. Different FCs can have the same LSP assigned. In this example, only the FC EF has an LSP assigned.

Traffic statistics can be displayed per service per SDP. These statistics are not per LSP, so no distinction can be made between the different FCs.

```
*A:PE-1# show service id 2 sdp 122 detail
---snip---
Statistics            :
```

```
I. Fwd. Pkts.      : 9003                    I. Dro. Pkts.     : 0
I. Fwd. Octs.      : 882180                  I. Dro. Octs.     : 0
E. Fwd. Pkts.      : 9001                    E. Fwd. Octets    : 882060
---snip---
```

Note: In this configuration example, the service using the SDP with CBF is a VPLS. The SDP can be used in a similar way by a VPRN.

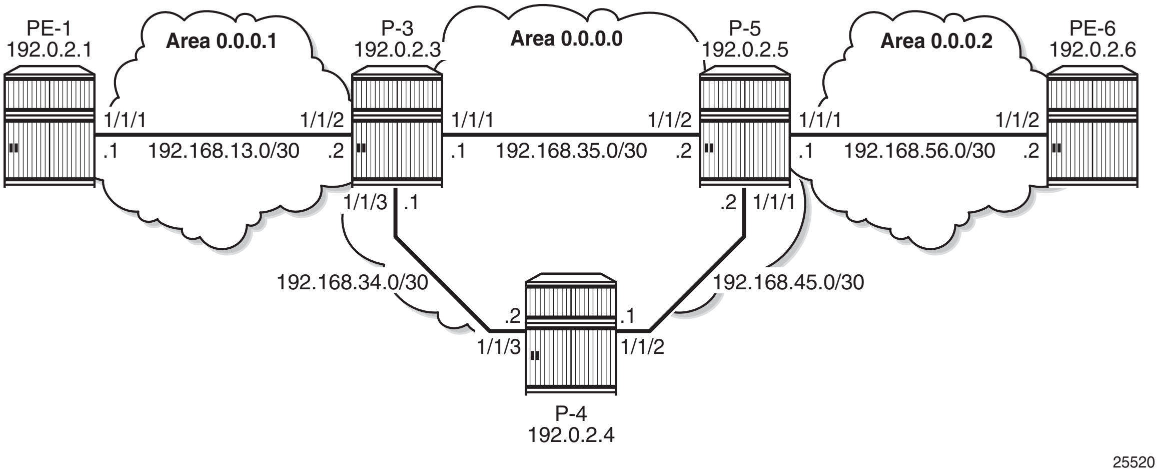
Note: An epipe service using CBF must have ingress shared-queuing enabled on access/SAP to work correctly.

CBF of LDP Prefix Packets over IGP Shortcuts

For this implementation, a more complex test topology is needed. Different OSPF areas are used.

In this example, traffic will be sent from VPRN 3 in PE-1 in OSPF area 0.0.0.1 to VPRN 3 in PE-6 in OSPF area 0.0.0.2. Most nodes used in the previous use case are reused, but some reconfiguration is required (router interfaces, OSPF areas, LSPs). For simplicity, PE-2 is not used anymore. The test topology is shown in Figure 197.

Figure 197 **Test Topology for CBF of LDP Prefix Packets over IGP Shortcuts**



Initial Configuration

All nodes have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces
- IGP (here: OSPF) - PE-1 is now in OSPF area 0.0.0.1 instead of 0.0.0.0 in the preceding use case. P-3 and P-5 have interfaces in different areas:

```
*A:P-3# configure router
        ospf
            traffic-engineering
            area 0.0.0.0
                interface "system"
                exit
                interface "int-P-3-P-4"
                    interface-type point-to-point
                exit
                interface "int-P-3-P-5"
                    interface-type point-to-point
                exit
            exit
            area 0.0.0.1
                interface "int-P-3-PE-1"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
```

- MPLS enabled on all router interfaces
- RSVP enabled
- RSVP LSPs:
  - Between PE-1 and P-3 in area 0.0.0.1 and between PE-6 and P-5 in area 0.0.0.2.
  - Between P-3 and P-5 in area 0.0.0.0, there are two LSPs in each direction: one LSP uses the direct strict path and the other LSP uses an indirect strict path via P-4. On P-3, the LSP "LSP-P-3-P-4-P-5" with indirect path will be used as the default LSP, while the LSP "LSP-P-3-P-5" with the direct path will be used for traffic with FC EF. For simplicity, there are no LSPs configured to or from P-4.

```
*A:P-3# configure router
        mpls
            path "path-P-3-PE-1"
                hop 10 192.168.13.1 strict
                no shutdown
            exit
            path "path-P-3-P-5"
                hop 10 192.168.35.2 strict
                no shutdown
```

```
    exit
    path "path-P-3-P-4-P-5"
        hop 10 192.168.34.2 strict
        hop 20 192.168.45.2 strict
        no shutdown
    exit
    lsp "LSP-P-3-PE-1"
        to 192.0.2.1
        primary "path-P-3-PE-1"
        exit
        no shutdown
    exit
    lsp "LSP-P-3-P-5"
        to 192.0.2.5
        primary "path-P-3-P-5"
        exit
        no shutdown
    exit
    lsp "LSP-P-3-P-4-P-5"
        to 192.0.2.5
        primary "path-P-3-P-4-P-5"
        exit
        no shutdown
    exit
    no shutdown
exit
```

Configure IGP Shortcuts

IGP shortcut or forwarding adjacency must be enabled in one or more IGP instances. In the example, IGP shortcuts (RSVP shortcuts) are configured on all nodes:

- IGP shortcut:

```
*A:PE-1# configure router ospf rsvp-shortcut
```

- Forwarding adjacency (not configured in the example):

```
*A:PE-1# configure router ospf advertise-tunnel-link
```

By default, all LSPs are eligible for IGP shortcut. However, it is possible to exclude a specific RSVP LSP from being used in IGP shortcut as follows (this is not required in this example):

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-P-3" no igp-shortcut
```

For more information on IGP Shortcuts, see chapter “IGP Shortcuts”.

Configure ECMP

ECMP must be enabled in the global routing instance. Nokia recommends that ECMP is set to a value that is at least equal to the number of FCs used by the packets forwarded using the LDP over RSVP tunnel.

In this case, two traffic flows are distinguished: traffic with FC EF takes one LSP while all other traffic takes the default LSP. Here, ECMP should be at least equal to 2. For simplicity, ECMP equal to 2 is only configured on P-3 and P-5:

```
*A:P-3# configure router ecmp 2
```

➡ **Note:** The selection of ECMP LSPs is done regardless of class-forwarding assignments, if any. The total number of LSPs with same cost, the number of those which have class-forwarding assignments, and the max-ecmp-routes parameter value influence the final set of LSPs, the consistency (from a CBF perspective) of which will be verified. These three factors must be carefully configured so as to ensure that the final set is consistent from a CBF perspective.

Enable LDP over RSVP

LDP over RSVP must be enabled between all PE routers and all P routers within the same area and between all P routers in the area 0:

```
*A:PE-1# configure router
        ldp
            targeted-session
                peer 192.0.2.3
                    tunneling
                    exit
                exit
            exit

*A:P-3# configure router
        ldp
            targeted-session
                peer 192.0.2.1
                    tunneling
                    exit
                exit
                peer 192.0.2.5
                    tunneling
                    exit
                exit
            exit
```

Note: The LSP names configured in the tunneling context (if any) are not directly used by LDP when the rsvp-shortcut option is enabled. With IGP shortcuts, the set of tunnel next-hops is always provided by IGP in the routing table manager (RTM). The class-based forwarding rules will not apply to these named LSPs unless they are populated by IGP in the RTM as next-hops for a prefix.

Note: The option prefer-tunnel-in-tunnel must be disabled (which is the default) for CBF to apply to LDP prefixes which are the endpoints of tunnels.

There is no need to enable LDP over RSVP on the LSPs. It is enabled by default, as can be verified with the following command:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2-EF" detail | match LdpOverRsvp
LdpOverRsvp : Enabled                         VprnAutoBind    : Enabled
```

For more information on LDP over RSVP, see chapter “LDP over RSVP Using OSPF as IGP”.

Configure CBF of LDP Prefix Packets over IGP Shortcuts

Enable CBF of LDP prefix packets over IGP shortcuts with the following command in the LDP context on P-3 and P-5:

```
*A:P-3# configure router ldp class-forwarding
```

In this example, the PEs only have one path to the P routers. There is no need to enable class-forwarding.

When CBF is enabled, LDP prefixes resolved to a set of ECMP tunnel next-hops will have their packets forwarded to the LSP configured to carry the forwarding class that the packet was classified to at the ingress SAP, access interface, or network interface.

In release 13.0.R6, this command is applicable for LSRs forwarding LDP forwarding equivalence class (FEC) prefix packets over a set of MPLS LSPs using IGP shortcuts. It is not supported for LER forwarding.

Note: The command configure router ldp class-forwarding applies to the following contexts:

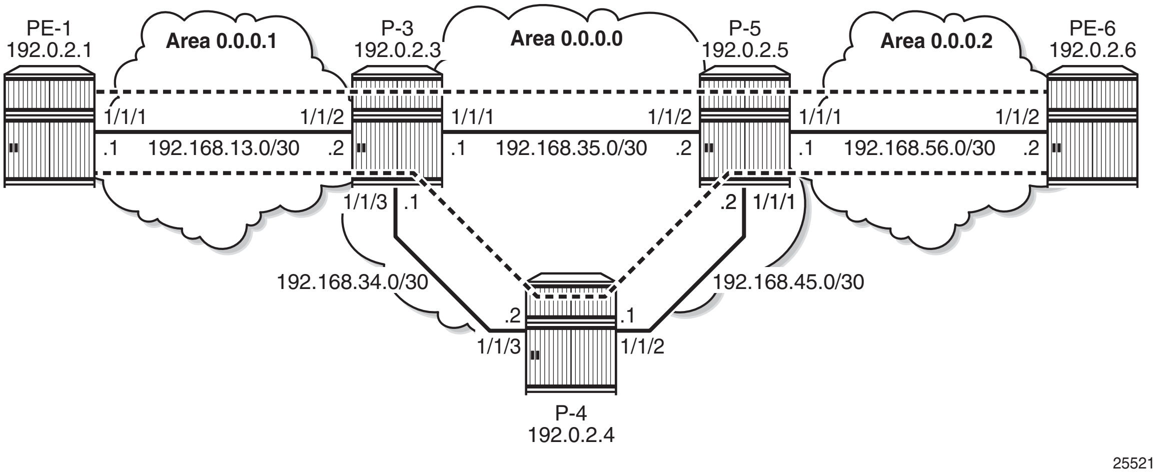
1. LER forwarding of packets of VPRN and L2 services that use auto-binding to LDP when the LDP FEC is resolved to a set of MPLS LSPs using IGP shortcuts. This is not supported in 13.0.R6.
2. LER forwarding of shortcut packets over LDP FEC that is resolved to a set of MPLS LSPs using IGP shortcuts. This is not supported in 13.0.R6.
3. LSR forwarding LDP FEC prefix packets over a set of MPLS LSPs using IGP shortcuts. This is supported in 13.0.R6.

This command does not apply to the following contexts:

- LER forwarding of VPRN and L2 services over a user-provisioned SDP of type LDP when the LDP FEC is resolved to a set of MPLS LSPs using IGP shortcuts.

In this use case, the direct link from P-3 to P-5 should only be used for traffic with FC EF. The default LSP will take the longer path from P-3 via P-4 to P-5. The two different paths are shown in Figure 198.

Figure 198 **Different Paths for Different FCs**



The LSPs can be configured as follows:

```
*A:P-3# configure router
        mpls
           lsp "LSP-P-3-P-5"
                class-forwarding
                    fc ef
                exit
            exit
            lsp "LSP-P-3-P-4-P-5"
                class-forwarding
```

```
                    default-lsp
                exit
            exit
```

All traffic with an FC that is not mapped to a specific LSP will be mapped to the default LSP.   The default LSP will also be used for traffic with FC EF, in case the dedicated LSP for that traffic is no longer available. It is possible to configure more than one default LSP, but only one will be used at a time.

The same LSP can be assigned to one or more FCs and be the default LSP at the same time.

```
*A:P-3# configure router
        mpls
            lsp "LSP-P-3-P-4-P-5"
                class-forwarding
                    fc be
                    default-lsp
                exit
            exit
```

The consistency of the configuration among the tunnel next-hops of an LDP FEC can only be verified by LDP at the time the FEC is resolved to IGP shortcuts. An example of an inconsistent configuration would be when class-forwarding is enabled while all tunnel next-hops for an LDP FEC have neither **fc** nor **default-lsp** assigned to them. LDP will then revert to ECMP routing for that FEC.

In this example, only P-3 and P-5 have ECMP equal to 2 and class-forwarding enabled.

Multiple LSPs can have the same FC assigned. Only one of these LSPs will be used to forward packets of this FC. That LSP is the one with the lowest tunnel ID.

Multiple LSPs can have the default-lsp configuration assigned, but only one of those will be the default LSP carrying all the traffic that should get the default treatment. That LSP is the one with the lowest tunnel ID.

If at least one LSP (amongst the ECMP set of LSPs) has an FC assigned, but no LSP has the default-lsp configuration, a single LSP will be automatically designated by LDP, even if all eight FCs have been mapped. A default LSP is needed to carry packets of an FC for which no explicit mapping to an LSP exists. The LSP with the lowest tunnel ID will be selected.

Note: If none of the LSPs has an FC or default LSP configuration, the set is inconsistent and no CBF happens.

When the active LDP bindings are displayed, FECs resolved with CBF can be recognized by the (C) after the prefix:

```
*A:P-3# show router ldp bindings active prefixes prefix 192.0.2.6/32
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262136
192.0.2.5                                   LspId 2

192.0.2.6/32                                Push              --      262136
192.0.2.5                                   LspId 3

192.0.2.6/32(C)                             Swap            262133    262136
192.0.2.5                                   LspId 2

192.0.2.6/32(C)                             Swap            262133    262136
192.0.2.5                                   LspId 3

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:P-3#
```

Only the entries that correspond to swap operations get the (C) added. The entries that correspond to push operations do not get the (C) because CBF is not supported for LERs in 13.0.R6.

Traffic that is sent from PE-1 to PE-6 will be forwarded on P-3 based on the FC, while traffic that originates on P-3 will not be subjected to CBF.

LSP 2 uses the direct path from P-3 to P-5 while LSP 3 uses the indirect path from P-3 to P-5 via P-4. This can be verified in the tunnel table:

```
*A:P-3# show router tunnel-table
===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  1         7        192.168.13.1   16777215
192.0.2.1/32      ldp       MPLS  65537     9        192.0.2.1      65535
```

```
192.0.2.5/32       rsvp       MPLS  2          7        192.168.35.2   16777215
192.0.2.5/32       rsvp       MPLS  3          7        192.168.34.2   16777215
192.0.2.5/32       ldp        MPLS  65538      9        192.0.2.5      65535
192.0.2.6/32       ldp        MPLS  65539      9        192.0.2.5      131070
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:P-3#
```

The LSP that uses the direct path to P-5 has the lowest tunnel ID. When both LSPs are configured as default LSP, the LSP with the lowest tunnel ID will effectively be used as the default LSP. In the current configuration, only the LSP using the indirect path is configured as the default LSP.

Verify CBF

Traffic will be sent from a VPRN on PE-1 to a VPRN on PE-6 and back.

Configure VPRN

On PE-1 and PE-6, a VPRN service needs to be configured. BGP will be used to exchange VPN routes.

Import and export policies are configured as follows:

```
*A:PE-1# configure router
        policy-options
            begin
            community "VPN3" members "target:64496:3"
            policy-statement "VPN3-export"
                entry 10
                    from
                        protocol direct
                    exit
                    to
                        protocol bgp-vpn
                    exit
                    action accept
                        community add "VPN3"
                    exit
                exit
            exit
            policy-statement "VPN3-import"
                entry 10
                    from
                        protocol bgp-vpn
                        community "VPN3"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

The VPRN is configured using auto-bind-tunnel with resolution-filter ldp, as follows:

```
*A:PE-1# configure service
        vprn 3 customer 1 create
            vrf-import "VPN3-import"
            vrf-export "VPN3-export"
            route-distinguisher 64496:31
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            interface "loopback3" create
                address 172.31.1.3/32
                loopback
            exit
            no shutdown
        exit
```

For BGP, P-5 is used as route reflector. The address family is VPN-IPv4.

```
*A:PE-1# configure router
        autonomous-system 64496
        bgp
            family vpn-ipv4
            group "internal"
                peer-as 64496
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
        exit

*A:P-5# configure router
        bgp
        autonomous-system 64496
            family vpn-ipv4
            cluster 1.1.1.1
            group "internal"
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
        exit
```

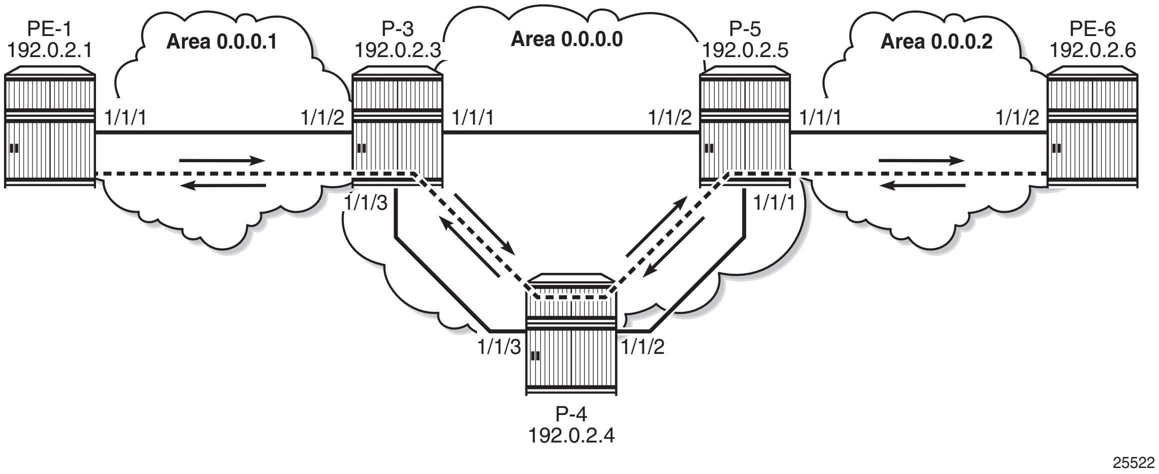
The routes are learned, as can be verified with the following command:

```
*A:PE-1# show router 3 route-table
===============================================================================
Route Table (Service: 3)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.31.1.3/32                                 Local   Local     00h21m08s  0
       loopback3                                                    0
172.31.6.3/32                                 Remote  BGP VPN   00h10m47s  170
       192.0.2.6 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

Verify Traffic Flow - Default LSP

When ping messages are sent from the VPRN on PE-1 to the VPRN on PE-6, the traffic will be forwarded according to the FC at the LSRs. The FC is not manipulated yet. For P-3, the traffic will be sent on the default LSP to P-5 via P-4. On node P-3, the ingress port for traffic destined to PE-6 is 1/1/2 and the egress port to P-4 is 1/1/3. The ping replies will follow the reverse path. The traffic flow is shown in Figure 199.

Figure 199 **Traffic on Default LSP**



```
*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 1000
*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress          Ingress          Egress          Egress
Id                      Packets           Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                        25             2548              24            2322
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress          Ingress          Egress          Egress
Id                      Packets           Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      1029           116738            1030          116844
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress          Ingress          Egress          Egress
Id                      Packets           Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                      1037           118074            1040          118648
===============================================================================
*A:P-3#
```

With the same commands, the traffic flow on the other nodes can be verified. P-4 has traffic on port 1/1/3 (to P-3) and on port 1/1/2 (to P-5). P-5 has traffic on port 1/1/1 (to P-4) and port 1/1/3 (to PE-6).

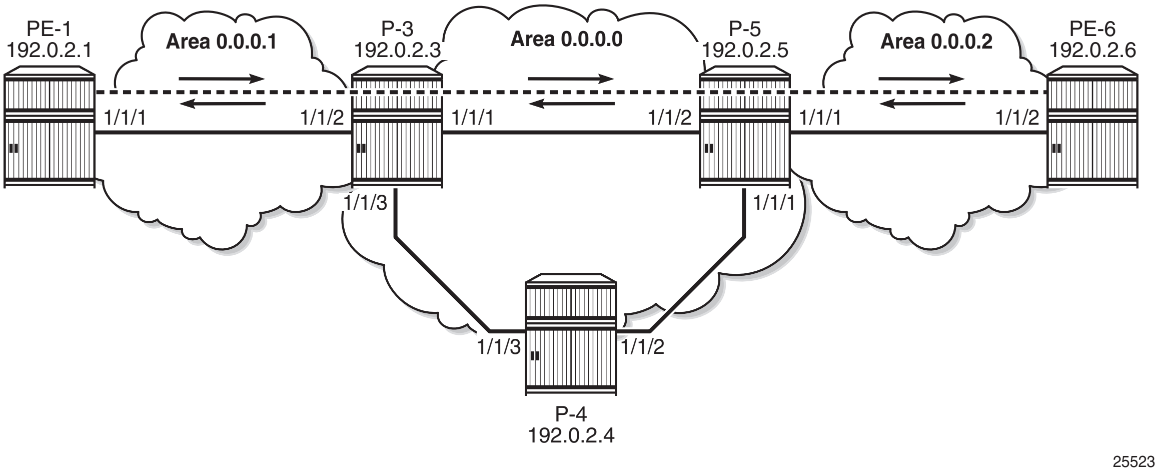
Verify Traffic Flow - LSP for FC EF

On nodes P-3 and P-5, the FC is set to EF at the ingress of the interface to the PE router.

```
*A:P-3# configure qos
        network 2 create
            ingress
                default-action fc ef profile in
            exit
        exit
*A:P-3# configure router interface "int-P-3-PE-1" qos 2
```

When the FC is modified to EF, the traffic takes the direct path between P-3 and P-5. On node P-3, traffic will be sent to and received from P-5 on port 1/1/1, as shown in Figure 200.

Figure 200 **Traffic with FC EF on Direct Path**



```
*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 1000
*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                        Ingress         Ingress          Egress          Egress
Id                          Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                          1015          115276            1015          115276
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                        Ingress         Ingress          Egress          Egress
Id                          Packets          Octets         Packets          Octets
```

```
-------------------------------------------------------------------------------
1/1/2                      1024          116136            1024          116004
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress        Egress           Egress
Id                      Packets          Octets        Packets          Octets
-------------------------------------------------------------------------------
1/1/3                        22            1834              23            2062
===============================================================================
*A:P-3#
```

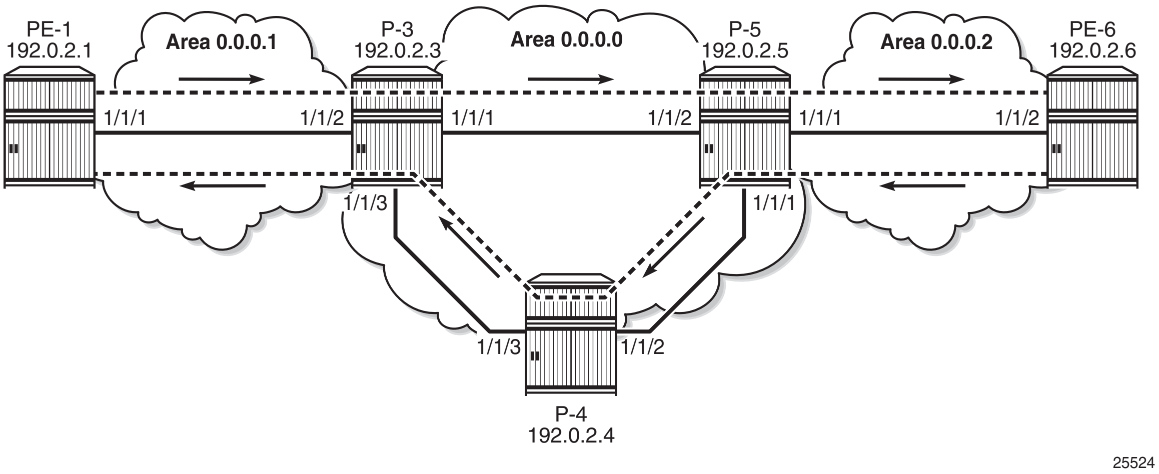
The original configuration is restored by removing the QoS network policy from the interface to the PE as follows:

```
*A:P-3# configure router interface "int-P-3-PE-1" no qos
```

Verify Traffic Flow - Multiple LSPs Configured as Default LSP on P-3

The FC is not manipulated anymore. On P-3, LSP "LSP-P-3-P-4-P-5" was configured as default LSP. Additionally, also LSP "LSP-P-3-P-5" is configured as a default LSP. So "LSP-P-3-P-5" becomes the used default LSP, because the tunnel ID (equal to 2) is lower than the tunnel ID of the LSP using the long path from P-3 to P-5 via P-4 (equal to 3). When this is configured on P-3, without changing anything to the configuration at P-5, the behavior is asymmetric. Incoming traffic on port 1/1/2 is forwarded to P-5 via port 1/1/1, while the ping replies still take the long path using ingress port 1/1/3 on P-3 The traffic flow is shown in Figure 201.

Figure 201 Traffic on Default LSP with Lowest ID on P-3



```
*A:P-3# configure router
        mpls
            lsp "LSP-P-3-P-5"
                class-forwarding
                    default-lsp
                exit
*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 1000
*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                        22            1904            1028          116819
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      1029          116813            1025          116289
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress          Egress          Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                      1031          117375              32            3520
===============================================================================
*A:P-3#
*A:P-3# show router tunnel-table
```

```
===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  1         7        192.168.13.1   16777215
192.0.2.1/32      ldp       MPLS  65550     9        192.0.2.1      65535
192.0.2.5/32      rsvp      MPLS  2         7        192.168.35.2   16777215
192.0.2.5/32      rsvp      MPLS  3         7        192.168.34.2   16777215
192.0.2.5/32      ldp       MPLS  65553     9        192.0.2.5      65535
192.0.2.6/32      ldp       MPLS  65554     9        192.0.2.5      131070
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:P-3#
```

The original configuration is restored. The LSP with the direct path will only be used for traffic with FC EF ("LSP-P-3-P-5" remained an EF LSP even when it became a default LSP):

```
*A:P-3# configure router mpls lsp "LSP-P-3-P-5" class-forwarding no default-lsp
```

Verify Traffic Flow - No Default LSP Configured

When the class-forwarding configuration for the different LSPs always has one or more FCs and no default LSP, the system will select a default LSP itself. This can be tested by removing the default LSP configuration from all LSPs on P-3 and P-5 and adding FC AF instead. FC AF does not correspond to the FC of the ping messages; therefore, a default LSP will be required.

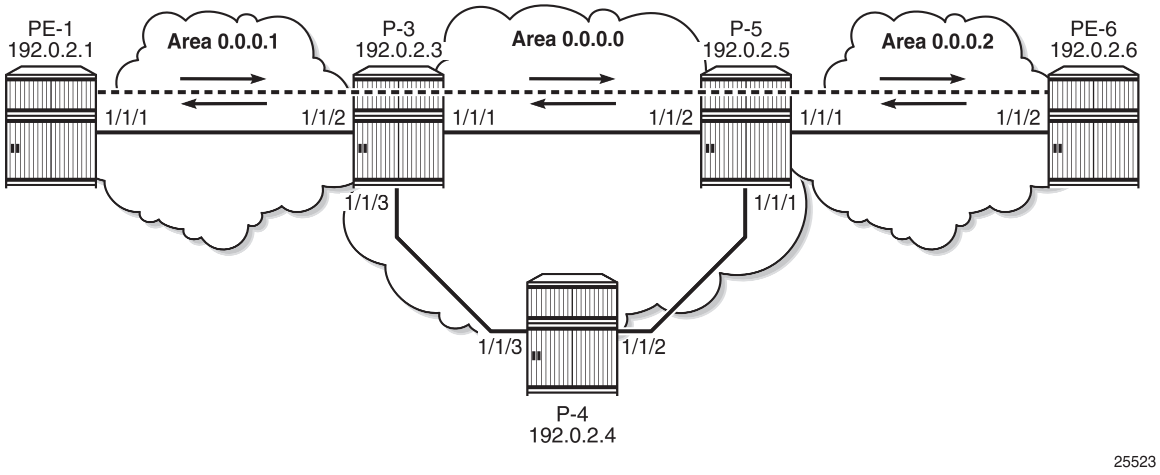
```
*A:P-3# configure router mpls lsp "LSP-P-3-P-4-P-5" class-forwarding no default-lsp
*A:P-3# configure router mpls lsp "LSP-P-3-P-4-P-5" class-forwarding fc af
*A:P-5# configure router mpls lsp "LSP-P-5-P-4-P-3" class-forwarding no default-lsp
*A:P-5# configure router mpls lsp "LSP-P-5-P-4-P-3" class-forwarding fc af
```

The actual CBF configuration does not include any default LSPs anymore:

```
*A:P-3>config>router>mpls# info
----------------------------------------------
---snip---
            lsp "LSP-P-3-P-5"
                to 192.0.2.5
                class-forwarding
                    fc ef
                exit
                primary "path-P-3-P-5"
                exit
                no shutdown
            exit
            lsp "LSP-P-3-P-4-P-5"
                to 192.0.2.5
                class-forwarding
                    fc af
                exit
                primary "path-P-3-P-4-P-5"
                exit
                no shutdown
            exit
```

The system selects the LSP with the lowest tunnel ID as the default LSP, in this case the LSP using the direct path between P-3 and P-5, as shown in Figure 202.

Figure 202 **Traffic on System-Selected Default LSPs with Lowest Tunnel ID**



The tunnel ID can be verified in the tunnel table. The tunnel table for P-3 has already been shown in the preceding section.

On P-5, the LSP to P-3 with the tunnel ID 2 has next hop 192.168.35.1, as can be verified in the tunnel table. This corresponds to LSP "LSP-P-5-P-3". The LSP with tunnel ID 3 and next hop 192.168.45.1 corresponds to LSP "LSP-P-5-P-4-P-3".

```
*A:P-5# show router tunnel-table
===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65552     9        192.0.2.3      131070
192.0.2.3/32      rsvp      MPLS  2         7        192.168.35.1   16777215
192.0.2.3/32      rsvp      MPLS  3         7        192.168.45.1   16777215
192.0.2.3/32      ldp       MPLS  65551     9        192.0.2.3      65535
192.0.2.6/32      rsvp      MPLS  1         7        192.168.56.2   16777215
192.0.2.6/32      ldp       MPLS  65553     9        192.0.2.6      65535
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
```

```
===============================================================================
*A:P-5#
```

On P-3, incoming traffic from PE-1 will be forwarded on port 1/1/1 toward P-5.

```
*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 1000

*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets        Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      1024          116120           1022          115840
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets        Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      1026          116221           1028          116325
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress          Egress
Id                      Packets          Octets        Packets          Octets
-------------------------------------------------------------------------------
1/1/3                        26            2142             25            2174
===============================================================================
*A:P-3#
```

No CBF in Case of Inconsistent Configuration

The FC LSP and default LSPs are removed from P-3 and P-5. Class-forwarding remains enabled, but the configuration is inconsistent and the forwarding behavior reverts to ECMP routing.

```
*A:P-3# configure router mpls lsp "LSP-P-3-P-4-P-5" class-forwarding no default-lsp
*A:P-3# configure router mpls lsp "LSP-P-3-P-4-P-5" class-forwarding no fc af
*A:P-3# configure router mpls lsp "LSP-P-3-P-5" class-forwarding no fc ef
*A:P-5# configure router mpls lsp "LSP-P-5-P-4-P-3" class-forwarding no default-lsp
*A:P-5# configure router mpls lsp "LSP-P-5-P-4-P-3" class-forwarding no fc af
*A:P-5# configure router mpls lsp "LSP-P-5-P-3" class-forwarding no fc ef
*A:P-3>config>router>mpls# info
----------------------------------------------
---snip---
            lsp "LSP-P-3-P-5"
                to 192.0.2.5
                class-forwarding
                exit
                primary "path-P-3-P-5"
                exit
                no shutdown
            exit
            lsp "LSP-P-3-P-4-P-5"
                to 192.0.2.5
                class-forwarding
                exit
                primary "path-P-3-P-4-P-5"
                exit
                no shutdown
            exit
            no shutdown
```

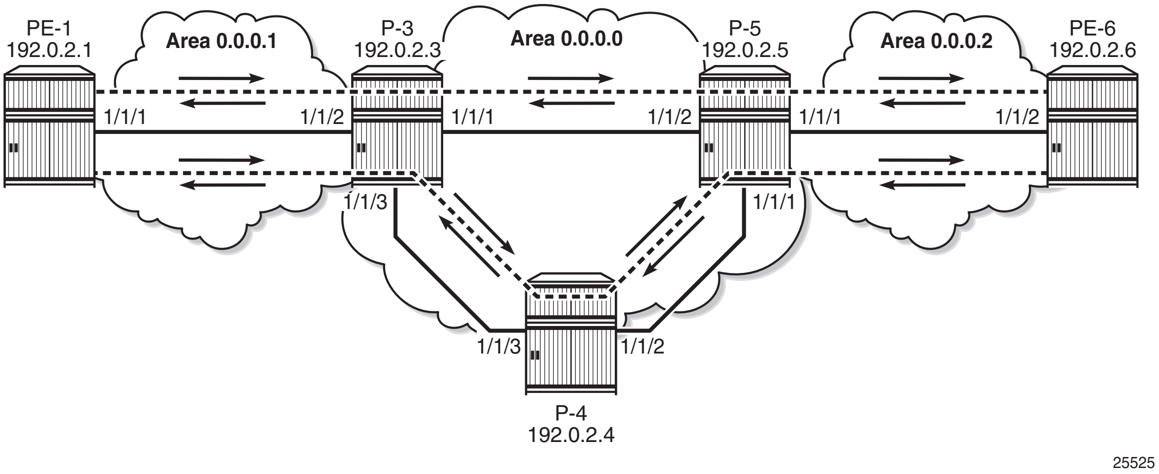
ECMP equals 2, so the traffic flows will be distributed over the two LSPs between P-3 and P-5, as can be verified in the routing tables at P-3 and P-5:

```
*A:P-3# show router route-table
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      03h05m18s  10
       192.0.2.1 (tunneled:RSVP:1)                                  65535
192.0.2.3/32                                  Local   Local     04h56m20s  0
       system                                                       0
192.0.2.4/32                                  Remote  OSPF      04h40m38s  10
       192.168.34.2                                                 10
192.0.2.5/32                                  Remote  OSPF      00h00m06s  10
       192.0.2.5 (tunneled:RSVP:2)                                  65535
192.0.2.5/32                                  Remote  OSPF      00h00m06s  10
       192.0.2.5 (tunneled:RSVP:3)                                  65535
192.0.2.6/32                                  Remote  OSPF      00h00m06s  10
       192.0.2.5 (tunneled:RSVP:2)                                  131070
192.0.2.6/32                                  Remote  OSPF      00h00m06s  10
       192.0.2.5 (tunneled:RSVP:3)                                  131070
---snip---

*A:P-5# show router route-table 192.0.2.1/32
===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      00h01m06s  10
       192.0.2.3 (tunneled:RSVP:2)                                  131070
192.0.2.1/32                                  Remote  OSPF      00h01m06s  10
       192.0.2.3 (tunneled:RSVP:3)                                  131070
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:P-5#
```

The ECMP hashing algorithm will hash on the source and destination IP address. As long as they are the same for each packet, the traffic will be sent on the same link. Therefore, different traffic flows need to be generated in parallel from PE-1 to PE-6. It is possible to create additional loopback interfaces in VPRN 3, but for this test, the system IP addresses of PE-1 and PE-6 can also be used. The ping messages are not sent after one another, but parallel in two different sessions with PE-1.

Figure 203 shows that different traffic flows are sprayed over the different LSPs. The returning traffic need not take the same path.

Figure 203 Inconsistent CBF Configuration. Revert to ECMP Forwarding



```
*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 2000
*A:PE-1# ping 192.0.2.6 rapid count 2000
*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress          Ingress           Egress          Egress
Id                      Packets           Octets          Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      2052           216350               54            4745
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress          Ingress           Egress          Egress
Id                      Packets           Octets          Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      4054           445019             4055          444877
```

```
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                      2048          232178            4042          443622
===============================================================================
*A:P-3#


*A:P-3# clear port 1/1/[1..3] statistics
*A:PE-1# ping router 3 172.31.6.3 rapid count 2000
*A:PE-1# ping 192.0.2.6 rapid count 2000
*A:P-3# show port 1/1/1 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                      2052          216350              54            4745
===============================================================================
*A:P-3# show port 1/1/2 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      4054          445019            4055          444877
===============================================================================
*A:P-3# show port 1/1/3 statistics
===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/3                      2048          232178            4042          443622
===============================================================================
*A:P-3#
```

In this case, all traffic from PE-1 to PE-6 for both traffic flows takes the LSP "LSP-P-3-P-4-P-5". The returning traffic is distributed over the LSPs "LSP-P-5-P-3" and "LSP-P-5-P-4-P-3". ECMP works, but there are not enough traffic flows to show it more clearly.

Conclusion

Within a service, traffic with different FCs can be forwarded using different paths, depending on the requirements for latency and jitter. This is also the case for services traversing multiple domains.

CBF is a local decision and need not be enabled on all intermediate routers. This allows for better interoperability.

DiffServ Traffic Engineering

This chapter provides information about DiffServ traffic engineering.

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS Release 14.0.R2. The CLI in the current edition corresponds to SR OS Release 15.0.R1.

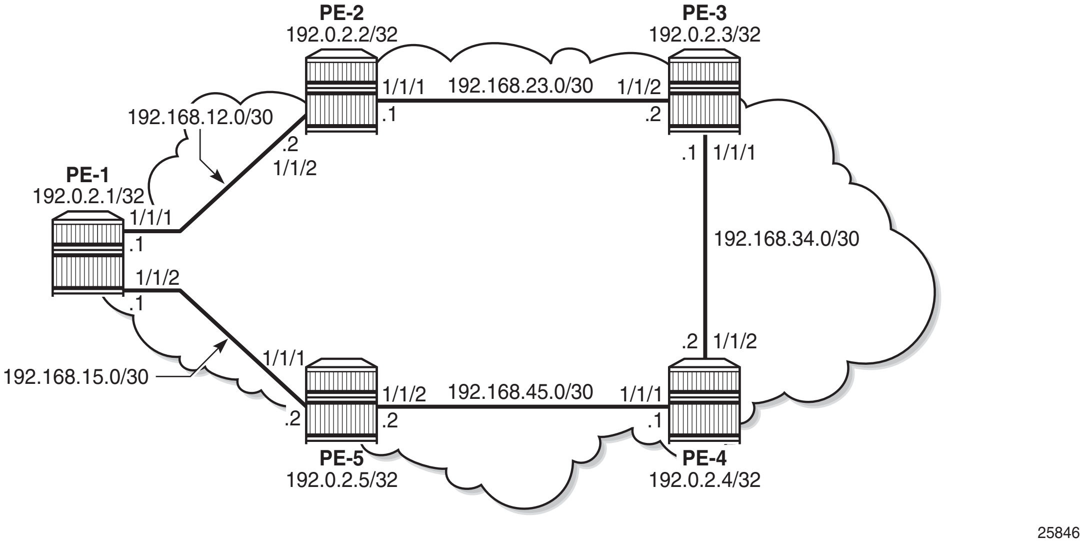
Overview

Differentiated Services (DiffServ) is a mechanism to classify and manage network traffic to provide Quality of Service (QoS). DiffServ Traffic Engineering (DiffServ TE) reserves bandwidth for Label Switched Paths (LSPs) on ReSource reserVation Protocol (RSVP) interfaces on a per TE class basis.

Example Topology

Figure 204 shows the example topology that contains five 7750 SRs in a ring topology.

Figure 204 **Example Topology**



Definitions

The following definitions are used in this chapter:

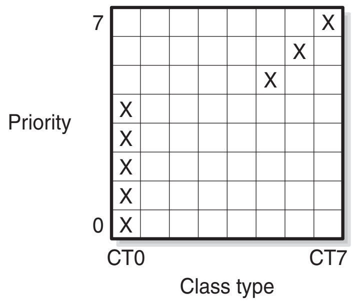
- Forwarding Classes (FCs) classify micro-flows into macro-flows. FCs can be mapped to Class Types (CTs).
- A CT is macro-flow crossing a link governed by a specific Bandwidth Constraint (BC). The BC is defined on a per-link and per-CT basis. A CT can be considered as a network-wide FC, advertised by the IGP (OSPF opaque link state advertisement (LSA), IS-IS TE Type Length Value (TLV)).
  - IGP-TE can reserve bandwidth per CT on a link (BC).
  - RSVP-TE can reserve bandwidth per LSP path, based on TE class.
- A TE class is a combination of a CT and a preemption priority.

There are eight FCs that can be mapped to CTs. The CTs range from CT0 (lowest) to CT7 (highest) and each gets a percentage of the bandwidth of the link. Each CT has eight different priority levels that are used for preemption. Even though there are 64 different potential combinations of CT and priority, only eight different combinations can be defined for TE classes. All CTs and priorities must be manually configured.

The system allows up to eight TE classes to be configured. The more TE classes are defined, the more RSVP LSPs need to be configured for each service. TE classes are consistently configured on all TE-aware Label Switching Routers (LSRs) throughout the network and advertised through the IGP.

The following shows a DiffServ TE configuration where each CT can reserve up to 10% of the maximum reservable bandwidth meaning that 20% of the bandwidth is not allocated to any CT. MPLS needs to be shut down when DiffServ TE is configured. Figure 205 shows the mapping of the TE classes.

Figure 205 **Mapping of TE Classes**



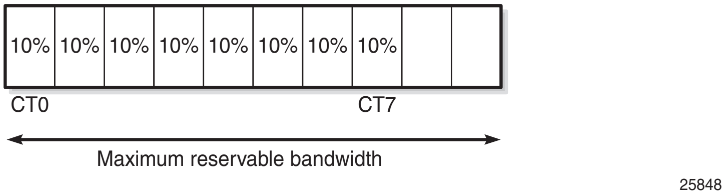
25847

```
*A:PE-1# configure
    router
        mpls
            shutdown
        exit
        rsvp
            preemption-timer 5
            diffserv-te mam
               class-type-bw ct0 10 ct1 10 ct2 10 ct3 10 ct4 10 ct5 10 ct6 10 ct7 10
                te-class 0 class-type 0 priority 0
                te-class 1 class-type 0 priority 1
                te-class 2 class-type 0 priority 2
                te-class 3 class-type 0 priority 3
                te-class 4 class-type 0 priority 4
                te-class 5 class-type 5 priority 5
                te-class 6 class-type 6 priority 6
                te-class 7 class-type 7 priority 7
                fc af class-type 0
                fc be class-type 0
                fc ef class-type 5
                fc h1 class-type 6
                fc h2 class-type 0
                fc l1 class-type 0
                fc l2 class-type 0
                fc nc class-type 7
            exit
        exit
        mpls
            no shutdown
        exit
```

This configuration is applied on all TE-aware LSRs for consistency.

Figure 206 shows the bandwidth reservation for the CTs.

Figure 206 **Bandwidth Reservation for the CTs**



Eight TE classes are defined, each with a different priority, using four CTs (0, 5, 6 and 7). There is no need to assign any bandwidth to the unused CTs. The configuration in this example is not a recommendation. In this example, the BC for each CT is 10% of the maximum reservable bandwidth.

On each node, the RSVP status shows whether DiffServ TE is enabled and how it is configured, as follows:

```
*A:PE-1# show router rsvp status

===============================================================================
RSVP Status
===============================================================================
Admin Status       : Up                  Oper Status        : Up
Keep Multiplier    : 3                   Refresh Time       : 30 sec
Message Pacing     : Disabled            Pacing Period      : 100 msec
Max Packet Burst   : 650 msgs            Refresh Bypass     : Disabled
Rapid Retransmit   : 5 hmsec             Rapid Retry Limit  : 3
Graceful Shutdown  : Disabled            SoftPreemptionTimer: 5 sec
GR Max Recovery    : 300 sec             GR Max Restart     : 120 sec
Implicit Null Label: Disabled            Node-id in RRO     : Exclude
P2P Merge Point Ab*: Disabled            P2MP Merge Point A*: Disabled
DiffServTE AdmModel: Mam                 Entropy Label      : Disabled
Percent Link Bw CT0: 10                  Percent Link Bw CT4: 10
Percent Link Bw CT1: 10                  Percent Link Bw CT5: 10
Percent Link Bw CT2: 10                  Percent Link Bw CT6: 10
Percent Link Bw CT3: 10                  Percent Link Bw CT7: 10
TE0  -> Class Type : 0                   Priority           : 0
TE1  -> Class Type : 0                   Priority           : 1
TE2  -> Class Type : 0                   Priority           : 2
TE3  -> Class Type : 0                   Priority           : 3
TE4  -> Class Type : 0                   Priority           : 4
TE5  -> Class Type : 5                   Priority           : 5
TE6  -> Class Type : 6                   Priority           : 6
TE7  -> Class Type : 7                   Priority           : 7
FCName             : af                  Class Type         : 0
FCName             : be                  Class Type         : 0
FCName            : ef                Class Type        : 5
FCName            : h1                Class Type        : 6
FCName            : h2                Class Type        : 0
FCName            : l1                Class Type        : 0
FCName            : l2                Class Type        : 0
FCName            : nc                Class Type        : 7
IgpThresholdUpdate : Disabled
Up Thresholds(%)   : 0 15 30 45 60 75 80 85 90 95 96 97 98 99 100
Down Thresholds(%) : 100 99 98 97 96 95 90 85 80 75 60 45 30 15 0
Update Timer       : N/A
Update on CAC Fail : Disabled
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

The OSPF LSAs will contain BC information, as shown in the following output taken from PE-1:

```
*A:PE-1# show router ospf opaque-database adv-router 192.0.2.3 detail

===============================================================================
Rtr Base OSPFv2 Instance 0 Opaque Link State Database (type: All) (detail)
===============================================================================
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0              Adv Router Id    : 192.0.2.3
Link State Id    : 1.0.0.1              LSA Type         : Area Opaque
Sequence No      : 0x80000006           Checksum         : 0x922c
Age              : 1993                 Length           : 28
Options          :  E
Advertisement    : Traffic Engineering
     ROUTER-ID TLV  (0001) Len   4 : 192.0.2.3
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0              Adv Router Id    : 192.0.2.3
Link State Id    : 1.0.0.3              LSA Type         : Area Opaque
Sequence No      : 0x80000001           Checksum         : 0xf162
Age              : 281                  Length           : 164
Options          :  E
Advertisement    : Traffic Engineering
     LINK INFO TLV  (0002) Len 140 :
       Sub-TLV: 1     Len: 1     LINK_TYPE    : 1
       Sub-TLV: 2     Len: 4     LINK_ID      : 192.0.2.2
       Sub-TLV: 3     Len: 4     LOC_IP_ADDR  : 192.168.23.2
       Sub-TLV: 4     Len: 4     REM_IP_ADDR  : 192.168.23.1
       Sub-TLV: 5     Len: 4     TE_METRIC    : 10
       Sub-TLV: 6     Len: 4     MAX_BDWTH    : 10000000 Kbps
       Sub-TLV: 7     Len: 4     RSRVBL_BDWTH : 10000000 Kbps
       Sub-TLV: 8     Len: 32    UNRSRVD_CLS0 :
         P0: 1000000 Kbps P1: 1000000 Kbps P2: 1000000 Kbps P3: 1000000 Kbps
         P4: 1000000 Kbps P5: 1000000 Kbps P6: 1000000 Kbps P7: 1000000 Kbps
       Sub-TLV: 9     Len: 4     ADMIN_GROUP  : 0 None
       Sub-TLV: 17    Len: 36    TELK_BW_CONST:
         BW Model : MAM
         BC0: 1000000 Kbps BC1: 1000000 Kbps BC2: 1000000 Kbps BC3: 1000000 Kbps
         BC4: 1000000 Kbps BC5: 1000000 Kbps BC6: 1000000 Kbps BC7: 1000000 Kbps
```

```
-------------------------------------------------------------------------------
---snip---
```

In the preceding output, only the output for the interface between PE-2 and PE-3 is shown; the output is similar for the other interfaces. On each interface between nodes, there will be eight BCs for the eight CTs: from BC0 to BC7. In this example, each of the BCs has the same constraint of 1 Gb/s, which corresponds to 10% of the 10 Gb/s interfaces. As long as no LSP is configured with a CT and a bandwidth, no bandwidth will be reserved. The BCs for an interface, such as the interface between PE-1 and PE-2, can be shown as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
Interface         : int-PE-1-PE-2
Port ID           : 1/1/1
Admin State       : Up                    Oper State        : Up
Active Sessions   : 0                     Active Resvs      : 0
Total Sessions    : 0
Subscription      : 100 %                 Port Speed        : 10000 Mbps
Total BW          : 10000 Mbps            Aggregate         : Dsabl
Hello Interval    : 3000 ms               Hello Timeouts    : 0
Key Type Auth     : Disabled
Keychain Auth     : Disabled
Auth Rx Seq Num   : n/a                   Auth Key Id       : n/a
Auth Tx Seq Num   : n/a                   Auth Win Size     : n/a
Refresh Reduc.    : Disabled              Reliable Deli.    : Disabled
Bfd Enabled       : No                    Graceful Shut.    : Disabled
ImplicitNullLabel : Disabled*             GR helper         : Disabled

Percent Link Bandwidth for Class Types*
Link Bw CT0       : 10                    Link Bw CT4       : 10
Link Bw CT1       : 10                    Link Bw CT5       : 10
Link Bw CT2       : 10                    Link Bw CT6       : 10
Link Bw CT3       : 10                    Link Bw CT7       : 10

Bandwidth Constraints for Class Types (Kbps)
BC0               : 1000000               BC4               : 1000000
BC1               : 1000000               BC5               : 1000000
BC2               : 1000000               BC6               : 1000000
BC3               : 1000000               BC7               : 1000000

Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE1->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE2->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE3->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE4->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE5->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE6->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
TE7->  Resv. Bw   : 0                     Unresv. Bw        : 1000000
```

```
IGP Update
Up Thresholds(%)    : 0 15 30 45 60 75 80 85 90 95 96 97 98 99 100  *
Down Thresholds(%) : 100 99 98 97 96 95 90 85 80 75 60 45 30 15 0  *
IGP Update Pending : No
Next Update        : N/A
No Neighbors.
* indicates inherited values
===============================================================================
*A:PE-1#
```

In this example, all BCs for the CTs are equal to 1 Gb/s, which is 10% of the maximum reservable bandwidth of 10 Gb/s. Currently no bandwidth is reserved for any of the TE classes. The unreserved bandwidth equals 1 Gb/s for TE0 through TE7.

The maximum bandwidth that can be allocated depends on the bandwidth of the link and the subscription percentage. When the subscription percentage is doubled to 200%, the BCs will be doubled too, as follows:

```
*A:PE-1# configure router rsvp interface "int-PE-1-PE-2" subscription 200
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
---snip---
Percent Link Bandwidth for Class Types*
Link Bw CT0       : 10                  Link Bw CT4       : 10
Link Bw CT1       : 10                  Link Bw CT5       : 10
Link Bw CT2       : 10                  Link Bw CT6       : 10
Link Bw CT3       : 10                  Link Bw CT7       : 10

Bandwidth Constraints for Class Types (Kbps)
BC0               : 2000000             BC4               : 2000000
BC1               : 2000000             BC5               : 2000000
BC2               : 2000000             BC6               : 2000000
BC3               : 2000000             BC7               : 2000000

Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE3->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE4->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE5->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 2000000
---snip---
```

The subscription percentage is restored to its default value of 100% as follows:

```
*A:PE-1# configure router rsvp interface "int-PE-1-PE-2" no subscription
```

Bandwidth Constraint Models

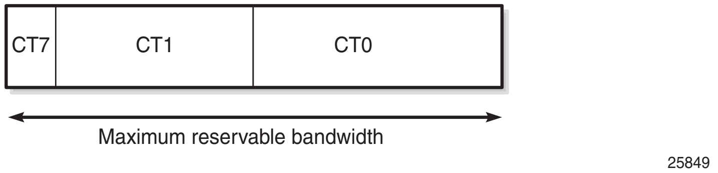
Two models are available for the bandwidth calculation that is required during the LSP setup: the Maximum Allocation Model (MAM) and the Russian Doll Model (RDM). Table 22 shows a comparison between the two models.

Table 22 **Comparison Bandwidth Constraint Models**

Maximum Allocation Model (MAM)	Russian Doll Model (RDM)
Fixed BC per CT. No bandwidth sharing between CTs.	Maps one BC to one or more CTs. Lower CTs are allowed to reserve from the unused bandwidth of the pools defined for higher CTs.
Achieves isolation between CTs and guaranteed bandwidth to CTs without the need for preemption.	No isolation between CTs. Requires preemption to guarantee bandwidth to CTs other than the premium.
Bandwidth may be wasted.	Efficient use of bandwidth.
Easy to manage.	More complex.

Figure 207 shows the reserved bandwidth for the different class types according to the MAM model. In this example, there is only bandwidth reserved for CT0, CT1, and CT2.

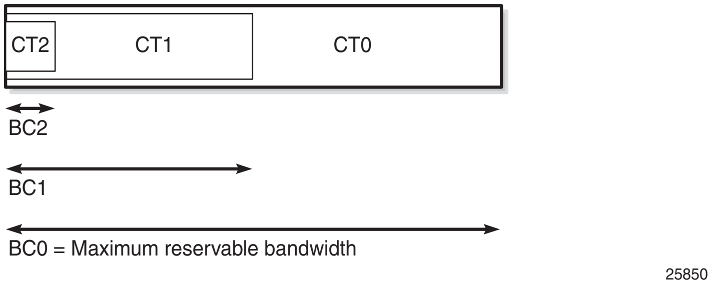
Figure 207 **Bandwidth Reservation in Maximum Allocation Model for Three CTs**



Bandwidth that is reserved for a specific CT cannot be used by any other CT. Therefore, bandwidth may be wasted.

The Russian Doll Model is more flexible: when CT1 has some spare bandwidth that might be used by CT0, this is allowed. Depending on the configured setup priority and hold priority, this may be reversed when CT1 requires the bandwidth. The bandwidth reservation in the Russian doll model is shown in Figure 208.

Figure 208 **Bandwidth Reservation in Russian Doll Model for Three CTs**



Backup Class Types

The main CT is defined at LSP level or primary path level. The main CT is used at the first attempt for the initial establishment and re-signal Make-Before-Break (MBB) of the LSP primary path. Re-signaling of the LSP path can be triggered manually or timer-based. Subsequent retries use the backup CT, which is configured on the primary path level. Secondary paths are always signaled with the main CT. There is no verification whether the backup CT is lower than the main CT. This applies to CSPF and non-CSPF LSPs. An example of an LSP with main CT1 and backup CT0 is as follows:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-withBackupCT"
                to 192.0.2.3
                cspf
                class-type 1
                primary "dyn"
                    bandwidth 50
                    priority 4 4
                    backup-class-type 0
                exit
                no shutdown
            exit
```

Possible triggers for using the backup CT are:

- Local interface failure or control plane failure (hello timeout)
- Received Resv message with the local-protection-in-use flag set (global revertive trigger)
- Received Patherr message with Fast ReRoute (FRR) protection active notification (global revertive trigger)
- Received Patherr message with error code 34 (Reroute) and value 1 (soft preemption trigger)
- Received Patherr message with Preemption pending flag set (soft preemption trigger)
- Received ResvTear message

When the reservable bandwidth for a CT (including the bandwidth for the inner dolls in case of RDM) is insufficient, this does not trigger the backup CT to be used. If possible, an alternate path will be used for the LSP requiring this bandwidth.

Priorities

Two different priorities are linked to an LSP in a range from 0 to 7, where 0 is the highest priority and 7 the lowest. These values are important when preemption occurs, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" primary "dyn" priority
  - no priority
  - priority <setup-priority> <hold-priority>

 <setup-priority>     : [0..7]
 <hold-priority>      : [0..7]
```

The following shows an LSP with both priorities equal to 4:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3"
                to 192.0.2.3
                cspf
                primary "dyn"
                    bandwidth 50
                    priority 4 4
                exit
                no shutdown
            exit
```

- The first priority in the configuration is the setup priority. When an LSP is signaled and there is not enough bandwidth available on the egress Label Edge Router (eLER) or LSR, the LSP can preempt an established LSP with a hold priority lower than this setup priority. For a setup priority of 4, existing LSPs with a hold priority of 5, 6, or 7 can be preempted in case of insufficient bandwidth.
- The second priority in the configuration is the hold priority. When this LSP is established and a new LSP needs to be established, and there is insufficient bandwidth, this LSP can only be preempted by an LSP with a higher setup priority than this hold priority. For a hold priority of 4, the LSP can be preempted by any LSP with a setup priority of 0, 1, 2, or 3.

The default values are a setup priority of 7 and a hold priority of 0. A low setup priority of 7 means the LSP cannot preempt any LSP. A high hold priority of 0 implies that the LSP cannot be preempted by any other LSP.

The setup priority needs to be lower than or equal to the hold priority to avoid preemption loops. Nokia recommends that the setup priority and the hold priority are set to equal values.

Bandwidth, CT information, and priorities are shown as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
Path LSP ID      : 61442
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262142
---snip---

Neg MTU          : 1564                 Oper MTU             : 1564
Bandwidth        : 50 Mbps              Oper Bandwidth       : 50 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Record Route     : Record               Oper Record Route    : Record
Record Label     : Record               Oper Record Label    : Record
Setup Priority   : 4                    Oper Setup Priority  : 4
Hold Priority    : 4                    Oper Hold Priority   : 4
Class Type       : 0                    Oper CT              : 0
```

```
Backup CT        : None
MainCT Retry     : n/a
    Rem          :
MainCT Retry     : 0
    Limit        :
---snip---
```

When the LSP is being established, the path message contains the setup and hold priorities, and the required bandwidth, as follows:

```
*A:PE-1# debug router rsvp packet path detail

1 2017/03/30 13:17:10.56 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.3
          TTL:255, Checksum:0x9f8d, Flags:0x0
Session    - EndPt:192.0.2.3, TunnId:4, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PE-1-PE-3::dyn
             SetupPri:4, HoldPri:4, Flags:0x46
RSVPHop    - Ctype:1, Addr:192.168.12.1, LIH:2
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:61442
SendTSpec  - Ctype:QOS, CDR:50.000 Mbps, PBS:50.000 Mbps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.12.1, Flags:0x0
ERO        - IPv4Prefix 192.168.12.2/32, Strict
             IPv4Prefix 192.168.23.2/32, Strict
"
```

As soon as the LSP is established, the bandwidth is reserved on the interface int-PE-1-PE-2 on PE-1 in TE class 4 (configured as a combination of CT0 and priority 4), as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
Interface         : int-PE-1-PE-2
Port ID           : 1/1/1
Admin State       : Up                    Oper State         : Up
Active Sessions   : 1                     Active Resvs       : 1
Total Sessions    : 1
Subscription      : 100 %                 Port Speed         : 10000 Mbps
Total BW          : 10000 Mbps            Aggregate          : Dsabl
---snip---

Percent Link Bandwidth for Class Types*
Link Bw CT0       : 10                    Link Bw CT4        : 10
Link Bw CT1       : 10                    Link Bw CT5        : 10
Link Bw CT2       : 10                    Link Bw CT6        : 10
Link Bw CT3       : 10                    Link Bw CT7        : 10
```

```
Bandwidth Constraints for Class Types (Kbps)
BC0               : 1000000             BC4               : 1000000
BC1               : 1000000             BC5               : 1000000
BC2               : 1000000             BC6               : 1000000
BC3               : 1000000             BC7               : 1000000

Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE3->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE4->  Resv. Bw   : 50000               Unresv. Bw        : 950000
TE5->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
---snip---
```

Configuration

The example topology consists of five 7750 SRs in a ring topology, as shown in Figure 204.

Initial Configuration

The nodes have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces. For PE-1:

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-5"
            address 192.168.15.1/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.1/32
        exit
```

- IGP: OSPF (alternatively, IS-IS could have been used) with TE enabled. For PE-1:

```
*A:PE-1# configure
    router
```

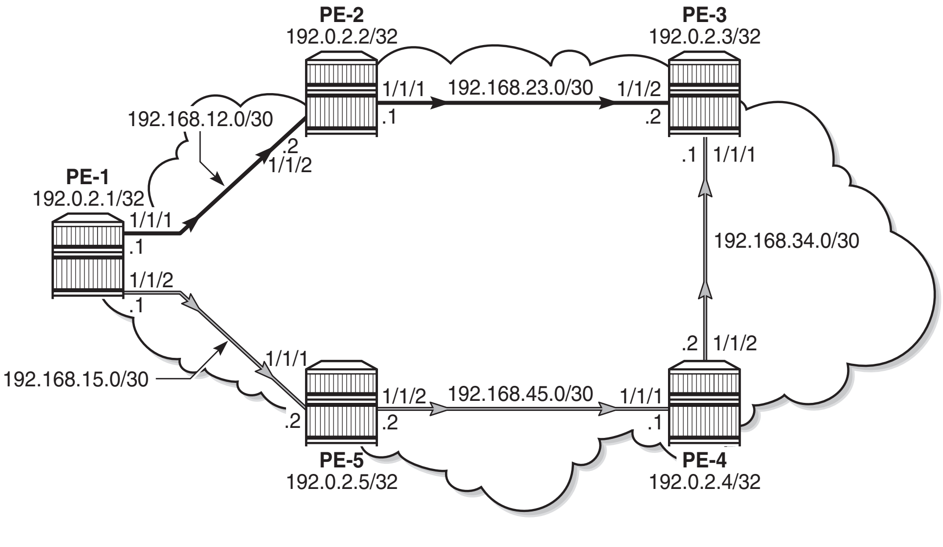
```
            ospf
                traffic-engineering
                area 0.0.0.0
                    interface "system"
                    exit
                    interface "int-PE-1-PE-2"
                        interface-type point-to-point
                    exit
                    interface "int-PE-1-PE-5"
                        interface-type point-to-point
                    exit
                exit
                no shutdown
            exit
```

- MPLS and RSVP enabled on all interfaces. For PE-1:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-5"
            exit
            no shutdown
        exit
        rsvp
            no shutdown
        exit
```

LSPs will be established from PE-1 to PE-3 with the short path via PE-2 as preferred. If insufficient bandwidth is available on the short path via PE-2, the longer path via PE-5 and PE-4 is taken, as shown in Figure 209.

Figure 209 **Paths from PE-1 to PE-3**



Initially, the default BC model, which is MAM, is enabled. Different LSPs will be created with different class type and priority. When an LSP is established, the bandwidth reservation is verified on the interfaces of PE-1. The same LSPs will later be established for the second BC model, RDM, where the bandwidth reservation is more efficient. For simplicity, FRR is not enabled on the LSPs.

Maximum Allocation Model

Enable DiffServ MAM

DiffServ TE can only be configured when MPLS is shutdown. This is something to consider during migration. An error is raised when MPLS is not shutdown, as follows:

```
*A:PE-1# configure router rsvp diffserv-te mam
MINOR: RSVP #1005 Invalid operation for RSVP instance -
 Diffserv can be enabled only when MPLS is shutdown
```

The DiffServ TE configuration must be consistent on all the nodes in the setup. In this example, DiffServ TE is configured as follows:

```
*A:PE-1# configure
    router
        mpls
            shutdown
        exit
        rsvp
            diffserv-te mam
                class-type-bw ct0 50 ct1 40 ct2 0 ct3 0 ct4 0 ct5 0 ct6 0 ct7 10
                te-class 0 class-type 0 priority 7
                te-class 1 class-type 0 priority 4
                te-class 2 class-type 1 priority 7
                te-class 3 class-type 1 priority 4
                te-class 4 class-type 2 priority 7
                te-class 5 class-type 2 priority 2
                fc af class-type 1
                fc be class-type 0
                fc nc class-type 2
            exit
        exit
        mpls
            no shutdown
        exit
```

The bandwidth percentage for each CT must be configured. For unused CTs, the bandwidth percentage must be set to 0. An error is raised if unused CTs are missing , as follows:

```
*A:PE-1# configure router rsvp diffserv-te class-type-bw ct0 50 ct1 50
                                                         ^
Error: Missing parameter
```

The sum of bandwidth percentages can be lower than, but must not exceed 100%, as follows:

```
*A:PE-1# configure router rsvp diffserv-te class-type-
bw ct0 50 ct1 50 ct2 0 ct3 0 ct4 0 ct5 0 ct6 0 ct7 10
MINOR: RSVP #1005 Invalid operation for RSVP instance -
 Total CT percent (110) exceeds 100
```

Fewer than eight classes can be configured, as in this example.

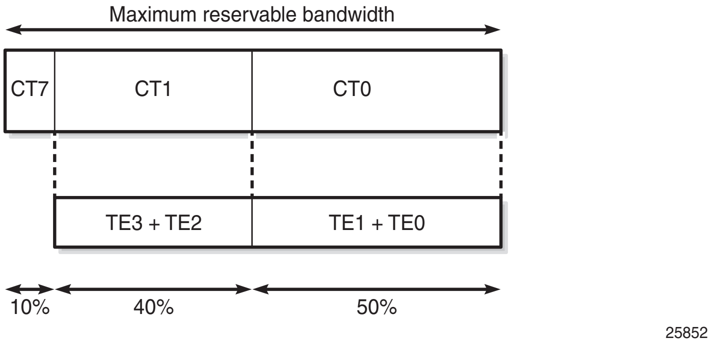
In the example, only three CTs are used by the TE classes: CT0, CT1, and CT2. However, 10% of the maximum reservable bandwidth is allocated to CT7. Because the MAM model does not allow bandwidth allocated to a CT to be used by other CTs, only 90% of the bandwidth can be reserved: 50% to be divided between TE0 and TE1, and 40% to be divided between TE2 and TE3. TE4 and TE5 do not have any bandwidth allocated. The bandwidth allocated to CT7 is completely wasted. This is just an example, not a recommendation.

The same settings will be repeated in the RDM model, where the bandwidth will not be wasted. The following bandwidth information can be seen on any interface:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
---snip---
Percent Link Bandwidth for Class Types*
Link Bw CT0       : 50                Link Bw CT4       : 0
Link Bw CT1       : 40                Link Bw CT5       : 0
Link Bw CT2       : 0                 Link Bw CT6       : 0
Link Bw CT3       : 0                 Link Bw CT7       : 10
Bandwidth Constraints for Class Types (Kbps)
BC0               : 5000000           BC4               : 0
BC1               : 4000000           BC5               : 0
BC2               : 0                 BC6               : 0
BC3               : 0                 BC7               : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                 Unresv. Bw        : 5000000
TE1->  Resv. Bw   : 0                 Unresv. Bw        : 5000000
TE2->  Resv. Bw   : 0                 Unresv. Bw        : 4000000
TE3->  Resv. Bw   : 0                 Unresv. Bw        : 4000000
TE4->  Resv. Bw   : 0                 Unresv. Bw        : 0
TE5->  Resv. Bw   : 0                 Unresv. Bw        : 0
TE6->  Resv. Bw   : 0                 Unresv. Bw        : 0
TE7->  Resv. Bw   : 0                 Unresv. Bw        : 0
---snip---
```

Figure 210 shows the bandwidth allocation for the CTs and TE classes.

Figure 210 **MAM Bandwidth Allocation**



Establishing LSPs

TE class 5 corresponds to CT2 and priority 2. No bandwidth can be reserved for an RSVP LSP with CT2, setup priority 2, and hold priority 2. This can be verified by configuring an empty path and an LSP, as follows:

```
*A:PE-1# configure
    router
        mpls
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-3-TE5"
                to 192.0.2.3
                cspf
                class-type 2
                primary "dyn"
                    bandwidth 1000
                    priority 2 2
                exit
                no shutdown
            exit
```

CSPF must be enabled. The class type is by default CT0, but can be changed to CT2 by configuration. The class type can be configured in the LSP context, as shown here, or in the primary path context. The setup priority and hold priority are configured in the primary path context.

The LSP cannot be established, because no bandwidth is allocated to TE class 5, as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE5" path detail | match "Failure Code"
Failure Code      : noCspfRouteToDestination
```

No bandwidth is reserved on the RSVP interfaces, as follows:

```
*A:PE-1# show router rsvp interface

===============================================================================
RSVP Interfaces
===============================================================================
Interface                        Total    Active    Total BW  Resv BW   Adm Opr
                                 Sessions Sessions  (Mbps)    (Mbps)
-------------------------------------------------------------------------------
system                           -        -         -         -         Up  Up
int-PE-1-PE-2                    0        0         10000     0         Up  Up
int-PE-1-PE-5                    0        0         10000     0         Up  Up
-------------------------------------------------------------------------------
Interfaces : 3
===============================================================================
*A:PE-1#
```

Bandwidth can be reserved for an LSP with CT1 and priorities 4, as for the following LSP:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-TE3"
                to 192.0.2.3
                cspf
                primary "dyn"
                    bandwidth 2000
                    priority 4 4
                    class-type 1
                exit
                no shutdown
            exit
```

In the example, the CT is configured in the primary path context whereas the CT in the previous example was configured in the LSP context.

The path is set up via PE-2 to PE-3, as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3-TE3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3
```

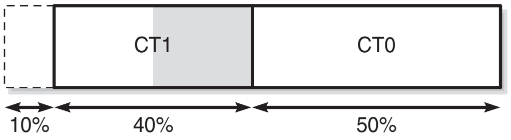
```
Path LSP ID       : 28160
From              : 192.0.2.1           To                     : 192.0.2.3
Admin State       : Up                  Oper State             : Up
Path Name         : dyn                 Path Type              : Primary
Path Admin        : Up                  Path Oper              : Up
Out Interface     : 1/1/1               Out Label              : 262143
---snip---
Actual Hops       :
    192.168.12.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 262143
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 262143
---snip---
```

Bandwidth is reserved in TE class 3, as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail
===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
---snip---
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw    : 0                  Unresv. Bw         : 5000000
TE1->  Resv. Bw    : 0                  Unresv. Bw         : 5000000
TE2->  Resv. Bw    : 0                  Unresv. Bw         : 2000000
TE3->  Resv. Bw    : 2000000            Unresv. Bw         : 2000000
TE4->  Resv. Bw    : 0                  Unresv. Bw         : 0
TE5->  Resv. Bw    : 0                  Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                  Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                  Unresv. Bw         : 0
---snip---
```

Figure 211 shows the bandwidth reservation for CT1 on interface int-PE-1-PE-2 on PE-1 and on interface int-PE-2-PE-3 on PE-2.

Figure 211 **Reserved and Unreserved Bandwidth**



25853

An additional LSP is configured with CT1 and priority 4 and with CT0 as backup CT. The backup CT will not be used when the amount of unreserved bandwidth for CT1 is insufficient, as in the following case where int-PE-1-PE-2 and int-PE-1-PE-5 have insufficient unreserved bandwidth for CT1:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-TE3-backupTE1"
                to 192.0.2.3
                cspf
                primary "dyn"
                    bandwidth 5000
                    priority 4 4
                    class-type 1
                    backup-class-type 0
                exit
                no shutdown
            exit
```

The LSP will not come up, as follows:

```
*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To                Tun     Fastfail  Adm  Opr
                                                     Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-PE-3-TE5                  192.0.2.3         3       No        Up   Dwn
LSP-PE-1-PE-3-TE3                  192.0.2.3         4       No        Up   Up
LSP-PE-1-PE-3-TE3-backupTE1        192.0.2.3         5       No        Up   Dwn
-------------------------------------------------------------------------------
LSPs : 3
```

The bandwidth requirement is lowered, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE3-
backupTE1" primary "dyn" bandwidth 2500
```

Interface int-PE-1-PE-2 does not have sufficient bandwidth for CT1, but the longer path via PE-5 and PE-4 has sufficient unreserved bandwidth for CT1. The LSP will be operationally up, as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE3-backupTE1" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3-TE3-backupTE1 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE3-backupTE1 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3-backupTE1
Path LSP ID      : 23552
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
```

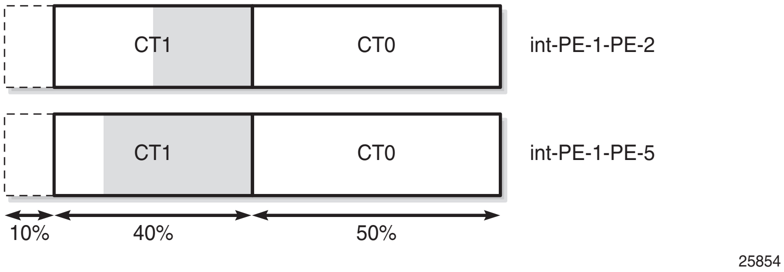
```
Out Interface    : 1/1/2                Out Label            : 262143
---snip---
Setup Priority   : 4                    Oper Setup Priority  : 4
Hold Priority    : 4                    Oper Hold Priority   : 4
Class Type       : 1                    Oper CT              : 1
Backup CT        : 0
---snip---
Actual Hops      :
    192.168.15.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.15.2 (192.0.2.5)                     Record Label        : 262143
 -> 192.168.45.1 (192.0.2.4)                     Record Label        : 262143
 -> 192.168.34.1 (192.0.2.3)                     Record Label        : 262142
---snip---
```

The bandwidth reservation on RSVP interface int-PE-1-PE-2 remains unchanged, because the bandwidth is reserved on int-PE-1-PE-5, as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-5" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-5
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-5
-------------------------------------------------------------------------------
Interface         : int-PE-1-PE-5
Port ID           : 1/1/2
---snip---
Percent Link Bandwidth for Class Types*
Link Bw CT0       : 50                  Link Bw CT4       : 0
Link Bw CT1       : 40                  Link Bw CT5       : 0
Link Bw CT2       : 0                   Link Bw CT6       : 0
Link Bw CT3       : 0                   Link Bw CT7       : 10
Bandwidth Constraints for Class Types (Kbps)
BC0               : 5000000             BC4               : 0
BC1               : 4000000             BC5               : 0
BC2               : 0                   BC6               : 0
BC3               : 0                   BC7               : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 5000000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 5000000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 1500000
TE3->  Resv. Bw   : 2500000             Unresv. Bw        : 1500000
TE4->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE5->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 0
---snip---
```

Figure 212 shows the reserved and unreserved bandwidth on the RSVP interfaces on PE-1.

Figure 212 **Reserved and Unreserved Bandwidth on PE-1**



Trigger Backup Class-Type

This mechanism is described for MAM, but it is also supported in RDM.

Port 1/1/2 is shut down. The long path via PE-5 and PE-4 can no longer be used. However, the LSP has a backup CT (CT0), which is triggered by the port being down, as follows:

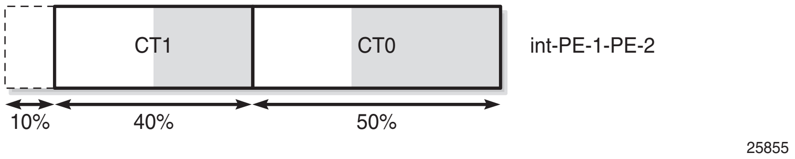
```
*A:PE-1# configure port 1/1/2 shutdown
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE3-backupTE1" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3-TE3-backupTE1 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE3-backupTE1 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3-backupTE1
Path LSP ID      : 23554
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262142
---snip---
Setup Priority   : 4                    Oper Setup Priority  : 4
Hold Priority    : 4                    Oper Hold Priority   : 4
Class Type       : 1                    Oper CT              : 0
Backup CT        : 0
---snip---
Actual Hops      :
    192.168.12.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 262142
 -> 192.168.23.2 (192.0.2.3)                    Record Label        : 262142
```

The bandwidth for this LSP is reserved in TE class 1, as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
Interface          : int-PE-1-PE-2
Port ID            : 1/1/1
Admin State        : Up                    Oper State         : Up
Active Sessions    : 2                     Active Resvs       : 2
Total Sessions     : 2
---snip---
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw    : 0                     Unresv. Bw         : 2500000
TE1->  Resv. Bw    : 2500000               Unresv. Bw         : 2500000
TE2->  Resv. Bw    : 0                     Unresv. Bw         : 2000000
TE3->  Resv. Bw    : 2000000               Unresv. Bw         : 2000000
TE4->  Resv. Bw    : 0                     Unresv. Bw         : 0
TE5->  Resv. Bw    : 0                     Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                     Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                     Unresv. Bw         : 0
---snip---
```

Figure 213 shows the bandwidth reservation on interface int-PE-1-PE-2. The bandwidth reservation on interface int-PE-2-PE-3 on PE-2 is identical.

Figure 213 **Bandwidth Reservation**



The preceding examples illustrate that bandwidth can be wasted in the MAM model. The bandwidth allocated to CT7 cannot be used because there is no TE class configured with CT7. The bandwidth cannot be shared between CTs. The next section explains how the same LSPs will be used in the RDM model. They will be established one-by-one and, therefore, they are shut down, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE5" shutdown
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE3" shutdown
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE3-backupTE1" shutdown
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE1" shutdown

*A:PE-1# configure port 1/1/2 no shutdown
```

Russian Doll Model

Enable DiffServ RDM

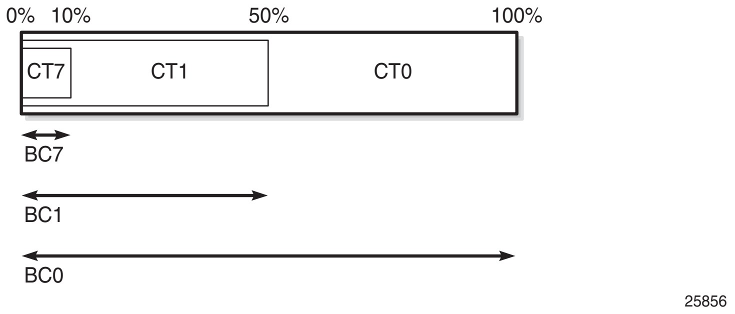
The DiffServ TE configuration needs to be consistent on all the nodes in the network, as follows:

```
*A:PE-1# configure router
        mpls
            shutdown
        exit
        rsvp
            diffserv-te rdm
                class-type-bw ct0 50 ct1 40 ct2 0 ct3 0 ct4 0 ct5 0 ct6 0 ct7 10
                te-class 0 class-type 0 priority 7
                te-class 1 class-type 0 priority 4
                te-class 2 class-type 1 priority 7
                te-class 3 class-type 1 priority 4
                te-class 4 class-type 2 priority 7
                te-class 5 class-type 2 priority 2
                fc af class-type 1
                fc be class-type 0
                fc nc class-type 2
            exit
        exit
        mpls
            no shutdown
        exit
```

In this example, three FCs are mapped to CTs. FC BE corresponds to CT0, FC AF to CT1, and FC NC to CT2. CT0 can be mapped to TE class 0 for priority 7, and to TE class 1 for priority 4. The mapping is similar for CT1 (with priorities 7 and 4), and CT2 (with priorities 7 or 2).

The RDM model allows the outer dolls (lower CT) to use the unused bandwidth allocated to the inner dolls (higher CT), as shown in Figure 214.

Figure 214 **Russian Doll Model for Three Class Types**



The calculation of the BCs takes into account the BCs of the inner dolls, as shown in the OSPF LSAs in the opaque database, as follows:

```
*A:PE-1# show router ospf opaque-database adv-router 192.0.2.1 detail

===============================================================================
Rtr Base OSPFv2 Instance 0 Opaque Link State Database (type: All) (detail)
===============================================================================
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0             Adv Router Id    : 192.0.2.1
Link State Id    : 1.0.0.1             LSA Type         : Area Opaque
Sequence No      : 0x80000003          Checksum         : 0x9035
Age              : 945                 Length           : 28
Options          :  E
Advertisement    : Traffic Engineering
     ROUTER-ID TLV  (0001) Len   4 : 192.0.2.1
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0             Adv Router Id    : 192.0.2.1
Link State Id    : 1.0.0.3             LSA Type         : Area Opaque
Sequence No      : 0x8000000d          Checksum         : 0x25f4
Age              : 15                  Length           : 164
Options          :  E
Advertisement    : Traffic Engineering
     LINK INFO TLV  (0002) Len 140 :
       Sub-TLV: 1     Len: 1     LINK_TYPE    : 1
       Sub-TLV: 2     Len: 4     LINK_ID      : 192.0.2.2
       Sub-TLV: 3     Len: 4     LOC_IP_ADDR  : 192.168.12.1
       Sub-TLV: 4     Len: 4     REM_IP_ADDR  : 192.168.12.2
       Sub-TLV: 5     Len: 4     TE_METRIC    : 10
       Sub-TLV: 6     Len: 4     MAX_BDWTH    : 10000000 Kbps
       Sub-TLV: 7     Len: 4     RSRVBL_BDWTH : 10000000 Kbps
       Sub-TLV: 8     Len: 32    UNRSRVD_CLS0 :
         P0: 10000000 Kbps P1: 10000000 Kbps P2: 5000000 Kbps P3: 5000000 Kbps
         P4: 1000000 Kbps P5: 1000000 Kbps P6:       0 Kbps P7:       0 Kbps
       Sub-TLV: 9     Len: 4     ADMIN_GROUP  : 0 None
       Sub-TLV: 17    Len: 36    TELK_BW_CONST:
         BW Model : RDM
         BC0: 10000000 Kbps BC1: 5000000 Kbps BC2: 1000000 Kbps BC3: 1000000 Kbps
         BC4: 1000000 Kbps BC5: 1000000 Kbps BC6: 1000000 Kbps BC7: 1000000 Kbps
-------------------------------------------------------------------------------
---snip---
```

Six TE classes are defined:

- TE0 and TE1 are defined for CT0. They can reserve all the available bandwidth, if it is not required by the other TE classes (100% = 50% for CT0 + 40% for CT1 + 10% for CT7)
- TE2 and TE3 are defined for CT1. They can reserve 50% of the bandwidth (50% = 40% for CT1 + 10% for CT7)

- TE4 and TE5 are defined for CT2. They can reserve 10% of the bandwidth, even though the configured bandwidth percentage for CT2 is 0. The 10% allocated to higher class CT7 can be used.

Bandwidth is more efficiently used in RDM than in MAM.

The BCs and bandwidth per TE class type show that bandwidth can be shared with the outer dolls, as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

================================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
================================================================================
--------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
--------------------------------------------------------------------------------
Interface         : int-PE-1-PE-2
Port ID           : 1/1/1
Admin State       : Up                  Oper State        : Up
Active Sessions   : 0                   Active Resvs      : 0
Total Sessions    : 0
Subscription      : 100 %               Port Speed        : 10000 Mbps
Total BW          : 10000 Mbps          Aggregate         : Dsabl
---snip---

Percent Link Bandwidth for Class Types*
Link Bw CT0       : 50                  Link Bw CT4       : 0
Link Bw CT1       : 40                  Link Bw CT5       : 0
Link Bw CT2       : 0                   Link Bw CT6       : 0
Link Bw CT3       : 0                   Link Bw CT7       : 10
Bandwidth Constraints for Class Types (Kbps)
BC0               : 10000000            BC4               : 1000000
BC1               : 5000000             BC5               : 1000000
BC2               : 1000000             BC6               : 1000000
BC3               : 1000000             BC7               : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 10000000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 10000000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 5000000
TE3->  Resv. Bw   : 0                   Unresv. Bw        : 5000000
TE4->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE5->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 0
---snip---
```

Establishing LSPs

LSP-PE-1-PE-3-TE5 could not be established in the MAM model, because there was no bandwidth assigned to TE5 (CT2). However, in the RDM model, TE5 can use the bandwidth of the inner doll CT7 and the LSP will be operationally up, as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE5" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3-TE5 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE5 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE5
Path LSP ID      : 50176
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262143
---snip---
Setup Priority   : 2                    Oper Setup Priority  : 2
Hold Priority    : 2                    Oper Hold Priority   : 2
Class Type       : 2                    Oper CT              : 2
Backup CT        : None
---snip---
Actual Hops      :
    192.168.12.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 262143
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 262143
---snip---
```

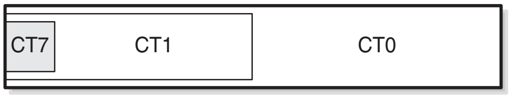
The bandwidth reservation on interface int-PE-1-PE-2 is as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
---snip---

Percent Link Bandwidth for Class Types*
Link Bw CT0       : 50                  Link Bw CT4       : 0
Link Bw CT1       : 40                  Link Bw CT5       : 0
Link Bw CT2       : 0                   Link Bw CT6       : 0
Link Bw CT3       : 0                   Link Bw CT7       : 10
Bandwidth Constraints for Class Types (Kbps)
BC0               : 10000000            BC4               : 1000000
BC1               : 5000000             BC5               : 1000000
BC2               : 1000000             BC6               : 1000000
BC3               : 1000000             BC7               : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 9000000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 9000000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 4000000
TE3->  Resv. Bw   : 0                   Unresv. Bw        : 4000000
```

```
TE4->  Resv. Bw    : 0                   Unresv. Bw         : 0
TE5->  Resv. Bw    : 1000000             Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                   Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                   Unresv. Bw         : 0
---snip---
```

This LSP uses all the available bandwidth for CT7. Because TE5 is defined with the best priority (2) of all TE classes, this LSP will not be preempted when a new LSP is enabled. Therefore, this bandwidth is subtracted from the amount of unreserved bandwidth. The remaining unreserved bandwidth is for CT0 and CT1 only. LSPs with other CTs cannot be established on this interface. Figure 215 shows the reserved bandwidth on interface int-PE-1-PE-2 for this LSP.

Figure 215 **Reserved Bandwidth for LSP with CT2 (One Session)**



25857

Another LSP is established: LSP-PE-1-PE-3-TE3, with CT1 and priority 4, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE3" no shutdown
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3-TE3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE3 Path dyn
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-3-TE3
Path LSP ID       : 28162
From              : 192.0.2.1            To                   : 192.0.2.3
Admin State       : Up                   Oper State           : Up
Path Name         : dyn                  Path Type            : Primary
Path Admin        : Up                   Path Oper            : Up
Out Interface     : 1/1/1                Out Label            : 262142
---snip---
Setup Priority    : 4                    Oper Setup Priority  : 4
Hold Priority     : 4                    Oper Hold Priority   : 4
Class Type        : 1                    Oper CT              : 1
Backup CT         : None
---snip---
Actual Hops       :
    192.168.12.1 (192.0.2.1)                      Record Label        : N/A
```

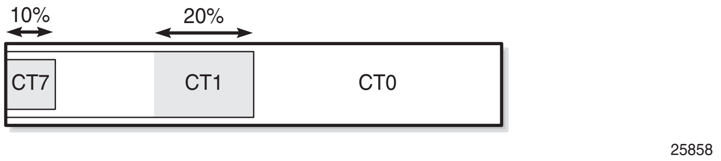
```
 -> 192.168.12.2 (192.0.2.2)                  Record Label        : 262142
 -> 192.168.23.2 (192.0.2.3)                  Record Label        : 262142
---snip---
```

The bandwidth reservation on RSVP interface int-PE-1-PE-2 is as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail

===============================================================================
RSVP Interface (Detailed) : int-PE-1-PE-2
===============================================================================
-------------------------------------------------------------------------------
Interface : int-PE-1-PE-2
-------------------------------------------------------------------------------
Interface          : int-PE-1-PE-2
Port ID            : 1/1/1
Admin State        : Up                    Oper State         : Up
Active Sessions    : 2                     Active Resvs       : 2
Total Sessions     : 2
---snip---

Bandwidth Constraints for Class Types (Kbps)
BC0                : 10000000              BC4                : 1000000
BC1                : 5000000               BC5                : 1000000
BC2                : 1000000               BC6                : 1000000
BC3                : 1000000               BC7                : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw    : 0                     Unresv. Bw         : 7000000
TE1->  Resv. Bw    : 0                     Unresv. Bw         : 7000000
TE2->  Resv. Bw    : 0                     Unresv. Bw         : 2000000
TE3->  Resv. Bw    : 2000000               Unresv. Bw         : 2000000
TE4->  Resv. Bw    : 0                     Unresv. Bw         : 0
TE5->  Resv. Bw    : 1000000               Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                     Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                     Unresv. Bw         : 0
---snip---
```

Figure 216 shows the bandwidth reservation for the two active sessions.

Figure 216 **Bandwidth Reservation for LSP with CT2 and LSP with CT1 (2 Sessions)**



Another LSP is established for CT1, requesting more bandwidth than the short path via PE-2 has available. Therefore, the longer path via PE-5 and PE-4 is set up for this LSP, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3-TE3-backupTE1" no shutdown
```

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3-TE3-backupTE1" path detail

================================================================================
MPLS LSP LSP-PE-1-PE-3-TE3-backupTE1 Path  (Detail)
================================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
================================================================================
--------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3-TE3-backupTE1 Path dyn
--------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3-backupTE1
Path LSP ID      : 23556
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/2                Out Label            : 262143
---snip---
Setup Priority   : 4                    Oper Setup Priority  : 4
Hold Priority    : 4                    Oper Hold Priority   : 4
Class Type       : 1                    Oper CT              : 1
Backup CT        : 0
---snip---
Actual Hops      :
    192.168.15.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.15.2 (192.0.2.5)                     Record Label        : 262143
 -> 192.168.45.1 (192.0.2.4)                     Record Label        : 262143
 -> 192.168.34.1 (192.0.2.3)                     Record Label        : 262141
---snip---
```

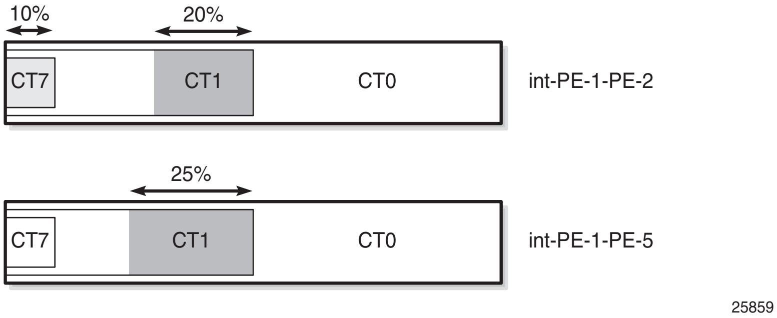
The bandwidth for this LSP is reserved on interface int-PE-1-PE-5, because the amount of unreserved bandwidth for TE3 is insufficient and inner dolls cannot use bandwidth assigned to outer dolls. Inner dolls are of higher priority than outer dolls, as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-5" detail
---snip---
Bandwidth Constraints for Class Types (Kbps)
BC0               : 10000000          BC4                : 1000000
BC1               : 5000000           BC5                : 1000000
BC2               : 1000000           BC6                : 1000000
BC3               : 1000000           BC7                : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw         : 7500000
TE1->  Resv. Bw   : 0                   Unresv. Bw         : 7500000
TE2->  Resv. Bw   : 0                   Unresv. Bw         : 2500000
TE3->  Resv. Bw   : 2500000             Unresv. Bw         : 2500000
TE4->  Resv. Bw   : 0                   Unresv. Bw         : 1000000
TE5->  Resv. Bw   : 0                   Unresv. Bw         : 1000000
TE6->  Resv. Bw   : 0                   Unresv. Bw         : 0
TE7->  Resv. Bw   : 0                   Unresv. Bw         : 0
```

```
---snip---
```

Figure 217 shows the reserved bandwidth on both interfaces of PE-1.

Figure 217 **Reserved Bandwidth on Both Interfaces of PE-1 (3 Sessions)**



The following LSP is configured on PE-1:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-TE1"
                to 192.0.2.3
                cspf
                primary "dyn"
                    bandwidth 3000
                    priority 4 4
                exit
                no shutdown
            exit
```

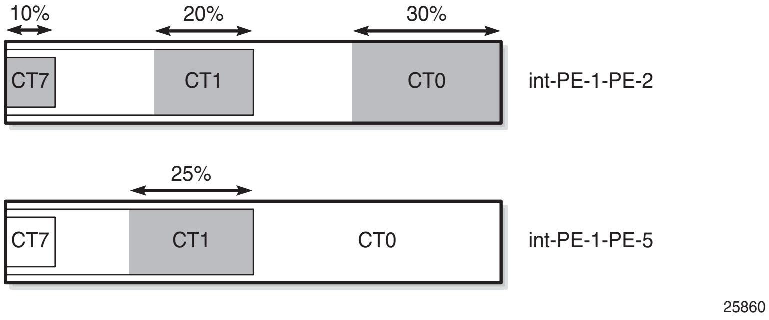
The class type is by default 0. CT0 and priority 4 corresponds to TE1. There is sufficient bandwidth available on the short path via PE-2. The bandwidth reservation on RSVP interface int-PE-1-PE-2 is as follows:

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail
---snip---
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw    : 0                       Unresv. Bw         : 4000000
TE1->  Resv. Bw    : 3000000                 Unresv. Bw         : 4000000
TE2->  Resv. Bw    : 0                       Unresv. Bw         : 2000000
TE3->  Resv. Bw    : 2000000                 Unresv. Bw         : 2000000
TE4->  Resv. Bw    : 0                       Unresv. Bw         : 0
TE5->  Resv. Bw    : 1000000                 Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                       Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                       Unresv. Bw         : 0
---snip---
```

None of the established LSPs can be preempted. Therefore, the sum of the reserved and unreserved bandwidth does not exceed the total bandwidth.

Figure 218 shows the bandwidth reservation on both interfaces.

Figure 218 **Reserved Bandwidth on Both Interfaces on PE-1 (4 Sessions)**



The following LSP with CT0 and priority 7 is configured on PE-1:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-TE0"
                to 192.0.2.3
                cspf
                primary "dyn"
                    bandwidth 4000
                    priority 7 7
                exit
                no shutdown
            exit
```

The bandwidth is reserved in TE class 0. There is sufficient bandwidth on the short path to PE-3. The bandwidth is now reserved for 100%, as follows:

```
*A:PE-1# show router rsvp interface

===============================================================================
RSVP Interfaces
===============================================================================
Interface                        Total    Active   Total BW   Resv BW    Adm Opr
                                 Sessions Sessions (Mbps)     (Mbps)
-------------------------------------------------------------------------------
system                           -        -        -          -          Up  Up
int-PE-1-PE-2                    4        4        10000      10000      Up  Up
int-PE-1-PE-5                    1        1        10000      2500       Up  Up
-------------------------------------------------------------------------------
Interfaces : 3
```

The bandwidth reservation on int-PE-1-PE-2 is as follows:

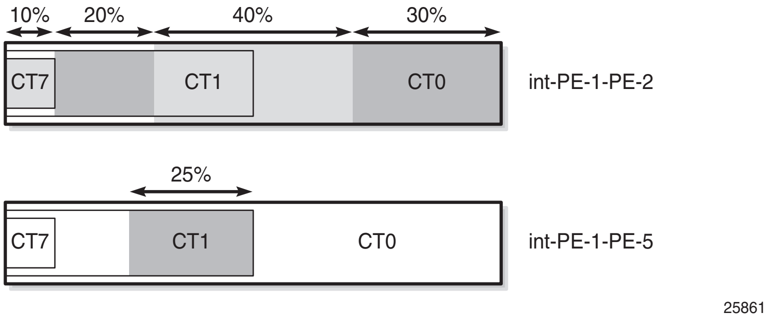
```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail
---snip---
Percent Link Bandwidth for Class Types*
```

```
Link Bw CT0        : 50                  Link Bw CT4        : 0
Link Bw CT1        : 40                  Link Bw CT5        : 0
Link Bw CT2        : 0                   Link Bw CT6        : 0
Link Bw CT3        : 0                   Link Bw CT7        : 10
Bandwidth Constraints for Class Types (Kbps)
BC0                : 10000000            BC4                : 1000000
BC1                : 5000000             BC5                : 1000000
BC2                : 1000000             BC6                : 1000000
BC3                : 1000000             BC7                : 1000000
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw    : 4000000             Unresv. Bw         : 0
TE1->  Resv. Bw    : 3000000             Unresv. Bw         : 4000000
TE2->  Resv. Bw    : 0                   Unresv. Bw         : 0
TE3->  Resv. Bw    : 2000000             Unresv. Bw         : 2000000
TE4->  Resv. Bw    : 0                   Unresv. Bw         : 0
TE5->  Resv. Bw    : 1000000             Unresv. Bw         : 0
TE6->  Resv. Bw    : 0                   Unresv. Bw         : 0
TE7->  Resv. Bw    : 0                   Unresv. Bw         : 0
---snip---
```

Even though the sum of the reserved bandwidth equals the maximum reservable bandwidth on the link, there is still unreserved bandwidth for specific TE classes. When an additional LSP is established requiring bandwidth in TE3 or TE1 (which have setup priority 4), it can preempt another LSP with a lower hold priority. LSPs requiring bandwidth in TE class TE2 have a setup priority 7 and cannot preempt any other LSP. The setup priority in TE1 and TE3 is 4, which is higher than the hold priority in TE2 and TE0 (7 is the lowest priority). There are no LSPs in TE2, so the only LSPs to preempt have bandwidth reserved in TE0.

Figure 219 shows the bandwidth reservation on the interfaces of PE-1.

Figure 219 **Reserved Bandwidth on Both Interfaces of PE-1 (5 Sessions)**



Preemption

The following LSP is configured with CT1, setup priority 4, and hold priority 4, which corresponds to TE class 3:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3-TE3-2nd"
                to 192.0.2.3
                cspf
                class-type 1
                primary "dyn"
                    bandwidth 750
                    priority 4 4
                exit
                no shutdown
            exit
```

Because the setup priority 4 exceeds the hold priority 7 of LSP-PE-1-PE-3-TE0, this LSP will preempt the existing one. The following output shows that the next hop for LSP-PE-1-PE-3-TE0 is 192.168.15.2 (PE-5), while the next hop for LSP-PE-1-PE-3-TE3-2nd is 192.168.12.2 (PE-2):

```
*A:PE-1# show router mpls lsp path

===============================================================================
MPLS LSP  Path
===============================================================================
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE5    To                   : 192.0.2.3
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.12.2    Primary    1/1/1     Up   Up
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3    To                   : 192.0.2.3
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.12.2    Primary    1/1/1     Up   Up
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3-   To                   : 192.0.2.3
                   backupTE1
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.15.2    Primary    1/1/2     Up   Up
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE1    To                   : 192.0.2.3
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.12.2    Primary    1/1/1     Up   Up
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE0    To                   : 192.0.2.3
Adm State        : Up                   Oper State           : Up
-------------------------------------------------------------------------------
```

```
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.15.2    Primary    1/1/2     Up   Up
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3-TE3-    To                   : 192.0.2.3
                   2nd
Adm State        : Up                    Oper State           : Up
-------------------------------------------------------------------------------
Path Name                        Next Hop        Type       Out I/F   Adm  Opr
-------------------------------------------------------------------------------
dyn                              192.168.12.2    Primary    1/1/1     Up   Up
===============================================================================
*A:PE-1#
```

The bandwidth reservation on RSVP interface int-PE-1-PE-2 is as follows:

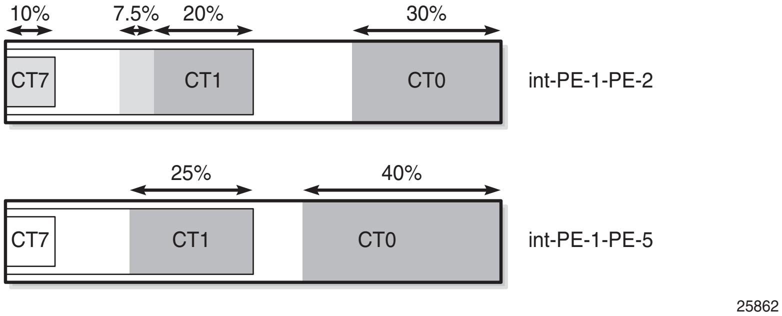
```
*A:PE-1# show router rsvp interface "int-PE-1-PE-2" detail
---snip---
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 0                   Unresv. Bw        : 3250000
TE1->  Resv. Bw   : 3000000             Unresv. Bw        : 3250000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 1250000
TE3->  Resv. Bw   : 2750000             Unresv. Bw        : 1250000
TE4->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE5->  Resv. Bw   : 1000000             Unresv. Bw        : 0
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 0
---snip---
```

The bandwidth reservation on RSVP interface int-PE-1-PE-5 is as follows

```
*A:PE-1# show router rsvp interface "int-PE-1-PE-5" detail
---snip---
Bandwidth for TE Class Types (Kbps)
TE0->  Resv. Bw   : 4000000             Unresv. Bw        : 3500000
TE1->  Resv. Bw   : 0                   Unresv. Bw        : 7500000
TE2->  Resv. Bw   : 0                   Unresv. Bw        : 2500000
TE3->  Resv. Bw   : 2500000             Unresv. Bw        : 2500000
TE4->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE5->  Resv. Bw   : 0                   Unresv. Bw        : 1000000
TE6->  Resv. Bw   : 0                   Unresv. Bw        : 0
TE7->  Resv. Bw   : 0                   Unresv. Bw        : 0
---snip---
```

Figure 220 shows the bandwidth reservation on both interfaces on PE-1 for the six sessions.

Figure 220 **Reserved Bandwidth on Both Interfaces on PE-1 (6 Sessions)**



Preemption can also be within the same CT. An LSP with CT0 and priority 4 (TE1) could have preempted the LSP with CT0 and priority 7 (TE0) equally well.

Bandwidth Availability Check

A tools command can be launched to verify the available bandwidth toward a node for a specific class type (by default CT0) and priority (by default setup priority 7 and hold priority 0). The options for this command are as follows:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3
  - cspf to <ip-addr> [from <ip-addr>] [bandwidth <bandwidth>]
    [include-bitmap <bitmap>] [exclude-bitmap <bitmap>] [hop-limit <limit>]
    [exclude-address <excl-addr> [<excl-addr>...(upto 8 max)]]
    [use-te-metric] [strict-srlg] [srlg-group <grp-id>...(upto 8 max)]
    [exclude-node <excl-node-id> [<excl-node-id>..(upto 8 max)]]
    [skip-interface <interface-name>] [ds-class-type <class-type>]
    [cspf-reqtype <req-type>] [least-fill-min-thd <thd>] [setup-priority <val>]
    [hold-priority <val>]

 <ip-addr>            : a.b.c.d
 <bandwidth>          : [1..100000] in Mbps
 <bitmap>             : [0..4294967295] - accepted in decimal, hex(0x) or binary(0b)
 <limit>              : [2..255]
 <excl-addr>          : a.b.c.d (outbound interface)
 <use-te-metric>      : keyword
 <strict-srlg>        : keyword
 <grp-id>             : [0..4294967295]
 <excl-node-id>       : [a.b.c.d] (outbound interface)
 <interface-name>     : [max 32 chars]
 <class-type>         : [0..7]
 <req-type>           : all|random|least-fill : keywords
 <thd>                : [1..100]
 <priority>           : [0..7]
```

The following verifies whether an LSP can be set up from PE-1 to PE-3 requesting 100 Mb/s with CT0 (default) and both priorities equal to 4:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 bandwidth 100 setup-
priority 4 hold-priority 4
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(DiffServ = RDM) class: 0 , setup Priority 4, Ho
ld Priority 4 TE Class: 1
CSPF Path
To        : 192.0.2.3
Path 1    : (cost 20)
    Src:   192.0.2.1   (= Rtr)
    Egr:   192.168.12.1     -> Ingr:   192.168.12.2    Rtr:   192.0.2.2   (met 10)
    Egr:   192.168.23.1     -> Ingr:   192.168.23.2    Rtr:   192.0.2.3   (met 10)
    Dst:   192.0.2.3   (= Rtr)
*A:PE-1#
```

The short path via PE-2 has sufficient bandwidth for an LSP with these TE requirements (TE class 1 with CT0 and both priorities 4). This is different for TE class 5 (CT2 and priorities 2), where the bandwidth is completely reserved. The following shows that the longer path via PE-5 and PE-4 must be taken:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 bandwidth 100 ds-class-
type 2 setup-priority 2 hold-priority 2
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(DiffServ = RDM) class: 2 , setup Priority 2,
                                    Hold Priority 2 TE Class: 5

CSPF Path
To        : 192.0.2.3
Path 1    : (cost 30)
    Src:   192.0.2.1   (= Rtr)
    Egr:   192.168.15.1     -> Ingr:   192.168.15.2    Rtr:   192.0.2.5   (met 10)
    Egr:   192.168.45.2     -> Ingr:   192.168.45.1    Rtr:   192.0.2.4   (met 10)
    Egr:   192.168.34.2     -> Ingr:   192.168.34.1    Rtr:   192.0.2.3   (met 10)
    Dst:   192.0.2.3   (= Rtr)
*A:PE-1#
```

This tools command can only be launched when a TE class is defined with the requested CT and priority. An error is raised when the request cannot be fulfilled, as follows:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 setup-priority 5
MINOR: CLI No Te class mapped to Class Type 0 , Setup Priority 5.
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 setup-priority 4
MINOR: CLI No Te class mapped to Class Type 0 , Hold Priority 0.
```

Conclusion

DiffServ TE enforces different BCs for different classes of traffic. DiffServ TE controls overbooking and supports preemption. Two BC models are described in this chapter: MAM and RDM.

Entropy Label

This chapter provides information about the Entropy Label (EL).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R4. The CLI in the current edition corresponds to SR OS Release 15.0.R1.

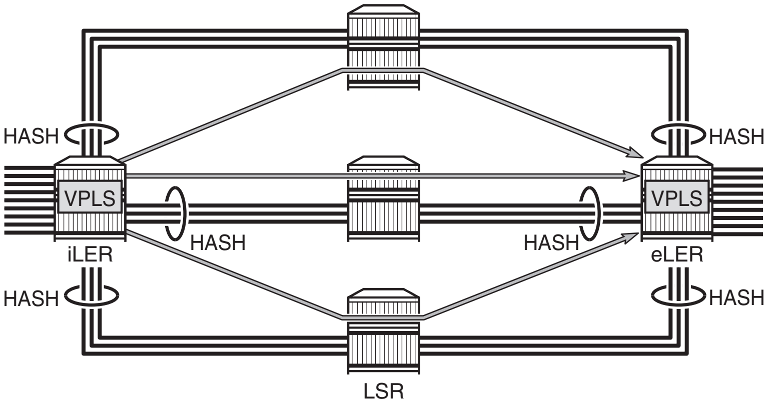
RFC 6391 hash label or flow-aware transport label is supported in SR OS Release 8.0.R1, and later. RFC 6790 Entropy Labels (ELs) are supported on RSVP and BGP tunnels in release 14.0.R1, and later, and supported on LDP tunnels in release 14.0.R4, and later.

Overview

Entropy is the degree of disorder or uncertainty in a system. SR OS supports both the MPLS EL and the hash label, but they are mutually exclusive. These labels allow Label Switched Routers (LSRs) to load-balance labeled packets in a granular way without the need to inspect the IPv4 or IPv6 header. The hash label is applicable to services such as Epipe VLL, VPLS, IES (spoke-SDP), and VPRN services. The main advantage of the EL compared to the hash label is that the EL can be applied to a wider range of services, such as EVPN, BGP VPWS, Fpipe VLL, Ipipe VLL, H-VPLS, and BGP RFC 3107 tunnels.

Figure 221 shows that different flows from an ingress Label Edge Router (iLER) are load-balanced across different paths in the MPLS network toward the egress Label Edge Router (eLER).

Figure 221 **Load-Balancing of Flows Based on Hash Label or Entropy Label**

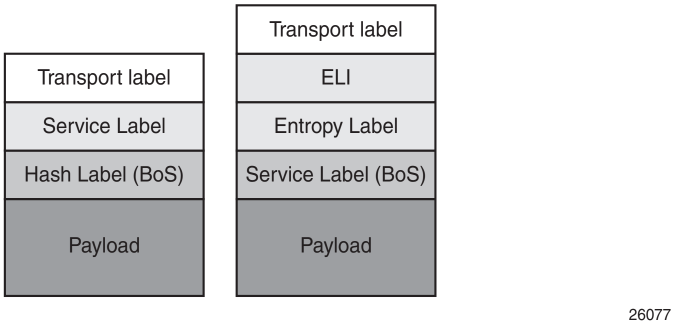


The EL is inserted below the innermost Tunnel Label (TL) in the label stack, along with the Entropy Label Indicator (ELI), which is a special purpose MPLS label with a value of 7 to indicate that the next label in the stack is an EL. As with the hash label, the value of the EL is calculated based on a hash of the packet payload header (IP and Layer 4). The EL is inserted as deep as possible in the stack to ensure preservation for as far as possible through the network. When the EL and ELI are present, load-balancing on the transit LSRs can be configured to only take into account the EL label for Link Aggregation Group (LAG) and Equal Cost Multi-Path (ECMP). Load-balancing on the LSR can be configured to take into account the IP header also, but that is not required when the EL and ELI are present. The eLER removes the EL and ELI before forwarding the packet to its final destination.

The EL requires two additional labels in the label stack, which might result in an unsupported label stack depth in an intermediate (possibly third-party) LSR. SR OS allows control of the EL and ELI insertion and accounting of extra labels in the tunnel table. The supported label stack depth is 12 in SR OS release 14.0, including transport, service, hash, and OAM labels.

Figure 222 shows a comparison between a label stack with a hash label and a label stack with an EL and ELI.

Figure 222 **Label Stack with Hash Label versus Label Stack with EL and ELI**

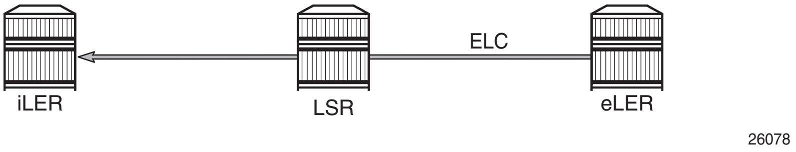


Hash labels or ELs are inserted by the iLER, based on the incoming packet header. The iLER creates a unique label per conversation, based on the incoming packet header (IP and Layer 4). The hash label is a single label that is inserted at the bottom of the stack, below the service label. The EL is inserted together with the ELI below the innermost tunnel label, but above the service label. The entropy and hash label value is outside of the range for MPLS label values, because the most significant bit equals 1 for entropy and hash labels and 0 for MPLS labels.

The EL is used as part of the LSR hashing algorithm for spraying packets over multi-port LAGs, ECMP, or BGP tunnels with multiple downstream interfaces. The packet ordering is preserved because one label is used per conversation. Hashing based on a label stack containing an EL per service will become more granular than when based on a standard label stack alone.

Figure 223 shows how the eLER signals its EL capability to the iLER.

Figure 223 **Downstream LERs Signal EL Capability to ILER**



The Entropy Label Capability (ELC) is signaled by the eLER and indicates the ability to receive and process the ELs. This can be advertised for LDP and RSVP. However, ELC signaling is not supported for BGP tunnels, because no agreed standard exists in the IETF. RFC 6790 introduced the ELC BGP attribute that can be signaled by the eLER to indicate that it supports EL. According to RFC 6790, LSRs incapable of processing ELs must remove the ELC BGP attribute, but this requirement could not be guaranteed; therefore, the ELC BGP attribute has been deprecated in RFC 7447.

As a workaround, at the iLER, an override of the ELC for a BGP tunnel can be manually configured. After this ELC has been overridden, the BGP sender assumes that the receiver is capable of receiving and processing the ELs, regardless of the signaled ELC. The iLER inserts an EL on a tunnel for which the ELC is confirmed by the downstream peer or when the ELC is overridden by configuration.

ELC signaling can be enabled for LDP on the LERs, as follows:

```
*A:PE-2# configure router ldp entropy-label-capability
```

ELC signaling for RSVP can be enabled as follows:

```
*A:PE-2# configure router rsvp entropy-label-capability
```

As previously described, the lack of an IETF standard for BGP tunnels means that ELC is ignored by the receiving LER, and no EL is inserted. Therefore, an override is required that assumes that the far-end LER has ELC, and allows insertion of the EL. The override is enabled as follows:

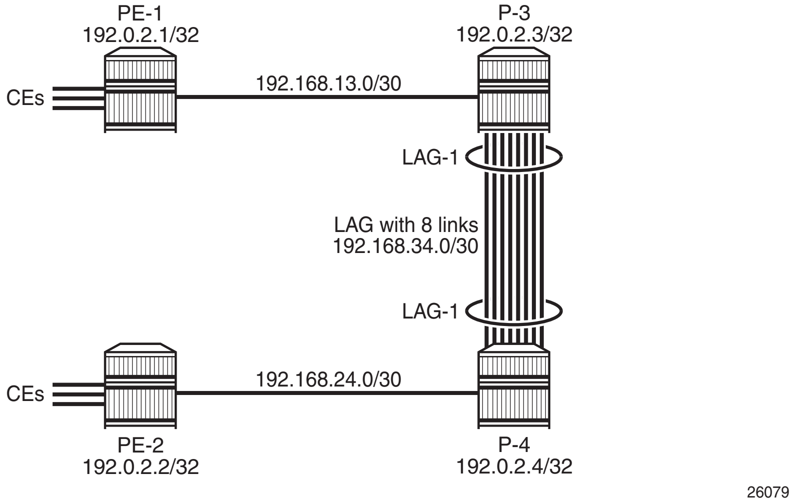
```
*A:PE-1# configure router bgp override-tunnel-elc
```

When the ELC is overridden for BGP, the iLER assumes that the receiver can handle the EL.

Configuration

In this section, an EVPN-VPLS will be configured on PE-1 and PE-2 to illustrate LAG hashing based on EL. Figure 224 shows the topology used for this example. Load-balancing of the traffic will be performed in the P-routers that are connected by a LAG with eight links.

Figure 224 **Example Topology**



The initial configuration of the PE/P-routers includes the following:

- Cards, MDAs, ports
- LAG with eight network links between P-3 and P-4
- Router interfaces
- IGP (IS-IS or OSPF)
- LDP enabled on all interfaces
- MPLS enabled on all interfaces. RSVP enabled.
- iBGP configured with P-3 as route reflector (RR)

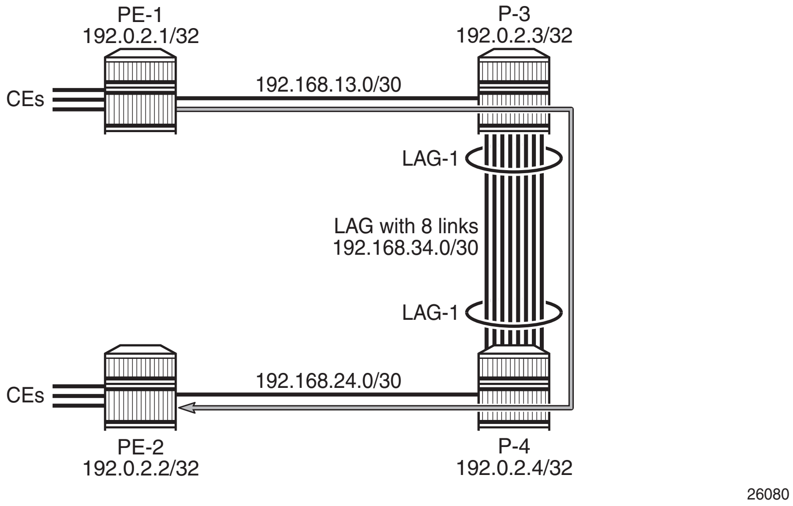
For the configuration of EVPN-MPLS, the BGP configuration needs to include the address family EVPN, as follows. See chapter *EVPN for MPLS Tunnels* for more information.

```
configure
    router
        autonomous-system 64500
        bgp
            min-route-advertisement 1
            rapid-update evpn
            group "iBGP"
                family evpn
                peer-as 64500
                neighbor 192.0.2.3
                exit
            exit
        exit
```

EVPN Service Using RSVP Tunnel with EL

Figure 225 shows LSP "LSP-PE-1-PE-2" from PE-1 to PE-2 via core routers P-3 and P-4.

Figure 225 **RSVP LSP "LSP-PE-1-PE-2" from PE-1 to PE-2 via P-3 and P-4**



The LSP "LSP-PE-1-PE-2" is configured on PE-1, as follows:

```
configure
    router
        mpls
            path "path-PE-1-PE-2"
                hop 10 192.168.13.2 strict
                hop 20 192.168.34.2 strict
                hop 30 192.168.24.1 strict
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                primary "path-PE-1-PE-2"
                exit
                no shutdown
            exit
        exit
    exit
exit
```

The configuration for LSP "LSP-PE-2-PE-1" on PE-2 is similar.

ELC is disabled by default, and needs to be enabled for RSVP on the eLERs, as follows:

```
*A:PE-2# configure router rsvp entropy-label-capability
```

The EL can be disabled (force-disable) or enabled in the MPLS context on the iLER, as follows:

```
*A:PE-1# configure router mpls entropy-label
  - entropy-label rsvp-te <rsvp-te>

 <rsvp-te>            : <force-disable | enable>
```

The EL can also be disabled (force-disable) or enabled per LSP, but there is a third option to inherit the EL settings from the MPLS context, as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" entropy-label
  - entropy-label {force-disable | inherit | enable}

 <force-disable | i*> : force-disable|inherit|enable
```

By default, the EL on the LSP inherits the EL settings in the MPLS context, as follows:

```
*A:PE-1# configure router mpls
*A:PE-1>config>router>mpls#  info detail
----------------------------------------------
            ---snip---
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                ---snip---
                entropy-label inherit
                ---snip---
```

When LSP templates are used, EL can be configured within the LSP template context for single-hop and mesh point-to-point LSPs, as follows:

```
*A:PE-1# configure router mpls lsp-template "LSPtemplate1" one-hop-p2p entropy-
label
  - entropy-label {force-disable | inherit | enable}

 <force-disable | i*> : force-disable|inherit|enable


*A:PE-1# configure router mpls lsp-template "LSPtemplate2" mesh-p2p entropy-label
  - entropy-label {force-disable | inherit | enable}

 <force-disable | i*> : force-disable|inherit|enable
```

When the EL settings are modified, for example, from inherit to enabled, the changes only take effect after the LSP has been cleared or MPLS has been bounced, using shutdown/no shutdown. The following message is raised when the configuration is modified for an LSP in no shutdown state:

```
*A:PE-1>config>router>mpls>lsp# entropy-label enable
INFO: MPLS #1029 Entropy Label change is not operational - LSP must be bounced
```

The following command shows that EL is enabled in MPLS:

```
*A:PE-1# show router mpls status | match Label
Entropy Label RSVP-TE     : Enabled
*A:PE-1#
```

The following command shows that EL is enabled in RSVP:

```
*A:PE-1# show router rsvp status | match Label
Implicit Null Label: Disabled           Node-id in RRO     : Exclude
DiffServTE AdmModel: Basic              Entropy Label      : Enabled
*A:PE-1#
```

The following command shows that EL is enabled and operational in LSP "LSP-PE-1-PE-2":

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" detail | match Label
Entropy Label   : Enabled                   Oper Entropy Label   : Enabled
*A:PE-1#
```

In this case, the EL was configured to be enabled on the LSP, instead of the default behavior, which would have been displayed as "Inherited", as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" detail | match Label
Entropy Label   : Inherited                   Oper Entropy Label   : Enabled
*A:PE-1#
```

The following command shows that the LSP "LSP-PE-1-PE-2" has the flag "entropy-label-capable" in the RSVP tunnel table:

```
*A:PE-1# show router tunnel-table protocol rsvp detail

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination      : 192.0.2.2/32
NextHop          : 192.168.13.2
Tunnel Flags     : exclude-for-lfa entropy-label-capable
Age              : 00h14m59s
CBF Classes      : (Not Specified)
Owner            : rsvp                 Encap            : MPLS
Tunnel ID        : 1                    Preference       : 7
Tunnel Label     : 262137               Tunnel Metric    : 16777215
Tunnel MTU       : 1496                 Max Label Stack  : 1
LSP ID           : 49666                Bypass Label     : 0
LSP Bandwidth    : 0                    LSP Weight       : 0
-------------------------------------------------------------------------------
Number of tunnel-table entries          : 1
Number of tunnel-table entries with LFA : 0
===============================================================================
*A:PE-1#
```

EL is supported on many Layer 2 and Layer 3 services. In this example, VPLS 1 is configured on PE-1 with BGP EVPN-MPLS, as follows:

```
configure
    service
        vpls 1 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 1
                mpls
                    entropy-label
                    auto-bind-tunnel
                        resolution-filter
                            rsvp
                        exit
                        resolution filter
                    exit
                    no shutdown
                exit
            exit
            sap 1/1/2:1 create
            exit
            no shutdown
```

Auto-bind-tunnel is resolved to the RSVP LSP "LSP-PE-1-PE-2". A similar configuration is applied on PE-2.

The following command shows that the EL is enabled for BGP-EVPN, as follows:

```
*A:PE-1# show service id 1 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
---snip---

===============================================================================
BGP EVPN MPLS Information
===============================================================================
Admin Status       : Enabled
Force Vlan Fwding  : Disabled           Control Word       : Disabled
Split Horizon Group: (Not Specified)
Ingress Rep BUM Lbl: Disabled           Max Ecmp Routes    : 0
Ingress Ucast Lbl  : 262138             Ingress Mcast Lbl  : 262138
Entropy Label      : Enabled
RestProtSrcMacAct  : none
Evpn Mpls Encap    : Enabled            Evpn MplsoUdp      : Disabled
===============================================================================

===============================================================================
BGP EVPN MPLS Auto Bind Tunnel Information
===============================================================================
Resolution         : filter
Filter Tunnel Types: rsvp
===============================================================================
*A:PE-1#
```

The iLER PE-1 will add the EL and ELI to the label stack. Traffic load-balancing will be performed in P-3 where the traffic will be sprayed over all eight links of the LAG. By default, the load-balancing settings on P-3 are as follows:

```
*A:P-3>config>system>load-balancing# info detail
----------------------------------------------
            no l4-load-balancing
            no lsr-load-balancing
            no system-ip-load-balancing
            no mc-enh-load-balancing
            no service-id-lag-hashing
----------------------------------------------
```

When the EL and ELI are included in the stack, only the EL label is used for the load-balancing. It is not required to inspect the IP header or any other header. The options for LSR load-balancing are the following:

```
*A:P-3# configure system load-balancing lsr-load-balancing
  - lsr-load-balancing <hashing-algorithm>
  - no lsr-load-balancing

 <hashing-algorithm>  : lbl-only | lbl-ip | ip-only | eth-encap-ip | lbl-ip-l4-teid
```

The option **lbl-only** is preferred when the EL is present, and is configured as follows:

```
*A:P-3# configure system load-balancing lsr-load-balancing lbl-only
```

For the reverse path, PE-2 is the iLER, and P-4 will do the spraying of the packets over all links of the LAG. Therefore, LSR load-balancing with lbl-only is also configured on P-4.

EVPN Service Using LDP Tunnel with EL

ELC is disabled by default for LDP, as follows:

```
*A:PE-1# show router ldp status

===============================================================================
LDP Status for IPv4 LSR ID 192.0.2.1
               IPv6 LSR ID ::
===============================================================================
---snip---
Admin State        : Up
IPv4 Oper State    : Up                   IPv6 Oper State      : Down
---snip---

Entropy Label Capa*: False
---snip---
```

The command to enable ELC is as follows:

```
*A:PE-1# configure router ldp entropy-label-capability
```

In the list of LDP active bindings, the egress labels that are pushed by iLER PE-1 on traffic toward an eLER capable of handling an EL get the indication "e", as follows:

```
*A:PE-1# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop             262143      --
  --                                          --

192.0.2.2/32                                Push              --      262140e
192.168.13.2                                1/1/1

192.0.2.2/32                                Swap            262132    262140
192.168.13.2                                1/1/1

192.0.2.3/32                                Push              --      262143
192.168.13.2                                1/1/1

192.0.2.4/32                                Push              --      262141
192.168.13.2                                1/1/1

192.0.2.4/32                                Swap            262133    262141
192.168.13.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 6
===============================================================================
*A:PE-1#
```

Because the iLER adds the EL, only labels which are pushed get the "e"-indication. Labels which are swapped or popped do not get this label.

The details of the tunnel table for LDP show that the LDP tunnel toward PE-2 has tunnel flag entropy-label-capable, as follows:

```
*A:PE-1# show router tunnel-table protocol ldp detail

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination      : 192.0.2.2/32
NextHop          : 192.168.13.2
Tunnel Flags     : entropy-label-capable
Age              : 00h00m30s
CBF Classes      : (Not Specified)
Owner            : ldp                  Encap            : MPLS
Tunnel ID        : 65539                Preference       : 9
Tunnel Label     : 262140               Tunnel Metric    : 30
Tunnel MTU       : 1496                 Max Label Stack  : 1
-------------------------------------------------------------------------------
---snip---
```

The following EVPN-VPLS uses an LDP transport tunnel and has EL enabled:

```
configure
    service
        vpls 2 customer 1 create
            bgp
            exit
            bgp-evpn
                evi 2
                mpls
                    entropy-label
                    auto-bind-tunnel
                        resolution-filter
                            ldp
                        exit
                        resolution filter
                    exit
                    no shutdown
                exit
            exit
            sap 1/1/2:16 create
            exit
            no shutdown
```

The configuration on PE-1 and PE-2 is similar.

EL is enabled in this service, as follows:

```
*A:PE-2# show service id 2 bgp-evpn

===============================================================================
BGP EVPN Table
===============================================================================
---snip---
===============================================================================
BGP EVPN MPLS Information
```

```
===============================================================================
Admin Status       : Enabled
Force Vlan Fwding  : Disabled           Control Word       : Disabled
Split Horizon Group: (Not Specified)
Ingress Rep BUM Lbl: Disabled           Max Ecmp Routes    : 0
Ingress Ucast Lbl  : 262130             Ingress Mcast Lbl  : 262130
Entropy Label      : Enabled
RestProtSrcMacAct  : none
Evpn Mpls Encap    : Enabled            Evpn MplsoUdp      : Disabled
===============================================================================
===============================================================================
BGP EVPN MPLS Auto Bind Tunnel Information
===============================================================================
Resolution         : filter
Filter Tunnel Types: ldp
===============================================================================
*A:PE-2#
```

LSR Load-Sharing Based on EL

The LSR load-sharing based on the EL results in a granular load-sharing, as shown in the following test results. In each direction, 9000 flows were generated; each flow with four unique variables: source IP address, destination IP address, source UDP port, and destination UDP port. For this test, an EVPN-VPLS with RSVP tunnel was used. Each flow gets a unique EL and the P-nodes distribute the load over all eight links in the LAG, based on the hashing algorithm "lbl-only", as follows:

```
A:P-4# monitor lag 1 rate

===============================================================================
Monitor statistics for LAG ID 1
===============================================================================
Port-id   Input       Input       Output      Output      Input      Output
          Bytes       Packets     Bytes       Packets     Errors     Errors
-------------------------------------------------------------------------------
At time t = 15 sec (Mode: Rate)
-------------------------------------------------------------------------------
2/1/12    93943884    61572       94542017    61964       0          0
  % Util  76.14       --          76.62       --          --         --
2/1/13    94156935    61711       94535383    61959       0          0
  % Util  76.31       --          76.61       --          --         --
2/1/14    94042970    61637       94005352    61612       0          0
  % Util  76.22       --          76.19       --          --         --
2/1/15    94075016    61657       93999248    61608       0          0
  % Util  76.24       --          76.18       --          --         --
2/1/16    94162530    61715       93992635    61603       0          0
  % Util  76.31       --          76.17       --          --         --
2/1/17    94030277    61628       93986023    61599       0          0
  % Util  76.21       --          76.17       --          --         --
2/1/18    93993653    61604       94502319    61938       0          0
  % Util  76.18       --          76.59       --          --         --
2/1/19    93765261    61455       94495707    61933       0          0
  % Util  75.99       --          76.58       --          --         --
```

```
-------------------------------------------------------------------------------
Totals    752170526   492979      754058684   494216      0          0
  % Util  76.20       --          76.39       --          --         --
```

When one of the links is unavailable, the traffic is redistributed over the remaining seven links, as follows:

```
A:P-4# monitor lag 1 rate

===============================================================================
Monitor statistics for LAG ID 1
===============================================================================
Port-id   Input       Input       Output      Output      Input      Output
          Bytes       Packets     Bytes       Packets     Errors     Errors
-------------------------------------------------------------------------------
At time t = 18 sec (Mode: Rate)
-------------------------------------------------------------------------------
2/1/12    107659513   70560       108164110   70891       0          0
  % Util  87.25       --          87.66       --          --         --
2/1/13    107683907   70576       108245984   70944       0          0
  % Util  87.27       --          87.73       --          --         --
2/1/14    107717479   70598       107593365   70516       0          0
  % Util  87.30       --          87.20       --          --         --
2/1/15    107213899   70267       107511469   70462       0          0
  % Util  86.89       --          87.13       --          --         --
2/1/16    107426013   70407       107674751   70570       0          0
  % Util  87.06       --          87.26       --          --         --
2/1/17    107387355   70381       107757155   70624       0          0
  % Util  87.03       --          87.33       --          --         --
2/1/18    107086224   70184       108326862   70997       0          0
  % Util  86.79       --          87.79       --          --         --
-------------------------------------------------------------------------------
Totals    752174390   492973      755273696   495004      0          0
  % Util  87.08       --          87.44       --          --         --
```

The flows are evenly spread in each direction.

Conclusion

The EL is a standard-compliant way to maintain packet ordering within a conversation while load-balancing across multiple ECMP paths or links in a LAG. The EL is supported for both Layer 2 and Layer 3 services and is, therefore, a more general solution than the MPLS hash label.

IGP Shortcuts

This chapter provides information about IGP shortcuts.

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS when the feature is not related to BGP.There are no other prerequisites for this configuration. This chapter was initially written for SR OS release 12.0.R3, but the CLI in the current edition corresponds to SR OS release 16.0.R3.

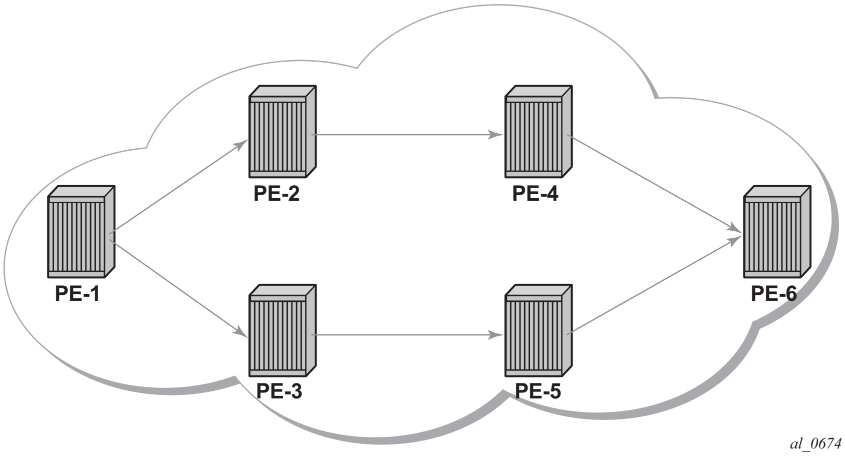
Overview

Interior Gateway Protocols (IGPs) are routing protocols that operate inside an Autonomous System (AS). An AS is a network domain that is managed under a single administration. Because the scope of operation of an IGP is usually within an AS, IGPs are also called intra-AS protocols. The purpose of an IGP is to provide reachability information to destination nodes that are inside the domain. IGPs can be one or more of a variety of protocols, including routing protocols such as Routing Information Protocol (RIP) version 1 or 2, Open Shortest Path First (OSPF), and Intermediate System to Intermediate System (IS-IS).

IGPs such as OSPF and IS-IS are link-state protocols that use a Shortest Path First (SPF) algorithm to compute the shortest path tree to all nodes in a network. The results of such computations indicate the destination node, next hop address, and output interface, where the output interface is a physical interface. Optionally, Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) can be included in the SPF algorithm on the node performing the calculations, as LSPs behave as logical interfaces directly connected to remote nodes in the network. Because the SPF algorithm treats the LSPs in the same way as a physical interface (being a potential output interface), the computation results could be to select a destination node together with an output LSP, using the LSP as a shortcut through the network to the destination node.

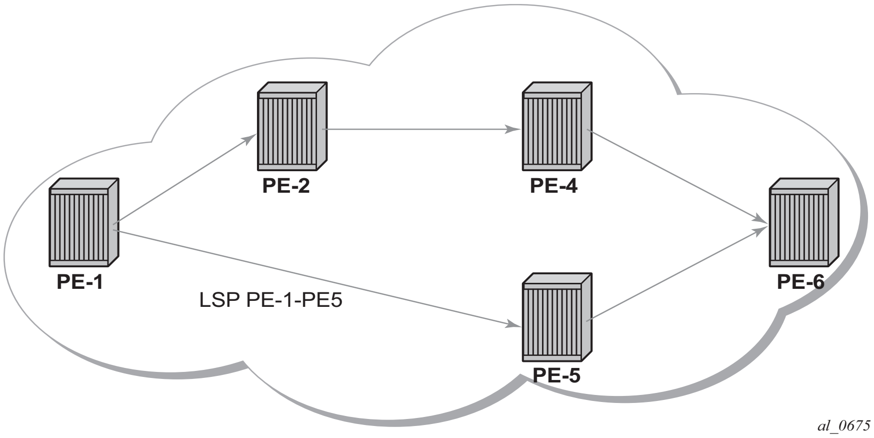
Figure 226 shows a normal SPF tree sourced by PE-1 (Provider Edge-1).

Figure 226 **Normal SPF Tree Sourced by PE-1**



If there is an LSP that connects PE-1 to PE-5, and IGP shortcuts are configured on PE-1, the SPF tree will be as shown in Figure 227.

Figure 227 **SPF Tree Sourced by PE-1 Using LSP Shortcuts**

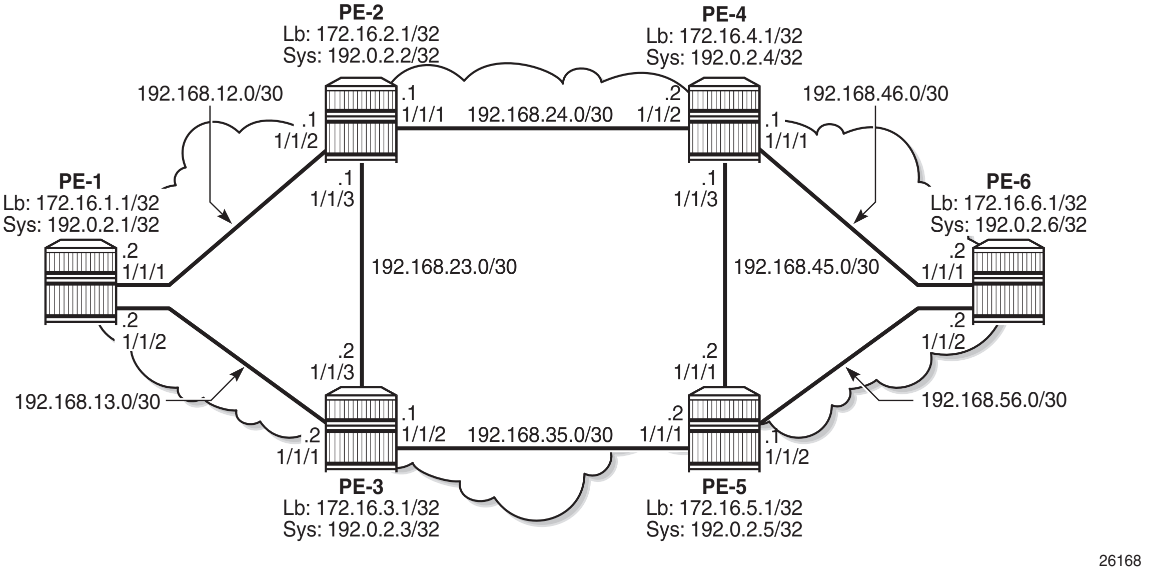


IGP shortcuts are enabled on a per router basis; SPF computations are independent and irrelevant to other routers, so there is no need to enable shortcuts on every single router.

The example topology used in this example is shown in Figure 228. The setup consists of six 7750 service routers. There is a single AS and a single IGP area. The following configuration tasks should be completed first:

- IS-IS or OSPF on all interfaces within the AS (configuration has been done using IS-IS but using OSPF shows exactly the same behavior).
- Label Distribution Protocol (LDP) and Resource Reservation Protocol (RSVP) on all interfaces within the AS.

Figure 228 **Example Topology**



In all figures, **Lb** stands for Loopback and **Sys** stands for the system IP addresses.

Configuration

The first step is to configure the IGP (IS-IS) on all nodes, where IS-IS redistributes route reachability to all routers. To facilitate IS-IS configuration, all routers are L2-L1 capable within the same IS-IS area-id so there is only a single topology area in the network (all routers share the same topology). Traffic engineering (TE) is enabled on the IGP as it is a requirement for RSVP. The metric is using the default values: because no reference bandwidth command is used, the default metric of 10 is applicable on all interfaces. The configuration for PE-2 is as follows.

```
*A:PE-2# configure
    router
        interface "int-PE-2-PE-1"
            address 192.168.12.2/30
            port 1/1/2
        exit
        interface "int-PE-2-PE-3"
            address 192.168.23.1/30
            port 1/1/3
        exit
        interface "int-PE-2-PE-4"
            address 192.168.24.1/30
            port 1/1/1
        exit
        interface "system"
            address 192.0.2.2/32
        exit
        isis
            area-id 49.0001
            traffic-engineering
            interface "system"
                passive
            exit
            interface "int-PE-2-PE-1"
                interface-type point-to-point
            exit
            interface "int-PE-2-PE-4"
                interface-type point-to-point
            exit
            interface "int-PE-2-PE-3"
                interface-type point-to-point
            exit
            no shutdown
```

The configuration for the other nodes is similar. The IP addresses can be derived from Figure 228.

The global route table (GRT) for PE-2 is as follows:

```
*A:PE-2# show router route-table

===============================================================================
Route Table (Router: Base)
```

```
===============================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                   Remote  ISIS      00h00m46s  15
       192.168.12.1                                                 10
192.0.2.2/32                                   Local   Local     00h02m00s  0
       system                                                       0
192.0.2.3/32                                   Remote  ISIS      00h00m38s  15
       192.168.23.2                                                 10
192.0.2.4/32                                   Remote  ISIS      00h00m23s  15
       192.168.24.2                                                 10
192.0.2.5/32                                   Remote  ISIS      00h00m18s  15
       192.168.23.2                                                 20
192.0.2.6/32                                   Remote  ISIS      00h00m08s  15
       192.168.24.2                                                 20
192.168.12.0/30                                Local   Local     00h02m00s  0
       int-PE-2-PE-1                                                0
192.168.13.0/30                                Remote  ISIS      00h00m46s  15
       192.168.12.1                                                 20
192.168.23.0/30                                Local   Local     00h02m00s  0
       int-PE-2-PE-3                                                0
192.168.24.0/30                                Local   Local     00h02m00s  0
       int-PE-2-PE-4                                                0
192.168.35.0/30                                Remote  ISIS      00h00m38s  15
       192.168.23.2                                                 20
192.168.45.0/30                                Remote  ISIS      00h00m23s  15
       192.168.24.2                                                 20
192.168.46.0/30                                Remote  ISIS      00h00m23s  15
       192.168.24.2                                                 20
192.168.56.0/30                                Remote  ISIS      00h00m17s  15
       192.168.23.2                                                 30
-------------------------------------------------------------------------------
No. of Routes: 14
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-2#
```

LDP and RSVP Shortcuts

Interface Label Distribution Protocol (iLDP) is enabled on all interfaces (except system interfaces, which is not allowed) in all routers. The configuration on all nodes is similar and the IP addresses are derived from Figure 228. The configuration of PE-4 is as follows:

```
*A:PE-4# configure
    router
        ldp
            interface-parameters
                interface "int-PE-4-PE-2" dual-stack
                    ipv4
```

```
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-PE-4-PE-5" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-PE-4-PE-6" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

With iLDP enabled, PE-4 establishes iLDP sessions with its directly connected neighbors, as follows:

```
*A:PE-4# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   1205      1204      0d 00:53:02
192.0.2.5:0         Link      Established   1198      1197      0d 00:52:55
192.0.2.6:0         Link      Established   181       183       0d 00:07:43
------------------------------------------------------------------------------
No. of IPv4 Sessions: 3
==============================================================================
*A:PE-4#
```

The following tunnel table shows that there is a Label Switched Path (LSP) to every other router. The reason is that the LDP label distribution mode is downstream unsolicited (DU) by default, originating label bindings for system addresses only (which are used by iLDP as transport address by default). The command also shows the preference of the LSPs (where the preference is 9 for LDP) and the metric of the LSPs (the metric is inherited from the IGP, each hop counts as a metric of 10), as follows. The metric to destinations PE-1 and PE-3 is 20 because there are two hops in between (PE-4 is two hops away from PE-1 and PE-3).

```
*A:PE-4# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65538     9        192.168.24.1   20
192.0.2.2/32      ldp       MPLS  65537     9        192.168.24.1   10
```

```
192.0.2.3/32       ldp       MPLS  65539     9        192.168.24.1   20
192.0.2.5/32       ldp       MPLS  65540     9        192.168.45.2   10
192.0.2.6/32       ldp       MPLS  65545     9        192.168.46.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-4#
```

In order to configure RSVP shortcuts, RSVP must be enabled on all interfaces where traffic engineering is required, but in this example, MPLS and RSVP are enabled on all interfaces of the network. By default, MPLS is enabled on the system interface, therefore, it need not be configured explicitly. When RSVP is in **no shutdown**, it is automatically configured on the interfaces where MPLS is configured. The configuration for PE-6 is as follows.

```
*A:PE-6# configure router mpls no shutdown
*A:PE-6# configure router rsvp no shutdown
```

```
*A:PE-6# configure
    router
        mpls
            interface "int-PE-6-PE-4"
            exit
            interface "int-PE-6-PE-5"
            exit
            no shutdown
        exit
        rsvp no shutdown
```

The configuration of the other nodes is similar. The IP addresses can be derived from Figure 228. Because there are no RSVP LSPs configured yet, the tunnel table has no RSVP LSPs and only contains LDP LSPs.

LDP Static Route (IP Tunneled in LDP Tunnel)

Using LDP LSP shortcuts for static route resolution enables forwarding of IPv4 packets over LDP LSPs instead of using a regular IP next hop. In other words, the traffic to the resolved static routes is forwarded using MPLS LDP LSP rather than plain IP.

The configuration defines a static route pointing to the destination PE (remote loopback, which is an indirect next hop in the example), and explicitly indicates that it should use LDP rather than IGP. Taking PE-1 and PE-6 as an example, two loopback interfaces are configured (172.16.X.1/32), where X = PE number, and a static route is defined according to the preceding explanation. The following shows the configuration on PE-1.

```
*A:PE-1# configure
    router
        interface "loopback"
            address 172.16.1.1/32
            loopback
        exit
        static-route-entry 172.16.6.1/32
            indirect 192.0.2.6
                tunnel-next-hop
                    resolution-filter
                        ldp
                    exit
                    disallow-igp
                    resolution filter
                exit
                no shutdown
            exit
        exit
```

Looking at the GRT or forwarding information base (FIB), there are two new entries corresponding to the two configured loopback interfaces. One entry is has the protocol set to local (the local loopback on the PE), and the other entry has the protocol set to static, where the next hop is reached using an LDP LSP.

```
*A:PE-1# show router fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
172.16.1.1/32                                               LOCAL
    172.16.1.1 (loopback)
172.16.6.1/32                                               STATIC
    192.0.2.6 (Transport:LDP)
---snip---
```

The following output shows that a **ping** sourced by the loopback interface on PE-1 is able to reach the loopback interface on PE-6, and **traceroute** demonstrates that the traffic is following the LDP LSP. The **ping** and **traceroute** traffic cannot follow the IGP path because the static route command states that the IGP is disallowed when no LDP LSP toward PE-6 is available (also, the loopback interfaces are not enabled on IS-IS).

```
*A:PE-1# ping 172.16.6.1 source 172.16.1.1
PING 172.16.6.1 56 data bytes
64 bytes from 172.16.6.1: icmp_seq=1 ttl=64 time=2.03ms.
64 bytes from 172.16.6.1: icmp_seq=2 ttl=64 time=2.16ms.
64 bytes from 172.16.6.1: icmp_seq=3 ttl=64 time=2.01ms.
64 bytes from 172.16.6.1: icmp_seq=4 ttl=64 time=2.78ms.
64 bytes from 172.16.6.1: icmp_seq=5 ttl=64 time=3.18ms.

---- 172.16.6.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
```

```
round-trip min = 2.01ms, avg = 2.43ms, max = 3.18ms, stddev = 0.466ms


*A:PE-1# traceroute 172.16.6.1 source 172.16.1.1
traceroute to 172.16.6.1 from 172.16.1.1, 30 hops max, 40 byte packets
  1  0.0.0.0  * * *
  2  0.0.0.0  * * *
  3  172.16.6.1 (172.16.6.1)    1.69 ms  3.24 ms  2.46 ms
```

With the **traceroute** command, there are three hops from PE-1 to PE-6. There is no information regarding IP for the first two hops because the traffic is encapsulated in an MPLS LSP. The reason why the hops are displayed even when there is an MPLS LSP tunnel is because by default, the SR router propagates (copies) the Time To Live (TTL) from the IP header in the MPLS header. This is known as uniform mode.

However, a service provider might not want to show how many MPLS hops (nodes) there are in their network if a **traceroute** command is executed from outside their network. To prevent internal hops being shown, **no propagate** commands are needed in the LDP configuration, as follows. This is known as pipe mode.

```
*A:PE-1# configure
    router
        ldp
            no shortcut-local-ttl-propagate
            no shortcut-transit-ttl-propagate
        exit
```

When TTL propagation is disabled, the hops are not displayed any longer when running the **traceroute** command.

```
*A:PE-1# traceroute 172.16.6.1 source 172.16.1.1
traceroute to 172.16.6.1 from 172.16.1.1, 30 hops max, 40 byte packets
  1  172.16.6.1 (172.16.6.1)    2.36 ms  2.24 ms  2.25 ms
```

For more information about uniform mode and pipe mode, see the Tunneling of ICMP Reply Packets over MPLS LSPs chapter.

RSVP Static Route (IP Tunneled in RSVP Tunnel)

Using RSVP LSP shortcuts for static route resolution enables forwarding of IPv4 packets over RSVP LSPs instead of using a regular IP next hop. In other words, the traffic to the resolved static routes is forwarded using an MPLS RSVP LSP rather than plain IP.

The configuration defines a static route pointing to a destination PE (remote loopback, which is an indirect next hop in the example), and explicitly indicates that it should use RSVP rather than IGP. Taking PE-6 and PE-1 as an example, two loopback interfaces are configured (172.16.X.1/32), where X = PE number, and a static route is defined according to the preceding explanation. The following shows the configuration on PE-6.

```
*A:PE-6# configure
    router
        interface "loopback"
            address 172.16.6.1/32
            loopback
            no shutdown
        exit
        static-route-entry 172.16.1.1/32
            indirect 192.0.2.1
                tunnel-next-hop
                    resolution-filter
                        rsvp-te
                        exit
                    exit
                    disallow-igp
                    resolution filter
                exit
                no shutdown
            exit
        exit
```

Also, an RSVP LSP needs to be configured with the system interface of PE-1 as the destination:

```
*A:PE-6# configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-6-PE-1"
                to 192.0.2.1
                primary "loose"
                exit
                no shutdown
            exit
```

In the LSP tunnel table, an RSVP LSP is created:

```
*A:PE-6# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  1         7        192.168.46.1   30
```

```
192.0.2.1/32      ldp       MPLS  65553     9        192.168.46.1   30
192.0.2.2/32      ldp       MPLS  65552     9        192.168.46.1   20
192.0.2.3/32      ldp       MPLS  65540     9        192.168.56.1   20
192.0.2.4/32      ldp       MPLS  65551     9        192.168.46.1   10
192.0.2.5/32      ldp       MPLS  65541     9        192.168.56.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-6#
```

The default RSVP preference is 7 (preferred over that of LDP, which is 9) and the metric reflects that this LSP spans 3 hops (for a dynamic LSP not using constrained shortest path first (CSPF), the metric is inherited from IGP). See the RSVP Shortcut for IGP Route Resolution section for more details about the metric applied in LSPs.

The RSVP LSP is used to resolve the indirect next hop (PE-1 system address) in the static route (the LSP used is identified with the tunnel ID, in this case 1), therefore, the route for prefix 172.16.1.1 in the GRT looks as follows:

```
*A:PE-6# show router route-table 172.16.1.1

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.1/32                                 Remote  Static    00h00m21s  5
       192.0.2.1 (tunneled:RSVP:1)                                  1
-------------------------------------------------------------------------------
No. of Routes: 1
```

As in the LDP shortcut with static route example, between PE-6 and PE-1, TTL propagation is disabled, as follows.

```
*A:PE-6# configure
    router
        mpls
            no shortcut-local-ttl-propagate
            no shortcut-transit-ttl-propagate
        exit
```

The output is the following when running a traceroute:

```
*A:PE-6# traceroute 172.16.1.1 source 172.16.6.1
traceroute to 172.16.1.1 from 172.16.6.1, 30 hops max, 40 byte packets
  1  172.16.1.1 (172.16.1.1)    2.22 ms  2.04 ms  2.18 ms
```

The two static routes that have been defined to use the LDP and RSVP shortcuts follow the static routes default values and have a preference of 5 and a metric of 1.

LDP Shortcut for IGP Route Resolution

Using LDP shortcuts for IGP route resolution enables forwarding of packets to IGP learned routes over an LDP LSP. The default is to disable the LDP shortcut across all interfaces in the node.

When LDP shortcuts are enabled, LDP populates the Route Table Manager (RTM) with next hop entries corresponding to all prefixes for which it activated an LDP Forwarding Equivalence Class (FEC). For a prefix, two route entries are populated in RTM. One corresponds to the LDP shortcut next hop and has an owner of LDP. The other one is the regular IP next hop. The LDP shortcut next hop always takes preference over the regular IP next hop for forwarding user packets and specific control packets over an outgoing interface to the route next hop.

When LDP has activated a FEC for a prefix and programmed the RTM, it also programs the ingress tunnel table in the line card with the LDP tunnel information.

When an IPv4 packet is received on an ingress network interface, a subscriber Internet Enhanced Service (IES) interface, or a regular IES interface, the lookup of the packet by the ingress line card results in the packet being sent labeled with the label stack corresponding to the Next Hop Label Forwarding Entry (NHLFE) of the LDP LSP when the preferred RTM entry corresponds to an LDP shortcut. If the preferred RTM entry corresponds to an IP next hop, the IPv4 packet is forwarded unlabeled. The activation of the FEC by LDP is done by performing an exact match with an IGP route prefix in RTM, but it can also be done by performing a longest prefix match with an IGP route in RTM if the aggregate-prefix-match option is enabled globally in LDP.

Handling of Control Packets

All control plane packets will not see the LDP shortcut route entry in RTM with the exception of the following control packets which will be forwarded over an LDP shortcut when enabled:

- A locally generated or in transit ICMP ping and UDP traceroute of an IGP route. The transit message appears as a user packet to the ingress LER node.
- A locally generated response to a received ICMP ping or UDP traceroute message.

All other control plane packets that require an RTM lookup and have knowledge of which destination is reachable over the LDP shortcut will continue to be forwarded over the IP next hop route in RTM.

Handling of Multicast Packets

LDP shortcuts apply to unicast FEC types and are used for forwarding IP unicast packets in the data path. IP multicast packets forwarded over an multicast Label Distribution Protocol (mLDP) Point-to-Multi-Point (P2MP) LSP make use of a multicast FEC and thus cannot make use of the LDP unicast shortcut.

ECMP Considerations

When Equal Cost Multi-Path (ECMP) is enabled and multiple equal cost next hops exist for the IGP route, the ingress line card will spray the packets for this route based on the hashing routine supported for IPv4 packets. When the preferred RTM entry corresponds to an LDP shortcut route, spraying is performed across the multiple next hops for the LDP FEC. The FEC next hops can either be direct link LDP neighbors, or T-LDP (targeted LDP) neighbors reachable over RSVP LSPs in the case of LDP-over-RSVP, but not both. This is as per ECMP for LDP in the existing implementation. When the preferred RTM entry corresponds to a regular IP route, spraying will be performed across regular IP next hops for the prefix. Spraying across regular IP next hops and LDP shortcut next hops concurrently is not supported.

Configuring IGP LDP shortcuts is straightforward, and only applies to the node where there is interest to provision the LDP shortcut. In this example, only PE-1 is provisioned with LDP shortcuts, as follows:

```
*A:PE-1#configure router ldp-shortcut
```

Now, all tunnel LSPs that resolve an IGP next hop will replace the IP next hops, as shown in the following output:

```
*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     01h31m15s  0
       system                                                       0
192.0.2.2/32                                  Remote  LDP       00h04m15s  9
       192.168.12.2 (tunneled)                                     10
192.0.2.3/32                                  Remote  LDP       00h04m15s  9
       192.168.13.2 (tunneled)                                     10
192.0.2.4/32                                  Remote  LDP       00h04m15s  9
       192.168.12.2 (tunneled)                                     20
192.0.2.5/32                                  Remote  LDP       00h04m15s  9
       192.168.13.2 (tunneled)                                     20
192.0.2.6/32                                  Remote  LDP       00h04m15s  9
```

```
       192.168.12.2 (tunneled)                                      30
192.168.12.0/30                               Local   Local     01h31m15s  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     01h31m15s  0
       int-PE-1-PE-3                                                0
192.168.23.0/30                               Remote  ISIS      01h29m45s  15
       192.168.12.2                                                 20
192.168.24.0/30                               Remote  ISIS      01h29m45s  15
       192.168.12.2                                                 20
192.168.35.0/30                               Remote  ISIS      01h29m37s  15
       192.168.13.2                                                 20
192.168.45.0/30                               Remote  ISIS      01h29m22s  15
       192.168.12.2                                                 30
192.168.46.0/30                               Remote  ISIS      01h29m22s  15
       192.168.12.2                                                 30
192.168.56.0/30                               Remote  ISIS      01h29m16s  15
       192.168.13.2                                                 30
-------------------------------------------------------------------------------
No. of Routes: 14
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#


*A:PE-1# show router fib 1
===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                                LOCAL
    192.0.2.1 (system)
192.0.2.2/32                                                LDP
    192.0.2.2 (Transport:LDP)
192.0.2.3/32                                                LDP
    192.0.2.3 (Transport:LDP)
192.0.2.4/32                                                LDP
    192.0.2.4 (Transport:LDP)
192.0.2.5/32                                                LDP
    192.0.2.5 (Transport:LDP)
192.0.2.6/32                                                LDP
    192.0.2.6 (Transport:LDP)
192.168.12.0/30                                             LOCAL
    192.168.12.0 (int-PE-1-PE-2)
192.168.13.0/30                                             LOCAL
    192.168.13.0 (int-PE-1-PE-3)
192.168.23.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.168.24.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.168.35.0/30                                             ISIS
    192.168.13.2 (int-PE-1-PE-3)
192.168.45.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.168.46.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
```

```
192.168.56.0/30                                             ISIS
    192.168.13.2 (int-PE-1-PE-3)
-------------------------------------------------------------------------------
Total Entries : 14
-------------------------------------------------------------------------------
```

Applying LDP IGP shortcuts only on PE-1 implies that IP traffic from PE-1 to any of the system addresses of the rest of nodes will use the LDP shortcut, however, the traffic replied from any PE back to PE-1 will be native IP because IGP shortcuts have not been provisioned in the other nodes.

RSVP Shortcut for IGP Route Resolution

Using RSVP LSP shortcuts when resolving IGP routes enables forwarding of packets to IGP learned routes over an RSVP LSP. The use of RSVP shortcut for resolving IGP routes is enabled at the IS-IS (or OSPF) routing protocol level or at the LSP level, and instructs IS-IS and OSPF to include RSVP LSPs originating on this node and terminating on the system address (router ID) of a remote node and considers them as direct links. RSVP LSPs with a destination address corresponding to an interface address or any other loopback interface address of a remote node are automatically not considered by IS-IS or OSPF.

By default, **rsvp-shortcut** is disabled in all IGP instances.

RSVP LSPs are included in the IGP SPF computation with the following characteristics:

- RSVP LSP is modeled as a point-to-point link IP interface and its metric is used in the computation of the shortest path of IGP routes
- Next hop and interface include the NHLFE of the shortcut LSP when the IGP path cost using the RSVP LSP is the best.
- Shortcuts are not used when the destination RSVP LSP is in a different IGP area. In addition, IGP adjacencies across an RSVP LSP are not supported.

RSVP shortcut is enabled at IGP instance level as follows:

```
configure
    router
        isis
            igp-shortcut
                tunnel-next-hop
                    family ipv4
                        resolution filter
                        resolution-filter
                            rsvp
                        exit
                    exit
                exit
                no shutdown
            exit
```

The configuration can be done at the IGP level or per LSP level. When **rsvp-shortcut** is enabled at the IGP instance level, all RSVP LSPs originating on this node are eligible by default. The user can, however, exclude a specific RSVP LSP from being used as a shortcut for resolving IGP routes by entering the command

```
*A:PE-6# configure router mpls lsp "LSP-PE-6-PE-1" no igp-shortcut
```

As RSVP shortcuts can coexist with LDP shortcuts or IP next hops, SPF computation and path selection follows the procedures in RFC 3906:

- SPF picks the RSVP shortcut next hop if there is an RSVP LSP directly to that address regardless of the path cost compared to the IGP next hop.
- SPF picks the RSVP shortcut next hop or the IGP next hop based on path lowest cost if there is an IGP path to the prefix that does not go via the tail-end of the LSP.
- If the IGP next hop is picked, then it can be an LDP shortcut next hop or a regular IP next hop. The LDP shortcut next hop always has preference over the regular IP next hop.

Handling of Control Packets

All control plane packets requiring an RTM lookup and whose destination is reachable over the RSVP shortcut are forwarded over the shortcut. This is because RTM keeps a single route entry for each prefix, except if there is ECMP over different outgoing interfaces. Interface bound control packets are not impacted by the RSVP shortcut because RSVP LSPs with a destination address different than the router ID are not included by IGP in its SPF calculation.

RSVP shortcuts for IGP shortcut resolution should only be used with CSPF LSPs or with fully explicit path non-CSPF LSPs. RSVP hop-by-hop Path messages will try to use the shortcut and consequently LSPs without CSPF enabled, or that use a loose/empty hop path, will not come up. However, LSPs with CSPF enabled or using a strict hop path will come up. This is because in the former case, the RTM lookup to get the next hop results in using the shortcut and so the path messages are sent directly to the destination of the LSP, where they are dropped. With CSPF enabled, the next hop (and the entire path) is provided by CSPF and the path messages are sent unlabeled to the directly connected neighbor which corresponds to the next hop of the destination of the LSP. Similar processing occurs if a strict hop path is used, as is the case in the following example.

Handling of Multicast Packets

IP multicast packets cannot be forwarded over an RSVP shortcut, they can only be forwarded over an RSVP P2MP LSP. However, RSVP shortcut routes appear in RTM and are seen by all applications when they are the best route. When the Reverse Path Forwarding (RPF) check for the source of the multicast packet matches an RSVP shortcut route, the check will pass if both the RSVP shortcut and the multicast-import options are enabled in the IGP, as follows, because the RTM is populated with next hops only and not with tunnels (RPFs will fail for source prefixes resolved to a tunnel NH).

```
*A:PE-6# configure router isis multicast-import
  - multicast-import [both]
  - multicast-import [ipv4]
  - multicast-import [ipv6]
  - no multicast-import [both]
  - no multicast-import [ipv4]
  - no multicast-import [ipv6]
```

The unicast RTM can still use the tunnel next hop for the same prefix. SPF keeps track of both the direct first hop and the tunneled first hop of a node which is added to the Dijkstra tree.

ECMP Considerations

When ECMP is enabled and multiple equal cost paths exist for the route over a set of tunnel next hops based on the hashing routine supported for IPv4 packets, there are two possibilities:

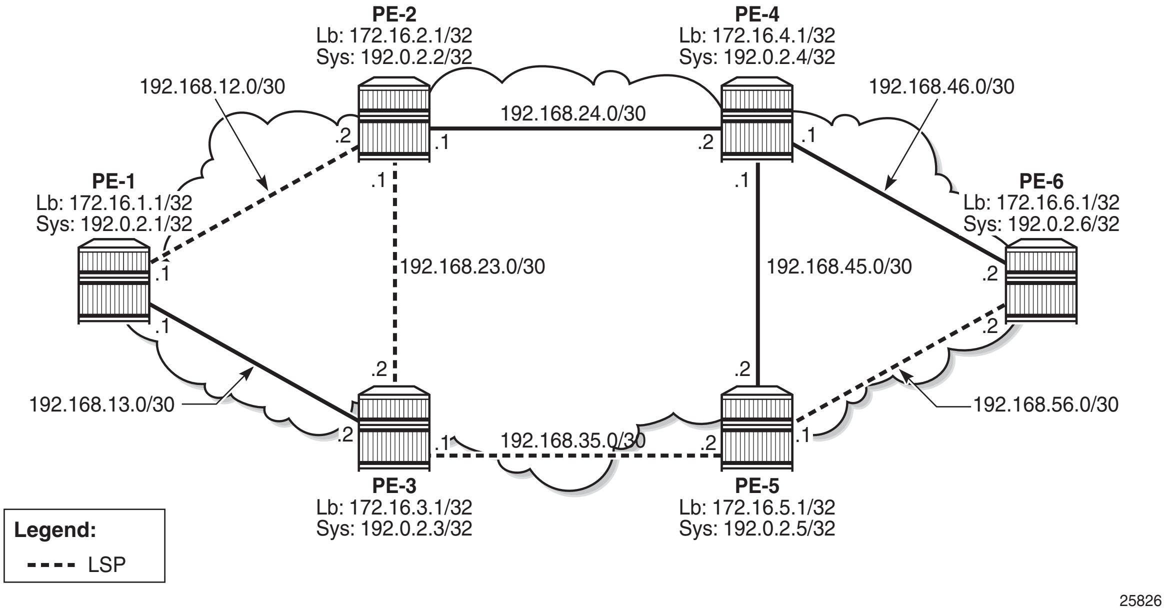
- Destination is tunnel endpoint: the system selects the tunnel with lowest tunnel ID (IP next hop is never used).
- Destination is different from the tunnel endpoint: it selects tunnel endpoints when the LSP metric is not greater than the IGP cost and it prefers tunnel endpoint over IP next hop.

ECMP is not performed across the IP and tunnel next hops simultaneously.

RSVP Shortcuts Configuration

Configuring RSVP LSP shortcuts is straightforward, and only applies to the node where there is interest to provision the RSVP shortcut. Two LSPs, from PE-6 to PE-1 and from PE-1 to PE-6, with strict hops, are provisioned according to Figure 229.

Figure 229 **LSPs Between PE-1 and PE-6**



The configuration on PE-1 and PE-6 is similar (replacing the IP addresses), so only the configuration for PE-6 is shown:

```
*A:PE-6#
configure
    router
        isis
            igp-shortcut
                tunnel-next-hop
                    family ipv4
                        resolution filter
                        resolution-filter
                            rsvp
                        exit
                    exit
                exit
                no shutdown
            exit



configure
    router
        mpls
            path "path-to-PE-1"
                hop 10 192.0.2.5 strict
                hop 20 192.0.2.3 strict
                hop 30 192.0.2.2 strict
                hop 40 192.0.2.1 strict
                no shutdown
            exit
            lsp "LSP-PE-6-PE-1-strict"
```

```
                to 192.0.2.1
                primary "path-to-PE-1"
                exit
                no shutdown
            exit
```

The GRT output shows the change in the next hop, using an RSVP shortcut:

```
*A:PE-6# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h00m34s  15
       192.0.2.1 (tunneled:RSVP:2)                                  16777215
192.0.2.2/32                                  Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 20
192.0.2.3/32                                  Remote  ISIS      00h00m34s  15
       192.168.56.1                                                 20
192.0.2.4/32                                  Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 10
192.0.2.5/32                                  Remote  ISIS      00h00m34s  15
       192.168.56.1                                                 10
192.0.2.6/32                                  Local   Local     01h45m01s  0
       system                                                       0
192.168.12.0/30                               Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 30
192.168.13.0/30                               Remote  ISIS      00h00m34s  15
       192.168.56.1                                                 30
192.168.23.0/30                               Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 30
192.168.24.0/30                               Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 20
192.168.35.0/30                               Remote  ISIS      00h00m34s  15
       192.168.56.1                                                 20
192.168.45.0/30                               Remote  ISIS      00h00m34s  15
       192.168.46.1                                                 20
192.168.46.0/30                               Local   Local     01h45m02s  0
       int-PE-6-PE-4                                                0
192.168.56.0/30                               Local   Local     01h45m01s  0
       int-PE-6-PE-5                                                0
-------------------------------------------------------------------------------
No. of Routes: 14
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-6#
```

The RSVP LSP in the output has a metric of 16777215, the LSP administrative metric matches the maximum value allowed for an IS-IS link using the wide-metric (24-bit value with a range of [0 — 16777215]). The following metric rules apply:

- A dynamic strict path non-CSPF LSP has the maximum metric (16777215).
- A dynamic CSPF LSP has a metric equal to the cumulative IGP cost.
  - If the user enabled the use of the TE metric on this LSP (**configure router mpls lsp cspf use-te-metric**), then the metric for the LSP is the maximum (16777215).
  - If the user enabled the use of the TE metric on this LSP (**configure router mpls lsp cspf use-te-metric**) and provisioned a specific metric on the lsp (configure router mpls lsp metric <metric>:<0..16777215>), then the metric for the LSP is the one provisioned. When configuring the metric of an LSP, the parameter “use-te-metric” is not required.
- A static LSP has a maximum metric (16777215).
- Manual and dynamic bypass LSPs have the maximum metric (16777215).

The RSVP shortcuts section detailed the importance of the LSP metric when using CSPF LSPs or when importing RSVP tunnel links into the IGP. The LSP metric can be inherited from the IGP, or can be manually modified by configuring a specific LSP metric or relative metric offset. Because IP and LDP FECs resolve to RSVP LSPs when the metric is equal or lower compared to the regular routing metric, configuring a specific static LSP metric (lower than the IGP metric) or relative metric offset is strongly recommended when using RSVP shortcuts, so that the GRT and LDP FEC resolution will always prefer the RSVP LSP shortcuts when the CSPF path computation is not using the shortest path.

For the preceding example, the first rule applies.

Advertising RSVP LSP Tunnel Links in the IGP: Forwarding Adjacency Feature

If configured, an RSVP LSP can also be advertised into the IGP similar to regular links so that other routers in the network can include that RSVP LSP into their SPF computations. The forwarding adjacency feature can be enabled independently from the RSVP shortcut feature in CLI. If both are configured for an IGP instance, the forwarding adjacency takes precedence. An RSVP LSP must exist in the reverse direction in order for the advertised link to pass the bi-directional link check and be usable by other routers in the network. However, this is not required for the node which originates the LSP. The LSP is advertised as an unnumbered point-to-point link and the Link State Protocol data unit (LSP) and Link State Advertisement (LSA) have no traffic engineering opaque sub-TLVs as per RFC 3906.

Reusing the RSVP IGP shortcuts set up previously (PE-1 and PE-6 RSVP IGP shortcut example according to Figure 229), the outcome is a route linked with an RSVP LSP as next hop, as follows:

```
*A:PE-6# show router route-table 192.0.2.1/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h02m37s  15
       192.0.2.1 (tunneled:RSVP:2)                                16777215
-------------------------------------------------------------------------------
No. of Routes: 1
```

The route tunneled through RSVP has a metric of 16777215, so it is not used by PE-6 GRT to reach any other routes because the metric is very high. After enabling the forwarding adjacency feature (tunnel links) to use shortcuts in the configuration, PE-1 and PE-6 have a direct connection through the RSVP LSP (as a virtual link). This configuration command must be executed in both routers.

```
configure router isis advertise-tunnel-link
```

When the shortcut is advertised by IS-IS, the route will disappear from the RTM because the metric of the shortcut is greater than the IGP cost.

```
*A:PE-6# show router route-table 192.0.2.1/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h00m04s  15
       192.168.46.1                                               30
-------------------------------------------------------------------------------
No. of Routes: 1
```

If the LSP is reconfigured to use a metric equal to or smaller than the IGP cost, the router PE-6 will use the RSVP shortcut again. In the example, the LSP is reconfigured with a metric of 30:

```
*A:PE-6# configure router mpls lsp "LSP-PE-6-PE-1-strict" metric 30
```

Now the shortcut shows up as the preferred next hop to reach PE-1 from PE-6.

```
*A:PE-6# show router route-table 192.0.2.1/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
```

```
192.0.2.1/32                                   Remote  ISIS      00h00m06s  15
       192.0.2.1 (tunneled:RSVP:2)                                   30
--------------------------------------------------------------------------------
No. of Routes: 1
```

As explained earlier, this could be combined with ECMP, so if ECMP is configured to 2, the system shows the two equal cost paths.

```
*A:PE-6# configure router ecmp 2

*A:PE-6# show router route-table 192.0.2.1/32
================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Remote  ISIS      00h00m05s  15
       192.0.2.1 (tunneled:RSVP:2)                                   30
192.0.2.1/32                                   Remote  ISIS      00h00m05s  15
       192.168.46.1                                                  30
--------------------------------------------------------------------------------
No. of Routes: 2
```

The GRT on PE-4 displays the route to reach PE-1 (192.0.2.1/32) with a metric of 20 via PE-2 as next hop. Although PE-6 is announcing the RSVP LSP-PE-6-PE-1 to the other routers, the LSP shortcut is not used by PE-4, because the metric to reach PE-6 (10) plus the metric of the LSP shortcut from PE-6 to PE-1 (metric 30) is greater than 20.

```
*A:PE-4# show router route-table

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  20
192.0.2.2/32                                   Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  10
192.0.2.3/32                                   Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  20
192.0.2.4/32                                   Local   Local     01h58m26s  0
       system                                                        0
192.0.2.5/32                                   Remote  ISIS      01h57m08s  15
       192.168.45.2                                                  10
192.0.2.6/32                                   Remote  ISIS      01h57m01s  15
       192.168.46.2                                                  10
192.168.12.0/30                                Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  20
192.168.13.0/30                                Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  30
192.168.23.0/30                                Remote  ISIS      01h57m14s  15
       192.168.24.1                                                  20
```

```
192.168.24.0/30                               Local   Local     01h58m26s  0
       int-PE-4-PE-2                                                0
192.168.35.0/30                               Remote  ISIS      01h57m08s  15
       192.168.45.2                                                 20
192.168.45.0/30                               Local   Local     01h58m26s  0
       int-PE-4-PE-5                                                0
192.168.46.0/30                               Local   Local     01h58m26s  0
       int-PE-4-PE-6                                                0
192.168.56.0/30                               Remote  ISIS      01h57m08s  15
       192.168.45.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 14
```

If the metric of the LSP LSP-PE-6-PE-1 is modified to a value between 1 and 9, there is a better metric (less than 20) so that PE-4 will change the next hop via PE-6. First the metric of the LSP is modified to 9:

```
*A:PE-6# configure router mpls lsp "LSP-PE-6-PE-1-strict" metric 9
```

The GRT on PE-4 shows that the next hop to reach PE-1 has changed, from next hop PE-2 to next hop PE-6 (therefore, using the LSP shortcut), and the metric is 19 (10 to reach PE-6 plus metric 9 of the LSP PE-6-PE-1 shortcut):

```
*A:PE-4# show router route-table 192.0.2.1/32

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h00m07s  15
       192.168.46.2                                                 19
-------------------------------------------------------------------------------
No. of Routes: 1
```

Because the metric of the LSP shortcut was modified to a value of 9, the GRT of PE-6 shows that the next hops of several routes have changed and are also using the shortcut LSP PE-6-PE-1 because the metric is better than the regular IS-IS metric. IGP shortcuts will not be used to resolve prefixes downstream of the LSP endpoint when the LSP metric is higher than the underlying IGP cumulative metric.

```
*A:PE-6# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                  9
192.0.2.2/32                                  Remote  ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                  19
```

```
192.0.2.3/32                                        Remote   ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                           19
192.0.2.4/32                                        Remote   ISIS      00h11m25s  15
       192.168.46.1                                                          10
192.0.2.5/32                                        Remote   ISIS      00h11m25s  15
       192.168.56.1                                                          10
192.0.2.6/32                                        Local    Local     02h01m04s  0
       system                                                                0
192.168.12.0/30                                     Remote   ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                           19
192.168.13.0/30                                     Remote   ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                           19
192.168.23.0/30                                     Remote   ISIS      00h01m15s  15
       192.0.2.1 (tunneled:RSVP:2)                                           29
192.168.24.0/30                                     Remote   ISIS      00h11m25s  15
       192.168.46.1                                                          20
192.168.35.0/30                                     Remote   ISIS      00h11m25s  15
       192.168.56.1                                                          20
192.168.45.0/30                                     Remote   ISIS      00h10m34s  15
       192.168.46.1                                                          20
192.168.46.0/30                                     Local    Local     02h01m04s  0
       int-PE-6-PE-4                                                         0
192.168.56.0/30                                     Local    Local     02h01m04s  0
       int-PE-6-PE-5                                                         0
-------------------------------------------------------------------------------
No. of Routes: 14
```

There are also cases where an LDP FEC can resolve to an RSVP LSP, if the user enables the LDP-over-RSVP feature or IGP shortcut feature when **prefer-tunnel-in-tunnel** is enabled in LDP and the endpoint of the RSVP LSP matches the FEC prefix. For those cases, the metric to the prefix is the sum of the RSVP LSP metric and the remaining IGP path cost.

Table 23 provides a summary of the outcome when configuring the forwarding adjacency, LDPoRSVP and RSVP shortcut options at both the IGP instance level and at the LSP level.

Table 23 **RSVP LSP Role As Outcome of LSP Level and IGP Level Configuration Options**

	IGP Instance Level Configurations					
LSP Level Configuration	advertise-tunnel-link enabled/ rsvp-shortcut enabled/ldp-over-rsvp enabled	advertise-tunnel-link enabled/ rsvp-shortcut enabled/ldp-over-rsvp disabled	advertise-tunnel-link enabled/ rsvp-shortcut disabled/ldp-over-rsvp disabled	advertise-tunnel-link disabled/ rsvp-shortcut disabled/ldp-over-rsvp disabled	advertise-tunnel-link disabled/ rsvp-shortcut enabled/ldp-over-rsvp enabled	advertise-tunnel-link disabled/ rsvp-shortcut disabled/ldp-over-rsvp enabled

Table 23 RSVP LSP Role As Outcome of LSP Level and IGP Level Configuration Options

	IGP Instance Level Configurations					
igp-shortcut enabled/ldp-over-rsvp enabled	Forwarding Adjacency	Forwarding Adjacency	Forwarding Adjacency	None	IGP Shortcut	LDP-over-RSVP
igp-shortcut enabled/ldp-over-rsvp disabled	Forwarding Adjacency	Forwarding Adjacency	Forwarding Adjacency	None	IGP Shortcut	None
igp-shortcut disabled/ldp-over-rsvp enabled	None	None	None	None	None	LDP-over-RSVP
igp-shortcut disabled/ldp-over-rsvp disabled	None	None	None	None	None	None

LSP Relative Metric

It is possible to use relative metrics for IGP shortcuts as per RFC 3906, *Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels*, with the following command:

```
*A:PE-6# configure router mpls lsp "LSP-PE-6-PE-1-strict" igp-shortcut relative-metric
  - igp-shortcut [lfa-protect | lfa-only]
  - igp-shortcut relative-metric [offset]
  - no igp-shortcut

 <lfa-protect>         : keyword
 <lfa-only>            : keyword
 <relative-metric>     : keyword
 <offset>              : [-10..10]
```

When this feature is enabled, IGP applies the shortest IGP cost between the endpoints of the LSP, plus the value of a configured offset when computing the cost of the prefix that is resolved to the LSP.

The offset value is optional and can have a value between -10 and 10, and defaults to zero (0). An offset value of zero (0) is used when the relative metric option is enabled without specifying the offset parameter value. The minimum net cost for the prefix is capped to the value of one (1) after applying the offset:

Prefix cost = max (1, IGP Cost + relative metric offset)

The **relative-metric** option is ignored when **advertise-tunnel-link** is enabled in IS-IS or OSPF. In that case, the IGP advertises the LSP as a P2P unnumbered link using the LSP operational metric.

The **relative-metric** option is mutually exclusive with the **lfa-protect** (Loop-Free Alternate (LFA)) or the **lfa-only** options. An LSP with **relative-metric** option enabled cannot be included in the LFA SPF and vice versa when the **rsvp-shortcut** option is enabled in the IGP (see chapter LDP/IP FRR LFA for IGP Shortcut Using IS-IS/ OSPF for more information).

The offset can be used to enforce the preference of the shortcut path over the other paths for the prefix. Using an example, a new CSPF LSP with empty path and relative metric of -10 is created between PE-6 and PE-1. Whereas the operational or absolute metric is 30 (IGP cost and populated in the Tunnel Table Manager, TTM), the metric that the RTM shows is 20 after applying the offset:

```
*A:PE-6# configure
    router
        mpls
            lsp "LSP-PE-6-PE-1-loose"
                to 192.0.2.1
                cspf
                igp-shortcut relative-metric -10
                primary "loose"
                exit
                no shutdown
            exit
```

```
*A:PE-6# show router tunnel-table 192.0.2.1

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  2         7        192.168.56.1   16777215
192.0.2.1/32      rsvp      MPLS  3         7        192.168.56.1   30
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-6#
```

```
*A:PE-6# show router route-table 192.0.2.1

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h00m07s  15
       192.0.2.1 (tunneled:RSVP:3)                                20
```

```
-------------------------------------------------------------------------------
No. of Routes: 1
```

LDP/IP FRR LFA for IGP Shortcut Using IS-IS/OSPF

MPLS LDP/IP FRR LFA for IGP shortcuts allows for the use of RSVP LSP-based IGP shortcuts as Loop-Free Alternate (LFA) backups, this way expanding the coverage of the IP Fast-Reroute (FRR) and the LDP FRR capabilities for IS-IS and OSPF prefixes. For a detailed description about IP and LDP FRR, see chapter MPLS LDP FRR using ISIS as IGP.

When an RSVP LSP is used as a shortcut by IS-IS or OSPF, it is included by the SPF as a P2P link and it can also be optionally advertised into the rest of the network by the IGP.

Two LSP-level configuration options are provided:

- The **lfa-protect** option includes the RSVP LSP in both the main SPF and the LFA SPFs. If the prefix primary next hop (NH) is tunneled, no LFA NH is computed. The protection in this case is provided by RSVP FRR. If the prefix primary NH is direct, then an LFA NH is computed. A direct LFA NH is preferred over a tunneled LFA NH. Within each LFA NH type, node protection is preferred over link protection. The configuration command is:

```
configure router mpls lsp <lsp-name> igp-shortcut lfa-protect
```

- The **lfa-only** option includes the LSP in the LFA SPFs only so that the introduction of IGP shortcuts does not impact the main SPF decision. The prefix primary NH is always direct and the prefix LFA NH is computed. A direct LFA NH is preferred over a tunneled LFA NH. Within each LFA NH type, node protection is preferred over link protection. The configuration command is:

```
configure router mpls lsp <lsp-name> igp-shortcut lfa-only
```

LDP/IP FRR is a local decision, so it can be enabled per node and there are no interoperability issues with other nodes. In the topology, PE-2 is provisioned with IS-IS LFA (OSPF configuration for the rest of this section is similar):

```
A:PE-2# configure router isis loopfree-alternate
```

The second item to configure is whether LDP or IP FRR is provisioned. To configure IP FRR, the command is:

```
A:PE-2# configure router ip-fast-reroute
```

To configure LDP FRR, the following command is used:

```
A:PE-2# configure router ldp fast-reroute
```

Although not shown, it is recommended to enable IGP-LDP synchronization per interface to avoid possible traffic black-holes.

LFA is enabled in all routers of the topology. The following command shows the LFA coverage on PE-2 where four nodes out of five are protected (80%) and seven of the ten prefixes are protected (70%). IPv4 prefixes are protected (IPv6 is not configured). The following output shows L1 and L2 because this node is provisioned as an L1-L2 IS-IS router.

```
*A:PE-2# show router isis lfa-coverage

===============================================================================
Router Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      4/5(80%)       7/10(70%)           0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      4/5(80%)       7/10(70%)           0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-2#
```

PE-2, PE-3, PE-4, and PE-5 share the same results, whereas only PE-1 and PE-6 have a coverage of 100% as shown in the following output.

```
*A:PE-1# show router isis lfa-coverage

===============================================================================
Router Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      5/5(100%)      11/11(100%)         0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      5/5(100%)      11/11(100%)         0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-1#
```

Taking a deeper look into the IS-IS LFA on PE-2, it can be seen that the node which is not protected is PE-4 (system address 192.0.2.4, because it is the one missing):

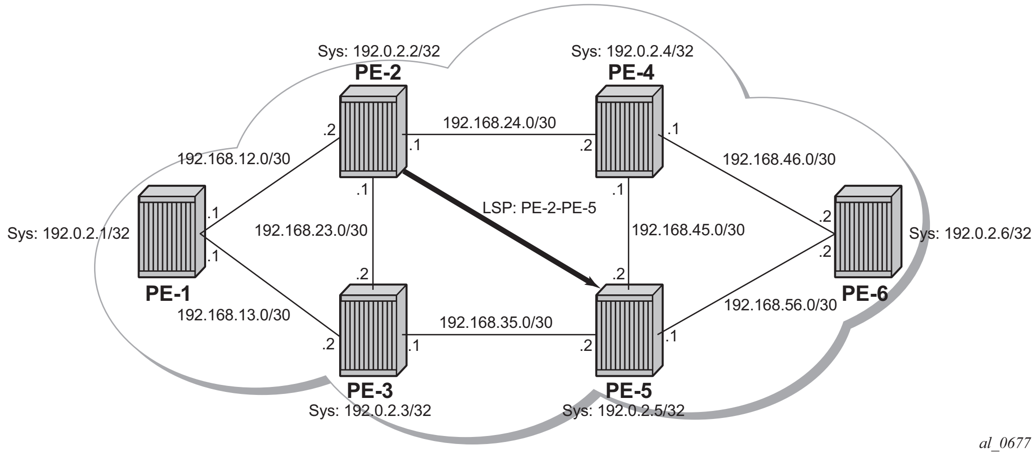
```
*A:PE-2# show router route-table alternative | match LFA pre-lines 2
```

```
192.0.2.1/32                                  Remote  ISIS      02h22m46s  15
       192.168.12.1                                                 10
       192.168.23.2 (LFA)                                           20
192.0.2.3/32                                  Remote  ISIS      02h22m38s  15
       192.168.23.2                                                 10
       192.168.12.1 (LFA)                                           20
192.0.2.5/32                                  Remote  ISIS      02h22m18s  15
       192.168.23.2                                                 20
       192.168.24.2 (LFA)                                           20
192.0.2.6/32                                  Remote  ISIS      02h22m08s  15
       192.168.24.2                                                 20
       192.168.23.2 (LFA)                                           30
192.168.13.0/30                               Remote  ISIS      02h22m46s  15
       192.168.12.1                                                 20
       192.168.23.2 (LFA)                                           30
192.168.35.0/30                               Remote  ISIS      02h22m38s  15
       192.168.23.2                                                 20
       192.168.12.1 (LFA)                                           30
192.168.56.0/30                               Remote  ISIS      02h22m17s  15
       192.168.23.2                                                 30
       192.168.24.2 (LFA)                                           30
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
*A:PE-2#
```

LFA is improved by taking advantage of RSVP shortcuts when it is properly provisioned. The reason why PE-4 cannot be protected with an LFA path is because the direct NH is using the direct link between PE-2 and PE-4 (the shortest IGP) and the intended LFA path through PE-3 is not valid (when LFA tries to find an alternate path via PE-3, the IGP cost from PE-3 to PE-4 is the same going via PE-5 as the path back via PE-2, invalidating that LFA calculation because there is a loop). This is normal because PE-2, PE-3, PE-4 and PE-5 are forming a ring. LFA coverage is increased by adding a link between PE-2 and PE-5, which can be done using a physical link or a virtual link with an RSVP shortcut. From the two possible options (**lfa-only** and **lfa-protect**), a new LSP “LSP-PE-2-PE-5” is configured with **igp-shortcut lfa-only**.

```
*A:PE-2# configure
    router
        mpls
            path "path-to-PE-5"
                hop 10 192.0.2.3 strict
                hop 20 192.0.2.5 strict
                no shutdown
            exit
            lsp "LSP-PE-2-PE-5"
                to 192.0.2.5
                igp-shortcut lfa-only
                primary "path-to-PE-5"
                exit
                no shutdown
            exit
```

Figure 230 **RSVP Shortcuts LFA Use Case Example**



Now the LFA coverage is 100% on PE-2 as shown by the following output:

```
*A:PE-2# show router isis lfa-coverage
===============================================================================
Router Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      5/5(100%)      10/10(100%)         0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      5/5(100%)      10/10(100%)         0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-2#
```

The GRT details the prefix information after the new LFA calculation using the **lfa-only** option (the shortcut is used by LFA SPF). The metric from PE-2 to PE-4 is the maximum plus the IGP cost (16777215 + 10) and the shortcut is also used to protect the rest of the previously unprotected prefixes:

```
*A:PE-2# show router route-table alternative | match LFA pre-lines 2
192.0.2.1/32                                  Remote  ISIS      02h27m59s  15
       192.168.12.1                                                 10
       192.168.23.2 (LFA)                                           20
192.0.2.3/32                                  Remote  ISIS      02h27m51s  15
       192.168.23.2                                                 10
       192.168.12.1 (LFA)                                           20
192.0.2.4/32                                  Remote  ISIS      02h27m36s  15
       192.168.24.2                                                 10
       192.0.2.5 (LFA) (tunneled:RSVP:1)                            16777225
```

```
192.0.2.5/32                                  Remote  ISIS      02h27m31s  15
       192.168.23.2                                                 20
       192.168.24.2 (LFA)                                           20
192.0.2.6/32                                  Remote  ISIS      02h27m21s  15
       192.168.24.2                                                 20
       192.168.23.2 (LFA)                                           30
192.168.13.0/30                               Remote  ISIS      02h27m59s  15
       192.168.12.1                                                 20
       192.168.23.2 (LFA)                                           30
192.168.35.0/30                               Remote  ISIS      02h27m51s  15
       192.168.23.2                                                 20
       192.168.12.1 (LFA)                                           30
192.168.45.0/30                               Remote  ISIS      02h27m36s  15
       192.168.24.2                                                 20
       192.0.2.5 (LFA) (tunneled:RSVP:1)                            16777235
192.168.46.0/30                               Remote  ISIS      02h27m36s  15
       192.168.24.2                                                 20
       192.0.2.5 (LFA) (tunneled:RSVP:1)                            16777235
192.168.56.0/30                               Remote  ISIS      02h27m30s  15
       192.168.23.2                                                 30
       192.168.24.2 (LFA)                                           30
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
*A:PE-2#
```

The tunnel table shows the RSVP LSP used as a shortcut and its operational metric.

```
*A:PE-2# show router tunnel-table 192.0.2.5

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.5/32      rsvp      MPLS  1         7        192.168.23.2   16777215
192.0.2.5/32      ldp       MPLS  65540     9        192.168.23.2   20
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

If the LSP “LSP-PE-2-PE-5” is provisioned with **lfa-protect** instead of **lfa-only**, the result is that the LSP "LSP-PE-2-PE-5“ is used by normal SPF to define the primary NH and it is not used by LFA SPF anymore.

```
*A:PE-2# configure router mpls lsp "LSP-PE-2-PE-5" igp-shortcut lfa-protect
```

The coverage when lfa-protect is used also shows a 100% for nodes and 100% for prefixes, as follows.

```
*A:PE-2# show router isis lfa-coverage

===============================================================================
Router Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      5/5(100%)      9/9(100%)           0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      5/5(100%)      9/9(100%)           0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-2#
```

In this case, the GRT looks as follows, the main difference being that now PE-5 (192.0.2.5) has a direct shortcut from PE-2:

```
*A:PE-2# show router route-table alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      02h32m17s  15
       192.168.12.1                                                 10
       192.168.23.2 (LFA)                                           20
192.0.2.2/32                                  Local   Local     02h33m31s  0
       system                                                       0
192.0.2.3/32                                  Remote  ISIS      02h32m09s  15
       192.168.23.2                                                 10
       192.168.12.1 (LFA)                                           20
192.0.2.4/32                                  Remote  ISIS      02h31m54s  15
       192.168.24.2                                                 10
       192.0.2.5 (LFA) (tunneled:RSVP:1)                            16777225
192.0.2.5/32                                  Remote  ISIS      00h01m56s  15
       192.0.2.5 (tunneled:RSVP:1)                                  16777215
192.0.2.6/32                                  Remote  ISIS      02h31m39s  15
       192.168.24.2                                                 20
       192.168.23.2 (LFA)                                           30
192.168.12.0/30                               Local   Local     02h33m31s  0
       int-PE-2-PE-1                                                0
192.168.13.0/30                               Remote  ISIS      02h32m17s  15
       192.168.12.1                                                 20
       192.168.23.2 (LFA)                                           30
192.168.23.0/30                               Local   Local     02h33m31s  0
       int-PE-2-PE-3                                                0
192.168.24.0/30                               Local   Local     02h33m31s  0
       int-PE-2-PE-4                                                0
192.168.35.0/30                               Remote  ISIS      02h32m09s  15
       192.168.23.2                                                 20
       192.168.12.1 (LFA)                                           30
192.168.45.0/30                               Remote  ISIS      02h31m54s  15
       192.168.24.2                                                 20
       192.0.2.5 (LFA) (tunneled:RSVP:1)                           16777235
192.168.46.0/30                               Remote  ISIS      02h31m54s  15
       192.168.24.2                                                20
       192.0.2.5 (LFA) (tunneled:RSVP:1)                           16777235
192.168.56.0/30                               Remote  ISIS      00h01m56s  15
       192.168.24.2                                                30
       192.168.23.2 (LFA)                                          40
-------------------------------------------------------------------------------
No. of Routes: 14
```

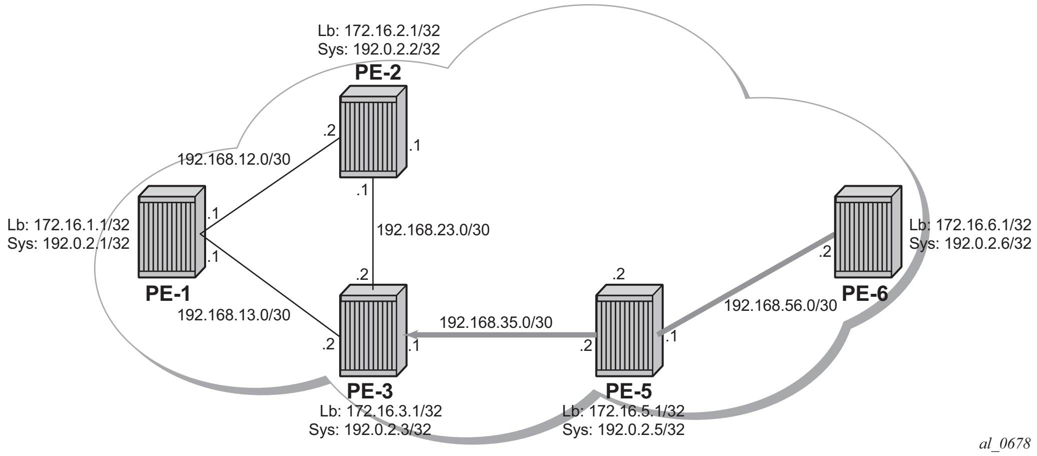
Rules Determining the Installation of Shortcuts into RTM

Although it was already mentioned in the RSVP-TE LSP shortcut for IGP route resolution section, the rules determining how shortcuts are installed into RTM are (sorted by higher priority):

- RSVP shortcut.
- LDP shortcut.
- IGP route with regular IP next hop.
- The implementation is compliant with RFC 3906.

To check the rules, the network configuration is iLDP in all interfaces with LDP shortcuts enabled, there is also an RSVP LSP from PE-6 to PE-3 available but RSVP shortcuts are disabled. The topology is shown in Figure 231.

Figure 231 **Network Topology to Verify Installation of Shortcuts into RTM**



The following RSVP LSP is needed between PE-6 and PE-3.

```
configure router ldp-shortcut
```

```
*A:PE-6#
configure
    router
        isis
            igp-shortcut
                shutdown
                tunnel-next-hop
                    family ipv4
                        resolution disabled
                        resolution-filter
                            no rsvp
                        exit
                    exit
                exit
            exit

configure
    router
        mpls
            path “loose”
                no shutdown
            exit
            lsp “LSP-PE-6-PE-3”
                to 192.0.2.3
                cspf
                primary “loose”
                exit
                no shutdown
            exit
```

The routes in the routing table on PE-6 are the following:

```
*A:PE-6# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  LDP       00h00m03s  9
       192.168.56.1 (tunneled)                                      30
192.0.2.2/32                                  Remote  LDP       00h00m03s  9
       192.168.56.1 (tunneled)                                      30
192.0.2.3/32                                  Remote  LDP       00h00m03s  9
       192.168.56.1 (tunneled)                                      20
192.0.2.5/32                                  Remote  LDP       00h20m48s  9
       192.168.56.1 (tunneled)                                      10
192.0.2.6/32                                  Local   Local     02h57m46s  0
       system                                                       0
192.168.12.0/30                               Remote  ISIS      00h00m03s  15
       192.168.56.1                                                 40
192.168.13.0/30                               Remote  ISIS      00h00m03s  15
       192.168.56.1                                                 30
192.168.23.0/30                               Remote  ISIS      00h00m03s  15
       192.168.56.1                                                 30
192.168.35.0/30                               Remote  ISIS      00h00m03s  15
       192.168.56.1                                                 20
```

```
192.168.56.0/30                               Local   Local     02h57m46s  0
       int-PE-6-PE-5                                                0
-------------------------------------------------------------------------------
No. of Routes: 10
```

The tunnel table shows the LSPs available for the shortcuts, and therefore, these are used in the GRT for LDP (but not for RSVP):

```
*A:PE-6# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65564     9        192.168.56.1   30
192.0.2.2/32      ldp       MPLS  65565     9        192.168.56.1   30
192.0.2.3/32      rsvp      MPLS  4         7        192.168.56.1   20
192.0.2.3/32      ldp       MPLS  65566     9        192.168.56.1   20
192.0.2.5/32      ldp       MPLS  65541     9        192.168.56.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-6#
```

So far, LDP shortcuts are preferred over the IGP next hops for the system addresses (router ID). After enabling RSVP shortcuts in the IS-IS context, the changes in the GRT are:

```
*A:PE-6# configure
    router
        isis
            igp-shortcut
                tunnel-next-hop
                    family ipv4
                        resolution filter
                        resolution-filter
                            rsvp
                        exit
                    exit
                exit
                no shutdown
            exit
```

```
*A:PE-6# show router route-table next-hop-type tunneled

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  30
```

```
192.0.2.2/32                                  Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  30
192.0.2.3/32                                  Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  20
192.0.2.5/32                                  Remote  LDP       00h23m42s  9
       192.168.56.1 (tunneled)                                      10
192.168.12.0/30                               Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  40
192.168.13.0/30                               Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  30
192.168.23.0/30                               Remote  ISIS      00h00m08s  15
       192.0.2.3 (tunneled:RSVP:4)                                  30
-------------------------------------------------------------------------------
No. of Routes: 7
```

The GRT shows that PE-6 is using an LDP shortcut to reach PE-5, but PE-6 is using the RSVP shortcut to reach not only PE-3’s system address, but also PE-1 and PE-2 routes (including all interfaces) which were behind the RSVP LSP shortcut.

In summary, the behavior is:

1. When resolving a prefix, SPF picks the RSVP shortcut next hop if there is an RSVP LSP directly to that address regardless of the IGP path cost compared to the IGP next hop. When multiple RSVP LSPs to that address exist and all have the same lowest metric, if ECMP is enabled on the system, the LSP with the lowest tunnel ID is chosen. In this example, if LSP “LSP-PE-6-PE-3” is provisioned with a metric of 100 (IGP metric is 20), the GRT shows that the PE-3 system address is reachable via the LSP.

```
*A:PE-6# show router route-table 192.0.2.3

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.3/32                                  Remote  ISIS      00h00m12s  15
       192.0.2.3 (tunneled:RSVP:4)                                  100
-------------------------------------------------------------------------------
No. of Routes: 1
```

2. SPF also picks the RSVP LSP shortcut if both the LSP path and the IGP path to the prefix are via the tail-end of the LSP. This is regardless of the path cost compared to the IGP next hop. When paths over multiple RSVP shortcuts have the same lowest cost, if ECMP is enabled on the system, the LSP with the lowest tunnel ID is chosen. In this example, 192.168.13.0 and 192.168.23.0 are using the shortcut but 192.168.12.0 is not.

```
*A:PE-6# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
---snip---

192.168.12.0/30                               Remote  ISIS      00h00m46s  15
       192.168.56.1                                                 40
192.168.13.0/30                               Remote  ISIS      00h00m46s  15
       192.0.2.3 (tunneled:RSVP:4)                                  110
192.168.23.0/30                               Remote  ISIS      00h00m46s  15
       192.0.2.3 (tunneled:RSVP:4)                                  110

---snip---
-------------------------------------------------------------------------------
No. of Routes: 10
```

LDP/RSVP LSP Shortcut for BGP NH Resolution

Using LDP/RSVP LSP shortcuts for resolving BGP next hops allows IPv4 packet forwarding to routes resolved via a BGP next hop using an LDP/RSVP LSP instead of using a regular IP next hop. In the network topology of Figure 228, both PE-3 and PE-6 have a single peer configured, initially without any shortcuts enabled under the BGP context. Also, one static route is configured in PE-3 and PE-6 and that is redistributed into BGP. The relevant configuration on PE-3 is the following:

```
*A:PE-3# configure
    router
        interface "static-route"
            address 172.16.33.1/30
            port 1/1/4:33
        exit
        autonomous-system 65536

        static-route-entry 10.10.10.0/24
            next hop 172.16.33.2
                no shutdown
            exit
        exit

        policy-options
            begin
            policy-statement "static-routes"
                description "export static-routes for I-BGP"
                entry 10
                    from
                        protocol static
                    exit
                    to
                        protocol bgp
                    exit
                    action accept
                        next-hop-self
                    exit
                exit
            exit
            commit
        exit

        bgp
            export "static-routes"
            group "ibgp"
                peer-as 65536
                neighbor 192.0.2.6
                exit
            exit
        exit
```

Checking the static route received on PE-6 via BGP, the next hop is the PE-3 system address:

```
*A:PE-6# show router bgp routes 10.10.10.0/24 detail

===============================================================================
 BGP Router ID:192.0.2.6        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : ISIS                   Res. Metric    : 20
Res. Nexthop   : 192.168.56.1
Local Pref.    : 100                    Interface Name : int-PE-6-PE-5
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  Incomplete
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : n/a
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h00m10s

Modified Attributes
```

```
Network         : 10.10.10.0/24
Nexthop         : 192.0.2.3
Path Id         : None
From            : 192.0.2.3
Res. Protocol   : ISIS                   Res. Metric    : 20
Res. Nexthop    : 192.168.56.1
---snip---
-------------------------------------------------------------------------------
Routes : 1
```

Three of the BGP peering configuration possibilities are LDP, RSVP, or BGP. The other resolution filter options are related to segment routing and are beyond the scope of this chapter. In case both LDP and RSVP are included in the filter, RSVP is preferred. Disabling the IGP is also allowed (meaning that unless there is a shortcut, the BGP peering will not fall back to IGP):

```
*A:PE-6# configure router bgp next-hop-resolution shortcut-tunnel family ipv4
resolution
  - resolution {any|filter|disabled}

*A:PE-6# configure router bgp next-hop-resolution shortcut-tunnel family ipv4
resolution-filter
  - resolution-filter

 [no] bgp             - Use BGP tunnelling for next hop resolution
 [no] ldp             - Use LDP tunnelling for next hop resolution
 [no] rsvp            - Use RSVP tunnelling for next hop resolution
 [no] sr-isis         - Use sr-isis for next hop resolution
 [no] sr-ospf         - Use sr-ospf for next hop resolution
 [no] sr-policy       - Use sr-policy for next hop resolution
 [no] sr-te           - Use sr-te for next hop resolution

*A:PE-6#
```

When enabling LDP shortcuts on PE-6, the output changes showing the detail of the received BGP route indicating that the next hop is resolved using LDP:

```
*A:PE-6# configure
    router
        bgp
            next-hop-resolution
                shortcut-tunnel
                    family ipv4
                        resolution-filter
                            ldp
                        exit
                        resolution filter
                    exit
                exit
            exit


*A:PE-6# show router bgp routes 10.10.10.0/24 detail

===============================================================================
 BGP Router ID:192.0.2.6        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : LDP                    Res. Metric    : 20
Res. Nexthop   : 192.168.56.1 (LDP)
Local Pref.    : 100                    Interface Name : int-PE-6-PE-5
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : None
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.3
Fwd Class      : None                   Priority       : None
Flags          : Used  Valid  Best  Incomplete
Route Source   : Internal
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : n/a
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0
Add Paths Send : Default
Last Modified  : 00h20m53s

Modified Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : LDP                    Res. Metric    : 20
Res. Nexthop   : 192.168.56.1 (LDP)
---snip---
-------------------------------------------------------------------------------
Routes : 1
```

The GRT output command also shows that the route is reachable using LDP (indicated as tunneled):

```
*A:PE-6# show router route-table next-hop-type tunneled

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
```

```
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.10.10.0/24                                 Remote  BGP       00h20m30s  170
       192.0.2.3 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
```

The previously created LSP LSP-PE-6-PE-3 is up and running:

```
*A:PE-6# show router mpls lsp "LSP-PE-6-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-6-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-6-PE-3 Path loose
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-6-PE-3
From             : 192.0.2.6           To                   : 192.0.2.3
Admin State      : Up                  Oper State           : Up
Path Name        : loose
Path LSP ID      : 8704                Path Type            : Primary
Path Admin       : Up                  Path Oper            : Up
Out Interface    : 1/1/1               Out Label            : 524281

---snip---

Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.56.2 (192.0.2.6)                     Record Label        : N/A
 -> 192.168.56.1 (192.0.2.5)                     Record Label        : 524281
 -> 192.168.35.1 (192.0.2.3)                     Record Label        : 524284
Computed Hops    :
    192.168.56.2(S)
 -> 192.168.56.1(S)
 -> 192.168.35.1(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 20
===============================================================================
*A:PE-6#
```

After adding **resolution-filter rsvp** to the shortcut-tunnel configuration in the bgp context, the output shows that the BGP peer is reachable using an RSVP LSP (switched from LDP to RSVP because RSVP is preferred):

```
*A:PE-6# configure
    router
        bgp
            next-hop-resolution
```

```
                shortcut-tunnel
                    family ipv4
                        resolution-filter
                            ldp
                            rsvp
                        exit
                        resolution filter
                    exit
                exit
            exit
```

```
*A:PE-6# show router bgp routes 10.10.10.0/24 detail

===============================================================================
 BGP Router ID:192.0.2.6        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : RSVP                   Res. Metric    : 100
Res. Nexthop   : 192.168.56.1 (RSVP LSP: 4)

---snip---
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
```

The GRT output command also shows that the route is reachable using RSVP (indicated as tunneled:RSVP:4):

```
*A:PE-6# show router route-table next-hop-type tunneled

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.10.10.0/24                                 Remote  BGP       00h16m09s  170
       192.0.2.3 (tunneled:RSVP:4)                                  0
-------------------------------------------------------------------------------
No. of Routes: 1
```

If the RSVP LSP is **shutdown**, the system reverts back to the LDP LSP:

```
*A:PE-6# configure router mpls lsp "LSP-PE-6-PE-3" shutdown
```

```
*A:PE-6# show router bgp routes 10.10.10.0/24 detail

===============================================================================
 BGP Router ID:192.0.2.6        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : LDP                     Res. Metric    : 20
Res. Nexthop   : 192.168.56.1 (LDP)

---snip---

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
```

When the shortcut tunnel with **resolution-filter rsvp** is enabled at the BGP level, all RSVP LSPs originating on this node are eligible to be used by default as long as the destination address of the LSP corresponds to that of the BGP next hop for that prefix. It is also possible to exclude a specific RSVP LSP from BGP next hop resolution, similar to the exclusion of a specific RSVP LSP being used as a shortcut for resolving IGP routes. In this example, if the RSVP LSP LSP-PE-6-PE-3 is excluded to be eligible for BGP next hop resolution, it reverts back to LDP.

```
*A:PE-6# configure
    router
        mpls
            lsp "LSP-PE-6-PE-3"
                no bgp-shortcut
                no shutdown
            exit


*A:PE-6# show router route-table 10.10.10.0

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
10.10.10.0/24                                 Remote  BGP       00h04m56s  170
       192.0.2.3 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 1
```

If the configuration is using **disallow-igp**, and neither LDP nor RSVP LSPs are available, the remote route received via BGP is removed from the GRT although the BGP peer session remains up. A field in the detailed show BGP route output indicates that the next hop is “Unresolved”:

```
*A:PE-6# configure
    router
        bgp
            next-hop-resolution
                shortcut-tunnel
                    family ipv4
                        resolution-filter
                            ldp
                            rsvp
                        exit
                        disallow-igp
                        resolution filter
                    exit
                exit
            exit
```

```
*A:PE-6# configure router ldp shutdown
```

```
*A:PE-6# show router bgp routes 10.10.10.0/24 detail

===============================================================================
 BGP Router ID:192.0.2.6        AS:65536       Local AS:65536
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Original Attributes

Network        : 10.10.10.0/24
Nexthop        : 192.0.2.3
Path Id        : None
From           : 192.0.2.3
Res. Protocol  : INVALID                Res. Metric    : 0
Res. Nexthop   : Unresolved

---snip---

Flags          : Invalid  Incomplete  Nexthop-Unresolved
---snip---

-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
Routes : 1
```

Because the route is unresolved, it does not appear in the GRT:

```
*A:PE-6# show router route-table 10.10.10.0

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
No. of Routes: 0
```

MPLS/GRE Shortcut for BGP NH Resolution within a VRF

Using RSVP/LDP or GRE shortcuts for resolving BGP next hops within a Virtual Private Routed Network (VPRN), also known as auto-bind-tunnel, allows a VPRN service to automatically resolve the BGP next hop for VPRN routes to an MPLS LSP or a GRE tunnel. Three possible mechanisms to provide transport tunnels for forwarding traffic between PE routers within an RFC 4364, *BGP/MPLS IP Virtual Private Networks (VPNs)*, network exist:

- RSVP-TE protocol to create tunnel LSPs between PE routers.
- LDP protocol to create tunnel LSPs between PE routers.
- GRE tunnels between PE routers.

These transport tunneling mechanisms provide the flexibility to use dynamically created LSPs where the service tunnels are automatically bound (the **auto-bind-tunnel** feature), and the ability to provide certain VPN services with their own transport tunnels by explicitly binding SDPs if desired. All services using the auto-bind-tunnel feature use the same set of LSPs, which does not allow for alternate tunneling mechanisms (like GRE) or the ability to craft sets of LSPs with bandwidth reservations for specific customers, as is available with explicit SDPs for the service.

The auto-bind-tunnel configuration is as follows:

```
*A:PE-2# configure service vprn 1 auto-bind-tunnel resolution
  - resolution {disabled|any|filter}

 <disabled|any|filt*> : disabled|any|filter


*A:PE-2# configure service vprn 1 auto-bind-tunnel resolution-filter
  - resolution-filter

 [no] bgp             - Enable/disable setting BGP type for auto-bind-tunnel
 [no] gre             - Enable/disable setting GRE type for auto-bind-tunnel
 [no] ldp             - Enable/disable setting LDP type for auto-bind-tunnel
 [no] rsvp            - Enable/disable setting RSVP-TE type for auto-bind-tunnel
 [no] sr-isis         - Enable/disable setting SR-ISIS type for auto-bind-tunnel
 [no] sr-ospf          - Enable/disable setting SR-OSPF type for auto-bind-tunnel
 [no] sr-te            - Enable/disable setting SR-TE type for auto-bind-tunnel
 [no] udp              - Enable/disable setting UDP type for auto-bind-tunnel
```

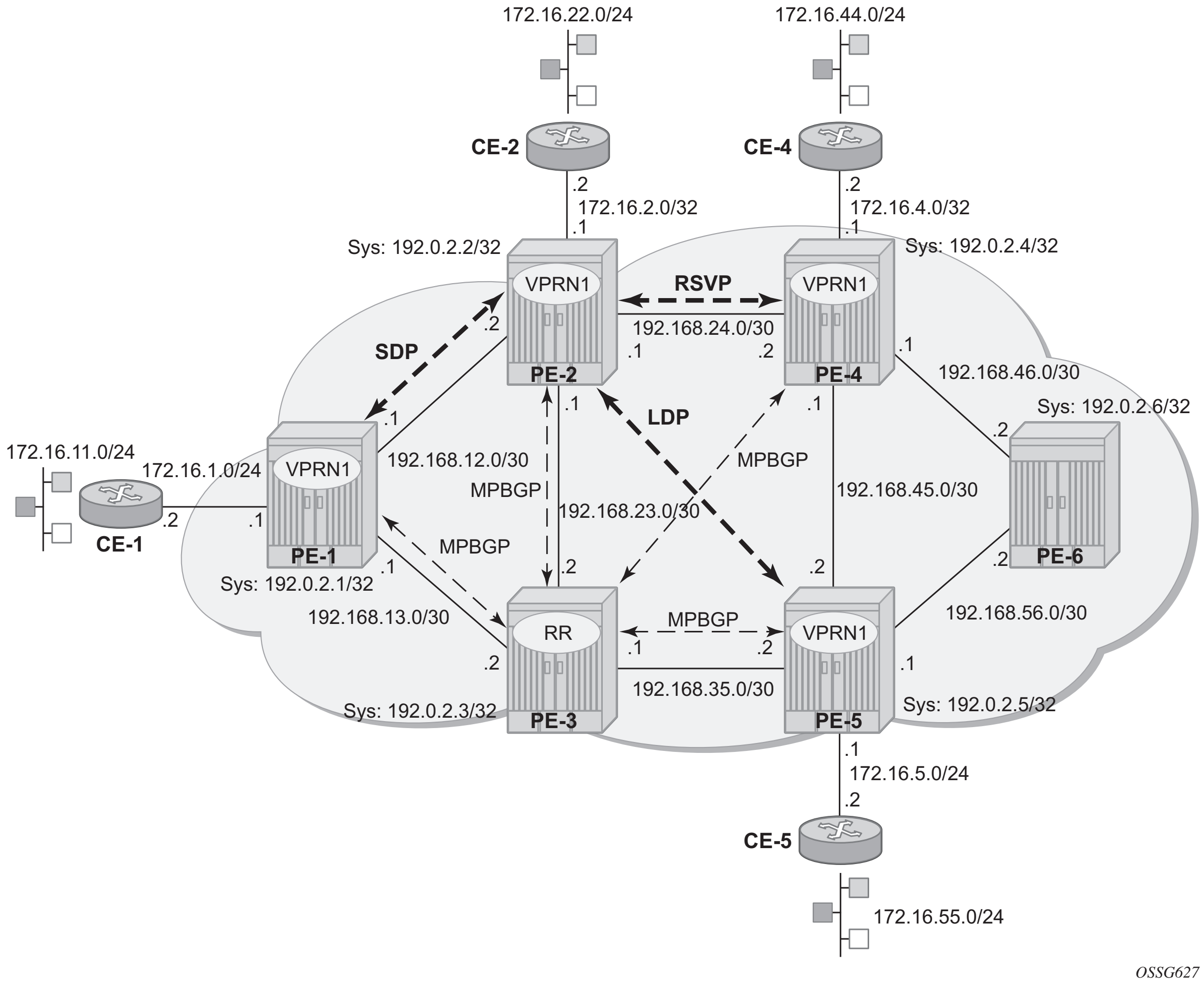
Parameter descriptions:

- **ldp** — Specifies LDP-based LSPs should be used to resolve the BGP next hop for VPRN routes in an associated VPRN instance.
- **gre** — Specifies GRE-based tunnels to be used to resolve the BGP next hop for VPRN routes in an associated VPRN instance. GRE is out of the scope regarding shortcuts, refer to SR OS documentation for further details.
- **rsvp** — Specifies RSVP-TE LSPs should be used to resolve the BGP next hop for VPRN routes in an associated VPRN instance.
- the remaining parameters are beyond the scope of this chapter.

In all cases, if an explicit spoke-SDP is specified in the VPRN, it is always preferred over automatically selected tunnels (even if the SDP is down, the route becomes inactive; there is no fallback to the automatic selection).

The network is configured according to the topology shown in Figure 232. Four PEs (PE-1, PE-2, PE-4, and PE-5) are connected forming a meshed IP-VPN (named VPRN 1), using a route reflector on PE-3 for MP-BGP peering. All PEs have LDP tunnels enabled so at a minimum all can establish LDP shortcut tunnels to the others. In order to have not only LDP but also RSVP-TE LSPs and static SDPs (using an RSVP LSP) in the network, a mix of tunneling methods is configured. For brevity, the configuration of PE-2 only is given, providing the details about the shortcuts created by auto-bind-tunnel. PE-2 has a static SDP (RSVP-based) with PE-1, an RSVP LSP with PE-4, and an LDP LSP with PE-5. Every PE has a CE connected, so each PE has an interface connected to the CE as well as a static route to a CE LAN (although redistribution routing policies are needed, they are not shown for brevity).

Figure 232 **Shortcuts Within a VRF Topology Network**



On PE-2, VPRN1 is configured as follows:

```
*A:PE-2# configure
    service
        sdp 1 mpls create
            far-end 192.0.2.1
            lsp "LSP-PE-2-PE-1"
            no shutdown
        exit
        vprn 1 name "VPRN 1" customer 1 create
            vrf-import "VPN1-import"
            vrf-export "VPN1-export"
            route-distinguisher 65536:1
            auto-bind-tunnel
                resolution-filter
                    gre
                    ldp
                    rsvp
                exit
                resolution filter
            exit
            interface "to-CE-2" create
                address 172.16.2.1/24
                sap 1/1/4:1 create
                exit
            exit
            static-route-entry 172.16.22.0/24
                next-hop 172.16.2.2
                    no shutdown
                exit
            exit
            spoke-sdp 1 create
            exit
            no shutdown
        exit
```

As previously mentioned, regarding IP-VPN meshed connectivity, the configuration shows that there is a static SDP 1 (pointing to PE-1), and the rest of the configuration is just **auto-bind-tunnel.** On PE-2, the connectivity toward the other PEs in the network can be verified by checking VPRN 1:

```
*A:PE-2# show router 1 route-table

================================================================================
Route Table (Service: 1)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
172.16.1.0/24                                  Remote  BGP VPN   00h12m26s  170
       192.0.2.1 (tunneled)                                          0
172.16.2.0/24                                  Local   Local     00h20m27s  0
       to-CE-2                                                       0
172.16.4.0/24                                  Remote  BGP VPN   00h00m27s  170
       192.0.2.4 (tunneled:RSVP:3)                                   0
172.16.5.0/24                                  Remote  BGP VPN   00h12m26s  170
       192.0.2.5 (tunneled)                                          0
172.16.11.0/24                                 Remote  BGP VPN   00h12m26s  170
       192.0.2.1 (tunneled)                                          0
172.16.22.0/24                                 Remote  Static    00h20m27s  5
       172.16.2.2                                                    1
172.16.44.0/24                                 Remote  BGP VPN   00h00m27s  170
       192.0.2.4 (tunneled:RSVP:3)                                   0
172.16.55.0/24                                 Remote  BGP VPN   00h12m26s  170
       192.0.2.5 (tunneled)                                          0
--------------------------------------------------------------------------------
No. of Routes: 8
```

As can be seen, there are eight routes because every PE has two routes (one direct PE-CE interface and one static route), so six routes are received from other PEs via MP-BGP. The VPRN 1 routing table can be understood by looking at the tunnel table (active LSPs for remote system IDs):

```
*A:PE-2# show router tunnel-table
```

```
===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref      Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      sdp       MPLS  1         5        192.0.2.1      0
192.0.2.1/32      rsvp      MPLS  2         7        192.168.12.1   10
192.0.2.1/32      ldp       MPLS  65537     9        192.168.12.1   10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.23.2   10
192.0.2.4/32      rsvp      MPLS  3         7        192.168.24.2   10
192.0.2.4/32      rsvp      MPLS  4         7        192.168.24.2   16777215
192.0.2.4/32      ldp       MPLS  65545     9        192.168.24.2   10
192.0.2.5/32      ldp       MPLS  65542     9        192.168.23.2   20
192.0.2.6/32      ldp       MPLS  65546     9        192.168.24.2   20
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#
```

The tunnel table shows one entry per LSP per remote PE. The following tunnel selection rules apply:

- SDP has the lowest (best) preference, followed by RSVP and then by LDP.
- If the preference is the same, the lowest metric is selected (ECMP is possible with LDP).

PE-2 has three possibilities to reach PE-1 (192.0.2.1): an SDP tunnel ID 1 with preference 5, an RSVP tunnel ID 1 with preference 7, and an LDP LSP with preference 9. Because SDP tunnel ID 1 has the lowest preference, it is the chosen option. PE-2 has three possibilities to reach PE-4 (192.0.2.4): an RSVP tunnel ID 3 with preference 7 and metric 10, an RSVP tunnel ID 4 with preference 7 and metric 16777215, and an LDP LSP with preference 9; so RSVP tunnel ID 3 is selected. PE-2 only has one option to reach PE-5 and PE-6 (192.0.2.5 and 192.0.2.6) using an LDP LSP.

The following FIB for router VPRN 1 on PE-2 provides more detailed information on the tunneling:

```
*A:PE-2# show router 1 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
172.16.1.0/24                                               BGP_VPN
  192.0.2.1 (VPRN Label:524281 Transport:SDP:1)
172.16.2.0/24                                               LOCAL
  172.16.2.0 (to-CE-2)
```

```
172.16.4.0/24                                               BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:RSVP LSP:3)
172.16.5.0/24                                               BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
172.16.11.0/24                                              BGP_VPN
  192.0.2.1 (VPRN Label:524281 Transport:SDP:1)
172.16.22.0/24                                              STATIC
  172.16.2.2 (to-CE-2)
172.16.44.0/24                                              BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:RSVP LSP:3)
172.16.55.0/24                                              BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 8
-------------------------------------------------------------------------------
```

The FIB shows the chosen transport tunnel, specifying SDP ID, RSVP Tunnel ID, and LDP, as well as service label information linked to the routes.

Static SDP tunnels are preferred over dynamic tunnels (RSVP or LDP auto-bind-tunnel). When the static SDP 1 is shut down or the LSP goes down (there is no fallback to dynamic tunneling), the associated routes are removed:

```
*A:PE-2# configure service sdp 1 shutdown
```

```
*A:PE-2# show router 1 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
172.16.2.0/24                                               LOCAL
  172.16.2.0 (to-CE-2)
172.16.4.0/24                                               BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:RSVP LSP:3)
172.16.5.0/24                                               BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
172.16.22.0/24                                              STATIC
  172.16.2.2 (to-CE-2)
172.16.44.0/24                                              BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:RSVP LSP:3)
172.16.55.0/24                                              BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 6
```

To avoid this fallback issue, the configuration is modified and the manual spoke-SDPs are removed from the configuration of PE-1 and PE-2; the rest of the configuration remains the same. Now the connectivity between PE-1 and PE-2 is using an RSVP LSP, as shown in the PE-1 following output (RSVP LSP which was used by SDP 1 has disappeared):

```
*A:PE-1# configure service vprn 1 no spoke-sdp 1
```

```
*A:PE-2# configure service vprn 1 no spoke-sdp 1
```

```
*A:PE-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
172.16.1.0/24                                 Local   Local     00h33m21s  0
       to-CE-1                                                      0
172.16.2.0/24                                 Remote  BGP VPN   00h00m24s  170
       192.0.2.2 (tunneled:RSVP:3)                                  0
172.16.4.0/24                                 Remote  BGP VPN   00h13m29s  170
       192.0.2.4 (tunneled)                                         0
172.16.5.0/24                                 Remote  BGP VPN   00h25m09s  170
       192.0.2.5 (tunneled)                                         0
172.16.11.0/24                                Remote  Static    00h33m21s  5
       172.16.1.2                                                   1
172.16.22.0/24                                Remote  BGP VPN   00h00m24s  170
       192.0.2.2 (tunneled:RSVP:3)                                  0
172.16.44.0/24                                Remote  BGP VPN   00h13m29s  170
       192.0.2.4 (tunneled)                                         0
172.16.55.0/24                                Remote  BGP VPN   00h25m09s  170
       192.0.2.5 (tunneled)                                         0
-------------------------------------------------------------------------------
No. of Routes: 8
```

If RSVP is disabled, the connectivity falls back to LDP as the output shows:

```
*A:PE-1# configure router mpls shutdown
```

```
*A:PE-1# show router 1 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
172.16.1.0/24                                               LOCAL
  172.16.1.0 (to-CE-1)
172.16.2.0/24                                               BGP_VPN
  192.0.2.2 (VPRN Label:524280 Transport:LDP)
172.16.4.0/24                                               BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:LDP)
172.16.5.0/24                                               BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
172.16.11.0/24                                              STATIC
  172.16.1.2 (to-CE-1)
172.16.22.0/24                                              BGP_VPN
  192.0.2.2 (VPRN Label:524280 Transport:LDP)
172.16.44.0/24                                              BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:LDP)
```

```
172.16.55.0/24                                              BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:LDP)
-------------------------------------------------------------------------------
Total Entries : 8
-------------------------------------------------------------------------------
```

If LDP is disabled, the connectivity falls back to GRE as the output shows:

```
*A:PE-1# configure router ldp shutdown
```

```
*A:PE-1# show router 1 fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
172.16.1.0/24                                               LOCAL
  172.16.1.0 (to-CE-1)
172.16.2.0/24                                               BGP_VPN
  192.0.2.2 (VPRN Label:524280 Transport:GRE)
172.16.4.0/24                                               BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:GRE)
172.16.5.0/24                                               BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:GRE)
172.16.11.0/24                                              STATIC
  172.16.1.2 (to-CE-1)
172.16.22.0/24                                              BGP_VPN
  192.0.2.2 (VPRN Label:524280 Transport:GRE)
172.16.44.0/24                                              BGP_VPN
  192.0.2.4 (VPRN Label:524280 Transport:GRE)
172.16.55.0/24                                              BGP_VPN
  192.0.2.5 (VPRN Label:524281 Transport:GRE)
-------------------------------------------------------------------------------
Total Entries : 8
-------------------------------------------------------------------------------
```

Conclusion

IGP shortcuts provide a variety of shortcuts in IP, MPLS, and IP-VPN scenarios to customers who want to use new options for building routing topologies. Because IGP shortcuts are enabled on a per router basis, SPF computations are independent and irrelevant to other routers, so there is no need to enable shortcuts globally. This network example shows the configuration of IGP shortcuts together with the associated show outputs which can be used for verification and troubleshooting.

Inter-Area TE Point-to-Point LSPs

This chapter describes inter-area Traffic Engineering (TE) Point-to-Point (P2P) Label Switched Paths (LSPs) configurations.

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS release 11.0.R4, but the CLI in the current edition corresponds to SR OS release 16.0.R3.

Overview

Multi-Protocol Label Switching with Traffic Engineering (MPLS TE) is implemented on a wide scale in current Internet Service Provider (ISP) networks to steer traffic across their backbones to facilitate efficient use of available bandwidth between the routers and to guarantee fast convergence in case a link or node fails.

Regular TE LSPs in MPLS networks are confined to only a single Interior Gateway Protocol (IGP) area or level. This is because the head-end has information in the TE database of only the local area for Open Shortest Path First (OSPF) or level for Intermediate System to Intermediate System (IS-IS). As the name implies, inter-area TE LSPs can cross the area or level borders of the IGP.

Inter-Area TE LSP Based On Explicit Route Expansion

Inter-area TE LSP using Explicit Route Object (ERO) expansion enables the head-end to calculate the ERO path within its own area or level and keep the remaining Area Border Routers (ABRs) of other areas/levels as loose hops in the ERO path. On receiving a PATH message with a loose hop ERO and based on local configuration, each ABR does a partial Constrained Shortest Path First (CSPF) calculation to the next ABR or a full CSPF calculation to reach the destination.

Automatic selection of ABRs is supported so that the head-end node can work with an empty primary path. When the **to** field of an LSP definition is in an area/level different from the head-end node, CSPF will automatically compute the segment to the exit ABR router which advertised the prefix and which currently is the best path for resolving the prefix in the Route Table Manager (RTM).

ABR Protection

Link and node protection within the respective areas are supported through the TE capabilities of the IGP and Resource Reservation Protocol (RSVP) in each area. To support ABR node protection, a bypass is required from the Point of Local Repair (PLR; node prior to ABR) to the Merge Point (MP; next-hop node to ABR).

Two methods are possible: dynamic ABR protection and static ABR protection. Static ABR protection uses Manual Bypass Tunnels (MBTs), statically configured by the operator between the PLR and the MP. For dynamic ABR protection, node ID propagation and signaling of an eXclude Route Object (XRO) in RSVP PATH messages must both be supported.

Because the Record Route Object (RRO) Node ID sub-object description in RFC 4561 (*Definition of a Record Route Object (RRO) Node-Id Sub-Object*) is not clear about the format of the included node address (S), interface address (I) and label (L), the system supports multiple formats: IL, SL, ISL, SIL, SLI, ILSL and SLIL. The system uses the SLIL (node-address, label, interface-address, label) format to include the node ID itself.

The exclude route object (XRO) inclusion (RFC 4874, *Exclude Routes - Extension to Resource ReserVation Protocol-Traffic Engineering*) in bypass RSVP PATH messages is required to exclude the protected ABR from the bypass path. The XRO object contains the ABR system IP address.

Example Topology

The example topology in this chapter contains ten nodes in three areas, as shown in Figure 233.

Figure 233 **Inter-Area TE LSP Setup**

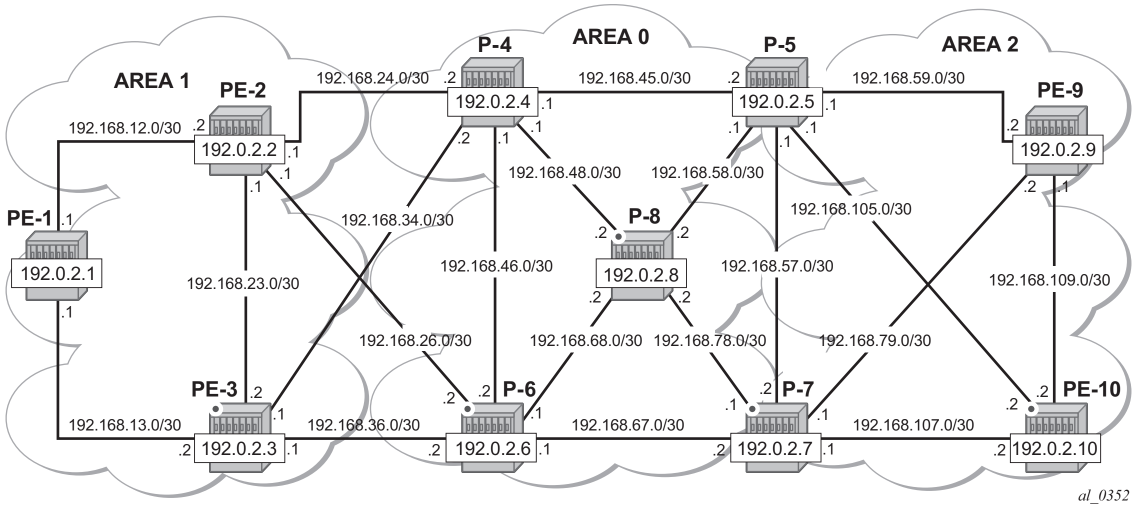
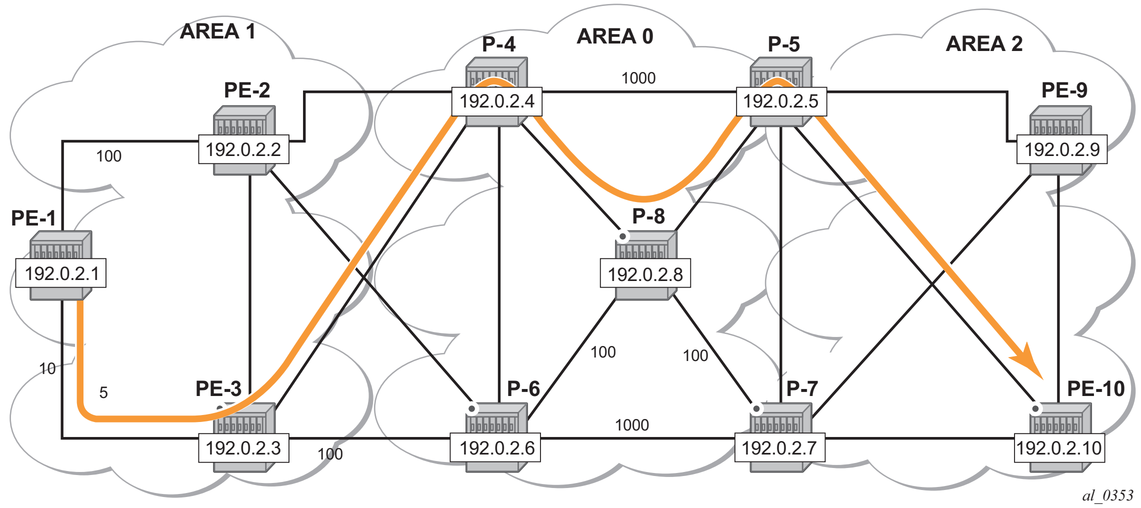


Figure 234 shows the LSP path intended to be set up through the network. An empty MPLS path is used. At the head-end node PE-1, the destination address PE-10 is learned via ABR node P-4 and ABR node P-5.

Figure 234 **Inter-Area TE LSP Path**



Configuration

The following base configuration has been implemented on the nodes:

- Cards, MDAs, and ports configured
- Interfaces configured
- IGP areas configured and converged
- Traffic Engineering configured for the IGP
- MPLS and RSVP configured on all links in the network

OSPF or IS-IS can be configured as the IGP; OSPF is used in this chapter.

The following output shows the opaque database of PE-1:

```
A:PE-1# show router ospf opaque-database

===============================================================================
OSPF Opaque Link State Database (Type : All)
===============================================================================
Type  Id              Link State Id     Adv Rtr Id      Age  Sequence   Cksum
-------------------------------------------------------------------------------
Area  0.0.0.1         1.0.0.1           192.0.2.1       1503 0x80000002 0x9234
Area  0.0.0.1         1.0.0.3           192.0.2.1       1470 0x80000001 0x9b45
Area  0.0.0.1         1.0.0.4           192.0.2.1       1465 0x80000001 0xe9f2
Area  0.0.0.1         1.0.0.1           192.0.2.2       1498 0x80000002 0x962e
Area  0.0.0.1         1.0.0.3           192.0.2.2       1470 0x80000001 0xdce8
Area  0.0.0.1         1.0.0.4           192.0.2.2       1470 0x80000001 0x833b
Area  0.0.0.1         1.0.0.5           192.0.2.2       1471 0x80000001 0x637b
Area  0.0.0.1         1.0.0.6           192.0.2.2       1471 0x80000001 0x665f
Area  0.0.0.1         1.0.0.1           192.0.2.3       1504 0x80000002 0x9a28
Area  0.0.0.1         1.0.0.3           192.0.2.3       1472 0x80000001 0x6d43
Area  0.0.0.1         1.0.0.4           192.0.2.3       1481 0x80000001 0x1495
Area  0.0.0.1         1.0.0.5           192.0.2.3       1470 0x80000001 0x9f3c
Area  0.0.0.1         1.0.0.6           192.0.2.3       1472 0x80000001 0x4283
Area  0.0.0.1         1.0.0.1           192.0.2.4       1482 0x80000002 0x9e22
Area  0.0.0.1         1.0.0.6           192.0.2.4       1471 0x80000001 0x7e44
Area  0.0.0.1         1.0.0.7           192.0.2.4       1473 0x80000001 0x218b
Area  0.0.0.1         1.0.0.1           192.0.2.6       1584 0x80000002 0xa616
Area  0.0.0.1         1.0.0.6           192.0.2.6       1467 0x80000001 0xf6c5
Area  0.0.0.1         1.0.0.7           192.0.2.6       1482 0x80000001 0x990d
-------------------------------------------------------------------------------
No. of Opaque LSAs: 19
===============================================================================
A:PE-1#
```

The information is only about routers that are part of area 0.0.0.1. PE-1 cannot calculate an end-to-end CSPF path to node PE-10 because this would require TE topology information from area 0.0.0.0 and area 0.0.0.2.

Each node announces its router ID and each attached link that is part of that area, resulting in 19 opaque LSAs in area 0.0.0.1. The system interfaces of P-4 and P-6 are configured in backbone area 0.0.0.0, not in area 0.0.0.1.

In Figure 234, the LSP passes through node PE-3 and node P-8. In order to prefer a dynamic path from PE-1 to P-4 via PE-3 rather than via PE-2, it is necessary to configure on PE-1 a lower IGP metric on the interface to PE-3 (the default metric is derived from the interface speed; in this case the metric is 10 by default).

```
*A:PE-1# configure router ospf area 1 interface "int-PE-1-PE-3" metric 5
```

Similarly, in the core, the IGP metric between P-4 and P-5, and between P-6 and P-7 is increased to force the LSP to pass through the core P-8 node.

```
*A:P-4# configure router ospf area 0 interface "int-P-4-P-5" metric 1000
```

```
*A:P-6# configure router ospf area 0 interface "int-P-6-P-7" metric 1000
```

Other metrics have also been manipulated as shown on Figure 234.

MPLS Path Configuration

An empty MPLS path is sufficient on the head-end node PE-1, because automatic ABR selection is performed. Using an empty MPLS path will ease the provisioning process and brings consistency because this empty MPLS path can be used for both intra and inter-area/level type LSPs.

```
*A:PE-1# configure router mpls path “dyn” no shutdown
```

MPLS LSP Configuration

On PE-1, the following LSP to PE-10 is configured with the previously created MPLS path as primary path. CSPF and fast reroute (FRR) facility are enabled on the LSP.

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-10"
                to 192.0.2.10
                cspf
                fast-reroute facility
                exit
                primary "dyn"
                exit
                no shutdown
            exit
```

At this stage, the LSP is in an operational Down state with a failure code of badNode at failure node 192.168.34.2 (ABR P-4), as follows.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-10" path detail
---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-10 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-10
From             : 192.0.2.1            To                   : 192.0.2.10
Admin State      : Up                   Oper State           : Down
Path Name        : dyn
Path LSP ID      : 4610                 Path Type            : Primary
Path Admin       : Up                   Path Oper            : Down
---snip---

Failure Code     : badNode
Failure Node     : 192.168.34.2
---snip---
```

In order to get around the intra-area CSPF confinement, the ERO-expansion feature is enabled on all ABR nodes.

```
*A:P-4# configure router mpls cspf-on-loose-hop
*A:P-6# configure router mpls cspf-on-loose-hop

*A:P-7# configure router mpls cspf-on-loose-hop
*A:P-5# configure router mpls cspf-on-loose-hop
```

Cspf-on-loose-hop is only required if FRR or TE parameters are configured on the LSP. If any of these parameters is configured on the LSP while one of the ABRs along the path is not configured with cspf-on-loose-hop, the LSP will stay operationally down with failure code: badNode and an indication of the interface address of the failure node.

The LSP path can also contain other strict and/or loose hops. However, **cspf-on-loose-hop** must be configured in the MPLS context whenever loose hops are configured in the MPLS path. This command enables ERO expansion and is only required for inter-area LSPs on all possible ABR nodes and all nodes not belonging to the area where the iLER is located, which have a loose hop reference in the MPLS path. However, for simplicity and consistency, it can be configured on all nodes without having a negative effect.

The following trace shows the ERO calculation on the head-end to the first ABR.

```
*A:PE-1# debug router rsvp packet path detail

4 2018/09/11 10:25:02.22 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
```

```
Send PATH From:192.0.2.1, To:192.0.2.10
          TTL:255, Checksum:0x3b17, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PE-1-PE-10::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
RSVPHop    - Ctype:1, Addr:192.168.13.1, LIH:3
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:4614
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.13.1, Flags:0x0
ERO        - IPv4Prefix 192.168.13.2/32, Strict
             IPv4Prefix 192.168.34.2/32, Strict
             IPv4Prefix 192.0.2.10/32, Loose
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"
```

On the ABR P-4, the ERO is expanded to include the nodes of area 0.0.0.0 of which P-4 is also part. The RRO contains all the hops the PATH message has passed so far.

```
*A:P-4# debug router rsvp packet path detail

4 2018/09/11 10:24:40.42 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.10
          TTL:253, Checksum:0xc992, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PE-1-PE-10::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
RSVPHop    - Ctype:1, Addr:192.168.48.1, LIH:4
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:4614
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.48.1, Flags:0x0
             IpAddr:192.168.34.1, Flags:0x0
             IpAddr:192.168.13.1, Flags:0x0
ERO        - IPv4Prefix 192.168.48.2/32, Strict
             IPv4Prefix 192.168.58.1/32, Strict
             IPv4Prefix 192.0.2.10/32, Loose
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"
```

Finally, the P-5 ABR will expand the ERO to the final destination PE-10:

```
*A:P-5# debug router rsvp packet path detail

8 2018/09/11 10:25:03.52 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.10
          TTL:251, Checksum:0x71f1, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
```

```
SessAttr   - Name:LSP-PE-1-PE-10::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
RSVPHop    - Ctype:1, Addr:192.168.105.1, LIH:5
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:4614
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.105.1, Flags:0x0
             IpAddr:192.168.58.2, Flags:0x0
             IpAddr:192.168.48.1, Flags:0x0
             IpAddr:192.168.34.1, Flags:0x0
             IpAddr:192.168.13.1, Flags:0x0
ERO        - IPv4Prefix 192.168.105.2/32, Strict
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"
```

The MPLS LSP is now operational up and the LSP path can be shown in detail on the head-end, PE-1:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-10" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-10 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-10 Path dyn
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-10
From              : 192.0.2.1           To                    : 192.0.2.10
Admin State       : Up                  Oper State            : Up
Path Name         : dyn
Path LSP ID       : 4614                Path Type             : Primary
Path Admin        : Up                  Path Oper             : Up
Out Interface     : 1/1/2               Out Label             : 524287
---snip---

CSPF              : Enabled             Oper CSPF             : Enabled
Least Fill        : Disabled            Oper LeastFill        : Disabled
FRR               : Enabled             Oper FRR              : Enabled
FRR NodeProtect   : Enabled             Oper FRR NP           : Enabled
FR Hop Limit      : 16                  Oper FRHopLimit       : 16
FR Prop Admin Gr*: Disabled             Oper FRPropAdmGrp     : Disabled
Propogate Adm Grp: Disabled             Oper Prop Adm Grp     : Disabled
Inter-area        : True
---snip---

Explicit Hops     :
    No Hops Specified
Actual Hops       :
    192.168.13.1 (192.0.2.1) @ n                Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                  Record Label        : 524287
 -> 192.168.34.2 (192.0.2.4) @ n                Record Label        : 524287
 -> 192.168.48.2  @                             Record Label        : 524287
 -> 192.168.58.1  @                             Record Label        : 524287
 -> 192.168.105.2                               Record Label        : 524287

Computed Hops    :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.34.2(SA)
 -> 192.0.2.10(L)
Resignal Eligible: False
Last Resignal    : n/a                 CSPF Metric          : 15
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

ABR Node Protection

The LSP is configured with facility FRR protection; link and node protection are established within each area, as shown in the preceding output. Node protection is available for nodes PE-3 in area 1 (bypass originating in PE-1), and P-8 in area 0 (bypass originating in P-4), but not for the ABRs P-4 and P-5. No bypass tunnels for node protection originate in PLRs PE-3 (for ABR P-4) or P-8 (for ABR P-5). The bypass tunnels originating in PE-3 and P-8 only offer link protection. Dynamic ABR node protection requires the setup of a bypass tunnel from the PLR (node just upstream of the ABR) to the MP (node just downstream of the ABR). The following two things are required to establish a bypass tunnel for an ABR:

- The PLR node (part of area x) needs to know the system IP address of the MP node (part of area y) to set up the bypass. For this reason, the node ID of the MP must be included in the RESV message so that the PLR can link the manual bypass tunnel to the primary path to protect the ABR. By default, the node ID is not included in the RESV message, but it can be configured on the MPs as follows: **configure router rsvp node-id-in-rro include**.
- The other ABR node receiving the RSVP bypass PATH message for the protected ABR needs to do an ERO expansion toward the MP node. For this reason, the XRO object is included in the RSVP bypass PATH message, containing the node ID of the protected ABR. As an example, the following bypass PATH message is shown on node PE-3.

The XRO object includes the system IP address of the protected ABR node P-4 and the ERO object has MP node P-8 as loose destination:

```
*A:PE-3# debug router rsvp packet path detail

1 2018/09/11 10:26:58.62 UTC MINOR: DEBUG #2001 Base RSVP
```

```
"RSVP: PATH Msg
Send PATH From:192.0.2.3, To:192.0.2.8
          TTL:17, Checksum:0xfddd, Flags:0x0
Session    - EndPt:192.0.2.8, TunnId:61442, ExtTunnId:192.0.2.3
SessAttr   - Name:bypass-node192.0.2.4-61442
             SetupPri:7, HoldPri:0, Flags:0x2
RSVPHop    - Ctype:1, Addr:192.168.36.1, LIH:3
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.3, LspId:4
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.36.1, Flags:0x0
ERO        - IPv4Prefix 192.168.36.2/32, Strict
             IPv4Prefix 192.0.2.8/32, Loose
XRO        - IPv4Prefix: 192.0.2.4/32, Attribute: Node, LBit: Exclude
AdSpec     - General BreakBit:0, NumISHops:0, PathBwEstimate:0
                     MinPathLatency:4294967295, CompPathMTU:1564
             Controlled BreakBit:0
"
```

Node-ID Inclusion in the RESV Message

P-8 will be the MP for the bypass of ABR P-4 and PE-10 will be the MP for the bypass of ABR P-5. So P-8 and PE-10 need to include their node ID in the RESV message, inside the Record Route Object (RRO).

```
*A:P-8# configure router rsvp node-id-in-rro include
*A:PE-10# configure router rsvp node-id-in-rro include
```

The default is **node-id-in-rro exclude**. As an example, the RESV message received on PLR node PE-3 is as follows. The RRO contains the MP node P-8 information in SLIL format:

```
*A:PE-3# debug router rsvp packet resv detail

7 2018/09/11 10:27:49.62 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESV Msg
Send RESV From:192.168.13.2, To:192.168.13.1
          TTL:255, Checksum:0xfcb3, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
RSVPHop    - Ctype:1, Addr:192.168.13.2, LIH:3
TimeValue  - RefreshPeriod:30
Style      - SE
FlowSpec   - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1560, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.1, LspId:4614, Label:524287
RRO        - ---snip---
             SystemIp:192.0.2.8, Flags:0x29
             Label:524287, Flags:0x1
             InterfaceIp:192.168.48.2, Flags:0x9
             Label:524287, Flags:0x1
---snip---
```

"

Bypass Configuration for ABR Protection

Because dynamic ABR protection is supported and used in this example, no explicit Manual Bypass Tunnels (MBTs) are configured to protect the ABRs. Each PLR first checks if an MBT tunnel exists between the PLR and the MP matching the constraints and protecting the ABR. If no MBT is available, the PLR will signal a bypass tunnel in a dynamic way toward the MP node.

Figure 235 shows the two dynamic ABR node protections that are signaled for this LSP.

Figure 235 **ABR Protection**

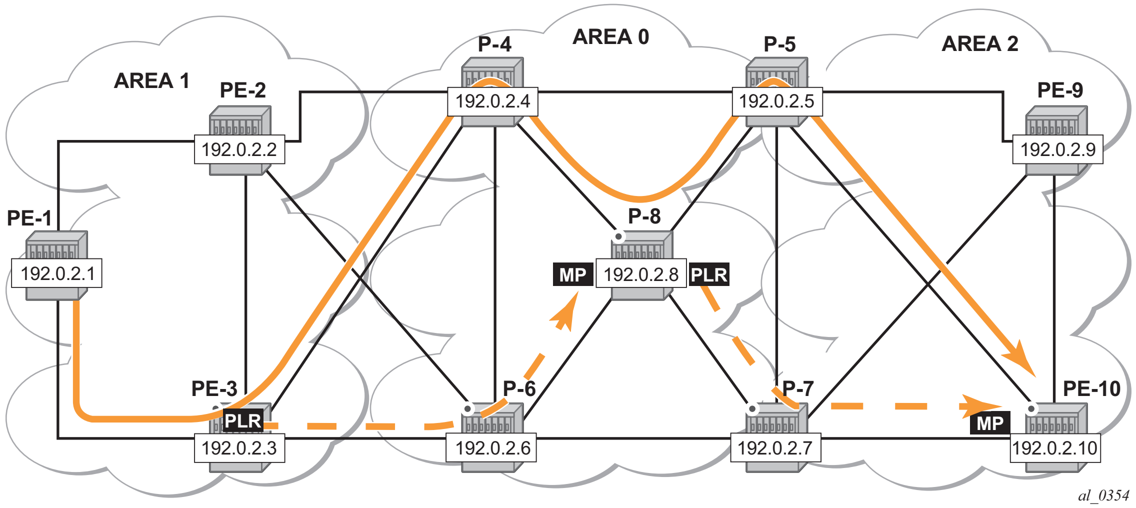
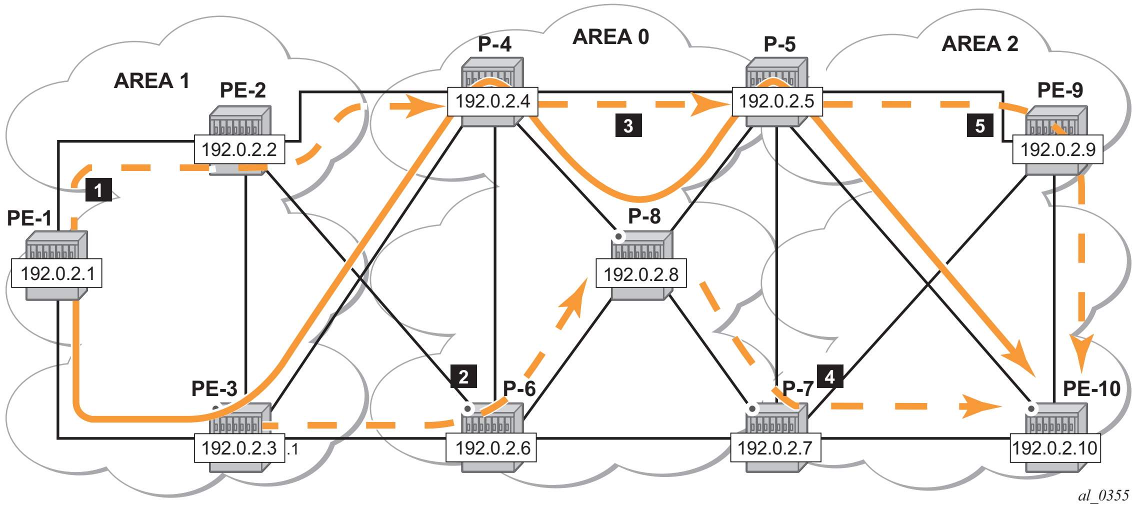


Figure 236 shows the complete picture of all the FRR protections and indicates each node/link protection in the setup.

Figure 236 Protection of All Nodes/Links Along the LSP Path



This can be seen in the detailed show output of the LSP path:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-10" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-10 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-10 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-10
From             : 192.0.2.1            To                     : 192.0.2.10
Admin State      : Up                   Oper State             : Up
Path Name        : dyn
Path LSP ID      : 4614                 Path Type              : Primary
Path Admin       : Up                   Path Oper              : Up
---snip---

Inter-area       : True
---snip---

Actual Hops      :
    192.168.13.1 (192.0.2.1) @ n                 Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @ n                 Record Label        : 524287
 -> 192.168.34.2 (192.0.2.4) @ n                 Record Label        : 524287
 -> 192.0.2.8 (192.0.2.8) @ n                    Record Label        : 524287
 -> 192.168.48.2  @ n                            Record Label        : 524287
 -> 192.0.2.5 (192.0.2.5) @                      Record Label        : 524287
 -> 192.168.58.1  @                              Record Label        : 524287
 -> 192.0.2.10 (192.0.2.10)                      Record Label        : 524287
 -> 192.168.105.2                                Record Label        : 524287
---snip---
```

- The first bypass originates in PE-1 and protects node PE-3.
- The second bypass originates in PE-3 and protects node P-4.
- The third bypass originates in P-4 and protects node P-8.
- The fourth bypass originates in P-8 and protects node P-5. There are two entries for P-8: hop 192.0.2.8 and hop 192.168.48.2.
- The fifth bypass originates in P-5 and protects the link between P-5 and PE-10. There are two entries for P-5: hop 192.0.2.5 and hop 192.168.58.1.

There are two entries for P-8, P-5 and PE-10 in the ‘Actual Hops’ section in the previous output: one for the interface IP address and one for the system IP address. This is a consequence of configuring **node-id-in-rro include** on P-8, P-5, and PE-10.

The **node-id-in-rro include** command is not mandatory for this example on ABR node P-5, but to be able to cover cases where a new LSP is established in the network and P-5 acts as an MP node while the corresponding PLR node for that new LSP is in another area. This RSVP command can be executed on all possible MP nodes in the network.

The following command shows the details of the bypass tunnel from PE-3 to PE-8, protecting PE-4:

```
*A:PE-3# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-node192.0.2.4-61442
-------------------------------------------------------------------------------
To             : 192.0.2.8          State              : Up
Out I/F        : 1/1/2              Out Label          : 524287
Up Time        : 0d 00:01:17        Active Time        : n/a
Reserved BW    : 0 Kbps             Protected LSP Count : 1
Type           : Dynamic            Bypass Path Cost    : 100
Setup Priority : 7                  Hold Priority       : 0
Class Type     : 0
Exclude Node   : 192.0.2.4          Inter-Area          : True
Computed Hops  :
    192.168.36.1(S)                 Egress Admin Groups : None
 -> 192.168.36.2(SA)                Egress Admin Groups : None
 -> 192.0.2.8(L)                    Egress Admin Groups : None
Actual Hops    :
    192.168.36.1 (192.0.2.3)        Record Label        : N/A
 -> 192.168.36.2 (192.0.2.6)        Record Label        : 524287
 -> 192.0.2.8 (192.0.2.8)           Record Label        : 524287
 -> 192.168.68.2                    Record Label        : 524287
```

```
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-10::dyn
From           : 192.0.2.1           To                  : 192.0.2.10
Avoid Node/Hop : 192.0.2.4           Downstream Label    : 524287
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-3#
```

The LSP could be protected with one or more additional secondary paths, pre-signaled or not, but this is outside the scope of this chapter.

When a link or node failure occurs along the LSP path, FRR protection kicks in and end-to-end path re-optimization is executed: a PATHERR message is forwarded to the head-end. Upon receiving the PATHERR message, the head-end calculates a new path.

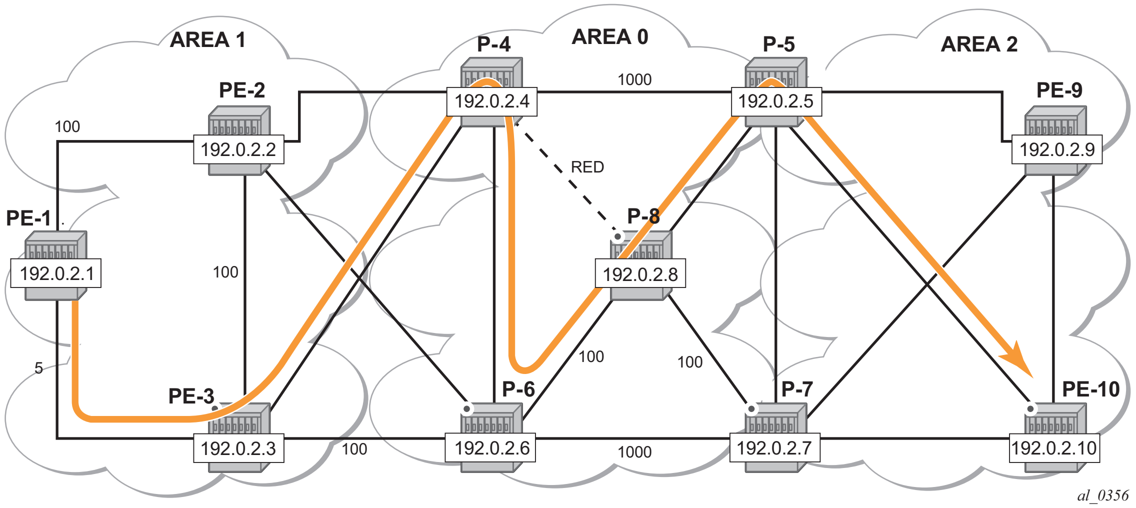
Admin Groups

The use of administrative groups is described in the RSVP Point-to-Point LSPs chapter.

To support admin groups for inter-area LSPs, the ingress node PE-1 must propagate the admin groups within the Session Attribute object (SA) of the PATH message so that the ABRs along the path receive the admin group restrictions they have to take into account when further expanding the ERO in the PATH message.

In Figure 236 the LSP path avoids the link between P-4 and P-8. This is implemented by assigning admin group “red” to the link between P-4 and P-8 and then configuring the LSP to exclude the admin group “red”.

Figure 237 **Admin Group Example**



Admin Group Configuration

On P-4, configure admin group “red” and assign a group value. In this example, group value 11 is used, but this can be any value between 0 and 31. Assign admin group “red” to the link to P-8.

This admin group configuration is required on P-4 and on iLER PE-1. However, it is good practice to configure the admin group on all the nodes.

```
*A:Px# configure router if-attribute admin-group red value 11
```

```
*A:P-4# configure
    router
        mpls
            interface "int-P-4-P-8"
                admin-group "red"
            exit
```

On PE-1, change the LSP configuration as follows:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-10"
                exclude "red"
                propagate-admin-group
            exit
```

It is possible to have the same admin group constraint applied to the FRR bypass tunnels in the PLRs, but that is not the case here. The bypass tunnels ignore any admin group constraint. The **propagate-admin-group** command is required to include the admin group properties in the SA object of the PATH message. The admin group value is mapped to a 32-bitmap. In this example, value 11 means that the 12th bit is set, which means in binary 100000000000 or hex 0x800.

```
*A:PE-1# debug router rsvp packet path detail

1 2018/09/11 10:29:25.22 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.10
          TTL:255, Checksum:0x3301, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PE-1-PE-10::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
             Ctype:RA, ExcAny:0x800, IncAny:0x0, IncAll:0x0
RSVPHop    - Ctype:1, Addr:192.168.13.1, LIH:3
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:4618
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.13.1, Flags:0x0
ERO        - IPv4Prefix 192.168.13.2/32, Strict
             IPv4Prefix 192.168.34.2/32, Strict
             IPv4Prefix 192.0.2.10/32, Loose
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"
```

The following two sets of output show that when P-4 expands the ERO it now excludes the link to node P-8 for the path calculation and the path is set up through P-6, P-8 and P-5.

```
*A:P-4# debug router rsvp packet path detail

8 2018/09/11 10:30:01.43 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.10
          TTL:253, Checksum:0xa1ba, Flags:0x0
Session    - EndPt:192.0.2.10, TunnId:1, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PE-1-PE-10::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
             Ctype:RA, ExcAny:0x800, IncAny:0x0, IncAll:0x0
RSVPHop    - Ctype:1, Addr:192.168.46.1, LIH:3
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.1, LspId:4618
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.46.1, Flags:0x0
             IpAddr:192.168.34.1, Flags:0x0
             IpAddr:192.168.13.1, Flags:0x0
ERO        - IPv4Prefix 192.168.46.2/32, Strict
             IPv4Prefix 192.168.68.2/32, Strict
             IPv4Prefix 192.168.58.1/32, Strict
             IPv4Prefix 192.0.2.10/32, Loose
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"

*A:PE-1# show router mpls lsp "LSP-PE-1-PE-10" path detail

---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-10 Path dyn
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-10
From              : 192.0.2.1            To                   : 192.0.2.10
Admin State       : Up                   Oper State           : Up
Path Name         : dyn
Path LSP ID       : 4618                 Path Type            : Primary
Path Admin        : Up                   Path Oper            : Up
---snip---

FR Prop Admin Gr*: Disabled              Oper FRPropAdmGrp    : Disabled
Propogate Adm Grp: Enabled               Oper Prop Adm Grp    : Enabled
Inter-area        : True
---snip---
Include Groups    :                      Oper Include Groups  :
None                                            None
Exclude Groups    :                      Oper Exclude Groups  :
red                                             red
---snip---
Actual Hops       :
    192.168.13.1 (192.0.2.1) @ n                 Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @ n                 Record Label        : 524286
 -> 192.168.34.2 (192.0.2.4) @ n                 Record Label        : 524286
 -> 192.168.46.2  @ n                            Record Label        : 524287
 -> 192.0.2.8 (192.0.2.8) @ n                    Record Label        : 524286
 -> 192.168.68.2  @ n                            Record Label        : 524286
 -> 192.0.2.5 (192.0.2.5) @                      Record Label        : 524286
 -> 192.168.58.1  @                              Record Label        : 524286
 -> 192.0.2.10 (192.0.2.10)                      Record Label        : 524284
 -> 192.168.105.2                                Record Label        : 524284
---snip---
```

Shared Risk Link Groups (SRLG)

Shared risk link groups are described in chapter Shared Risk Link Groups for RSVP-Based LSP.

SRLGs are also supported in the context of inter-area TE LSPs. SRLGs refer to situations where links in a network share a common fiber (or a common physical attribute). If one link fails, other links in the group may fail as well. Links in the group have fate sharing.

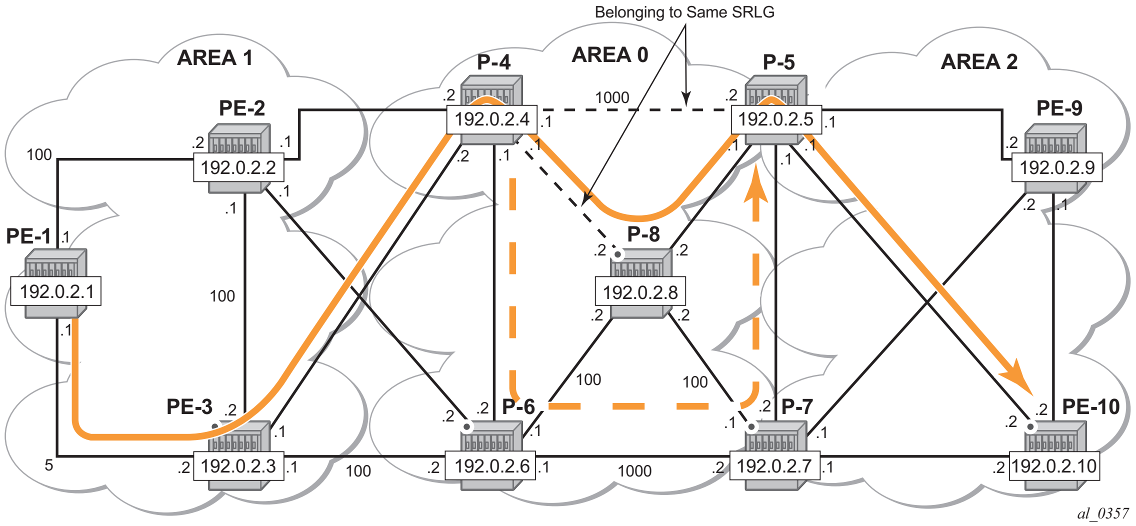
The MPLS TE SRLG feature enhances backup tunnel path selection so that a backup tunnel avoids using links that are in the same SRLG.

Consider the setup in Figure 238, where an inter-area LSP is set up from PE-1 to PE-10 and the path goes through P-8 because of a lower IGP metric. To protect against a node failure of P-8, P-4 (PLR) would normally set up an FRR backup directly to P-5 (MP), because of the lower IGP metric (P-4 to P-5:1000) compared to the IGP traffic via P-6 (P-4 to P-6 to P-7 to P-5:1020).

However, imagine that in this setup the link between P-4 and P-5 link and the link between P-4 and P-8 are part of the same transmission bundle. In this case, a cut of that fiber bundle will bring down both the primary and the backup path.

This can be avoided by configuring these two links in the same SRLG group and enabling **srlg-frr strict** on P-4. In that case, the backup will be set up via P-6 as indicated by the dashed line in Figure 238.

Figure 238 **Share Risk Link Groups**



SRLG Configuration

On P-4, an SRLG group is configured, **srlg-frr strict** is enabled and the links to P-5 and to P-8 are added to this SRLG group.

The SRLG group configuration is required on all nodes that use SRLG groups and on the ABR used by the inter-area TE LSP. In this example, it is configured on all nodes.

Note: Enabling or disabling SRLG for FRR is system-wide and requires the MPLS routing instance to be manually set to shutdown and then to no shutdown to activate the change. This may cause service outage. Nokia recommends that the operator incorporates the SRLG into the initial network design and implementation to minimize the traffic loss. In this case, it is sufficient to disable and re-enable RSVP on PE-1.

```
*A:Px# configure router if-attribute srlg-group bundle-red value 1
```

```
*A:P-4# configure
    router
        mpls
            srlg-frr strict
            interface "int-P-4-P-5"
                srlg-group "bundle-red"
            exit
            interface "int-P-4-P-8"
                srlg-group "bundle-red"
            exit
```

```
*A:PE-1# configure router rsvp shutdown
*A:PE-1# configure router rsvp no shutdown
```

LSP Configuration

Remove the admin group restriction from the LSP.

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-10"
                no exclude "red"
                no propagate-admin-group
            exit
```

Now check the LSP path on PE-1 and verify that FRR protection is in place.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-10"  path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-10 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-10 Path dyn
```

```
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-10
From              : 192.0.2.1               To                    : 192.0.2.10
Admin State       : Up                      Oper State            : Up
Path Name         : dyn
Path LSP ID       : 4628                    Path Type             : Primary
Path Admin        : Up                      Path Oper             : Up
---snip---

CSPF              : Enabled                 Oper CSPF             : Enabled
Least Fill        : Disabled                Oper LeastFill        : Disabled
FRR               : Enabled                 Oper FRR              : Enabled
FRR NodeProtect   : Enabled                 Oper FRR NP           : Enabled
FR Hop Limit      : 16                      Oper FRHopLimit       : 16
FR Prop Admin Gr*: Disabled                 Oper FRPropAdmGrp     : Disabled
Propogate Adm Grp: Disabled                 Oper Prop Adm Grp     : Disabled
Inter-area        : True

---snip---
Actual Hops       :
    192.168.13.1 (192.0.2.1) @ n                     Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @ n                     Record Label        : 524287
 -> 192.168.34.2 (192.0.2.4) @ n                     Record Label        : 524287
 -> 192.0.2.8 (192.0.2.8) @ n                        Record Label        : 524287
 -> 192.168.48.2  @ n                                Record Label        : 524287
 -> 192.0.2.5 (192.0.2.5) @                          Record Label        : 524287
 -> 192.168.58.1  @                                  Record Label        : 524287
 -> 192.0.2.10 (192.0.2.10)                          Record Label        : 524287
 -> 192.168.105.2                                    Record Label        : 524287
Computed Hops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.34.2(SA)
 -> 192.0.2.10(L)
Resignal Eligible: False
Last Resignal     : n/a                     CSPF Metric           : 15
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

On P-4, the SRLG configuration is checked as follows:

```
*A:P-4# show router if-attribute srlg-group

======================================================================
Interface Srlg Groups
======================================================================
Group Name                      Group Value      Penalty Weight
----------------------------------------------------------------------
bundle-red                      1                0
----------------------------------------------------------------------
No. of Groups: 1
======================================================================
*A:P-4#


*A:P-4# show router mpls interface

===============================================================================
MPLS Interfaces
===============================================================================
Interface                           Port-id          Adm    Opr    TE-metric
-------------------------------------------------------------------------------
system                              system           Up     Up     None
  Admin Groups                      None
  SRLG Groups                       None
int-P-4-P-5                         1/1/1            Up     Up     None
  Admin Groups                      None
  SRLG Groups                       bundle-red
int-P-4-P-6                         1/1/3            Up     Up     None
  Admin Groups                      None
  SRLG Groups                       None
int-P-4-P-8                         1/2/1            Up     Up     None
  Admin Groups                      red
  SRLG Groups                       bundle-red
int-P-4-PE-2                        1/1/2            Up     Up     None
  Admin Groups                      None
  SRLG Groups                       None
int-P-4-PE-3                        1/1/4            Up     Up     None
  Admin Groups                      None
  SRLG Groups                       None
-------------------------------------------------------------------------------
Interfaces : 6
===============================================================================
*A:P-4#
```

On PE-4, it is verified that the bypass tunnel is set up via P-6 rather than via P-5, as follows:

```
*A:P-4# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-node192.0.2.8-61447
-------------------------------------------------------------------------------
To             : 192.168.57.1        State              : Up
Out I/F        : 1/1/3               Out Label          : 524284
Up Time        : 0d 00:01:17         Active Time        : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 2
Type           : Dynamic             Bypass Path Cost   : 1020
Setup Priority : 7                   Hold Priority      : 0
Class Type     : 0
Exclude Node   : None                Inter-Area         : False
Computed Hops  :
    192.168.46.1(S)                  Egress Admin Groups : None
 -> 192.168.46.2(S)                  Egress Admin Groups : None
 -> 192.168.67.2(S)                  Egress Admin Groups : None
 -> 192.168.57.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.46.1 (192.0.2.4)         Record Label       : N/A
 -> 192.168.46.2 (192.0.2.6)         Record Label       : 524284
 -> 192.168.67.2 (192.0.2.7)         Record Label       : 524287
 -> 192.168.57.1 (192.0.2.5)         Record Label       : 524282
Last Resignal  :
Attempted At   : n/a                 Resignal Reason    : n/a
```

```
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-10::dyn
From           : 192.0.2.1           To                  : 192.0.2.10
Avoid Node/Hop : 192.0.2.8           Downstream Label    : 524287
Bandwidth      : 0 Kbps

===============================================================================
*A:P-4#
```

Conclusion

Inter-area TE P2P LSPs can be set up based on ERO expansion. With this feature, the head-end does a partial CSPF calculation to its local ABR. On receiving a PATH message with a loose hop ERO, this ABR does a partial or full CSPF calculation to the next ABR to reach the final destination.

FRR protection within the area is available. FRR node protection of the ABR is possible through an MBT on the PLR (node just upstream of the ABR) to the MP (node just downstream of the ABR) or through a dynamically signaled bypass tunnel on the PLR. Dynamic ABR node protection requires that the node ID of the MP node is propagated in the RESV message and that an XRO object is included in the bypass PATH message which makes it possible for the ABR to calculate a path to an MP node.

TE features such as BW, path prioritization, path pre-emption, and graceful shutdown are supported, as well as propagation of the session attribute with affinity along the LSP path (admin groups) and SRLG.

LDP FEC to BGP Label Route Stitching

This chapter provides information about LDP FEC to BGP label route stitching.

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS routers and was initially written for SR OS Release 13.0.R7. The CLI in the current edition corresponds to SR OS Release 14.0.R1. Label Distribution Protocol (LDP) Forwarding Equivalence Class (FEC) to Border Gateway Protocol (BGP) label route stitching was first implemented in SR OS release 8.0.

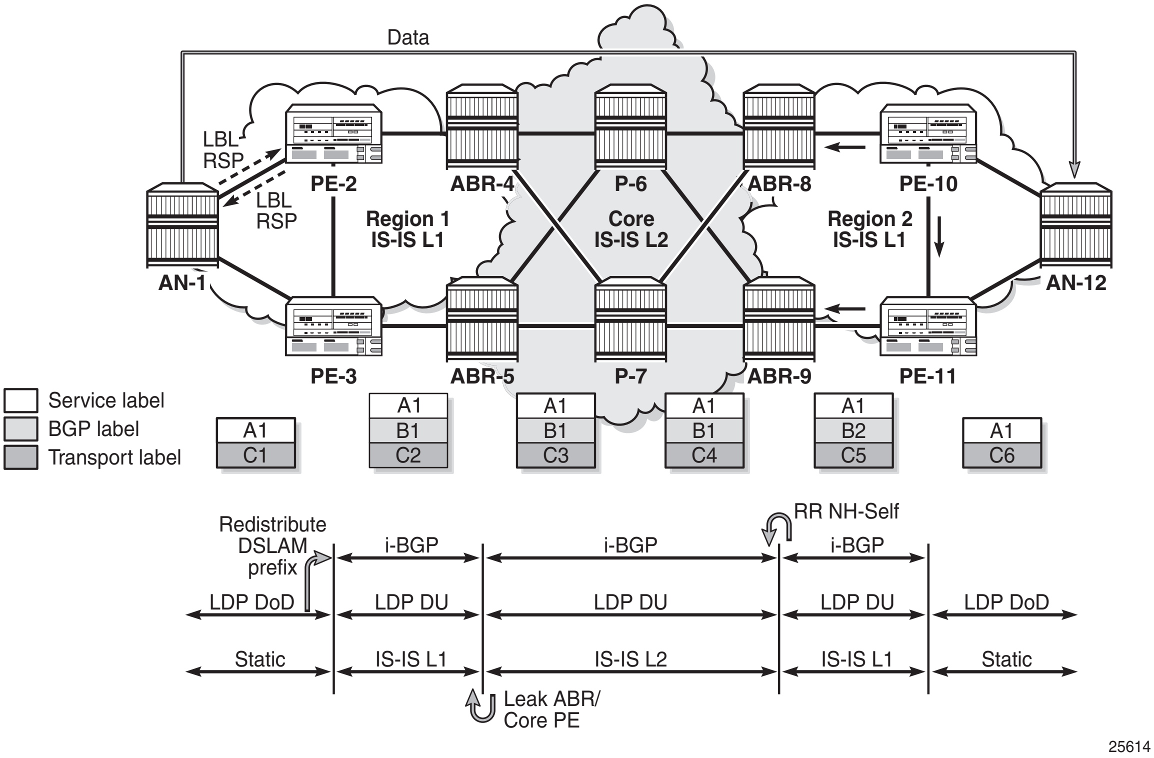
Overview

Stitching of an LDP FEC to a BGP labeled route allows LDP-capable PE devices, such as Digital Subscriber Line Access Multiplexers (DSLAMs), to offer services to LDP-capable PE devices in other areas or domains without the need to support BGP labeled routes. This feature is used in a large network to provide services across multiple areas or Autonomous Systems (ASs).

When BGP is used to distribute a particular route, it can at the same time be used to distribute a Multi-Protocol Label Switching (MPLS) label that is mapped to that route. The label mapping information for a particular route is appended to the same BGP update message that is used to distribute the route. This is described in RFC 3701, *Carrying Label Information in BGPv4*.

Figure 239 shows a network with a core area and regional areas. The components of the network are defined in the paragraphs that follow. For simplification, the control plane is displayed from right to left and the data plane from left to right.

Figure 239 **LDP FEC to BGP Label Route Stitching**



The Access Nodes (ANs) are DSLAMs that support LDP. In seamless MPLS networks, LDP Downstream-on-Demand (DoD) label advertisement can be used between the ANs and their next-hop PEs. Usually, MPLS routers implement LDP Downstream Unsolicited (DU) label distribution, advertising MPLS labels for all routes in their Routing Information Base (RIB). The ANs do not need to have LDP bindings for all prefixes in the network. The ANs will request the LDP labels they need. LDP DoD improves scalability in large networks.

BGP Route Reflectors (RRs) can be used to improve scalability. The RR can be any node; it does not need to be an Area Border Router (ABR) as in Figure 239. If the RR is not in the forwarding path, it does not need to be capable of forwarding MPLS packets.

There are different areas for IS-IS: routers in the core network have level 2 (L2) capability, whereas the routers in the regional areas have level 1 (L1) capability and the ABRs have both. In each ABR, an IS-IS export policy is defined to leak the routes from the core to the regional networks.

Passing L1 routes (regional) into L2 (core) is inherent to IS-IS and cannot be controlled through policy. Passing L2 routes to L1 can be controlled through policy.

Only nodes within a regional area, and the ABR nodes in the same area, exchange LDP FECs. PE routers in a regional area learn the reachability of PE routers in other regional areas by way of RFC 3107 BGP labeled routes redistributed by the remote nodes.

The label stack contains three labels for packets sent in an Epipe service between the access nodes:

- The DSLAMs push a service label to the packets sent in the Epipe service. The service label remains unchanged end-to-end between the DSLAMs. The service label is popped by the remote DSLAM and is the inner label of the label stack.
- The BGP label is the middle label of the label stack and should be regarded as a transport label. The transport label stack contains two labels: BGP and LDP transport label. BGP labeled routes are not supported on the DSLAMs. The BGP label is pushed by the PE nearest to the local DSLAM and is swapped at the BGP next hop, which can be a BGP peer configured with next-hop-self or the PE that is the remote endpoint of the BGP tunnel. The BGP label is popped by the PE at the end of the BGP tunnel.
- The DSLAMs push an LDP transport label to the packets sent to the remote DSLAM. At the PE nearest to the local DSLAM, the LDP transport label is stitched to the BGP label. At the same time, that same PE pushes the LDP transport label to reach the BGP next hop. The LDP transport label is swapped in every Label Switching Router (LSR) and popped by the PE nearest to the remote DSLAM. That PE also pops the BGP label, which is stitched to the LDP transport label that is pushed to the packets sent to the remote DSLAM. This LDP label is the top label of the label stack.

When PE-2 is an ingress Label Edge Router (iLER) sending a service packet to the remote PE, PE-2 inserts the BGP route label to reach the remote PE and an LDP label to reach the next-hop router. In Figure 239, this is the remote ABR because it has set next-hop-self (NH-Self).

The access node AN-1, which is a DSLAM, can behave as a PE router for Epipe services. It will need to establish a pseudowire (PW) to a PE in a different regional area via LSR PE-2. In this case, PE-2 performs the following actions:

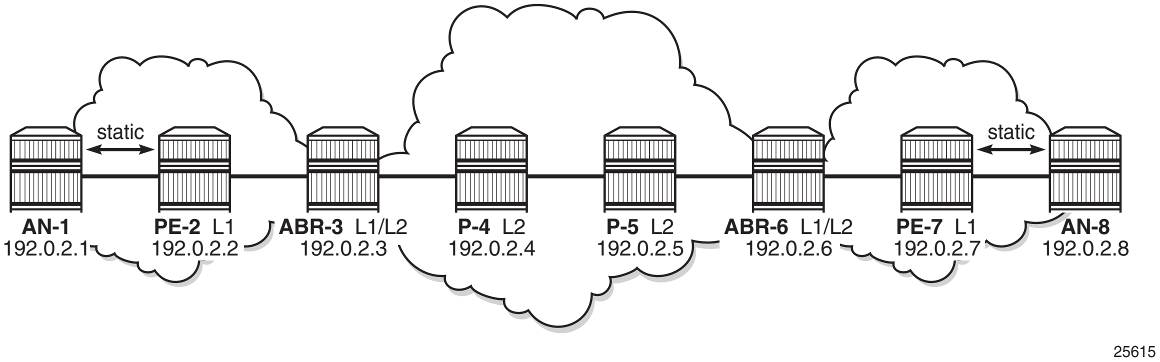
- Translates the LDP FEC it learned from AN-1 into a BGP labeled route and redistributes it using iBGP within its area. This is in addition to redistributing the FEC to its LDP neighbors in the same area.
- Translates the BGP labeled routes it learns through iBGP into an LDP FEC and redistributes it to its LDP neighbors in the same area. AN-1 requests the LDP FEC of the remote DSLAM (AN-12) using LDP DoD.

- When a data packet is received from AN-1 with destination AN-12, PE-2 swaps the LDP label into a BGP label and pushes the LDP label to reach the BGP next hop. When a data packet with destination AN-1 is received on PE-2 from the local ABR (ABR-4), the top transport label (LDP) is removed and the BGP label is swapped for the LDP label corresponding to AN-1.

Configuration

Figure 240 shows the example topology that is used in this section. An Epipe will be established between the access nodes AN-1 and AN-8. PE-2 and PE-7 will stitch the LDP FECs to BGP label routes. In the regional areas, IS-IS L1 capability is used whereas in the core area, IS-IS L2 is used. The ABR nodes support both IS-IS L1 and L2 and export routes from L2 to L1. Static routes are configured between the access nodes and the next-hop PEs.

Figure 240 **Example Topology**



Initial Configuration

Note: In the example topology, all nodes are 7750 SRs, while the ANs should be access devices, such as DSLAMs. The limitation of this approach is that the ANs (SRs) in this setup can only request a label for the directly connected PE and not for their remote peer AN; however, DSLAMs do not have this limitation. Consequently, the Epipe service in this configuration will be operationally down because the transport tunnel is down.

All nodes have the following initial configuration:

- Cards, media dependent adapters (MDAs), ports

- Router interfaces

Note: The IP addresses for the link between node A and node B are in subnet 192.168.AB.0/0. The node with the lowest ID has IP address 192.168.AB.1/30 and the node with the highest ID has IP address 192.168.AB.2/30.

```
*A:PE-2# configure
    router
        interface "int-PE-2-ABR-3"
            address 192.168.23.1/30
            port 1/1/1
        exit
        interface "int-PE-2-AN-1"
            address 192.168.12.2/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.2/32
        exit
```

- Static routes are configured between AN-1 and PE-2 and between PE-7 and AN-8:

```
*A:AN-1# configure
    router
        static-route-entry 0.0.0.0/0
            next-hop 192.168.12.2
                no shutdown
            exit
        exit
```

```
*A:PE-2# configure
    router
        static-route-entry 192.0.2.1/32
            next-hop 192.168.12.1
                no shutdown
            exit
        exit
```

- IS-IS (alternatively, OSPF could have been used)
    - PE-2 and PE-7 have L1 capability.

```
*A:PE-2# configure
    router
        isis
            level-capability level-1
            area-id 49.0001
            interface "system"
            exit
            interface "int-PE-2-ABR-3"
                interface-type point-to-point
            exit
            no shutdown
        exit
```

  - P-4 and P-5 have L2 capability.
  - ABR-3 and ABR-6 have L1 capability on the interfaces toward the PE routers in the regional areas and L2 capability on the interfaces toward the P routers in the core area. A policy is applied to export the system IP addresses from L2 to L1:

```
*A:ABR-3# configure
    router
        policy-options
            begin
            prefix-list "systemIP"
                prefix 192.0.2.0/24 longer
            exit
            policy-statement "export-L2-to-L1"
                entry 10
                    from
                        protocol isis
                        prefix-list "systemIP"
                        level 2
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        isis
            area-id 49.0001
            export "export-L2-to-L1"
            interface "system"
            exit
            interface "int-ABR-3-PE-2"
                level-capability level-1
                interface-type point-to-point
            exit
            interface "int-ABR-3-P-4"
                level-capability level-2
                interface-type point-to-point
            exit
        exit
```

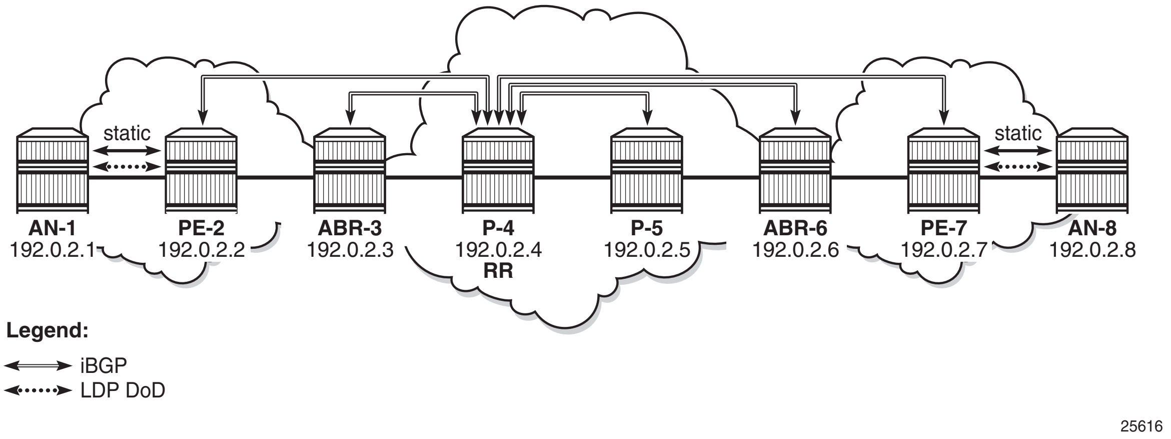
- LDP
  - Link LDP is enabled on all router interfaces on all nodes, including the ANs.
  - On PE-2 and PE-7, DoD is enabled in the session parameters for the peering sessions with the ANs:

```
*A:PE-2# configure
    router
        ldp
            session-parameters
                peer 192.0.2.1
                    dod-label-distribution
                exit
            exit
            interface-parameters
                interface "int-PE-2-AN-1
                exit
                interface "int-PE-2-ABR-3"
                exit
            exit
```

Configure BGP

BGP is configured on all nodes except the ANs. Figure 241 shows that P-4 is the RR.

Figure 241 **BGP Enabled with P-4 as RR**



The initial BGP configuration on PE-2 is the following:

```
*A:PE-2# configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                peer-as 64496
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
        exit
```

The configuration is identical for ABR-3, P-5, ABR-6, and PE-7. The initial BGP configuration on the RR P-4 is:

```
*A:P-4# configure
    router
        autonomous-system 64496
        bgp
            cluster 1.1.1.1
            group "internal"
                peer-as 64496
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.5
                exit
                neighbor 192.0.2.6
                exit
                neighbor 192.0.2.7
                exit
            exit
            no shutdown
        exit
```

This BGP configuration is incomplete: for labeled IPv4 BGP peering sessions, an additional address family will be configured on PE-2 and PE-7, as well as on RR P-4 for neighbors PE-2 and PE-7. The configuration is shown in the following section. The prefixes for AN-1 and AN-8 will be advertised in the labeled IPv4 BGP sessions only, not in IPv4 BGP sessions.

Export Policies for BGP and LDP

LDP FEC to BGP label route stitching is established by configuring separate tunnel table route export policies in both protocols. At the local next-hop PE, the LDP FEC of the local AN must be translated into a BGP label and at the remote PE, the BGP label must be translated into an LDP FEC.

An export policy for the export from LDP to BGP must be defined on the PE nodes.

```
*A:PE-2# configure
    router
        policy-options
            begin
            prefix-list "local-AN"
                prefix 192.0.2.1/32 exact
            exit
            prefix-list "remote-AN"
                prefix 192.0.2.8/32 exact
            exit
            policy-statement "export-BGP"
                entry 10
                    from
                        protocol ldp
                        prefix-list "local-AN"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

On PE-7, the policy statement is identical, but the prefix lists are different.

This export policy must be applied in the BGP context: either in the general settings or per group or per neighbor.

```
*A:PE-2# configure
    router
        bgp
            group "internal"
                export "export-BGP"
            exit
```

In a similar way, BGP labels must be exported to LDP on the PE routers. The export policy is configured as follows, with the prefix list already defined earlier:

```
*A:PE-2# configure
    router
        policy-options
            begin
            prefix-list "remote-AN"
                prefix 192.0.2.8/32 exact
            exit
            policy-statement "export-LDP"
                entry 10
                    from
                        protocol bgp-label
                        prefix-list "remote-AN"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

This export policy is applied in the LDP context, as follows:

```
*A:PE-2# configure
    router
        ldp
            export-tunnel-table "export-LDP"
        exit
```

Advertise Labels in BGP Updates

BGP should evaluate the activated /32 LDP prefixes in the export policy. This needs to be configured on the endpoints of the BGP tunnel on PE-2 and PE-7, as follows:

```
*A:PE-2/7# configure
    router
        bgp
            group "internal"
                neighbor 192.0.2.4
```

```
                    family label-ipv4
                    advertise-ldp-prefix
                exit
```

On RR P-4, the family label-ipv4 is enabled and the LDP prefix is advertised toward the clients PE-2 and PE-7, as follows.

```
*A:P-4# configure
    router
        bgp
            group "internal"
                neighbor 192.0.2.2
                    family label-ipv4
                    advertise-ldp-prefix
                exit
                neighbor 192.0.2.7
                    family label-ipv4
                    advertise-ldp-prefix
                exit
            exit
```

Configuring address **family label-ipv4** and the **advertise-ldp-prefix** argument implies that all activated /32 LDP FEC prefixes will be sent to the remote BGP peer as an RFC 3107 formatted label.

Configuring address **family label-ipv4** without the **advertise-ldp-prefix** argument implies that only core IPv4 routes learned from the Route Table Manager (RTM) are advertised as RFC 3107 BGP labeled routes to this neighbor. No stitching of LDP FEC to the BGP labeled route will be performed for this neighbor, even if the same prefix was learned from LDP.

The BGP open messages contain address family AFI=1 and SAFI=1 between the RR and peers for address family IPv4, that is used for IPv4 unicast. See Cap_Code MP-BGP. Bytes 0x0 0x1 (AFI=1) 0x0 0x1 (SAFI=1).

```
*A:ABR-3# show debug
debug
    router "Base"
        bgp
            open
            update
        exit
    exit
exit
```

```
2 2017/03/30 08:02:25.73 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.4 - Received BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.4: Opt Length 16
   Opt Para: Type CAPABILITY: Length = 14: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x1
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf0
"
```

Between peers that advertise the labels, AFI=1 and SAFI=4, the address family is labeled IPv4 unicast. The following BGP open message is seen on PE-2:

```
8 2017/03/30 08:06:24.34 UTC MINOR: DEBUG #2001 Base BGP
"BGP: OPEN
Peer 1: 192.0.2.4 - Received BGP OPEN: Version 4
   AS Num 64496: Holdtime 90: BGP_ID 192.0.2.4: Opt Length 16
   Opt Para: Type CAPABILITY: Length = 14: Data:
     Cap_Code MP-BGP: Length 4
       Bytes: 0x0 0x1 0x0 0x4
     Cap_Code ROUTE-REFRESH: Length 0
     Cap_Code 4-OCTET-ASN: Length 4
       Bytes: 0x0 0x0 0xfb 0xf0
"
```

No BGP update messages are sent to ABR-3. Prefix 192.0.2.8 is advertised as a labeled IPv4 route from PE-7 to P-4 and forwarded by P-4 to its other labeled IPv4 client, PE-2, but it is not sent to BGP IPv4 clients, such as ABR-3.

The BGP update messages between labeled IPv4 peers contain label information, for example, for prefix 192.0.2.8/32. The address family is IPV4-Labeled and the label is 262136. The following BGP update for prefix 192.0.2.8/32 is received on PE-2:

```
9 2017/03/30 08:06:56.26 UTC MINOR: DEBUG #2001 Base Peer 1: 192.0.2.4
"Peer 1: 192.0.2.4: UPDATE
Peer 1: 192.0.2.4 - Received BGP UPDATE:
    Withdrawn Length = 0
    Total Path Attr Length = 63
    Flag: 0x90 Type: 14 Len: 17 Multiprotocol Reachable NLRI:
        Address Family IPV4-Labeled
        NextHop len 4 NextHop 192.0.2.7
        192.0.2.8/32 Label 262136
    Flag: 0x40 Type: 1 Len: 1 Origin: 0
    Flag: 0x40 Type: 2 Len: 0 AS Path:
    Flag: 0x40 Type: 3 Len: 4 Nexthop: 192.0.2.7
    Flag: 0x80 Type: 4 Len: 4 MED: 1
    Flag: 0x40 Type: 5 Len: 4 Local Preference: 100
    Flag: 0x80 Type: 9 Len: 4 Originator ID: 192.0.2.7
    Flag: 0x80 Type: 10 Len: 4 Cluster ID:
        1.1.1.1
"
```

After applying the export policy from BGP to LDP, enabling the address family labeled IPv4 in BGP, and advertising labels for the LDP FEC prefixes, LDP will look for BGP route entries in the tunnel table. If a /32 BGP labeled route matches a prefix entry in the export policy, LDP originates an LDP FEC for this prefix, stitches it to the BGP labeled route, and redistributes the LDP FEC to its BGP neighbors. This can be shown on PE-7, as follows.

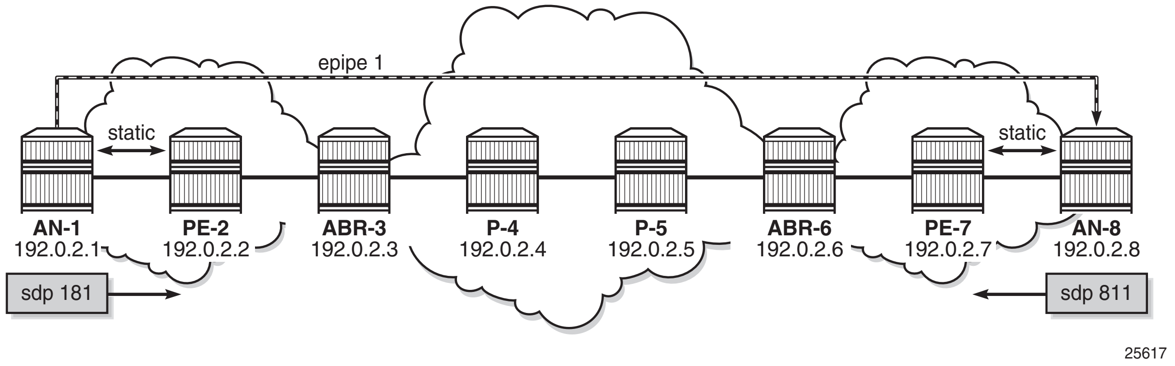
```
*A:PE-7# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.8                     262143         262136         InternalLdp
-------------------------------------------------------------------------------
Total Labels allocated:   1
===============================================================================
*A:PE-7#
```

The label received from AN-8 is 262143. The label origin is **InternalLdp**. This LDP label is stitched to BGP label 262136 that will be advertised by PE-7 to its BGP labeled IPv4 peers: PE-7 advertises to RR P-4 and P-4 advertises this route to PE-2. Traffic sent from AN-1 toward AN-8 will be forwarded from PE-2 to its BGP NH PE-7 using BGP label 262136. In PE-7, the BGP label is stitched to LDP label 262143 that will be used to forward the packet to AN-8.

Configure SDP and Epipe

An end-to-end Epipe service is established between AN-1 and AN-8, as shown in Figure 242.

Figure 242 **End-to-End Epipe Service**



Note: : In this setup, ANs are simulated by 7750 SRs. Due to this limitation, the SDP used by the Epipe service will not become operational. 7750 SR only supports single-hop DoD, which implies that AN-1 can only request a label for the LSR ID of the directly connected router, PE-2, not of remote nodes, such as AN-8. Similarly, AN-8 cannot request a label for AN-1. Therefore, it is not possible to have an LDP LSP between the ANs and the SDP will be down because there is no transport tunnel.

The SDP is configured on AN-1, as follows:

```
*A:AN-1# configure
    service
        sdp 181 mpls create
            far-end 192.0.2.8
            ldp
            no shutdown
        exit
```

Epipe 1 is configured on AN-1, as follows:

```
*A:AN-1# configure
    service
        epipe 1 customer 1 create
            sap 1/2/1:1 create
            exit
            spoke-sdp 181:1 create
            exit
            no shutdown
        exit
```

The configuration of the SDP and Epipe on AN-8 is similar.

The SDP is down because there is no transport tunnel, which can be shown as follows:

```
*A:AN-1# show service sdp detail

===============================================================================
Services: Service Destination Points Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 181  -192.0.2.8
-------------------------------------------------------------------------------
Description           : (Not Specified)
SDP Id               : 181                   SDP Source         : manual
Admin Path MTU       : 0                     Oper Path MTU      : 0
Delivery             : MPLS
Far End              : 192.0.2.8
Tunnel Far End       : 192.0.2.8             LSP Types          : LDP

Admin State          : Up                    Oper State         : Down
Signaling            : TLDP                  Metric             : 0
---snip---
Flags                : TranspTunnDown
```

```
---snip---
```

A targeted LDP session is established between AN-1 and AN-8, which can be shown as follows:

```
*A:AN-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   34389     34348     1d 01:29:46
192.0.2.8:0         Targeted  Established   1967      1969      0d 02:55:36
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
```

LDP FEC Resolution at PE-2 for Traffic from AN-8 to AN-1

The following steps occur at PE-2 for the LDP FEC resolution for traffic from AN-1 toward AN-8. The situation is similar for PE-7.

Step 1. After receiving an LDP label binding message for LDP FEC for the system address of AN-1 (192.0.2.1/32), PE-2 installs this prefix in the Label Forwarding Information Base (LFIB). PE-2 programs a push and a swap Next Hop Label Forwarding Entry (NHLFE) in the egress data path to forward packets to prefix 192.0.2.1/32.

Note: PE-2 installs this LDP FEC in the LFIB only if there is an exact match of the prefix 192.0.2.1/32 in the routing table or a longest match of the prefix in the routing table, in case aggregate-prefix-match is configured on PE-2. The advertising LDP neighbor (AN-1) must be the next hop to reach the FEC prefix.

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262143
192.168.12.1                                1/1/2

192.0.2.1/32                                Swap            262142    262143
192.168.12.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
```

Step 2. PE-2 programs a tunnel entry for prefix 192.0.2.1/32 in the tunnel table.

```
*A:PE-2# show router tunnel-table 192.0.2.1/32

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        192.168.12.1   1
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
```

Step 3. PE-2 advertises a new FEC label binding for prefix 192.0.2.1/32 toward all its LDP neighbors. The result can be shown on ABR-3, as follows:

```
*A:ABR-3# show router ldp bindings prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                      IngLbl                    EgrLbl
Peer                                        EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                  --                      262142
192.0.2.2:0                                   --
  --
```

```
-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 1
===============================================================================
*A:ABR-3#
```

Step 4. When BGP learns the LDP FEC via the tunnel table and the FEC prefix exists in the BGP route policy, PE-2 originates a BGP labeled route toward all its neighbors that have the advertise label for LDP FEC prefixes enabled. The following output shows the BGP labeled route to RR P-4 for prefix 192.0.2.1/32.

```
*A:PE-2# show router bgp routes label-ipv4 hunt

===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
-------------------------------------------------------------------------------
RIB In Entries
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
RIB Out Entries
-------------------------------------------------------------------------------
Network        : 192.0.2.1/32
Nexthop        : 192.0.2.2
Path Id        : None
To             : 192.0.2.4
Res. Nexthop   : n/a
Local Pref.    : 100                    Interface Name : NotAvailable
Aggregator AS  : None                   Aggregator     : None
Atomic Aggr.   : Not Atomic             MED            : 1
AIGP Metric    : None
Connector      : None
Community      : No Community Members
Cluster        : No Cluster Members
Originator Id  : None                   Peer Router Id : 192.0.2.4
IPv4 Label     : 262136
Origin         : IGP
AS-Path        : No As-Path
Route Tag      : 0
Neighbor-AS    : N/A
Orig Validation: NotFound
Source Class   : 0                      Dest Class     : 0

-------------------------------------------------------------------------------
Routes : 3
===============================================================================
*A:PE-2
```

BGP Labeled Route Resolution at PE-2 for Traffic from AN-1 to AN-8

The following steps occur at PE-2 for the BGP labeled route resolution for traffic from AN-1 toward AN-8. The situation is similar for PE-7.

Step 1. When there is an LDP LSP to the BGP neighbor advertising the route (PE-7) and PE-2 has received a BGP labeled route via iBGP for AN-8, PE-2 installs the prefix 192.0.2.8/32 in BGP. The LDP tunnel toward PE-7 is shown, then the BGP labeled IPv4 route toward AN-8, as advertised by PE-7.

```
*A:PE-2# show router tunnel-table 192.0.2.7

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.7/32      ldp       MPLS  65542     9        192.168.23.2   50
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-2#


*A:PE-2# show router bgp routes 192.0.2.8/32 label-ipv4

===============================================================================
 BGP Router ID:192.0.2.2       AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.8/32                                       100         1
      192.0.2.7                                          None        262136
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:PE-2#
```

The BGP label for traffic toward AN-8 is 262136. This is the middle label in the label stack. The next hop is PE-7.

Step 2. PE-2 programs a swap NHLFE in the egress data path to forward packets to 192.0.2.8/32, as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.8/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.8/32(B)                             Swap            262135    262136
192.0.2.7                                   LspId 65542

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 1
===============================================================================
*A:PE-2#
```

The (B) indicates that 192.0.2.8/32 is a BGP next hop. The ingress label is the LDP transport label from AN-1 for prefix 192.0.2.8/32. The LSP ID 65542 corresponds to the LDP LSP toward egress next-hop PE-7, as shown earlier in the tunnel table. The BGP egress label for traffic toward AN-8 is 262136.

Step 3. PE-2 programs a tunnel table entry for 192.0.2.8/32.

```
*A:PE-2# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        192.168.12.1   1
192.0.2.3/32      ldp       MPLS  65538     9        192.168.23.2   10
192.0.2.4/32      ldp       MPLS  65539     9        192.168.23.2   20
192.0.2.5/32      ldp       MPLS  65540     9        192.168.23.2   30
192.0.2.6/32      ldp       MPLS  65541     9        192.168.23.2   40
192.0.2.7/32      ldp       MPLS  65542     9        192.168.23.2   50
192.0.2.8/32      bgp       MPLS  262145    12       192.0.2.7      1000
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
```

```
===============================================================================
*A:PE-2#
```

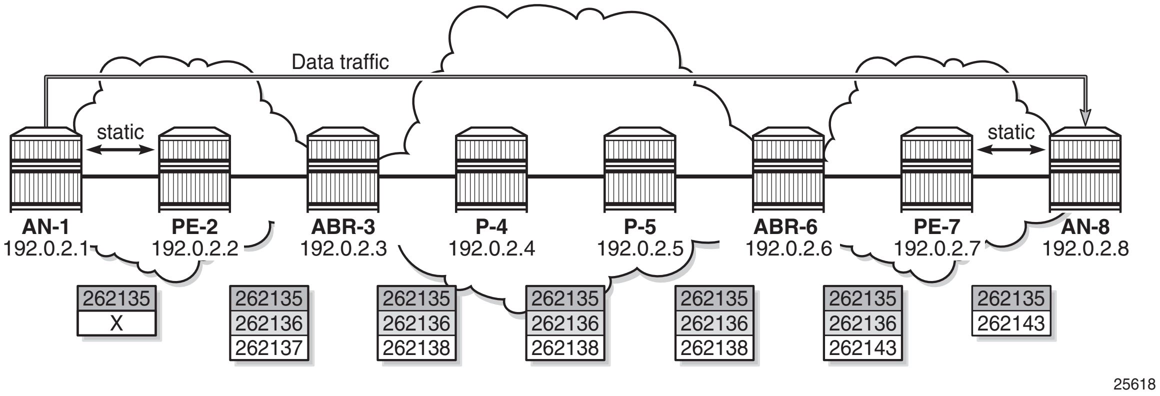
This is the only BGP tunnel in the tunnel table; all tunnels toward the other nodes are LDP tunnels. LDP routes have preference over BGP labeled routes, but there is no LDP route toward 192.0.2.8/32. Therefore, the BGP tunnel will be used for traffic destined to AN-8.

Step 4. PE-2 advertises a new FEC label binding for prefix 192.0.2.8/32 toward AN-1. This is only done after AN-1 requests a label for prefix 192.0.2.8/32, because LDP DoD is enabled. This is possible if the ANs are DSLAMs, but not in this setup with SRs.

Data Plane Overview for PE-2

Figure 243 shows the label stacks that are used for traffic from AN-1 to AN-8.

Figure 243 **Label Stacks for Traffic from AN-1 to AN-8**



Note: The LDP transport label that is pushed by AN-1 is not known because of the single-hop LDP DoD implementation in 7750 SR. AN-1 cannot request the LDP label for AN-8. Therefore, the LDP transport label is represented by "X".

The service label added for Epipe 1 on AN-1 for egress traffic to AN-8 is 262135. Ingress traffic on AN-1 has service label 262135. This can be shown as follows:

```
*A:AN-1# show service id 1 labels

===============================================================================
Martini Service Labels
===============================================================================
Svc Id     Sdp Binding         Type  I.Lbl                  E.Lbl
```

```
-------------------------------------------------------------------------------
1         181:1            Spok  262135              262135
-------------------------------------------------------------------------------
Number of Bound SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:AN-1#
```

This service label remains unchanged end-to-end.

As shown earlier, the (middle) BGP label for traffic with destination AN-8 is 262136, as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.8/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.8/32(B)                             Swap            262135    262136
192.0.2.7                                   LspId 65542

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 1
===============================================================================
*A:PE-2#
```

The next hop is PE-7, which is the PE nearest to AN-8. The BGP label will not be swapped between PE-2 and PE-7 because there is no intermediate node that has set next-hop-self. An intermediate node with next-hop-self would become the next hop instead of PE-7. The BGP label is only added or removed by the next-hop PE.

On PE-2, when a service packet with destination AN-8 is received, the ingress LDP transport label X is swapped into BGP label 262136. To reach PE-7, which is the BGP next hop for traffic toward AN-8, another LDP transport label 262137 is pushed to the packet, as follows:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.7/32
```

```
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.7/32                                Push              --      262137
192.168.23.2                                1/1/1

192.0.2.7/32                                Swap            262137    262137
192.168.23.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-2#
```

The next hop is ABR-3, where the ingress label 262137 is swapped to egress label 262138, as follows:

```
*A:ABR-3# show router ldp bindings active prefixes prefix 192.0.2.7/32
---snip---
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.7/32                                Push              --      262138
192.168.34.2                                1/1/1

192.0.2.7/32                                Swap            262137    262138
192.168.34.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
```

In the subsequent LSRs, the transport label is swapped, as follows:

On P-4:

```
*A:P-4# show router ldp bindings active prefixes prefix 192.0.2.7/32
---snip---
```

```
192.0.2.7/32                                Swap            262138    262138
192.168.45.2                                1/1/1

-------------------------------------------------------------------------------
```

On P-5:

```
*A:P-5# show router ldp bindings active prefixes prefix 192.0.2.7/32
---snip---

192.0.2.7/32                                Swap            262138    262138
192.168.56.2                                1/1/1

-------------------------------------------------------------------------------
```

On ABR-6:

```
*A:ABR-6# show router ldp bindings active prefixes prefix 192.0.2.7/32
---snip---
192.0.2.7/32                                Swap            262138    262143
192.168.67.2                                1/1/1

-------------------------------------------------------------------------------
```

On PE-7, the LDP label 262143 is popped, as follows:

```
*A:PE-7# show router ldp bindings active prefixes prefix 192.0.2.7/32
---anip---
192.0.2.7/32                                Pop             262143      --
  --                                          --

-------------------------------------------------------------------------------
```

The BGP label is also popped and mapped onto LDP label 262143 that will be pushed by PE-7 on packets toward AN-8.

```
*A:PE-7# show router ldp bindings active prefixes prefix 192.0.2.8/32

---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.8/32                                Push              --      262143
192.168.78.2                                1/1/1

192.0.2.8/32                                Swap            262137    262143
192.168.78.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
```

OAM

The following operations, administration, and maintenance (OAM) commands can be launched to validate an LDP FEC stitched to a BGP IPv4 labeled route and vice versa.

```
*A:PE-2# oam lsp-ping bgp-label prefix 192.0.2.8/32
LSP-PING 192.0.2.8/32: 80 bytes MPLS payload
Seq=1, send from intf int-PE-2-ABR-3, reply from 192.0.2.8
       udp-data-len=32 ttl=255 rtt=6.27ms rc=4 (NoFECMapping)

---- LSP 192.0.2.8/32 PING Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 6.27ms, avg = 6.27ms, max = 6.27ms, stddev = 0.000ms
```

In a similar way, LSP trace can validate LDP FEC to BGP label route stitching:

```
*A:PE-2# oam lsp-trace bgp-label prefix 192.0.2.8/32
lsp-trace to 192.0.2.8/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.3  rtt=0.696ms rc=8(DSRtrMatchLabel)
2  192.0.2.4  rtt=3.08ms rc=8(DSRtrMatchLabel)
3  192.0.2.5  rtt=3.33ms rc=8(DSRtrMatchLabel)
4  192.0.2.6  rtt=4.78ms rc=8(DSRtrMatchLabel)
5  192.0.2.7  rtt=5.76ms rc=8(DSRtrMatchLabel) rsc=1
6  192.0.2.8  rtt=6.38ms rc=4(NoFECMapping) rsc=1
```

The detailed output includes the BGP label to LDP label mapping information at the PE:

```
*A:PE-2# oam lsp-trace bgp-label prefix 192.0.2.8/32 detail
lsp-trace to 192.0.2.8/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.3  rtt=1.52ms rc=8(DSRtrMatchLabel)
2  192.0.2.4  rtt=0.955ms rc=8(DSRtrMatchLabel)
3  192.0.2.5  rtt=1.52ms rc=8(DSRtrMatchLabel)
4  192.0.2.6  rtt=3.55ms rc=8(DSRtrMatchLabel)
5  192.0.2.7  rtt=4.47ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.78.2 ifaddr=192.168.78.2 iftype=ipv4Numbered MRU=1560
           label[1]=262143 protocol=3(LDP)
6  192.0.2.8  rtt=5.37ms rc=4(NoFECMapping) rsc=1
```

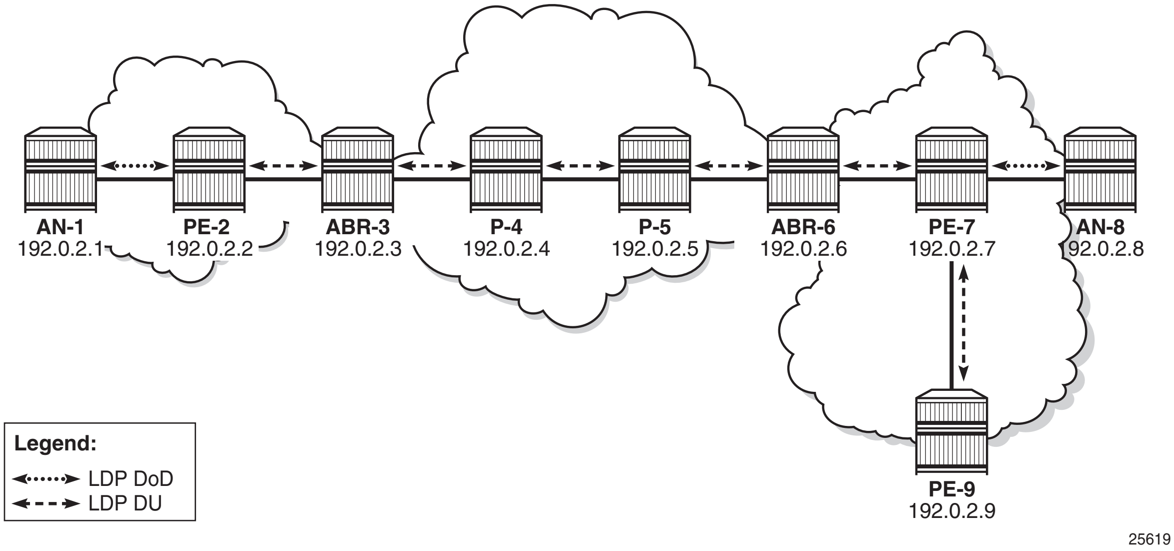
Block BGP Label Bindings to LDP DU Peers

On a PE, labeled BGP prefixes are exported to LDP to allow LDP DoD peers to request these labels. LDP DU peers will also get all labeled BGP prefixes if not explicitly blocked by an LDP export policy, based on prefix lists. This can result in a high administrative and operational effort in large networks.

Blocking BGP label bindings to LDP DU peers is less labor-intensive because per-peer export policies are re-evaluated on NH type change (such as from BGP to LDP or to "unresolved state"), not only on a configuration change.

Figure 244 shows the extended topology used for this configuration. The additional PE router, PE-9, does not need to know the BGP labeled prefixes. LDP DU is used between PE-7 and PE-9.

Figure 244 **Block BGP Label Bindings to LDP DU Peer PE-9**



Blocking BGP label bindings to LDP DU peers can be achieved in two ways:

1. LDP export policy based on prefix list.
2. LDP export policy based on BGP NH type change. No prefix list is required.

To compare the two, both are described.

LDP Export Policy Based on Prefix List

Before applying the policy to block BGP label bindings from PE-7 to PE-9, the LDP bindings on PE-9 for prefix 192.0.2.1 are the following:

```
*A:PE-9# show router ldp bindings prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.9)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                      IngLbl                    EgrLbl
Peer                                        EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                  --                      262135
192.0.2.7:0                                   --
  --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 1
===============================================================================
*A:PE-9#
```

The following policy created on PE-7 is based on a prefix list that only contains the system address of the remote AN: 192.0.2.1.

```
*A:PE-7# configure
    router
        policy-options
            begin
            prefix-list "remote-AN"
                prefix 192.0.2.1/32 exact
            exit
            policy-statement "block-BGP-bindings-remote-AN"
                entry 10
                    from
                        prefix-list "remote-AN"
                    exit
                    action drop
                    exit
                exit
            exit
            commit
```

The policy is applied on PE-7 in the LDP session-parameters context for peer 192.0.2.9.

```
*A:PE-7# configure
    router
        ldp
            session-parameters
                peer 192.0.2.9
                    export-prefixes "block-BGP-bindings-remote-AN"
                exit
            exit
        exit
```

After the policy is applied, there are no LDP bindings for prefix 192.0.2.1 on PE-9:

```
*A:PE-9# show router ldp bindings prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.9)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                        IngLbl                    EgrLbl
Peer                                          EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-9#
```

The original situation is restored by removing the export prefixes in the LDP session-parameters context on PE-7.

```
*A:PE-7# configure router ldp session-parameters peer 192.0.2.9 no export-prefixes
```

```
*A:PE-9# show router ldp bindings prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.9)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                        IngLbl                    EgrLbl
Peer                                          EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                    --                      262135
192.0.2.7:0                                     --
  --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 1
===============================================================================
```

LDP Export Policy Based on BGP NH Type Change

The "from protocol bgp" argument will have a different meaning in the context of per-peer and targeted export policies. For those types of policies, policies are re-evaluated on NH type change; for example, from BGP to LDP or from LDP to "unresolved state". This requires less configuration because no prefix list needs to be specified. The following policy is configured on PE-7.

```
*A:PE-7# configure
    router
        policy-options
            begin
            policy-statement "block-BGP-to-LDP-DU"
                entry 10
                    from
                        protocol bgp
                    exit
                    action drop
                    exit
                exit
            exit
            commit
        exit
```

The policy is applied in the LDP session-parameter context for peer 192.0.2.9.

```
*A:PE-7# configure
    router
        ldp
            session-parameters
                peer 192.0.2.9
                    export-prefixes "block-BGP-to-LDP-DU"
                exit
            exit
```

PE-7 will not send BGP label mapping information for prefix 192.0.2.1/32 to PE-9, or for any other prefix of a remote AN. In this example, AN-1 with prefix 192.0.2.1/32 is the only remote AN for PE-7.

```
*A:PE-9# show router ldp bindings prefixes prefix 192.0.2.1/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.9)
             (IPv6 LSR ID ::)
===============================================================================
Legend: U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        LF - Lower FEC, UF - Upper FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                      IngLbl                    EgrLbl
Peer                                        EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
```

```
No Matching Entries Found
===============================================================================
*A:PE-9#
```

Conclusion

LDP FEC to BGP label route stitching allows LDP-capable PE devices, such as DSLAMs, to offer services to LDP-capable PE devices in other areas or domains without the need to support BGP labeled routes. This feature can be used in a seamless MPLS environment.

LDP over RSVP Using OSPF as IGP

This chapter provides information about label distribution protocol (LDP) over resource reservation protocol with traffic engineering (RSVP-TE), also called LDPoRSVP, that uses RSVP label switched paths (LSPs) as a transport vehicle to carry the packets using LDP LSPs.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Additional Topics
- Conclusion

Applicability

This chapter was initially written for SR OS release7.0.R5, but the CLI in this edition corresponds to release 16.0.R3. There are no prerequisites.

Overview

Introduction

Only user packets are tunneled over RSVP LSPs; targeted LDP (T-LDP) control messages are still sent unlabeled using the interior gateway protocol (IGP) shortest path. Since LDP does not have traffic engineering (TE), it can now benefit from the RSVP-TE features. LDP FRR is loopfree alternate (LFA), but with LDPoRSVP, it can use RSVP FRR detour or bypass tunnels.

The main advantage of LDPoRSVP is seen in large networks. A full mesh of intra-area RSVP LSPs between PE nodes (which in some cases is not scalable) is not needed anymore. While a label edge router (LER) may not have that many tunnels, any transit node will potentially have thousands of LSPs, and if each transit node also has to deal with detour tunnels or bypass tunnels, this number can make the label switching router (LSR) overly burdened.

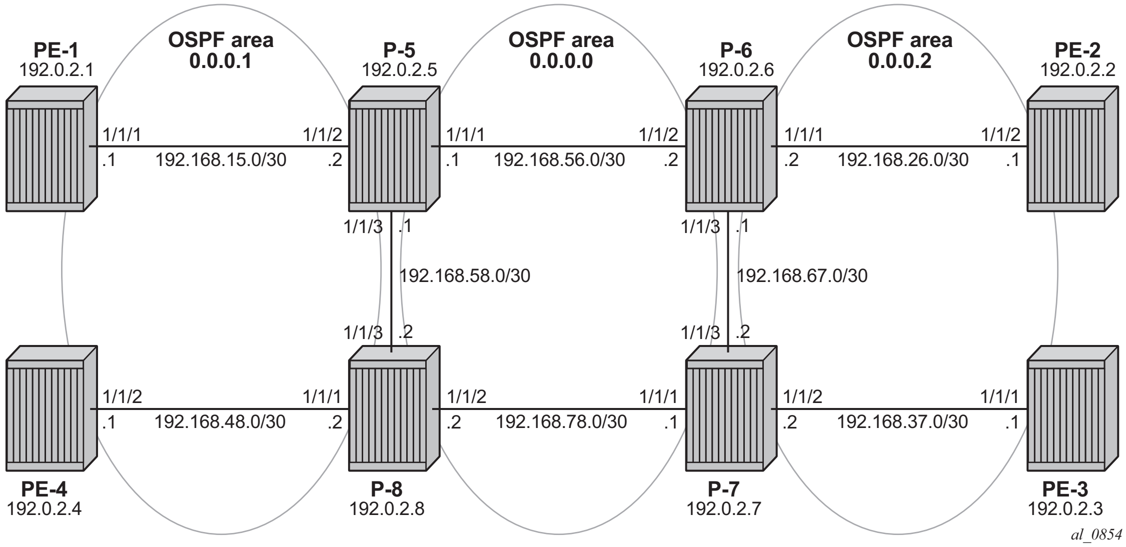
LDPoRSVP can be configured in an intra-area domain and an inter-area domain. Any router in an area can be a stitching point for LDP over RSVP. LDPoRSVP introduces a tunnel-in-tunnel tunnel type (in addition to the existing LDP tunnel type and RSVP tunnel type). If multiple tunnel types match the destination PE forwarding equivalence class (FEC) lookup, LDP will prefer an LDP tunnel over an LDPoRSVP tunnel by default.

First, it is important to understand how LDP FEC resolution is working (with LDPoRSVP enabled). A more detailed description can be found later on in this chapter. The ingress LER receives an LDP label message including a FEC with prefix **P** and label **L** from a peer by a T-LDP session. LDP tries to resolve prefix **P** by performing a lookup in the Routing Table Manager (RTM). The result of this is a Next Hop (NH) to the destination PE, either an intra-area PE (intra-area context) or an Area Border Router (ABR) (inter-area context). When the NH matches the targeted LDP peer, LDP performs a second lookup for that NH in the tunnel table manager (TTM) which returns a user configured RSVP LSP with the best metric. If there are multiple configured RSVP LSPs with the best metric, LDP selects the first available RSVP LSP. If all user configured RSVP LSPs are down, no more action is taken. If the user did not configure any RSVP LSPs under the T-LDP context, the lookup in TTM will return the first available RSVP LSP which terminates on the ABR (inter-area) or intra-area PE with the lowest metric.

If the lookup in TTM results in no RSVP LSP, the system can fall back to link-level LDP (iLDP). In that way, it is possible that the NH is reachable using iLDP. Accordingly, the egress label will be installed on the ingress LER.

Figure 245 shows the example topology with four PE routers and four P routers.

Figure 245 **Initial Example Topology**



OSPF area 0.0.0.1 and OSPF area 0.0.0.2 are two metro areas, connected to each other via a core area, represented by OSPF backbone area (area 0.0.0.0). Therefore, P-5, P-6, P-7, and P-8 are all acting as area border routers (ABRs). LDPoRSVP principles will be shown for intra-area PE communication (PE-1 <=> PE-4) and inter-area communication (PE-1 <=> PE-2).

Configuration

Step 1. Configuring the IP/MPLS network.

The system addresses and IP interface addresses are configured according to Figure 245. An interior gateway protocol (IGP) is needed to distribute routing information on all routers. In this case, the IGP is Open Shortest Path First (OSPF) using the backbone area 0.0.0.0 in the core and normal areas (area 0.0.0.1 and area 0.0.0.2) in the two metro regions, connected toward the backbone area via ABRs. A configuration example is shown for PE-1 and P-5. A similar configuration can be derived for the other P and PE nodes.

```
*A:PE-1# configure
    router
        ospf
            traffic-engineering
            area 0.0.0.1
                interface "system"
                exit
                interface "int-PE-1-P-5"
                    interface-type point-to-point
                exit
            exit


*A:P-5# configure
    router
        ospf
            traffic-engineering
            area 0.0.0.0
                interface "system"
                exit
                interface "int-P-5-P-6"
                    interface-type point-to-point
                exit
                interface "int-P-5-P-8"
                    interface-type point-to-point
                exit
            exit
            area 0.0.0.1
                interface "int-P-5-PE-1"
                    interface-type point-to-point
                exit
            exit
```

Because Fast Reroute (FRR) will be enabled on the RSVP LSPs in the core area, Traffic Engineering (TE) is needed on the IGP. By doing this, OSPF will generate opaque link state advertisements (LSAs) which are collected in a Traffic Engineering Database (TED), separate from the traditional OSPF topology database. OSPF interfaces are set up as type point-to-point to improve convergence, no Designated Router/Backup Designated Router (DR/BDR) election process is performed. Convergence is beyond the scope of this chapter.

On all nodes originating and terminating a T-LDP session, an explicit **ldp-over-rsvp** parameter must be configured to enable this OSPF instance for LDPoRSVP, as follows:

```
A:PE-[1..4]# configure router ospf ldp-over-rsvp
A:P-[5..8]# configure router ospf ldp-over-rsvp
```

To verify that OSPF neighbors are up (state:full), the **show router ospf neighbor** command is executed. To check if IP interface addresses/ subnets are known on all PEs, **show router route-table** or **show router fib** *IOM-card-slot* will display the content of the forwarding information base (FIB).

```
*A:PE-1# show router ospf neighbor

===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ   TTL
 Area-Id
-------------------------------------------------------------------------------
int-PE-1-P-5                     192.0.2.5       Full       1    0       38
 0.0.0.1
-------------------------------------------------------------------------------
No. of Neighbors: 1
===============================================================================
*A:PE-1#


*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     01h10m49s  0
       system                                                       0
192.0.2.2/32                                  Remote  OSPF      00h01m01s  10
       192.168.15.2                                                30
192.0.2.3/32                                  Remote  OSPF      00h00m48s  10
       192.168.15.2                                                40
192.0.2.4/32                                  Remote  OSPF      00h00m33s  10
       192.168.15.2                                                30
192.0.2.5/32                                  Remote  OSPF      00h01m19s  10
       192.168.15.2                                                10
```

```
192.0.2.6/32                                  Remote  OSPF      00h01m13s  10
       192.168.15.2                                                 20
192.0.2.7/32                                  Remote  OSPF      00h00m48s  10
       192.168.15.2                                                 30
192.0.2.8/32                                  Remote  OSPF      00h00m33s  10
       192.168.15.2                                                 20
192.168.15.0/30                               Local   Local     01h10m37s  0
       int-PE-1-P-5                                                 0
192.168.26.0/30                               Remote  OSPF      00h01m07s  10
       192.168.15.2                                                 30
192.168.37.0/30                               Remote  OSPF      00h00m48s  10
       192.168.15.2                                                 40
192.168.48.0/30                               Remote  OSPF      00h00m33s  10
       192.168.15.2                                                 30
192.168.56.0/30                               Remote  OSPF      00h01m19s  10
       192.168.15.2                                                 20
192.168.58.0/30                               Remote  OSPF      00h01m19s  10
       192.168.15.2                                                 20
192.168.67.0/30                               Remote  OSPF      00h01m07s  10
       192.168.15.2                                                 30
192.168.78.0/30                               Remote  OSPF      00h00m39s  10
       192.168.15.2                                                 30
-------------------------------------------------------------------------------
No. of Routes: 16
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#

*A:PE-1# show router fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                                LOCAL
    192.0.2.1 (system)
192.0.2.2/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.3/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.4/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.5/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.6/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.7/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.0.2.8/32                                                OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.15.0/30                                             LOCAL
    192.168.15.0 (int-PE-1-P-5)
192.168.26.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
```

```
192.168.37.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.48.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.56.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.58.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.67.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
192.168.78.0/30                                             OSPF
    192.168.15.2 (int-PE-1-P-5)
-------------------------------------------------------------------------------
Total Entries : 16
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The next step in the process of setting up the IP/MPLS network, is enabling the IP interfaces in the MPLS and RSVP context on all involved nodes (PE and P nodes). By default, the system interface is put automatically within the MPLS/RSVP context. When an interface is put in the MPLS context, the system also copies it into the RSVP context. Explicit enabling of MPLS and RSVP context is done by the **no shutdown** command. The following output displays the MPLS/RSVP configuration for PE-1.

```
*A:PE-1# configure router rsvp no shutdown


*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-P-5"
            exit
            no shutdown
```

Step 2. Configure the RSVP LSPs.

In both metro areas, RSVP LSPs are set up from all PEs toward the ABRs, no intra-area PE-PE RSVP LSPs are needed. In the core/backbone, a full RSVP LSP mesh is required. To simplify the RSVP LSP configuration, no fast reroute is enabled on the RSVP LSPs in the metro areas, only in the backbone area. All RSVP paths are configured as **strict paths.** As an example, the configuration for PE-1 and P-5 is as follows:

```
*A:PE-1# configure
    router
        mpls
            path "path-PE-1-P-5"
                hop 1 192.168.15.2 strict
                no shutdown
            exit
            path "path-PE-1-P-5-P-8"
                hop 10 192.168.15.2 strict
                hop 20 192.168.58.2 strict
                no shutdown
            exit
            lsp "LSP-PE-1-P-5"
                to 192.0.2.5
                primary "path-PE-1-P-5"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-P-8"
                to 192.0.2.8
                primary "path-PE-1-P-5-P-8"
                exit
                no shutdown
            exit
```

```
*A:P-5# configure
    router
        mpls
            path "path-P-5-P-6"
                hop 10 192.168.56.2 strict
                no shutdown
            exit
            path "path-P-5-P-8"
                hop 10 192.168.58.2 strict
                no shutdown
            exit
            path "path-P-5-P-6-P-7"
                hop 10 192.168.56.2 strict
                hop 20 192.168.67.2 strict
                no shutdown
            exit
            path "path-P-5-PE-1"
                hop 10 192.168.15.1 strict
                no shutdown
            exit
            path "path-P-5-P-8-PE-4"
                hop 10 192.168.58.2 strict
                hop 20 192.168.48.1 strict
                no shutdown
            exit
            lsp "LSP-P-5-PE-1"
                to 192.0.2.1
                primary "path-P-5-PE-1"
                exit
                no shutdown
            exit
            lsp "LSP-P-5-PE-4"
                to 192.0.2.4
                primary "path-P-5-P-8-PE-4"
                exit
                no shutdown
            exit
            lsp "LSP-P-5-P-6"
                to 192.0.2.6
                cspf
                fast-reroute facility
                exit
                primary "path-P-5-P-6"
                exit
                no shutdown
            exit
            lsp "LSP-P-5-P-7"
                to 192.0.2.7
                cspf
                fast-reroute facility
                exit
                primary "path-P-5-P-6-P-7"
                exit
                no shutdown
            exit
            lsp "LSP-P-5-P-8"
                to 192.0.2.8
                cspf
                fast-reroute facility
                exit
                primary "path-P-5-P-8"
                exit
                no shutdown
            exit
```

To display the state of RSVP LSPs, several show commands can be used. A command to show the TTM is **show router tunnel-table** with parameter **rsvp** to reference to the RSVP LSP signaling protocol. By default, an RSVP LSP has preference **7**.

```
*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To                Tun     Fastfail  Adm  Opr
                                                     Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-P-5                       192.0.2.5         1       No        Up   Up
LSP-PE-1-P-8                       192.0.2.8         2       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 2
===============================================================================
*A:PE-1#

*A:PE-1# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.5/32      rsvp      MPLS  1         7        192.168.15.2   16777215
192.0.2.8/32      rsvp      MPLS  2         7        192.168.15.2   16777215
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

On ABR P-5:

```
*A:P-5# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To             Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-P-5-PE-1                       192.0.2.1        1       No        Up   Up
LSP-P-5-PE-4                       192.0.2.4        2       No        Up   Up
LSP-P-5-P-6                        192.0.2.6        3       Yes       Up   Up
LSP-P-5-P-7                        192.0.2.7        4       Yes       Up   Up
LSP-P-5-P-8                        192.0.2.8        5       Yes       Up   Up
-------------------------------------------------------------------------------
LSPs : 5
===============================================================================
*A:P-5#
```

```
*A:P-5# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      rsvp      MPLS  1         7        192.168.15.1   16777215
192.0.2.4/32      rsvp      MPLS  2         7        192.168.58.2   16777215
192.0.2.6/32      rsvp      MPLS  3         7        192.168.56.2   10
192.0.2.7/32      rsvp      MPLS  4         7        192.168.56.2   20
192.0.2.8/32      rsvp      MPLS  5         7        192.168.58.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:P-5#
```

By default, the metric for strict LSPs configured without Constrained Shortest Path First (CSPF) (RSVP LSPs in metro areas) is infinite (value = 16777215). The LSP metric for CSPF LSPs (RSVP LSPs in the core area) follows the IGP cost. LSP metrics can be explicitly set on the LSP level, see also in the Additional Topics section.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-P-5" metric
  - metric <metric>
  - no metric

 <metric>             : <0..16777215>
```

Whenever an RSVP LSP comes up, it is by default eligible for LDPoRSVP, meaning that RSVP will signal to the relevant IGP (OSPF in this case) that the LSP should be included in the IGP Shortest Path First (SPF) run. The destination of the LSP (192.0.2.5) will be considered as a potential endpoint in the FEC resolution. With the **info detail** command, all default settings of a context are shown.

```
A:PE-1# configure router mpls lsp "LSP-PE-1-P-5"
A:PE-1>config>router>mpls>lsp# info detail
----------------------------------------------
                to 192.0.2.5

---snip---
                ldp-over-rsvp include

---snip---


*A:PE-1# show router mpls lsp "LSP-PE-1-P-5" detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-PE-1-P-5
LSP Type        : RegularLsp                LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.1                 To                   : 192.0.2.5
Adm State       : Up                        Oper State           : Up

---snip---

LdpOverRsvp : Enabled

---snip---

Primary(a)  : path-PE-1-P-5                    Up Time        : 0d 00:04:55
Bandwidth   : 0 Mbps
===============================================================================
*A:PE-1#
```

To make an RSVP LSP ineligible for LDPoRSVP, use the **exclude** command.

```
A:PE-1# configure router mpls lsp <LSP-name> ldp-over-rsvp exclude
```

Step 3. Create T-LDP sessions according to RSVP LSPs.

It is a must that when configuring an RSVP LSP eligible for LDPoRSVP, also a T-LDP session is initiated. This must be done on all PE and P nodes.

```
*A:PE-1# configure
```

```
    router
        ldp
            targeted-session
                peer 192.0.2.5
                exit
                peer 192.0.2.8
                exit
            exit
```

```
*A:PE-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.5:0         Targeted  Established   15        16        0d 00:01:04
192.0.2.8:0         Targeted  Established   5         7         0d 00:00:19
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-1#
```

Step 4. Enable LDPoRSVP.

This is done using the **tunneling** keyword inside the T-LDP session context. This configuration is needed on all PE and ABR nodes.

```
*A:PE-1# configure
    router
        ldp
            targeted-session
                peer 192.0.2.5
                    tunneling
                    exit
                exit
                peer 192.0.2.8
                    tunneling
                    exit
                exit
            exit
```

As a result of the **tunneling** command, the LDPoRSVP process of FEC resolving is initiated. As already stated in the introduction, FEC resolution is a three-step process. First run an SPF calculation to the destination, then select an endpoint close to that destination followed by a tunnel to that endpoint. The next two steps go more into detail on this FEC resolution process. Step 5 will handle inter-area FEC resolving and Step 6 will handle intra-area FEC resolving.

Step 5. Inter-area FEC resolving (ingress LER is PE-1, egress LER is PE-2)

i. Verification endpoint nodes and associated RSVP tunnels.

The first thing to do in the inter-area FEC resolving process is for PE-1 to perform an SPF calculation toward PE-2 with the purpose to search for an eligible endpoint, as close as possible to PE-2. An endpoint is eligible when a T-LDP session exists between PE-1 and the endpoint node, tunneling is configured on the endpoint node, PE-1 received a label for the destination FEC from the endpoint and an RSVP LSP exists between PE-1 and endpoint node that can be used for LDPoRSVP.

Endpoint node in OSPF area 1 can be either P-5 or P-8 (only those nodes have a T-LDP session toward PE-1). With **show router ldp bindings active prefixes prefix 192.0.2.2/32**, it can be concluded that P-5 will be the endpoint node (EgrNextHop).

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
  --                                        524268
192.0.2.5                                   LspId 1

192.0.2.2/32                                Swap
524281                                      524268
192.0.2.5                                   LspId 1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#


*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                          To               Tun     Fastfail  Adm  Opr
                                                   Id      Config
-------------------------------------------------------------------------------
```

```
LSP-PE-1-P-5                      192.0.2.5        1        No         Up   Up
LSP-PE-1-P-8                      192.0.2.8        2        No         Up   Up
-------------------------------------------------------------------------------
LSPs : 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

---snip---

192.0.2.5/32      rsvp      MPLS  1         7        192.168.15.2   16777215

---snip---
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

Endpoint node in OSPF area 0 can be either P-6, P-7 or P-8 (only those nodes have a T-LDP session toward P-5). With **show router ldp bindings active prefixes prefix 192.0.2.2/32**, it can be concluded that P-6 will be the endpoint node.

```
*A:P-5# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
  --                                        524270
192.0.2.6                                   LspId 3

192.0.2.2/32                                Swap
524268                                      524270
192.0.2.6                                   LspId 3

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-5#
```

```
*A:P-5# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-P-5-PE-1                       192.0.2.1        1       No        Up   Up
LSP-P-5-PE-4                       192.0.2.4        2       No        Up   Up
LSP-P-5-P-6                        192.0.2.6        3       Yes       Up   Up
LSP-P-5-P-7                        192.0.2.7        4       Yes       Up   Up
LSP-P-5-P-8                        192.0.2.8        5       Yes       Up   Up
-------------------------------------------------------------------------------
LSPs : 5
===============================================================================
*A:P-5#

*A:P-5# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

---snip---

192.0.2.6/32      rsvp      MPLS  3         7        192.168.56.2   10

---snip---

-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:P-5#
```

On node P-6, the same commands can be repeated for the final destination node PE-2. Also there, an RSVP LSP toward PE-2 will be used as transport tunnel for user packets.

```
*A:P-6# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.6)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
```

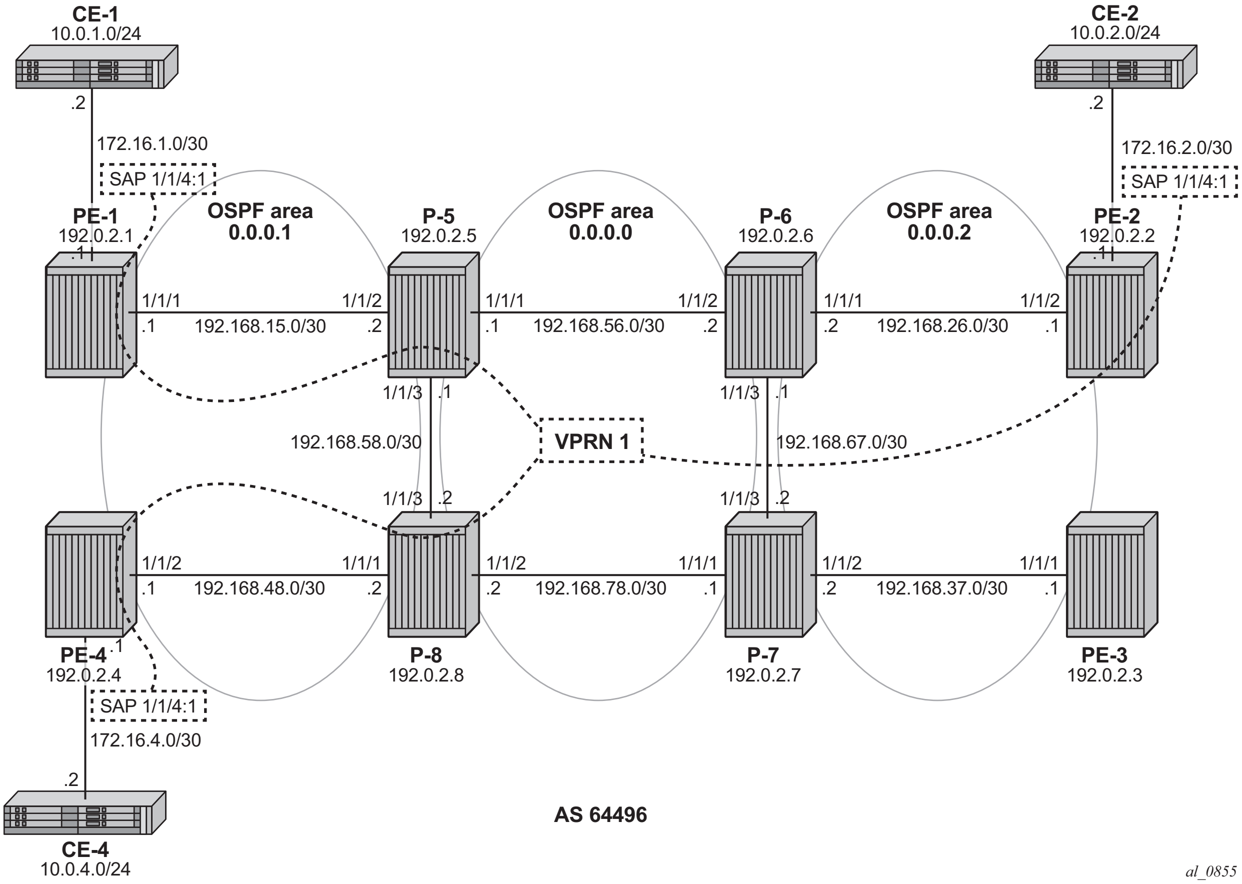
```
  --                                        524285
192.0.2.2                                   LspId 4

192.0.2.2/32                                Swap
524270                                      524285
192.0.2.2                                   LspId 4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-6#

*A:P-6# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To                 Tun     Fastfail  Adm  Opr
                                                      Id      Config
-------------------------------------------------------------------------------
LSP-P-6-P-5                        192.0.2.5          1       Yes       Up   Up
LSP-P-6-P-7                        192.0.2.7          2       Yes       Up   Up
LSP-P-6-P-8                        192.0.2.8          3       Yes       Up   Up
LSP-P-6-PE-2                       192.0.2.2          4       No        Up   Up
LSP-P-6-PE-3                       192.0.2.3          5       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 5
===============================================================================
A:P-6#

*A:P-6# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

---snip---

192.0.2.2/32      rsvp      MPLS  4         7        192.168.26.1   16777215

---snip---

-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:P-6#
```

Nodes P-5 and P-6 behave as stitching nodes to stitch RSVP LSPs. P-5 will stitch LSP-PE-1-P-5 and LSP-P-5-P-6 together while P-6 node will stitch LSP-P-5-P-6 and LSP-P-6-PE-2 together.

When the endpoints are defined, one corresponding RSVP LSP to those endpoints will be chosen (when ECMP equals 1). Selection criteria are as follows. When RSVP LSPs are configured under the T-LDP **tunneling** command (maximum 4), the one with the lowest LSP metric will be selected. When no RSVP LSPs are configured under the T-LDP **tunneling** command, LDP checks the TTM for all available RSVP LSPs. The RSVP LSP with the lowest metric and operational state up will be selected.

ii. Traffic verification using a virtual private routed network (VPRN) service.

Figure 246 **VPRN 1 with LDPoRSVP and No Intra-Area PE Connectivity**



VPRN service 1 is set up between three PE nodes (PE-1, PE-2, and PE-4) using the **auto-bind-tunnel resolution-filter ldp resolution filter** command. See also Figure 246 for the exact addressing scheme.

```
*A:PE-1# configure
    service
        vprn 1 name "VPRN 1" customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "int-PE-1-CE-1" create
                address 172.16.1.1/30
                sap 1/1/4:1 create
                exit
            exit
            static-route-entry 10.0.1.0/24
                next-hop 172.16.1.2
                    no shutdown
                exit
            exit
            no shutdown
```

In order to distribute VPRN information (VPN-IPv4 routes and VPRN service labels) across the service provider network, Multi-Protocol Border Gateway Protocol (MP-BGP) is needed. MP-BGP is configured on PE-1, PE-2, and PE-4 with P-5 (192.0.2.5) being the Route Reflector (RR). In this way, no full BGP mesh between the three PE-nodes is needed, only a BGP peering toward RR.

```
*A:PE-1# configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                family ipv4 vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.5
                exit
            exit
            no shutdown
```

```
*A:P-5# configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                family ipv4 vpn-ipv4
                peer-as 64496
                cluster 1.1.1.1
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.2
                exit
                neighbor 192.0.2.4
                exit
            exit
            no shutdown
```

If user traffic is monitored between PE-1 (ingress LER) and PE-2 (egress LER), three labels should be seen. The outer label is the transport label distributed using the RSVP protocol, the inner label is the service label distributed using MP-BGP. LDPoRSVP will add an extra MPLS label between transport and service label (distributed using LDP). This middle label is used to tell the endpoint nodes (P-5 and P-6 acting as ABR) what to do. The transport label stack contains two labels: an RSVP label and an LDP label.

Translated into show commands for traffic on ingress port 1/1/2 on ABR P-5 (PE-1<=>P-5 link):

RSVP transport label 524287 is added as the top label on each user packet

```
*A:PE-1# show router rsvp session lsp-name "LSP-PE-1-P-5::path-PE-1-P-5" detail

===============================================================================
RSVP Sessions (Detailed)
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-PE-1-P-5::path-PE-1-P-5
-------------------------------------------------------------------------------
From            : 192.0.2.1               To             : 192.0.2.5
Tunnel ID       : 1                       LSP ID         : 25088
Style           : SE                      State          : Up
Session Type    : Originate
In Interface    : n/a                     Out Interface  : 1/1/1
In IF Name      : n/a
Out IF Name     : int-PE-1-P-5
In Label        : n/a                     Out Label      : 524287
Previous Hop    : n/a                     Next Hop       : 192.168.15.2
---snip---
```

LDP label 524268 is added as the middle label on each user packet

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.2/32
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
  --                                        524268
192.0.2.5                                   LspId 1

192.0.2.2/32                                Swap
524281                                      524268
192.0.2.5                                   LspId 1
```

```
-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

Service label 524277 is added as the inner MP-BGP label on each user packet

Note: This label will not change at endpoint nodes (P-5 and P-6). Ingress LER (PE-1) will push the service label to the user packet while the egress LER (PE-2) will pop the service label.

```
*A:PE-1# show router bgp neighbor 192.0.2.5 received-routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64496:1:10.0.1.0/24                                100         None
      192.0.2.1                                          None        524277
      No As-Path
u*>i  64496:1:10.0.2.0/24                                100         None
      192.0.2.2                                          None        524277
      No As-Path
u*>i  64496:1:10.0.4.0/24                                100         None
      192.0.2.4                                          None        524277
      No As-Path
i     64496:1:172.16.1.0/30                              100         None
      192.0.2.1                                          None        524277
      No As-Path
u*>i  64496:1:172.16.2.0/30                              100         None
      192.0.2.2                                          None        524277
      No As-Path
u*>i  64496:1:172.16.4.0/30                              100         None
      192.0.2.4                                          None        524277
      No As-Path
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

Translated into show commands for traffic on ingress port 1/1/2 on ABR P-6 (P-5<=>P-6 link):

RSVP transport label 524284 is added as the top label on each user packet.

```
*A:P-5# show router rsvp session lsp-name "LSP-P-5-P-6::path-P-5-P-6" detail

===============================================================================
RSVP Sessions (Detailed)
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-P-5-P-6::path-P-5-P-6
-------------------------------------------------------------------------------
From            : 192.0.2.5               To            : 192.0.2.6
Tunnel ID       : 3                       LSP ID        : 15360
Style           : SE                      State         : Up
Session Type    : Originate
In Interface    : n/a                     Out Interface  : 1/1/1
In IF Name      : n/a
Out IF Name     : int-P-5-P-6
In Label        : n/a                     Out Label      : 524284
Previous Hop    : n/a                     Next Hop       : 192.168.56.2
---snip---
```

LDP label 524270 is added as the middle label on each user packet.

```
*A:P-5# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
  --                                        524270
192.0.2.6                                   LspId 3

192.0.2.2/32                                Swap
524268                                      524270
192.0.2.6                                   LspId 3

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-5#
```

Service label 524277 is added as the inner MP-BGP label on each user packet.

Translated into show commands for traffic on ingress port 1/1/2 on node PE-2 (P-6<=>PE-2 link).

RSVP transport label 524287 is added as the outer label on each user packet.

```
*A:P-6# show router rsvp session lsp-name "LSP-P-6-PE-2::path-P-6-PE-2" detail

===============================================================================
RSVP Sessions (Detailed)
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-P-6-PE-2::path-P-6-PE-2
-------------------------------------------------------------------------------
From            : 192.0.2.6             To            : 192.0.2.2
Tunnel ID       : 4                     LSP ID        : 31744
Style           : SE                    State         : Up
Session Type    : Originate
In Interface    : n/a                   Out Interface  : 1/1/1
In IF Name      : n/a
Out IF Name     : int-P-6-PE-2
In Label        : n/a                   Out Label     : 524287
Previous Hop    : n/a                   Next Hop      : 192.168.26.1
---snip---
```

iii. LDP label 524285 is added as the middle label on each user packet.

```
*A:P-6# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.6)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push
  --                                        524285
192.0.2.2                                   LspId 4

192.0.2.2/32                                Swap
524270                                      524285
192.0.2.2                                   LspId 4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-6#
```

Service label 524277 is added as the inner MP-BGP label on each user packet.

Step 6. Intra-area FEC resolving (ingress LER is PE-1, egress LER is PE-4).

i. Verification endpoint node and associated RSVP tunnel.

The first thing to do in the intra-area FEC resolving process is for PE-1 to perform an SPF calculation toward PE-4 to search for an eligible endpoint, as close as possible to PE-4. An endpoint is eligible when a T-LDP session exists between PE-1 and the endpoint node, tunneling is configured on the endpoint node, PE-1 received a label for the destination FEC from the endpoint and an RSVP LSP exists between PE-1 and endpoint node that can be used for LDPoRSVP.

First endpoint node in OSPF area 1 can be either P-5 or P-8 (only those nodes have a T-LDP session toward PE-1). With **show router ldp bindings active prefixes prefix 192.0.2.4/32** it can be concluded that P-5 will be the endpoint node.

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.4/32
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push
  --                                        524269
192.0.2.5                                   LspId 1

192.0.2.4/32                                Swap
524283                                      524269
192.0.2.5                                   LspId 1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-P-5                       192.0.2.5        1       No        Up   Up
LSP-PE-1-P-8                       192.0.2.8        2       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router tunnel-table
```

```
===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

192.0.2.5/32      rsvp      MPLS  1         7        192.168.15.2   16777215

---snip---

===============================================================================
*A:PE-1#
```

On node P-5, the same commands can be repeated for the final destination node (PE-4). Also there, an RSVP LSP toward PE-4 will be used as transport tunnel for user packets.

```
*A:P-5# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push
  --                                        524285
192.0.2.4                                   LspId 2

192.0.2.4/32                                Swap
524269                                      524285
192.0.2.4                                   LspId 2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-5#


*A:P-5# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-P-5-PE-1                       192.0.2.1        1       No        Up   Up
LSP-P-5-PE-4                       192.0.2.4        2       No        Up   Up
LSP-P-5-P-6                        192.0.2.6        3       Yes       Up   Up
LSP-P-5-P-7                        192.0.2.7        4       Yes       Up   Up
LSP-P-5-P-8                        192.0.2.8        5       Yes       Up   Up
```

```
-------------------------------------------------------------------------------
LSPs : 5
===============================================================================
*A:P-5#


*A:P-5# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

---snip---

192.0.2.4/32      rsvp      MPLS  2         7        192.168.58.2   16777215

---snip---

===============================================================================
*A:P-5#
```

P-5 node acts as a stitching node to stitch RSVP LSPs. P-5 will stitch LSP-PE-1-P-5 and LSP-P-5-PE-4 together.

When the endpoint node (P-5) is defined, the corresponding RSVP LSP to this endpoint will be chosen. Selection criteria are as follows (when ECMP=1). When RSVP LSPs are configured under the T-LDP **tunneling** command (maximum 4), the one with the lowest LSP metric will be selected. When no RSVP LSPs are configured under the T-LDP **tunneling** command, LDP checks TTM for all available RSVP LSPs. The RSVP LSP with the lowest metric and operational state **up** will be selected.

ii. Traffic verification using a VPRN service (see VPRN 1 with LDPoRSVP and No Intra-Area PE Connectivity).

If user traffic between PE-1 (ingress LER) and PE-4 (egress LER) is monitored, three labels are seen. The outer label is the transport label (distributed using RSVP protocol), the inner label is the service label (distributed using MP-BGP). LDPoRSVP will add an extra MPLS transport label between outer and inner label (distributed using LDP). This middle label is used to tell the endpoint node (P-5) what to do.

Translated into show commands for traffic on ingress port 1/1/2 on P-5 node (PE-1<=>P-5 link):

Transport label 524287 is added as the top RSVP label on each user packet.

```
*A:PE-1# show router rsvp session lsp-name "LSP-PE-1-P-5::path-PE-1-P-5" detail

===============================================================================
RSVP Sessions (Detailed)
```

```
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-PE-1-P-5::path-PE-1-P-5
-------------------------------------------------------------------------------
From            : 192.0.2.1             To             : 192.0.2.5
Tunnel ID       : 1                     LSP ID         : 25088
Style           : SE                    State          : Up
Session Type    : Originate
In Interface    : n/a                   Out Interface  : 1/1/1
In IF Name      : n/a
Out IF Name     : int-PE-1-P-5
In Label        : n/a                   Out Label      : 524287
Previous Hop    : n/a                   Next Hop       : 192.168.15.2
---snip---
```

LDPoRSVP label 524269 is added as the middle LDP label on each user packet.

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push
  --                                        524269
192.0.2.5                                   LspId 1

192.0.2.4/32                                Swap
524283                                      524269
192.0.2.5                                   LspId 1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

Service label 524277 is added as the inner MP-BGP label on each user packet.

Note: This label will not change at endpoint node (P-5). Ingress LER (PE-1) will push the service label to the user packet while the egress LER (PE-4) will pop the service label.

```
*A:PE-1# show router bgp neighbor 192.0.2.5 received-routes vpn-ipv4
```

```
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64496:1:10.0.1.0/24                                100         None
      192.0.2.1                                          None        524277
      No As-Path
u*>i  64496:1:10.0.2.0/24                                100         None
      192.0.2.2                                          None        524277
      No As-Path
u*>i  64496:1:10.0.4.0/24                                100         None
      192.0.2.4                                          None        524277
      No As-Path
i     64496:1:172.16.1.0/30                              100         None
      192.0.2.1                                          None        524277
      No As-Path
u*>i  64496:1:172.16.2.0/30                              100         None
      192.0.2.2                                          None        524277
      No As-Path
u*>i  64496:1:172.16.4.0/30                              100         None
      192.0.2.4                                          None        524277
      No As-Path
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

Translated into show commands for traffic on ingress port 1/1/2 on node PE-4 (PE-4<=>P-8 link):

P-5 pushes RSVP transport label 524284 as the top label on each user packet. This RSVP transport label is swapped by P-8 to label 524287.

```
*A:P-5# show router mpls lsp "LSP-P-5-PE-4" path detail

===============================================================================
MPLS LSP LSP-P-5-PE-4 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-P-5-PE-4 Path path-P-5-P-8-PE-4
-------------------------------------------------------------------------------
```

```
LSP Name         : LSP-P-5-PE-4
From             : 192.0.2.5              To                   : 192.0.2.4
Admin State      : Up                     Oper State           : Up
Path Name        : path-P-5-P-8-PE-4
Path LSP ID      : 52224                  Path Type            : Primary
Path Admin       : Up                     Path Oper            : Up
Out Interface    : 1/1/3                  Out Label            : 524284
---snip---

Explicit Hops    :
    192.168.58.2(S)    -> 192.168.48.1(S)
Actual Hops      :
    192.168.58.1 (192.0.2.5)                       Record Label        : N/A
 -> 192.168.58.2 (192.0.2.8)                       Record Label        : 524284
 -> 192.168.48.1                                   Record Label        : 524287
Resignal Eligible: False
Last Resignal    : n/a                    CSPF Metric          : 0
===============================================================================
*A:P-5#
```

Note: show router rsvp session lsp-name LSP-P-5-PE-4::path-P-5-P-8-PE-4 detail cannot be used because it only shows the outgoing RSVP label toward node P-8. On node P-8, RSVP transport label 524284 will be swapped into RSVP transport label 524287 for the link P-8 <=> PE-4.

LDP label 524285 is added as the middle label on each user packet.

```
*A:P-5# show router ldp bindings active prefixes prefix 192.0.2.4/32
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push
  --                                        524285
192.0.2.4                                   LspId 2

192.0.2.4/32                                Swap
524269                                      524285
192.0.2.4                                   LspId 2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-5#
```

Service label 524277 is added as the inner MP-BGP label on each user packet.

Additional Topics

prefer-tunnel-in-tunnel

If the next-hop router advertised the same FEC over link-level LDP (iLDP), LDP will prefer the iLDP tunnel by default unless the user explicitly changed the default preference using the **prefer-tunnel-in-tunnel** command. In this case an LDPoRSVP tunnel will have precedence.

Until now, no RSVP LSPs are configured inside the **ldp targeted-session peer tunneling** context. Therefore, two additional strict non-CSPF RSVP LSPs are added between ingress LER PE-1 and egress LER P-5. Both LSPs will have an explicit metric setting and will be applied inside the **ldp tunneling** context. On the Layer 3 interface between PE-1 and P-5, iLDP is enabled.

```
A:PE-1# configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-P-5" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

```
A:P-5# configure
    router
        ldp
            interface-parameters
                interface "int-P-5-PE-1" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-P-5-metric100"
                to 192.0.2.5
                metric 100
                primary "path-PE-1-P-5"
                exit
                no shutdown
            exit
            lsp "LSP-PE-1-P-5-metric200"
                to 192.0.2.5
                metric 200
                primary "path-PE-1-P-5"
                exit
                no shutdown
            exit

*A:PE-1# configure
    router
        ldp
            targeted-session
                peer 192.0.2.5
                    tunneling
                        lsp "LSP-PE-1-P-5-metric100"
                        lsp "LSP-PE-1-P-5-metric200"
                    exit
                exit
            exit
```

The following tunnel table on node PE-1 contains four tunnels toward P-5: one LDP tunnel and three RSVP tunnels:

```
*A:PE-1# show router tunnel-table

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------

---snip---

192.0.2.5/32      rsvp      MPLS  3         7        192.168.15.2   100
192.0.2.5/32      rsvp      MPLS  4         7        192.168.15.2   200
192.0.2.5/32      rsvp      MPLS  1         7        192.168.15.2   16777215
192.0.2.5/32      ldp       MPLS  65537     9        192.168.15.2   10

---snip---

===============================================================================
```

Tunnel ID 1 is a reference to LSP-PE-1-P-5. Tunnel ID 3 is a reference to LSP-PE-1-P-5-metric100. Tunnel ID 4 is a reference to LSP-PE-1-P-5-metric200 and owner LDP is a reference to iLDP.

Taken into account the FEC resolution rules, iLDP prevails (no LDPoRSVP tunnel will be used).

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
```

```
================================================================================
LDP IPv4 Prefix Bindings (Active)
================================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
--------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524271
192.168.15.2                                1/1/1

192.0.2.5/32                                Swap
524284                                      524271
192.168.15.2                                1/1/1

--------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
================================================================================
*A:PE-1#
```

This behavior can be changed by setting the **prefer-tunnel-in-tunnel** command in the LDP context. Now, the LDPoRSVP tunnel with the best (= lowest) metric is taken.

```
*A:PE-1# configure router ldp prefer-tunnel-in-tunnel
```

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
================================================================================
---snip---
================================================================================
LDP IPv4 Prefix Bindings (Active)
================================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
--------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524271
192.0.2.5                                   LspId 3

192.0.2.5/32                                Swap
524284                                      524271
192.0.2.5                                   LspId 3

--------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
================================================================================
*A:PE-1#
```

```
*A:PE-1# show router mpls lsp

================================================================================
MPLS LSPs (Originating)
================================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-P-5                       192.0.2.5        1       No        Up   Up
LSP-PE-1-P-8                       192.0.2.8        2       No        Up   Up
LSP-PE-1-P-5-metric100             192.0.2.5        3       No        Up   Up
LSP-PE-1-P-5-metric200             192.0.2.5        4       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 4
===============================================================================
*A:PE-1#
```

If the LSP-PE-1-P-5-metric100 is shut down, then the LSP-PE-1-P-5-metric200 will become active.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-P-5-metric100" shutdown
```

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524271
192.0.2.5                                   LspId 4

192.0.2.5/32                                Swap
524284                                      524271
192.0.2.5                                   LspId 4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#*
```

```
*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-P-5                       192.0.2.5        1       No        Up   Up
LSP-PE-1-P-8                       192.0.2.8        2       No        Up   Up
LSP-PE-1-P-5-metric100             192.0.2.5        3       No        Dwn  Dwn
LSP-PE-1-P-5-metric200             192.0.2.5        4       No        Up   Up
-------------------------------------------------------------------------------
```

```
LSPs : 4
===============================================================================
*A:PE-1#
```

If LSP-PE-1-P-5-metric200 is shut down, iLDP resumes.

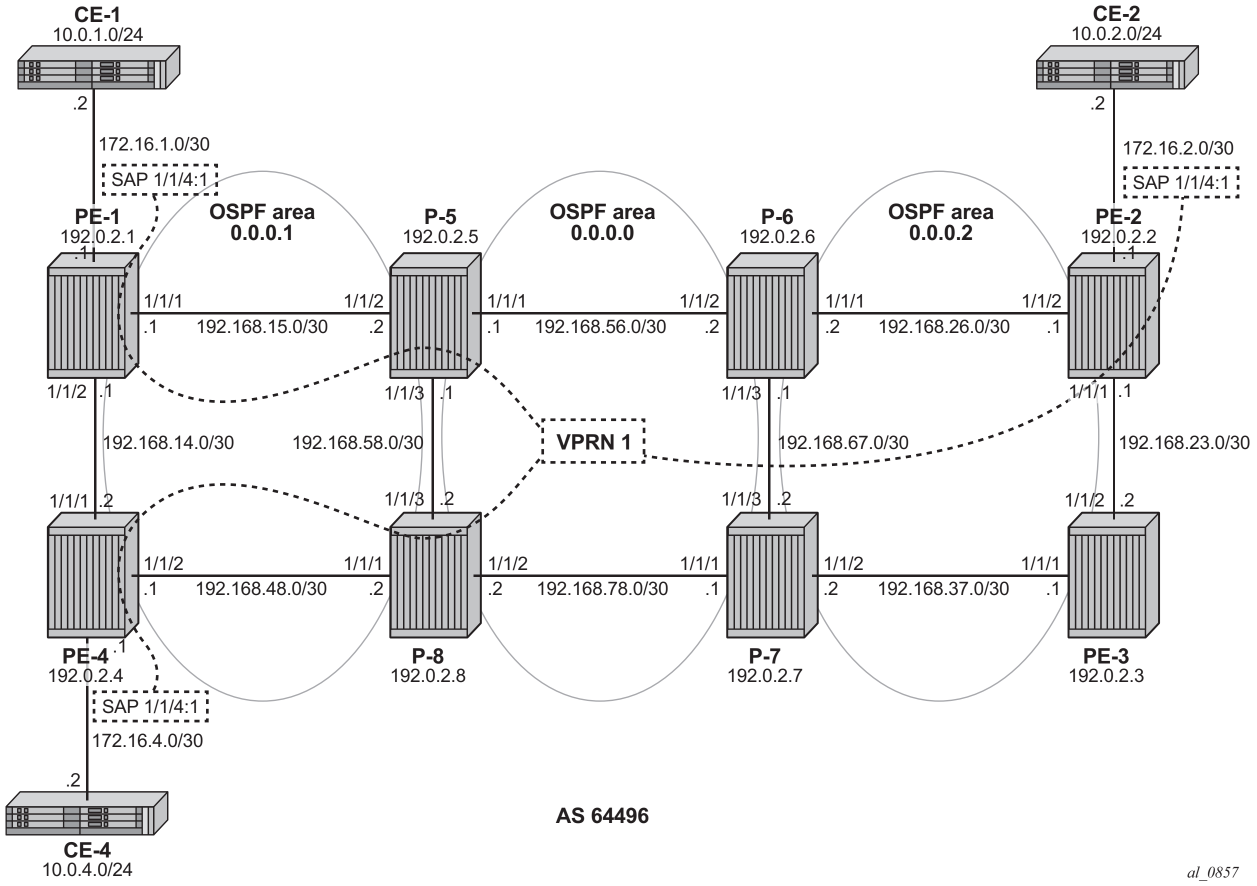
```
*A:PE-1# configure router mpls lsp "LSP-PE-1-P-5-metric200" shutdown
```

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524271
192.168.15.2                                1/1/1

192.0.2.5/32                                Swap
524284                                      524271
192.168.15.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

Intra-PE Connectivity Changes LDPoRSVP Behavior

Figure 247 shows two metro areas; both of the intra PEs are physically connected with each other. Compared with the previous figures, PE-1 is directly connected to PE-4 and PE-2 is directly connected to PE-3 (up to the OSPF level).

Figure 247 **VPRN 1 with LDPoRSVP and Intra-Area PE Connectivity**



The SPF path calculation on PE-1 toward destination (PE-4) will not point to node P-5 anymore (as was seen before), but will now point directly to PE-4 (shortest, lowest IGP metric). As a conclusion, it can be said that when possible intra-area endpoint nodes are not part of the calculated SPF path, LDPoRSVP will be not be preferred anymore. For this situation, it is advisable to configure iLDP on the intra-PE interfaces to have a fallback mechanism.

This is configured on PE-1 and PE-4 as follows:

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-4"
            address 192.168.14.1/30
            port 1/1/2
        exit
        ospf
            area 0.0.0.1
                interface "int-PE-1-PE-4"
                    interface-type point-to-point
                exit
            exit
        exit

*A:PE-4# configure
    router
        interface "int-PE-4-PE-1"
            address 192.168.14.2/30
            port 1/1/1
        exit
        ospf
            area 0.0.0.1
                interface "int-PE-4-PE-1"
                    interface-type point-to-point
                exit
            exit
        exit
```

From the moment iLDP is configured, an LDP LSP is set up. Intra-area PE traffic will flow over this LDP LSP.

```
*A:PE-1# configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-4" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit

*A:PE-4# configure
    router
        ldp
            interface-parameters
                interface "int-PE-4-PE-1" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
```

```
*A:PE-1# show router tunnel-table 192.0.2.4/32
===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.4/32      ldp       MPLS  65538     9        192.168.14.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

If user traffic is monitored, between PE-1 (ingress LER) and PE-4 (egress LER) only two labels are seen. The outer label is the transport label (distributed using LDP), the inner label is the service label (distributed using MP-BGP). No LDPoRSVP label is present anymore. Translated into show commands for traffic on ingress port 1/1/1 on PE-4 node (PE-1<=>PE-4 link):

LDP transport label 524285 is added as the outer label on each user packet.

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push
  --                                        524285
192.168.14.2                                1/1/2

192.0.2.4/32                                Swap
524283                                      524285
192.168.14.2                                1/1/2’

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

Service label 524277 is added as the inner MP-BGP label on each user packet.

```
*A:PE-1# show router bgp neighbor 192.0.2.5 received-routes vpn-ipv4
===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64496:1:10.0.1.0/24                                100         None
```

```
      192.0.2.1                                                None        524277
      No As-Path
u*>i  64496:1:10.0.2.0/24                                      100         None
      192.0.2.2                                                None        524277
      No As-Path
u*>i  64496:1:10.0.4.0/24                                      100         None
      192.0.2.4                                                None        524277
      No As-Path
i     64496:1:172.16.1.0/30                                    100         None
      192.0.2.1                                                None        524277
      No As-Path
u*>i  64496:1:172.16.2.0/30                                    100         None
      192.0.2.2                                                None        524277
      No As-Path
u*>i  64496:1:172.16.4.0/30                                    100         None
      192.0.2.4                                                None        524277
      No As-Path
-------------------------------------------------------------------------------
Routes : 6
===============================================================================
*A:PE-1#
```

Conclusion

LDPoRSVP allows tunneling of user packets toward an LDP far-end destination inside an RSVP LSP (with the benefits of RSVP LSPs, fast-reroute (FRR) and traffic engineering (TE)). The main application of this feature is for deployment of MPLS based services, for example, VPRN, virtual leased line (VLL), and virtual private LAN service (VPLS) services, in large networks where a full mesh of LSPs reaches the limits of scalability.

LDP Point-to-Point LSPs

This chapter provides information about label distribution protocol (LDP) point-to-point label switched paths (LSPs)

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS and was originally written for SR OS release 7.0.R5. The output in the current edition corresponds to SR OS release 16.0.R3. There are no pre-requisites or conditions on the hardware for this configuration.

Overview

Due to the connectionless nature of the network layer protocol IP, packets travel through the network on a hop-by-hop basis with routing decisions made at each node. As a result, hyperaggregation of data on certain links may occur and it may impact the provider's ability to provide guaranteed service levels across the network end-to-end. To address these shortcomings, Multi-Protocol Label Switching (MPLS) was developed. The technology provides the capability to establish connection-oriented paths, called Label Switched Paths (LSPs), over a connectionless (IP) network. The LSP offers a mechanism to engineer network traffic independently from the underlying network routing protocol (mostly IP) to improve the network resiliency and recovery options and to permit delivery of new services that are not readily supported by conventional IP routing techniques (Layer 2 IP Virtual Private Networks (VPNs)). These benefits are essential for today's communication network explaining the wide deployment base of the MPLS technology.

RFC 3031, *Multiprotocol Label Switching Architecture*, specifies the MPLS architecture whereas this chapter describes the configuration and troubleshooting of point-to-point LSPs on SR OS.

Packet Forwarding

When a packet of a connectionless network layer protocol travels from one router to the next, each router in the network makes an independent forwarding decision by performing the following basic tasks: first analyzing the packet's header, then referencing the local routing table to find the longest match based on the destination address in the IP header, and finally sending out the packet on the selected interface. In other terms, the first function partitions the entire set of possible packets into a set of Forwarding Equivalence Classes (FECs). All packets associated to a particular FEC will be forwarded along the same logical path to the same destination. The second function maps each FEC to a next hop destination router. Each router along the packet's path performs these actions.

On the other hand, in MPLS, the assignment of a packet to a particular FEC is done just once, when the packet enters the network. In turn, the FEC is mapped to an LSP, which is established prior to packet forwarding. An MPLS label, representing the FEC to which the packet is assigned, is attached to the packet (push operation) and once labeled, the packet is forwarded to the next hop router along that LSP path. At subsequent hops, no analysis of the packet's network layer header is needed. Instead, the label is used as an index into a table which specifies the next hop and a new label. The old label is replaced with the new label (swap operation), and the packet is forwarded to its next hop. At the MPLS network egress, the label is removed from the packet (pop operation). If this router is the destination (based on the remaining packet), the packet is handed to the receiving application (such as a Virtual Private LAN Service (VPLS) domain). If this router is not the destination of the packet, the packet will be sent into a new MPLS tunnel or forwarded by conventional IP forwarding towards the Layer 3 destination

Terminology

Figure 248 **Generic MPLS Network, MPLS Label Operations**

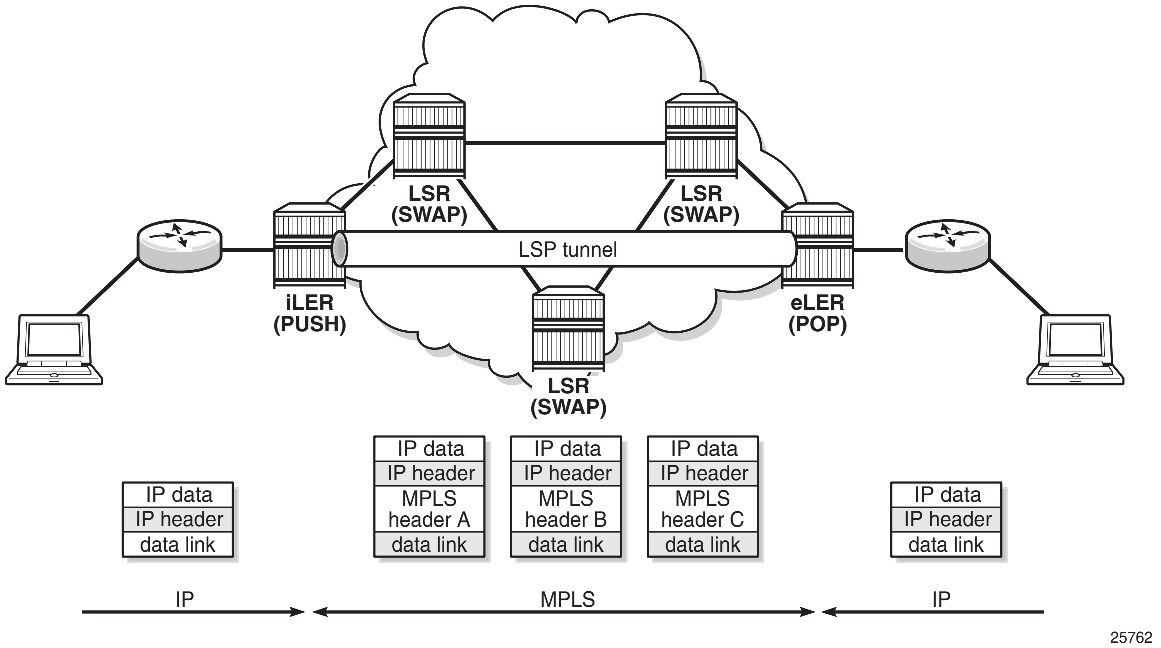


Figure 248 shows a general network topology clarifying the MPLS-related terms. A Label Edge Router (LER) is a device at the edge of an MPLS network, with at least one interface outside the MPLS domain. A router is usually defined as an LER based on its position relative to a particular LSP. The MPLS router at the head-end of an LSP is called the ingress Label Edge Router (iLER). The MPLS router at the tail-end of an LSP is called the egress Label Edge Router (eLER). The iLER receives unlabeled packets from outside the MPLS domain, then applies MPLS labels to the packets, and forwards the labeled packets into the MPLS domain. The eLER receives labeled packets from the MPLS domain, then removes the labels, and forwards unlabeled packets outside the MPLS domain. The eLER can signal an implicit-null label (numeric value 3). This informs the previous hop to send MPLS packets without an outer label and so is known as Penultimate Hop Popping (PHP).

A Label Switching Router (LSR) is a device internal to an MPLS network, with all interfaces inside the MPLS domain. These devices switch labeled packets inside the MPLS domain. In the core of the network, LSRs ignore the packet's network layer (IP) header and simply forward the packet using the MPLS label swapping mechanism.

LSP Establishment

Prior to packet forwarding, the LSP must be established. In order to do so, labels need to be distributed for the path. Labels are usually distributed by a downstream router in the upstream direction (relative to the data flow). There are a number of ways used for label distribution: static, LDP, and RSVP. For static P2P LSPs, see chapter Static Point-to-Point LSPs; for RSVP-TE P2P LSPs, see chapter RSVP Point-to-Point LSPs.

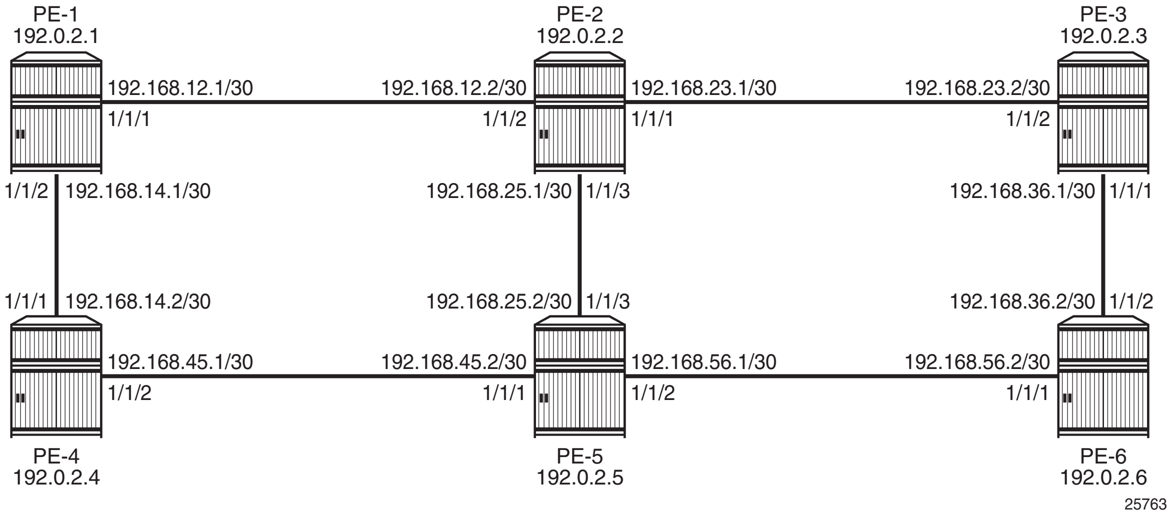
LDP (RFC 5036, *LDP Specification*) can be considered as an extension to the network interior gateway protocol (IGP). As routers become aware of new destination networks, they advertise labels in the upstream direction that will allow upstream routers to reach the destination.

Fast reroute (FRR) allows for establishing backup paths before a failure takes place. This way traffic can flow almost continuously, without waiting for routing protocol convergence; see chapter MPLS LDP FRR using ISIS as IGP.

Example Topology

The example topology is shown in Figure 249. The setup consists of six 7750 SR nodes located in a single autonomous system.

Figure 249 **MPLS Example Topology**



Configuration

As a general prerequisite for the configuration of MPLS LSPs, a correctly working Interior Gateway Protocol (IGP) is required. Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS) can be used as IGP.

LDP is a simple label distribution protocol with basic MPLS functionality (no traffic engineering). Fast Reroute is supported; see chapter MPLS LDP FRR using ISIS as IGP. LDP relies on the underlying routing information provided by an IGP in order to forward labeled packets. Each LDP configured LSR will originate a label for its system address and a label for each FEC for which it has a next hop that is external to the MPLS domain, without the explicit need to manually configure the LSPs. When deviations from this default behavior are desired, import and export policies can be applied.

The configuration is as simple as enabling the LDP protocol instance and adding all network interfaces, for each node. The configuration on node PE-1 is as follows; similar configurations apply on the other nodes.

```
*A:PE-1# configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
                interface "int-PE-1-PE-4" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
```

The **show router ldp discovery** and **show router ldp session** commands can be used to verify the LDP hello adjacencies and sessions. The adjacency type (AdjType) needs to be **Link** while the state should be **Established**. In this example, only IPv4 addresses are used, so the output can be limited to IPv4 only by adding the keyword **ipv4**.

```
*A:PE-1# show router ldp discovery ipv4

===============================================================================
LDP IPv4 Hello Adjacencies
===============================================================================
Interface Name                   Local Addr                              State
AdjType                          Peer Addr
-------------------------------------------------------------------------------
int-PE-1-PE-2                    192.0.2.1:0                             Estab
```

```
link                             192.0.2.2:0

int-PE-1-PE-4                    192.0.2.1:0                            Estab
link                             192.0.2.4:0

-------------------------------------------------------------------------------
No. of IPv4 Hello Adjacencies: 2
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.2:0         Link      Established   35        37        0d 00:01:19
192.0.2.4:0         Link      Established   29        31        0d 00:00:58
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
*A:PE-1#
```

The **show router ldp bindings prefixes** command displays the contents of the LIB (Label Information Base) and contains all labels locally generated (IngLbl) and those received from any LDP neighbors (EgrLbl), whether they are in use or not. The following output is for IPv4 prefixes:

```
*A:PE-1# show router ldp bindings prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix
Peer                                        FEC-Flags
IgrLbl                                      EgrLbl
EgrNextHop                                  EgrIntf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32
192.0.2.2:0
524287U                                       --
  --                                          --

192.0.2.1/32
192.0.2.4:0
524287U                                       --
```

```
  --                                           --

192.0.2.2/32
192.0.2.2:0
  --                                         524287
192.168.12.2                                 1/1/1

192.0.2.2/32
192.0.2.4:0
524286U                                      524285
  --                                           --

192.0.2.3/32
192.0.2.2:0
524285N                                      524285
192.168.12.2                                 1/1/1

192.0.2.3/32
192.0.2.4:0
524285U                                      524284
  --                                           --

192.0.2.4/32
192.0.2.2:0
524284U                                      524284
  --                                           --

192.0.2.4/32
192.0.2.4:0
  --                                         524287
192.168.14.2                                 1/1/2

192.0.2.5/32
192.0.2.2:0
524283N                                      524283
192.168.12.2                                 1/1/1

192.0.2.5/32
192.0.2.4:0
524283U                                      524283
  --                                           --

192.0.2.6/32
192.0.2.2:0
524282N                                      524282
192.168.12.2                                 1/1/1

192.0.2.6/32
192.0.2.4:0
524282U                                      524282
  --                                           --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 12
===============================================================================
*A:PE-1#
```

The **show router ldp bindings active prefixes** command displays the content of the Label Forwarding Information Base (LFIB) and contains all active labels and the associated label actions used for label switching packets. The active LDP bindings for IPv4 prefixes are the following:

```
*A:PE-1# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop
524287                                        --
  --                                          --

192.0.2.2/32                                Push
  --                                        524287
192.168.12.2                                1/1/1

192.0.2.2/32                                Swap
524286                                      524287
192.168.12.2                                1/1/1

192.0.2.3/32                                Push
  --                                        524285
192.168.12.2                                1/1/1

192.0.2.3/32                                Swap
524285                                      524285
192.168.12.2                                1/1/1

192.0.2.4/32                                Push
  --                                        524287
192.168.14.2                                1/1/2

192.0.2.4/32                                Swap
524284                                      524287
192.168.14.2                                1/1/2

192.0.2.5/32                                Push
  --                                        524283
```

```
192.168.12.2                                1/1/1

192.0.2.5/32                                Swap
524283                                      524283
192.168.12.2                                1/1/1

192.0.2.6/32                                Push
 --                                         524282
192.168.12.2                                1/1/1

192.0.2.6/32                                Swap
524282                                      524282
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 11
===============================================================================
*A:PE-1#
```

In the tunnel table, there are LDP LSPs to all other nodes:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.12.2   20
192.0.2.4/32      ldp       MPLS  65539     9        192.168.14.2   10
192.0.2.5/32      ldp       MPLS  65540     9        192.168.12.2   20
192.0.2.6/32      ldp       MPLS  65541     9        192.168.12.2   30
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

In order to signal PHP with LDP, implicit-null must be configured on the eLER.

```
*A:PE-6# configure router ldp implicit-null-label
```

The implicit-null is signaled immediately, all related labels are withdrawn and re-advertised with label value of 3. The new label would show up on PE-5 as a swap from the ingress label to an egress label of 3, although label 3 is not pushed on to the frame.

```
*A:PE-5# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
```

```
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push
  --                                        3
192.168.56.2                                1/1/2

192.0.2.6/32                                Swap
524282                                      3
192.168.56.2                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
```

Import and Export Policies

The default label handling behavior is to originate label bindings for the system address and to propagate all FECs received. If this is not the desired behavior, an import/export policy can be applied. An LDP import policy impacts inbound filtering; an LDP export policy impacts outbound filtering. An export policy may be configured to control the set of LDP label bindings advertised by the LER (sending to LDP peers). As such, export policies are used to include additional FECs rather than filtering FECs from those advertised. An import policy can be used to control for which FECs a router will generate labels (accepting from LDP peers). This functionality is not unique to LDP; it can be used for RSVP-TE, OSPF, and IS-IS as well as others.

The policy can be global or LDP peer FEC prefix filtering, both for import and export. LDP peer FEC prefix filtering uses a similar policy context as the LDP global policies and works in addition to these global policies.

```
*A:PE-1# tree flat detail | match import-pref
configure router ldp session-parameters peer import-prefixes <policy-name>
                     [<policy-name>...(up to 5 max)]
configure router ldp session-parameters peer no import-prefixes
configure router ldp targeted-session import-prefixes <policy-name>
                     [<policy-name>...(up to 5 max)]
configure router ldp targeted-session no import-prefixes
*A:PE-1#
```

```
*A:PE-1# tree flat detail | match export-pref
configure router ldp session-parameters peer export-prefixes <policy-name>
                     [<policy-name>...(up to 5 max)]
configure router ldp session-parameters peer no export-prefixes
configure router ldp targeted-session export-prefixes <policy-name>
                     [<policy-name>...(up to 5 max)]
```

```
configure router ldp targeted-session no export-prefixes
*A:PE-1#
```

By default, no labels are generated for directly connected (local) interfaces. To change this behavior, an export policy is created and applied to the LDP instance. There is no configuration difference in defining an import and export policy.

A policy starts with the keyword **begin** and contains a list of entries (of which each has a number), and ends with the keyword **commit**. An entry typically contains matching criteria (however, it is not required in cases where everything matches) and a corresponding action. Entries without an action are considered incomplete and are rendered inactive. When processing the policy, the router executes the specified action on the first matching statement; it does not process any further matches. For this reason, entries must be sequenced correctly from most to least specific.

The configuration of the LDP export policy for local interfaces is as follows:

```
*A:PE-1# configure
    router
        policy-options
            begin
            policy-statement "LDP-export"
                entry 10
                    from
                        protocol direct
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

There are 11 active LDP bindings before applying the export policy, as shown earlier.

The LDP export or import policy is applied to the LDP instance on the router, with the **export** or **import** keyword.

```
*A:PE-1# configure router ldp export "LDP-export"
```

When the export policy is applied, the active LDP binding table contains additional entries: the local interfaces of PE-x.

```
*A:PE-1# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop
524287                                        --
  --                                          --

192.0.2.2/32                                Push
  --                                        524287
192.168.12.2                                1/1/1

192.0.2.2/32                                Swap
524286                                      524287
192.168.12.2                                1/1/1

---snip---

192.168.12.0/30                             Pop
524281                                        --
  --                                          --

192.168.14.0/30                             Pop
524275                                        --
  --                                          --

192.168.23.0/30                             Swap
524280                                      524280
192.168.12.2                                1/1/1

192.168.25.0/30                             Swap
524279                                      524279
192.168.12.2                                1/1/1

192.168.36.0/30                             Swap
524276                                      524276
192.168.12.2                                1/1/1

192.168.45.0/30                             Swap
524278                                      524278
192.168.14.2                                1/1/2

192.168.56.0/30                             Swap
524277                                      524277
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 18
===============================================================================
*A:PE-1#
```

OAM

The following operations, administration, and maintenance operations can be launched on an LDP LSP:

- oam lsp-ping
- oam lsp-trace

The following options are supported for LSP ping:

```
*A:PE-1# oam lsp-ping
 - lsp-ping <lsp-name> [path <path-name>]
  - lsp-ping bgp-label prefix <ip-prefix/prefix-length> [path-destination <ip-address>
                             [interface <if-name> | next-hop <ip-address>]]
  - lsp-ping prefix <ip-prefix/prefix-length> [path-destination <ip-address>
                    [interface <if-name> | next-hop <ip-address>]]
  - lsp-ping sr-isis prefix <ip-prefix/prefix-length> [igp-instance <igp-instance>]
                            [path-destination <ip-address> [interface <if-name> |
                            next-hop <ip-address>]]
  - lsp-ping sr-ospf prefix <ip-prefix/prefix-length> [igp-instance <igp-instance>]
                            [path-destination <ip-address> [interface <if-name> |
                            next-hop <ip-address>]]
  - lsp-ping sr-te <lsp-name> [path <path-name>] [path-destination <ip-address>
                   [interface <if-name> | next-hop <ip-address>]]
  - lsp-ping static <lsp-name> [assoc-channel <ipv4|non-ip|none>] [dest-global-id
                    <global-id> dest-node-id <node-id>] [force]
                    [path-type <active|working|protect>]
  - options common to all lsp-ping cases:  [detail] [fc <fc-name> [profile <in|out>]]
                   [interval <interval>] [send-count <send-count>] [size <octets>]
                   [src-ip-address <ip-address>] [timeout <timeout>]
                   [ttl <label-ttl>]
```

As an example, an LSP ping is sent from PE-1 to PE-6:

```
*A:PE-1# oam lsp-ping prefix 192.0.2.6/32
LSP-PING 192.0.2.6/32: 80 bytes MPLS payload
Seq=1, send from intf int-PE-1-PE-2, reply from 192.0.2.6
       udp-data-len=32 ttl=255 rtt=2.56ms rc=3 (EgressRtr)

---- LSP 192.0.2.6/32 PING Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 2.56ms, avg = 2.56ms, max = 2.56ms, stddev = 0.000ms
*A:PE-1#
```

An LSP trace is sent from PE-1 to PE-6:

```
*A:PE-1# oam lsp-trace prefix 192.0.2.6/32
lsp-trace to 192.0.2.6/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.2  rtt=1.34ms rc=8(DSRtrMatchLabel) rsc=1
2  192.0.2.3  rtt=2.04ms rc=8(DSRtrMatchLabel) rsc=1
3  192.0.2.6  rtt=2.91ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

The return code (**rc**) is 8 for the LSRs and 3 for the eLER.

The detailed output for this LSP trace includes the interface IP address, the interface type, maximum receive unit (MRU), label and protocol; as follows:

```
*A:PE-1# oam lsp-trace prefix 192.0.2.6/32 detail
```

```
lsp-trace to 192.0.2.6/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.2  rtt=1.33ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.23.2 ifaddr=192.168.23.2 iftype=ipv4Numbered MRU=1564
           label[1]=524282 protocol=3(LDP)
2  192.0.2.3  rtt=2.24ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.36.2 ifaddr=192.168.36.2 iftype=ipv4Numbered MRU=1564
           label[1]=3 protocol=3(LDP)
3  192.0.2.6  rtt=3.05ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

LDP Statistics

LDP-related statistics can be collected in files. First a file needs to be configured.

```
*A:PE-1# configure
    log
        file-id 1
            location cf1:
            rollover 5 retention 1
        exit
```

The next step is to configure an accounting policy that will define which statistics should be recorded, for example as follows:

```
*A:PE-1# configure
    log
        accounting-policy 1
            record combined-ldp-lsp-egress
            to file 1
            no shutdown
        exit
```

The collection of statistics for prefix 192.0.2.6/32 is enabled on PE-1 in the LDP context, as follows:

```
*A:PE-1# configure
    router
        ldp
            egress-statistics
                fec-prefix 192.0.2.6/32
                    collect-stats
                    accounting-policy 1
                    no shutdown
                exit
            exit
```

The following FEC egress statistics can be displayed:

```
*A:PE-1# show router ldp fec-egress-stats
  - fec-egress-stats [<ip-prefix/ip-prefix-length>]
  - fec-egress-stats [active] [family]
```

```
 <ip-prefix/ip-pref*> : ipv4-prefix    - a.b.c.d
                        ipv4-prefix-le - [0..32]
                        ipv6-prefix    - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
 <active>             : keyword
 <family>             : ipv4|ipv6

*A:PE-1#
```

The FEC egress stats for prefix 192.0.2.6/32 can be retrieved as follows:

```
*A:PE-1# show router ldp fec-egress-stats 192.0.2.6/32
===============================================================================
LDP IPv4 FEC Egress Statistics
===============================================================================
-------------------------------------------------------------------------------
FEC Prefix/Mask    : 192.0.2.6/32
-------------------------------------------------------------------------------
Collect Stats      : Enabled               Accounting Plcy.     : 1
Admin State        : Up
FC BE
InProf Pkts        : 0                     OutProf Pkts         : 7
InProf Octets      : 0                     OutProf Octets       : 858
FC L2
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC AF
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC L1
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC H2
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC EF
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC H1
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
FC NC
InProf Pkts        : 0                     OutProf Pkts         : 0
InProf Octets      : 0                     OutProf Octets       : 0
===============================================================================
LDP IPv4 FEC Egress Statistics: 1
===============================================================================
*A:PE-1#
```

Statistics can be cleared as follows:

```
*A:PE-1# clear router ldp fec-egress-statistics 192.0.2.6/32
```

Debug

LDP debugging can be configured per LDP interface or per LDP peer, as follows:

```
*A:PE-1# debug router ldp
  - ldp
  - no ldp

 [no] interface       + Enable/disable and configure debugging for an LDP interface
 [no] peer            + Enable/disable and configure debugging for an LDP peer

*A:PE-1#
```

A particular peer is specified by its IPv4 or IPv6 address. It is possible to configure debugging for specific LDP events: bindings or messages, as follows:

```
*A:PE-1# debug router ldp peer 192.0.2.2
  - no peer <ip-address>
  - peer <ip-address>

 <ip-address>         : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D

 [no] event           + Configure debugging for specific LDP events
 [no] packet          + Enable/disable debugging for specific LDP packets


*A:PE-1# debug router ldp peer 192.0.2.2 event
  - event
  - no event

 [no] bindings        - Enable/disable debugging for LDP bindings
 [no] messages        - Enable/disable debugging for LDP messages
*A:PE-1#
```

It is also possible to configure debugging for specific packets, such as label packets:

```
*A:PE-1# debug router ldp peer 192.0.2.2 packet
  - no packet
  - packet

 [no] hello           - Enable/disable debugging for LDP Hello packets
 [no] init            - Enable/disable debugging for LDP Init packets
 [no] keepalive       - Enable/disable debugging for LDP Keepalive packets
 [no] label           - Enable/disable debugging for LDP Label packets
```

The following debugging is configured on PE-1:

```
*A:PE-1# debug router ldp peer 192.0.2.2 packet label detail
```

Some label mapping packets sent to peer 192.0.2.2 are the following, with the label mapping for prefixes 192.0.2.1/32 and 192.0.2.4/32:

```
2 2018/09/06 14:18:59.71 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 7) to 192.0.2.2:0
Protocol version = 1
Label 524284 advertised for the following FECs
Prefix Address Family = 1 Prefix = 192.0.2.4/32
"
1 2018/09/06 14:18:59.71 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 6) to 192.0.2.2:0
Protocol version = 1
Label 524274 advertised for the following FECs
Prefix Address Family = 1 Prefix = 192.0.2.1/32
"
```

Conclusion

MPLS provides the capability to establish connection-oriented paths over a connectionless network. LDP point-to-point LSPs are dynamically signaled and FRR is supported. This can greatly improve network resiliency. In this chapter, the configuration of several LDP point-to-point LSP features is given together with the associated show output which can be used to verify and troubleshoot.

LDP-IGP Synchronization

This chapter provides information about LDP-IGP Synchronization

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS Release 14.0.R6, but the CLI in the current edition is based on SR OS Release 15.0.R1.

Label Distribution Protocol - Interior Gateway Protocol (LDP-IGP) synchronization based on RFC 5443 is supported in SR OS Release 6.0, and later. LDP end-of-Label Information Base (LIB), as described in RFC 5919, is supported in SR OS Release 14.0.R1, and later.

Overview

Within an MPLS network using LDP, it is common practice to enable a synchronization timer between LDP and the IGP to give both the IGP and LDP time to converge after a link is restored. Without LDP-IGP synchronization, the IGP and LDP converge independently. Because the IGP converges before LDP, traffic can be black-holed until LDP has converged. When the IGP converges after link restoration and a new next hop is available, this change in next hop causes LDP to stop using the LDP labels for the alternate path. After the adjacency with the new next hop is established, labels are allocated for the new shortest (primary) path. These new labels are not yet signaled by LDP, causing the traffic to be black-holed for all or part of the FECs until LDP converges.

LDP-IGP synchronization based on RFC 5443 consists of temporarily setting the runtime IGP cost of a restored link to infinity to give time for both IGP and LDP to converge. When the LDP synchronization timer expires, the runtime IGP cost is restored to the configured IGP cost and IGP will re-advertise it and use this for the next shortest path first (SPF) computation. The value for infinity of the IGP cost for a router interface depends on the IGP: 0xFFFF (65535) for OSPF, 0x3F (63) for IS-IS regular metric, and 0xFFFFFE (16777214) for IS-IS wide metric. LDP-IGP synchronization is not supported on RIP interfaces.

When the system converges, the IGP starts the LDP synchronization timer when the LDP session to the neighbor is established over the interface. The LDP synchronization timer is running during the exchange of label FEC bindings over the interface. When the LDP synchronization timer expires, the IGP announces the new best next hop and LDP uses this next hop if the label bindings for the neighbor's FEC are available. However, the LDP synchronization timer does not guarantee that all FEC bindings will be exchanged when the timer expires. Operators do not want to configure very large timers on every node, which may result in long synchronization times. The end-of-LIB option (RFC 5919) reduces the synchronization time; therefore, operators can configure large synchronization timers that will be aborted when the end-of-LIB notification has been received from a downstream node.

By default, LDP-IGP synchronization is enabled for IS-IS and for OSPF, as follows:

```
*A:PE-1# configure router isis
*A:PE-1>config>router>isis# info detail | match ldp-sync
            no disable-ldp-sync

*A:PE-1>config>router>ospf# info detail | match ldp-sync
            no disable-ldp-sync
```

By default, LDP synchronization is disabled (out-of-service) on each interface, as follows:

```
*A:PE-1# show router isis interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync        : outOfService                    Ldp Sync Wait   : Disabled
Ldp Timer State : Disabled                        Ldp Tm Left     : 0

*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync         : outOfService         Ldp Sync Wait    : Disabled
Ldp Timer State  : Disabled             Ldp Tm Left      : 0
```

In SR OS Release 14.0.R1, and later, LDP end-of-LIB is supported, as defined in RFC 5919. LDP end-of-LIB allows a downstream node to notify its upstream peer that the node has advertised its entire LIB to its upstream peer, which can terminate the LDP synchronization timer. LDP end-of-LIB notifications use a FEC TLV with the type wildcard FEC element for all negotiated FEC types. LDP end-of-LIB is sent even if the system has no label bindings to advertise. Each node notifies its peer nodes that it is safe to send LDP end-of-LIB notifications even if the node is not configured to process them. The node sends an unrecognized notification capability TLV (RFC 5919) in the initialization message, indicating that it will ignore notification messages that carry status TLV with a non-fatal status code unknown to it.

The LDP synchronization timer is configured in seconds with a maximum of 1800 seconds on a per interface basis, as follows:

```
*A:PE-1# configure router interface "int-PE-1-P-2" ldp-sync-timer
  - ldp-sync-timer <seconds> [end-of-lib]
  - no ldp-sync-timer

 <seconds>            : [1..1800]
 <end-of-lib>         : keyword
```

As an example, an LDP synchronization timer of 300 seconds can be configured on interface "int-PE-1-P-2", with or without the LDP end-of-LIB option, as follows:

```
*A:PE-1# configure router interface "int-PE-1-P-2" ldp-sync-timer 300
```

```
*A:PE-1# configure router interface "int-PE-1-P-2" ldp-sync-timer 300 end-of-lib
```

- When the end-of-LIB option is not configured, the LDP synchronization timer is started when the LDP hello adjacency comes up over the interface. Any received LDP end-of-LIB message is ignored.
- When the end-of-LIB option is configured, the receiving node behaves as follows:
  - The LDP synchronization timer is started when the LDP hello adjacency comes up over the interface.
  - When LDP end-of-LIB type wildcard FEC messages have been received for all negotiated FEC types for a certain session to an LDP peer for the IGP interface, the LDP synchronization timer is terminated and the system restores the IGP link cost.
  - If the LDP synchronization timer expires before the LDP end-of-LIB messages are received for all negotiated FEC types, the system restores the IGP link cost.
  - All unexpected LDP end-of-LIB messages are dropped.
- When the end-of-LIB option is configured, the sending node will advertise an LDP end-of-LIB message for all FECs (prefix and P2MP FECs) after all FECs are sent for all peers that have advertised the unrecognized notification capability TLV.

When a user changes the IGP cost of an interface, the new value is advertised at the next flooding of link attributes by the IGP. If the LDP synchronization timer is running, the new cost value will only be advertised after the timer expires. However, the following commands can be used to terminate the LDP-IGP synchronization, causing the new IGP cost value to be advertised instantly:

```
*A:PE-1# tools perform router isis ldp-sync-exit
*A:PE-1# tools perform router ospf ldp-sync-exit
*A:PE-1# configure router interface "int-PE-1-P-2" no ldp-sync-timer
*A:PE-1# configure router ospf disable-ldp-sync
*A:PE-1# configure router isis disable-ldp-sync
```

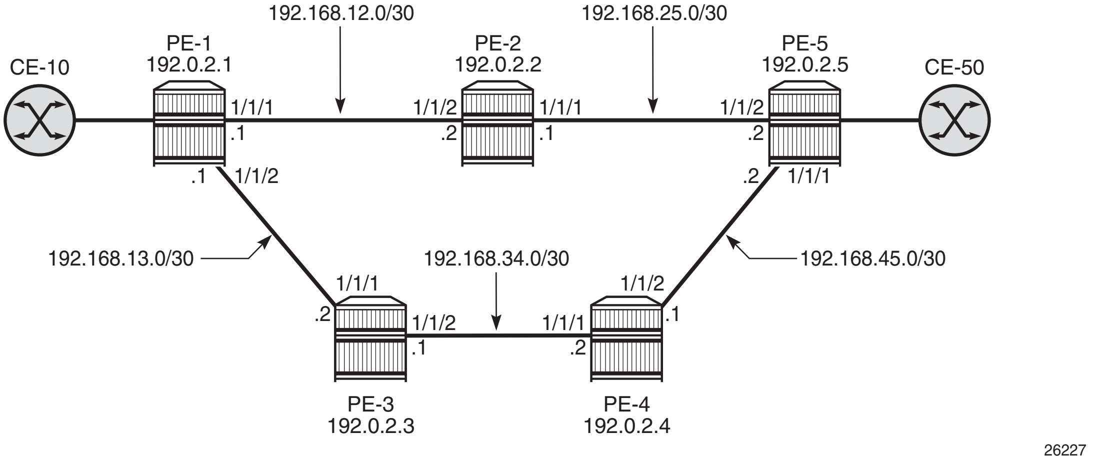
The first two commands do not modify the configuration; they terminate the LDP synchronization timer and restore the actual cost of the IGP interface. The last three commands disable the LDP-IGP configuration entirely, either from the interface or globally for the IGP (OSPF or IS-IS).

If the user changes the value of the LDP synchronization timer parameter, the new value will take effect at the next synchronization event. If the timer is still running, it will continue to use the previous value.

Configuration

Figure 250 shows the example topology.

Figure 250 **Example Topology**



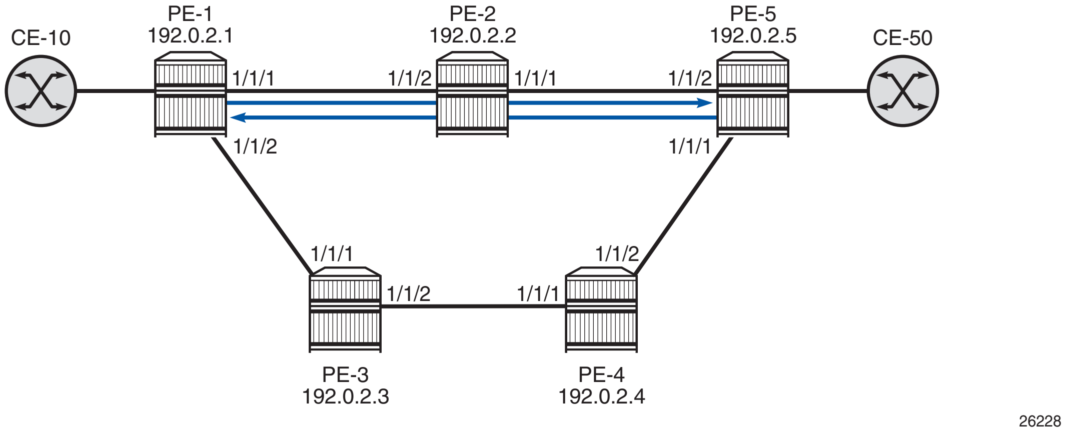
The initial configuration on these nodes includes the following:

- Cards, MDAs, ports

- Router interfaces
- IGP: OSPF on all interfaces between the five P/PE routers (alternatively, IS-IS can be configured)
- LDP on all interfaces (LDP link adjacencies)
- Services on the PEs; for example, an Epipe between PE-1 and PE-5 (LDP targeted adjacencies)
- In this test topology, CE-10 and CE-50 correspond to VPRN 10 on PE-1 and PE-5 using a hairpin to loop the traffic back to the node.

Default IGP metrics are used on the interfaces and, under normal conditions, traffic between CE-10 and CE-50 is sent over the shortest path via P-2, as shown in Figure 251.

Figure 251 **Shortest Path between PE-1 and PE-5**



LDP-IGP Synchronization without LDP End-of-LIB

LDP-IGP synchronization is, by default, globally enabled for OSPF and IS-IS, but disabled on every interface. In this example, LDP-IGP synchronization will be configured with an LDP synchronization timer of 300 seconds on all the interfaces in all the nodes, as follows:

```
*A:PE-1# configure router interface "int-PE-1-P-2" ldp-sync-timer 300
*A:PE-1# configure router interface "int-PE-1-P-3" ldp-sync-timer 300

*A:P-2# configure router interface "int-P-2-PE-1" ldp-sync-timer 300
*A:P-2# configure router interface "int-P-2-PE-5" ldp-sync-timer 300
```

The configuration is similar on the other nodes. With this configuration, a restored interface will temporarily get an IGP cost of infinity; therefore, the link will not be used for data traffic until the LDP synchronization timer terminates (when it expires after 300 seconds or when it is terminated manually). To simulate a link failure, port 1/1/1 is disabled (shutdown) and re-enabled (no shutdown) on PE-1, as follows:

```
*A:PE-1# configure port 1/1/1 shutdown
*A:PE-1# configure port 1/1/1 no shutdown
```

The LDP synchronization timer is not started before the LDP hello adjacency is established. The following output shows the port re-enabled, but before the LDP adjacency is established (Ldp Timer State = Wait for Ldp Adj.):

```
*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync         : inService            Ldp Sync Wait    : Disabled
Ldp Timer State  : Wait for Ldp Adj.    Ldp Tm Left      : 0
```

The following debug messages for OSPF show that the OSPF interface state is up (point-to-point), the LDP synchronization timer is state is updated to “waiting for adjacency”, and afterward the LDP state is updated to “LDP interface has adjacency”, as follows:

```
256 2017/03/30 08:07:26.60 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
IF 192.168.12.1 Idx 2 Event: IF_UP state: from DOWN to PTP"

259 2017/03/30 08:07:26.59 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
Updated the LDP Sync Timer state for I/F 2 to WAIT_FOR_ADJ"

263 2017/03/30 08:07:30.70 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
OSPF I/F 2 LDP state: new LDP_INTF_HAS_ADJ old LDP_INTF_DOWN"
```

When the LDP hello adjacency is established, the interface between PE-1 and P-2 gets an IGP cost of infinity and the LDP synchronization timer is started, as follows:

```
264 2017/03/30 08:07:30.70 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
Updated the LDP Sync Timer state for I/F 2 to TMR_ACTIVE"
```

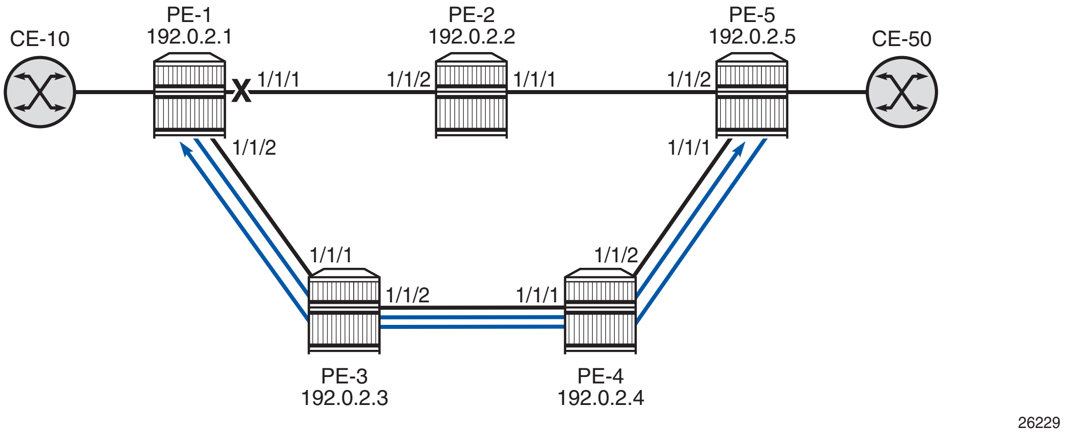
LDP bindings are exchanged as follows, but no message indicates the end-of-LIB (and if it were sent by P-2, it would be ignored by PE-1). The LDP synchronization timer is not automatically terminated when the LDP bindings are received, because the configuration does not include the end-of-LIB option.

```
265 2017/03/30 08:07:30.97 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262140 for Prefix Address Family = 1 Prefix = 192.0.2.3/
32 to peer 192.0.2.2:0."
```

```
266 2017/03/30 08:07:30.97 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262139 for Prefix Address Family = 1 Prefix = 192.0.2.4/
32 to peer 192.0.2.2:0."

267 2017/03/30 08:07:30.97 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262138 for Prefix Address Family = 1 Prefix = 192.0.2.5/
32 to peer 192.0.2.2:0."

268 2017/03/30 08:07:31.10 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262143 for Prefix Address Family = 1 Prefix = 192.0.2.1/
32 to peer 192.0.2.2:0."
```

As long as the LDP synchronization timer is not terminated, traffic between CE-10 and CE-50 is redirected to the path via P-3 and P-4, as shown in Figure 252.

Figure 252 **Rerouting via P-3 and P-4 until LDP Synchronization Timer Terminates**



The following commands for the OSPF interfaces between PE-1 and P-2 show the LDP synchronization timer status (active), LDP synchronization waiting state (enabled; therefore, traffic is rerouted), and the remaining time:

```
*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync          : inService            Ldp Sync Wait     : Enabled
Ldp Timer State   : Timer Active         Ldp Tm Left       : 299

*A:P-2# show router ospf interface "int-P-2-PE-1" detail | match Ldp
Ldp Sync          : inService            Ldp Sync Wait     : Enabled
Ldp Timer State   : Timer Active         Ldp Tm Left       : 294
```

The restored interface between PE-1 and P-2 will have an infinite IGP cost, so will not be used for data traffic as long as the LDP synchronization timer is active. All traffic between the CEs takes the path via P-3 and P-4, which can be verified as follows. The port statistics are cleared and 1000 ICMP echo requests are sent by CE-10 to CE-50. On PE-1, port 1/1/1 is used toward P-2 and port 1/1/2 is used toward P-3. All traffic is expected to take the path toward P-3. However, there will be some IGP and LDP signaling on all interfaces, so the packet count will be slightly greater than 1000, as follows:

```
*A:PE-1# clear port 1/1/[1..2] statistics

*A:PE-1# ping router 10 172.16.10.2 rapid count 1000
PING 172.16.10.2 56 data bytes
---snip---
---- 172.16.10.2 PING Statistics ----
1000 packets transmitted, 1000 packets received, 0.00% packet loss
round-trip min = 1.61ms, avg = 1.80ms, max = 3.59ms, stddev = 0.213ms

*A:PE-1# show port port 1/1/[1..2] statistics

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/1                        13            1476              14            1549
===============================================================================

===============================================================================
Port Statistics on Slot 1
===============================================================================
Port                    Ingress         Ingress         Egress           Egress
Id                      Packets          Octets         Packets          Octets
-------------------------------------------------------------------------------
1/1/2                      1033          127055            1034          127158
===============================================================================
```

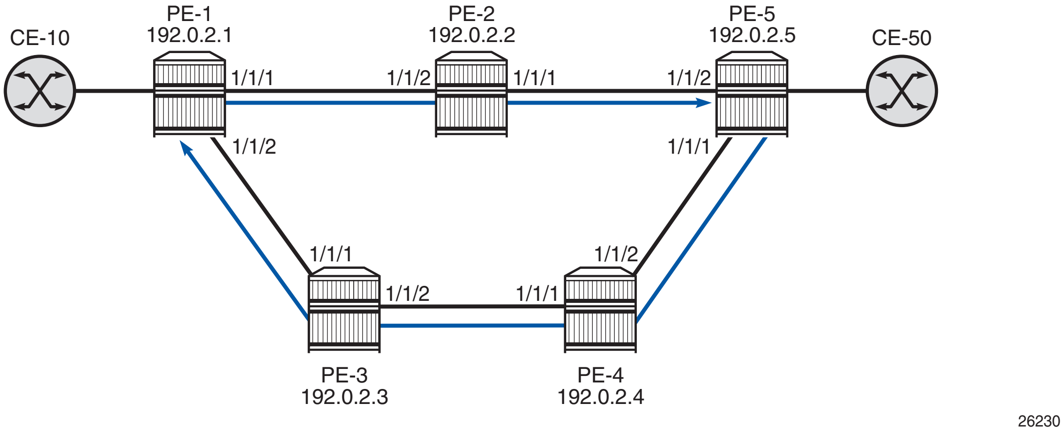
The port statistics on the other nodes will also show that these packets are sent via P-3 and P-4 instead of via P-2.

Even though the LIB was exchanged within seconds, the restored link only gets its normal IGP cost after the LDP synchronization timer has terminated. This can be done manually for a specific IGP (in this example, for OSPF on interface "int-PE-1-P-2" on PE-1) as follows:

```
*A:PE-1# tools perform router ospf ldp-sync-exit
Done.
*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync         : inService           Ldp Sync Wait    : Disabled
Ldp Timer State  : Manual Exit          Ldp Tm Left      : 0
```

The LDP synchronization timer can be configured independently for each IGP on each interface. The LDP synchronization timer for OSPF on interface "int-PE-1-P-2" is terminated manually (Ldp Timer State = Manual Exit; Ldp Sync Wait = Disabled; Ldp Tm Left = 0). Traffic from CE-10 to CE-50 can use interface "int-PE-1-P-2" because that interface has its configured (default) IGP cost. However, traffic from CE-50 to CE-10 will not use interface "int-P-2-PE-1" because that interface still has an infinite IGP cost as long as the LDP synchronization timer is not terminated; therefore, traffic toward CE-10 will pass via P-3 instead. This leads to an asymmetric traffic flow: the shortest path from CE-10 to CE-50 is via P-2, while the shortest path from CE-50 to CE-10 is via P-4 and P-3, as shown in Figure 253.

Figure 253 **Restored Link with One LDP Synchronization Timer Terminated**



When the second LDP synchronization timer is also terminated, the shortest path is via P-2 for all traffic between CE-10 and CE-50.

The LDP synchronization timer needs to be configured to a value that is long enough to prevent traffic being black-holed, but not too long to cause unnecessary suboptimal routing after the LIB has been exchanged and before the termination of the LDP synchronization timer. The end-of-LIB option reduces the LDP synchronization time when the configured LDP synchronization timer is longer than required for the exchange of the LIB, as described in the next section.

LDP synchronization is disabled on the interfaces of PE-1, as follows:

```
*A:PE-1# configure router interface "int-PE-1-P-2" no ldp-sync-timer
*A:PE-1# configure router interface "int-PE-1-P-3" no ldp-sync-timer
```

Similar commands to disable LDP synchronization on an interface can be configured on the other nodes.

LDP-IGP Synchronization with LDP End-of-LIB

The LDP synchronization is configured with the end-of-LIB option on all interfaces on all nodes; for example, for PE-1, as follows:

```
*A:PE-1# configure router interface "int-PE-1-P-2" ldp-sync-timer 300 end-of-lib
*A:PE-1# configure router interface "int-PE-1-P-3" ldp-sync-timer 300 end-of-lib
```

The configuration on the other nodes is similar.

A link failure is simulated by disabling and re-enabling port 1/1/1 on PE-1. Initially, the LDP synchronization timer state is waiting for LDP adjacency, as follows:

```
*A:PE-1# configure port 1/1/1 shutdown
*A:PE-1# configure port 1/1/1 no shutdown
*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync         : inService            Ldp Sync Wait    : Enabled
Ldp Timer State  : Wait for Ldp Adj.    Ldp Tm Left      : 0
```

After the LDP hello adjacency is established on the restored link, the LDP synchronization timer is started and PE-1 sends all LDP bindings to its peer P-2, as follows:

```
157 2017/03/30 09:05:59.41 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
OSPF I/F 2 LDP state: new LDP_INTF_HAS_ADJ old LDP_INTF_DOWN"


158 2017/03/30 09:05:59.41 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
Updated the LDP Sync Timer state for I/F 2 to TMR_ACTIVE"


160 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262143 for Prefix Address Family = 1
Prefix = 192.0.2.1/32 to peer 192.0.2.2:0."


161 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262138 for Prefix Address Family = 1
Prefix = 192.0.2.5/32 to peer 192.0.2.2:0."


162 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262139 for Prefix Address Family = 1
Prefix = 192.0.2.4/32 to peer 192.0.2.2:0."


163 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262140 for Prefix Address Family = 1
Prefix = 192.0.2.3/32 to peer 192.0.2.2:0."
```

```
164 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
Updated the LDP Sync Timer state for I/F 2 to EXCH_DONE"

165 2017/03/30 09:05:59.78 UTC MINOR: DEBUG #2001 Base OSPFv2
"OSPFv2: INTF
OSPF I/F 2 LDP state: new LDP_LBL_EXCH_DONE old LDP_INTF_HAS_ADJ"
```

When a downstream node has sent its entire LIB to its upstream peer, the node sends an end-of-LIB (RFC 5919) notification. When the upstream peer receives an end-of-LIB notification from its downstream peer, LDP is considered to be fully operational for the link. LDP triggers the IGP to advertise the link with normal cost instead of infinity and transit traffic can be sent on the restored link. The LDP synchronization timer state changes to label exchange done, as in the preceding debug messages and in the following show command output:

```
*A:PE-1# show router ospf interface "int-PE-1-P-2" detail | match Ldp
Ldp Sync         : inService            Ldp Sync Wait    : Disabled
Ldp Timer State  : Label Exchg. Done    Ldp Tm Left      : 0
```

The LDP synchronization timer is terminated when the entire LIB is exchanged. In this example setup, the LDP synchronization time is reduced from 300 seconds to less than 10 seconds after enabling LDP end-of-LIB.

Conclusion

LDP-IGP synchronization (RFC 5443) allows directly connected nodes to delay the use of a restored link for transit IP packets until the LDP labels have been exchanged. RFC 5919 adds the end-of-LIB option that reduces the LDP synchronization time to the minimum, so operators can configure large values for the LDP synchronization timer.

LDP-SR Stitching for IPv4 Prefixes (IS-IS)

This chapter provides information about LDP-SR Stitching for IPv4 Prefixes (IS-IS).

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS Release 14.0.R5. The CLI in the current edition is based on SR OS release 15.0.R1.

Overview

Segment Routing (SR) allows for the construction of source-routed Label Switched Paths (LSPs) where the series of hops to be taken through the network are indicated by one or more Segment Identifiers (SIDs) assigned at the ingress PE. In the case of an MPLS data plane, these SIDs are MPLS labels learned through extensions to the OSPF/IS-IS control plane. SR provides benefits to the MPLS data plane, such as high scalability (due to lack of soft-state), traffic engineering capability, and topology-independent fast reroute.

When SR is configured in an IP/MPLS network that runs the Label Distribution Protocol (LDP), it is possible that SR and LDP will coexist, in which case preference for LDP or SR is a local matter at the LSP head end. It is equally possible that not all devices will have the capability to support SR, in which case some kind of interworking between SR and LDP is necessary to create an end-to-end LSP. Fast reroute coverage can also benefit from this SR-LDP interworking function, where SR is used to increase Loop Free Alternate (LFA) coverage using Remote or Directed LFA.

This chapter describes the configuration requirements for the interworking of LDP and SR to form a single end-to-end LSP when using IS-IS as an IGP. The chapter shows how this interworking function can be used to extend fast reroute coverage for LDP-based LSPs.

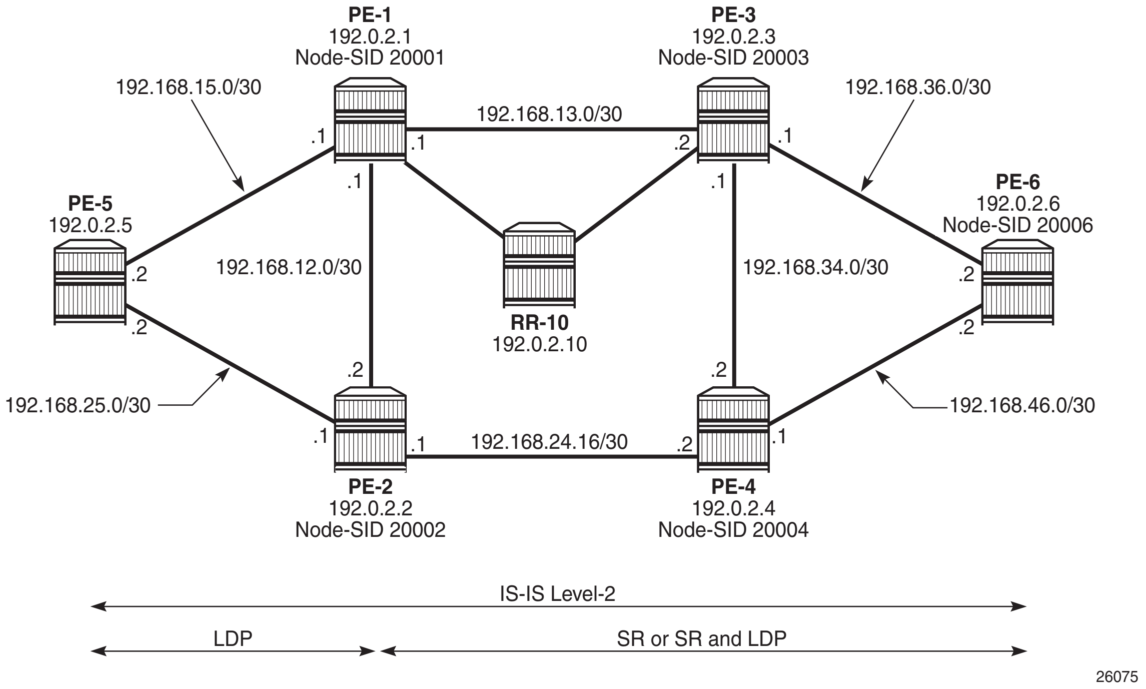
Configuration

Example Topology

The topology shown in Figure 254 provides an example of SR-LDP interworking. All routers within the topology form part of Autonomous System 64496 and are IBGP clients of RR-10 for the VPN-IPv4 address family. All routers in the topology belong to the same IS-IS Level-2 area, and all link metrics are set to 100. RR-10 does not participate in any MPLS data plane, and signals the IS-IS overload bit to avoid being used for transit traffic.

PE-5 is a router that does not support SR and, therefore, runs only LDP to its connected peers PE-1 and PE-2. PE-1, PE-2, PE-3, PE-4, and PE-6 are capable of running both SR and LDP, but are initially configured to only run SR with the associated node-SIDs shown in Figure 254. When explicitly described, LDP will be enabled in conjunction with SR on these routers to show the difference between the two approaches, and to show how SR can be used as a fast reroute backup for SR primary LSPs.

Figure 254 **Example Topology**



The LDP configuration at PE-1 toward PE-5 is shown in the following output. The configuration at PE-2 is similar with the only exception being IP addressing.

```
configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-5" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
            targeted-session
            exit
            no shutdown
        exit
```

PE-1, PE-2, PE-3, PE-4, and PE-6 run SR. The following output provides an example of the relevant SR configuration parameters at PE-1, with similar configurations on the remaining SR routers. For a description of these parameters, see chapter Segment Routing with IS-IS Control Plane.

```
configure
    router Base
        mpls-labels
            sr-labels start 20000 end 20099
        exit
        isis 0
            advertise-router-capability as
            interface "system"
                level-capability level-2
                ipv4-node-sid label 20001
                passive
                no shutdown
            exit
            segment-routing
                prefix-sid-range start-label 20000 max-index 99
                no shutdown
            exit
            no shutdown
        exit
```

SR Mapping Server

An SR Mapping Server (SR-MS) is an integral part of SR-LDP interoperability and has the responsibility for advertising prefixes to SID/label mappings on behalf of routers that do not support SR. When using IS-IS, a SID/label-binding TLV (TLV 149) containing a prefix-SID sub-TLV is used to advertise one or more SID index/labels and one or more prefixes. In the example topology, PE-4 is selected as the SR-MS and will advertise a prefix SID for the non-SR-capable router, PE-5.

The following output provides an example of the configuration required to implement SR-MS functionality. Under the **segment-routing** node, a **mapping-server** context is created that allows for origination of a SID/label-binding TLV and prefix-SID sub-TLV. The syntax begins with **sid-map node-sid** and is followed by an index. In SR OS, prefix-SIDs are always advertised with an index value (as opposed to an absolute label value), and the formula {start-label + SID index} is used to derive the label value. In this example, **index 5** is used and, therefore, the derived label value is {20000+5} 20005 for the PE-5 prefix 192.0.2.5/32.

An optional **range** argument allows for advertisement of a contiguous range of prefixes and associated SIDs using the configured index/prefix as the beginning of the range. Non-contiguous ranges require multiple entries and are advertised as separate SID/label-binding TLVs. An additional optional **set-flags s** argument can also be used to set the S-flag, which controls the flooding scope. When set, the flooding scope is the entire IS-IS domain. When not set, the flooding scope is the IS-IS level into which the TLV was advertised.

```
configure
    router
        isis 0
            segment-routing
                prefix-sid-range start-label 20000 max-index 99
                mapping-server
                    sid-map node-sid index 5 prefix 192.0.2.5/32
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The relevant part of the IS-IS LSP generated by PE-4, showing the SID/label-binding TLV, is shown in the following output:

```
*A:PE-4# show router isis database PE-4.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================
---snip---
```

```
TLVs :
  Area Addresses:
    Area Address : (3) 49.0001
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-4
  Router ID   :
    Router ID   : 192.0.2.4
  Router Cap : 192.0.2.4, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF
  SID Label Binding:
    Prefix: 192.0.2.5/32 Range:1 Weight:0 bFlgs:v4 SID:5 Algo:0 pFlgs:N
---snip---
```

At other routers within the SR domain, the presence of the advertised prefix can be validated as shown in the following output taken at PE-1. The SRMS field is set to Y for prefix 192.0.2.5/32, indicating that the prefix was advertised by an SR-MS. (In the case of IS-IS, the prefix-SID is a sub-TLV of the SID/label-binding TLV and the "N" (node-SID) flag is set; therefore, it can be recognized as being advertised by a mapping server.) The Y is followed by an "(S)" flag, indicating that the SRMS prefix-SID is selected to be programmed. This indication is provided in case there are multiple advertisements for the same prefix and/or node-SID from different SR mapping servers that result in some kind of conflict or inconsistency. If there are multiple mapping servers advertising the same prefix-SID, the advertising router with the lowest system/router ID is preferred.

```
*A:PE-1# show router isis prefix-sids

===============================================================================
Rtr Base ISIS Instance 0 Prefix/SID Table
===============================================================================
Prefix                            SID        Lvl/Typ    SRMS   AdvRtr
                                                         MT     Flags
-------------------------------------------------------------------------------
192.0.2.1/32                      1          2/Int.      N     PE-1
                                                            0       NnP
192.0.2.2/32                      2          2/Int.      N     PE-2
                                                            0       NnP
192.0.2.3/32                      3          2/Int.      N     PE-3
                                                            0       NnP
192.0.2.4/32                      4          2/Int.      N     PE-4
                                                            0       NnP
192.0.2.5/32                      5          2/Int.      Y(S)  PE-4
                                                            0       NnP
192.0.2.6/32                      6          2/Int.      N     PE-6
                                                            0       NnP
-------------------------------------------------------------------------------
No. of Prefix/SIDs: 6 (6 unique)
-------------------------------------------------------------------------------
SRMS : Y/N  = prefix SID advertised by SR Mapping Server (Y) or not (N)
       S    = SRMS prefix SID is selected to be programmed
Flags: R    = Re-advertisement
```

```
       N    = Node-SID
       nP   = no penultimate hop POP
       E    = Explicit-Null
       V    = Prefix-SID carries a value
       L    = value/index has local significance
===============================================================================
```

SR-LDP Interworking

Interworking SR and LDP essentially consists of stitching an LDP FEC and an SR node-SID route for the same prefix. In the example topology, PE-1 and PE-2 will act as the SR-LDP interworking nodes.

In the LDP-to-SR data plane direction, LDP uses an **export-tunnel-table** command under the **ldp** context to reference a policy that defines which prefixes should be redistributed from the IS-IS/SR domain into LDP. When applied, the LDP process monitors the tunnel-table until it locates a /32 SR tunnel of type sr-isis that matches a prefix defined in the export policy. LDP then programs an LDP Incoming Label Map (ILM) entry and stitches it to the SR node-SID tunnel endpoint. The LDP process also originates a FEC for the prefix and advertises that FEC to its peers.

The following output provides an example of the route policy and application of the policy at PE-1. In this policy, PE-1 advertises LDP FECs to PE-5 for PE-3 (192.0.2.3), PE-4 (192.0.2.4), and PE-6 (192.0.2.6), provided that PE-1 has a /32 SR tunnel of type sr-isis in the tunnel-table for those same prefixes. PE-1 also programs an LDP ILM entry for each prefix and stitches it to the appropriate SR tunnel. An identical configuration exists at PE-2.

```
configure
    router
        policy-options
            begin
            prefix-list "sr-domain"
                prefix 192.0.2.3/32 exact
                prefix 192.0.2.4/32 exact
                prefix 192.0.2.6/32 exact
            exit
            policy-statement "SR-to-LDP"
                entry 10
                    from
                        protocol isis
                        prefix-list "sr-domain"
                    exit
                    to
                        protocol ldp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
        ldp
            export-tunnel-table "SR-to-LDP"
            interface-parameters
                interface "int-PE-1-PE-5" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
            no shutdown
        exit
```

In the SR-to-LDP data plane direction, the **export-tunnel-table ldp** command within the **segment-routing** context is the only required configuration. Unlike the LDP-to-SR data plane direction, where policy is used to control which prefixes are stitched, in the SR-to-LDP direction, no policy is explicitly referenced because the SR-MS provides a network-wide policy for the prefixes that SR needs to stitch to a corresponding LDP FEC. With the **export-tunnel-table ldp** command applied, whenever a /32 LDP tunnel destination matches a prefix for which a prefix-SID sub-TLV was received from a mapping server, the SR ILM is stitched to the corresponding LDP tunnel endpoint.

The following output shows the configuration applied at PE-1 to implement SR-to-LDP data plane interworking:

```
configure
    router
        isis 0
            segment-routing
                export-tunnel-table ldp
                no shutdown
            exit
            no shutdown
        exit
```

With the required configuration in the SR-LDP interworking routers (PE-1 and PE-2), it is possible to validate the correct ILM entries. In the LDP-to-SR data plane direction, the following output shows the active LDP bindings at PE-1. Each of the entries for PE-3 (192.0.2.3), PE-4 (192.0.2.4), and PE-6 (192.0.2.6) have an "(I)" flag to indicate that the prefix has an SR-ISIS next-hop. Each entry also has an ingress label and an egress label. The ingress label represents the LDP FEC advertised for the corresponding prefix (in this case, advertised only to PE-5). The egress label represents the SR node-SID for the same prefix. Therefore, a mapping exists between LDP FEC and SR node-SID.

```
*A:PE-1# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
```

```
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop             262143      --
  --                                          --

192.0.2.3/32(I)                             Swap            262135    20003
192.168.13.2                                1/1/1:100

192.0.2.4/32(I)                             Swap            262136    20004
192.168.12.2                                1/1/3:100

192.0.2.5/32                                Push              --      262143
192.168.15.2                                1/1/2:100

192.0.2.6/32(I)                             Swap            262134    20006
192.168.13.2                                1/1/1:100

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 5
===============================================================================
```

In the SR-to-LDP data plane direction, the following output, taken at PE-1, shows a dump of the SR database for next-hops resolved to LDP. There is a single entry with index 5 (label value 20005) advertised by the SR-MS for the PE-5 prefix 192.0.2.5. The final line of the entry shows that an LDP FEC is the SID next-hop for SR-LDP stitching. The tunnel LSP ID is 65537. The tunnel-table verifies that this is an LDP tunnel to PE-5 (192.0.2.5).

```
*A:PE-1# tools dump router isis sr-database nh-type ldp detail
===============================================================================
Rtr Base ISIS Instance 0 SR Database
===============================================================================
-------------------------------------------------------------------------------
SID 5
-------------------------------------------------------------------------------
Label              : 20005                Adv System Id       : 1920.0000.2005
Prefix             : 192.0.2.5
Route Level        : 2                    MT Id               : 0
Rtm Preference     : 18                   Ttm Preference      : 0
Metric             : 0                    Last Action         : AddTnl
Num Ip NextHop     : 0                    Num SR-Tnl NextHop  : 1
```

```
Mtu                : 0
Mtu Prim           : 0                      Mtu Backup          : 0
Exclude from LFA   : 0                      Remote LFA          : 0
Duplicate Pending  : 0                      Tunnel Active State : Reported/Ack
SR Error           : SR_ERR_OK

LDP Next-Hop IP    : 192.0.2.5
Tunnel LSP Id      : 65537                  Tunnel Type         : 2

-------------------------------------------------------------------------------
No. of Entries: 1
-------------------------------------------------------------------------------
LDP = LDP FEC is the SID NH for SR-LDP stitching
===============================================================================
*A:PE-1#
```

To verify that the data plane is intact from end-to-end, a VPRN service is configured at the non-SR-capable PE-5 and the SR-capable PE-6, each with a locally configured subnet that is used to test IP connectivity. The configuration of the VPRN at PE-5 is shown in the following output. The **auto-bind-tunnel** configuration uses a resolution filter allowing only **ldp** to be used to resolve BGP next-hops for VPN-IPv4 routes. Usually, this could be configured for **resolution any**, but this configuration shows that LDP is being used. The local IP address at PE-5 is 172.31.5.1/24.

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "Local-Subnet" create
                address 172.31.5.1/24
                sap 1/2/1:1 create
                exit
            exit
            no shutdown
        exit
```

The configuration of the VPRN at PE-6 is shown in the following output. Again, the **auto-bind-tunnel** configuration uses a resolution filter, but this time it is configured for **sr-isis**. It could be set to **resolution any**, so that the tunnel-table preference would resolve an LSP with the lowest preference/metric, but the resolution filter configuration again shows that SR is being used. The **auto-bind-tunnel** context allows the transport mechanism to be a local decision at service level. The local IP address at PE-6 is 172.31.6.1/24.

Note: An alternative approach would be to configure the auto-bind-tunnel context for **resolution any**, then modify the tunnel-table preference for SR using the **tunnel-table-pref** command in the **segment-routing** context.

```
configure
    service
        vprn 1 customer 1 create
            route-distinguisher 64496:1
            auto-bind-tunnel
                resolution-filter
                    sr-isis
                exit
                resolution filter
            exit
            vrf-target target:64496:1
            interface "Local-Subnet" create
                address 172.31.6.1/24
                sap 1/2/1:1 create
                exit
            exit
            no shutdown
        exit
```

A VPRN ping between 172.31.5.1 at PE-5 and 172.31.6.1 at PE-6 verifies that the data plane is intact:

```
*A:PE-5# ping router 1 172.31.6.1 source 172.31.5.1
PING 172.31.6.1 56 data bytes
64 bytes from 172.31.6.1: icmp_seq=1 ttl=64 time=7.05ms.
64 bytes from 172.31.6.1: icmp_seq=2 ttl=64 time=6.93ms.
64 bytes from 172.31.6.1: icmp_seq=3 ttl=64 time=6.88ms.
64 bytes from 172.31.6.1: icmp_seq=4 ttl=64 time=6.54ms.
64 bytes from 172.31.6.1: icmp_seq=5 ttl=64 time=9.98ms.
---- 172.31.6.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 6.54ms, avg = 7.47ms, max = 9.98ms, stddev = 1.26ms
```

SR and LDP Coexistence

The previous example demonstrates the use of SR-LDP interworking when the SR domain runs only SR. A more common scenario is that SR will coexist with LDP, because LDP is already deployed and the SR deployment will be added. In this sub-section, PE-1, PE-2, PE-3, PE-4, and PE-6 are configured to run LDP in conjunction with SR. PE-5 remains the same as the previous sub-section in that it runs only LDP to its connected peers PE-1 and PE-2.

In the SR-to-LDP data plane direction, there is no notable change when LDP coexists in the SR domain. Whenever a /32 LDP tunnel destination matches a prefix for which a prefix-SID sub-TLV was received from a mapping server, the SR ILM is stitched to the corresponding LDP tunnel endpoint.

In the LDP-to-SR data plane direction, there is a significant change. If only SR is running within the SR domain, the LDP process monitors the tunnel-table and when a /32 SR tunnel of type sr-isis is found that matches a prefix in the (**export-tunnel-table**) export policy, LDP programs an LDP ILM and stitches it to the SR node-SID tunnel endpoint. However, if an LDP FEC exists for the same /32 prefix, SR OS will resolve the LDP ILM entry to the LDP FEC. This is because LDP attempts to resolve the prefix in the route table first before looking in the tunnel-table and, therefore, prefers the LDP tunnel to the SR tunnel.

The following output is taken at PE-1 when LDP and SR coexist in the SR domain. The previous version of this output (when LDP was not running in the SR domain) showed the prefixes for PE-3, PE-4, and PE-6 as known via an SR-ISIS next-hop, and the egress labels as node-SIDs. When LDP is active in conjunction with SR, the egress labels resolve to an LDP FEC.

```
*A:PE-1# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Pop             262143      --
  --                                          --

192.0.2.2/32                                Push              --      262142
192.168.12.2                                1/1/3:100

192.0.2.2/32                                Swap            262139    262142
192.168.12.2                                1/1/3:100

192.0.2.3/32                                Push              --      262139
192.168.13.2                                1/1/1:100
```

```
192.0.2.3/32                                Swap            262135    262139
192.168.13.2                                1/1/1:100

192.0.2.4/32                                Push              --      262136
192.168.12.2                                1/1/3:100

192.0.2.4/32                                Swap            262136    262136
192.168.12.2                                1/1/3:100

192.0.2.5/32                                Push              --      262143
192.168.15.2                                1/1/2:100

192.0.2.5/32                                Swap            262138    262143
192.168.15.2                                1/1/2:100

192.0.2.6/32                                Push              --      262134
192.168.13.2                                1/1/1:100

192.0.2.6/32                                Swap            262134    262134
192.168.13.2                                1/1/1:100

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 11
===============================================================================
```

That the LDP-to-SR data path resolves to LDP FECs rather than SR tunnels may result in an asymmetric data path. Taking the previously used VPRN service between PE-5 and PE-6 as an example:

- Traffic from PE-6 to PE-5 will use SR between PE-6 and one of the SR-LDP interworking gateways at PE-1 or PE-2, after which it will use LDP.
- Traffic from PE-5 to PE-6 will use LDP between ingress and egress. The interworking function between SR and LDP has no effect.

Both directions still use an MPLS data plane. However, the MPLS control plane differs in each direction.

LDP Fast Reroute Using SR Tunnels

With the ability to interwork LDP and SR, primary LSPs signaled using LDP can select a remote LFA SR tunnel as backup. This provides the potential to increase fast reroute coverage. As with any other backup or fast reroute mechanism, the SR backup tunnel can be installed in the forwarding database before any failure, but can only be activated when the failure of the primary path has been detected.

The ability to detect a failure quickly forms a significant part of the overall reconvergence time and may require the use of failure detection mechanisms, such as Bidirectional Forwarding Detection (BFD), the 802.3ah Ethernet in the First Mile (EFM), or just Loss of Signal (LoS). These mechanisms are beyond the scope of this chapter.

To use SR as a backup for LDP, the **fast-reroute backup-sr-tunnel** command must be configured in the **ldp** context. The **export-tunnel-table** command previously described should also be present, and should reference a policy including all of the prefixes for which backup is required. There is no requirement for an SR-MS when using SR tunnels for LDP backup, nor is there a requirement to enable SR-to-LDP interworking using the **export-tunnel-table ldp** command within the **segment-routing** context.

The following output shows the configuration applied at PE-6. When this configuration is applied, if the LFA SPF does not find an adjacent IP next-hop prefix for an LDP FEC, but can compute a remote LFA tunnel next-hop, LDP programs the LDP FEC using an LDP Next-Hop Label Forwarding Entry (NHLFE), and a backup next-hop using an LDP NHLFE pointing to the SR tunnel endpoint. The LDP packet is not tunneled over the SR tunnel, but rather the LDP label is stitched to the segment-routing label stack. This behavior is similar to the LDP-SR interworking function previously described within this chapter, but is modified such that the stitching of an LDP ILM entry to an SR tunnel only takes place if no adjacent LFA next-hop could be found for the prefix.

```
configure
    router Base
        isis
            loopfree-alternate remote-lfa
        exit
        ldp
            export-tunnel-table "SR-to-LDP"
            fast-reroute backup-sr-tunnel
        exit
        policy-options
            begin
            prefix-list "sr-domain"
                prefix 192.0.2.0/24 longer
            exit
            policy-statement "SR-to-LDP"
                entry 10
                    from
                        protocol isis
                        prefix-list "sr-domain"
                    exit
                    to
                        protocol ldp
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

With the preceding configuration in place at PE-6, it is possible to verify whether a backup exists for a specific prefix, using the command shown in the following output. In this example, the backup is displayed for the PE-6 adjacent neighbor PE-3 (192.0.2.3). There are two LSPs for the prefix 192.0.2.3/32; one is known via LDP and one is known via SR-ISIS, indicated in the protocol column. The entries are defined as follows:

- The first line of the LDP entry is the primary LSP with a next-hop of 192.168.36.1 using interface 1/1/2:100 (direct to PE-3).
- The second line of the LDP entry is the backup indicated by a "(B)" flag, with a next-hop of 192.168.46.1 using interface 1/1/1:100 (via PE-4). This backup is a basic LFA, which is possible to compute due to the example topology, or more explicitly the triangular mesh between PE-6, PE-4, and PE-3. Due to this topology, if the link between PE6 and PE3 fails, PE-6 can forward packets destined for PE-3 toward PE-4. PE-4 will then forward them directly toward PE-3, not return them to PE-6 (which would create a transient micro-loop until the next SPF is run).
- The first line of the SR-ISIS entry is the primary LSP with a next-hop of 192.168.36.1 using interface 1/1/2:100 (direct to PE-3).
- The second line of the SR-ISIS entry is the backup LSP indicated by the "(B)" flag, with a next-hop of 192.168.46.1 using interface 1/1/1:100 (via PE-4). Both the primary and backup LSPs use the label 20003, representing the PE-3 node-SID. As with the LDP backup entry, the SR-ISIS backup is a basic LFA.

```
*A:PE-6# show router fp-tunnel-table 1 192.0.2.3/32

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.3/32                                 LDP                  -
  262139
  192.168.36.1                                                 1/1/2:100
  262137
  192.168.46.1(B)                                              1/1/1:100
192.0.2.3/32                                 SR-ISIS-0            -
  20003
  192.168.36.1                                                 1/1/2:100
  20003
  192.168.46.1(B)                                              1/1/1:100
-------------------------------------------------------------------------------
Total Entries : 2
```

```
-------------------------------------------------------------------------------
===============================================================================
```

To show the benefits that SR provides in increasing fast reroute coverage, the link between PE-4 and PE-3 is removed from the example topology, creating a ring topology. With this link removed, it is no longer possible for PE-6 to compute a basic LFA to PE-3 for the link between PE-6 and PE-3. If that link failed and PE-6 forwarded packets destined for PE-3 toward PE-4, PE-4 would return them to PE-6 until the next SPF was complete. Therefore, a backup tunnel is needed to a place in the network that will not loop packets back; essentially a remote LFA.

The following output at PE-6 shows the primary and backup LSPs for PE-3 (192.0.2.3) with the modified topology. Again there are two LSPs: one known through via LDP and one known via SR-ISIS. The entries are defined as follows:

- The first line of the LDP entry is the primary LSP with a next-hop of 192.168.36.1 using interface 1/1/2:100 (direct to PE-3).
- The second line of the LDP entry is the backup indicated by a "(B)" flag with a next-hop of 192.0.2.3 (PE-3), which uses an SR tunnel. The label of "3" (implicit-null) indicates that the LDP label is not tunneled through the SR tunnel, but rather popped before the primary LDP LSP is stitched to the backup SR LSP.
- The first line of the SR-ISIS entry is the primary LSP with a next-hop of 192.168.36.1 using interface 1/1/2:100 (direct to PE-3). This LSP assigns a single label of value 20003, representing the node-SID of PE-3.
- The second line of the SR-ISIS entry is the backup indicated by a "(B)" flag with a next-hop of 192.168.46.1 using interface 1/1/1:100 (via PE-4). There are two labels assigned to this backup tunnel. The upper label has a value of 20002, which represents the node-SID of PE-2. This is the remote LFA "PQ-node". The second label has a value of 20003, which represents the node-SID of the destination, PE-3.

  When this backup tunnel is operational, PE-6 encapsulates traffic destined for PE-3 to a point in the network where it will not be looped back toward the source. In the example topology, that node is PE-2. When traffic arrives at PE-2, it pops the top label (20002) and forwards traffic for PE-3 (with label 20003) on the shortest path toward the destination.

```
*A:PE-6# show router fp-tunnel-table 1 192.0.2.3/32

===============================================================================
IPv4 Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                               Protocol   Tunnel-ID
    Lbl                   NextHop            Intf/Tunnel
-------------------------------------------------------------------------------
Destination                                  Protocol            Tunnel-ID
```

```
  Lbl
    NextHop                                                         Intf/Tunnel
--------------------------------------------------------------------------------
192.0.2.3/32                                 LDP                   -
  262139
  192.168.36.1                                                  1/1/2:100
  3
  192.0.2.3(B)                                                  SR
192.0.2.3/32                                 SR-ISIS-0             -
  20003
  192.168.36.1                                                  1/1/2:100
  20003/20002
  192.168.46.1(B)                                               1/1/1:100
--------------------------------------------------------------------------------
Total Entries : 2
--------------------------------------------------------------------------------
================================================================================
```

Conclusion

The SR control plane can (and likely will) coexist with other MPLS control plane clients, such as RSVP, LDP, or BGP. It is possible that these control plane clients will operate independently. However, where a mix of SR-capable and non-SR-capable routers exist within the same domain, SR-LDP interworking is necessary to form an end-to-end LSP. This chapter shows how that is possible using one or more SR mapping servers and one or more interworking routers.

SR-LDP interworking also provides an opportunity to increase fast reroute coverage in LDP-based networks. Before the introduction of SR-LDP interworking, a remote LFA could only be constructed using LDP-over-RSVP, which required the RSVP LSP to be manually configured and placed. When SR-LDP interworking is used, primary LDP LSPs can use a backup tunnel to a remote LFA signaled using SR. This requires no manual configuration, which provides the potential to greatly increase fast reroute coverage with minimal effort.

MPLS LDP FRR using ISIS as IGP

This chapter describes Multi- Protocol Label Switching (MPLS) Label Distribution Protocol (LDP) Fast Reroute (FRR) using Intermediate System to Intermediate System (IS-IS) as the Interior Gateway Protocol (IGP).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS release 9.0.R6, but the CLI in the current edition corresponds to SR OS release 16.0.R3. There are no prerequisites for this configuration.

Overview

LDP FRR improves convergence in case of a single link or single node failure in the network. Convergence times will be in the order of tens of milliseconds. This is important to some application services, such as voice over IP (VoIP), which are sensitive to traffic loss when running over the MPLS network.

Without FRR, link and/or node failures inside an MPLS LDP network result in traffic loss in the order of hundreds of milliseconds. The reason for that is that LDP depends on the convergence of the underlying IGP (IS-IS sending link state PDUs (LSPs) in this case). After IGP convergence, LDP itself needs to compute new primary Next Hop Label Forwarding Entries (NHLFEs) for all affected Forwarding Equivalence Classes (FECs). Finally, the different Label Forwarding Information Bases (LFIBs) are updated.

When FRR is configured on a node, the node pre-computes primary NHLFEs for all FECs and, in addition, it will pre-compute backup NHLFEs for all FECs. The backup NHLFE corresponds to the label received for the same FEC from a Loop-Free Alternate (LFA) next hop (see also RFC 5286, *Basic Specification for IP Fast Reroute: Loop-Free Alternates*). Both primary NHLFEs and backup NHLFEs are programmed in the IOM/IMM, which makes it possible to converge very quickly.

The SR OS software has implemented Inequality 1 (link criterion) and Inequality 3 (node criterion) of RFC 5286. Similar to the Shortest Path Tree (SPT) computation that is part of standard link-state routing functionality, also the LFA next hop computation is based on the IGP metric.

The underlying LFA formulas appear in the following format:

Inequality 1:

- SP(backup NHR, D) < {SP(backup NHR, S) + SP(S, D)}

Inequality 3:

- SP(backup NHR, D) < {SP(backup NHR, PN) + SP(PN, D)}

In these inequalities ‘SP’ is ‘shortest IGP metric path’, ‘NHR’ is ‘next hop router’, ‘D’ is ‘destination’, ‘S’ is ‘source node or upstream node doing the actual LFA next-hop computation’, and ‘PN’ is ‘protected node’. The inequality 3 rule is stricter than the inequality 1 rule. See Additional Topics for a practical example on these inequalities.

Configuration

This section provides information to configure:

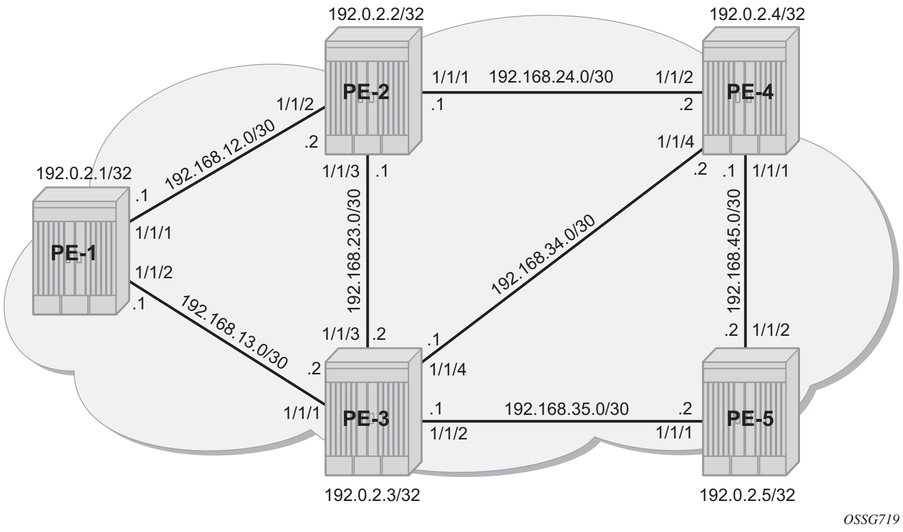
- Configure the IP/MPLS network
- Enable LDP FRR and Verification
- Enable Synchronization Timer
- Data Path Verification

Additional topics include:

- Metric Change
- IS-IS Overload Bit

Figure 255 shows the example topology with five PEs in the same autonomous system.

Figure 255 **Initial Example Topology**



Configure the IP/MPLS network

The system addresses and IP interface addresses are configured according to Figure 255. An interior gateway protocol (IGP) is needed to distribute routing information on all PEs. In our case, the IGP is IS-IS where each PE is acting as a level 2 router. A configuration example is shown for PE-1. Similar configurations can be derived for the other PEs.

```
*A:PE-1# configure
    router
        isis
            level-capability level-2
            level 2
                wide-metrics-only
            exit
            interface "system"
            exit
            interface "int-PE-1-PE-2"
                interface-type point-to-point
            exit
            interface "int-PE-1-PE-3"
                interface-type point-to-point
            exit
            no shutdown
```

IS-IS interfaces are set up as type point-to-point to improve convergence because no Designated Router/Backup Designated Router (DR/BDR) election process is done. To verify that IS-IS adjacencies are up, **show router isis adjacency** is performed. To check if IP interface addresses/subnets are known on all PEs, **show router route-table** or **show router fib** *slot-number* will display the content of the forwarding information base (FIB).

```
*A:PE-1# show router isis adjacency

================================================================================
Rtr Base ISIS Instance 0 Adjacency
================================================================================
System ID                Usage State Hold Interface                     MT-ID
--------------------------------------------------------------------------------
PE-2                     L2    Up    26   int-PE-1-PE-2                 0
PE-3                     L2    Up    20   int-PE-1-PE-3                 0
--------------------------------------------------------------------------------
Adjacencies : 2
================================================================================
*A:PE-1#
```

```
*A:PE-1# show router route-table

================================================================================
Route Table (Router: Base)
================================================================================
Dest Prefix[Flags]                             Type    Proto     Age        Pref
      Next Hop[Interface Name]                                     Metric
--------------------------------------------------------------------------------
192.0.2.1/32                                   Local   Local     07d23h34m  0
       system                                                        0
192.0.2.2/32                                   Remote  ISIS      07d23h34m  18
       192.168.12.2                                                  10
192.0.2.3/32                                   Remote  ISIS      07d23h17m  18
       192.168.13.2                                                  10
192.0.2.4/32                                   Remote  ISIS      07d22h58m  18
       192.168.12.2                                                  20
192.0.2.5/32                                   Remote  ISIS      07d23h17m  18
       192.168.13.2                                                  20
192.168.12.0/30                                Local   Local     07d23h34m  0
       int-PE-1-PE-2                                                 0
192.168.13.0/30                                Local   Local     07d23h34m  0
       int-PE-1-PE-3                                                 0
192.168.23.0/30                                Remote  ISIS      07d23h16m  18
       192.168.12.2                                                  20
192.168.24.0/30                                Remote  ISIS      07d23h34m  18
       192.168.12.2                                                  20
192.168.34.0/30                                Remote  ISIS      07d23h17m  18
       192.168.13.2                                                  20
192.168.35.0/30                                Remote  ISIS      07d23h17m  18
       192.168.13.2                                                  20
192.168.45.0/30                                Remote  ISIS      03h59m10s  18
       192.168.12.2                                                  30
--------------------------------------------------------------------------------
No. of Routes: 12
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                                LOCAL
    192.0.2.1 (system)
192.0.2.2/32                                                ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.0.2.3/32                                                ISIS
    192.168.13.2 (int-PE-1-PE-3)
192.0.2.4/32                                                ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.0.2.5/32                                                ISIS
    192.168.13.2 (int-PE-1-PE-3)
192.168.12.0/30                                             LOCAL
    192.168.12.0 (int-PE-1-PE-2)
192.168.13.0/30                                             LOCAL
    192.168.13.0 (int-PE-1-PE-3)
192.168.23.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.168.24.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
192.168.34.0/30                                             ISIS
    192.168.13.2 (int-PE-1-PE-3)
192.168.35.0/30                                             ISIS
    192.168.13.2 (int-PE-1-PE-3)
192.168.45.0/30                                             ISIS
    192.168.12.2 (int-PE-1-PE-2)
-------------------------------------------------------------------------------
Total Entries : 12
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

Initially, the default IS-IS Level 2 metric is applied on all interfaces (value 10).

```
*A:PE-1# show router isis status | match "L2 Default Metric"
L2 Default Metric    : 10
```

The next step in the process of setting up the IP/MPLS network is setting up interface-LDP sessions on all interfaces. If the keyword **dual-stack** and **ipv4 no shutdown** is not included in the command, as is the case for interface “int-PE-1-PE-2”, it will be added automatically.

```
*A:PE-1# configure
    router
        ldp
```

```
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                    ipv4
                        no shutdown
                    exit
                exit
                interface "int-PE-1-PE-3" dual-stack
                    ipv4
                        no shutdown
                    exit
                exit
            exit
            targeted-session
            exit
            no shutdown
        exit all
```

There is now a full mesh of LDP label switched paths (LSPs) set up between all system interfaces of the PEs, and the tunnel table on PE-1 looks as follows:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        192.168.12.2   10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.13.2   10
192.0.2.4/32      ldp       MPLS  65539     9        192.168.12.2   20
192.0.2.5/32      ldp       MPLS  65540     9        192.168.13.2   20
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The LDP LSP metric follows the IGP cost. Optionally, LSP metrics can be applied but that is beyond the scope for this chapter.

Enable LDP FRR and Verification

Because LDP FRR is using LFA next-hop pre-computation by the IGP (as described in RFC 5286), the IGP CLI configuration is as follows:

```
*A:PE-1# configure router isis loopfree-alternate

*A:PE-1# show router isis status | match Loopfree
Loopfree-Alternate   : Enabled
```

After enabling LFA inside the IGP context, FRR needs to be enabled within the LDP context:

```
*A:PE-1# configure router ldp fast-reroute

*A:PE-1# show router ldp status | match FRR
FRR                : Enabled              Mcast Upstream FRR    : Disabled
Mcast Upst ASBR FRR: Disabled
```

This chapter describes FRR for unicast LDP. For multicast upstream FRR, see the Multicast Label Distribution Protocol chapter. After these two CLI commands, the software pre-computes for each LDP FEC in the network both a primary and a backup NHLFE and uploads it to the IOM/IMM. The primary NHLFE corresponds to the label of the FEC received from the primary next-hop as per standard LDP resolution of the FEC prefix in the Routing Table Manager (RTM). The backup NHLFE corresponds to the label received for the same FEC from an LFA next-hop.

For point-to-point interfaces, when multiple LFA next hops are found for a primary next-hop, the following selection criteria are used:

- It will pick the node-protect type in favor of the link-protect type.
- If there is more than one LFA next-hop within the selected type, then it will pick one based on the lowest cost.
- If more than one LFA next-hop with the same cost, SPF will select the first one. This is not a deterministic selection and will vary following each SPF calculation.

Several show commands are possible to display LFA information:

The **show router isis statistics** command displays the number of LFA runs on a specific node.

```
*A:PE-1# show router isis statistics

===============================================================================
Rtr Base ISIS Instance 0 Statistics
===============================================================================
---snip---
LFA Statistics
LFA Runs          : 20
 Last scheduled   : 09/11/2018 08:58:54
Partial LFA Runs  : 1
 Last scheduled   : 09/11/2018 08:53:11

RLFA Statistics
RLFA Runs         : 0
---snip---
```

Remote LFA (RLFA) is used in segment routing and described in chapter Segment Routing with IS-IS Control Plane.

The **show router isis lfa-coverage** command performs a mathematical calculation between the number of nodes and IPv4/IPv6 routes in the network versus present LFA next-hop protections. In the example topology (see Figure 255), all IS-IS links have a default level 2 metric of 10. This results in all four nodes and all IS-IS routes learned by PE-1 being 100% LFA protected (link or node). See the following output:

```
*A:PE-1# show router isis lfa-coverage

===============================================================================
Rtr Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      0/0(0%)        9/9(100%)           0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      4/4(100%)      9/9(100%)           0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-1#
```

The **show router isis topology lfa detail** command shows the LFA protection type (link or node).

```
*A:PE-1# show router isis topology lfa detail

===============================================================================
Rtr Base ISIS Instance 0 Topology Table
===============================================================================
-------------------------------------------------------------------------------
IS-IS IP paths (MT-ID 0), Level 2
-------------------------------------------------------------------------------
Node      : PE-2.00                           Metric      : 10
Interface : int-PE-1-PE-2                     SNPA        : none
Nexthop   : PE-2

LFA intf  : int-PE-1-PE-3                     LFA Metric  : 20
LFA nh    : PE-3                              LFA type    : linkProtection

Node      : PE-3.00                           Metric      : 10
Interface : int-PE-1-PE-3                     SNPA        : none
Nexthop   : PE-3

LFA intf  : int-PE-1-PE-2                     LFA Metric  : 20
LFA nh    : PE-2                              LFA type    : linkProtection

Node      : PE-4.00                           Metric      : 20
Interface : int-PE-1-PE-2                     SNPA        : none
Nexthop   : PE-2

LFA intf  : int-PE-1-PE-3                     LFA Metric  : 20
LFA nh    : PE-3                              LFA type    : nodeProtection
```

```
Node      : PE-5.00                              Metric      : 20
Interface : int-PE-1-PE-3                        SNPA        : none
Nexthop   : PE-3

LFA intf  : int-PE-1-PE-2                        LFA Metric  : 30
LFA nh    : PE-2                                 LFA type    : linkProtection

===============================================================================
*A:PE-1#
```

The **show router route-table** command adds an ‘L’ flag as reference that the associated prefix is having also an LFA next hop available. For detailed interface address information used by the LFA calculation, use the **show router route-table alternative** or **show router isis alternative** command. The output on PE-1 for PE-4 looks as follows:

```
*A:PE-1# show router route-table 192.0.2.4

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.4/32 [L]                              Remote  ISIS      07d23h01m  18
       192.168.12.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#


*A:PE-1# show router route-table alternative 192.0.2.4

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.4/32                                  Remote  ISIS      07d23h02m  18
       192.168.12.2                                                 20
       192.168.13.2 (LFA)                                           20
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router isis routes 192.0.2.4 alternative

===============================================================================
Rtr Base ISIS Instance 0 Route Table (alternative)
===============================================================================
Prefix[Flags]                     Metric     Lvl/Typ      Ver.  SysID/Hostname
  NextHop                                                 MT     AdminTag/SID[F]
Alt-Nexthop                                               Alt-   Alt-Type
                                                         Metric
-------------------------------------------------------------------------------
192.0.2.4/32                      20         2/Int.       5     PE-2
   192.168.12.2                                             0        0
   192.168.13.2(L)                                         20       NP
-------------------------------------------------------------------------------
No. of Routes: 1 (1 path)
-------------------------------------------------------------------------------
Flags        : L = Loop-Free Alternate nexthop
Alt-Type     : LP = linkProtection, NP = nodeProtection
SID[F]       : R  = Re-advertisement
               N  = Node-SID
               nP = no penultimate hop POP
               E  = Explicit-Null
               V  = Prefix-SID carries a value
               L  = value/index has local significance
===============================================================================
*A:PE-1#
```

On PE-1, PE-4 (192.0.2.4/32) has a primary SPF next-hop pointing toward PE-2 (192.168.12.2) and an LFA next-hop pointing toward PE-3 (192.168.13.2).

The Inequality 3 formula on PE-1 for prefix 192.0.2.4/32 results in the following:

Inequality 3:

- [SP(backup NHR, D) < {SP(backup NHR, PN) + SP(PN, D)}] or
  [SP (PE-3, PE-4) < {SP (PE-3, PE-2) + SP(PE-2, PE-4)}] or
  [10 < {10 + 10}]

This means that Inequality 3 is met. The calculated LFA next-hop for prefix 192.0.2.4/32 on PE-1 is protecting node PE-2, see Figure 255 for a graphical representation.

The **show router ldp bindings** command displays the Label Information Base (LIB). A BU flag is present in case the associated label is used as backup NHLFE for the prefix. As an example, a display on PE-1 for prefix PE-4 is as follows.

This is only possible because the SR OS LDP implementation is using liberal retention mode which means that every label mapping received by a peer is retained regardless of whether the LSR is the next hop for the advertised mapping.

```
*A:PE-1# show router ldp bindings prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
```

```
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix
Peer                                        FEC-Flags
IgrLbl                                      EgrLbl
EgrNextHop                                  EgrIntf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32
192.0.2.2:0
524284N                                     524284
192.168.12.2                                1/1/1

192.0.2.4/32
192.0.2.3:0
524284U                                     524284BU
192.168.13.2                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 2
===============================================================================
*A:PE-1#
```

The **show router ldp bindings active** command displays the label forwarding information base (LFIB). Also, the BU flag is present and, in addition, a reference to the label action itself: **pop** for eLER, **push** for iLER and **swap** for LSR.

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop      (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
```

```
192.0.2.4/32                                Push
  --                                        524284
192.168.12.2                                1/1/1

192.0.2.4/32                                Push
  --                                        524284BU
192.168.13.2                                1/1/2

192.0.2.4/32                                Swap
524284                                      524284
192.168.12.2                                1/1/1

192.0.2.4/32                                Swap
524284                                      524284BU
192.168.13.2                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-1#
```

Enable Synchronization Timer

Within an MPLS network using LDP, it is common practice to enable a synchronization timer between LDP and the IGP. Also when LDP FRR is enabled, a situation can occur in which a synchronization timer between IGP and LDP will help: the revert scenario. When the interface for the previous primary next hop is restored, IGP may re-converge before LDP completed the FEC exchange with its neighbor over that interface. This may cause LDP to remove the LFA next hop from the FEC and blackhole traffic.

In order to avoid traffic being blackholed, it is recommended to first enable IGP-LDP synchronization on the interface. The time is expressed in seconds and can have a value between 1 and 1800 seconds. It is also possible to configure an end-of-LIB option to optimize the synchronization time, see the *LDP-IGP Synchronization* chapter. In this example, the LDP synchronization timer is enabled as follows:

```
*A:PE-1# configure router interface "int-PE-1-PE-2" ldp-sync-timer 10
*A:PE-1# configure router interface "int-PE-1-PE-3" ldp-sync-timer 10
```

The configuration on the other nodes is similar.

When this timer is enabled, it means that when an interface is restored, the IGP will advertise this link in the network with an infinite metric. The **ldp-sync-timer** is started, LDP adjacencies are brought up together with a label exchange. After the **ldp-sync-timer** expires, the normal metric is advertised in the network again.

Data Path Verification

Data path verification is performed using a Layer 2 Epipe service. Traffic generator ports are connected toward PE-1 and PE-5, and an Epipe service is created using an MPLS LDP based Service Distribution Path (SDP) on both PE-1 and PE-5.

```
*A:PE-1# configure
    service
        sdp 15 mpls create
            far-end 192.0.2.5
            ldp
            no shutdown
        exit
        epipe 1 name "Epipe 1" customer 1 create
            service-mtu 1450
            sap 1/1/3:1 create
            exit
            spoke-sdp 15:1 create
            exit
            no shutdown
```

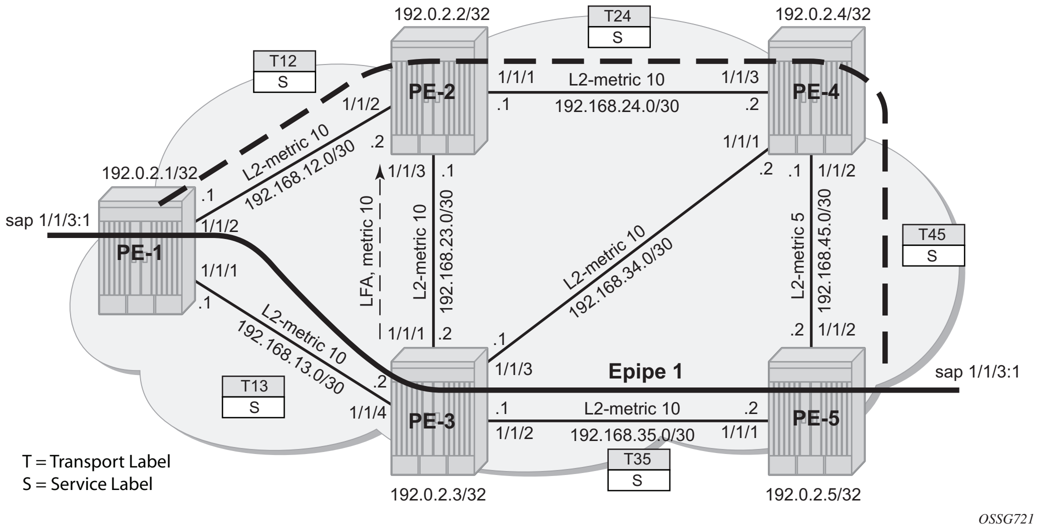
A similar service configuration is configured on PE-5.

The IS-IS Level 2 metric value on the interface between PE-4 and PE-5 is decreased to 5, see Figure 256.

```
*A:PE-4# configure router isis interface int-PE-4-PE-5 level 2 metric 5

*A:PE-5# configure router isis interface int-PE-5-PE-4 level 2 metric 5
```

Figure 256 shows the preferred data path for Epipe 1 via PE-3 and the LFA for PE-5 that is protecting node PE-3.

Figure 256 **Data Verification in the Direction from PE-1 to PE-5 Using Epipe Service**



In this setup, the following LFA for prefix PE-5 from PE-1 is protecting the node PE-3:

```
*A:PE-1# show router isis topology lfa detail

===============================================================================
Rtr Base ISIS Instance 0 Topology Table
===============================================================================
-------------------------------------------------------------------------------
IS-IS IP paths (MT-ID 0),   Level 2
-------------------------------------------------------------------------------
---snip---
Node      : PE-5.00                           Metric      : 20
Interface : int-PE-1-PE-3                     SNPA        : none
Nexthop   : PE-3

LFA intf  : int-PE-1-PE-2                     LFA Metric  : 25
LFA nh    : PE-2                              LFA type    : nodeProtection

===============================================================================
*A:PE-1#
```

```
*A:PE-1# show router isis routes alternative 192.0.2.5

===============================================================================
Rtr Base ISIS Instance 0 Route Table (alternative)
===============================================================================
Prefix[Flags]                     Metric     Lvl/Typ     Ver.  SysID/Hostname
  NextHop                                                MT     AdminTag/SID[F]
Alt-Nexthop                                              Alt-   Alt-Type
                                                        Metric
-------------------------------------------------------------------------------
192.0.2.5/32                      20         2/Int.      6     PE-3
   192.168.13.2                                            0       0
   192.168.12.2(L)                                         25      NP
--------------------------------------------------------------------------------
No. of Routes: 1 (1 path)
--------------------------------------------------------------------------------
Flags        : L = Loop-Free Alternate nexthop
Alt-Type     : LP = linkProtection, NP = nodeProtection
SID[F]       : R  = Re-advertisement
               N  = Node-SID
               nP = no penultimate hop POP
               E  = Explicit-Null
               V  = Prefix-SID carries a value
               L  = value/index has local significance
================================================================================
*A:PE-1#
```

In normal conditions, MPLS traffic from PE-1 toward PE-5 over Epipe 1 will have two MPLS labels: an outer (transport) label given by LDP protocol, swapped on each intermediate LSR and an inner (service) label given by T-LDP, the same end-to-end. See the following show commands.

The T-LDP service label is S (524282):

```
*A:PE-1# show router ldp bindings services service-id 1

================================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
================================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        S - Status Signaled Up,  D - Status Signaled Down, e - Label ELC
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
Service Type:
        E - Epipe Service, V - VPLS Service, M - Mirror Service
        A - Apipe Service, F - Fpipe Service, I - IES Service, R - VPRN service
        P - Ipipe Service, C - Cpipe Service
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
================================================================================
LDP Service FEC 128 Bindings
================================================================================
Type                                           VCId       SDPId            LMTU
Peer                                           SvcId      IngLbl           RMTU
                                                          EgrLbl
--------------------------------------------------------------------------------
E-Eth                                          1          15               1436
192.0.2.5:0                                    1          524282U          1436
                                                          524282S

--------------------------------------------------------------------------------
No. of VC Labels: 1
================================================================================
================================================================================
LDP Service FEC 129 Bindings
================================================================================
SAII                                             AGII        IngLbl    LMTU
TAII                                             Type        EgrLbl    RMTU
```

```
Peer                                            SvcId       SDPId
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
*A:PE-1#
```

The transport LDP label between PE-1 and PE-3 for prefix 192.0.2.5/32 is T13 (524283):

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524283
192.168.13.2                                1/1/2

192.0.2.5/32                                Push
  --                                        524283BU
192.168.12.2                                1/1/1

192.0.2.5/32                                Swap
524283                                      524283
192.168.13.2                                1/1/2

192.0.2.5/32                                Swap
524283                                      524283BU
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-1#
```

The transport LDP label between PE-3 and PE-5 for prefix 192.0.2.5/32 is T35 (524287):

```
*A:PE-3# show router ldp bindings active prefixes prefix 192.0.2.5/32
```

```
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524287
192.168.35.2                                1/1/2

192.0.2.5/32                                Push
  --                                        524283BU
192.168.34.2                                1/1/4

192.0.2.5/32                                Swap
524283                                      524287
192.168.35.2                                1/1/2

192.0.2.5/32                                Swap
524283                                      524283BU
192.168.34.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 4
===============================================================================
*A:PE-3#
```

When PE-3 reboots, PE-1 performs an immediate swap to LFA next-hop for prefix 192.0.2.5/32 bypassing PE-3. The service label remains the same; only the transport labels can change on the network ports PE-1 <=> PE-2, PE-2 <=> PE-4 and PE-4 <=> PE-5. Refer to the following show commands.

Note: The LDP FRR MPLS label stack will never contain more than two labels. This is different from RSVP-TE FRR facility mode which uses a three-label MPLS stack.

The T-LDP service label is S (524282):

```
*A:PE-1# show router ldp bindings services service-id 1

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        S - Status Signaled Up,  D - Status Signaled Down, e - Label ELC
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
Service Type:
        E - Epipe Service, V - VPLS Service, M - Mirror Service
        A - Apipe Service, F - Fpipe Service, I - IES Service, R - VPRN service
        P - Ipipe Service, C - Cpipe Service
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, M - Community Mismatch, BA - ASBR Backup FEC
===============================================================================
LDP Service FEC 128 Bindings
===============================================================================
Type                                          VCId       SDPId             LMTU
Peer                                          SvcId      IngLbl            RMTU
                                                         EgrLbl
-------------------------------------------------------------------------------
E-Eth                                         1          15                1436
192.0.2.5:0                                   1          524282U           1436
                                                         524282S

-------------------------------------------------------------------------------
No. of VC Labels: 1
===============================================================================
---snip---
```

The transport LDP label value between PE-1 and PE-2 for prefix 192.0.2.5/32 is the same label (previously tagged as BU) as before the node failure event: T12 (524283):

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524283
192.168.12.2                                1/1/1

192.0.2.5/32                                Swap
524283                                      524283
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

The transport LDP label between PE-2 and PE-4 for prefix 192.0.2.5/32 is T24 (524283):

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524283
192.168.24.2                                1/1/1

192.0.2.5/32                                Swap
524283                                      524283
192.168.24.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-2#
```

The transport LDP label between PE-4 and PE-5 for prefix 192.0.2.5/32 is T45 (524287):

```
*A:PE-4# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op
IngLbl                                      EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push
  --                                        524287
192.168.45.2                                1/1/1

192.0.2.5/32                                Swap
524283                                      524287
192.168.45.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-4#
```

Additional Topics

Metric Change

Restore all the level 2 metrics back to their default value (10) before applying further changes:

```
*A:PE-4# configure router isis interface "int-PE-4-PE-5" level 2 no metric
```

```
*A:PE-5# configure router isis interface "int-PE-5-PE-4" level 2 no metric
```

Suppose that the IS-IS level 2 metric between PE-2 and PE-3 changes to 30, then 100% LFA coverage is no longer possible. The IS-IS level 2 metric is modified as follows:

```
*A:PE-3# configure router isis interface "int-PE-3-PE-2" level 2 metric 30
```

```
*A:PE-2# configure router isis interface "int-PE-2-PE-3" level 2 metric 30
```

On PE-1, Inequality 3 formula will find LFA next-hop coverages for prefix PE-4 and PE-5. Inequality formula 1 will find LFA next-hop coverages for prefix PE-4, PE-5, and the subnet between PE-4 and PE-5.

Both inequality formulas are visualized in Figure 258 and Figure 257 for prefix 192.0.2.5/32 (= PE-5) on PE-1 which serves as the source node for LFA next-hop computation.

Figure 257 **LFA Computation, Inequality 1 for Prefix PE-5 (D) on PE-1 (S)**

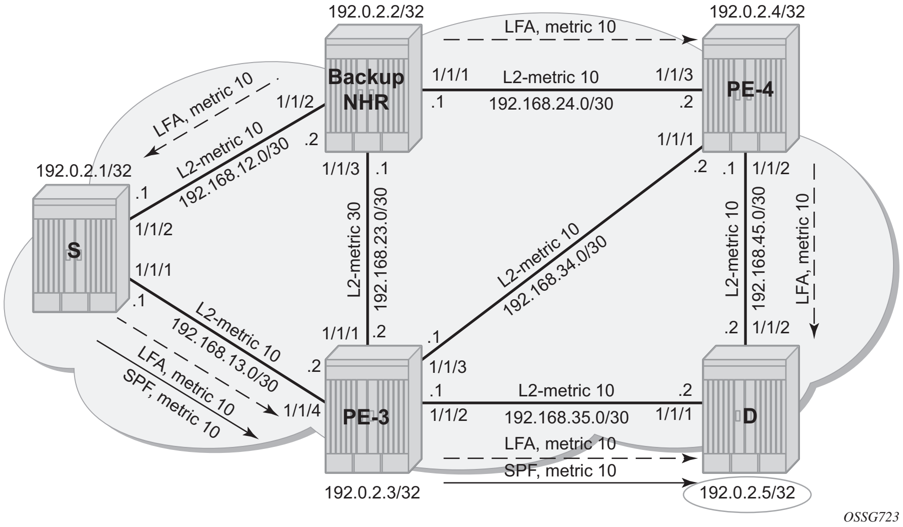
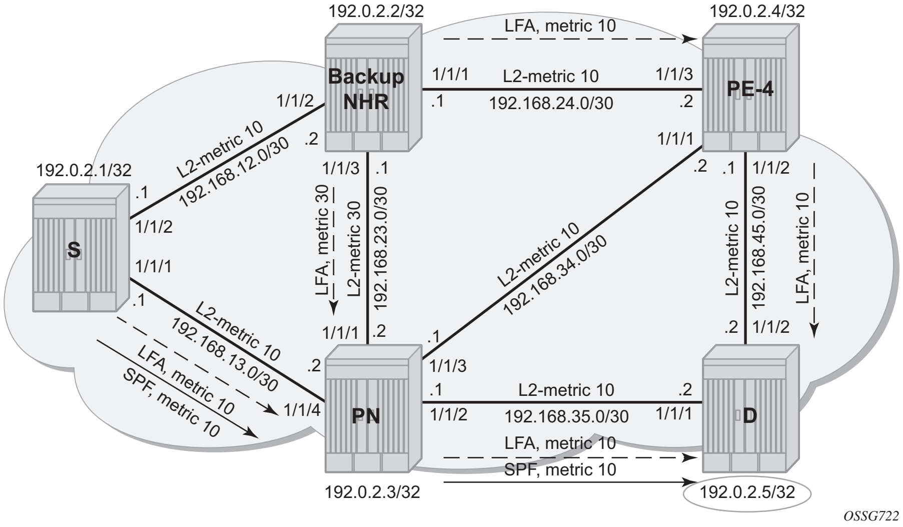


Figure 258 **LFA Computation, Inequality 3 for Prefix PE-5 (D) on PE-1 (S)**



Inequality 3 formula:

- [SP(backup NHR, D) < {SP(backup NHR, PN) + SP(PN, D)}]

For a node LFA next-hop calculation of prefix 192.0.2.5/32 (D) on PE-1, this means that the shortest path from backup next-hop router PE-2 toward destination PE-5 must be smaller than the sum of the shortest path from backup next-hop router PE-2 toward protected node PE-3 with the shortest path from protected node PE-3 to destination PE-5

The shortest path from backup next-hop router PE-2 toward destination PE-5 is going via PE-4, using IS-IS level 2 metric 10 for interface “int-PE-2-PE-4” and IS-IS level 2 metric 10 for interface “int-PE-4-PE-5”. The shortest path from backup next-hop router (PE-2) toward protected node (PE-3) uses IS-IS level 2 metric 30 for interface “int-PE-2-PE-3”. The shortest path from protected node (PE-3) to destination (PE-5) uses IS-IS level 2 metric 10 for interface “int-PE-3-PE-5”. The calculation is as follows:

```
Prefix 192.0.2.5/32: SP (PE-2, PE-5) < SP (PE-2, PE-3) + SP (PE-3, PE-5)
                            10 + 10  < 30 + 10 => OK
```

Inequality 1 formula:

- SP(backup NHR, D) < {SP(backup NHR, S) + SP(S, D)}

For a link LFA next-hop calculation of prefix 192.0.2.5/32 (D) on PE-1, this means that the shortest path from backup next-hop router PE-2 toward destination PE-5 must be smaller than the sum of the shortest path from backup next-hop router PE-2 toward source PE-1 with the shortest path from source PE-1 to destination PE-5.

The shortest path from backup next-hop router PE-2 toward destination PE-5 is going over PE-4, using IS-IS level 2 metric 10 for interface “int-PE-2-PE-4” and IS-IS level 2 metric 10 for interface “int-PE-4-PE-5”. The shortest path from backup next-hop router PE-2 toward source PE-1 uses IS-IS level 2 metric 10 for interface “int-PE-2-PE-1”. The shortest path from source PE-1 to destination PE-5 follows the normal SPF calculation, going over PE-3, using IS-IS level 2 metric 10 for interface “int-PE-1-PE-3”, and IS-IS level 2 metric 10 for interface “int-PE-3-PE-5”.

The calculation is as follows:

```
Prefix 192.0.2.5/32 :  SP(PE-2,PE-5) < SP(PE-2,PE-1) + SP(PE-1,PE-5)
                          10 + 10    <    10 + (10 + 10)            => OK
```

For completeness, all the other Inequality 1 calculations on PE-1 are as follows:

```
Prefix 192.0.2.2/32 :  SP(PE-3,PE-2) < SP(PE-3,PE-1) + SP(PE-1,PE-2)
                              30     <      10 + 10                       => NOK
Prefix 192.0.2.3/32 :  SP(PE-2,PE-3) < SP(PE-2,PE-1) + SP(PE-1,PE-3)
                              30     <      10 + 10                       => NOK
Prefix 192.0.2.4/32 :  SP(PE-3,PE-4) < SP(PE-3,PE-1) + SP(PE-1,PE-4)
                              10     <      10 + (10 + 10)                => OK
Prefix 192.168.23.0/30 :  SP(PE-3,D) < SP(PE-3,PE-1) + SP(PE-1,D)
                              30     <      10 + (10 + 10)                => NOK
```

```
Prefix 192.168.24.0/30 :  SP(PE-3,D) < SP(PE-3,PE-1) + SP(PE-1,D)
                           30 + 10   <      10 + (10 + 10)                => NOK
Prefix 192.168.34.0/30 :  SP(PE-2,D) < SP(PE-2,PE-1) + SP(PE-1,D)
                           30 + 10   <      10 + (10 + 10)                => NOK
Prefix 192.168.35.0/30 :  SP(PE-2,D) < SP(PE-2,PE-1) + SP(PE-1,D)
                           30 + 10   <      10 + (10 + 10)                => NOK
Prefix 192.168.45.0/30 :  SP(PE-3,D) < SP(PE-3,PE-1) + SP(PE-1,D)
                           10 + 10   <      10 + (10 + 10 + 10)           => OK
```

Considering all inequality 1 calculations, only three of these are valid (OK).

In SR OS, the following summary command exists for LFA coverage on the router:

```
*A:PE-1# show router isis lfa-coverage

===============================================================================
LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      0/0(0%)        3/9(33%)            0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      2/4(50%)       3/9(33%)            0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-1#
```

Restore all the IS-IS level 2 metrics back to the default value as follows:

```
*A:PE-2# configure router isis interface "int-PE-2-PE-3" level 2 no metric

*A:PE-3# configure router isis interface "int-PE-3-PE-2" level 2 no metric
```

IS-IS Overload Bit

As stated in RFC 3137, O*SPF Stub Router Advertisement*, sometimes it is useful and desirable for a router not to be a transit node. For those cases, it is also desirable not to have that router used as transit node during the LFA next-hop computation. Within the IS-IS protocol, this is achieved by setting the overload bit. When other routers detect that this bit is set, they will only use this router for packets destined to the overloaded router's directly connected networks and IP prefixes.

As an example, setting the IS-IS overload condition for a specific time on PE-2 provides following result on PE-1:

```
*A:PE-2# configure router isis overload
  - no overload
  - overload [timeout <seconds>] [max-metric]

 <seconds>            : [60..1800]

*A:PE-2# configure router isis overload timeout 60

*A:PE-1# show router isis lfa-coverage

===============================================================================
Rtr Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      0/0(0%)        3/9(33%)            0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      1/4(25%)       3/9(33%)            0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-1#

*A:PE-1# show router route-table alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     07d23h49m  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS      07d23h48m  18
       192.168.12.2                                                 10
       192.168.13.2 (LFA)                                           20
192.0.2.3/32                                  Remote  ISIS      00h01m59s  18
       192.168.13.2                                                 10
192.0.2.4/32                                  Remote  ISIS      00h00m40s  18
       192.168.13.2                                                 20
192.0.2.5/32                                  Remote  ISIS      00h01m59s  18
       192.168.13.2                                                 20
192.168.12.0/30                               Local   Local     07d23h48m  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     07d23h48m  0
       int-PE-1-PE-3                                                0
192.168.23.0/30                               Remote  ISIS      00h01m11s  18
       192.168.12.2                                                 20
       192.168.13.2 (LFA)                                           30
192.168.24.0/30                               Remote  ISIS      07d23h48m  18
       192.168.12.2                                                 20
       192.168.13.2 (LFA)                                           30
192.168.34.0/30                               Remote  ISIS      00h01m59s  18
       192.168.13.2                                                 20
192.168.35.0/30                               Remote  ISIS      00h01m59s  18
```

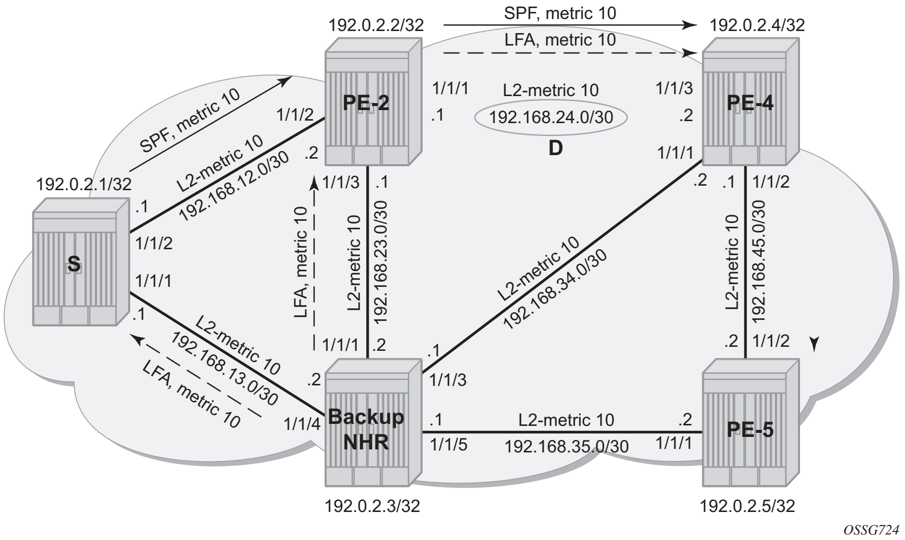
```
       192.168.13.2                                                 20
192.168.45.0/30                               Remote  ISIS      00h00m40s  18
       192.168.13.2                                                 25
-------------------------------------------------------------------------------
No. of Routes: 12
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-1#

*A:PE-1# show router isis routes alternative

===============================================================================
Rtr Base ISIS Instance 0 Route Table (alternative)
===============================================================================
Prefix[Flags]                     Metric     Lvl/Typ     Ver.  SysID/Hostname
  NextHop                                                MT     AdminTag/SID[F]
Alt-Nexthop                                              Alt-   Alt-Type
                                                        Metric
-------------------------------------------------------------------------------
192.0.2.1/32                      0          2/Int.      4     PE-1
   0.0.0.0                                                 0       0
192.0.2.2/32                      10         2/Int.      6     PE-2
   192.168.12.2                                            0       0
   192.168.13.2(L)                                        20      LP
192.0.2.3/32                      10         2/Int.      24    PE-3
   192.168.13.2                                            0       0
192.0.2.4/32                      20         2/Int.      29    PE-3
   192.168.13.2                                            0       0
192.0.2.5/32                      20         2/Int.      25    PE-3
   192.168.13.2                                            0       0
192.168.12.0/30                   10         2/Int.      4     PE-1
   0.0.0.0                                                 0       0
192.168.13.0/30                   10         2/Int.      24    PE-1
   0.0.0.0                                                 0       0
192.168.23.0/30                   20         2/Int.      27    PE-2
   192.168.12.2                                            0       0
   192.168.13.2(L)                                        30      LP
192.168.24.0/30                   20         2/Int.      6     PE-2
   192.168.12.2                                            0       0
   192.168.13.2(L)                                        30      LP
192.168.34.0/30                   20         2/Int.      24    PE-3
   192.168.13.2                                            0       0
192.168.35.0/30                   20         2/Int.      24    PE-3
   192.168.13.2                                            0       0
192.168.45.0/30                   30         2/Int.      29    PE-3
   192.168.13.2                                            0       0
-------------------------------------------------------------------------------
No. of Routes: 12 (12 paths)
-------------------------------------------------------------------------------
Flags        : L = Loop-Free Alternate nexthop
Alt-Type     : LP = linkProtection, NP = nodeProtection
SID[F]       : R  = Re-advertisement
               N  = Node-SID
               nP = no penultimate hop POP
               E  = Explicit-Null
               V  = Prefix-SID carries a value
```

```
             L  = value/index has local significance
===============================================================================
*A:PE-1#
```

On PE-1, only three inequality 1 calculations are possible, as seen in the previous show commands. The inequality 1 calculation on PE-1 for destination 192.168.24.0/30 is as follows:

```
         [SP(backup NHR,D) < {SP(backup NHR,S) + SP(S,D)}]
               SP(PE-3,D) <  SP(PE-3,PE-1) + SP(PE-1,D)
                 10 + 10  <      10 + (10 + 10)                => OK
```

Figure 259 **IS-IS Overload on PE-2, Inequality 1 for 192.168.24.0/30 (D) on PE-1 (S)**



Conclusion

In production MPLS networks where FRR needs to be deployed, a trade off must be made between RSVP-TE FRR versus LDP FRR. The two main advantages of using LDP FRR compared to RSVP FRR are the simple configuration and the fact that LFA next-hop calculation is a local decision, which means there are no interoperability issues when working in a multi-vendor environment. The main disadvantage of using LDP FRR is that LFA next-hop calculation has to deal with the source-route paradigm (inequality formulas exclude a path going over the original source router).

MPLS Transport Profile

This chapter provides information about Multi-Protocol Label Switching Transport Profile (MPLS-TP).

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS and was initially written for SR OS release 12.0.R2. The CLI in the current edition corresponds to 16.0.R3.

The reader should be familiar with the configuration of IP/MPLS and Virtual Leased Line (VLL) services in SR OS.

MPLS-TP was first introduced in SR OS release 11.0.R4 and further enhancements were added in subsequent releases.

Overview

MPLS-TP is intended to allow MPLS to be operated in a manner similar to existing transport technologies, with static configuration of transport paths (particularly with no requirement for a dynamic control plane), proactive in-band and on-demand Operations, Administration, and Maintenance (OAM), and protection mechanisms that do not rely on a control plane (for example, Resource Reservation Protocol with Traffic Engineering (RSVP-TE)) to operate. SR OS routers can operate both as a Label Edge Router (LER) and Label Switching Router (LSR) for MPLS-TP LSPs, and as a Terminating Provider Edge (T-PE) and Switching Provider Edge (S-PE) for pseudowires (PWs) with MPLS-TP OAM. The SR OS router can therefore act as a node within an MPLS-TP network, or as a gateway between MPLS-TP and IP/MPLS domains.

MPLS can provide a network layer with packet transport services. In some operational environments, it is desirable that the operation and maintenance of such an MPLS based packet transport network follows the operational models typically used in traditional optical transport networks (for example with SONET, SDH) while providing additional OAM, survivability and other maintenance functions targeted at that environment.

MPLS-TP defines a profile of MPLS targeted at transport applications. This profile defines specific MPLS characteristics and extensions required to meet the transport requirements, while retaining compliance with the standard IETF MPLS architecture and label switching paradigm. The basic architecture and requirements for MPLS-TP are described by the IETF in RFC 5654, RFC 5921, and RFC 5960, in order to meet two objectives:

- To enable MPLS to be deployed in a transport network and operated in a manner similar to existing transport technologies.
- To enable MPLS to support packet transport services with a similar degree of predictability to that found in existing transport networks.

In order to meet these objectives, MPLS-TP has a number of high-level characteristics:

- MPLS-TP, including resilience and protection, operates in the absence of an IP control plane and IP. MPLS-TP does not modify the MPLS forwarding architecture, which is based on existing pseudowire and LSP constructs. Point-to-point LSPs may be unidirectional or bi-directional. Bi-directional LSPs must be congruent (i.e. co-routed and follow the same path in each direction) and are the only supported type on SR OS. MPLS-TP is only supported on static LSPs and pseudowires (PWs). Also, there is no LSP merging.
- LSP and pseudowire monitoring is achieved using in-band OAM and does not rely on control plane or IP routing functions to determine the health of a path, for example, LDP hello failures do not trigger protection.

The system supports MPLS-TP on LSPs and PWs with static labels. MPLS-TP is not supported on dynamically signaled LSPs and PWs, although switching a static MPLS-TP PW to a Targeted LDP (T-LDP) signaled PW is supported. MPLS-TP is supported for Epipe, Apipe, and Cpipe VLLs, and Epipe spoke SDP termination on IES, VPRN, and VPLS. Static PWs may use SDPs in addition to static MPLS-TP LSPs or RSVP-TE LSPs.

The following MPLS-TP OAM and protection mechanisms defined by the IETF are supported:

- MPLS-TP generic associated channel for LSPs and PWs (RFC 5586)
- MPLS-TP identifiers (RFC 6370)

- Proactive Continuity Check (CC), Connectivity Verification (CV), and Remote Defect Indicator (RDI) using Bi-directional Forwarding Detection (BFD) for LSPs (RFC 6428)
- On-demand CV for LSPs and PWs using LSP ping and LSP trace (RFC 6426)
- 1-for-1 linear protection for LSPs (RFC 6378)
- Static PW status signaling (RFC 6478)

The system can play the role of an LER and an LSR for static MPLS-TP LSPs, and a PE/T-PE and an S-PE for static MPLS-TP PWs. It can also act as an S-PE for MPLS-TP segments between an MPLS network that strictly follows the transport profile and an MPLS network that supports both MPLS-TP and dynamic IP/MPLS.

Configuration

This section details the configuration steps for some simple MPLS-TP examples.

Figure 260 shows the example topology. It consists of four nodes and two Epipe VLL services. One service is used to transport traffic across a network domain consisting of only static MPLS-TP LSPs (Epipe 10) from PE-1 to PE-2. The other Epipe (Epipe 20) is used to transport traffic from PE-1 in the MPLS-TP domain to a VPLS service on PE-4 in an IP/MPLS domain. A static MPLS-TP LSP exists between PE-1 and PE-2, while a dynamic RSVP-TE LSP exists between PE-2 and PE-4.

Figure 260 **MPLS-TP Example Network Showing LSPs**

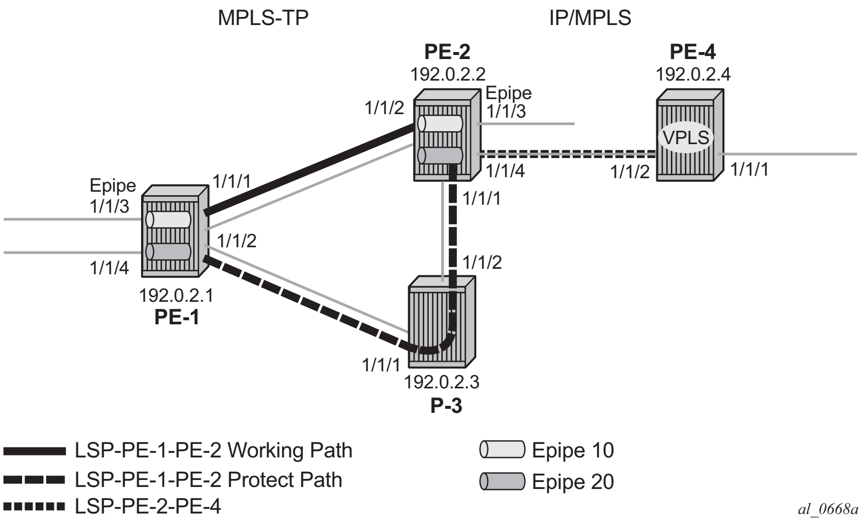


Figure 261 shows further details of the logical architecture of the services in the example network. The Epipe spoke-SDPs use the static MPLS-TP transport LSP between PE-1 and PE-2, and the dynamically signaled RSVP-TE LSP between PE-2 and PE-4. The MPLS-TP LSP is protected using 1:1 linear protection, with a working path from PE-1 to PE-2, and a protect path from PE-1, through LSR P-3, to PE-2. The Ethernet PW for Epipe 10 connects an Ethernet SAP on port 1/1/3 on PE-1 to an Ethernet SAP on port 1/1/3 on PE-2. The PW for Epipe 20 connects an Ethernet SAP on port 1/1/4 on PE-1 to the VPLS on PE-4 and is switched between a static MPLS-TP segment and a dynamic targeted LDP (T-LDP) segment at PE-2. PE-2 thus acts as a gateway between the MPLS-TP domain and the IP/MPLS domain.

***Figure 261* MPLS-TP Example Network Showing Services Detail**

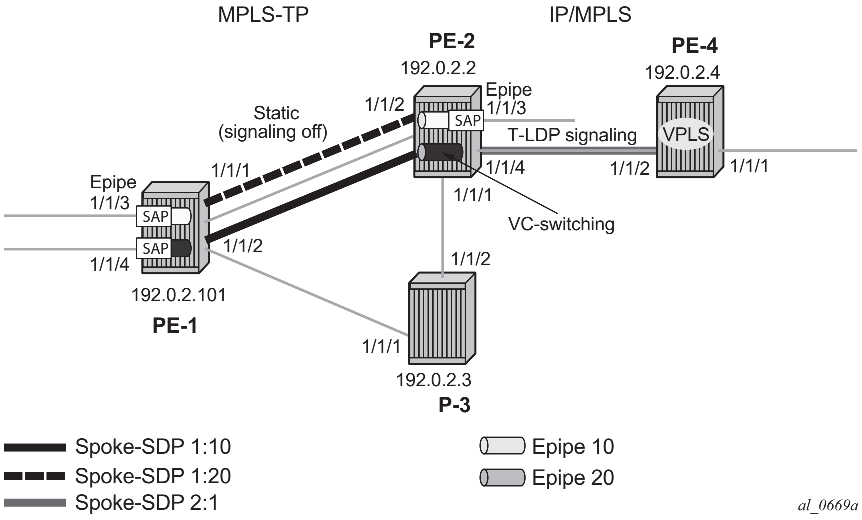
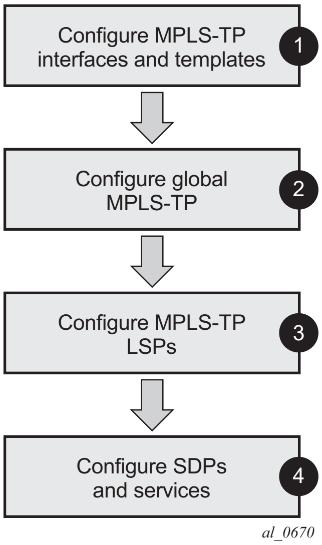


Figure 262 shows the configuration process to be followed when setting up MPLS-TP.

Figure 262 **MPLS-TP Configuration Steps**


Configure MPLS-TP interfaces and templates
1
Configure global MPLS-TP
2
Configure MPLS-TP LSPs
3
Configure SDPs and services
4
al_0670

Configure MPLS-TP Interfaces and Templates

MPLS-TP LSPs can use either numbered or unnumbered network IP interfaces, or unnumbered network interfaces that have been configured to operate without relying on IP routing. This non-IP interface type does not have an IP address associated with it and may be configured to have either a unicast, broadcast or multicast MAC address. The intent of using a broadcast or multicast MAC address is to enable a standard set of MAC addresses to be configured for a network without requiring any changes to the configuration of neighboring router interfaces each time an interface to which a router is connected is changed. If a broadcast or multicast MAC address is used, then the operator should take care that only a point-to-point link is connected to the Ethernet port used by the interface. Otherwise, MPLS-TP packets may be replicated to each remote port to which the link is connected.

The non-IP network interface type is known as an unnumbered MPLS-TP interface. Only MPLS-TP can use this interface type; other IP protocols are blocked from using it. Also, Address Resolution Protocol (ARP) is not used for next hop resolution. This example uses unnumbered MPLS-TP interfaces.

Unnumbered MPLS-TP interfaces are configured on each network-facing interface for the nodes in the MPLS-TP domain, as shown in the following output. This is done using the **unnumbered-mpls-tp** keyword at create time. In addition, the **static-arp unnumbered** command is used to set the next-hop unicast, broadcast, or multicast MAC address of the interface. The system interface should also be configured. Numbered IP Network interfaces, bound to port 1/1/4 of PE-2 and port 1/1/2 of PE-4 are used for the IP/MPLS portion of the network in Figure 260.

```
*A:PE-1# configure
    router
        interface "int-PE-1-P-3" unnumbered-mpls-tp
            port 1/1/2
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "int-PE-1-PE-2" unnumbered-mpls-tp
            port 1/1/1
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        autonomous-system 64511
```

```
*A:PE-2# configure
    router
        interface "int-PE-2-P-3" unnumbered-mpls-tp
            port 1/1/1
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "int-PE-2-PE-1" unnumbered-mpls-tp
            port 1/1/2
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "int-PE-2-PE-4"
            address 192.168.24.1/30
            port 1/1/4
        exit
        interface "system"
            address 192.0.2.2/32
        exit
        autonomous-system 65535
        static-route-entry 192.0.2.4/32
            next-hop 192.168.24.2
                no shutdown
            exit
        exit
```

```
*A:P-3# configure
    router
        interface "int-P-3-PE-1" unnumbered-mpls-tp
            port 1/1/1
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "int-P-3-PE-2" unnumbered-mpls-tp
            port 1/1/2
            static-arp unnumbered 01:00:5e:90:00:00
            no shutdown
        exit
        interface "system"
            address 192.0.2.3/32
        exit
        autonomous-system 65535
```

```
*A:PE-4# configure
    router
        interface "int-PE-4-PE-2"
            address 192.168.24.2/30
            port 1/1/2
        exit
        interface "system"
            address 192.0.2.4/32
        exit
        autonomous-system 65535
        static-route-entry 192.0.2.2/32
            next-hop 192.168.24.1
                no shutdown
            exit
        exit
```

Next, MPLS should be configured on each of the interfaces to be used by MPLS-TP. As an example, only the configuration on PE-1 is shown although a similar configuration is provisioned on PE-2 and P-3.

```
*A:PE-1# configure
    router
        mpls
            mpls-tp
            exit
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-P-3"
            exit
            no shutdown
        exit
```

PE-4 is an IP/MPLS only node so there is no MPLS TP configuration

```
*A:PE-4# configure
    router
        mpls
            interface "int-PE-4-PE-2"
            exit
            no shutdown
        exit
```

The **mpls** context must be in the **no shutdown** state to enable MPLS-TP.

Static labels are used by MPLS-TP LSPs and PWs. By default, SR OS splits the full range in a static and a dynamic range, and these ranges are as follows:

```
*A:PE-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           505856      505856
    Seg-Route   0           0           -           0           0
============================================================================
*A:PE-1#
```

This can be modified by configuration. To reserve 2000 labels starting from label 32, the following command is launched:

```
*A:PE-1# configure router mpls-labels static-label-range 2000
```

As a result, the range from label 32 to 2031 is reserved for static labels.

```
*A:PE-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          2031        -           2000        2000
Dynamic         2032        524287      0           522256      522256
    Seg-Route   0           0           -           0           0
============================================================================
*A:PE-1#
```

In this case, there is no need to modify the range for static labels. The labels that will be chosen are in the default range.

Next, one or more BFD templates are configured on the LERs. These templates are used to define BFD state machine parameters used for BFD Continuity Check (CC) on an LSP, including the transmit and receive timer intervals (in milliseconds). CPM network processor BFD is required if timer intervals as short as 10ms are to be used, but depending on the platform, 100ms BFD may use CPU based BFD (as shown in the example here).

```
*A:PE-1# configure router bfd
  - bfd

      abort           - Discard the changes that have been made to bfd template
                        during a session
      begin           - Switch to edit mode for bfd template - use commit to save
                        or abort to discard the changes made in a session
```

```
 [no] bfd-template    + Configure a bfd template
      commit          - Save the changes made to bfd template during a session

*A:PE-1# configure router bfd bfd-template
  - bfd-template <[32 chars max]>
  - no bfd-template <[32 chars max]>

 [no] echo-receive    - Configure echo receive interval
 [no] multiplier      - Configure multiplier
 [no] receive-interv* - Configure receive interval
 [no] transmit-inter* - Configure transmit interval
 [no] type            - Configure the bfd session endpoint type
```

A subset of these parameters is used by MPLS-TP BFD sessions, as follows:

- **transmit-interval** *transmit-interval* and the **receive-interval** *receive-interval* — These are the transmit and receive timers for BFD packets. For MPLS-TP, these are the timers used by BFD CC packets. Values are in milliseconds: 10ms to 100,000ms, with 1ms granularity. Default 10ms for CPM3 or higher, 1 sec for other hardware. The minimum interval that can be supported is hardware dependent. For MPLS-TP BFD Connectivity Verification (CV) packets, a transmit interval of 1 sec is used.
- **multiplier** *multiplier* — Integer 3 – 20. Default: 3. The configured parameter is used for MPLS-TP CC BFD sessions. It is ignored for MPLS-TP combined CC/CV BFD sessions, and the default of 3 is used.
- **type cpm-np** — This selects the CPM network processor as the local termination point for the BFD session. This is used by default for MPLS-TP. Type CPM-NP is needed to configure a transmit interval down to 10ms.

The following CLI illustrates the BFD template configuration at PE-1. Default parameters are sufficient, so only the BFD template name is configured. BFD templates use a begin/commit model for configuration. Create or modify a template with the **begin** statement. Changes to an existing template or the creation of a new template is not effected until the **commit** statement is entered.

```
*A:PE-1# configure
    router
        bfd
            begin
            bfd-template "tp-bfd"
            exit
            commit
```

The following **info detail** command shows the values that are assigned by default.

```
*A:PE-1>config>router>bfd# info detail
----------------------------------------------
            bfd-template "tp-bfd"
                no type
                transmit-interval 100
                receive-interval 100
                multiplier 3
                echo-receive 100
            exit
----------------------------------------------
```

Configure Global MPLS-TP Parameters

MPLS-TP global parameters are configured in the **config>router>mpls>mpls-tp** context. The MPLS-TP global parameters include the MPLS-TP identifiers for the node and the range of tunnel identifiers that should be reserved for MPLS-TP LSPs.

Node identifiers include the global ID and the node ID. The node ID may be defined as an unsigned integer or use dotted quad notation (a.b.c.d), but the node ID does not have to be a routable IP address.

The CLI tree for configuring the MPLS-TP identifiers for a node is as follows:

```
*A:PE-1# configure router mpls mpls-tp
  - mpls-tp
  - no mpls-tp

 [no] global-id       - Global id for MPLS-TP
 [no] node-id         - Node id for MPLS-TP local router
 [no] oam-template    + Configure a MPLS-TP OAM Template
 [no] protection-tem* + Configure a MPLS-TP Protection Template
 [no] shutdown        - Administratively enable/disable the MPLS-TP
 [no] tp-tunnel-id-r* - Configure MPLS-TP tunnel id range on the ingress router
 [no] transit-path    + Configure a MPLS-TP Transit Path
```

The default value for the global ID is 0. This is used if the global ID is not configured. If an operator expects that inter-domain LSPs will be configured, then it is recommended to set the global ID to the local Autonomous System Number (ASN) of the node, as configured under **config>router**, to ensure that the combination of global ID and node ID is globally unique. If two-byte ASNs are used, then the two most significant bytes of the global ID are padded with zeros.

The default value of the **node-id** is the system interface IPv4 address. The MPLS-TP context cannot be administratively enabled unless at least a system interface IPv4 address is configured, because MPLS requires that this value be configured.

In order to change the values, **config>router>mpls>mpls-tp** must be in the shutdown state. This will bring down all of the MPLS-TP LSPs on the node. New values are propagated to the system when a **no shutdown** is performed.

The following CLI shows the MPLS-TP node identifier configuration for PE-1. A similar configuration is implemented in all routers in this chapter, except that the node IDs must be different (PE-2 is 10.0.0. 2 and P-3 is 10.0.0. 3). In this example, the global ID for PE-2 and P-3 equals 65535.

```
*A:PE-1# configure
    router
        mpls
            mpls-tp
                global-id 64511
                node-id 10.0.0.1
            exit
```

```
*A:PE-2# configure
    router
        mpls
            mpls-tp
                global-id 65535
                node-id 10.0.0.2
            exit
```

Next, protection and OAM templates should be configured at the MPLS-TP LERs. A protection template defines the parameters for the linear protection state coordination mechanism. MPLS-TP Linear Protection is specified in RFC 6378. It provides protection for an LSP using a working and a protect path. A Protection State Coordination (PSC) protocol is used by the LERs at each end of the protected LSP to coordinate whether the working or protect path is used for forwarding. BFD is run on both the working and protect paths.

The linear protection parameters include revertive or non-revertive behavior, the **wait-to-restore** timer, the **rapid-psc-timer** and the **slow-psc-timer**. The **wait-to-restore** timer (in seconds) defines the time to wait before reverting to the working path if, on restoration of connectivity, the revertive behavior is selected.

The following command is used to configure the protection template:

```
*A:PE-1>config>router>mpls>tp$ protection-template
  - no protection-template <[32 chars max]>
  - protection-template <[32 chars max]>


      rapid-psc-timer - Configure the rapid Protection Switch Coordination (PSC) timer
 [no] revertive       - Enable/Disable the template's revertive mode
      slow-psc-timer  - Configure the slow Protection Switch Coordination (PSC) timer
 [no] wait-to-restore - Configure the WTR timer for the template
```

See the CLI command descriptions in the MPLS User Guide for further details of these commands.

The OAM template defines generic proactive OAM parameters, such as BFD hold down and hold up timer values (which can be used to introduce some hysteresis if BFD bounces) and the BFD template to use.

The following command is used to configure the OAM template:

```
*A:PE-1# configure router mpls mpls-tp oam-template
  - no oam-template <template-name>
  - oam-template <template-name>

 <template-name>      : [32 chars max]

 [no] bfd-template - Configure the Bidirectional Forwarding Detection (BFD) template
 [no] hold-time-down  - Configure hold-down dampening timer
 [no] hold-time-up    - Configure the hold-up dampening timer
```

See the CLI command descriptions in the MPLS User Guide for further details of these commands.

MPLS-TP requires the reservation of a tunnel ID range, dedicated for the use of MPLS-TP LSPs. This range is reserved using the following CLI tree:

```
*A:PE-1# configure router mpls mpls-tp tp-tunnel-id-range
  - tp-tunnel-id-range <min> <max>
  - no tp-tunnel-id-range

 <min>                : [1..61440]
 <max>                : [1..61440]
```

PE-1 and PE-2 have the same protection and OAM templates configured, as follows:

```
*A:PE-1# configure
    router
        mpls
            mpls-tp
                tp-tunnel-id-range 100 1000
                protection-template "tp-protect"
                exit
                oam-template "tp-oam"
                    bfd-template "tp-bfd"
                exit
                no shutdown
            exit
```

Configure MPLS-TP LSPs

When the global MPLS-TP parameters have been configured, the system is ready to configure MPLS-TP LSPs. An MPLS-TP LSP is configured under the **config>router>mpls>lsp** context.

Because LSP labels are statically configured, both ends of the LSP must be explicitly configured. The LSP paths must also be explicitly configured in the LSR nodes. MPLS-TP LSPs must use the **mpls-tp** keyword including a source tunnel number at creation time.

The following commands are used to configure an MPLS-TP LSP at an LER:

```
configure
   router
      mpls
         lsp <lsp-name> mpls-tp <src-tunnel-num>]
            to node-id {<a.b.c.d> | <1.. .4,294,967,295>}
            dest-global-id <global-id>
            dest-tunnel-number <tunnel-num>
            [no] working-tp-path
                lsp-num <lsp-num>
                in-label <in-label>
                out-label <out-label> out-link <if-name> [next-hop <ipv4-address>]
                [no] mep
                   [no] oam-template <name>
                   [no] bfd-enable [cc | cc-cv]
                   [no] shutdown
                   exit
                [no] shutdown
                exit
             [no] protect-tp-path
                lsp-num <lsp-num>
                in-label <in-label>
                out-label <out-label> out-link <if-name> [next-hop <ipv4-address> ]
                [no] mep
                   [no] protection-template <name>
                   [no] oam-template <name>
                   [no] bfd-enable [cc | cc-cv]
                   [no] shutdown
                   exit
                [no] shutdown
                exit
```

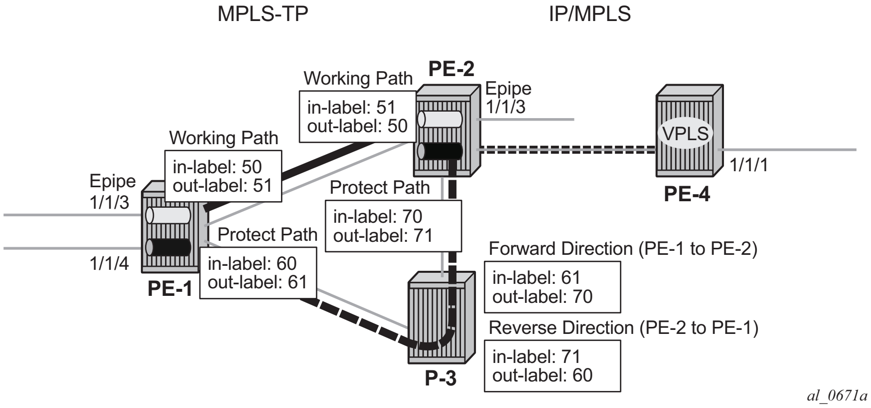
See the CLI command descriptions in the MPLS User Guide for further details of these commands.

A working path and a protect path for LSP LSP-PE-1-P-2 must be configured between PE-1 and PE-2. Each LSP is configured with the full set of MPLS-TP identifiers required to build the LSP ID. Each working path and protect path must have an incoming label, outgoing label and outgoing link configured.

Each working path and protect path also includes a Maintenance entity group End-Point (MEP) configuration, under which the applicable OAM template is configured. BFD is also enabled under the MEP context for the path. In this example, BFD operating in CC mode is enabled on the working and protect paths. The protection template, containing parameters for linear protection, is only applied under the protect path context.

Figure 263 shows the LSP working and protect path label values configured at PE-1, PE-2, and P-3. At each node, the outgoing label must match the incoming label on the next hop for a specific direction. At the LERs (PE-1 and PE-2), the incoming and outgoing label values for each LSP path are configured together. At the LSR (P-3), the label values for the label mapping between ingress and egress for each direction of the path (that is, forward and reverse) are configured together.

Figure 263 **LSP Path Label Value Configurations**



The following shows the LER LSP configuration of PE-1 and PE-2.

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-2" mpls-tp 100
                to node-id 10.0.0.2
                dest-global-id 65535
                dest-tunnel-number 100
                working-tp-path
                    in-label 50
                    out-label 51 out-link "int-PE-1-PE-2"
                    mep
                        oam-template "tp-oam"
                        bfd-enable cc
                        no shutdown
                    exit
                    no shutdown
                exit
                protect-tp-path
                    in-label 60
                    out-label 61 out-link "int-PE-1-P-3"
                    mep
                        protection-template "tp-protect"
                        oam-template "tp-oam"
                        bfd-enable cc
                        no shutdown
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit

*A:PE-2# configure
    router
        mpls
           lsp "LSP-PE-1-PE-2" mpls-tp 100
                to node-id 10.0.0.1
                dest-global-id 64511
                dest-tunnel-number 100
                working-tp-path
                    in-label 51
                    out-label 50 out-link "int-PE-2-PE-1"
                    mep
                        oam-template "tp-oam"
                        bfd-enable cc
                        no shutdown
                    exit
                    no shutdown
                exit
                protect-tp-path
                    in-label 70
                    out-label 71 out-link "int-PE-2-P-3"
                    mep
                        protection-template "tp-protect"
                        oam-template "tp-oam"
                        bfd-enable cc
                        no shutdown
                    exit
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
```

This example requires a protect path to be switched via P-3, therefore, a transit path must be configured in P-3. The CLI tree for configuring MPLS-TP transit paths is as follows:

```
configure
   router
      mpls
         mpls-tp
            transit-path <path-name>
               [no] path-id {lsp-num <lsp-num>|working-path|protect-path

                  [src-global-id <global-id>]
                  src-node-id {<ipv4address> | <1.. .4,294,967,295>}
                  src-tunnel-num <tunnel-num>
                  [dest-global-id <global-id>]
                  dest-node-id {<ipv4address> | <1.. .4,294,967,295>}
                  [dest-tunnel-num <tunnel-num>]}
                  lsp-num <lsp-num>
               forward-path
                  in-label <in-label> out-label <out-label>
                  out-link <if-name> [next-hop <ipv4-next-hop>]
                  exit
               reverse-path
                  in-label <in-label> out-label <out-label>
                  [out-link <if-name> [next-hop <ipv4-next-hop>]]
                   exit
               [no] shutdown
```

See the CLI command descriptions in the MPLS User Guide for further details of these commands.

The CLI configuration for the forward and reverse directions of the transit path (that is, the protect path of the LSP) at P-3 is as follows:

```
*A:P-3# configure
    router
        mpls
            mpls-tp
                transit-path "LSP-PE-1-PE-2"
                    forward-path
                        in-label 61 out-label 70 out-link "int-P-3-PE-2"
                    exit
                    reverse-path
                        in-label 71 out-label 60 out-link "int-P-3-PE-1"
                    exit
                    path-id src-global-id 64511 src-node-id 10.0.0.1 src-tunnel-num
                            100 dest-global-id 65535 dest-node-id 10.0.0.2
                            dest-tunnel-num 100 lsp-num 2
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The example also requires an LSP across the IP/MPLS network to backhaul traffic from PE-2 at the edge of the MPLS-TP network to the VPLS service hosted in PE-4. An RSVP LSP is configured at PE-2 for this purpose, as follows:

```
*A:PE-2# configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-PE-2-PE-4"
                to 192.0.2.4
                primary "loose"
                exit
                no shutdown
            exit


*A:PE-2# configure router rsvp no shutdown
```

Create a T-LDP session toward PE-4. LDP over RSVP is preferred (prefer-tunnel-in-tunnel).

```
*A:PE-2# configure
    router
        ldp
            prefer-tunnel-in-tunnel
            interface-parameters
            exit
            targeted-session
                peer 192.0.2.4
                exit
            exit
        exit
```

A similar configuration is implemented in PE-4.

At this point in the configuration process, it is recommended to verify the MPLS-TP LSP configuration and operation of BFD and linear protection.

First, check that the BFD sessions on both the working and protect paths are up:

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
wp::LSP-PE-1-PE-2                                    Up         4806       4799
  65535::10.0.0.2                                     3          100        100
  mplsTp                                        central          N/A        N/A
pp::LSP-PE-1-PE-2                                    Up         1781       1437
  65535::10.0.0.2                                     3          100        100
  mplsTp                                        central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 2
===============================================================================
*A:PE-1#
```

Next, check the currently active path. This can be done using the **oam lsp-trace** command. The static option must be specified for MPLS-TP LSPs.

```
*A:PE-1# oam lsp-trace static "LSP-PE-1-PE-2"
lsp-trace to LSP-PE-1-PE-2: 0 hops min, 0 hops max, 100 byte packets
1  GlobalId 65535 NodeId 10.0.0.2
   rtt=0.564ms rc=3(EgressRtr)
```

This shows that data packets currently follow the working path of the LSP (no transit node is shown).

In order to test the operation of linear protection, the port used by the working path can be shutdown, and the BFD session state checked again:

```
*A:PE-1# configure port 1/1/1 shutdown
```

```
*A:PE-1# show router bfd session

===============================================================================
Legend:
  Session Id = Interface Name | LSP Name | Prefix | RSVP Sess Name | Service Id
  wp = Working path   pp = Protecting path
===============================================================================
BFD Session
===============================================================================
Session Id                                        State      Tx Pkts    Rx Pkts
  Rem Addr/Info/SdpId:VcId                      Multipl     Tx Intvl   Rx Intvl
  Protocols                                        Type     LAG Port     LAG ID
-------------------------------------------------------------------------------
wp::LSP-PE-1-PE-2                                  Down         6822       6814
  65535::10.0.0.2                                     3         1000        100
  mplsTp                                        central          N/A        N/A
pp::LSP-PE-1-PE-2                                    Up         3796       3452
  65535::10.0.0.2                                     3          100        100
  mplsTp                                        central          N/A        N/A
-------------------------------------------------------------------------------
No. of BFD sessions: 2
```

Execute LSP trace again to check that the LSP has failed over to use the protect path:

```
*A:PE-1# oam lsp-trace static  "LSP-PE-1-PE-2"
lsp-trace to LSP-PE-1-PE-2: 0 hops min, 0 hops max, 100 byte packets
1  GlobalId 65535 NodeId 10.0.0.3
   rtt=0.585ms rc=8(DSRtrMatchLabel)
2  GlobalId 65535 NodeId 10.0.0.2
   rtt=1.07ms rc=3(EgressRtr)
```

This shows that packets are now forwarded via the protect path through P-3, which has node ID 10.0.0.3.

Finally bring the LSP back to the working path by bringing port 1/1/1 up, and either waiting for the LSP to revert to the working path or forcing it onto the working path and clearing the revert timer by executing a **tools** command as follows:

```
*A:PE-1# configure port 1/1/1 no shutdown
```

```
*A:PE-1# tools perform router mpls tp-tunnel force "LSP-PE-1-PE-2"
```

```
*A:PE-1# tools perform router mpls tp-tunnel clear "LSP-PE-1-PE-2"
```

```
*A:PE-1# oam lsp-trace static "LSP-PE-1-PE-2"
lsp-trace to LSP-PE-1-PE-2: 0 hops min, 0 hops max, 100 byte packets
1  GlobalId 65535 NodeId 10.0.0.2
   rtt=0.582ms rc=3(EgressRtr)
```

Configure SDPs and Services

Services can be configured to use MPLS-TP LSPs when the LSP configuration is completed. SDPs and services are configured in a similar manner to those using static-labeled pseudowires without MPLS-TP.

Distributed services are configured to use MPLS-TP with the following steps:

- Configure an SDP with signaling off. With signaling off, the SDP far-end may then be configured as an MPLS-TP node-id or an IPv4 address. SDP keep-alive should be disabled.
- Configure the service, including the spoke-SDP using the SDP. To use MPLS-TP, the spoke-SDP must have statically assigned ingress and egress labels, the control-word must be enabled, and it must have an MPLS-TP identifier for the PW (the PW Path ID) configured. This is comprised of two parts, a Source Attachment Individual Identifier (SAII) and a Target Attachment Individual Identifier (TAII), both of which must be configured. Control channel status signaling may also be configured to support PW status signaling on the static MPLS-TP PW.

In this example, an SDP is configured to use the MPLS-TP LSP from PE-1 to PE-2, which will act as a transport for the static MPLS-TP PWs corresponding to Epipe 10 and Epipe 20. Another SDP is configured for the targeted LDP (T-LDP) PW segment corresponding to Epipe 20 between PE-2 and PE-4.

The following CLI shows the SDP between PE-1 and PE-2 and the SDP between PE-2 and PE-4:

```
*A:PE-1# configure
    service
        sdp 1 mpls create
            signaling off
            far-end node-id 10.0.0.2 global-id 65535
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit


*A:PE-2# configure
    service
        sdp 1 mpls create
            signaling off
            far-end node-id 10.0.0.1 global-id 64511
            lsp "LSP-PE-1-PE-2"
            no shutdown
        exit
        sdp 2 mpls create
            far-end 192.0.2.4
            lsp "LSP-PE-2-PE-4"
            no shutdown
        exit

*A:PE-4# configure
    service
        sdp 2 mpls create
            far-end 192.0.2.2
            lsp "LSP-PE-4-PE-2"
            no shutdown
        exit
```

Next, configure the services that will use the MPLS-TP LSPs.

The service configuration CLI tree for an Epipe service using MPLS-TP is as follows:

```
— configure
      service
        epipe
          [no] spoke-sdp sdp-id[:vc-id]
               [no] hash-label
               [no] standby-signaling-slave
          [no] spoke-sdp sdp-id[:vc-id] [vc-type {ether|vlan}]
              [create] [vc-switching] [no-endpoint | {endpoint [icb]}]
            egress
              vc-label <out-label>
            ingress
              vc-label <in-label>
—         [no] control-word
            [no] pw-path-id
              agi <agi>
              saii-type2 <global-id:node-id:ac-id>
              taii-type2 <global-id:node-id:ac-id>
              exit
            control-channel-status
              [no] acknowledgment
              [no] refresh-timer <value>
              [no] request-timer <value> retry-timer <value> [timeout-multiplier <value>]
              [no] shutdown
```

See the CLI command descriptions in the user guides for further details of these commands.

The following CLI examples show the Epipe service configuration at PE-1, PE-2, and the VPLS spoke-SDP termination point at PE-4.

Epipe 10 belongs to customer 1, and Epipe 20 belongs to customer 2 in this example.

```
*A:PE-1# configure
```

```
    service
        epipe 10 name "Epipe 10" customer 1 create
            sap 1/1/3 create
            exit
            spoke-sdp 1:10 create
                ingress
                    vc-label 150
                exit
                egress
                    vc-label 151
                exit
                control-word
                pw-path-id
                    saii-type2 64511:10.0.0.1:1
                    taii-type2 65535:10.0.0.2:1
                exit
                control-channel-status
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        epipe 20 name "Epipe 20" customer 2 create
            sap 1/1/4 create
            exit
            spoke-sdp 1:20 create
                ingress
                    vc-label 200
                exit
                egress
                    vc-label 201
                exit
                control-word
                pw-path-id
                    saii-type2 64511:10.0.0.1:2
                    taii-type2 65535:10.0.0.2:2
                exit
                control-channel-status
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
```

At PE-2, Epipe 10 terminates on a SAP on port 1/1/3, while Epipe 20 is switched between a static MPLS-TP PW segment (spoke-SDP 1:20) and a T-LDP signaled PW segment (spoke-SDP 2:1) for backhaul to the remote PE-4 containing the VPLS service.

```
*A:PE-2# configure
    service
        epipe 10 name "Epipe 10" customer 1 create
            sap 1/1/3 create
            exit
            spoke-sdp 1:10 create
```

```
                ingress
                    vc-label 151
                exit
                egress
                    vc-label 150
                exit
                control-word
                pw-path-id
                    saii-type2 65535:10.0.0.2:1
                    taii-type2 64511:10.0.0.1:1
                exit
                control-channel-status
                    no shutdown
                exit
                no shutdown
            exit
            no shutdown
        exit
        epipe 20 name "Epipe 20" customer 2 vc-switching create
            spoke-sdp 1:20 create
                ingress
                    vc-label 201
                exit
                egress
                    vc-label 200
                exit
                control-word
                pw-path-id
                    saii-type2 65535:10.0.0.2:2
                    taii-type2 64511:10.0.0.1:2
                exit
                control-channel-status
                    no shutdown
                exit
                no shutdown
            exit
            spoke-sdp 2:1 create
                control-word
                no shutdown
            exit
            no shutdown
        exit
```

At PE-4, the T-LDP signaled PW segment for Epipe 20 is terminated on a VPLS service:

```
*A:PE-4# configure
    service
        vpls 1 name "VPLS 1" customer 2 create
            sap 1/1/1 create
            exit
            spoke-sdp 2:1 create
                control-word
                no shutdown
            exit
            no shutdown
        exit
```

Epipe 10 uses a static MPLS-TP PW from end to end, which can be tested using the Virtual Circuit Connectivity Verification **vccv-ping** command at PE-1, as follows:

```
*A:PE-1# oam vccv-ping static 1:10
VCCV-PING 1:10 84 bytes MPLS payload
Seq=1, send from intf int-PE-1-PE-2
       send from lsp LSP-PE-1-PE-2
       reply via Control Channel
       src id tlv received: GlobalId 65535 NodeId 10.0.0.2
       cv-data-len=44 rtt=0.597ms rc=3 (EgressRtr)

---- VCCV PING 1:10 Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 0.597ms, avg = 0.597ms, max = 0.597ms, stddev = 0.000ms
*A:PE-1#
```

The operation of control channel status signaling can also be verified for this Epipe, as follows:

Shut down the port the SAP on PE-2 is using:

```
*A:PE-2# configure port 1/1/3 shutdown
```

The PW peer status bits for the spoke-SDP for Epipe 10, signaled using control channel status signaling, can be displayed at node PE-1 using the following command (some of the show command output has been removed for brevity). The peer PW status bits are shown in bold in the following output.

```
*A:PE-1# show service id 10 sdp detail

===============================================================================
Services: Service Destination Points Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 1:10  -(10.0.0.2:65535)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 1:10                   Type             : Spoke
Spoke Descr     : (Not Specified)
VC Type            : Ether                  VC Tag           : n/a
Admin Path MTU     : 0                      Oper Path MTU    : 1556
Far End            : 10.0.0.2:65535         Tunnel Far End   : n/a
Oper Tunnel Far End: n/a
LSP Types          : MPLSTP

---snip---
Ingress Label      : 150                    Egress Label     : 151
---snip---

Local Pw Bits      : None
Peer Pw Bits       : lacIngressFault lacEgressFault
Peer Fault Ip      : None
Peer Vccv CV Bits  : None
Peer Vccv CC Bits  : None
---snip---
```

Epipe 20 uses a static MPLS-TP PW from PE-1 to PE-2, identified by a static PW Forwarding Equivalence Class (FEC), and a T-LDP segment with FEC128 from PE-2 to PE-4. Therefore, the target FEC used for a VCCV-ping command from PE-1 to PE-4 is different from the local FEC for the PW at PE-1. VCCV-trace provides a useful tool to test the resulting multi-segment PW (MS-PW), as follows. The same associated channel type must be used for both segments. This is the IPv4 channel.

```
*A:PE-1# oam vccv-trace static 1:20 assoc-channel ipv4 detail
VCCV-TRACE 1:20  with 116 bytes of MPLS payload
1  192.0.2.2 GlobalId 65535 NodeId 10.0.0.2
   rtt=0.599ms rc=8(DSRtrMatchLabel)
   Next segment: VcId=1 VcType=Ether Source=192.0.2.2 Remote=192.0.2.4
2  192.0.2.4  rtt=1.06ms rc=3(EgressRtr)
```

The system supports the interworking of control channel status on a static MPLS-TP PW segment with T-LDP-signaled PW status on a T-LDP PW segment. This can be tested as follows.

Shut down the port the spoke SDP on PE-4 is using:

```
*A:PE-4# configure port 1/1/2 shutdown
```

The PW peer status bits for the spoke-SDP for Epipe 20 can then be displayed at node PE-1 using the following command (some of the show command output has been removed for brevity). The peer PW status bits are shown in bold in the following output.

```
*A:PE-1# show service id 20 sdp detail

===============================================================================
Services: Service Destination Points Details
===============================================================================
-------------------------------------------------------------------------------
 Sdp Id 1:20  -(10.0.0.2:65535)
-------------------------------------------------------------------------------
Description     : (Not Specified)
SDP Id             : 1:20                   Type              : Spoke
Spoke Descr     : (Not Specified)
VC Type            : Ether                  VC Tag            : n/a
Admin Path MTU     : 0                      Oper Path MTU     : 1556
Delivery           : MPLS
Far End            : 10.0.0.2:65535         Tunnel Far End    : n/a
Oper Tunnel Far End: n/a
LSP Types          : MPLSTP

---snip---
Ingress Label      : 200                    Egress Label      : 201
---snip---

Local Pw Bits      : None
Peer Pw Bits       : psnIngressFault psnEgressFault
Peer Fault Ip      : None
Peer Vccv CV Bits  : None
Peer Vccv CC Bits  : None
```

```
---snip---
```

Conclusion

SR OS supports extensive MPLS Transport Profile (MPLS-TP) capabilities. MPLS-TP is intended to allow MPLS to be operated in a manner similar to existing transport technologies, with proactive in-band and on-demand operations and maintenance (OAM), and protection mechanisms that do not rely on a control plane to operate. The 7x50 can operate both as an LER and LSR for MPLS-TP LSPs, and as a T-PE and S-PE for PWs with MPLS-TP OAM. The 7x50 can therefore act as a node within an MPLS-TP network, or as a gateway between MPLS-TP and IP/MPLS domains.

This example has illustrated a simple configuration, demonstrating the role of the SR OS router as an LER and LSR for MPLS-TP LSPs, and how its already extensive multi-service capabilities can be extended over an MPLS-TP network and between MPLS-TP and IP/MPLS networks.

Multicast Label Distribution Protocol

This chapter provides information about multicast Label Distribution Protocol (mLDP).

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS and was initially written for release 13.0.R6. The CLI in this edition corresponds to release 15.0.R1.

In this chapter, the emphasis is on IPv4. However, multicast Label Distribution Protocol (mLDP) is also supported on IPv6 interfaces.

Overview

Multicast Label Distribution Protocol provides extensions to LDP for the setup of point-to-multipoint (P2MP) Label Switched Paths (LSPs) and multipoint-to-multipoint (MP2MP) LSPs in MPLS networks.

The protocol is described in RFC 6388 - *Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths*.

Multicast LSPs can be applied for IP multicast or support for multicast in BGP/MPLS Layer 3 Virtual Private Networks (L3 VPNs).

Compared to RSVP P2MP LSPs, mLDP P2MP LSPs are easier to configure and the setup direction is different. Whereas the RSVP P2MP LSPs are set up from the root node toward the leaf nodes, mLDP P2MP LSPs are set up from the leaf nodes toward the root node.

P2MP Terminology

The following terminology will be used.

Table 24 **Terminology**

Node	Description
Ingress / Root	P2MP LSPs have just one ingress (root) node. The root node receives IP multicast traffic and maps the traffic to a P2MP LSP (push). The node might perform MPLS multicast replication.
Egress / Leaf	P2MP LSPs have multiple egress (leaf) nodes. A leaf node removes data packets from a P2MP LSP (pop) for further processing. The node might perform IP multicast replication.
Transit	A transit Label Switching Router (LSR) can reach the root node via a directly connected upstream LSR. A transit LSR also has one or more directly connected downstream LSRs. The LSR swaps the MPLS label and might perform MPLS multicast replication.
Branch	A branch LSR is a transit LSR that has several directly connected LSRs. The LSR swaps the MPLS label and performs MPLS multicast replication.
Bud	A bud node is an egress node, but also a transit node. The node has directly connected receivers and also one or more directly connected downstream LSRs.

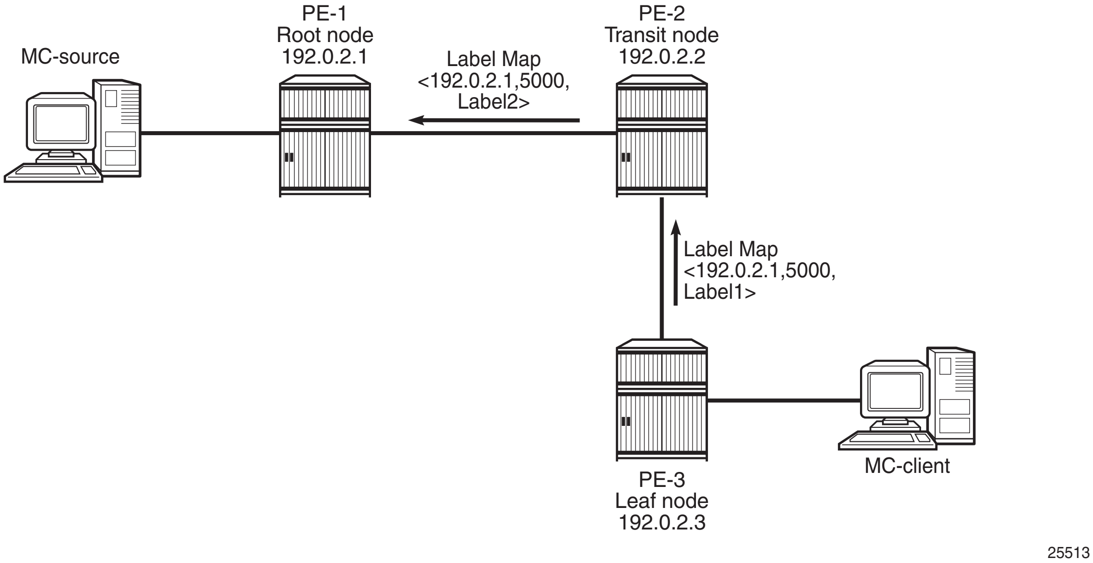
Setup of mLDP P2MP LSP

The setup of the P2MP LSP in the control plane is as follows.

1. The leaf node initiates a tree setup according to what is configured. Mandatory parameters are the IP address of the root and an opaque value. The leaf node sends an LDP label map message to its upstream hop toward the root node of the tree.
2. Each transit node receives the LDP label map message and sends another LDP label map message to its upstream hop toward the root node of the tree. Each label can be different.
3. The root node receives the LDP label map message.

The label map message contains the root node address, an opaque value, and a label. In the example in Figure 1, the root node address is 192.0.2.1 and the opaque value is 5000.

Figure 264 **Setup of mLDP P2MP LSP**



After the LDP label map messages are sent in the control plane, the nodes program pop, swap, or push entries for the corresponding labels in the data plane.

1. The leaf node programs a pop entry for the label sent upstream.
2. The transit node programs a swap entry for the label it sent upstream with the next-hop address and the label it received from the downstream node.
3. The root node programs a push entry and a next-hop address for the label it received from the downstream node.

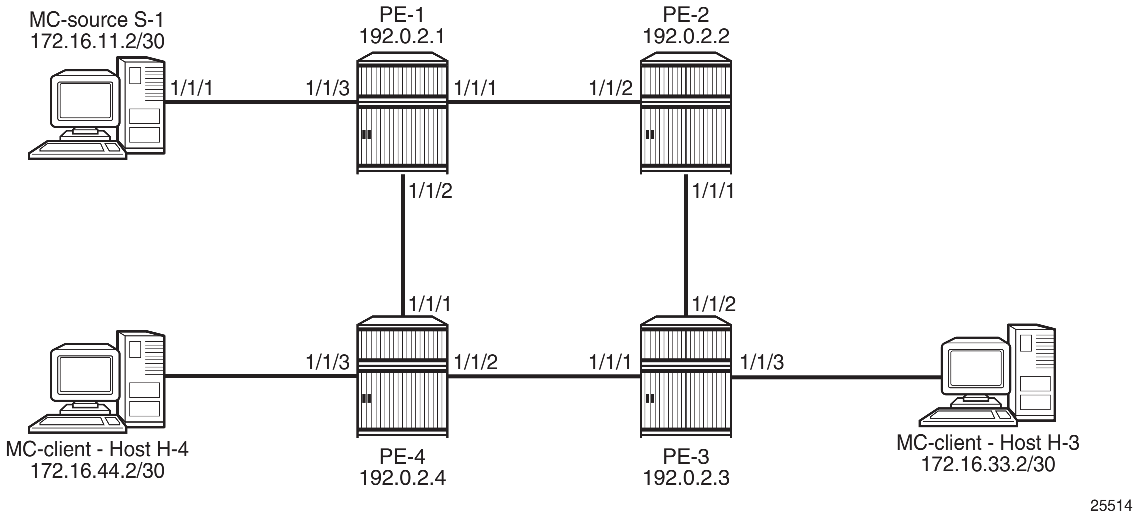
If multiple Equal Cost Multi-Path (ECMP) paths exist between two adjacent nodes, the upstream node of the multicast receiver must program all the entries in the forwarding plane. Only one entry must be active based on the ECMP hashing algorithm.

Configuration

The example setup shown in Figure 265 will be used. The multicast source S-1 is connected to root node PE-1. PE-2 or PE-4 will be the transit node for traffic destined to PE-3. There are two leaf nodes: PE-3 and PE-4. Multicast client H-3 is connected to PE-3, whereas multicast client H-4 is connected to PE-4.

Under normal circumstances, PE-2 is the transit node for traffic toward PE-3 and PE-4 is an egress node. If PE-4 is the transit node for traffic toward PE-3, while it also has a directly connected receiver, PE-4 is a bud node.

Figure 265 **Test Topology**



Configure LDP P2MP LSP

Initial Configuration

The PEs should have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces
- IGP (OSPF or IS-IS)

As an example, the router interfaces and OSPF configuration on PE-1 are as follows:

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
        exit
        interface "int-PE-1-PE-4"
            address 192.168.14.1/30
            port 1/1/2
        exit
        interface "int-PE-1-S-1"
            address 172.16.11.1/30
            port 1/1/3
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        ospf
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-1-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-1-PE-4"
                    interface-type point-to-point
                exit
                interface "int-PE-1-S-1"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
```

Enabling mLDP

When LDP is enabled, mLDP is enabled by default

The following command enables mLDP on a specific interface:

```
configure router ldp interface-parameters interface <ip-int-name> ipv4 fec-type-
capability p2mp-ipv4 enable
```

Enable LDP (including mLDP) on the router interfaces of PE-1, as follows:

```
*A:PE-1# configure
    router
        ldp
            interface-parameters
                interface "int-PE-1-PE-2"
                exit
                interface "int-PE-1-PE-4"
                exit
            exit
```

Verify that mLDP is enabled (P2MP: Enabled), as follows:

```
*A:PE-1# show router ldp status

===============================================================================
LDP Status for IPv4 LSR ID 192.0.2.1
               IPv6 LSR ID ::
===============================================================================
---snip---
Admin State        : Up
IPv4 Oper State    : Up                   IPv6 Oper State      : Down
IPv4 Up Time       : 0d 00:01:47          IPv6 Down Time       : 0d 00:01:47
IPv4 Oper Down Rea*: n/a                  IPv6 Oper Down Reason: systemIpDown
IPv4 Oper Down Eve*: 0                    IPv6 Oper Down Events: 0
---snip---
-------------------------------------------------------------------------------
Capabilities
-------------------------------------------------------------------------------
Dynamic            : Enabled              P2MP                 : Enabled
IPv4 Prefix Fec    : Enabled              IPv6 Prefix Fec      : Enabled
Service Fec128     : Enabled              Service Fec129       : Enabled
MP MBB             : Enabled              Overload             : Enabled
Unrecognized Notif*: Enabled
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

Verify that mLDP is enabled on the interface "int-PE-1-PE-2" (**IPv4 P2MP Fec Cap**), as follows:

```
*A:PE-1# show router ldp interface "int-PE-1-PE-2" detail

===============================================================================
LDP Interfaces
===============================================================================
===============================================================================
Interface "int-PE-1-PE-2"
===============================================================================
-------------------------------------------------------------------------------
BASE
-------------------------------------------------------------------------------
Admin State        : Up                  Oper State       : Up
BFD Status         : Disabled
-------------------------------------------------------------------------------
IPv4
-------------------------------------------------------------------------------
IPv4 Admin State   : Up                  IPv4 Oper State  : Up
Last Oper Chg      : 0d 00:01:19
Hold Time          : 15                  Hello Factor     : 3
Oper Hold Time     : 15
Keepalive Timeout  : 30                  Keepalive Factor : 3
Transport Addr     : System              Last Modified    : 03/25/17 14:02:42
Active Adjacencies : 1
Local LSR Type     : System
Local LSR          : None
IPv4 Pfx Fec Cap   : Enabled             IPv6 Pfx Fec Cap : Enabled
IPv4 P2MP Fec Cap  : Enabled             IPv6 P2MP Fec Cap: Enabled
===============================================================================
No. of Interfaces: 1
===============================================================================
*A:PE-1#
```

Disable mLDP on interface "int-PE-1-PE-2" and verify the LDP status again, as follows:

```
*A:PE-1# configure router
        ldp
            interface-parameters
                interface "int-PE-1-PE-2" dual-stack
                    ipv4
                        fec-type-capability
                            p2mp-ipv4 disable
                        exit all
```

```
*A:PE-1# show router ldp status

===============================================================================
LDP Status for IPv4 LSR ID 192.0.2.1
               IPv6 LSR ID ::
===============================================================================
---snip---
Admin State        : Up
IPv4 Oper State    : Up                    IPv6 Oper State      : Down
---snip---
-------------------------------------------------------------------------------
Capabilities
-------------------------------------------------------------------------------
Dynamic            : Enabled               P2MP                 : Enabled
IPv4 Prefix Fec    : Enabled               IPv6 Prefix Fec      : Enabled
Service Fec128     : Enabled               Service Fec129       : Enabled
MP MBB             : Enabled               Overload             : Enabled
Unrecognized Notif*: Enabled
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

P2MP LDP is still enabled on the router, but it is disabled on interface "int-PE-1-PE-2", which can be verified as follows:

```
*A:PE-1# show router ldp interface "int-PE-1-PE-2" detail

===============================================================================
LDP Interfaces
===============================================================================
===============================================================================
Interface "int-PE-1-PE-2"
===============================================================================
-------------------------------------------------------------------------------
BASE
-------------------------------------------------------------------------------
Admin State        : Up                   Oper State       : Up
BFD Status         : Disabled
-------------------------------------------------------------------------------
IPv4
```

```
-------------------------------------------------------------------------------
IPv4 Admin State   : Up                IPv4 Oper State  : Up
---snip---

IPv4 Pfx Fec Cap   : Enabled           IPv6 Pfx Fec Cap : Enabled
IPv4 P2MP Fec Cap  : Disabled          IPv6 P2MP Fec Cap: Enabled
===============================================================================
No. of Interfaces: 1
===============================================================================
*A:PE-1#
```

P2MP multicast forwarding can be disabled per LDP interface. Disabling P2MP multicast forwarding will not prevent LDP from exchanging P2MP FEC elements on that interface in the control plane. In the data plane, the forwarding plane is not programmed with the next hop on the outgoing interface that is P2MP disabled.

Configure Tunnel Interface on the Root and Leaf Nodes

Multicast LDP can be applied in different scenarios. In the following example, a tunnel interface is created on the root and leaf nodes. Other examples are Multicast Virtual Private Network (MVPN) with mLDP and dynamic PIM-mLDP mapping. In several ACG chapters on MVPN, mLDP is chosen; for example, in *Multicast VPN: Use of Wildcard Selective PMSI*.

A tunnel interface needs to be created on the root node, as follows:

```
*A:PE-1# configure router tunnel-interface ldp-p2mp 5000 sender 192.0.2.1 root-node
```

In this example, the tunnel interface gets interface index 73728, as follows:

```
*A:PE-1# show router tunnel-interface

===============================================================================
 P2MP-RSVP P2MP-LDP Tunnel-Interfaces
===============================================================================
LSP/LDP            Type        SenderAddr          IfIndex             RootNode
-------------------------------------------------------------------------------
5000               ldp         192.0.2.1           73728               Yes
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-1#
```

A similar command will be launched on the leaf nodes, but without the keyword **root-node**, as follows:

```
*A:PE-3# configure router tunnel-interface ldp-p2mp 5000 sender 192.0.2.1
*A:PE-4# configure router tunnel-interface ldp-p2mp 5000 sender 192.0.2.1
*A:PE-3# show router tunnel-interface
```

```
===============================================================================
 P2MP-RSVP P2MP-LDP Tunnel-Interfaces
===============================================================================
LSP/LDP            Type        SenderAddr          IfIndex             RootNode
-------------------------------------------------------------------------------
5000               ldp         192.0.2.1           73728               No
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-3#
```

A P2MP LSP ping can be sent to verify the P2MP LSP. The options for P2MP LSP ping are as follows:

```
*A:PE-1# oam p2mp-lsp-ping
  - p2mp-lsp-ping {<lsp-name> [p2mp-instance <instance-name>
                  [s2l-dest-address <ipv4-address> [... up to 5]]]
                  [ttl <label-ttl>]}
  - p2mp-lsp-ping {ldp <p2mp-identifier> [vpn-recursive-fec]
                  [sender-addr <ipv4-address>]
                  [leaf-addr <ipv4-address> [... up to 5]]}
  - p2mp-lsp-ping {ldp-ssm source <ip-address> group <ip-address>
                  [router <router-instance>|service-name <service-name>]
                  [sender-addr <ipv4-address>] [leaf-addr <ipv4-address>
                  [... up to 5]]}
  - options common to all p2mp-lsp-ping cases:  [fc <fc-name> [profile {in|out}]]
                                   [size <octets>] [timeout <timeout>] [detail]
 <lsp-name>           : [64 chars max]
 <instance-name>      : [32 chars max]
 <ipv4-address>       : a.b.c.d
 <in|out>             : in|out - Default: out
 <fc-name>            : be|l2|af|l1|h2|ef|h1|nc - Default: be
 <octets>             : [1..9198] - Default: 1
 <label-ttl>          : [1..255] - Default: 255
 <timeout>            : [1..120] seconds - Default: 10
 <detail>             : keyword - displays detailed information
 <p2mp-identifier>    : [1..4294967295]
 <ldp-ssm>            : keyword
 <ip-address>         : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
 <router-instance>    : <router-name>|<service-id>
                        router-name    - "Base"|"management"  Default - Base
                        service-id     - [1..2147483647]
 <service-name>       : [64 chars max]
 <vpn-recursive-fec>  : keyword -
 add a VPN Recursive FEC element to the launched                         packet (use
ful for pinging
```

Verify the P2MP LSP with the following ping command:

```
*A:PE-1# oam p2mp-lsp-ping ldp 5000
P2MP identifier 5000: 88 bytes MPLS payload

Total Leafs responded = 2
```

```
          round-trip min/avg/max  = 0.898 / 1.01 / 1.12 ms

Responses based on return code:
        EgressRtr(3)=2
```

Both leaf nodes have sent a reply. The return code 3 indicates that the replying router is an egress for the Forwarding Equivalence Class (FEC).

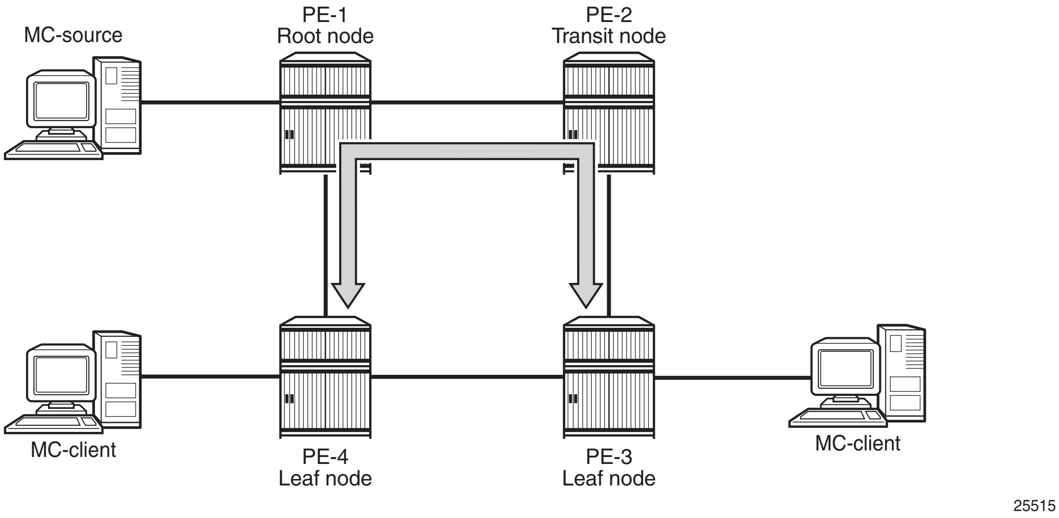
For a detailed output per leaf, use the following command:

```
*A:PE-1# oam p2mp-lsp-ping ldp 5000 detail
P2MP identifier 5000: 88 bytes MPLS payload

===============================================================================
Leaf Information
===============================================================================
From             RTT                    Return Code
-------------------------------------------------------------------------------
192.0.2.4        =0.889ms               EgressRtr(3)
192.0.2.3        =1.36ms                EgressRtr(3)
===============================================================================

Total Leafs responded = 2
          round-trip min/avg/max  = 0.889 / 1.12 / 1.36 ms

Responses based on return code:
        EgressRtr(3)=2
```

Verify LDP P2MP Bindings

The example LDP P2MP LSP is shown in Figure 266. In this case, PE-4 is only an egress node and not a bud node.

Figure 266 **LDP P2MP LSP**



Verify the LDP P2MP bindings on the leaf node PE-4, as follows.

The leaf node programs a pop entry for the label sent upstream.

```
*A:PE-4# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use,  N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Pop             262139      --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
---snip---

*A:PE-4# show router ldp bindings p2mp detail ipv4
```

```
===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
==============================================================================
LDP Generic IPv4 P2MP Bindings
==============================================================================
-------------------------------------------------------------------------------
P2MP Type      : 1                  P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.1:0
Ing Lbl        : 262139U
Egr Lbl        :   --
Egr Int/LspId  :   --
EgrNextHop     :   --
Egr. Flags     : None               Ing. Flags : None
===============================================================================
No. of Generic IPv4 P2MP Bindings: 1
==============================================================================
---snip---
```

PE-4 is only an egress node and not a transit node. There is no next hop.

Verify the LDP P2MP bindings on the leaf node PE-3, as follows:

```
*A:PE-3# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Pop             262139      --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
---snip---
```

Because PE-3 is an egress node, there is no next hop. The traffic toward PE-3 is sent via transit PE-2 and not via PE-4, as can be verified as follows:

```
*A:PE-3# show router ldp bindings p2mp detail ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
==============================================================================
LDP Generic IPv4 P2MP Bindings
==============================================================================
-------------------------------------------------------------------------------
P2MP Type      : 1                 P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.2:0
Ing Lbl        : 262139U
Egr Lbl        :   --
Egr Int/LspId  :   --
EgrNextHop     :   --
Egr. Flags     : None              Ing. Flags : None
===============================================================================
No. of Generic IPv4 P2MP Bindings: 1
==============================================================================
```

PE-2 has programmed a swap entry for the label it sent to its upstream node PE-1 with the next-hop address and the label it received from the downstream node, as follows:

```
*A:PE-2# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
```

```
-------------------------------------------------------------------------------
5000                                        Unknw
192.0.2.1                                   Swap            262139    262139
192.168.23.2                                1/1/1

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
---snip---


*A:PE-2# show router ldp bindings p2mp detail ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
==============================================================================
LDP Generic IPv4 P2MP Bindings
==============================================================================
-------------------------------------------------------------------------------
P2MP Type      : 1                  P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.1:0
Ing Lbl        : 262139U
Egr Lbl        :   --
Egr Int/LspId  :   --
EgrNextHop     :   --
Egr. Flags     : None               Ing. Flags : None
-------------------------------------------------------------------------------
P2MP Type      : 1                  P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.3:0
Ing Lbl        :   --
Egr Lbl        : 262139
Egr Int/LspId  : 1/1/1
EgrNextHop     : 192.168.23.2
Egr. Flags     : None               Ing. Flags : None
Egr If Name    : int-PE-2-PE-3
Metric         : 1                  Mtu        : 1564
===============================================================================
No. of Generic IPv4 P2MP Bindings: 2
==============================================================================
---snip---
```

The egress next hop is PE-3.

On the root node PE-1, there is MPLS multicast replication. One traffic stream goes via transit node PE-2 toward leaf node PE-3 and the other traffic stream goes directly toward leaf node PE-4. There are two push entries with the corresponding next-hop address, as follows:

```
*A:PE-1# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Push              --      262139
192.168.12.2                                1/1/1

5000                                        73728
192.0.2.1                                   Push              --      262139
192.168.14.2                                1/1/2

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
---snip---
```

```
*A:PE-1# show router ldp bindings p2mp detail ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
==============================================================================
LDP Generic IPv4 P2MP Bindings
==============================================================================
-------------------------------------------------------------------------------
P2MP Type      : 1                 P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.2:0
```

```
Ing Lbl        :   --
Egr Lbl        : 262139
Egr Int/LspId  : 1/1/1
EgrNextHop     : 192.168.12.2
Egr. Flags     : None               Ing. Flags : None
Egr If Name    : int-PE-1-PE-2
Metric         : 1                  Mtu        : 1564
-------------------------------------------------------------------------------
P2MP Type      : 1                  P2MP-Id    : 5000
Root-Addr      : 192.0.2.1
-------------------------------------------------------------------------------
Peer           : 192.0.2.4:0
Ing Lbl        :   --
Egr Lbl        : 262139
Egr Int/LspId  : 1/1/2
EgrNextHop     : 192.168.14.2
Egr. Flags     : None               Ing. Flags : None
Egr If Name    : int-PE-1-PE-4
Metric         : 1                  Mtu        : 1564
===============================================================================
No. of Generic IPv4 P2MP Bindings: 2
==============================================================================
---snip---
```

Tools Command

The following tools command can be launched on any of the nodes in the P2MP LSP.

For the ingress node PE-1, where one branch goes to transit node PE-2 (192.0.2.2) and another branch to leaf node PE-4 (192.0.2.4), the output is as follows:

```
*A:PE-1# tools dump router ldp fec p2mp-id 5000 root 192.0.2.1
P2MP: root: 192.0.2.1, T: 1, L: 4, TunnelId: 5000
  Create Time   : 03/25/17 14:05:29.747 (elapsed: 0d 00:01:44)
  Last Mod. Time: 03/25/17 14:05:39.017 (elapsed: 0d 00:01:35)
  FEC Flags     : Push Mttm
  TunlIfId      : 73728   (OperState : up)
  LSP ID        : 65540
  LSP ID Acct.  : 4
  isIngressMttm : Yes        HasLeaf     : No
  isIngrItermdte: No
  CanProgIngress: Yes
  InPhopFrr     : No
  isStitchedUpr : No
  RslvdPhop(p)  : 0.0.0.0:0 (seqNum 0)
  RslvdPhop(b)  : 0.0.0.0:0 (seqNum 0)
  pri Upstream  : None
  mbb Upstream  : None
  bkp Upstream  : None
  AdvInLabel(p) : 0
  AdvInLabel(b) : 0
  Num Resolved   Nhops    : 2
  Num MBB Req.   Nhops    : 0
  Num Programmed Nhops    : 2
    Programmed Nhop[01] : 192.0.2.2:0, OutLabel 262139
    Programmed Nhop[02] : 192.0.2.4:0, OutLabel 262139
  Metric        : 1          Mtu         : 1564

  Num of Peers : 2

  FEC Peer: 192.0.2.2:0
    Peer Flags: MPush  (0x800)
    ModTime   : 03/25/17 14:05:37.598 (elapsed.: 0d 00:01:36)

    ->Num Egress Labels:
      -> (Label: 262139    Status: UsePush)
      Flow Label Tx: no, Rx: no

    <-Num Ingress Labels:
      None

    <Resolved as Next Hop>
     Next Hop Info :
       metric: 1  mtu: 1564
       [01]: Next Hop: 192.168.12.2    Interface: 2
             owner   : 192.0.2.2:0  egress label: 262139

  FEC Peer: 192.0.2.4:0
    Peer Flags: MPush  (0x800)
    ModTime   : 03/25/17 14:05:39.019 (elapsed.: 0d 00:01:35)

    ->Num Egress Labels:
      -> (Label: 262139    Status: UsePush)
      Flow Label Tx: no, Rx: no

    <-Num Ingress Labels:
      None

    <Resolved as Next Hop>
     Next Hop Info :
       metric: 1  mtu: 1564
       [01]: Next Hop: 192.168.14.2    Interface: 3
             owner   : 192.0.2.4:0  egress label: 262139
*A:PE-1#
```

The labels that are pushed at PE-1 are 262139 for traffic to PE-2 and 262139 for traffic to PE-4.

On transit node PE-2, the incoming label 262139 is swapped to outgoing label 262139 toward PE-3, as follows:

```
*A:PE-2# tools dump router ldp fec p2mp-id 5000 root 192.0.2.1
P2MP: root: 192.0.2.1, T: 1, L: 4, TunnelId: 5000
  Create Time   : 03/25/17 14:05:27.209 (elapsed: 0d 00:01:36)
  Last Mod. Time: 03/25/17 14:05:27.209 (elapsed: 0d 00:01:36)
  FEC Flags     : Swap
  TunlIfId      : 0       (OperState : dn)
  LSP ID        : 0
  LSP ID Acct.  : 0
  isIngressMttm : No         HasLeaf     : No
  isIngrItermdte: No
  CanProgIngress: No
  InPhopFrr     : No
  isStitchedUpr : No
  RslvdPhop(p)  : 192.0.2.1:0 (seqNum 2)
  RslvdPhop(b)  : 0.0.0.0:0 (seqNum 0)
  pri Upstream  : 192.0.2.1:0, AdvLabel 262139
  mbb Upstream  : None
  bkp Upstream  : None
  AdvInLabel(p) : 262139
  AdvInLabel(b) : 0
  PrgInLabel(p) : 1
  Num Resolved   Nhops  : 1
  Num MBB Req.   Nhops  : 0
  Num Programmed Nhops  : 1
    Programmed Nhop[01] : 192.0.2.3:0, OutLabel 262139
  Metric        : 1          Mtu         : 1564

  Num of Peers : 2

  FEC Peer: 192.0.2.1:0
    Peer Flags: none  (0x0)
    ModTime   : 03/25/17 14:05:27.210 (elapsed.: 0d 00:01:36)

    ->Num Egress Labels:
      None

    <-Num Ingress Labels:
      <- (Label: 262139    Status: UseSwap)
      Rej Status: OK
      Flow Label Tx: no, Rx: no
      Flow Label Tx Sent: no, Rx Sent: no

    <Resolved as CUR Upstream>

  FEC Peer: 192.0.2.3:0
    Peer Flags: MSwap  (0x1000)
    ModTime   : 03/25/17 14:05:27.210 (elapsed.: 0d 00:01:36)

    ->Num Egress Labels:
      -> (Label: 262139    Status: UseSwap)
      Flow Label Tx: no, Rx: no

    <-Num Ingress Labels:
      None

    <Resolved as Next Hop>
     Next Hop Info :
       metric: 1  mtu: 1564
       [01]: Next Hop: 192.168.23.2    Interface: 3
             owner   : 192.0.2.3:0  egress label: 262139
*A:PE-2#
```

On leaf node PE-3, the incoming label from PE-2 (262139) is popped. There is no next hop.

```
*A:PE-3# tools dump router ldp fec p2mp-id 5000 root 192.0.2.1
P2MP: root: 192.0.2.1, T: 1, L: 4, TunnelId: 5000
  Create Time   : 03/25/17 14:05:30.157 (elapsed: 0d 00:01:36)
  Last Mod. Time: 03/25/17 14:05:30.157 (elapsed: 0d 00:01:36)
  FEC Flags     : Pop Mttm
  TunlIfId      : 73728   (OperState : up)
  LSP ID        : 0
  LSP ID Acct.  : 0
  isIngressMttm : No         HasLeaf     : Yes
  isIngrItermdte: No
  CanProgIngress: No
  InPhopFrr     : No
  isStitchedUpr : No
  RslvdPhop(p)  : 192.0.2.2:0 (seqNum 2)
  RslvdPhop(b)  : 0.0.0.0:0 (seqNum 0)
  pri Upstream  : 192.0.2.2:0, AdvLabel 262139
  mbb Upstream  : None
  bkp Upstream  : None
  AdvInLabel(p) : 262139
  AdvInLabel(b) : 0
  PrgInLabel(p) : 1
  Num Resolved   Nhops  : 1
  Num MBB Req.   Nhops  : 0
  Num Programmed Nhops  : 1
    Programmed Nhop[01] : 0.0.0.0:0, OutLabel 0 (Leaf)
  Metric        : 0          Mtu         : 0

  Num of Peers : 1

  FEC Peer: 192.0.2.2:0
    Peer Flags: none  (0x0)
    ModTime   : 03/25/17 14:05:30.158 (elapsed.: 0d 00:01:36)

    ->Num Egress Labels:
      None

    <-Num Ingress Labels:
      <- (Label: 262139    Status: UsePop)
      Rej Status: OK
      Flow Label Tx: no, Rx: no
      Flow Label Tx Sent: no, Rx Sent: no

    <Resolved as CUR Upstream>
*A:PE-3#
```

The output for leaf node PE-4 is similar.

Debug Commands

Debugging was enabled on the nodes when LDP was configured. To distinguish which messages are being logged for a debug command, the debug configuration is different for the nodes, as follows:

```
*A:PE-2# debug router ldp peer 192.0.2.1 event bindings
*A:PE-3# debug router ldp peer 192.0.2.2 packet label detail
*A:PE-4# debug router ldp peer 192.0.2.1 packet init detail
```

The following LDP messages were logged. The first two messages correspond to the label mapping messages to establish LDP bindings. The following message is sent from transit node PE-2 to root node PE-1.

```
*A:PE-2# debug router ldp peer 192.0.2.1 event bindings
```

```
7 2017/03/25 14:05:27.21 UTC MINOR: DEBUG #2001 Base LDP
"LDP: Binding
Sending Label mapping label 262139 for P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId:
 5000 to peer 192.0.2.1:0."
```

The following LDP message is sent by the leaf node PE-3 to the transit node PE-2.

```
*A:PE-3# debug router ldp peer 192.0.2.2 packet label detail
```

```
7 2017/03/25 14:05:30.15 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 30) to 192.0.2.2:0
Protocol version = 1
Label 262139 advertised for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
"
```

The following message shows the negotiation of capabilities when LDP bindings are initialized.

```
*A:PE-4# debug router ldp peer 192.0.2.1 packet init detail
```

```
1 2017/03/25 14:02:54.54 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Initialization packet (msgId 2) to 192.0.2.1:0
Protocol version = 1
Keepalive Timeout = 30  Label Advertisement = downStreamUnsolicited
Loop Detection = Off  PathVector Limit = 0  Max Pdu = 4096
P2MP Capability = yes
MP MBB Capability = yes
Overload Capability = yes
Dynamic Capability = yes
Unrecognized Notification Capability = yes
"
```

Configure Multicast LDP and Verify Traffic

Configure PIM and IGMP on the Root and Leaf Nodes

PIM needs to be enabled on the root node on the interface toward the multicast source S-1, as follows:

```
*A:PE-1# configure router pim interface "int-PE-1-S-1"
```

On the leaf nodes, PIM needs to be enabled (no shutdown), but no interfaces need to be assigned.

The IGMP configuration for root node PE-1 is needed to forward the incoming traffic for multicast group 232.1.1.1 from source 172.16.11.2 to the tunnel-interface. If IGMP is not configured, the incoming traffic on the interface toward the multicast source will be dropped, because no outgoing interface is defined.

```
*A:PE-1# configure
    router
        igmp
            tunnel-interface ldp-p2mp "5000" sender 192.0.2.1
                static
                    group 232.1.1.1
                        source 172.16.11.2
                    exit
                exit
            exit
        exit
```

The IGMP configuration for leaf node PE-3 is as follows:

```
*A:PE-3# configure
    router
        igmp
            interface "int-PE-3-H-3"
                static
                    group 232.1.1.1
                        source 172.16.11.2
                    exit
                exit
            exit
        exit
```

The incoming traffic from the tunnel interface will be forwarded to the outgoing interface toward the receiving multicast host H-3.

The IGMP configuration for leaf node PE-4 is similar.

At this point, the IGMP/PIM configuration on the root node is complete. This can be verified, as follows:

```
*A:PE-1# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address                Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
```

```
-------------------------------------------------------------------------------
232.1.1.1                  (S,G)                          int-PE-1-S-1   1
   172.16.11.2
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#

*A:PE-1# show router pim group detail
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.1.1.1
Source Address     : 172.16.11.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 172.16.11.2
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:39        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 172.16.11.2
Incoming Intf      : int-PE-1-S-1
Outgoing Intf List : mpls-if-73728

Curr Fwding Rate   : 6319.3 kbps
Forwarded Packets  : 150360             Discarded Packets  : 0
Forwarded Octets   : 6916560            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

The incoming interface is the interface facing the multicast source S-1. The outgoing interface is a reference to the tunnel interface. The name for the outgoing interface (mpls-if-73728) contains the tunnel interface index 73728 as in previous CLI output. The multicast source S-1 is already sending traffic, but the receivers cannot receive it yet.

The configuration on the leaf nodes is still incomplete. A multicast policy needs to be configured and applied first.

Configure and Apply Multicast Policy on Leaf Nodes

The leaf nodes need to get multicast traffic off the LDP P2MP LSP. Therefore, a multicast policy needs to be created and applied. For leaf node PE-3, this is as follows:

```
*A:PE-3# configure
    mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle1" create
                primary-tunnel-interface ldp-p2mp 5000 sender 192.0.2.1
                channel "232.1.1.1" create
                exit
            exit
        exit
```

```
*A:PE-3# configure router multicast-info-policy "p2mp-pol"
```

The configuration for leaf node PE-4 is identical.

Verify Multicast Traffic on Leaf Nodes

Verify the multicast traffic, as follows:

```
*A:PE-3# show router pim group

===============================================================================
Legend:  A = Active   S = Standby
===============================================================================
PIM Groups ipv4
===============================================================================
Group Address               Type                Spt Bit  Inc Intf       No.Oifs
   Source Address              RP                  State    Inc Intf(S)
-------------------------------------------------------------------------------
232.1.1.1                   (S,G)                        mpls-if-73728  1
   172.16.11.2
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

The multicast source S-1 sends a multicast stream with group address 232.1.1.1. The multicast traffic is received by the leaf nodes, which can be verified as follows:

```
*A:PE-3# show router pim group detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 232.1.1.1
Source Address     : 172.16.11.2
```

```
RP Address         : 0
Advt Router        :
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:36        Resolved By        : unresolved

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-3-H-3

Curr Fwding Rate   : 6235.2 kbps
Forwarded Packets  : 204013             Discarded Packets  : 0
Forwarded Octets   : 9384598            RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-3#
```

The incoming interface is from the tunnel interface, whereas the outgoing interface is toward the receiving multicast host H-3.

mLDP Fast Upstream Switchover

mLDP fast upstream switchover allows a downstream node of an mLDP FEC to perform a fast switchover and source the traffic from another upstream node. This switchover is necessary when IGP and LDP are converging due to a failure of the upstream LSR, which is the primary next hop of the root LSR for the P2MP FEC. There will be traffic duplication toward the node that has the upstream alternate backup (in this case to PE-3), but only one stream will be accepted. The multicast stream will be sent to the primary next hop as well as to the loopfree alternate backup. As long as there is no failure, the primary next hop accepts the traffic and forwards it. The backup rejects the traffic. When a failure occurs and the primary LDP session goes down, the backup will start accepting packets.

mLDP fast upstream switchover provides an upstream Fast Reroute (FRR) node-protection capability for the mLDP FEC packets. This multicast upstream FRR node protection is at the expense of traffic duplication from two different upstream nodes into the node that performs the fast upstream switchover. This feature is described in *draft-pdutta-mpls-mldp-up-redundancy*.

Multicast upstream FRR can be configured for mLDP, as follows:

```
*A:PE-1# configure router ospf loopfree-alternate
*A:PE-1# configure router ldp mcast-upstream-frr
```

This configuration can be repeated on some or all of the nodes. In this example, it is configured on all nodes. FRR for unicast can be configured in combination with this, but that is not required. FRR for unicast can be enabled as follows:

```
*A:PE-1# configure router ldp fast-reroute
*A:PE-1# configure router ip-fast-reroute
```

In this example, it is assumed that unicast IP and unicast LDP prefixes do not need to be protected. Therefore, unicast FRR remains disabled.

FRR can be verified as disabled for unicast (FRR) and enabled for multicast (Mcast Upstream FRR), as follows:

```
*A:PE-1# show router ldp status

===============================================================================
LDP Status for IPv4 LSR ID 192.0.2.1
               IPv6 LSR ID ::
===============================================================================
---snip---
Admin State        : Up
IPv4 Oper State    : Up                   IPv6 Oper State      : Down
---snip---

FRR                : Disabled             Mcast Upstream FRR   : Enabled
Mcast Upst ASBR FRR: Disabled
MP MBB Time        : 3
---snip---
-------------------------------------------------------------------------------
Capabilities
-------------------------------------------------------------------------------
Dynamic            : Enabled              P2MP                 : Enabled
IPv4 Prefix Fec    : Enabled              IPv6 Prefix Fec      : Enabled
Service Fec128     : Enabled              Service Fec129       : Enabled
MP MBB             : Enabled              Overload             : Enabled
Unrecognized Notif*: Enabled
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

Of the three nodes in the example topology that have upstream nodes, only PE-3 has an upstream alternate for FRR. PE-4 becomes a transit node for traffic destined to PE-3 (but PE-3 will drop it, until the primary LDP session fails). PE-3 sends a label mapping message to PE-4 for label 262138, as in the following trace message.

```
8 2017/03/25 14:10:34.07 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 63) to 192.0.2.4:0
Protocol version = 1
```

```
Label 262138 advertised for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
"
```

PE-4 will have an additional LDP P2MP binding where the label is swapped, as follows:

```
*A:PE-4# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Pop             262139      --
 --                                           --

5000                                        73728
192.0.2.1                                   Swap            262139    262138
192.168.34.1                                1/1/2

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
---snip---
```

PE-3 has an additional entry for the FRR backup that is available (BU - Alternate for Fast Re-Route). PE-3 will get duplicated traffic, but will reject all traffic from PE-4 and only accept traffic from PE-2 as long as there is no failover.

```
*A:PE-3# show router ldp bindings active p2mp

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Pop             262139      --
  --                                          --

5000                                        73728
192.0.2.1                                   Pop             262138BU    --
  --                                          --

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
---snip---
```

Because LoopFree Alternate (LFA) and ECMP are mutually exclusive, LFA is only useful when ECMP is disabled. When both are enabled, ECMP will have preference.

mLDP fast upstream switchover relies on the fast detection of loss of the LDP session to the upstream peer to which the primary Ingress Label Map (ILM) label had been advertised. As a result, Nokia recommends to perform the following:

1. Enable Bidirectional Forwarding Detection (BFD) on all LDP interfaces to upstream LSR nodes. When BFD detects the loss of the last adjacency to the upstream LSR, BFD brings down the LDP session immediately. The backup ILM is activated.
2. If there is a concurrent T-LDP adjacency to the same LSR node, enable BFD on the T-LDP peer as well as on the interface.
3. Enable the **ldp-sync-timer** option on all interfaces to the upstream LSR nodes.

   If the LDP session for the primary ILM to the upstream LSR goes down for any other reason than a failure of the interface or of the upstream LSR, routing and LDP will go out of sync. The backup ILM will remain activated until the Interior Gateway Protocol (IGP) seeks the next Shortest Path First (SPF). By enabling the **ldp-sync-timer**, this process is accelerated because the advertised link metric will get the maximum value as soon as the LDP session goes down. This triggers the IGP to calculate an SPF route. See chapter LDP-IGP Synchronization.

The FRR configuration can be removed, as follows:

```
*A:PE-1# configure router ldp no mcast-upstream-frr
*A:PE-1# configure router ospf no loopfree-alternate
```

Multipoint Make-Before-Break (MP MBB)

Multipoint MBB is performed when the best path to the root changes, but the existing path can still be used, such as when a link comes up or when the routing metric changes. The goal of MBB is to establish a new P2MP LSP before the old P2MP is removed, in order to avoid traffic loss.

Leaf or transit nodes must allocate a new label and program the ILM with a duplicate set of existing Next-Hop Label Forwarding Entries (NHLFEs) toward the upstream nodes. This may lead to traffic duplication for a short period of time.

Multipoint MBB is enabled by default, as follows:

```
*A:PE-3# show router ldp status

===============================================================================
LDP Status for IPv4 LSR ID 192.0.2.3
               IPv6 LSR ID ::
===============================================================================
---snip---
Admin State        : Up
IPv4 Oper State    : Up                   IPv6 Oper State      : Down
---snip---
MP MBB Time        : 3
---snip---
-------------------------------------------------------------------------------
Capabilities
-------------------------------------------------------------------------------
Dynamic            : Enabled              P2MP                 : Enabled
IPv4 Prefix Fec    : Enabled              IPv6 Prefix Fec      : Enabled
Service Fec128     : Enabled              Service Fec129       : Enabled
MP MBB             : Enabled              Overload             : Enabled
Unrecognized Notif*: Enabled
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-3#
```

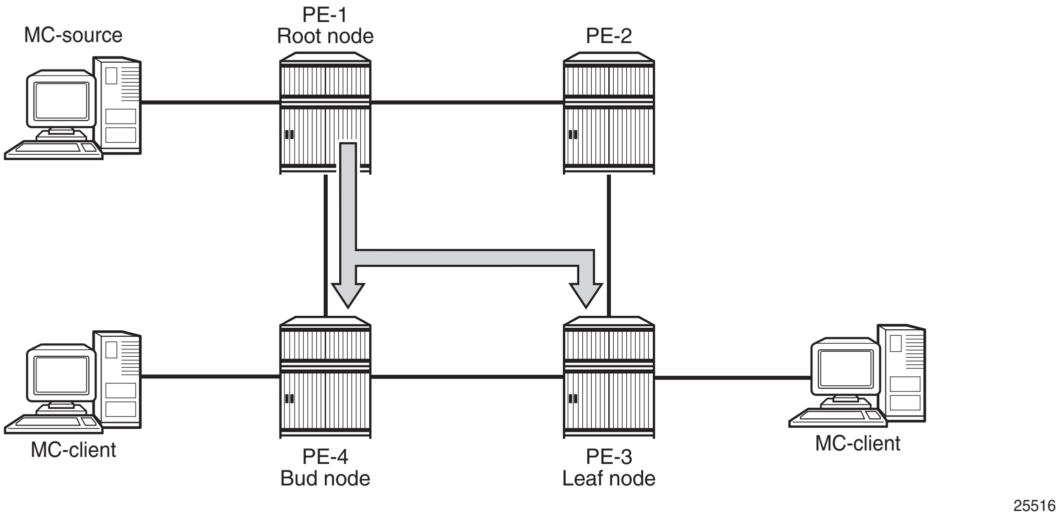
When the metric is increased on the interface (int-PE-3-PE-2) toward the active upstream node, PE-3 sends out an OSPF link status update. Traffic still arrives at PE-3 using the original P2MP LSP. The MBB P2MP LSP is set up.

PE-3 sends a label mapping message toward PE-4, including an MP status TLV carrying an MBB status code indicating that MBB procedures apply to the LSP. PE-4 sends an LDP notification toward PE-3, including an MP status TLV indicating that PE-4 has a state for the existing P2MP LSP.

PE-3 sends an LDP withdrawal message to PE-2. PE-2 replies with an LDP release message.

The multicast traffic arrives at PE-3 using the new LDP P2MP LSP. This way, PE-4 becomes a bud node, and PE-2 is not used for transit anymore; see Figure 267.

Figure 267 **New LDP P2MP LSP after Metric Change**



Originally, leaf node PE-3 preferred the route via PE-2 toward root node PE-1, as follows:

```
*A:PE-3# show router route-table 192.0.2.1

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      00h00m52s  10
       192.168.23.1                                                 20
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

Consequently, the label map messages were originally sent to PE-2, not to PE-4. PE-2 is the transit node for traffic destined to PE-3, as follows:

```
*A:PE-2# show router ldp bindings active p2mp ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        Unknw
192.0.2.1                                   Swap            262139    262139
192.168.23.2                                1/1/1

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
---snip---
```

The metric is changed on the interface between PE-3 and PE-2, as follows:

```
*A:PE-3# configure router ospf area 0 interface "int-PE-3-PE-2" metric 1000
```

The preferred route from leaf node PE-3 to root node PE-1 is now via PE-4, as follows:

```
*A:PE-3# show router route-table 192.0.2.1

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  OSPF      00h13m02s  10
       192.168.34.2                                                 20
-------------------------------------------------------------------------------
No. of Routes: 1
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:PE-3#
```

The leaf node PE-3 will prefer to set up a path from PE-4 rather than PE-2. PE-3 will send label mapping messages to PE-4. The old P2MP LSP will be used until the new P2MP LSP is set up. There will be no traffic interruption.

PE-3 will send a label withdrawal message to PE-2 and PE-2 will no longer be a transit node, as follows:

```
*A:PE-2# show router ldp bindings active p2mp

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
No Matching Entries Found
===============================================================================
---snip---
```

PE-4 is the transit node for traffic to PE-3, and also has a local multicast client H-4, so it is a bud node, as follows:

```
*A:PE-4# show router ldp bindings active p2mp

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op              IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Pop             262139      --
  --                                          --

5000                                        73728
192.0.2.1                                   Swap            262139    262138
192.168.34.1                                1/1/2

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 2
===============================================================================
---snip---
```

There is no traffic multiplication at the root node PE-1. All traffic goes to PE-4, as follows:

```
*A:PE-1# show router ldp bindings active p2mp

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP Generic IPv4 P2MP Bindings (Active)
===============================================================================
P2MP-Id                                     Interface
RootAddr                                    Op            IngLbl    EgrLbl
EgrNH                                       EgrIf/LspId
-------------------------------------------------------------------------------
5000                                        73728
192.0.2.1                                   Push              --      262139
192.168.14.2                                1/1/2

-------------------------------------------------------------------------------
No. of Generic IPv4 P2MP Active Bindings: 1
===============================================================================
---snip---
```

The switchover to this new P2MP LSP occurred without traffic loss.

The following debugging was enabled before the metric change:

```
*A:PE-3# debug router ldp peer 192.0.2.2 packet label detail
*A:PE-3# debug router ldp peer 192.0.2.4 packet label detail
*A:PE-3# debug router ldp peer 192.0.2.2 packet init detail
*A:PE-3# debug router ldp peer 192.0.2.4 packet init detail
```

The first trace message shows that label 262138 is advertised to PE-4. MBB is requested, as follows:

```
11 2017/03/25 14:12:16.89 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Mapping packet (msgId 76) to 192.0.2.4:0
Protocol version = 1
Label 262138 advertised for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
MP Status MBB = REQ
"
```

The next message is a notification from PE-4 confirming that there is no fatal error and MBB can be applied, as follows:

```
12 2017/03/25 14:12:16.89 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Notification packet (msgId 76) from 192.0.2.4:0
```

```
Protocol version = 1
Status Code = MPStatus (0x00000040) Non-fatal
Causing message Id = 0
Causing message type = NULL
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
MP Status MBB = ACK
"
```

The following message is a label withdraw message for label 262139 sent to PE-2:

```
13 2017/03/25 14:12:16.89 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Label Withdraw packet (msgId 76) to 192.0.2.2:0
Protocol version = 1
Label 262139 withdrawn for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
"
```

The last message is a label release message for label 262139 received from PE-2, as follows. This message is only sent after the new P2MP LSP is set up.

```
14 2017/03/25 14:12:17.01 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Label Release packet (msgId 75) from 192.0.2.2:0
Protocol version = 1
Label 262139 released for the following FECs
P2MP: root = 192.0.2.1, T: 1, L: 4, TunnelId: 5000
"
```

Conclusion

Multicast LDP provides extensions to the LDP protocol for the setup of P2MP and MP2MP LSPs in MPLS networks. mLDP is simple to configure compared to RSVP. FRR and MBB are supported for mLDP.

Path MTU Discovery

This chapter provides information about Path MTU Discovery.

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS Release 14.0.R7, but the CLI in the current edition corresponds to SR OS Release 15.0.R2.

Overview

The Maximum Transmission Unit (MTU) is the largest packet size (in bytes) that a network can transmit. IP datagrams larger than the MTU are fragmented into smaller packets before being sent. Table 25 describes the MTU types that are supported in SR OS at both port and service level.

Table 25 **MTU Types**

MTU type	Description
Port MTU	Maximum frame size on a physical wire
Service MTU	Maximum end-to-end frame size sent from the customer across an L2 VPN service
SDP path MTU	Maximum frame size of encapsulated packets sent over the SDP between service endpoints in IP/MPLS VPN
VC MTU	Maximum IP payload size that can be carried inside the tunnel. The VC MTU is derived from the service MTU and negotiated by T-LDP.
LSP path MTU	MTU value negotiated by RSVP path/resv messages

Table 25 **(Continued)MTU Types**

MTU type	Description
OSPF MTU	Maximum size of the OSPF packet
IP MTU	Used in L3 VPN services (IES or VPRN). Maximum IP packet size that L3 VPN customers can send across the provider network.

Table 26 lists the values for the MTU types for Ethernet frames. In SR OS, the MTU value never includes the Frame Check Sequence (FCS).

Table 26 **MTU Values for Ethernet Frames**

MTU type	Value
Access Port MTU	Configurable in the port context. Value should be greater than or equal to the sum of the service MTU and the port encapsulation overhead (0 for null, 4 for dot1q, 8 for QinQ).
Network Port MTU	Configurable in the port context. Value should be greater than or equal to the sum of the SDP path MTU, the MPLS labels (transport, service, hash (entropy), and OAM labels), and an Ethernet header (possibly with a VLAN tag for dot1q).
Service MTU	Configurable in the service context. Maximum payload (IP + Ethernet) that the service offers to the client. Only used in L2 services.
SDP path MTU	By default, not configured. Derived from the network port MTU. Should be, at a minimum, the value of the service MTU. Should be, at a maximum, the result of {network port MTU – 2 labels – Ethernet header }. However, the service will become operationally up when the SDP path MTU is higher. The SDP path MTU need not match on both sides of the SDP.
VC MTU	Not configurable. Derived from the service MTU and negotiated by T-LDP. The VC MTU value must match the other side. VC MTU = service MTU - 14 bytes (Ethernet header).
LSP path MTU	Derived from the port MTU of the network port

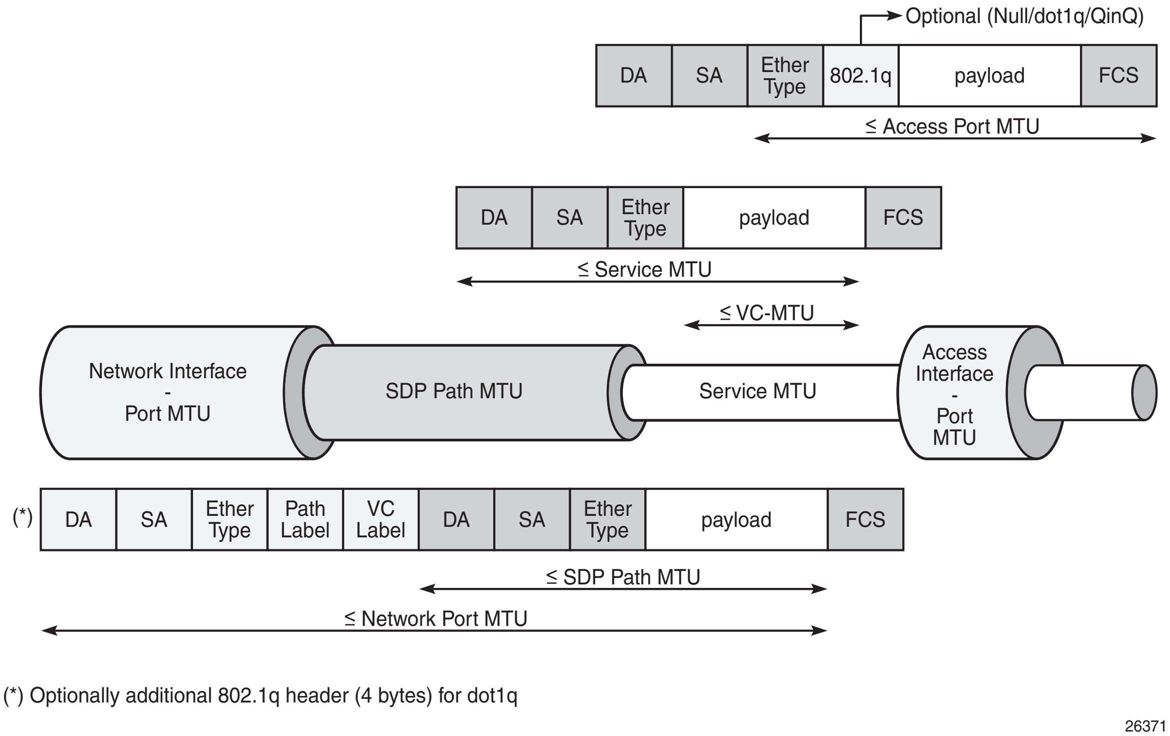
Table 26 **(Continued)MTU Values for Ethernet Frames**

MTU type	Value
OSPF MTU	MTU negotiated by OSPF and derived from the port MTU or administratively set
IP MTU	Configurable in the L3 routing interfaces

The values of the first five MTU types listed in Table 2 are important in getting L2 services to an operational state of up. For L3 services, the IP MTU is used instead of the service MTU.

Figure 268 shows the MTUs used for Ethernet frames in an L2 service, such as an Epipe or a VPLS service.

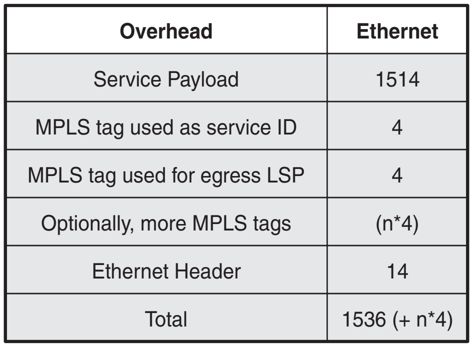
Figure 268 **L2 Services MTUs for Ethernet Frames**



The VC MTU contains the IP payload. The service MTU contains IP payload and Ethernet header. The SDP path MTU must be greater than or equal to the service MTU. Typically, the VLAN tags are stripped at the service ingress, unless VLAN range SAPs are defined and one VLAN tag is preserved. The physical port MTU on an Ethernet access interface needs to be set to at least 1514 for null encapsulation (1500 + 14 (Ethernet header)), at least 1518 for dot1q (1500 + 14 + 4 (dot1q)), and at least 1522 for QinQ (1500 + 14 + 4 + 4).

Figure 269 shows the minimum physical MTU on network interfaces for a router that needs to support services offering a 1514 byte service payload over MPLS for Ethernet.

Figure 269 **Minimum Network Port MTU for Ethernet Frames in MPLS Encapsulation**



Overhead	Ethernet
Service Payload	1514
MPLS tag used as service ID	4
MPLS tag used for egress LSP	4
Optionally, more MPLS tags	(n*4)
Ethernet Header	14
Total	1536 (+ n*4)

Maximum 12 MPLS labels.
Optionally 1 VLAN tag for dot1q.

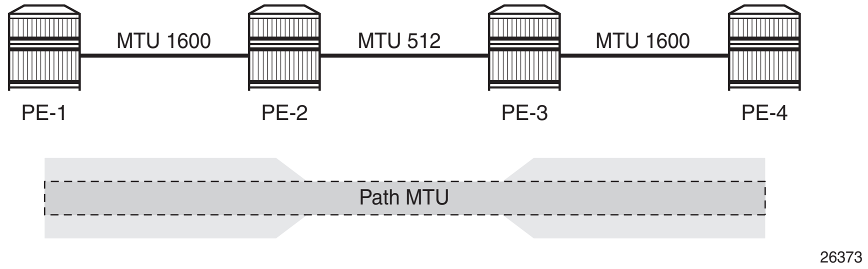
26372

The network port MTU must be at least the maximum service MTU to be supported plus the largest encapsulation type used. The SDP path MTU is at least equal to the service MTU, which is at a minimum 1514 for a service running on a typical Ethernet access interface. This is also valid when the access interface is dot1q or QinQ, because the VLAN tags are stripped at ingress and replaced by the appropriate VLAN tag at egress, unless VLAN range SAPs are defined, in which case one VLAN tag is preserved. The VC tag (service ID) adds a 4 byte service label, the MPLS path adds—at least—one 4 byte transport label, and the Ethernet header adds 14 bytes, for a total of at least 1536 for Ethernet encapsulation. For MPLS, the maximum label stack depth is 12.

The default behavior in SR OS is that the network port MTU is set to its maximum per MDA type, if the network port MTU is not explicitly configured. By default, the SDP path MTU is derived from the network port MTU. For example, when the network port is set to 1600, the SDP path MTU = 1600 (network port MTU) – 4 (MPLS service label) – 4 (MPLS path label) – 14 (Ethernet label) = 1578. However, the SDP path MTU is only accurate when the end-to-end path is considered and the lowest network port MTU in the path is taken.

Figure 270 shows that the path MTU is determined by the lowest MTU along the path that the service needs to transit. When IP hosts transmit IP datagrams to each other, the path MTU is the largest size for which no fragmentation is required along the path.

Figure 270 **Path MTU**



Path MTU Discovery (PMTUD)

PMTUD is a technique for dynamically discovering the MTU size on the network path between two IP hosts, to maximize packet efficiency and avoid packet fragmentation. PMTUD is standardized in RFC 1191 and for IPv6 in RFC 1981.

PMTUD can be enabled in LDP and BGP in the following contexts:

```
*A:PE-1# tree flat detail | match path-mtu-discovery
configure router bgp group neighbor no path-mtu-discovery
configure router bgp group neighbor path-mtu-discovery
configure router bgp group no path-mtu-discovery
configure router bgp group path-mtu-discovery
configure router bgp no path-mtu-discovery
configure router bgp path-mtu-discovery
configure router ldp tcp-session-parameters peer-transport no path-mtu-discovery
configure router ldp tcp-session-parameters peer-transport path-mtu-discovery
configure service vprn bgp group neighbor no path-mtu-discovery
configure service vprn bgp group neighbor path-mtu-discovery
configure service vprn bgp group no path-mtu-discovery
configure service vprn bgp group path-mtu-discovery
configure service vprn bgp no path-mtu-discovery
configure service vprn bgp path-mtu-discovery
```

PMTUD can be enabled in BGP at different levels: global, per group, or per neighbor. PMTUD can be enabled in BGP in the base router or in a VPRN. For LDP, PMTUD is enabled per peer.

PMTUD works by setting the Don't Fragment (DF) option bit in the IP header of outgoing packets. The source assumes initially that the path MTU is the MTU of its egress interface. Any device along the path with an MTU smaller than the IPv4 packet will drop the packet and notify the source by sending back an Internet Control Message Protocol (ICMP) "Fragmentation Needed" (type 3, code 4) error message containing its MTU. IPv6 packets larger than the MTU will also be dropped in which case an ICMPv6 error message "Packet Too Big" (type 2, code 0) containing its MTU will be sent back. The source can then reduce its path MTU to this received MTU. The process repeats until the MTU is small enough to traverse the entire path without fragmentation.

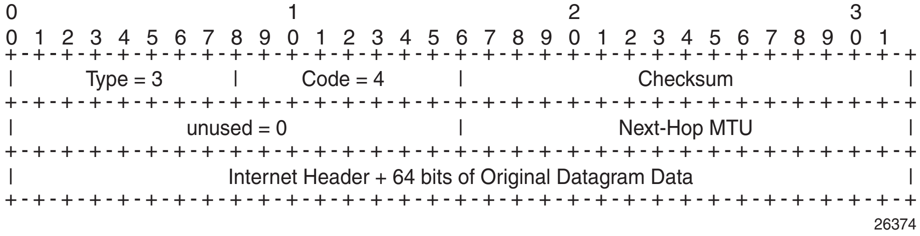
If the path MTU changes to a lower value after the connection is set up, the first larger packet will cause an ICMP error message and the new, lower path MTU will be determined.

PMTUD is used to determine the most efficient packet size for protocols or applications that may send large packets or large data transfers, including BGP updates, LDP, IGPs, FTP/TFTP/SCP transfers. With PMTUD enabled, each connection can start with the maximum MTU—based on egress MTU—then allow remote and/or transit routers to lower the effective MTU for the session if the current MTU is too large for one of their next hops. The path MTU is handled and tracked on a per session/connection basis.

All routers along the path must be able to send ICMP error messages of type 3 ("Destination Unreachable") and code 4 ("Fragmentation Needed").

Figure 271 shows the format of such an ICMP message. The next hop MTU is the MTU of the egress interface to the destination of the packet on the router that dropped the packet. The MTU is a count of the octets of the IP header and IP data, without lower-level headers.

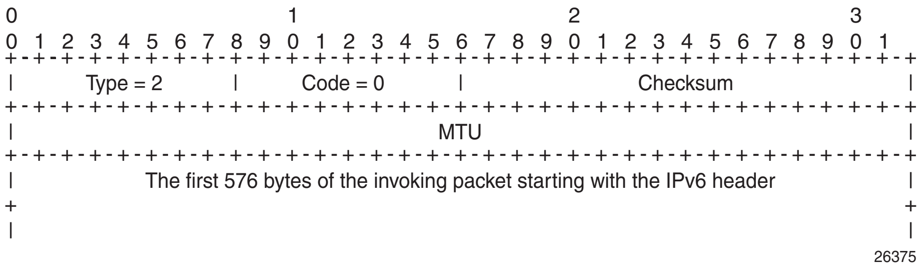
Figure 271 **ICMP "Destination Unreachable" Message - Fragmentation Needed**



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    Type = 3   |    Code = 4   |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          unused = 0           |         Next-Hop MTU          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Internet Header + 64 bits of Original Datagram Data      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The mechanism for IPv6 is similar, but the format of the ICMPv6 message is different. For IPv6, the router will send an ICMPv6 error message of type 2 ("Packet Too Big") and code 0, as shown in Figure 272. The MTU field is populated with the MTU of the egress interface to the destination of the packet on the router that dropped the packet. The MTU is a count of the octets of the IP header and IP data, but no lower-level headers.

Figure 272 **ICMPv6 "Packet Too Big" Message**



```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type = 2      |    Code = 0    |          Checksum         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                               MTU                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   The first 576 bytes of the invoking packet starting with the IPv6 header   |
+                                                                +
|                                                                |
```

When PMTUD is enabled, the IP MTU is initially set to the egress MTU size, based on the source IP interface for that session. When a node along the path is unable to forward a packet due to a smaller MTU, the node drops the packet and sends back an ICMP error message with the MTU of the egress interface. The node that receives the ICMP error message will adjust its MTU accordingly. The IP header and the following bytes of the original IP datagram should be used to determine which connection caused the error.

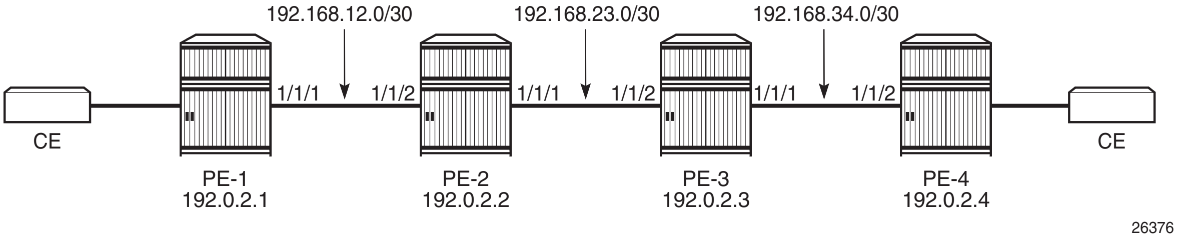
Configuration

The following examples are configured:

- PMTUD in LDP for an IPv4 peer
- PMTUD in LDP for an IPv6 peer
- PMTUD in BGP for an IPv4 peer
- PMTUD in BGP for an IPv6 peer

Figure 273 shows the example topology with four PE nodes in autonomous system 64496. The interfaces have IPv4 and IPv6 addresses, but in this figure, only the IPv4 addresses are shown.

Figure 273 **Example Topology**



The initial configuration on the nodes includes:

- Cards, MDAs, ports
- Router interfaces with IPv4 and IPv6 address
- IS-IS as IGP on all interfaces between the PEs (alternatively, OSPF can be used)
- LDP enabled on all interfaces between the PEs for IPv4 and IPv6

The initial configuration on PE-1 is as follows:

```
configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/1
            ipv6
                address 2001:db8:12::/127
            exit
        exit
        interface "system"
            address 192.0.2.1/32
            ipv6
                address 2001:db8::1/128
            exit
        exit
        isis
            area-id 49.0001
            ipv6-routing native
            interface "system"
            exit
            interface "int-PE-1-PE-2"
                interface-type point-to-point
            exit
            no shutdown
        exit
        ldp
            interface-parameters
                interface int-PE-1-PE-2 dual-stack
                    ipv4
                        no shutdown
                    exit
                    ipv6
```

```
                        no shutdown
                    exit
                exit
            exit
        exit
```

The configuration is similar on the other PEs.

In the example, the default service MTU is used (= 1514 bytes), the access port MTU is 1518 (dot1q encapsulation), and a network port MTU (1600) is set, high enough to support the service MTU (1514):

```
*A:PE-1# show port

===============================================================================
Ports on Slot 1
===============================================================================
Port          Admin Link Port    Cfg  Oper LAG/ Port Port Port   C/QS/S/XFP/
Id            State      State   MTU  MTU  Bndl Mode Encp Type   MDIMDX
-------------------------------------------------------------------------------
1/1/1         Up    Yes  Up      1600 1600    - netw null xgige  10GBASE-LR  *
---snip---
1/2/1         Up    Yes  Up      1518 1518    - accs dotq xgige  10GBASE-LR  *
---snip---
===============================================================================
```

The network port MTU on the link between PE-2 and PE-3 is configured to 512 for IPv4. For IPv6, this network port MTU on the link between PE-2 and PE-3 is reconfigured with a value of 1300.

The service MTU is 1514, the SAP MTU is 1518 (dot1q encapsulation on access port), and the SDP MTU is 1578 (= 1600 (network port MTU) – 14 (Ethernet) – 8 (2 MPLS labels: service label and transport label)), as shown for an Epipe service on PE-1. The configuration for SDP 14 is shown in section SDP Path MTU for IPv4; for Epipe 1, in section PMTUD for LDP IPv4. This SDP MTU does not consider the lowest network port MTU in the path, but only the local network MTU.

```
*A:PE-1# show service id 100 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
---snip---
MTU               : 1514
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/2/1:100                            q-tag        1518    1518    Up   Up
```

```
sdp:14:100 S(192.0.2.4)                  Spok         0       1578    Up   Up
===============================================================================
*A:PE-1#
```

SDP Path MTU for IPv4

The network port MTU is configured to 512 on the interfaces between PE-2 and PE-3, as follows:

```
*A:PE-2# configure port 1/1/1 ethernet mtu 512
*A:PE-3# configure port 1/1/2 ethernet mtu 512
```

On PE-1, SDP 14 is configured toward PE-4, as follows:

```
configure
    service
        sdp 14 mpls create
            far-end 192.0.2.4
            ldp
            no shutdown
        exit
```

The configuration is similar on PE-4, but with a far end of 192.0.2.1 instead.

The SDP path MTU is derived from the lowest network port MTU in the path: 512 – 14 (Ethernet header) – 4 (MPLS service label) – 4 (MPLS path label) = 490. This can be verified on PE-1 for the end-to-end path with the following OAM command that sends packets with an incrementing size: from 400 to 500 bytes in steps of 10 bytes. The packet with size 490 bytes gets a response, whereas the packet with size 500 gets a timeout.

```
*A:PE-1# oam sdp-mtu 14 size-inc 400 500 step 10
Size    Sent    Response
----------------------------
400     .       Success
410     .       Success
420     .       Success
430     .       Success
440     .       Success
450     .       Success
460     .       Success
470     .       Success
480     .       Success
490     .       Success
500     ...     Request Timeout

Maximum Response Size: 490
```

The next step is to repeat the OAM command to send packets with incrementing size from 490 to 500 in steps of 1:

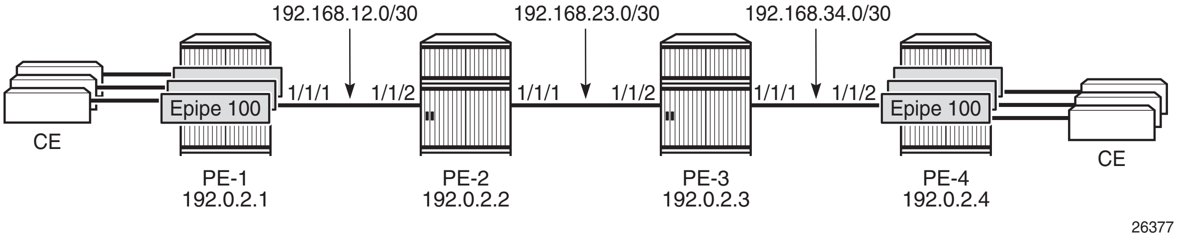
```
*A:PE-1# oam sdp-mtu 14 size-inc 490 500 step 1
Size    Sent    Response
----------------------------
490     .       Success
491     ...     Request Timeout

Maximum Response Size: 490
```

The SDP path MTU is 490 bytes.

PMTUD for LDP IPv4

Figure 274 shows that multiple Epipe services are configured on PE-1 and PE-4.

Figure 274 **Multiple Epipes Using LDP SDPs**



The following Epipes are configured on PE-1:

```
configure
    service
        epipe 100 customer 1 create
            sap 1/2/1:100 create
            exit
            spoke-sdp 14:100 create
            exit
            no shutdown
        exit
---snip---
        epipe 109 customer 1 create
            sap 1/2/1:109 create
            exit
            spoke-sdp 14:109 create
            exit
            no shutdown
        exit
```

The following configuration enables PMTUD for LDP IPv4 peer 192.0.2.4 on PE-1. The configuration is similar on PE-4.

```
configure
    router
```

```
        ldp
            tcp-session-parameters
                peer-transport 192.0.2.4
                    path-mtu-discovery
                exit
            exit
---snip---

*A:PE-1# show router ldp tcp-session-parameters ipv4

===============================================================================
LDP IPv4 TCP Session Parameters
===============================================================================
-------------------------------------------------------------------------------
Peer Transport: 192.0.2.4
-------------------------------------------------------------------------------
Authentication Key : Disabled           Path MTU Discovery : Enabled
Auth key chain     :                    Min-TTL            : 0
===============================================================================
No. of IPv4 Peers: 1
```

When LDP is disabled (**shutdown**) and re-enabled (**no shutdown)** on PE-1, all label mappings are signaled again.

```
*A:PE-1# configure router ldp shutdown
*A:PE-1# configure router ldp no shutdown
```

The size of the LDP label mapping messages may exceed the MTU between PE-2 and PE-3. The DF bit is set, so the packet is discarded at the egress of PE-2 to PE-3. PE-2 sends an ICMP error message of type 3 and code 4 to PE-1. The following ICMP error message is received on PE-1 when debugging is enabled for ICMP:

```
*A:PE-1# debug router ip icmp


2 2017/04/21 08:05:20.11 UTC MINOR: DEBUG #2001 Base PIP
"PIP: ICMP
instance 1 (Base), interface index 2 (int-PE-1-PE-2),
ICMP  ingressing on int-PE-1-PE-2:
   192.168.23.1 -> 192.0.2.1
   type: Destination Unreachable (3)  code: Fragmentation Needed and Don't Fragment
was Set (4)
"
```

On the egress interface "int-PE-2-PE-3" on PE-2, the network MTU is 512, the IP MTU is 498 (= 512 – 14 (Ethernet header)), and the TCP Maximum Segment Size (MSS) is 458 (= 498 – 20 (IP header) – 20 (TCP header)), as shown on PE-1:

```
*A:PE-1# show system connections port 646

===============================================================================
Connections
===============================================================================
Prot RecvQ   TxmtQ   Local Address                                   State
                MSS  Remote Address                                     vRtrID
```

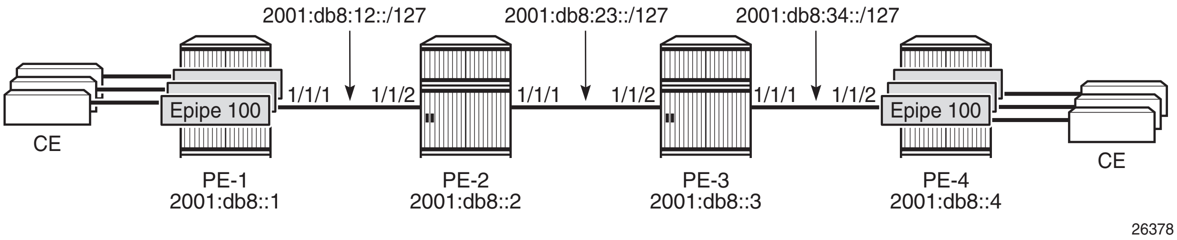
```
-------------------------------------------------------------------------------
TCP        0       0 192.0.2.1.646                                      LISTEN
              1024  0.0.0.0.0                                                1
TCP        0       0 192.0.2.1.646                                      ESTABLISH
              1024  192.0.2.2.50300                                          1
TCP        0       0 192.0.2.1.646                                      ESTABLISH
               458  192.0.2.4.50962                                          1
---snip---
```

TCP port 646 is used for LDP messages. The LDP TCP session with PE-2 keeps the (default) TCP MSS value of 1024, whereas the LDP TCP session with PE-4 has a reduced TCP MSS of 458 octets. PE-1 adapts the TCP MSS size to 458 and retransmits the LDP mapping messages to PE-4. With TCP MSS set to 458, no fragmentation is required along the path.

PMTUD for LDP IPv6

Multiple Epipes are configured between PE-1 and PE-4. Figure 275 shows the IPv6 addresses used.

Figure 275 **Multiple Epipes between PE-1 and PE-4 - IPv6**



The service configuration is the same as the preceding service configuration, but the far end of the SDP is an IPv6 address instead, as follows:

```
configure
    service
        sdp 146 mpls create
            far-end 2001:db8::4
            ldp
            no shutdown
        exit
```

With the configured network MTU of 512 on the link between PE-2 and PE-3, SDP 146 (for IPv6) is operationally down, whereas SDP 14 (for IPv4) is up, as follows:

```
*A:PE-1# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP    Sig
----------------------------------------------------------------------------
14     0       1578    192.0.2.4        Up   Up          MPLS    L      TLDP
146    0       1578                     Up   Down        MPLS    L      TLDP
                       2001:db8::4
----------------------------------------------------------------------------
Number of SDPs : 2
```

RFC 2460 IPv6 Specification states that links with a configurable MTU should have an MTU of at least 1280 octets; preferably 1500 or greater to accommodate possible tunneling encapsulations without the need for fragmentation.

In this example, the network MTU on the link between PE-2 and PE-3 is configured with a value of 1300 and SDP 146 will then be operationally up.

```
*A:PE-2# configure port 1/1/1 ethernet mtu 1300
*A:PE-3# configure port 1/1/2 ethernet mtu 1300
```

The SDP path MTU for SDP 146 is 1278 (= 1300 – 14 – 4 – 4). This can be verified on PE-1 with the following OAM command:

```
*A:PE-1# oam sdp-mtu 146 size-inc 1270 1280 step 1
Size    Sent    Response
----------------------------
1270    .       Success
1271    .       Success
1272    .       Success
1273    .       Success
1274    .       Success
1275    .       Success
1276    .       Success
1277    .       Success
1278    .       Success
1279    ...     Request Timeout

Maximum Response Size: 1278
```

PMTUD is enabled for LDP IPv6 peer 2001:db8::4 on PE-1, as follows:

```
configure
    router
        ldp
            tcp-session-parameters
                peer-transport  2001:db8::4
                    path-mtu-discovery
                exit
            exit
---snip---
```

```
*A:PE-1# show router ldp tcp-session-parameters ipv6

===============================================================================
LDP IPv6 TCP Session Parameters
===============================================================================
-------------------------------------------------------------------------------
Peer Transport: 2001:db8::4
-------------------------------------------------------------------------------
Authentication Key : Disabled           Path MTU Discovery : Enabled
Auth key chain     :                    Min-TTL            : 0
===============================================================================
No. of IPv6 Peers: 1
```

With an SDP path MTU of 1280 octets, it is extremely unlikely that LDP packets will exceed this size. An example of an ICMPv6 message that is sent when the packet is too big is shown for BGP in section PMTUD for BGP IPv6.

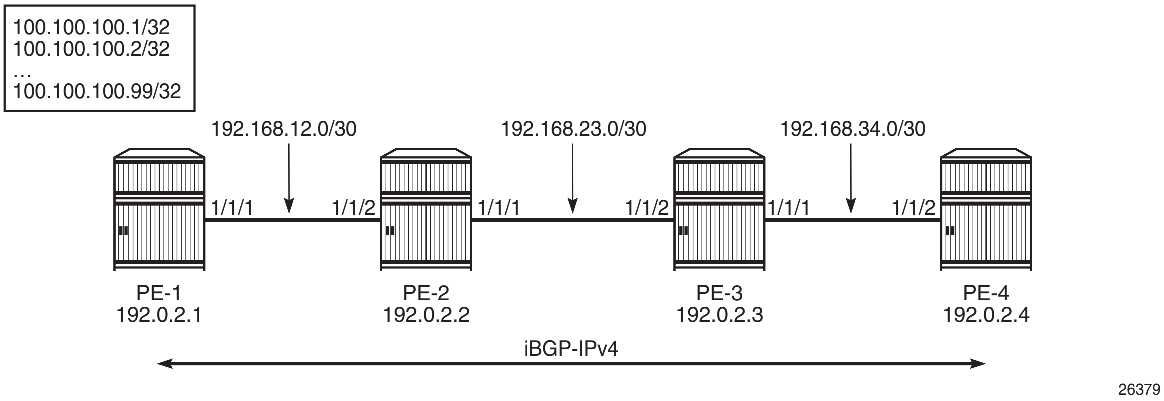
The TCP MSS for the IPv6 LDP connection between PE-1 and PE-4 is the default value of 1024 bytes. When the SDP path MTU is big enough for TCP segments with segments of 1024 bytes, the TCP MSS is set to 1024, unless **tcp-mss** is configured manually on the IPv6 interfaces. This TCP MSS value may change after an ICMPv6 "Packet Too Big" message is received on PE-1.

```
*A:PE-1# show system connections address 2001:db8::4 port 646

===============================================================================
Connections
===============================================================================
Prot RecvQ   TxmtQ   Local Address                                 State
                MSS  Remote Address                                   vRtrID
-------------------------------------------------------------------------------
TCP        0       0 2001:db8::1.646                              ESTABLISH
               1024  2001:db8::4.49715                                     1
-------------------------------------------------------------------------------
No. of Connections: 1
```

PMTUD for BGP IPv4

Figure 276 shows that a BGP session is established between PE-1 and PE-4 for address family IPv4. Static routes on PE-1 are exported as BGP routes to PE-4.

Figure 276 **BGP-IPv4**



The network port MTU on the link between PE-2 and PE-3 is set to 512 again:

```
*A:PE-2# configure port 1/1/1 ethernet mtu 512
*A:PE-3# configure port 1/1/2 ethernet mtu 512
```

BGP is configured for address family IPv4 on PE-1, as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            min-route-advertisement 1
            group "iBGP-IPv4"
                peer-as 64496
                neighbor 192.0.2.4
                    export "export-static"
                    path-mtu-discovery
                exit
            exit
        exit
        policy-options
            begin
            policy-statement "export-static"
                entry 10
                    from
                        protocol static
                    exit
                    action accept
                    exit
                exit
            exit
            commit
```

The export policy exports static routes as BGP routes to neighbor 192.0.2.4. PMTUD can be enabled in the global BGP context, per group, or per neighbor. In this example, PMTUD is enabled for neighbor 192.0.2.4. The configuration on PE-4 is similar, but with a neighbor 192.0.2.1 and without any export policy.

Also, a range of static routes is configured on PE-1 to ensure that the size of the BGP update messages will be larger than the SDP path MTU, as follows:

```
configure
    router
        static-route-entry 100.100.100.1/32 black-hole no shutdown
        static-route-entry 100.100.100.2/32 black-hole no shutdown
---snip---
        static-route-entry 100.100.100.99/32 black-hole no shutdown
        exit all
```

Debugging is enabled for ICMP, as follows:

```
*A:PE-1# debug router ip icmp
```

BGP is disabled and re-enabled to ensure that all BGP routes are re-advertised to PE-4. The BGP route update messages exceed the MTU on the egress port of PE-2 to PE-3, and PE-2 should have to fragment them to be able to forward them on the egress interface toward PE-3, but the DF bit is set. Therefore, PE-2 discards the packet and sends an ICMP error message to PE-1 of type 3 ("Destination Unreachable") and code 4 ("Fragmentation Needed and Don't Fragment was Set"). PE-1 receives the following ICMP error message:

```
1 2017/04/21 14:06:56.84 UTC MINOR: DEBUG #2001 Base PIP
"PIP: ICMP
instance 1 (Base), interface index 2 (int-PE-1-PE-2),
ICMP  ingressing on int-PE-1-PE-2:
   192.168.23.1 -> 192.0.2.1
   type: Destination Unreachable (3)  code: Fragmentation Needed and Don't Fragment
was Set (4)
"
```

The following output shows that the TCP MSS for the BGP connection between PE-1 and PE-4 is 458. TCP destination port 179 is used for BGP traffic.

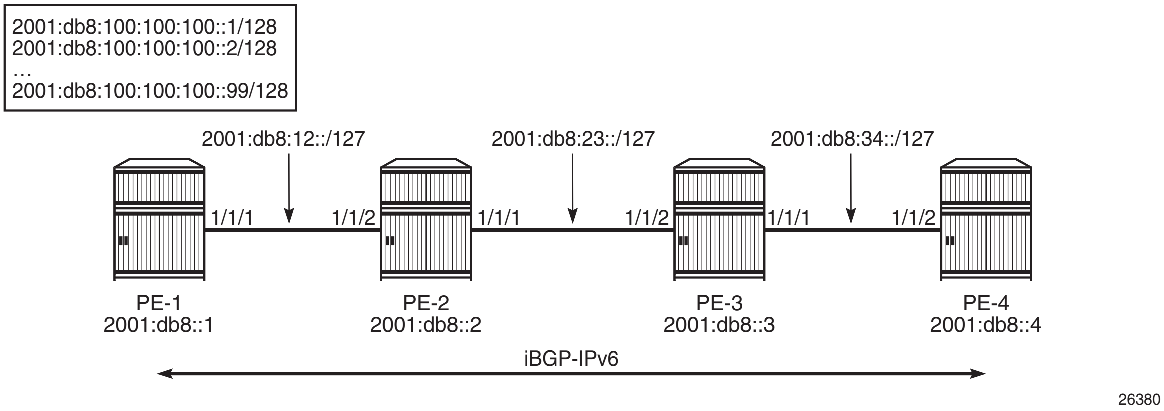
```
*A:PE-1# show system connections port 179

===============================================================================
Connections
===============================================================================
Prot RecvQ   TxmtQ   Local Address                                  State
                MSS  Remote Address                                      vRtrID
-------------------------------------------------------------------------------
TCP        0       0 0.0.0.0.179                                    LISTEN
               1024  0.0.0.0.0                                               1
TCP        0       0 192.0.2.1.50218                                ESTABLISH
                458  192.0.2.4.179                                           1
TCP        0       0 ::.179                                         LISTEN
               1024  ::.0                                                    1
-------------------------------------------------------------------------------
No. of Connections: 3
```

The TCP MSS is calculated as follows: 512 – 14 – 20 – 20 = 458, where 512 is the lowest network port MTU in the path, 14 bytes are used for the Ethernet header, 20 bytes for the IPv4 header, and 20 bytes for the TCP header.

PMTUD for BGP IPv6

Figure 277 shows that a BGP session is established between PE-1 and PE-4 for address family IPv6. PE-1 exports a range of IPv6 routes to PE-4.

Figure 277 **BGP-IPv6**



The network port MTU on the link between PE-2 and PE-3 is set to 1300 again:

```
*A:PE-2# configure port 1/1/1 ethernet mtu 1300
*A:PE-3# configure port 1/1/2 ethernet mtu 1300
```

The BGP configuration is similar for IPv6 to the configuration for IPv4, only the BGP address family and the neighbor addresses are different. The export policy is identical. PMTUD is enabled in the BGP group "iBGP-IPv6". The static routes have now IPv6 addresses. The configuration on PE-1 is as follows:

```
configure
    router
        autonomous-system 64496
        bgp
            group "iBGP-IPv6"
                family ipv6
                peer-as 64496
                path-mtu-discovery
                neighbor 2001:db8::4
                    export "export-static"
                exit
            exit
        exit
        static-route-entry 2001:db8:100:100:100::1/128 black-hole no shutdown
        static-route-entry 2001:db8:100:100:100::2/128 black-hole no shutdown
        ---snip---
        static-route-entry 2001:db8:100:100:100::99/128 black-hole no shutdown
```

The configuration on PE-4 resembles this configuration, but with a different neighbor address. When the group "iBGP-IPv6" is disabled and re-enabled, PE-1 advertises all the IPv6 routes to its peer 2001:db8::4. PE-2 cannot forward the large BGP messages and discards them. PE-2 sends an ICMPv6 error message to PE-1 indicating that the packet is too big (type 2, code 0). PE-1 receives the following ICMPv6 error message:

```
*A:PE-1# debug router ip icmp6


5 2017/04/21 14:25:20.60 UTC MINOR: DEBUG #2001 Base TIP
"TIP: ICMP6_PKT
ICMP6 ingressing on int-PE-1-PE-2 (Base):
   2001:db8:23:: -> 2001:db8::1
   Type: Packet Too Big (2)
   Code: No Code (0)
   MTU : 1286
"
```

The MTU is 1286 and includes the IP header and the IP data, but not the Ethernet header. The calculation is as follows: 1300 – 14 = 1286, where 1300 is the lowest network port MTU in the path and 14 bytes are used for the Ethernet header.

On PE-1, the TCP MSS for BGP traffic with destination address 2001:db8::4 is 1226, as follows:

```
*A:PE-1# show system connections port 179

===============================================================================
Connections
===============================================================================
Prot RecvQ   TxmtQ   Local Address                                    State
                MSS  Remote Address                                     vRtrID
-------------------------------------------------------------------------------
TCP        0       0 0.0.0.0.179                                      LISTEN
              1024  0.0.0.0.0                                              1
TCP        0       0 192.0.2.1.50218                                  ESTABLISH
               458  192.0.2.4.179                                          1
TCP        0       0 ::.179                                           LISTEN
              1024  ::.0                                                   1
TCP        0       0 2001:db8::1.50221                                ESTABLISH
              1226  2001:db8::4.179                                        1
-------------------------------------------------------------------------------
No. of Connections: 4
```

The TCP MSS is calculated as follows: 1300 – 14 – 40 – 20 = 1226, where 1300 is the lowest network port MTU in the path, 14 bytes are used for the Ethernet header, 40 bytes for the IPv6 header, and 20 bytes for the TCP header. This TCP MSS value is larger than the default value of 1024, so the ICMPv6 "Packet Too Big" message can result in a larger TCP MSS value.

Conclusion

PMTUD is a technique to determine the MTU size on the network path between two IP hosts, to maximize packet efficiency and avoid packet fragmentation. PMTUD can be enabled for LDP and BGP connections.

PCEP Support for RSVP-TE LSPs

This chapter provides information about PCEP Support for RSVP-TE LSPs.

Topics in this chapter include:

Applicability

The information and configuration in this chapter is based on SR OS Release 15.0.R1.

Overview

This chapter describes how to use an external controller to compute RSVP-TE LSPs.

Without external controller, the source-routed path computation of an RSVP-TE LSP is achieved by the head end router examining its own Traffic Engineering database (TE-DB) and computing an end-to-end path comprising a list of IP hops. For this to be achieved, the **cspf** keyword must be enabled within the LSP CLI construct of the LSP.

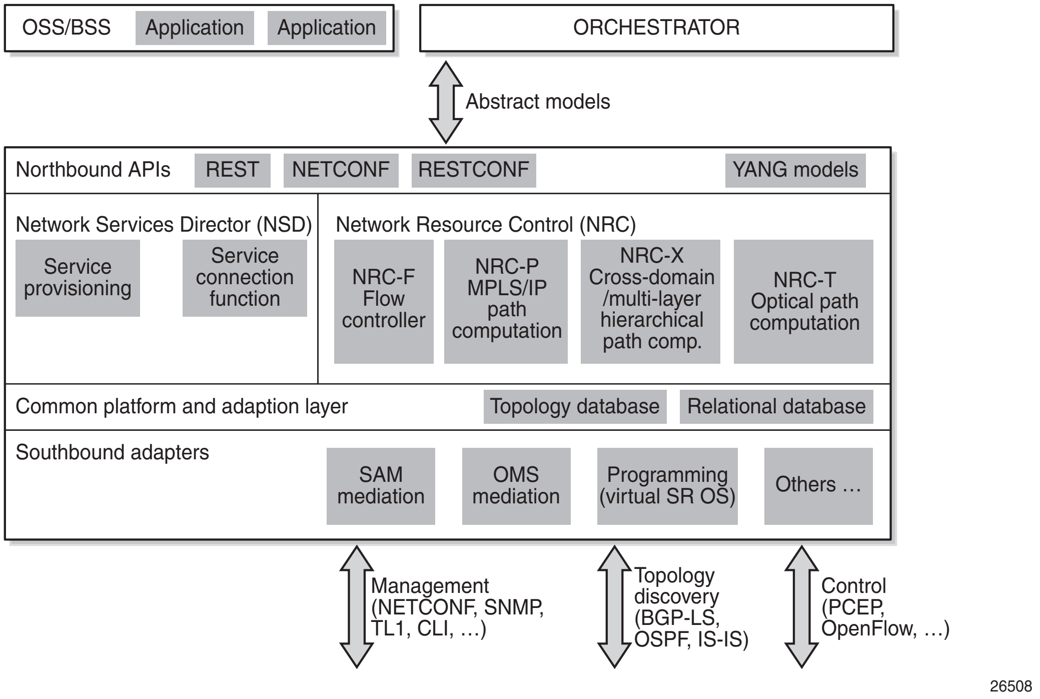
The computed path is inserted into the Explicit Route Object (ERO) of the RSVP Path message, and forwarded out of the interface toward the first hop router, determined by the first entry in the ERO. At each hop, the relevant router will examine the ERO within the Path message, and forward the message toward the next downstream router through the outgoing interface indicated by the top address in the ERO. The router then removes the top ERO entry and forwards the Path message. At the same time, the router creates an entry in the Record Route Object (RRO) that matches the address of the incoming interface (the address of the interface through which the Path message is received).

It is possible for a head end router to request an external controller to compute a path between head and tail end routers, rather than compute it locally. This is useful when, for example, the LSP is to be terminated on a router in a different routing domain from the source router, for which it has no view of the topology. The external controller must be aware of the end-to-end topology; it must have a complete TE topology database of all areas that it can use to compute an end-to-end path.

The external controller that computes a path is the Path Computation Element (PCE). In this case, it is the Network Resources Controller - Path (NRC-P), which runs within the Network Services Platform (NSP). The networking interface to the NRC-P is the Virtual Service Router - Network Resources Controller (VSR-NRC). The VSR-NRC is a virtual SR (vSR) OS instance that can run on a Linux server. The instance has a physical interface into the network, and collects topology information along with signaled path computation requests from head end routers.

Figure 278 shows a block diagram of the NSP layout. The NRC-P and its path computation elements are highlighted.

Figure 278 Network Services Platform Block Diagram


OSS/BSS
Application
Application
ORCHESTRATOR
Abstract models
Northbound APIs
REST
NETCONF
RESTCONF
YANG models
Network Services Director (NSD)
Service provisioning
Service connection function
Network Resource Control (NRC)
NRC-F Flow controller
NRC-P MPLS/IP path computation
NRC-X Cross-domain /multi-layer hierarchical path comp.
NRC-T Optical path computation
Common platform and adaption layer
Topology database
Relational database
Southbound adapters
SAM mediation
OMS mediation
Programming (virtual SR OS)
Others …
Management (NETCONF, SNMP, TL1, CLI, …)
Topology discovery (BGP-LS, OSPF, IS-IS)
Control (PCEP, OpenFlow, …)
26508

The creation of the vSR OS instance is outside the scope of this chapter. Also, there are several ways that the NRC-P can learn the network topology. For this chapter, it is just assumed that the NRC-P communication is configured, and that the complete topology database has been learned.

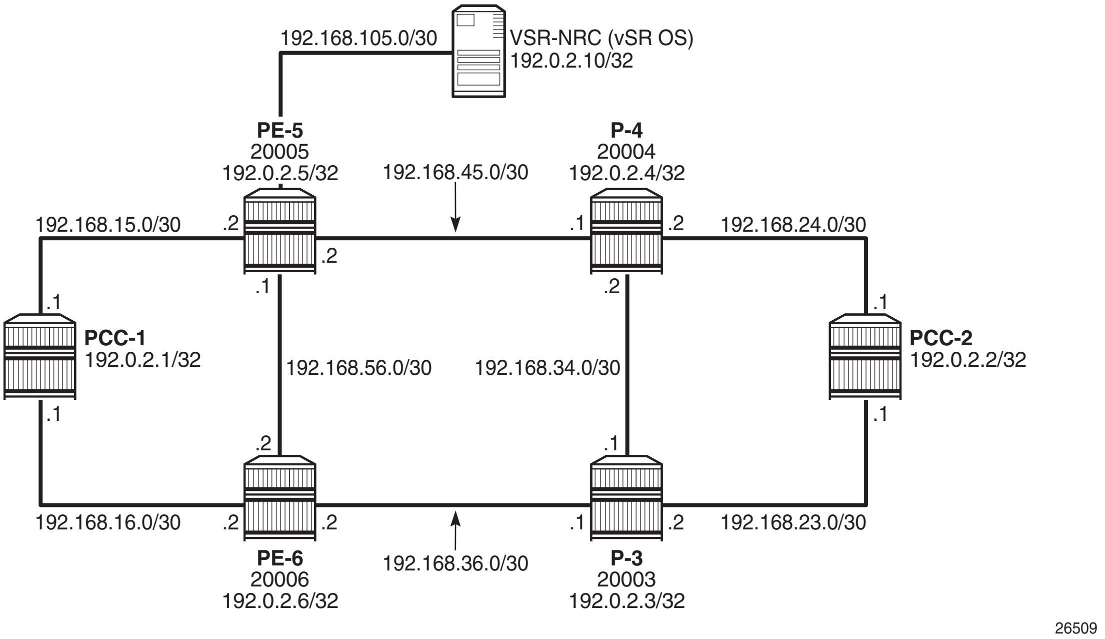
The following configuration describes how the NRC-P can compute an RSVP-TE LSP, for a:

- path with no constraints (zero hop path)
- path that is constrained using strict hops
- primary and secondary standby path, constrained and diversely-routed using admin groups

Configuration

Figure 279 shows the example topology. The VSR-NRC is connected to the network at PE-5. This vSR OS runs an IS-IS instance so that it is reachable by all routers in the network. For clarity, the NSP/NRC-P has been removed from the diagram. The NRC-P will be referred to as the Path Computation Element (PCE) throughout the remainder of the chapter.

Figure 279 **Example Topology**



Global IS-IS Configuration

The first step is to configure IS-IS on each router seen in Figure 279. All routers are members of a single level 2 area 49.0001.

The configuration for Path Computation Client PCC-1 to enable IS-IS is as follows:

```
A:PCC-1>config>router>isis# info
----------------------------------------------
            level-capability level-2
            area-id 49.0001
            level 2
                wide-metrics-only
            exit
            interface "system"
                level-capability level-2
                ipv4-node-sid index 420
                no shutdown
            exit
            interface "int-PCC-1-PE-5"
                level-capability level-2
                interface-type point-to-point
                level 2
                    metric 1000
                exit
                no shutdown
            exit
            interface "int-PCC-1-PE-6"
                level-capability level-2
                interface-type point-to-point
                level 2
                    metric 1000
                exit
                no shutdown
            exit
            no shutdown
```

The configuration for all other nodes is the same, apart from the IP addresses. The IP addresses can be derived from Figure 279.

Path Computation Element Protocol (PCEP)

The PCE is a vSR OS router instance serving as an interface between the physical network and the NRC-P. The instance has a direct physical connection to the network, and has a northbound interface toward the PCE within the NSP. The PCE communicates with its PCCs using the TCP-based protocol, PCEP. The TCP session is initiated by each client, but must be enabled on the PCE, as follows:

```
*A:PCE# configure router pcep
----------------------------------------------
```

```
            pce
                local-address 192.0.2.10
                no shutdown
            exit
----------------------------------------------
```

The local address is the system address, and is used as the source address for PCEP messaging between itself and the PCCs, when in-band communication is used. The management routing instance could also be used for out-of-band PCEP communication.

On each PCC, the PCE configuration specifies the VSR-NRC as the peer, using the local address of the PCE as the peer address, as follows:

```
A:PCC-1# configure router pcep
----------------------------------------------
            pcc
                local-address 192.0.2.1
                peer 192.0.2.10
                    no shutdown
                exit
                no shutdown
            exit
----------------------------------------------
```

Again, the local address is configured and is used as the source address for PCEP messages by the PCC. For in-band communication, the system address is used.

The following output shows the state of the PCEP sessions on the PCE. There are two sessions: one to each of PCC-1 and PCC-2.

```
*A:PCE# show router pcep pce peer

===============================================================================
PCEP Path Computation Element (PCE) Peer Info
===============================================================================
Peer                       Sync State             Oper Keepalive/Oper DeadTimer
-------------------------------------------------------------------------------
192.0.2.1:4189           done                   30/120
192.0.2.2:4189           done                   30/120
-------------------------------------------------------------------------------
No. of Peers: 2
===============================================================================
```

The PCEP session to PCC-1 is shown in more detail in the following output. The peer capabilities show that the computation of RSVP paths is supported. Stateful delegation capability is negotiated between the PCE and PCC. The PCC can delegate control of an LSP to the PCE so that if there is a requirement to modify the existing path of the LSP, the PCE will resignal a new path using a PCEP update message.

This PCEP session requires stateful PCE, so that the state of the LSP (both RSVP and Segment Routing TE (SR-TE) is reported to the PCE by the PCC. This state change could be a change in configuration, or any change due to a received PCE update.

```
*A:PCE# show router pcep pce peer 192.0.2.1

===============================================================================
PCEP Path Computation Element (PCE) Peer Info
===============================================================================
IP Address             : 192.0.2.1    Port                   : 4189
Sync State             : done
Peer Capabilities      : stateful-delegate stateful-pce segment-rt-path rsvp-
                         path
Speaker ID             : 2a:e1:ff:00:00:00
Session Establish Time : 8d 00:00:21
Oper Keepalive         : 30 seconds     Oper DeadTimer          : 120 seconds
===============================================================================
```

PCE Computed RSVP-TE LSP with Zero-hop Path

The following output shows an RSVP-TE LSP configured on PCC-1 with the tail end on PCC-2.

```
*A:PCC-1# configure router mpls
----------------------------------------------
    path "pce-controlled"
         no shutdown
    exit
    lsp "LSP-PCC-1-PCC-2"
                to 192.0.2.2
                cspf
                pce-computation
                pce-report enable
                pce-control
                primary "pce-controlled"
                exit
                no shutdown
            exit
```

The MPLS path is a loose path containing no hops and is applied as a primary path within the LSP construct.

The **pce-computation** command forces a PCEP Request by the router to the PCE for a valid path between PCC-1 and PCC-2, computed by the NRC-P. The PCE replies using a PCEP Reply with a valid path, or a no-path message if no valid path exists.

The PCC reports the state of the LSP to the PCE if the **pce-report enable** command is configured.

The pce-control command allows the router to delegate control of the LSP to the PCE. Because the PCE is aware of the full topology, if an event occurs that affects the state of the LSP, for example, a link or node failure, then the PCE will send a PCEP Update with a list of hops representing a new path (if a new path is available).

Debug: PCEP Messaging when LSP is Enabled

The following output shows the PCEP messaging in the form of debug, when the LSP is placed in a "no shutdown" state.

The PCC sends a PCReq message to the PCE, requesting the computation of an RSVP-TE Path. Because the PST(SegRt) field is set to zero, the path request is not a segment routing path request, so must be an RSVP-TE request. The PCEP LSP-ID (PLSP-ID) is set by the PCC. This value is used in all PCEP messages between PCE and PCC for the lifetime of the LSP. In this case, the value is set to 38. When the PCC or PCE sends a PCEP message with this value set, it refers to this specific LSP.

The source and destination addresses are 192.0.2.1 and 192.0.2.2, respectively. The LSP name and path name are shown in text form: **LSP-PCC-1-PCC-2::pce-controlled**.

```
10 2017/02/16 15:13:23.18 UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REQUEST ][001 23:53:12.130]
  Svec :{numOfReq 1}{nodeDiverse F linkDiverse F srlgDiverse F}
    Request: {id 29 PST(SegRt) 0 srcAddr 192.0.2.1 destAddr 192.0.2.2}
             {PLspId 38 tunnelId:3 lspId: 9736 lspName LSP-PCC-1-PCC-2::pce-
             controlled}
             {{bw 0 (0) isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
```

The following shows the PCReply received by the PCC from the PCE, showing that a valid path has been computed within the constraints of the path request, and contains a list of strict IPv4 hops (isLoose flag is set to F (false), for each hop).

```
12 2017/02/16 15:13:23.26 UTC MINOR: DEBUG #2001 Base PCC
"PCC: [RX-Msg: REPLY ][001 23:53:12.210]
[Peer 192.0.2.10]
  Request: {id 29} {} Response has calculated path
        {{Total Paths: 1}}
          Path: {PST(SegRt) 0}
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.45.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.24.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 1110 hop 3}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0}}}
```

Upon receipt of the PCReply, PCC-1 will signal the LSP with an RSVP Path message, as shown in the following output. The Path message contains an ERO comprising the same hops as those received in the PCReply.

```
29 2017/02/16 15:26:16.87 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.2
          TTL:255, Checksum:0xc440, Flags:0x1
MSG ID     - Flags:0x1, Epoch:10790745, MsgId:128
Session    - EndPt:192.0.2.2, TunnId:3, ExtTunnId:192.0.2.1
SessAttr   - Name:LSP-PCC-1-PCC-2::pce-controlled
             SetupPri:7, HoldPri:0, Flags:0x6
RSVPHop    - Ctype:1, Addr:192.168.15.1, LIH:2
TimeValue  - RefreshPeriod:180
SendTempl  - Sender:192.0.2.1, LspId:9740
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8686
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.15.1, Flags:0x0
ERO        - IPv4Prefix 192.168.15.2/32, Strict
             IPv4Prefix 192.168.45.1/32, Strict
             IPv4Prefix 192.168.24.1/32, Strict
"
```

PCC-1 receives an RSVP Resv message with the RRO containing a list of hops, with a label mapping per hop.

```
30 2017/02/16 15:26:16.88 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESV Msg
Recv RESV From:192.168.15.2, To:192.168.15.1
          TTL:255, Checksum:0xfa6a, Flags:0x1
MSG ID     - Flags:0x1, Epoch:14080719, MsgId:35
Session    - EndPt:192.0.2.2, TunnId:3, ExtTunnId:192.0.2.1
RSVPHop    - Ctype:1, Addr:192.168.15.2, LIH:2
TimeValue  - RefreshPeriod:180
Style      - SE
FlowSpec   - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8686, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.1, LspId:9740, Label:262120
LspAttr    - Attribute Flags TLV  0x400000
RRO        - InterfaceIp:192.168.15.2, Flags:0x0
             Label:262120, Flags:0x1
             InterfaceIp:192.168.45.1, Flags:0x0
             Label:262105, Flags:0x1
             InterfaceIp:192.168.24.1, Flags:0x0
             Label:262134, Flags:0x1
```

The PCC then sends a PC LSP State Report message to the PCE with the state of the LSP set to Admin=1 (up), and OperState = 2 (up and carrying traffic).

The ERO is a copy of the ERO contained in the PCReply, and the RRO is a copy of the RRO contained in the RSVP Resv message.

```
"PCC: [TX-Msg: REPORT ][002 00:06:05.830]
[Peer 192.0.2.10]
```

```
  Report : {srpId:0 PST(SegRt):0 PLspId:40 lspId: 9740 tunnelId:3}
    {Sync 0 Rem 0 AdminState 1 OperState 2 Delegate 1 Create 0}
    {srcAddr 192.0.2.1 destAddr 192.0.2.2 extTunnelId :: pathName LSP-PCC-1-PCC-2::
                                                         pce-controlled}
    {Binding Type: 0 Binding Val : 0}
    Lsp Constraints:
             {{bw 0 isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
      Ero Path:
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.45.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.24.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 1110 hop 4}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0}}}
      RRO:
            {{type IPv4 addr 192.168.15.1/32 Flag 0}}
            {{type IPv4 addr 192.168.15.2/32 Flag 0}}
            {{type Label label c-type 1 label 262120}}
            {{type IPv4 addr 192.168.45.1/32 Flag 0}}
            {{type Label label c-type 1 label 262105}}
            {{type IPv4 addr 192.168.24.1/32 Flag 0}}
            {{type Label label c-type 1 label 262134}}
      {Lsp Err NA RsvpErr 0 LspDbVersion 0}
```

When the LSP has connected, the following **show** command output shows the state of the LSP Path.

```
*A:PCC-1# show router mpls lsp "LSP-PCC-1-PCC-2" path detail

===============================================================================
MPLS LSP LSP-PCC-1-PCC-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PCC-1-PCC-2 Path pce-controlled
-------------------------------------------------------------------------------
LSP Name         : LSP-PCC-1-PCC-2
Path LSP ID      : 9740
From             : 192.0.2.1          To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : pce-controlled       Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1:220            Out Label            : 262114
Path Up Time     : 0d 00:09:06          Path Down Time       : 0d 00:00:00
Retry Limit      : 0                    Retry Timer          : 30 sec
Retry Attempt    : 0                    Next Retry In        : 0 sec
BFD Template     : None                 BFD Ping Interval    : 60
BFD Enable       : False

Adspec           : Disabled             Oper Adspec          : Disabled
```

```
CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
FRR              : Disabled             Oper FRR             : Disabled
Propogate Adm Grp: Disabled             Oper Prop Adm Grp    : Disabled
Inter-area       : False

PCE Updt ID      : 0

PCE Report       : Enabled              Oper PCE Report      : Enabled
PCE Control      : Enabled              Oper PCE Control     : Disabled
PCE Compute      : Enabled              Oper PCE Compute     : Disabled

Neg MTU          : 8686                 Oper MTU             : 8686
Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Record Route     : Record               Oper Record Route    : Record
Record Label     : Record               Oper Record Label    : Record
Setup Priority   : 7                    Oper Setup Priority  : 7
Hold Priority    : 0                    Oper Hold Priority   : 0
Class Type       : 0                    Oper CT              : 0
Backup CT        : None
MainCT Retry     : n/a
    Rem          :
MainCT Retry     : 0
    Limit        :
Include Groups   :                      Oper Include Groups  :
None                                           None
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None

Adaptive         : Enabled              Oper Metric          : 1110
Preference       : n/a
Path Trans       : 3                    CSPF Queries         : 2
Failure Code     : noError
Failure Node     : n/a
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.15.1 (192.0.2.1)                    Record Label        : N/A
 -> 192.168.15.2 (192.0.2.5)                    Record Label        : 262120
 -> 192.168.45.1 (192.0.2.4)                    Record Label        : 262105
 -> 192.168.24.1 (192.0.2.2)                    Record Label        : 262134
Computed Hops    :
    192.168.15.1(S)
 -> 192.168.15.2(S)
 -> 192.168.45.1(S)
 -> 192.168.24.1(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 1110
===============================================================================
```

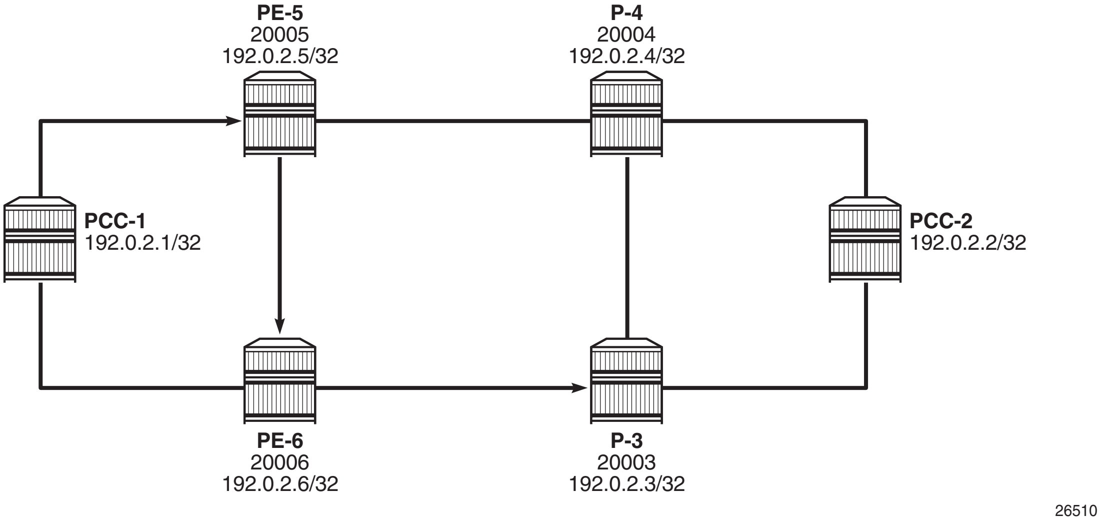
PCE-Computed RSVP-TE LSP with Strict Hop Path

It is possible for the PCC to influence the path computed by the PCE by including a primary path with one or more explicit hops. This path is translated into a list of hops in the Include Route Object (IRO) in the PCEP PC Request, which is sent from the PCC to the PCE at the time of the path request.

The PCE will take the hops listed in the IRO into account when computing an end-to-end path, as the following example shows.

Figure 280 shows an RSVP-TE LSP with source PCC-1 and destination PCC-2, which has a requirement to follow a path via PE-5, PE-6, and P-3.

Figure 280 **RSVP-TE LSP with Strict Hops**



The following output shows the path converted to a list of strict hops.

```
*A:PCC-1>config>router>mpls# path pce-controlled-strict
---------------------------------------------
                hop 1 192.0.2.5 strict
                hop 2 192.0.2.6 strict
                hop 3 192.0.2.3 strict
                no shutdown
```

The use of strict hops requires that each consecutive hop must be contiguous from the previous hop. Applying this path to an RSVP-TE LSP with PCEP commands included is as follows:

```
*A:PCC-1>config>router>mpls# lsp "PCC-1-PCC-2-RSVP-PCE-strict-001"
---------------------------------------------
                to 192.0.2.2
```

```
cspf
pce-computation
pce-report enable
pce-control
primary "pce-controlled-strict"
exit
no shutdown
```

The following **show** command output shows that the LSP path is connected.

```
*A:PCC-1# show router mpls lsp "PCC-1-PCC-2-RSVP-PCE-strict-001" path detail

===============================================================================
MPLS LSP PCC-1-PCC-2-RSVP-PCE-strict-001 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP PCC-1-PCC-2-RSVP-PCE-strict-001 Path pce-controlled-strict
-------------------------------------------------------------------------------
LSP Name         : PCC-1-PCC-2-RSVP-PCE-strict-001
Path LSP ID      : 31746
From             : 192.0.2.1          To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : pce-controlled-      Path Type            : Primary
                   strict
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1:220            Out Label            : 262088
Path Up Time     : 0d 00:00:53          Path Down Time       : 0d 00:00:00
Retry Limit      : 0                    Retry Timer          : 30 sec
Retry Attempt    : 0                    Next Retry In        : 0 sec
BFD Template     : None                 BFD Ping Interval    : 60
BFD Enable       : False

Adspec           : Disabled             Oper Adspec          : Disabled
CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
FRR              : Disabled             Oper FRR             : Disabled
Propogate Adm Grp: Disabled             Oper Prop Adm Grp    : Disabled
Inter-area       : False

PCE Updt ID      : 0

PCE Report       : Enabled              Oper PCE Report      : Enabled
PCE Control      : Enabled              Oper PCE Control     : Enabled
PCE Compute      : Enabled              Oper PCE Compute     : Enabled

Neg MTU          : 8686                 Oper MTU             : 8686
Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Record Route     : Record               Oper Record Route    : Record
Record Label     : Record               Oper Record Label    : Record
Setup Priority   : 7                    Oper Setup Priority  : 7
```

```
Hold Priority    : 0                    Oper Hold Priority   : 0
Class Type       : 0                    Oper CT              : 0
Backup CT        : None
MainCT Retry     : n/a
    Rem          :
MainCT Retry     : 0
    Limit        :
Include Groups   :                      Oper Include Groups  :
None                                           None
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None

Adaptive         : Enabled              Oper Metric          : 2200
Preference       : n/a
Path Trans       : 3                    CSPF Queries         : 0
Failure Code     : noError
Failure Node     : n/a
Explicit Hops    :
    192.0.2.5(S)      -> 192.0.2.6(S)      -> 192.0.2.3(S)
Actual Hops      :
    192.168.15.1 (192.0.2.1)                    Record Label        : N/A
 -> 192.168.15.2 (192.0.2.5)                    Record Label        : 262088
 -> 192.168.56.2 (192.0.2.6)                    Record Label        : 262088
 -> 192.168.36.1 (192.0.2.3)                    Record Label        : 262070
 -> 192.168.23.1 (192.0.2.2)                    Record Label        : 262112
Computed Hops    :
    192.168.15.2(S)
 -> 192.168.56.2(S)
 -> 192.168.36.1(S)
 -> 192.168.23.1(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 2200
```

The explicit hops of the path configuration are shown, with the "(S)" signifying that the hops are configured as strict hops. The "Actual Hops" show that the strict hops are enforced as per the RSVP-TE LSP configuration. The "Computed Hops" are taken from the Path object in PCEP Reply, as shown in the following debug output.

Debug: PCEP Messaging for Path Computation

The PCC sends a PCReq message to the PCE requesting the computation of an RSVP-TE Path. The source and destination addresses are 192.0.2.1 and 192.0.2.2, respectively. The LSP name and path name are shown in text form: **PCC-1-PCC-2-RSVP-PCE-strict-001::pce-controlled-strict**. The request contains an IRO, containing the configured hops from the MPLS path configuration. Each hop contains an isLoose flag set to F (false), which implies that the hop is strict (that is, not loose). The following debug output shows the PCEP messaging.

```
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REQUEST ]
  Svec :{numOfReq 1}{nodeDiverse F linkDiverse F srlgDiverse F}
    Request: {id 13281 PST(SegRt) 0 srcAddr 192.0.2.1 destAddr 192.0.2.2}
```

```
             {PLspId 13314 tunnelId:2 lspId: 31746 lspName PCC-1-PCC-2-RSVP-PCE-
strict-001::pce-controlled-strict}
             {{bw 0 (0) isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
             {{iro isOpt: F}}
                 {{ctype IPv4 addr 192.0.2.5/0 isLoose F}}
                 {{ctype IPv4 addr 192.0.2.6/0 isLoose F}}
                 {{ctype IPv4 addr 192.0.2.3/0 isLoose F}}
"
```

The PCReply received by PCC-1 from the PCE has computed a valid path. This is shown in the following output.

```
6 2017/03/15 15:36:05.55 UTC MINOR: DEBUG #2001 Base PCC
"PCC: [RX-Msg: REPLY ][022 01:06:53.520]
[Peer 192.0.2.10]
  Request: {id 13281} {} Response has calculated path
        {{Total Paths: 1}}
          Path: {PST(SegRt) 0}
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.56.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.36.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.23.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2200 hop 4}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0}}}
"
```

The path in the PCReply is the path replicated in the "Computed Hops" of the preceding **show router mpls lsp path detail** command output.

Upon receipt of the PCReply from the PCE, the PCC uses the computed hops from the PCReply in the ERO of the RSVP Path message.

```
7 2017/03/15 15:36:05.54 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.2
          TTL:255, Checksum:0x5311, Flags:0x1
MSG ID     - Flags:0x1, Epoch:2387139, MsgId:8165
Session    - EndPt:192.0.2.2, TunnId:2, ExtTunnId:192.0.2.1
SessAttr   - Name:PCC-1-PCC-2-RSVP-PCE-strict-001::pce-controlled-strict
             SetupPri:7, HoldPri:0, Flags:0x6
RSVPHop    - Ctype:1, Addr:192.168.15.1, LIH:2
TimeValue  - RefreshPeriod:180
SendTempl  - Sender:192.0.2.1, LspId:31746
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8686
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.15.1, Flags:0x0
ERO        - IPv4Prefix 192.168.15.2/32, Strict
             IPv4Prefix 192.168.56.2/32, Strict
             IPv4Prefix 192.168.36.1/32, Strict
             IPv4Prefix 192.168.23.1/32, Strict
"
```

A PCEP Report is then sent to the PCE when the path is connected.

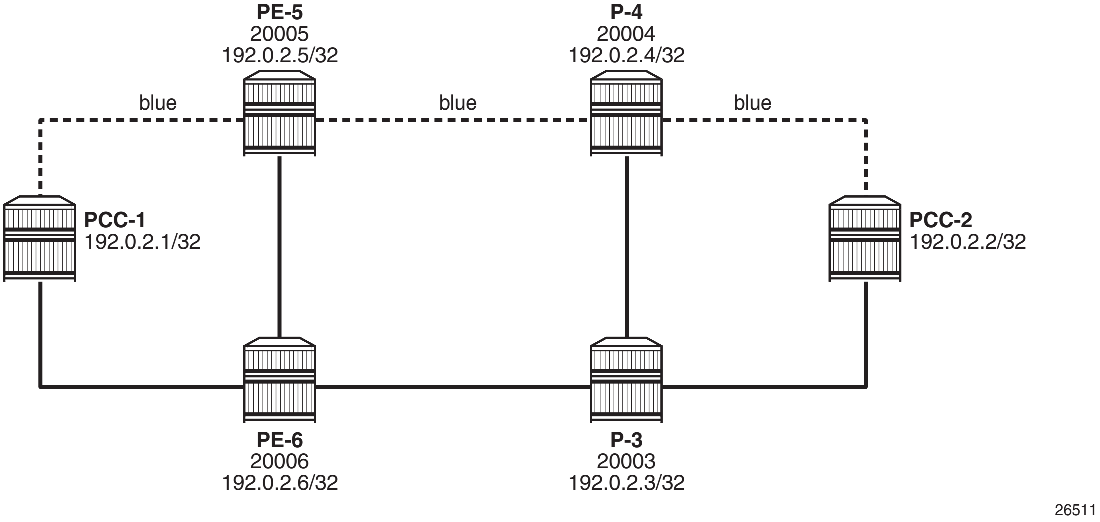
```
8 2017/03/15 15:36:05.56 UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REPORT ][022 01:06:53.540]
[Peer 192.0.2.10]
  Report : {srpId:0 PST(SegRt):0 PLspId:13314 lspId: 31746 tunnelId:2}
    {Sync 0 Rem 0 AdminState 1 OperState 2 Delegate 1 Create 0}
    {srcAddr 192.0.2.1 destAddr 192.0.2.2 extTunnelId ::
                   pathName PCC-1-PCC-2-rsvp-pce-strict-001::pce-controlled-strict}
    {Binding Type: 0 Binding Val : 0}
    Lsp Constraints:
             {{bw 0 isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
             {{iro isOpt: F}}
                 {{ctype IPv4 addr 192.0.2.5/0 isLoose F}}
                 {{ctype IPv4 addr 192.0.2.6/0 isLoose F}}
                 {{ctype IPv4 addr 192.0.2.3/0 isLoose F}}
      Ero Path:
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.56.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.36.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.23.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2200 hop 5}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 0 inclAll 0}}}
      RRO:
            {{type IPv4 addr 192.168.15.1/32 Flag 0}}
            {{type IPv4 addr 192.168.15.2/32 Flag 0}}
            {{type Label label c-type 1 label 262088}}
            {{type IPv4 addr 192.168.56.2/32 Flag 0}}
            {{type Label label c-type 1 label 262088}}
            {{type IPv4 addr 192.168.36.1/32 Flag 0}}
            {{type Label label c-type 1 label 262070}}
            {{type IPv4 addr 192.168.23.1/32 Flag 0}}
            {{type Label label c-type 1 label 262112}}
      {Lsp Err NA RsvpErr 0 LspDbVersion 0}
"
```

PCE-Computed RSVP-TE LSP with Primary and Secondary Paths using Admin Groups for Diversity

The PCE can compute primary and secondary paths on behalf of the PCC. Admin groups can be used to ensure that the zero-hop paths are diverse.

Figure 281 shows that there are two diverse paths between PCC-1 and PCC-2. The upper path via PE-5 and P-4 will have interfaces configured with an admin group called "blue".

Figure 281 **Admin Groups**



Admin groups are configured under the **configure router** context, and are applied to the router interface under the **configure router mpls** context. The following output is an example of an admin group created and applied on router PCC-1.

```
A:PCC-1>config>router# info
--------------------------------------------------
        if-attribute
            admin-group "blue" value 10
        exit
        mpls
            interface "int-PCC-1-PE-5"
                admin-group "blue"
                no shutdown
            exit
        exit
```

Similarly, for PE-5, the admin group configuration is as follows:

```
A:PE-5>config>router# info
--------------------------------------------------
        if-attribute
            admin-group "blue" value 10
        exit
        mpls
             interface "int-PE-5-PCC-1"
                admin-group "blue"
                no shutdown
            exit
            interface "int-PE-5-P-4"
                admin-group "blue"
                no shutdown
            exit
       exit
```

The admin group is also configured on P-4 and on PCC-2, so that there is a continuous path between PCC-1 and PCC-2 comprising interfaces that are included in the admin group. The **value** argument within the admin group configuration must be configured with the same value on each router; in this case, 10.

The presence of the admin group on each interface is advertised by IS-IS, so that each router is aware of all interfaces in the admin group. The PCE is also aware of the admin groups via its topology database.

In this example, a primary and secondary path will be used, so it is necessary to configure two separate MPLS path statements, as in the following output (primary and secondary paths of the same LSP cannot use the same MPLS path).

```
*A:PCC-1>config>router>mpls# info
----------------------------------------------
            path "pce-secondary"
                no shutdown
            exit
            path "pce-controlled"
                no shutdown
            exit
```

These path statements are applied within the configuration of the LSP and the admin group constraints are applied to each path, as in the following output.

```
            lsp "PCC-1-PCC-2-RSVP-PCE-ag-001"
                to 192.0.2.2
                cspf
                pce-computation
                pce-report enable
                pce-control
                primary "pce-controlled"
                    include "blue"
                exit
                secondary "pce-secondary"
                    standby
                    exclude "blue"
                exit
                no shutdown
```

The primary path must follow the path where the interfaces are included in the admin group "blue", whereas the secondary must not use any interface in the admin group; therefore, the **exclude "blue"** command within the **secondary** context. The secondary is configured as standby, so will be connected.

Debug: PCEP Requests for Primary and Secondary Paths

When the LSP is placed into a "no shutdown" state, the router initiates a separate PCEP Request for each of the primary and secondary paths, as shown in the following debug output.

```
#Primary Path Request
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REQUEST ][023 00:15:04.160]
  Svec :{numOfReq 1}{nodeDiverse F linkDiverse F srlgDiverse F}
    Request: {id 13284 PST(SegRt) 0 srcAddr 192.0.2.1 destAddr 192.0.2.2}
             {PLspId 13317 tunnelId:4 lspId: 49158
                     lspName PCC-1-PCC-2-rsvp-pce-ag -001::pce-controlled}
             {{bw 0 (0) isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 1024 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
"

#Secondary Path Request
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REQUEST ][023 00:15:04.160]
  Svec :{numOfReq 1}{nodeDiverse F linkDiverse F srlgDiverse F}
    Request: {id 13285 PST(SegRt) 0 srcAddr 192.0.2.1 destAddr 192.0.2.2}
             {PLspId 13318 tunnelId:4 lspId: 49154
                     lspName PCC-1-PCC-2-rsvp-pce-ag-001::pce-secondary}
             {{bw 0 (0) isOpt: F}}
             {{setup 7 hold 0 exclAny 1024 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
"
```

The PCEP message exchange can be identified by the request ID, and the textual LSP Name and path name are visible. The request message also shows that the admin group is signaled via the "inclAny 1024" and "exclAny 1024". The argument of 1024 corresponds to the original **value** of 10, where the admin group argument signaled = 2value = (210) = 1024, as configured within the **configure router if-attribute** context.

The following debug output corresponds to the PCReply to the PCReq for the primary path - they have the same request ID of 13284. A valid path has been computed and a list of strict hops (*isLoose* = *F*) is returned.

```
22 2017/03/16 14:44:16.29 UTC MINOR: DEBUG #2001 Base PCC
"PCC: [RX-Msg: REPLY ][023 00:15:04.260]
[Peer 192.0.2.10]
  Request: {id 13284} {} Response has calculated path
        {{Total Paths: 1}}
          Path: {PST(SegRt) 0}
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.45.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.24.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2100 hop 3}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 1024 inclAll 0}}}
```

```
"
```

When the PCEP Reply is received with a valid path, PCC-1 originates an RSVP Path message, as in the following debug output.

```
23 2017/03/16 14:44:16.29 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.2
          TTL:255, Checksum:0x6575, Flags:0x1
MSG ID     - Flags:0x1, Epoch:2387139, MsgId:8389
Session    - EndPt:192.0.2.2, TunnId:4, ExtTunnId:192.0.2.1
SessAttr   - Name:PCC-1-PCC-2-RSVP-PCE-ag-001::pce-controlled
             SetupPri:7, HoldPri:0, Flags:0x6
RSVPHop    - Ctype:1, Addr:192.168.15.1, LIH:2
TimeValue  - RefreshPeriod:180
SendTempl  - Sender:192.0.2.1, LspId:49158
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8686
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.15.1, Flags:0x0
ERO        - IPv4Prefix 192.168.15.2/32, Strict
             IPv4Prefix 192.168.45.1/32, Strict
             IPv4Prefix 192.168.24.1/32, Strict
"
```

An RSVP Resv message is received, as follows:

```
24 2017/03/16 14:44:16.30 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESV Msg
Recv RESV From:192.168.15.2, To:192.168.15.1
          TTL:255, Checksum:0xcda7, Flags:0x1
MSG ID     - Flags:0x1, Epoch:16280838, MsgId:185
Session    - EndPt:192.0.2.2, TunnId:4, ExtTunnId:192.0.2.1
FlowSpec   - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8686, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.1, LspId:49158, Label:262107
LspAttr    - Attribute Flags TLV  0x400000
RRO        - InterfaceIp:192.168.15.2, Flags:0x0
             Label:262107, Flags:0x1
             InterfaceIp:192.168.45.1, Flags:0x0
             Label:262090, Flags:0x1
             InterfaceIp:192.168.24.1, Flags:0x0
             Label:262136, Flags:0x1
"
```

PCC-1 now originates a PCEP Report and forwards it to the PCE, reporting the state of the LSP.

```
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REPORT ][023 00:15:04.270]
[Peer 192.0.2.10]
  Report : {srpId:0 PST(SegRt):0 PLspId:13317 lspId: 49158 tunnelId:4}
    {Sync 0 Rem 0 AdminState 1 OperState 2 Delegate 1 Create 0}
    {srcAddr 192.0.2.1 destAddr 192.0.2.2 extTunnelId :: pathName PCC-1-PCC-2-
rsvp-pce-ag-001::pce-controlled}
    {Binding Type: 0 Binding Val : 0}
    Lsp Constraints:
             {{bw 0 isOpt: F}}
             {{setup 7 hold 0 exclAny 0 inclAny 1024 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
      Ero Path:
            {{ctype IPv4 addr 192.168.15.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.45.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.24.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2100 hop 4}}}
            {{      {setup 7 hold 0 exclAny 0 inclAny 1024 inclAll 0}}}
      RRO:
            {{type IPv4 addr 192.168.15.1/32 Flag 0}}
            {{type IPv4 addr 192.168.15.2/32 Flag 0}}
            {{type Label label c-type 1 label 262107}}
            {{type IPv4 addr 192.168.45.1/32 Flag 0}}
            {{type Label label c-type 1 label 262090}}
            {{type IPv4 addr 192.168.24.1/32 Flag 0}}
            {{type Label label c-type 1 label 262136}}
      {Lsp Err NA RsvpErr 0 LspDbVersion 0}
"
```

A valid path for the secondary LSP path is shown in the second PCEP Reply with a request ID of 13285. The message also contains a non-zero value (set to 1024) for the **exclAny** parameter, so that the computation excludes the "blue" admin group.

The list of hops is shown in the following output.

```
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [RX-Msg: REPLY ][023 00:15:04.290]
[Peer 192.0.2.10]
  Request: {id 13285} {} Response has calculated path
        {{Total Paths: 1}}
          Path: {PST(SegRt) 0}
            {{ctype IPv4 addr 192.168.16.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.36.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.23.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2100 hop 3}}}
            {{      {setup 7 hold 0 exclAny 1024 inclAny 0 inclAll 0}}}
"
```

When the PCEP Reply is received, the list of hops is used within the ERO of the RSVP Path message.

```
UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.1, To:192.0.2.2
          TTL:255, Checksum:0xd66a, Flags:0x1
MSG ID     - Flags:0x1, Epoch:2387139, MsgId:8390
Session    - EndPt:192.0.2.2, TunnId:4, ExtTunnId:192.0.2.1
SessAttr   - Name:PCC-1-PCC-2-RSVP-PCE-ag-001::pce-secondary
             SetupPri:7, HoldPri:0, Flags:0x6
RSVPHop    - Ctype:1, Addr:192.168.16.1, LIH:3
TimeValue  - RefreshPeriod:180
SendTempl  - Sender:192.0.2.1, LspId:49154
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
```

```
             MPU:20, MTU:8690
LabelReq   - IfType:General, L3ProtID:2048
RRO        - IpAddr:192.168.16.1, Flags:0x0
ERO        - IPv4Prefix 192.168.16.2/32, Strict
             IPv4Prefix 192.168.36.1/32, Strict
             IPv4Prefix 192.168.23.1/32, Strict
"
```

An RSVP Resv message is received containing the RRO with the label allocations for each hop.

```
29 2017/03/16 14:44:16.32 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESV Msg
Recv RESV From:192.168.16.2, To:192.168.16.1
          TTL:255, Checksum:0xdf1f, Flags:0x1
MSG ID     - Flags:0x1, Epoch:5527111, MsgId:200
Session    - EndPt:192.0.2.2, TunnId:4, ExtTunnId:192.0.2.1
FlowSpec   - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:8690, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.1, LspId:49154, Label:262112
LspAttr    - Attribute Flags TLV  0x400000
RRO        - InterfaceIp:192.168.16.2, Flags:0x0
             Label:262112, Flags:0x1
             InterfaceIp:192.168.36.1, Flags:0x0
             Label:262068, Flags:0x1
             InterfaceIp:192.168.23.1, Flags:0x0
             Label:262111, Flags:0x1
"
```

When PCC-1 receives an RSVP Resv message in response to the Path message, it will send a PCEP Report to the PCE to report the state of the secondary LSP path.

```
UTC MINOR: DEBUG #2001 Base PCC
"PCC: [TX-Msg: REPORT ][023 00:15:04.300]
[Peer 192.0.2.10]
  Report : {srpId:0 PST(SegRt):0 PLspId:13318 lspId: 49154 tunnelId:4}
    {Sync 0 Rem 0 AdminState 1 OperState 1 Delegate 1 Create 0}
    {srcAddr 192.0.2.1 destAddr 192.0.2.2 extTunnelId :: pathName PCC-1-PCC-2-
rsvp-pce-ag-001::pce-secondary}
    {Binding Type: 0 Binding Val : 0}
    Lsp Constraints:
             {{bw 0 isOpt: F}}
             {{setup 7 hold 0 exclAny 1024 inclAny 0 inclAll 0 isOpt: F}}
             {{igp-met 16777215 B:F BVal:0 C:T isOpt: F}}
             {{hop-cnt 0 B:T BVal:255 C:T isOpt: F}}
      Ero Path:
            {{ctype IPv4 addr 192.168.16.2/32 isLoose F}}
            {{ctype IPv4 addr 192.168.36.1/32 isLoose F}}
            {{ctype IPv4 addr 192.168.23.1/32 isLoose F}}
            {{Attr: {bw 0 te-metric 0 igp 2100 hop 4}}}
            {{      {setup 7 hold 0 exclAny 1024 inclAny 0 inclAll 0}}}
      RRO:
            {{type IPv4 addr 192.168.16.1/32 Flag 0}}
            {{type IPv4 addr 192.168.16.2/32 Flag 0}}
            {{type Label label c-type 1 label 262112}}
            {{type IPv4 addr 192.168.36.1/32 Flag 0}}
            {{type Label label c-type 1 label 262068}}
            {{type IPv4 addr 192.168.23.1/32 Flag 0}}
            {{type Label label c-type 1 label 262111}}
      {Lsp Err NA RsvpErr 0 LspDbVersion 0}
```

Verification

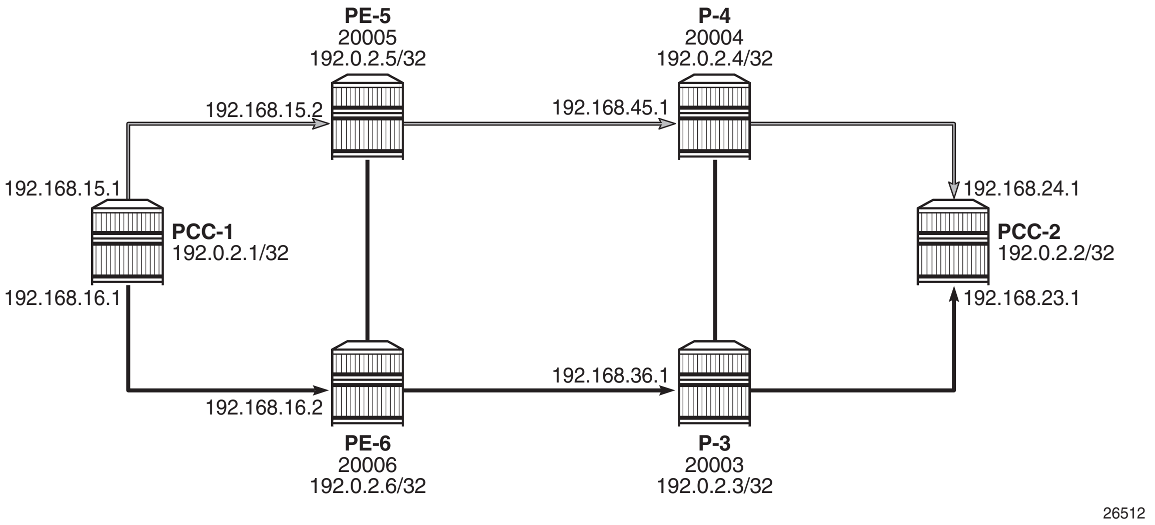
The following show command output shows the state of the primary path after connection.

```
A:PCC-1# show router mpls lsp "PCC-1-PCC-2-RSVP-PCE-ag-001" path "pce-
controlled" detail
===============================================================================
MPLS LSP PCC-1-PCC-2-RSVP-PCE-ag-001 Path pce-controlled (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP PCC-1-PCC-2-RSVP-PCE-ag-001 Path pce-controlled
-------------------------------------------------------------------------------
LSP Name         : PCC-1-PCC-2-RSVP-PCE-ag-001
Path LSP ID      : 49158
From             : 192.0.2.1          To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : pce-controlled       Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1:220            Out Label            : 262107
Path Up Time     : 0d 00:36:46          Path Down Time       : 0d 00:00:00
Retry Limit      : 0                    Retry Timer          : 30 sec
Retry Attempt    : 0                    Next Retry In        : 0 sec
BFD Template     : None                 BFD Ping Interval    : 60
BFD Enable       : False

Adspec           : Disabled             Oper Adspec          : Disabled
CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
FRR              : Disabled             Oper FRR             : Disabled
Propogate Adm Grp: Disabled             Oper Prop Adm Grp    : Disabled
Inter-area       : False

PCE Updt ID      : 0

PCE Report       : Enabled              Oper PCE Report      : Enabled
PCE Control      : Enabled              Oper PCE Control     : Enabled
PCE Compute      : Enabled              Oper PCE Compute     : Enabled

Neg MTU          : 8686                 Oper MTU             : 8686
Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Record Route     : Record               Oper Record Route    : Record
Record Label     : Record               Oper Record Label    : Record
```

```
Setup Priority   : 7                    Oper Setup Priority  : 7
Hold Priority    : 0                    Oper Hold Priority   : 0
Class Type       : 0                    Oper CT              : 0
Backup CT        : None
MainCT Retry     : n/a
    Rem          :
MainCT Retry     : 0
    Limit        :
Include Groups   :                      Oper Include Groups  :
blue                                           blue
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None

Adaptive         : Enabled              Oper Metric          : 2100
Preference       : n/a
Path Trans       : 5                    CSPF Queries         : 0
Failure Code     : noError
Failure Node     : n/a
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.15.1 (192.0.2.1)                    Record Label        : N/A
 -> 192.168.15.2 (192.0.2.5)                    Record Label        : 262107
 -> 192.168.45.1 (192.0.2.4)                    Record Label        : 262090
 -> 192.168.24.1 (192.0.2.2)                    Record Label        : 262136
Computed Hops    :
    192.168.15.2(S)
 -> 192.168.45.1(S)
 -> 192.168.24.1(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 2100
```

Comparing the Actual Hop with the addresses in Figure 282, the path signaled uses the router interfaces configured with the admin group "blue".

Figure 282 **Diverse Primary and Secondary Paths**



The following output shows that the Actual Hops for the secondary path excludes the admin group "blue", as in Figure 282. The output also shows that the "exclude groups" and "oper exclude groups" are marked with the admin group "blue".

```
A:PCC-1# show router mpls lsp "PCC-1-PCC-2-RSVP-PCE-ag-001" path "pce-
secondary" detail

===============================================================================
MPLS LSP PCC-1-PCC-2-RSVP-PCE-ag-001 Path pce-secondary (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP PCC-1-PCC-2-RSVP-PCE-ag-001 Path pce-secondary
-------------------------------------------------------------------------------
LSP Name         : PCC-1-PCC-2-RSVP-PCE-ag-001
Path LSP ID      : 49154
From             : 192.0.2.1          To                    : 192.0.2.2
Admin State      : Up                   Oper State            : Up
Path Name        : pce-secondary        Path Type             : Standby
Path Admin       : Up                   Path Oper             : Up
Out Interface    : 1/1/2                Out Label             : 262112
Path Up Time     : 0d 00:44:02          Path Down Time        : 0d 00:00:00
Retry Limit      : 0                    Retry Timer           : 30 sec
Retry Attempt    : 0                    Next Retry In         : 0 sec
```

```
BFD Template     : None                 BFD Ping Interval    : 60
BFD Enable       : False

Adspec           : Disabled             Oper Adspec          : Disabled
CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
Propogate Adm Grp: Disabled             Oper Prop Adm Grp    : Disabled
Inter-area       : False

PCE Updt ID      : 0

PCE Report       : Enabled              Oper PCE Report      : Enabled
PCE Control      : Enabled              Oper PCE Control     : Enabled
PCE Compute      : Enabled              Oper PCE Compute     : Enabled

Neg MTU          : 8690                 Oper MTU             : 8690
Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Record Route     : Record               Oper Record Route    : Record
Record Label     : Record               Oper Record Label    : Record
Setup Priority   : 7                    Oper Setup Priority  : 7
Hold Priority    : 0                    Oper Hold Priority   : 0
Class Type       : 0                    Oper CT              : 0
Include Groups   :                      Oper Include Groups  :
None                                           None
Exclude Groups   :                      Oper Exclude Groups  :
blue                                           blue

Adaptive         : Enabled              Oper Metric          : 2100
Preference       : 255
Path Trans       : 1                    CSPF Queries         : 0
Failure Code     : noError
Failure Node     : n/a
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.16.1 (192.0.2.1)                    Record Label        : N/A
 -> 192.168.16.2 (192.0.2.6)                    Record Label        : 262112
 -> 192.168.36.1 (192.0.2.3)                    Record Label        : 262068
 -> 192.168.23.1 (192.0.2.2)                    Record Label        : 262111
Computed Hops    :
    192.168.16.2(S)
 -> 192.168.36.1(S)
 -> 192.168.23.1(S)
Srlg             : Disabled
Srlg Disjoint    : False
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 2100
===============================================================================
```

Conclusion

PCEP support for RSVP-TE LSPs extends the use of MPLS labels into traffic engineering applications   This example provides the configuration for PCE controlled and computed RSVP-TE LSPs together with the associated commands and outputs that can be used for verifying and troubleshooting.

RSVP Point-to-Point LSPs

This chapter provides information about point-to-point label switched paths (LSPs) established using resource reservation protocol (RSVP) with or without traffic engineering (TE).

Topics in this chapter include:

Applicability

This chapter was initially written for SR OS release 7.0.R5, but the CLI in the current edition is based on SR OS release 16.0.R3. There are no prerequisites or conditions on the hardware for this configuration.

Overview

Due to the connectionless nature of the network layer protocol IP, packets travel through the network on a hop-by-hop basis with routing decisions made at each node. As a result, hyperaggregation of data on certain links may occur and it may impact the provider's ability to provide guaranteed service levels across the network end-to-end. To address these shortcomings, Multi-Protocol Label Switching (MPLS) was developed. The technology provides the capability to establish connection-oriented paths, called Label Switched Paths (LSPs), over a connectionless (IP) network. The LSP offers a mechanism to engineer network traffic independently from the underlying network routing protocol (mostly IP) to improve the network resiliency and recovery options and to permit delivery of new services that are not readily supported by conventional IP routing techniques, such as Layer 2 IP Virtual Private Networks (VPNs). These benefits are essential for today's communication network explaining the wide deployment base of the MPLS technology.

RFC 3031, *Multiprotocol Label Switching Architecture*, specifies the MPLS architecture whereas this document describes the configuration and troubleshooting of RSVP point-to-point LSPs on SR OS. Besides RSVP P2P LSPs, there are also Static Point-to-Point LSPs, LDP Point-to-Point LSPs, and Segment Routing (SR) LSPs (SR-ISIS, SR-OSPF, and SR-TE). For SR-ISIS, see chapter Segment Routing with IS-IS Control Plane.

Packet Forwarding

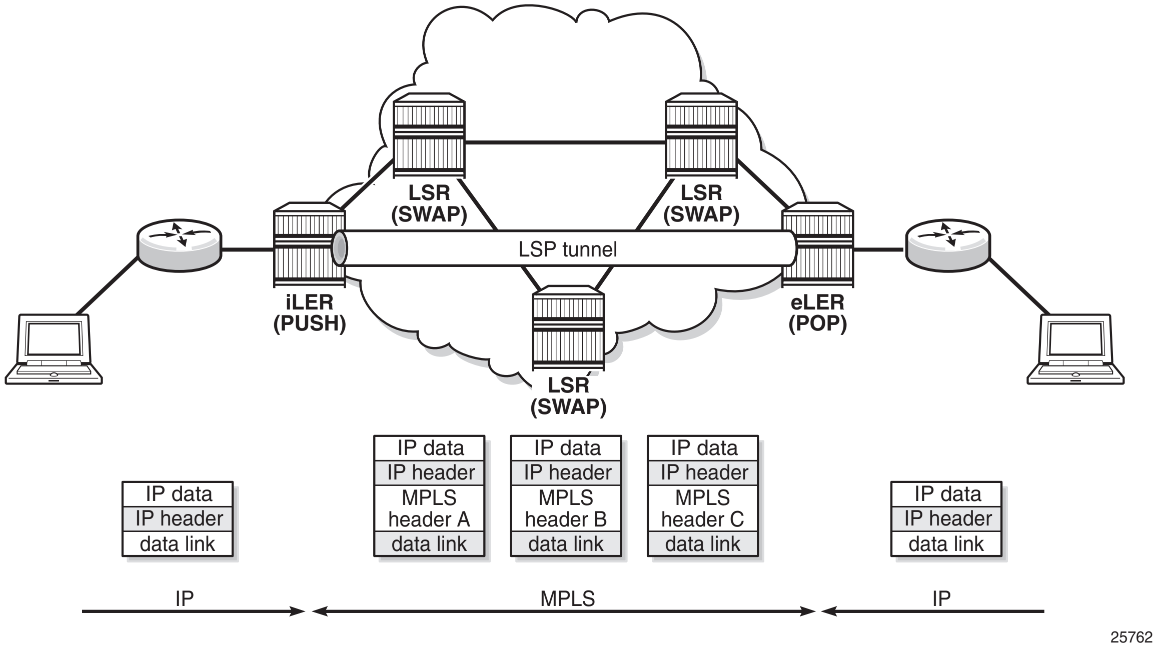
As a packet of a connectionless network layer protocol travels from one router to the next, each router in the network makes an independent forwarding decision by performing the following basic tasks: first analyzing the packet header, then referencing the local routing table to find the longest match based on the destination address in the IP header, and finally sending out the packet on the selected interface. In other words, the first function partitions the entire set of possible packets into a set of Forwarding Equivalence Classes (FECs). All packets associated to a particular FEC will be forwarded along the same logical path to the same destination. The second function maps each FEC to a next hop destination router. Each router along the data path performs these actions.

On the other hand, in MPLS, the assignment of a packet to a particular FEC is done just once, as the packet enters the network. In turn, the FEC is mapped to an LSP, which is pre-signaled prior to any data flowing. An MPLS label, representing the FEC to which the packet is assigned, is attached to the packet (push operation) and once labeled, the packet is forwarded to the next hop router along that LSP path. At subsequent hops, there is no further analysis of the packet network layer header. Instead, the label is used as an index into a table which specifies the next hop and a new label. The old label is replaced with the new label (swap operation), and the packet is forwarded to its next hop. At the MPLS network egress, the label is removed from the packet (pop operation). If this router is the destination (based on the remaining packet), the packet is handed to the receiving application (such as a Virtual Private LAN Service (VPLS) domain). If this router is not the final destination of the packet, the packet will be sent into a new MPLS tunnel or forwarded by conventional IP forwarding toward the Layer 3 destination.

Terminology

Figure 283 shows a general network topology clarifying the MPLS-related terms.

Figure 283 **Generic MPLS Network, MPLS Label Operations**



A Label Edge Router (LER) is a device at the edge of an MPLS network, with at least one interface outside the MPLS domain. A router is usually defined as an LER based on its position relative to a particular LSP. The MPLS router at the head-end of an LSP is called the ingress Label Edge Router (iLER). The MPLS router at the tail-end of an LSP is called the egress Label Edge Router (eLER). The iLER receives unlabeled packets from outside the MPLS domain, then applies MPLS labels to the packets, and forwards the labeled packets into the MPLS domain. The eLER receives labeled packets from the MPLS domain, then removes the labels, and forwards unlabeled packets outside the MPLS domain. The eLER can signal an implicit-null label (numeric value 3). This informs the previous hop to send MPLS packets without an outer label and is known as Penultimate Hop Popping (PHP).

A Label Switching Router (LSR) is a device internal to an MPLS network, with all interfaces inside the MPLS domain. These devices switch labeled packets inside the MPLS domain. In the core of the network, LSRs ignore the packet network layer (IP) header and simply forward the packet using the MPLS label swapping mechanism.

A single LSP is unidirectional. In common practice, because the bidirectional nature of most traffic flows is implied, the term LSP often is used to define the pair of LSPs that enable the bidirectional flow. For ease of terminology and discussion however, the LSP in this chapter is referred to as a single entity.

LSP Establishment

Prior to packet forwarding, the LSP must be established. In order to do so, labels need to be distributed for the path. Labels are usually distributed by a downstream router in the upstream direction (relative to the data flow). There are a number of ways used for label distribution: static, LDP, and RSVP. For static P2P LSPs, see chapter Static Point-to-Point LSPs; for LDP P2P LSPs, see chapter LDP Point-to-Point LSPs.

RSVP-TE (RFC 3209, *RSVP-TE: Extensions to RSVP for LSP Tunnels*) can be used to signal LSPs across the network. RSVP-TE is used for traffic engineering when the ingress router creates an LSP with specific constraints beyond the best route chosen by the IGP. RSVP-TE identifies the specific path desired for the LSP and may include resource requirements for the path.

The most important benefit of the label swapping mechanism RSVP-TE is its ability to map any type of user traffic to an LSP that has been specifically engineered to satisfy user traffic requirements. Customized LSPs may be created based on hop count, bandwidth requirements, administrative groups, or Shared Risk Link Groups (SRLGs). They can even be routed through a strict path with specific network links or nodes, as specified by the ingress node. This offers service providers precise control over the flow of traffic in their networks and results in a network that operates more efficiently and provides more predictable and scalable services. For information about SRLG, see chapter Shared Risk Link Groups for RSVP-Based LSP.

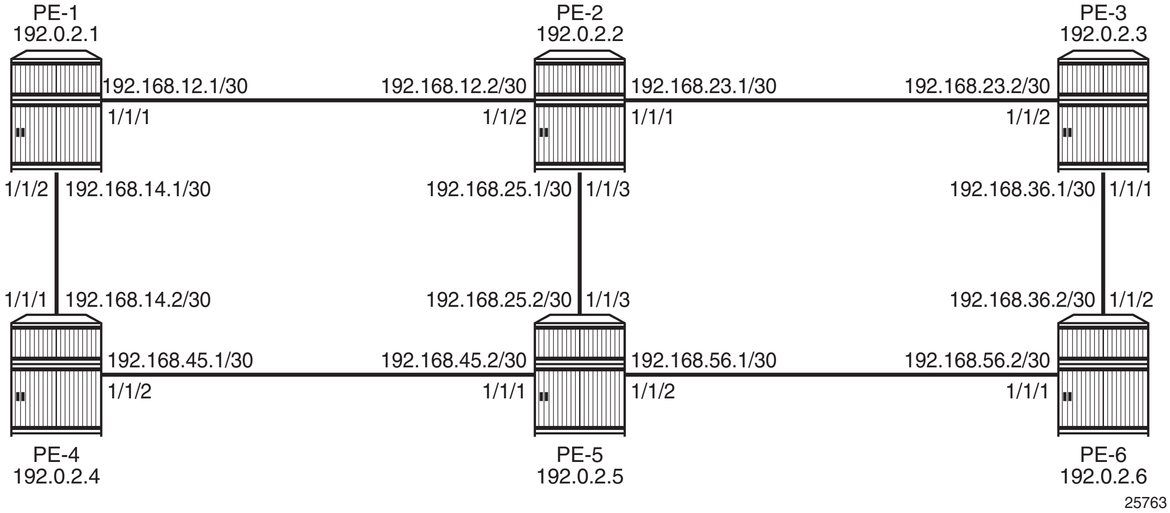
Fast reroute (FRR) allows to signal backup paths before a failure takes place. This allows traffic to flow almost continuously, without waiting for routing protocol convergence. There a two different methods for FRR for an RSVP-TE LSP: one-to-one and facility.

- FRR one-to-one defines detour tunnels toward the eLER for a particular LSP only. The advantage is that the detour tunnel is the best path to the eLER that avoids the node or link at the point of failure. The drawback is that when different LSPs would need the same detour, a dedicated RSVP-TE detour LSP needs to be signaled for each LSP.
- FRR facility defines local repair tunnels avoiding one particular node (the next hop in the data path) or one particular link (the next link in the data path), ignoring the eLER. These bypass tunnels originate in a point of local repair (PLR) and terminate in a merge point (MP) on the LSP. Bypass tunnels are shared between LSPs.

Example Topology

The example topology is displayed in Figure 284. The setup consists of six 7750 SR nodes located in a single autonomous system.

Figure 284 **MPLS Example Topology**



Configuration

In this chapter, RSVP LSPs are configured manually, but they can also be configured automatically using LSP templates; see chapter Automatic Creation of RSVP-TE LSPs.

As a general prerequisite for the configuration of MPLS LSPs, a correctly working Interior Gateway Protocol (IGP) is required. Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS-IS) can be used as IGP.

RSVP-TE, an extension of the original RSVP protocol, has two major benefits adding to the basic MPLS functionality. The first benefit is traffic engineering, which allows the ingress router to create an LSP with specific constraints beyond the best route chosen by the IGP. The second benefit is improved network resiliency when a link or node fails in the network, using FRR and secondary paths. FRR is also supported for LDP, see chapter MPLS LDP FRR using ISIS as IGP.

In this chapter, several RSVP-TE LSPs are configured:

- A simple LSP with a primary path that has strict hops and no specific TE constraints
- A simple LSP with a dynamic path without any configured hops is created. Initially, there are no constraints and the actual path is calculated based on the IGP best route.
- An LSP configured with constrained shortest path first (CSPF) that will use the TE metric, even though the IGP metric can also be used
- An LSP with fast reroute (FRR) one-to-one enabled
- An LSP with FRR facility enabled
- An LSP including an admin group "blue" and an LSP excluding admin group "red"
- An LSP with a hop limit configured

There is no configuration example with bandwidth constraints configured in this chapter. See chapter Automatic Bandwidth Adjustment in P2P LSPs for a configuration with bandwidth constraint with or without automatic adjustment.

Initially, no traffic engineering is enabled in the ISIS context, but it will be enabled when required. For RSVP LSPs, the MPLS instance needs to be enabled on each router and all network interfaces facing the MPLS domain. By default, the system interface is put automatically within the MPLS context. When adding interfaces to the MPLS instance, they are automatically added to the RSVP instance as well, but the instance itself is still in an administrative shutdown state. The next step is to enable the RSVP instance on all routers in the MPLS network. As a result, all interfaces facing the MPLS domain as well as the system interface are added to the MPLS and RSVP instance and both instances are in a no shutdown state. For PE-1, the following configuration is required:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-4"
            exit
            no shutdown
        exit
```

```
*A:PE-1# configure router rsvp no shutdown
```

Strict or Loose Path

On the iLER, first the definition of a path is required. A path is a sequence of MPLS routers (hops) through which the LSP using that path has to pass. It is not uniquely bound to a particular LSP; it can be used by any LSP originating in that node. A hop in a path can be strict or loose: strict or loose meaning that the LSP must take either a direct path from the previous hop router to this router (strict) or can traverse through other routers (loose). The hops not explicitly defined in the loose path definition are created by calculating the IGP shortest path. A third possibility is an empty path implying not a single node is required to be present in the LSP path and the shortest path from the IGP is used to define the LSP path. Other techniques, such as the use of admin groups or shared risk link groups, can also be used to influence the decision which hops to include in the path. Three paths will be configured, respectively:

1. Only strict hops
2. Mixed strict and loose hops
3. Empty path

To find a valid path, the last hop in the path sequence needs to be the system IP or an interface address of the terminating router (eLER). The IP addresses in the hop command can be the system IP addresses or the interface addresses of the node. However, it is recommended to use the system IP addresses with keyword **loose** as this allows more flexibility when finding new paths in failover scenarios (because the upstream node could use any of multiple paths to the system address, whereas specifying the interface address would restrict the upstream node to a single entry-point). The recommendation when using the keyword strict in the hop command context, is to use the physical link addresses. However, the last hop in the path should be a system address to make it appear in the list on the 5620 SAM (service-aware manager).

```
*A:PE-1# configure
    router
        mpls
            path "path-PE-1-PE-6-strict"
                hop 10 192.168.12.2 strict
                hop 20 192.168.25.2 strict
                hop 30 192.168.56.2 strict
                no shutdown
            exit
            path "path-PE-1-PE-6-semiLoose"
                hop 10 192.0.2.5 loose
                hop 20 192.168.56.2 strict
                no shutdown
            exit
            path "dyn"
                no shutdown
            exit
```

The paths can be checked with the **show router mpls path** command.

```
*A:PE-1# show router mpls path
===============================================================================
MPLS Path:
===============================================================================
Path Name                        Adm  Hop Index   IP Address       Strict/Loose
-------------------------------------------------------------------------------
path-PE-1-PE-6-strict            Up   10          192.168.12.2     Strict
                                      20          192.168.25.2     Strict
                                      30          192.168.56.2     Strict

path-PE-1-PE-6-semiLoose         Up   10          192.0.2.5        Loose
                                      20          192.168.56.2     Strict

dyn                              Up   no hops     n/a              n/a

-------------------------------------------------------------------------------
Total Paths : 3
===============================================================================
*A:PE-1#
```

Simple RSVP LSP with Strict Primary Path

The configuration of a simple LSP using RSVP signaling contains at least on the iLER:

- System IP address of the terminating node (to)
- Path to the eLER (primary)
- Administratively enabled (no shutdown)

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-6"
                to 192.0.2.6
                primary "path-PE-1-PE-6-strict"
                exit
                secondary "dyn"
                exit
                no shutdown
            exit
```

All the hops in the strict path are already defined and there is no need to look up the IGP best route. The configuration of secondary paths is optional. In case the primary path fails, the secondary path can be signaled to take over the traffic. It can even be signaled as standby while the primary path is operational for a faster switchover when the keyword **standby** is added, which is not the case here. The secondary path has no hops defined. The hops will be calculated based on the IGP best route. The nodes through which the LSP will pass (LSRs and eLER) require no additional configuration: enabling MPLS and RSVP on their interfaces suffices.

An overview of all LSPs configured on a particular node is given by the **show router mpls lsp** command. More details about a particular LSP can be retrieved by adding the keyword **detail** to the previous command.

```
*A:PE-1# show router mpls lsp

===============================================================================
MPLS LSPs (Originating)
===============================================================================
LSP Name                           To               Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-PE-1-PE-6                      192.0.2.6        1       No        Up   Up
-------------------------------------------------------------------------------
LSPs : 1
===============================================================================
*A:PE-1#


*A:PE-1# show router mpls lsp "LSP-PE-1-PE-6" detail

===============================================================================
MPLS LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name    : LSP-PE-1-PE-6
LSP Type    : RegularLsp                       LSP Tunnel ID  : 1
LSP Index   : 1                                TTM Tunnel Id  : 1
From        : 192.0.2.1                        To             : 192.0.2.6
Adm State   : Up                               Oper State     : Up
LSP Up Time : 0d 00:01:27                      LSP Down Time  : 0d 00:00:00
Transitions : 1                                Path Changes   : 1
Retry Limit : 0                                Retry Timer    : 30 sec
Signaling   : RSVP                             Resv. Style    : SE
Hop Limit   : 255                              Negotiated MTU : 1564
Adaptive    : Enabled                          ClassType      : 0
FastReroute : Disabled                         Oper FR        : Disabled
CSPF        : Disabled                         ADSPEC         : Disabled
Metric      : N/A
Load Bal Wt : N/A                              ClassForwarding: Disabled
Include Grps:                                  Exclude Grps   :
None                                              None
```

```
Least Fill  : Disabled
BFD Template: None                             BFD Ping Intvl : 60
BFD Enable  : False                            BFD Failure-action   : None

Revert Timer    : Disabled                  Next Revert In       : N/A
Entropy Label   : Enabled+                  Oper Entropy Label   : Enabled
Negotiated EL   : Disabled
Auto BW         : Disabled
LdpOverRsvp     : Enabled
VprnAutoBind    : Enabled
IGP Shortcut    : Enabled                   BGP Shortcut         : Enabled
IGP LFA         : Disabled                  IGP Rel Metric       : Disabled
BGPTransTun     : Enabled
Oper Metric     : 16777215
Prop Adm Grp    : Disabled
PCE Report      : Disabled+
PCE Compute     : Disabled                  PCE Control          : Disabled
Path Profile    : None
Admin Tags      : None

Primary(a)  : path-PE-1-PE-6-strict            Up Time        : 0d 00:01:27

Bandwidth   : 0 Mbps
Secondary   : dyn                              Down Time      : 0d 00:01:27

Bandwidth   : 0 Mbps
===============================================================================
*A:PE-1#
```

In each hop (originating, transit and terminate), the RSVP sessions can be verified as follows:

```
*A:PE-1# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
RSVP Session Name
    From                To              Tunnel ID   LSP ID       State
-------------------------------------------------------------------------------
LSP-PE-1-PE-6::path-PE-1-PE-6-strict
192.0.2.1           192.0.2.6           1           3584         Up

-------------------------------------------------------------------------------
Sessions : 1
===============================================================================
*A:PE-1#
```

The detailed output of this command includes among others the session type (here: originate), the incoming and outgoing labels, the previous and next hop, and - for originating LSPs - also the list of hops):

```
*A:PE-1# show router rsvp session detail

===============================================================================
RSVP Sessions (Detailed)
```

```
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-PE-1-PE-6::path-PE-1-PE-6-strict
-------------------------------------------------------------------------------
From            : 192.0.2.1             To             : 192.0.2.6
Tunnel ID       : 1                     LSP ID         : 3584
Style           : SE                    State          : Up
Session Type    : Originate
In Interface    : n/a                   Out Interface  : 1/1/1
In IF Name      : n/a
Out IF Name     : int-PE-1-PE-2
In Label        : n/a                   Out Label      : 524287
Previous Hop    : n/a                   Next Hop       : 192.168.12.2
Hops            :
    192.168.12.2(S)    -> 192.168.25.2(S)    -> 192.168.56.2(S)
SetupPriority   : 7                     Hold Priority  : 0
Class Type      : 0
SubGrpOrig ID   : 0                     SubGrpOrig Addr:
P2MP ID         : 0
FrrAvailType    : N/A
FrrSrlgStrict   : N/A                   SrlgDisjoint   : N/A

Path Recd       : 0                     Path Sent      : 1
Resv Recd       : 2                     Resv Sent      : 0
Summary msgs    :
SPath Recd      : 0                     SPath Sent     : 0
SResv Recd      : 0                     SResv Sent     : 0
LSP Attr Flags  : N/A
===============================================================================
*A:PE-1#
```

The following RSVP LSP is in the tunnel table on PE-1:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.6/32      rsvp      MPLS  1         7        192.168.12.2   16777215
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

In order to signal PHP with RSVP, implicit-null must be configured on the eLER (RSVP must be shut down to perform this command).

```
*A:PE-6# configure
    router
        rsvp
            shutdown
            implicit-null-label
            no shutdown
            exit
```

The implicit-null is signaled after re-enabling RSVP and is shown on PE-5 as an egress label of 3. This label is not actually sent toward PE-6.

```
*A:PE-5# show router rsvp session detail

===============================================================================
RSVP Sessions (Detailed)
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-PE-1-PE-6::path-PE-1-PE-6-strict
-------------------------------------------------------------------------------
From            : 192.0.2.1             To            : 192.0.2.6
Tunnel ID       : 1                     LSP ID        : 3588
Style           : SE                    State         : Up
Session Type    : Transit
In Interface    : 1/1/3                 Out Interface : 1/1/2
In IF Name      : int-PE-5-PE-2
Out IF Name     : int-PE-5-PE-6
In Label        : 524287                Out Label     : 3
Previous Hop    : 192.168.25.1          Next Hop      : 192.168.56.2
---snip---
```

The use of implicit-null can also be enabled/disabled on a per interface basis (either RSVP, or the interface within RSVP, must be shut down to perform this change).

```
A:PE-6>config>router>rsvp# interface "int-PE-6-PE-5"
A:PE-6>config>router>rsvp>if# implicit-null-label
  - implicit-null-label {<enable|disable>}
  - no implicit-null-label
 <<enable|disable>>   : keyword
```

In the remainder of the chapter, LSPs with empty paths will be used. LSP "LSP-PE-1-PE-6" is shut down.

Simple RSVP LSP with Dynamic Path

In this section, an LSP is configured from PE-1 to PE-3 with a dynamic path that is empty. There is no secondary path. LSP "LSP-PE-1-PE-3" is configured as follows:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3"
                to 192.0.2.3
                primary "dyn"
                exit
                no shutdown
            exit
```

Interfaces with a lower metric will be preferred over links with a high metric. The default IGP metric in this example is 10. The metric is lower for higher speed links, but can be configured manually; as follows:

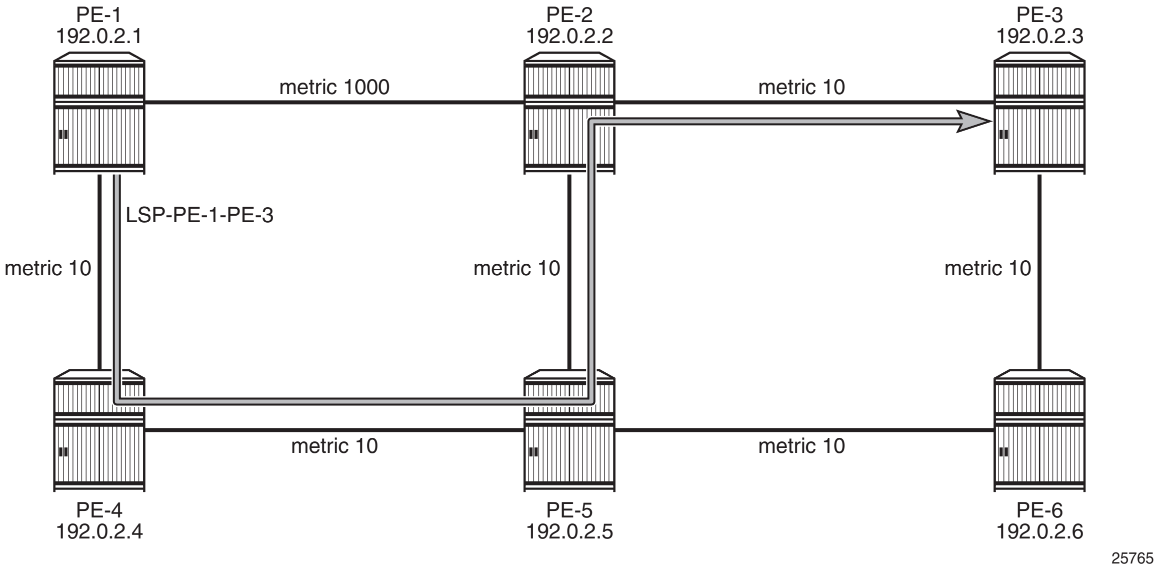
```
*A:PE-1# configure router isis interface "int-PE-1-PE-2" level 1 metric 1000
```

The link between PE-1 and PE-2 has a higher metric and will not be selected for forwarding traffic because the detour via PE-4 has a lower metric. The routing table shows that the route to prefix 192.0.2.3 has PE-4 as next hop instead of PE-2 and that the metric is 40:

```
*A:PE-1# show router route-table 192.0.2.3

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.3/32                                  Remote  ISIS        00h14m02s  15
       192.168.14.2                                                 40
-------------------------------------------------------------------------------
No. of Routes: 1
```

Figure 285 shows the path used by the LSP:

Figure 285 LSP with Dynamic Path Takes IGP Best Route



The actual hops can be verified in the following output. The path is dynamic, therefore, no explicit hops are configured.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
From             : 192.0.2.1           To                   : 192.0.2.3
Admin State      : Up                  Oper State           : Up
Path Name        : dyn
Path LSP ID      : 1024                Path Type            : Primary
Path Admin       : Up                  Path Oper            : Up
Out Interface    : 1/1/2               Out Label            : 524287

---snip---

Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.14.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)                     Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)                     Record Label        : 524287
 -> 192.168.25.1 (192.0.2.2)                     Record Label        : 524287
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524287
---snip---
```

The tunnel table shows the RSVP LSP with PE-4 as the next hop and a metric of 40:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.3/32      rsvp      MPLS  2         7        192.168.14.2   40
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

Another path with metric 40 is as follows: PE-1, PE-4, PE-5, PE-6, PE-3. This can equally well be selected.

RSVP-TE LSP with Dynamic Path

Traffic engineering is enabled in the ISIS context on all nodes; as follows:

```
*A:PE-1# configure router isis traffic-engineering
```

The LSP can be configured with constrained shortest path first (CSPF); as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" cspf
```

For this LSP, it will not make any difference. By default, the IGP metrics are used and the LSP path takes the IGP shortest path.

Besides IGP metrics, also TE metrics can be configured; as follows:

```
*A:PE-2# configure
    router
        mpls
            interface "int-PE-2-PE-1"
                te-metric 10
            exit
            interface "int-PE-2-PE-3"
                te-metric 500
            exit
            interface "int-PE-2-PE-5"
                te-metric 10
            exit
```

In this example, all interfaces on all PEs get a TE metric of 10, except for the interfaces between PE-2 and PE-3, which get a TE metric of 500. Even with these TE metrics configured, the LSP path will not change, because the IGP metric is used by default, as can be verified as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail
---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn
Path LSP ID      : 1026                 Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/2                Out Label            : 524286

---snip---
CSPF             : Enabled              Oper CSPF            : Enabled
---snip---

Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.14.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)                     Record Label        : 524286
 -> 192.168.45.2 (192.0.2.5)                     Record Label        : 524286
 -> 192.168.25.1 (192.0.2.2)                     Record Label        : 524287
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524286
---snip---
Last Resignal    : n/a                  CSPF Metric         : 40
---snip---
```

The RSVP LSP in the tunnel table has next hop PE-4 and a metric of 40:

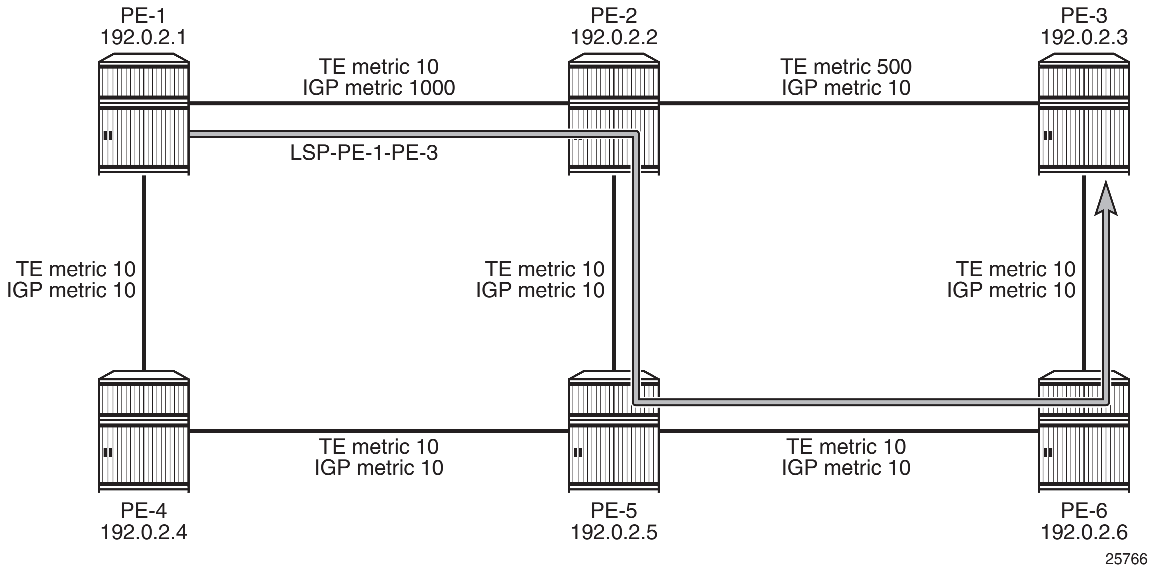
```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.3/32      rsvp      MPLS  2         7        192.168.14.2   40
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

To force the LSP to use the TE metric, the LSP is reconfigured as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" cspf use-te-metric
```

The LSP path is shown in Figure 286:

Figure 286 **RSVP-TE LSP with Dynamic Path Using TE Metric**



The LSP path goes from PE-1 to PE-2 and via PE-5 and PE-6 to PE-3, as can be seen in the following output. The CSPF metric is 40, which corresponds to the TE metric in this case:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail
---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
From             : 192.0.2.1             To                   : 192.0.2.3
Admin State      : Up                    Oper State           : Up
Path Name        : dyn
Path LSP ID      : 1028                  Path Type            : Primary
Path Admin       : Up                    Path Oper            : Up
Out Interface    : 1/1/1                 Out Label            : 524287
---snip---

CSPF             : Enabled               Oper CSPF            : Enabled
---snip---

Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1)                      Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                      Record Label        : 524287
 -> 192.168.25.2 (192.0.2.5)                      Record Label        : 524287
 -> 192.168.56.2 (192.0.2.6)                      Record Label        : 524286
 -> 192.168.36.1 (192.0.2.3)                      Record Label        : 524287
Computed Hops    :
    192.168.12.1(S)
 -> 192.168.12.2(S)
 -> 192.168.25.2(S)
 -> 192.168.56.2(S)
 -> 192.168.36.1(S)
Resignal Eligible: False
Last Resignal    : n/a                   CSPF Metric          : 40
Last MBB    :
 MBB Type        : ConfigChange          MBB State            : Success
 Ended At        : 09/07/2018 12:32:12  Old Metric           : 40
 Signaled BW     : 0 Mbps
===============================================================================
*A:PE-1#
```

The tunnel table shows the RSVP LSP with next hop PE-2 and a metric of 16777215 (infinity) because the IGP metric is not used:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.3/32      rsvp      MPLS  2         7        192.168.12.2   16777215
-------------------------------------------------------------------------------
```

```
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

It is also possible that the following path is taken: from PE-1 to PE-4 to PE-5 to PE-6 to PE-3. The CSPF metric is 40.

The IGP metric values are restored to their default value as follows:

```
*A:PE-1# configure router isis interface "int-PE-1-PE-2" level 1 no metric
*A:PE-2# configure router isis interface "int-PE-2-PE-1" level 1 no metric
```

The TE metrics are configured with the same value for all interfaces; as follows:

```
*A:PE-2# configure router mpls interface "int-PE-2-PE-3" te-metric 10
*A:PE-3# configure router mpls interface "int-PE-3-PE-2" te-metric 10
```

When all metrics have the same value, it does not matter whether CSPF uses the IGP or TE metric. CSPF will use the IGP metric after the following command is executed.

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" cspf
```

The primary path will go from PE-1 to PE-2 and then to PE-3 with a CSPF (IGP) metric of 20.

Fast Reroute for RSVP-TE LSPs

It is mandatory to have CSPF enabled for FRR.

Fast reroute can be configured on the RSVP LSP in two ways:

1. One-to-one: for each potential point of failure, the best detour tunnel to the eLER is signaled. This detour tunnel is signaled for this particular LSP only and cannot be shared among LSPs
2. Facility: local bypass tunnels are signaled from each point of local repair avoiding the next link or the next node. The bypass tunnels can be shared among LSPs.

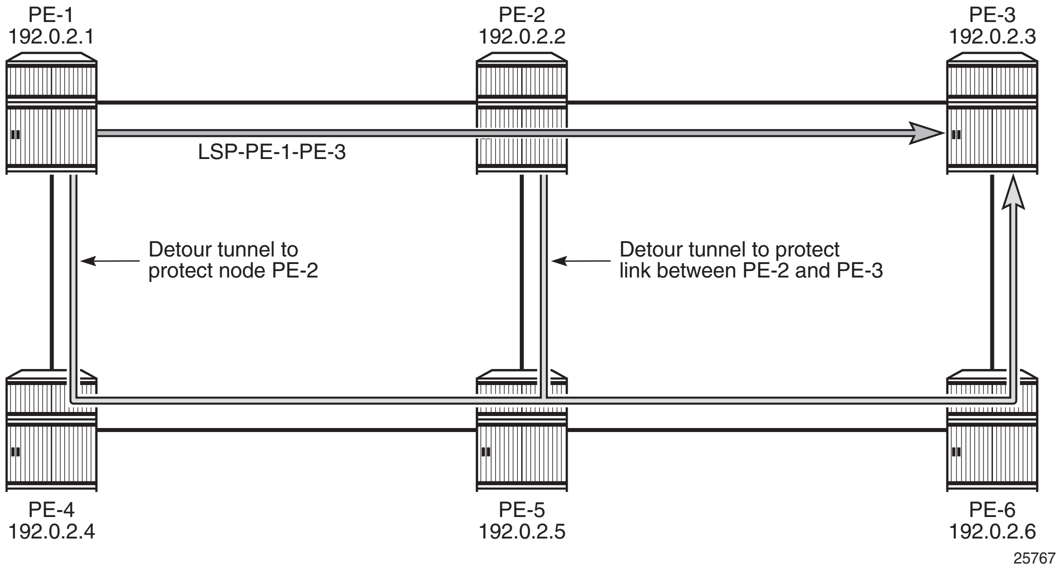
FRR One-to-One

The LSP "LSP-PE-1-PE-3" is configured with FRR one-to-one; as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" fast-reroute one-to-one
```

The preferred path from PE-1 to PE-3 is via PE-2. There will be two detour tunnels: one originating in PE-1 to protect node PE-2, and a second detour tunnel originating in PE-2 to protect the link between PE-2 and PE-3. Both detour tunnels use the same path from PE-5 to PE-3 and there is no need to signal this path twice. One detour tunnel terminates in PE-5, and the diverted traffic in this tunnel will be sent to PE-6 and PE-3 via the established detour tunnel. Depending on which detour tunnel is established first, the other detour tunnel terminates in PE-5. The preferred tunnel and the detour tunnels are shown in Figure 287:

Figure 287 **Fast Reroute One-to-One Detour Tunnels**



The protection can be seen in the list of actual hops in the path. In PE-1, a detour tunnel for node protection originates (indicated by @ n; see legend) and in PE-2 a detour tunnel for link protection (indicated by @):

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-3
From              : 192.0.2.1            To                    : 192.0.2.3
```

```
Admin State      : Up                   Oper State           : Up
Path Name        : dyn
Path LSP ID      : 1034                 Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 524286
---snip---

CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
FRR              : Enabled              Oper FRR             : Enabled
FRR NodeProtect  : Enabled              Oper FRR NP          : Enabled
FR Hop Limit     : 16                   Oper FRHopLimit      : 16
FR Prop Admin Gr*: Disabled             Oper FRPropAdmGrp    : Disabled
---snip---

Actual Hops      :
    192.168.12.1 (192.0.2.1) @ n                 Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @                   Record Label        : 524286
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524286
---snip---

Detour Status    : Standby              Detour Type          : Originate
Detour Avoid Nod*: 192.0.2.2            Detour Origin        : 192.0.2.1
Setup Priority   : 7                    Hold Priority        : 0
Class Type       : 0
Detour Active Ti*: n/a                  Detour Up Time       : 0d 00:00:42
In Interface     : n/a                  In Label             : n/a
Out Interface    : 1/1/2                Out Label            : 524287
NextHop          : 192.168.14.2
Explicit Hops    :
    192.168.14.1(S)
 -> 192.168.14.2(S)
 -> 192.168.45.2(S)
 -> 192.168.56.2(S)
 -> 192.168.36.1(S)
===============================================================================
```

The output also contains information about the detour tunnel originating in PE-1 that protects node PE-2. Because the detour tunnel is dedicated for this LSP, that information can be included in the LSP information.

The RSVP detour sessions can be retrieved in the originating, transit, and terminating nodes. On originating node PE-1:

```
*A:PE-1# show router rsvp session detour
```

```
===============================================================================
RSVP Sessions
===============================================================================
RSVP Session Name
    From                To              Tunnel ID   LSP ID       State
-------------------------------------------------------------------------------
LSP-PE-1-PE-3::dyn_detour
192.0.2.1           192.0.2.3           2           1034         Up

-------------------------------------------------------------------------------
Sessions : 1
```

In the transit/terminating node PE-5:

```
*A:PE-5# show router rsvp session detour-transit

===============================================================================
RSVP Sessions
===============================================================================
RSVP Session Name
    From                To              Tunnel ID   LSP ID       State
-------------------------------------------------------------------------------
LSP-PE-1-PE-3::dyn_detour
192.0.2.1           192.0.2.3           2           1034         Up

-------------------------------------------------------------------------------
Sessions : 1
===============================================================================

*A:PE-5# show router rsvp session detour-terminate

===============================================================================
RSVP Sessions
===============================================================================
RSVP Session Name
    From                To              Tunnel ID   LSP ID       State
-------------------------------------------------------------------------------
LSP-PE-1-PE-3::dyn_detour
192.0.2.1           192.0.2.3           2           1034         Up

-------------------------------------------------------------------------------
Sessions : 1
```

More detailed information can be retrieved as follows:

```
*A:PE-5# show router rsvp session detail

===============================================================================
RSVP Sessions (Detailed)
===============================================================================
-------------------------------------------------------------------------------
LSP : LSP-PE-1-PE-3::dyn_detour
-------------------------------------------------------------------------------
From            : 192.0.2.1             To             : 192.0.2.3
Tunnel ID       : 2                     LSP ID         : 1034
Style           : SE                    State          : Up
Session Type    : Transit (Detour)
In Interface    : 1/1/1                 Out Interface  : 1/1/2
In IF Name      : int-PE-5-PE-4
Out IF Name     : int-PE-5-PE-6
In Label        : 524287                Out Label      : 524287
Previous Hop    : 192.168.45.1          Next Hop       : 192.168.56.2
---snip---
-------------------------------------------------------------------------------
LSP : LSP-PE-1-PE-3::dyn_detour
-------------------------------------------------------------------------------
From            : 192.0.2.1             To             : 192.0.2.3
Tunnel ID       : 2                     LSP ID         : 1034
Style           : SE                    State          : Up
```

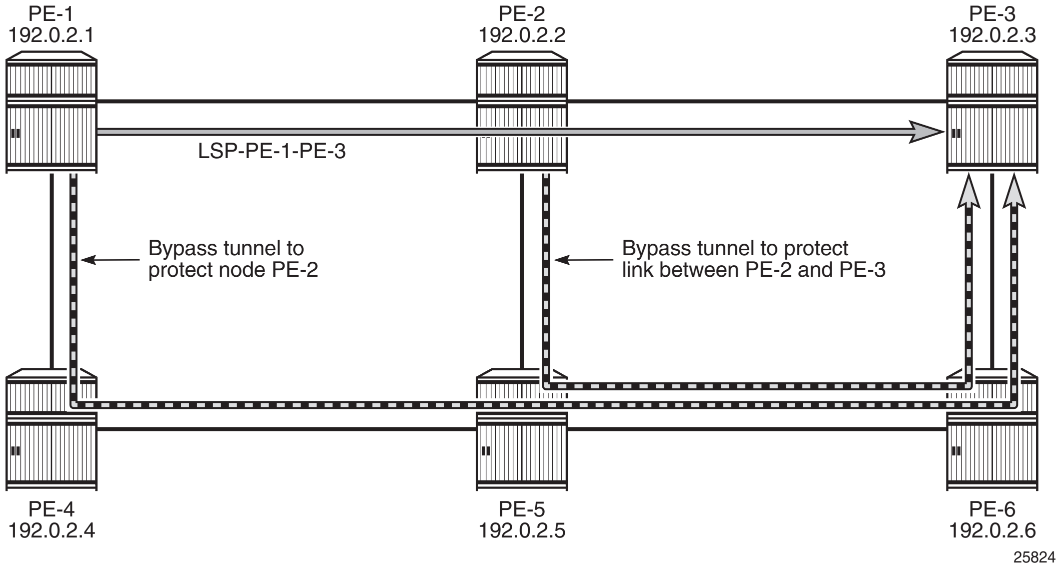
```
Session Type     : Terminate (Detour)
In Interface     : 1/1/3                 Out Interface  : 1/1/2
In IF Name       : int-PE-5-PE-2
Out IF Name      : int-PE-5-PE-6
In Label         : 524286                Out Label      : 524287
Previous Hop     : 192.168.25.1          Next Hop       : 192.168.56.2
---snip---
```

PE-5 is a transit node for the detour tunnel with previous hop PE-4 and a terminating node for the detour tunnel with previous hop PE-2. In both cases, the next hop is PE-6.

FRR Facility

The drawback of FRR one-to-one is that each LSP requires its own detour tunnels to be signaled. FRR facility does not have this issue, because it offers local repair for the next node or the next link that uses bypass tunnels that can be shared by LSPs. FRR facility bypass tunnels terminate in the merge point (MP), which is a hop in the primary path. FRR facility bypass tunnels for link protection terminate in the next hop in the primary path and FRR facility bypass tunnels for node protection terminate in the next hop of that next hop. FRR bypass tunnels are unaware of the final destination of the LSP and need not terminate in the final destination, but in this case they do, because the number of hops in the primary path is limited. Figure 288 shows the FRR facility bypass tunnels for LSP "LSP-PE-1-PE-3":

Figure 288 **Fast-Reroute Facility Bypass Tunnels**



Fast reroute facility is enabled on LSP "LSP-PE-1-PE-3" as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" fast-reroute facility
```

The LSP path detail output shows that there is a bypass tunnel available in PE-1 that offers node protection for the next node in the primary path: PE-2. In PE-2, there is a bypass tunnel offering link protection for the next link, which is the link between PE-2 and PE-3; as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn
Path LSP ID      : 1038                 Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 524286
---snip---

CSPF             : Enabled              Oper CSPF            : Enabled
Least Fill       : Disabled             Oper LeastFill       : Disabled
FRR              : Enabled              Oper FRR             : Enabled
FRR NodeProtect  : Enabled              Oper FRR NP          : Enabled
---snip---

Actual Hops      :
    192.168.12.1 (192.0.2.1) @ n                 Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @                   Record Label        : 524286
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524286
---snip---
```

In FRR facility mode, the bypass tunnels are shared. They are not included in the LSP information. The bypass tunnels can be retrieved as follows:

```
*A:PE-1# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-node192.0.2.2-61441
-------------------------------------------------------------------------------
To              : 192.168.36.1        State               : Up
Out I/F         : 1/1/2               Out Label           : 524287
Up Time         : 0d 00:19:28         Active Time         : n/a
```

```
Reserved BW    : 0 Kbps               Protected LSP Count : 1
Type           : Dynamic              Bypass Path Cost    : 40
Setup Priority : 7                    Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                 Inter-Area          : False
Computed Hops  :
    192.168.14.1(S)                   Egress Admin Groups : None
 -> 192.168.14.2(S)                   Egress Admin Groups : None
 -> 192.168.45.2(S)                   Egress Admin Groups : None
 -> 192.168.56.2(S)                   Egress Admin Groups : None
 -> 192.168.36.1(S)                   Egress Admin Groups : None
Actual Hops    :
    192.168.14.1 (192.0.2.1)          Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)          Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)          Record Label        : 524286
 -> 192.168.56.2 (192.0.2.6)          Record Label        : 524286
 -> 192.168.36.1 (192.0.2.3)          Record Label        : 524284
Last Resignal  :
Attempted At   : n/a                  Resignal Reason     : n/a
Resignal Status: n/a                  Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-3::dyn
From           : 192.0.2.1            To                  : 192.0.2.3
Avoid Node/Hop : 192.0.2.2            Downstream Label    : 524286
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-1#
```

This is the bypass tunnel that originates in PE-1 to protect (avoid) PE-2. In this example, there is only one LSP protected by this bypass tunnel, but the list of protected LSPs can be longer. The same command can be launched on PE-2, where a bypass tunnel originates that protects the link between PE-2 and PE-3.

The RSVP sessions can be displayed as follows:

```
*A:PE-3# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
RSVP Session Name
    From                To                Tunnel ID   LSP ID       State
-------------------------------------------------------------------------------
LSP-PE-1-PE-3::dyn
192.0.2.1           192.0.2.3           2           1038         Up

bypass-link192.168.23.2-61441
192.0.2.2           192.168.36.1        61441       2            Up

bypass-node192.0.2.2-61441
192.0.2.1           192.168.36.1        61441       2            Up

-------------------------------------------------------------------------------
Sessions : 3
```

In PE-3, there is an RSVP session for the regular LSP and two bypass tunnels. In this case, the bypass tunnels all go to PE-3, which is the terminating node for the LSP, but that need not be the case. All bypass tunnels are signaled from the point of local repair to the merge point on the LSP path.

To force a FRR facility switchover to a bypass tunnel, a failure is simulated as follows:

```
*A:PE-2# configure port 1/1/1 shutdown
```

The detailed output for the LSP path on PE-1 shows that the tunnel is locally repaired.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-1-PE-3
---snip---
Failure Code      : tunnelLocallyRepaired
Failure Node      : 192.0.2.2
Explicit Hops     :
    No Hops Specified
Actual Hops       :
    192.168.12.1 (192.0.2.1) @ n                Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @ #                Record Label        : 524286
 -> 192.168.23.2 (192.0.2.3)                    Record Label        : 524286
---snip---
```

The failure code is tunnelLocallyRepaired and next to the actual hop 192.168.12.2 (PE-2), the symbol # indicates that the detour is in use.

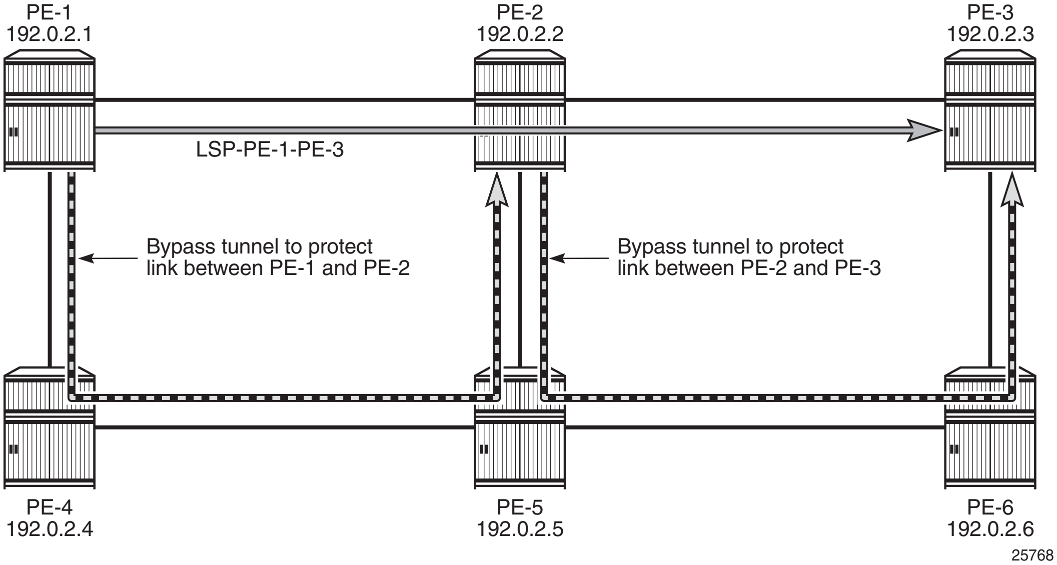
FRR Facility without Node Protection

Node protection is by default enabled, but can be disabled as follows:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-3"
                fast-reroute facility
                    no node-protect
                exit
```

As a result, there is only link protection. The bypass tunnels from PE-1 and PE-2 terminate in the next hop in the primary path, as shown in Figure 289:

Figure 289 **FRR Facility without Node Protection**



The LSP path detail output shows that there is no node protection (@):

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3
From             : 192.0.2.1            To                    : 192.0.2.3
Admin State      : Up                   Oper State            : Up
Path Name        : dyn
Path LSP ID      : 1042                 Path Type             : Primary
Path Admin       : Up                   Path Oper             : Up
Out Interface    : 1/1/1                Out Label             : 524287
---snip---

FRR              : Enabled              Oper FRR              : Enabled
FRR NodeProtect  : Disabled             Oper FRR NP           : Disabled
---snip---
Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @                   Record Label        : 524287
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524283
---snip---
```

The bypass tunnel originating in PE-1 is now terminating in PE-2 instead of PE-3; as follows:

```
*A:PE-1# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.12.2-61443
-------------------------------------------------------------------------------
To             : 192.168.25.1        State               : Up
Out I/F        : 1/1/2               Out Label           : 524287
Up Time        : 0d 00:01:04         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : Dynamic             Bypass Path Cost    : 30
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.14.1(S)                  Egress Admin Groups : None
 -> 192.168.14.2(S)                  Egress Admin Groups : None
 -> 192.168.45.2(S)                  Egress Admin Groups : None
 -> 192.168.25.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.14.1 (192.0.2.1)         Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)         Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)         Record Label        : 524284
 -> 192.168.25.1 (192.0.2.2)         Record Label        : 524282
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-3::dyn
From           : 192.0.2.1           To                  : 192.0.2.3
Avoid Node/Hop : 192.168.12.2        Downstream Label    : 524287
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-1#
```

In the remainder of this chapter, this LSP is no longer used. Therefore, the LSP is shut down:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-3" shutdown
```

Administrative Groups for RSVP-TE LSPs

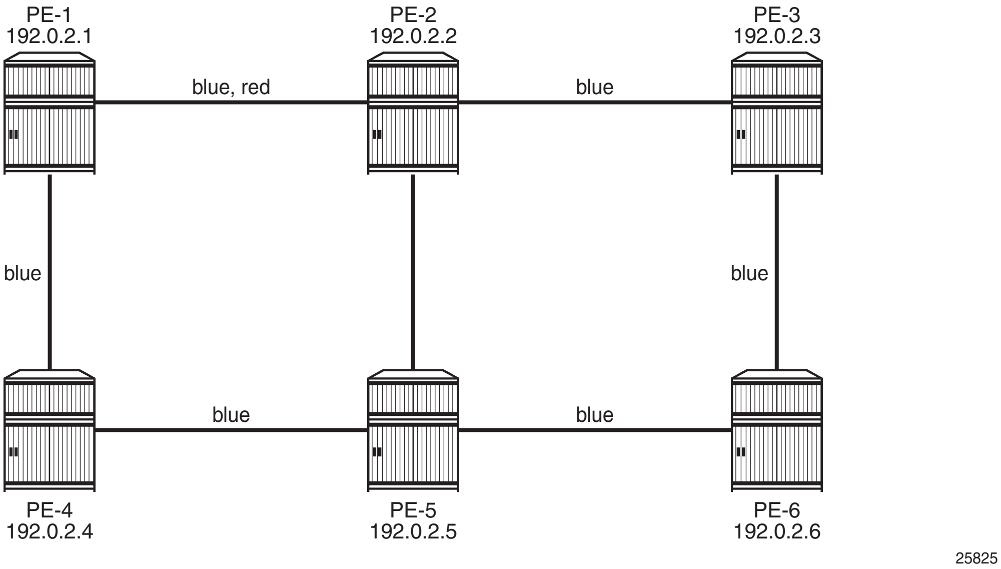
Administrative groups (link-coloring) can be used to calculate a path with the restriction to only include links of a particular admin group (color) or to exclude links of a particular admin group. Paths can be disjointed from each other, without the need for an explicit hops list.

Two admin groups are configured on all nodes; as follows:

```
*A:PE-1# configure router if-attribute admin-group "red" value 0
*A:PE-1# configure router if-attribute admin-group "blue" value 1
```

Admin group "blue" is assigned to all MPLS interfaces, except for the link between PE-2 and PE-5 while admin group "red" is only assigned to the link between PE-1 and PE-2; see Figure 290:

Figure 290 **Admin Groups 'Blue' and 'Red'**



The admin groups are assigned to the MPLS interfaces as follows:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
                admin-group "blue"
                admin-group "red"
            exit
            interface "int-PE-1-PE-4"
                admin-group "blue"
            exit
```

The configuration on the other nodes is similar.

To ensure that FRR bypass tunnels will adhere to the same admin group constraints as defined in the LSP, the following is configured on all nodes. It is required on all Points of Local Repair (PLRs):

```
*A:PE-1# configure router mpls admin-group-frr
```

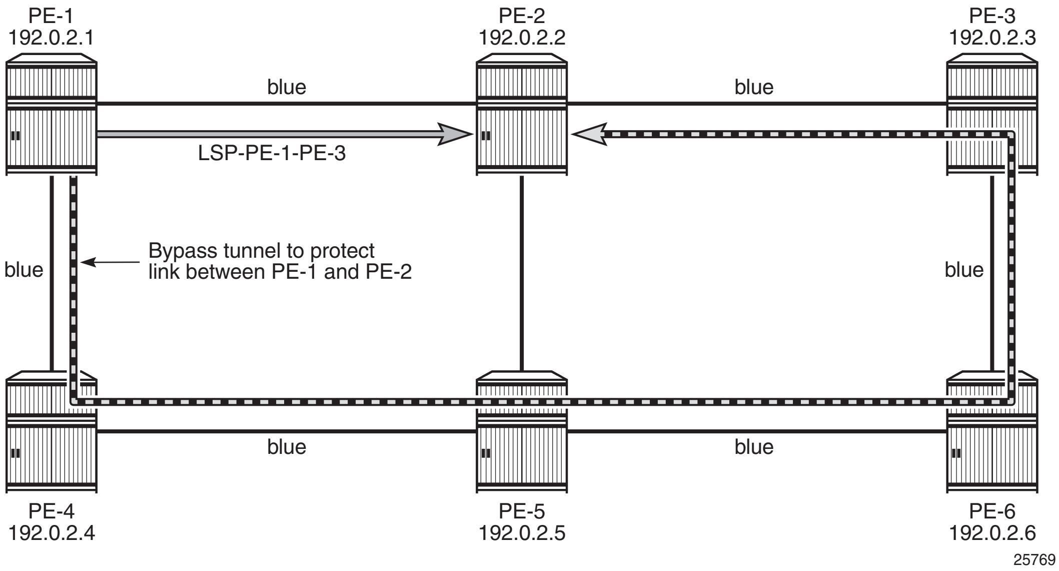
LSP Includes Admin Group 'blue'

LSP "LSP-PE-1-PE-2" is created on PE-1 with a dynamic primary path. FRR facility is enabled. The LSP includes admin group blue and both the primary path as the bypass tunnel must use links in admin group "blue" (**propagate-admin-group**). **Admin-group-frr** is enabled in the MPLS context, to ensure that the admin group restriction is respected for FRR.

```
*A:PE-1# configure
    router
        mpls
            admin-group-frr
            lsp "LSP-PE-1-PE-2"
                to 192.0.2.2
                cspf
                include "blue"
                propagate-admin-group
                fast-reroute facility
                    propagate-admin-group
                exit
                primary "dyn"
                exit
                no shutdown
            exit
```

The bypass tunnel cannot include the link between PE-2 and PE-5, because that link does not belong to admin group "blue". The LSP and its bypass tunnel are shown in Figure 291:

Figure 291 **LSP and Bypass within Admin Group 'Blue'**



The LSP path detailed information is as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-2
From             : 192.0.2.1            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : dyn
Path LSP ID      : 51200                Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 524287
---snip---

FR Prop Admin Gr*: Enabled              Oper FRPropAdmGrp    : Enabled
Propogate Adm Grp: Enabled              Oper Prop Adm Grp    : Enabled
---snip---

Include Groups   :                      Oper Include Groups  :
blue                                           blue
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None
---snip---

Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
```

```
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 524287
---snip---
```

There is a bypass tunnel originating in PE-1 that offers protection for the link between PE-1 and PE-2. More information about this bypass tunnel can be retrieved as follows:

```
*A:PE-1# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.12.2-61444
-------------------------------------------------------------------------------
To             : 192.168.23.1        State               : Up
Out I/F        : 1/1/2               Out Label           : 524287
Up Time        : 0d 00:09:40         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : Dynamic             Bypass Path Cost    : 50
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.14.1(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.14.2(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.45.2(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.56.2(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.36.1(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.23.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.14.1 (192.0.2.1)         Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)         Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)         Record Label        : 524287
 -> 192.168.56.2 (192.0.2.6)         Record Label        : 524287
 -> 192.168.36.1 (192.0.2.3)         Record Label        : 524287
 -> 192.168.23.1 (192.0.2.2)         Record Label        : 524286
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-2::dyn
From           : 192.0.2.1           To                  : 192.0.2.2
Avoid Node/Hop : 192.168.12.2        Downstream Label    : 524287
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-1#
```

All egress links are in admin group blue on the originating and transit nodes.

LSP Excludes Admin Group 'red'

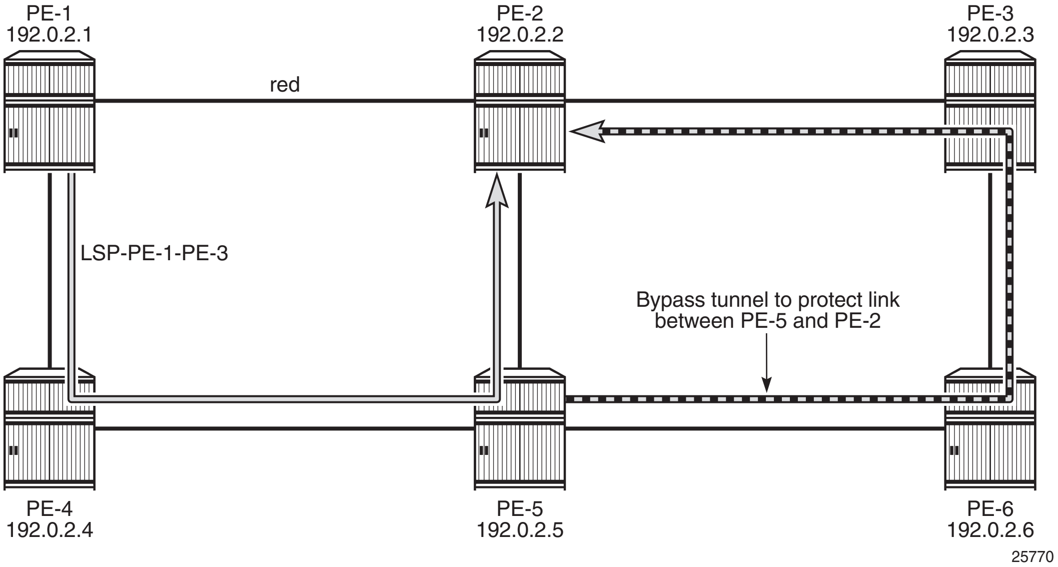
The LSP is reconfigured: instead of including admin group 'blue', it will exclude admin group 'red'. Nothing is changed to the configuration of FRR.

The MPLS configuration is modified as follows:

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-PE-1-PE-2"
                no include "blue"
                exclude "red"
            exit
```

The LSP cannot use the red link between PE-1 and PE-2. The path that avoids the red link, is from PE-1 via PE-4 and PE-5 to PE-2. On all PLRs, **admin-group-frr** is configured, which implies that the originating FRR bypass tunnels need to respect the admin-group constraint of the LSP. There can be no node protection for PE-4 or PE-5 without using the red link between PE-1 and PE-2. The only link that can be protected without using the red link between PE-1 and PE-2, is the link between PE-5 and PE-2. The LSP and the FRR bypass tunnel are shown in Figure 292:

Figure 292 **LSP and FRR Bypass Tunnel Excluding Admin Group 'Red'**



The LSP path can be verified as follows:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" path detail
---snip---
-------------------------------------------------------------------------------
```

```
LSP LSP-PE-1-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-2
From             : 192.0.2.1              To                   : 192.0.2.2
Admin State      : Up                     Oper State           : Up
---snip---

Include Groups   :                        Oper Include Groups  :
None                                             None
Exclude Groups   :                        Oper Exclude Groups  :
red                                              red
---snip---

Actual Hops      :
    192.168.14.1 (192.0.2.1)                         Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)                         Record Label        : 524286
 -> 192.168.45.2 (192.0.2.5) @                       Record Label        : 524286
 -> 192.168.25.1 (192.0.2.2)                         Record Label        : 524284
---snip---
```

There is only link protection for the link from PE-5 to PE-2. The bypass tunnel originates in PE-5 and has no links belonging to admin group 'red':

```
*A:PE-5# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.25.1-61468
-------------------------------------------------------------------------------
To             : 192.168.23.1        State               : Up
Out I/F        : 1/1/2               Out Label           : 524286
Up Time        : 0d 00:02:26         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : Dynamic             Bypass Path Cost    : 30
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.56.1(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.56.2(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.36.1(S)                  Egress Admin Groups :
                                     blue
 -> 192.168.23.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.56.1 (192.0.2.5)         Record Label        : N/A
 -> 192.168.56.2 (192.0.2.6)         Record Label        : 524286
 -> 192.168.36.1 (192.0.2.3)         Record Label        : 524286
 -> 192.168.23.1 (192.0.2.2)         Record Label        : 524283
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-1-PE-2::dyn
```

```
From            : 192.0.2.1           To                   : 192.0.2.2
Avoid Node/Hop  : 192.168.25.1        Downstream Label     : 524284
Bandwidth       : 0 Kbps

===============================================================================
*A:PE-5#
```

This configuration is preserved for the following example.

Hop Limit for RSVP-TE LSPs

Another constraint to influence the path selection, is hop limit. This can be configured on the LSP, on a secondary path, or on FRR in case the path should not contain too many hops. In this example, it will be configured on the LSP and later also for FRR on that LSP. By default, the LSP hop limit is 255, but it can be configured as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" hop-limit 5
```

This hop limit of 5 is enough for the path via PE-4 and PE-5, but it will not be sufficient when the link between PE-2 and PE-5 is down:

```
*A:PE-5# configure port 1/1/3 shutdown
```

In this case, the only possible path that excludes the 'red' link between PE-1 and PE-2, has to go to PE-2 via PE-4, PE-5, PE-6, and PE-3. There are too many hops. The FRR bypass tunnel can do a local repair, but no new LSP path can be signaled, with failure code: noCspfRouteToDestination:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" path detail
---snip---

Include Groups    :                       Oper Include Groups  :
None                                             None
Exclude Groups    :                       Oper Exclude Groups  :
red                                              red

Adaptive          : Enabled               Oper Metric          : 30
Preference        : n/a
Path Trans        : 4                     CSPF Queries         : 5
Failure Code      : tunnelLocallyRepaired
Failure Node      : 192.0.2.5
Explicit Hops     :
    No Hops Specified
Actual Hops       :
    192.168.14.1 (192.0.2.1)                       Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)                       Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5) @ #                   Record Label        : 524287
 -> 192.168.23.1 (192.0.2.2)                       Record Label        : 524287
---snip---
```

```
In Prog MBB :
 MBB Type        : GlobalRevert         Next Retry In        : 23 sec
 Started At      : 09/07/2018 13:47:18  Retry Attempt        : 1
 Failure Code    : noCspfRouteToDestina Failure Node         : 192.0.2.1
                   tion
 Signaled BW     : 0 Mbps
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

FRR tunnels also have a hop limit. The FRR hop limit is by default 16, but can be configured as follows:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" fast-reroute hop-limit 3
```

When the LSP is recalculated, it is impossible to establish the primary path with a hop limit of 5 and it is also impossible to establish a bypass tunnel protecting the link between PE-5 and PE-2 when the FRR hop limit is 3. The LSP will remain operationally down with failure code: noCspfRouteToDestination:

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
---snip---
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-2
From             : 192.0.2.1            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Down
Path Name        : dyn
Path LSP ID      : 51210                Path Type            : Primary
Path Admin       : Up                   Path Oper            : Down
Out Interface    : n/a                  Out Label            : n/a
Path Up Time     : 0d 00:00:00          Path Down Time       : 0d 00:00:40
---snip---

CSPF             : Enabled              Oper CSPF            : N/A
Least Fill       : Disabled             Oper LeastFill       : N/A
FRR              : Enabled              Oper FRR             : N/A
FRR NodeProtect  : Enabled              Oper FRR NP          : N/A
FR Hop Limit     : 3                    Oper FRHopLimit      : N/A
FR Prop Admin Gr*: Enabled              Oper FRPropAdmGrp    : N/A
Propogate Adm Grp: Enabled              Oper Prop Adm Grp    : N/A
---snip---

Neg MTU          : 0                    Oper MTU             : N/A
Bandwidth        : No Reservation       Oper Bandwidth       : N/A
Hop Limit        : 5                    Oper HopLimit        : N/A
---snip---
```

```
Include Groups    :                        Oper Include Groups  :
None                                              N/A
Exclude Groups    :                        Oper Exclude Groups  :
red                                               N/A
---snip---

Failure Code      : noCspfRouteToDestination
Failure Node      : 192.0.2.1
---snip---
```

For the remainder of the examples, FRR is disabled and the hop limit is restored to the default value, which is 255:

```
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" no fast-reroute
*A:PE-1# configure router mpls lsp "LSP-PE-1-PE-2" no hop-limit
```

The port is put in a no shutdown state:

```
*A:PE-5# configure port 1/1/3 no shutdown
```

Manual Resignal

Instead of waiting for the resignal timer to expire, one can manually trigger the resignal process.

The command to resignal the path "dyn" of LSP "LSP-PE-1-PE-2":

```
*A:PE-1# tools perform router mpls resignal lsp "LSP-PE-1-PE-2" path "dyn"
```

The command to resignal all RSVP LSPs originating at node PE-1:

```
*A:PE-1# tools perform router mpls resignal delay 0
```

This command can only be launched after the resignal timer is configured.

```
*A:PE-1# configure router mpls resignal-timer
  - no resignal-timer
  - resignal-timer <minutes>

 <minutes>            : [30..10080]
```

The configuration timer is configured to 30 minutes as follows:

```
*A:PE-1# configure router mpls resignal-timer 30
```

Whenever an LSP is resignaled, the resignal timer is restarted.

LSP OAM

The LSP diagnostics are modeled after ICMP echo request/reply which provides a mechanism to detect data plane failures in MPLS LSPs. For a given FEC, LSP ping verifies whether the packet reaches the egress label edge router (LER).

```
*A:PE-1# oam lsp-ping "LSP-PE-1-PE-2"
LSP-PING LSP-PE-1-PE-2: 92 bytes MPLS payload
Seq=1, send from intf int-PE-1-PE-4, reply from 192.0.2.2
       udp-data-len=32 ttl=255 rtt=1.09ms rc=3 (EgressRtr)

---- LSP LSP-PE-1-PE-2 PING Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 1.09ms, avg = 1.09ms, max = 1.09ms, stddev = 0.000ms
*A:PE-1#
```

In LSP traceroute mode, the packet is sent to the control plane of each transit label switched router (LSR) which performs various checks to see if it is actually a transit LSR for the path.

```
*A:PE-1# oam lsp-trace "LSP-PE-1-PE-2"
lsp-trace to LSP-PE-1-PE-2: 0 hops min, 0 hops max, 116 byte packets
1  192.0.2.4  rtt=0.679ms rc=8(DSRtrMatchLabel) rsc=1
2  192.0.2.5  rtt=1.10ms rc=8(DSRtrMatchLabel) rsc=1
3  192.0.2.2  rtt=1.88ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

```
*A:PE-1# oam lsp-trace "LSP-PE-1-PE-2" detail
lsp-trace to LSP-PE-1-PE-2: 0 hops min, 0 hops max, 116 byte packets
1  192.0.2.4  rtt=0.566ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.45.2 ifaddr=192.168.45.2 iftype=ipv4Numbered MRU=1564
           label[1]=524286 protocol=4(RSVP-TE)
2  192.0.2.5  rtt=2.00ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.25.1 ifaddr=192.168.25.1 iftype=ipv4Numbered MRU=1564
           label[1]=524286 protocol=4(RSVP-TE)
3  192.0.2.2  rtt=1.84ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

RSVP LSP Statistics

Statistics can be collected for RSVP LSPs. For each accounting record, a file ID is configured; as follows:

```
*A:PE-1# configure
    log
        file-id 2
            location cf1:
            rollover 5 retention 1
        exit
```

An accounting policy is configured for each record type; as follows:

```
*A:PE-1# configure
    log
        accounting-policy 2
            record combined-mpls-lsp-ingress
            to file 2
            no shutdown
        exit
```

The collection of statistics is enabled in the MPLS context as follows:

```
*A:PE-1# configure
    router
        mpls
            ingress-statistics
                lsp "LSP-PE-1-PE-2" sender 192.0.2.1
                    accounting-policy 2
                    no shutdown
                    collect-stats
                exit
            exit
```

To display the statistics, the following options are available for lsp-ingress-stats:

```
*A:PE-1# show router mpls lsp-ingress-stats
  - - lsp-ingress-stats [type <lsp-type>] [active]
                        [template-match <SessionNameString> [sender <ip-address>]]
  -  - lsp-ingress-stats lsp <lsp-name> sender <ip-address>

 <lsp-name>           : max 64 chars
 <ip-address>         : a.b.c.d
 <lsp-type>           : p2p|p2mp
 <active>             : match on all stats enabled lsp
 <template-match>     : match on p2p/p2mp stats template
 <SessionNameString>  : [Max 64 chars]
```

The following command retrieves the LSP ingress statistics for LSP "LPS-PE-1-PE-2" with sender 192.0.2.1:

```
*A:PE-1# show router mpls lsp-ingress-stats lsp "LSP-PE-1-PE-2" sender 192.0.2.1

===============================================================================
MPLS LSP Ingress Statistics
===============================================================================
-------------------------------------------------------------------------------
LSP Name      : LSP-PE-1-PE-2
Sender        : 192.0.2.1
-------------------------------------------------------------------------------
Collect Stats : Enabled                 Accting Plcy. : 2
Adm State     : Up                      PSB Match     : False
FC BE
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC L2
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC AF
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC L1
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC H2
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC EF
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC H1
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
FC NC
InProf Pkts   : 0                       OutProf Pkts  : 0
InProf Octets : 0                       OutProf Octets: 0
===============================================================================
*A:PE-1#
```

Statistics can be cleared as follows:

```
*A:PE-1# clear router mpls lsp-ingress-stats 192.0.2.1 lsp "LSP-PE-1-PE-2"
```

Debug

A wide range of debug tools are available which can be tuned to the specific information of importance for a certain troubleshooting task. In the **debug router mpls** context, the LSP object to trace or monitor can be selected by the following parameters:

- LSP name
- Source address of the LSP (the from parameter in the LSP definition)
- Termination point of the LSP (the to parameter in the LSP definition)
- Tunnel ID of the LSP
- LSP ID

```
*A:PE-1# debug router rsvp
  - no rsvp
  - rsvp [lsp name>] [sender <sender-address>] [endpoint <endpoint-address>]
         [tunnel-id <tunnel-id>] [lsp-id <lsp-id>] [interface <ip-int-name>]

 <name>                : [160 chars max]
 <sender-address>      : a.b.c.d
 <endpoint-address>    : a.b.c.d
 <tunnel-id>           : [0..4294967295]
 <lsp-id>              : [1..65535]
 <ip-int-name>         : [32 chars max]
 [no] event           + Enable/disable debugging for specific RSVP events
 [no] packet          + Enable/disable debugging for specific RSVP packets


*A:PE-1# debug router mpls
  - mpls [lsp <name>] [sender <source-address>] [endpoint <endpoint-address>]
    [tunnel-id <tunnel-id>] [lsp-id <lsp-id>]
  - no mpls

 <name>               : [160 chars max]
 <source-address>     : a.b.c.d
 <endpoint-address>   : a.b.c.d
 <tunnel-id>          : [0..4294967295]
 <lsp-id>             : [1..65535]

 [no] event           + Enable/disable debugging for specific MPLS events
```

In the **debug** command tree, the MPLS event type can be selected (tracing must be enabled):

```
*A:PE-1# debug router mpls lsp "LSP-PE-1-PE-2" event
  - event
  - no event

 [no] all             -  Enable/disable debugging for MPLS all
 [no] frr             -  Enable/disable debugging for MPLS frr
 [no] iom             -  Enable/disable debugging for MPLS iom
 [no] lsp-setup       -  Enable/disable debugging for MPLS lsp setup
 [no] mbb             -  Enable/disable debugging for MPLS mbb
 [no] misc            -  Enable/disable debugging for MPLS misc
 [no] pcc             -  Enable/disable debugging for MPLS PCC
 [no] te              -  Enable/disable debugging for MPLS TE
 [no] xc              -  Enable/disable debugging for MPLS xc
```

As an example, the **all** keyword is entered, logging all MPLS events related to the selected LSP:

```
*A:PE-1# debug router mpls lsp "LSP-PE-1-PE-2" event all
```

The last step is to create a log container which will gather all MPLS debugging information according to the criteria set in the debug context. The **from debug-trace** parameter must be configured but there are several options where the different captured entries will be stored: console, a syslog server, SNMP, local file on the compact flash card, a temporary circular memory buffer, or the telnet/SSH session from which you are logged into the node.

The ID of the log container is a local number without any other significance.

```
*A:PE-1# configure log log-id 2 to
  - to cli [<size>]
  - to console
  - to file <log-file-id>
  - to memory [<size>]
  - to netconf [<size>]
  - to session
  - to snmp [<size>]
  - to syslog <syslog-id>

 <console>            : keyword - specifies console as destination
 <syslog-id>          : [1..10]
 <snmp>               : keyword - specifies SNMP as destination
 <log-file-id>        : [1..99]
 <memory>             : keyword - specifies memory as destination
 <session>            : keyword - specifies telnet session as destination
 <netconf>            : keyword - specifies NETCONF as destination
 <cli>                : keyword - set the destination to any subscribed CLI session
 <size>               : [50..3000]
```

For this example, the temporary buffer (with adjustable size) is chosen, as follows:

```
configure
    log
        log-id 2
            from debug-trace
            to memory
            exit
```

All MPLS events related to the selected LSP are stored in the location (memory) specified. The content of this log container can be viewed through the **show log log-id 2** command. The following output is a subset of messages shown after port 1/1/2 on PE-2 is shut down, which causes LSP “LSP-PE-1-PE-2” to go down. The most recent message is on top.

```
*A:PE-1# show log log-id 2
===============================================================================
Event Log 2
===============================================================================
Description : (Not Specified)
Memory Log contents  [size=100   next event=34  (not wrapped)]

56 2018/09/07 13:49:24.350 UTC MINOR: DEBUG #2001 Base [TE]
"[TE]: CSPF[764/ISIS-0]
Use SRLG: (NOT-STRICT) EXCL color: 1 (00000001)
"

55 2018/09/07 13:49:24.350 UTC MINOR: DEBUG #2001 Base [TE]
"[TE]: CSPF[764/ISIS-0]
Max Hops: 254,  Use IGP Metric, "

54 2018/09/07 13:49:24.350 UTC MINOR: DEBUG #2001 Base [TE]
"[TE]: CSPF[764/ISIS-0]
INTRA-AREA LSP from 192.0.2.1 to 192.0.2.2 for a randomly selected lowest cost TE path
 with constraints"

53 2018/09/07 13:49:24.350 UTC MINOR: DEBUG #2001 Base MPLS
"MPLS: CSPF
Make CSPF request for LSP-PE-1-PE-2::dyn(LspId 51216)
From 192.0.2.1, To 192.0.2.2
Bandwidth 0 B/sec, Hop Limit 254, UseTeMetric 0, OutIntf 0
Include Colors 0x0, Exclude Colors 0x1
```

```
SRLG Usage - NOT-STRICT
SRLG Use DB - TEDB
DiffservClassType 0 TeClass 7, DsModel 300
CSPF Request Type Random ECMP, Leastfill Thresh 0, Leastfill Reopt Thres 0
IntraAreaOnly No, SkipUnnumbered No, PreferCurrHops No
Exclude List: None"
---snip---
```

Conclusion

MPLS provides the capability to establish connection oriented paths over a connectionless network. The LSP offers a mechanism to engineer network traffic on constraint-based paths rather than the IGP shortest path. This can greatly improve network resiliency. In this chapter, the configuration of several RSVP LSP features is given together with the associated show output which can be used to verify and troubleshoot.

RSVP Signaled Point-to-Multipoint LSPs

This chapter provides information about RSVP signaled point-to-multipoint LSPs.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was originally written for SR OS release 7.0.R5, but the CLI in the current edition corresponds to SR OS release 16.0.R3.

Overview

Point-to-MultiPoint (P2MP) Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) allow the source of multicast traffic to forward packets to one or many multicast receivers over a network without requiring a multicast protocol, such as Protocol Independent Multicast (PIM), to be configured in the network. A P2MP LSP tree is established in the control plane, and the path consists of a head-end node, one or many branch and bud nodes, and the leaf nodes. A bud node combines the roles of branch node and leaf node (for different source-to-leaf LSPs). Packets injected by the head-end node are replicated in the data plane at the branching nodes before they are delivered to the leaf nodes.

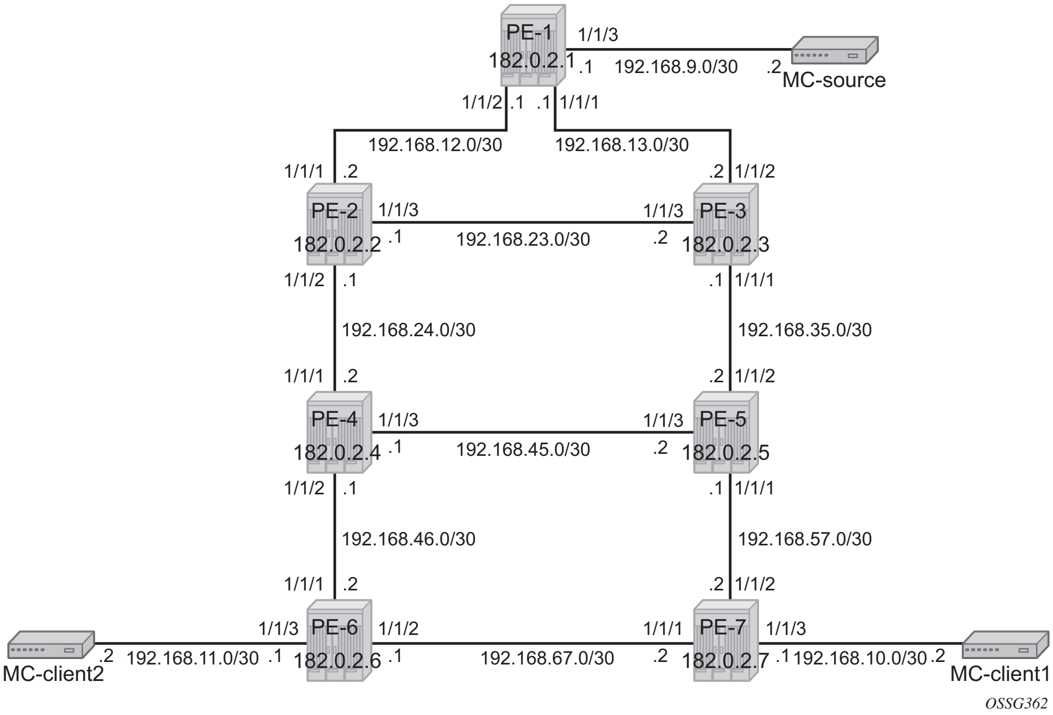
Similar to point-to-point (P2P) LSPs, also P2MP LSPs are unidirectional, originating on a head-end node (the ingress LER) and terminating on one or more leaf nodes (the egress LERs). Resource Reservation Protocol (RSVP) is used as signaling protocol. A P2MP LSP is modeled as a set of root-to-leaf sub LSPs (Source-to-Leaf: S2L). Each S2L is modeled as a point-to-point LSP in the control plane. This means that each S2L has its own PATH/RESV messages. This is called the de-aggregated method.

The forwarding of multicast packets to the LSP tree was initially based on static multicast routes, and has evolved to BGP-based VPN routes afterward (but the latter is beyond the scope of this chapter). In this example, forwarding multicast packets is done over P2MP RSVP LSPs in the base router instance.

RSVP signaled P2MP LSPs can have fast reroute (FRR) enabled, the facility method (one-to-many) with link protection is supported.

Figure 293 shows the P2MP example topology with seven PEs. The multicast source is connected to PE-1, multicast client 1 is attached to PE-7, and multicast client 2 to PE-6, as follows:

Figure 293 **P2MP Example Topology**



Configuration

The following sections describe the tasks which must be performed to configure RSVP signaled point-to-multipoint LSPs.

Configuring the IP/MPLS Network

The system addresses and Layer 3 interface addresses are configured according to Figure 293. An Interior Gateway Protocol (IGP) is needed to distribute routing information to all PEs. In this case, the IGP is OSPF using the backbone area 0.0.0.0. A configuration example is shown for PE-1. A similar configuration is needed on all PEs.

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            address 192.168.12.1/30
            port 1/1/2
        exit
        interface "int-PE-1-PE-3"
            address 192.168.13.1/30
            port 1/1/1
        exit
        interface "system"
            address 192.0.2.1/32
        exit
        ospf
            traffic-engineering
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-1-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-1-PE-3"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
```

Because fast reroute (FRR) is enabled for the P2MP LSP, Traffic Engineering (TE) is needed on the IGP. By doing this, OSPF will generate opaque LSAs which are collected in a Traffic Engineering Database (TED), separate from the traditional OSPF topology database. OSPF interfaces are set up as type *point-to-point* to improve convergence, because no Designated Router/Backup Designated Router (DR/BDR) election process is done. However, convergence is out of the scope of this chapter.

To verify that OSPF neighbors are up (state **Full**), **show router ospf neighbor** is performed. To check if Layer 3 interface addresses/subnets are known on all PEs, **show router route-table** or **show router fib** *iom-card-slot* will display the content of the forwarding information base (FIB).

```
*A:PE-1# show router ospf neighbor

===============================================================================
Rtr Base OSPFv2 Instance 0 Neighbors
===============================================================================
Interface-Name                   Rtr Id          State      Pri  RetxQ    TTL
   Area-Id
-------------------------------------------------------------------------------
int-PE-1-PE-2                    192.0.2.2       Full       1    0        32
   0.0.0.0
int-PE-1-PE-3                    192.0.2.3       Full       1    0        31
   0.0.0.0
-------------------------------------------------------------------------------
No. of Neighbors: 2
===============================================================================
*A:PE-1#


*A:PE-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
       Next Hop[Interface Name]                                     Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h01m19s  0
       system                                                       0
192.0.2.2/32                                  Remote  OSPF      00h00m48s  10
       192.168.12.2                                                 10
192.0.2.3/32                                  Remote  OSPF      00h00m37s  10
       192.168.13.2                                                 10
192.0.2.4/32                                  Remote  OSPF      00h00m29s  10
       192.168.12.2                                                 20
192.0.2.5/32                                  Remote  OSPF      00h00m24s  10
       192.168.13.2                                                 20
192.0.2.6/32                                  Remote  OSPF      00h00m14s  10
       192.168.12.2                                                 30
192.0.2.7/32                                  Remote  OSPF      00h00m07s  10
       192.168.13.2                                                 30
192.168.12.0/30                               Local   Local     00h01m19s  0
       int-PE-1-PE-2                                                0
192.168.13.0/30                               Local   Local     00h01m19s  0
       int-PE-1-PE-3                                                0
192.168.23.0/30                               Remote  OSPF      00h00m48s  10
       192.168.12.2                                                 20
192.168.24.0/30                               Remote  OSPF      00h00m48s  10
       192.168.12.2                                                 20
192.168.35.0/30                               Remote  OSPF      00h00m37s  10
       192.168.13.2                                                 20
192.168.45.0/30                               Remote  OSPF      00h00m29s  10
       192.168.12.2                                                 30
192.168.46.0/30                               Remote  OSPF      00h00m29s  10
       192.168.12.2                                                 30
192.168.57.0/30                               Remote  OSPF      00h00m24s  10
       192.168.13.2                                                 30
192.168.67.0/30                               Remote  OSPF      00h00m14s  10
       192.168.12.2                                                 40
-------------------------------------------------------------------------------
No. of Routes: 16
---snip---
```

```
*A:PE-1# show router fib 1

===============================================================================
FIB Display
===============================================================================
Prefix [Flags]                                              Protocol
    NextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                                LOCAL
    192.0.2.1 (system)
192.0.2.2/32                                                OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.0.2.3/32                                                OSPF
  192.168.13.2 (int-PE-1-PE-3)
192.0.2.4/32                                                OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.0.2.5/32                                                OSPF
  192.168.13.2 (int-PE-1-PE-3)
192.0.2.6/32                                                OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.0.2.7/32                                                OSPF
  192.168.13.2 (int-PE-1-PE-3)
192.168.12.0/30                                             LOCAL
  192.168.12.0 (int-PE-1-PE-2)
192.168.13.0/30                                             LOCAL
  192.168.13.0 (int-PE-1-PE-3)
192.168.23.0/30                                             OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.168.24.0/30                                             OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.168.35.0/30                                             OSPF
  192.168.13.2 (int-PE-1-PE-3)
192.168.45.0/30                                             OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.168.46.0/30                                             OSPF
  192.168.12.2 (int-PE-1-PE-2)
192.168.57.0/30                                             OSPF
  192.168.13.2 (int-PE-1-PE-3)
192.168.67.0/30                                             OSPF
  192.168.12.2 (int-PE-1-PE-2)
-------------------------------------------------------------------------------
Total Entries : 16
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

On PE-1, the interface toward the multicast source is configured in an IES service. This could have been on a router interface instead.

```
*A:PE-1# configure
    service
        ies 1 name "IES 1" customer 1 create
            interface "int-PE-1-MC-source" create
                address 192.168.9.1/30
                sap 1/1/3 create
                exit
            exit
            no shutdown
        exit
```

Similar IES services are configured on PE-7 and PE-6 for multicast client 1 and multicast client 2.

```
*A:PE-7# configure
    service
        ies 1 name "IES 1" customer 1 create
            interface "int-PE-7-MC-client1" create
                address 192.168.10.1/30
                sap 1/1/3 create
                exit
            exit
            no shutdown
        exit
```

```
*A:PE-6# configure
    service
        ies 1 name "IES 1" customer 1 create
            interface "int-PE-6-MC-client2" create
                address 192.168.11.1/30
                sap 1/1/3 create
                exit
            exit
            no shutdown
        exit
```

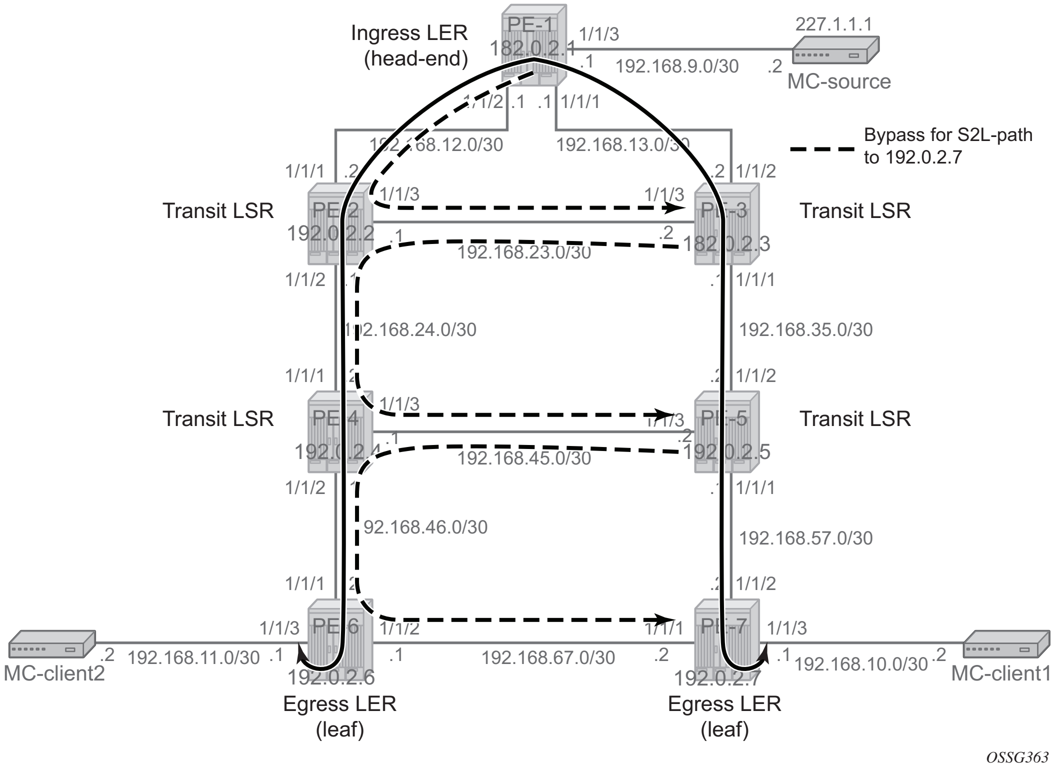
The next step in the process of setting up a P2MP LSP, is enabling the L3 interfaces in the MPLS and RSVP context on all involved PE nodes (from PE-1 to PE-7). By default, the system interface is put automatically within the MPLS/RSVP context. When an interface is put in the MPLS context, SR OS copies it also in the RSVP context. Explicit enabling of MPLS and RSVP context is done by the **no shutdown** command. The MPLS/RSVP configuration for PE-1 is as follows:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-3"
            exit
            no shutdown
        exit
        rsvp no shutdown
```

Configuring P2MP RSVP LSP

Figure 294 shows the P2MP LSP LSP-p2mp-1 with facility backup.

Figure 294 **P2MP LSP LSP-p2mp-1 with Bypass Tunnels**



A P2MP LSP (LSP-p2mp-1) will be set up from PE-1 acting as head-end node and PE-6 and PE-7 acting as leaf nodes. Because FRR is enabled, Constrained Shortest Path First (CSPF) is enabled to do route calculations on the Traffic Engineering Database (TED). FRR method **facility** is used without node protection, **facility** stands for one-to-many, meaning that one bypass tunnel can protect a set of primary LSPs with similar backup constraints. When a link failure occurs on one of the active S2L paths, the Point of Local Repair (PLR) node will push an additional MPLS label on the incoming MPLS packet before sending it into the bypass tunnel downstream toward the Merge Point (MP) node.

In the first example, the IGP OSPF will do the path calculation to the two destinations PE-6 and PE-7. The intermediate hops of the LSP are dynamically assigned by OSPF best route selection, so the S2L paths follow the IGP least cost path. Therefore, an MPLS path called “loose” is configured without specifying any hops.

```
*A:PE-1# configure
    router
        mpls
            path "loose"
                no shutdown
            exit
```

Creation of the P2MP LSP itself is done on the ingress LER or head-end node (PE-1 in the example) and can be seen in following CLI output. The name of the P2MP LSP is “LSP-p2mp-1”. A create-time keyword **p2mp-lsp** is added to the P2MP name to make a distinction in configuration between normal point-to-point LSPs and point-to-multipoint LSPs. A primary P2MP instance is initiated using the **primary-p2mp-instance** keyword accompanied with the P2MP instance name “p-LSP-p2mp-1”. Within this primary P2MP instance, the different S2Ls are defined using the **s2l-path** keyword. The same MPLS path name can be used for different S2Ls as long as the destination is different (**to** command).

```
*A:PE-1# configure
    router
        mpls
            lsp "LSP-p2mp-1" p2mp-lsp
                cspf
                fast-reroute facility
                    no node-protect
                exit
                primary-p2mp-instance "p-LSP-p2mp-1"
                    s2l-path "loose" to 192.0.2.6
                    exit
                    s2l-path "loose" to 192.0.2.7
                    exit
                exit
                no shutdown
            exit
```

On the head-end LER node of the P2MP LSP, several show commands can be used. A first set of show commands is used to verify the administrative and operational state of the P2MP LSP and its different S2L paths (including FRR bypass information). In this example, ”LSP-p2mp-1” P2MP LSP has two active S2L paths: one toward leaf node PE-6 and one to leaf node PE-7.

```
*A:PE-1# show router mpls p2mp-lsp

===============================================================================
MPLS P2MP LSPs (Originating)
===============================================================================
LSP Name                                            Tun     Fastfail  Adm  Opr
                                                    Id      Config
-------------------------------------------------------------------------------
LSP-p2mp-1                                          1       Yes       Up   Up
-------------------------------------------------------------------------------
LSPs : 1
===============================================================================


*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" detail

===============================================================================
MPLS P2MP LSPs (Originating) (Detail)
===============================================================================
Legend :
    + - Inherited
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name   : LSP-p2mp-1
LSP Type        : P2mpLsp                   LSP Tunnel ID        : 1
LSP Index       : 1                         TTM Tunnel Id        : 1
From            : 192.0.2.1
Adm State       : Up                        Oper State           : Up
LSP Up Time     : 0d 00:00:00               LSP Down Time        : 0d 00:00:00
Transitions     : 1                         Path Changes         : 1
Retry Limit     : 0                         Retry Timer          : 30 sec
Signaling       : RSVP                      Resv. Style          : SE
Hop Limit       : 255                       Negotiated MTU       : n/a
Adaptive        : Enabled                   ClassType            : 0
FastReroute     : Enabled                   Oper FR              : Enabled
FR Method       : Facility                  FR Hop Limit         : 16
FR Node Protect : Disabled                  FR Prop Adm Grp      : Disabled
FR Object       : Enabled
CSPF            : Enabled                   ADSPEC               : Disabled
Metric          : Disabled                  Use TE metric        : Disabled
Load Bal Wt     : N/A                       ClassForwarding      : Disabled
Include Grps    :                           Exclude Grps         :
None                                           None
Least Fill      : Disabled

Revert Timer    : Disabled                  Next Revert In       : N/A
Auto BW         : Disabled
LdpOverRsvp     : Disabled
VprnAutoBind    : Disabled
IGP Shortcut    : Disabled                  BGP Shortcut         : Disabled
IGP LFA         : Disabled                  IGP Rel Metric       : Disabled
BGPTransTun     : Disabled
Oper Metric     : Disabled
Prop Adm Grp    : Disabled

P2MPInstance    : p-LSP-p2mp-1
                                            P2MP-Inst-type       : Primary
S2L Cfg Counter : 2                         S2L Oper Counter     : 2
S2L-Name        : loose
                                            To                   : 192.0.2.6
S2L-Name        : loose
                                            To                   : 192.0.2.7
===============================================================================
*A:PE-1#


*A:PE-1# show router mpls p2mp-info

===============================================================================
MPLS P2MP Cross Connect Information
===============================================================================
-------------------------------------------------------------------------------
S2L:LSP-p2mp-1::loose
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1             Tunnel ID     : 1
P2MP ID              : 0                     Lsp ID        : 63488
To                   : 192.0.2.6
Out Interface        : 1/1/2                 Out Label     : 524287
Num. of S2ls         : 1
-------------------------------------------------------------------------------
```

```
S2L LSP-p2mp-1::loose
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1             Tunnel ID     : 1
P2MP ID              : 0                     Lsp ID        : 63488
To                   : 192.0.2.7
Out Interface        : 1/1/1                 Out Label     : 524287
Num. of S2ls         : 1
-------------------------------------------------------------------------------
P2MP Cross-connect instances : 2
===============================================================================
*A:PE-1#

*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1"
===============================================================================
MPLS P2MP Instance (Originating)
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name    : LSP-p2mp-1
P2MP ID         : 0                         LSP Tunnel ID        : 1
Adm State       : Up                        Oper State           : Up

P2MPInstance    : p-LSP-p2mp-1
                                            P2MP-Inst-type       : Primary
P2MP Inst Id    : 1                         P2MP Lsp Id          : 14336
Inst Admin      : Up                        Inst Oper            : Up
Inst Up Time    : 0d 00:00:00               Inst Dn Time         : 0d 00:00:00
Hop Limit       : 255                       Adaptive             : Enabled
Record Route    : Record                    Record Label         : Record
Include Grps    :                           Exclude Grps         :
None                                           None
Bandwidth       : No Reservation            Oper Bw              : 0 Mbps
S2L-Name        : loose
                                            To                   : 192.0.2.6
S2L Admin       : Up                        S2L Oper             : Up
S2L-Name        : loose
                                            To                   : 192.0.2.7
S2L Admin       : Up                        S2L Oper             : Up
-------------------------------------------------------------------------------
P2MP instances : 1
===============================================================================
*A:PE-1#
```

Note: As long as one S2L path is operationally up (show router mpls p2mp-lsp lsp-name p2mp-instance instance-name) , the Oper State of the P2MP LSP is Up.

FRR information can be displayed in detail for each S2L path. From this moment onward, the focus is on the S2L path toward PE-7. The following command shows that link protection is present for the link between PE-1 and PE-3, for the link between PE-3 and PE-5, and for the link between PE-5 and PE-7 (‘@’-reference inside show command).

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l loose 
to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-1 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                          # - Detour In Use
    b - Bandwidth Protected                       n - Node Protected
    S - Strict                                    L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-1
S2L LSP ID       : 19968
P2MP ID          : 0                    S2L Grp Id           : 2
Admin State      : Up                   Oper State           : Up
S2L State:       : Active                                    :
S2L Name         : loose
To               : 192.0.2.7
S2L Admin        : Up                   S2L Oper             : Up
OutInterface     : 1/1/1                Out Label            : 524287
S2L Up Time      : 0d 00:01:10          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 1                    CSPF Queries         : 1
Failure Code     : noError              Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.13.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                   Record Label        : 524287
 -> 192.168.35.2 (192.0.2.5) @                   Record Label        : 524287
 -> 192.168.57.2 (192.0.2.7)                     Record Label        : 524287
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.57.2(S)
LastResignal     : n/a
===============================================================================
*A:PE-1#
```

More in detail, **show router mpls bypass-tunnel** can be used. **Actual Hops** provides the explicit hops of the bypass tunnel used to avoid the direct link between PE-1 and PE-3. On node PE-1, the MPLS path from PE-1 to PE-3 via PE-2 is followed (see Figure 294).

```
*A:PE-1# show router mpls bypass-tunnel detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
---snip---
-------------------------------------------------------------------------------
bypass-link192.168.13.2-61442
-------------------------------------------------------------------------------
```

```
To              : 192.168.23.2        State               : Up
Out I/F         : 1/1/2               Out Label           : 524285
Up Time         : 0d 00:09:51         Active Time         : n/a
Reserved BW     : 0 Kbps              Protected LSP Count : 1
Type            : P2mp                Bypass Path Cost    : 20
Setup Priority  : 7                   Hold Priority       : 0
Class Type      : 0
Exclude Node    : None                Inter-Area          : False
Computed Hops   :
    192.168.12.1(S)                   Egress Admin Groups : None
 -> 192.168.12.2(S)                   Egress Admin Groups : None
 -> 192.168.23.2(S)                   Egress Admin Groups : None
Actual Hops     :
    192.168.12.1 (192.0.2.1)          Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)          Record Label        : 524285
 -> 192.168.23.2 (192.0.2.3)          Record Label        : 524284

Protected LSPs -
LSP Name        : LSP-p2mp-1::loose
From            : 192.0.2.1           To                  : 192.0.2.7
Avoid Node/Hop  : 192.168.13.2        Downstream Label    : 524287
Bandwidth       : 0 Kbps
-------------------------------------------------------------------------------
---snip---
```

On node PE-3, the MPLS path from PE-3 to PE-5 via PE-2 and PE-4 is followed (see Figure 294) to avoid the direct link between PE-3 and PE-5.

```
*A:PE-3# show router mpls bypass-tunnel protected-lsp p2mp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.35.2-61441
-------------------------------------------------------------------------------
To              : 192.168.45.2        State               : Up
Out I/F         : 1/1/3               Out Label           : 524286
Up Time         : 0d 00:13:14         Active Time         : n/a
Reserved BW     : 0 Kbps              Protected LSP Count : 1
Type            : P2mp                Bypass Path Cost    : 30
Setup Priority  : 7                   Hold Priority       : 0
Class Type      : 0
Exclude Node    : None                Inter-Area          : False
Computed Hops   :
    192.168.23.2(S)                   Egress Admin Groups : None
 -> 192.168.23.1(S)                   Egress Admin Groups : None
 -> 192.168.24.2(S)                   Egress Admin Groups : None
 -> 192.168.45.2(S)                   Egress Admin Groups : None
Actual Hops     :
    192.168.23.2 (192.0.2.3)          Record Label        : N/A
 -> 192.168.23.1 (192.0.2.2)          Record Label        : 524286
 -> 192.168.24.2 (192.0.2.4)          Record Label        : 524285
 -> 192.168.45.2 (192.0.2.5)          Record Label        : 524284

Protected LSPs -
LSP Name        : LSP-p2mp-1::loose
From            : 192.0.2.1           To                  : 192.0.2.7
```

```
Avoid Node/Hop : 192.168.35.2        Downstream Label    : 524287
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-3#
```

A similar output can be seen on PE-5 node also. To avoid the direct link from PE-5 to PE-7, the MPLS path from PE-5 to PE-7 via PE-4 and PE-6 is followed, as shown in Figure 294.

On the transit LSRs and egress LER/leaf node (see Figure 294), the **show router mpls p2mp-info** command can be used. See the show command on node PE-3 for the S2L path to 192.0.2.7. Similar outputs are possible for nodes PE-5 and PE-7.

```
*A:PE-3# show router mpls p2mp-info
  - p2mp-info [type {originate|transit|terminate}] [s2l-endpoint <ip-address>]

 <originate|transit*> : keywords
 <ip-address>         : [a.b.c.d]


*A:PE-3# show router mpls p2mp-info
===============================================================================
MPLS P2MP Cross Connect Information
===============================================================================
-------------------------------------------------------------------------------
S2L LSP-p2mp-1::loose
-------------------------------------------------------------------------------
Source IP Address    : 192.0.2.1             Tunnel ID     : 1
P2MP ID              : 0                     Lsp ID        : 14336
To                   : 192.0.2.7
Out Interface        : 1/1/1                 Out Label     : 524287
Num. of S2ls         : 1
-------------------------------------------------------------------------------
P2MP Cross-connect instances : 1
===============================================================================
*A:PE-3#
```

Mapping Multicast Traffic

To map multicast traffic into the LSP tree from the head-end node until leaf node, PIM and Internet Group Management Protocol (IGMP) configurations are needed on the head-end node (PE-1) and leaf nodes (PE-6 and PE-7) of the P2MP RSVP LSP. The intermediate nodes (transit LSR or branch LSR) do not need any explicit configuration for that.

Head-end Node (Ingress LER) PE-1

PIM must be enabled on the interface toward the multicast source and PIM must be enabled on the tunnel interface. A tunnel interface should be seen as an internal representation of a specific P2MP LSP. Creation is done within the PIM context using the **tunnel-interface rsvp-p2mp** command followed by the P2MP LSP name. This is configured as follows:

```
*A:PE-1# configure
    router
        pim
            interface "int-PE-1-MC-source"
            exit
            tunnel-interface rsvp-p2mp "LSP-p2mp-1"
```

For multicast packets received on an interface to pass through the data plane, a successful reverse path forwarding (RPF) check must be done on the source address, otherwise the packet will be dropped.

Besides enabling PIM on the tunnel interface, also IGMP is enabled to do a static <S,G> or <*,G> join of a multicast group address (227.1.1.1 in the example) to the tunnel interface/P2MP LSP. There is always a one-to-one mapping between <S,G> or <*,G> and a tunnel interface/P2MP LSP. In the example a < S,G > will be configured. A <*,G> join scenario is included in Additional Topics.

```
*A:PE-1# configure
    router
        igmp        …
            tunnel-interface rsvp-p2mp "LSP-p2mp-1"
                static
                    group 227.1.1.1
                        source 192.168.9.2
                    exit
                exit
            exit
            no shutdown
```

The **show router pim tunnel-interface** command shows you the admin state of the tunnel interface and an association to an internal local ifindex (**73728** in the example).

```
*A:PE-1# show router pim tunnel-interface

===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                        Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73728                    N/A                  Up   Up   Tx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-1#
```

The **show router igmp group** command provides the configured <S,G> entry and outgoing interface (= tunnel interface), represented by mpls-if-73728.

```
*A:PE-1# show router igmp group 227.1.1.1
===============================================================================
IGMP Interface Groups
===============================================================================

(192.168.9.2,227.1.1.1)                                   UpTime: 0d 00:11:10
    Fwd List  : mpls-if-73728
-------------------------------------------------------------------------------
Entries : 1
===============================================================================
IGMP Host Groups
===============================================================================
No Matching Entries
===============================================================================
IGMP SAP Groups
===============================================================================
No Matching Entries
===============================================================================
*A:PE-1#
```

Users can verify if multicast traffic is using P2MP LSP at the head-end node using the **show router pim group** *group-address* **detail** command.

```
*A:PE-1# show router pim group 227.1.1.1 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 227.1.1.1
Source Address     : 192.168.9.2
RP Address         : 0
Advt Router        : 192.0.2.1
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      : 192.168.9.2
MRIB Src Flags     : direct
Keepalive Timer    : Not Running
Up Time            : 0d 00:02:42        Resolved By        : rtable-u

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:00
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       : 192.168.9.2
Incoming Intf      : int-PE-1-MC-source
Outgoing Intf List : mpls-if-73728

Curr Fwding Rate   : 8352.8 kbps
Forwarded Packets  : 124213             Discarded Packets  : 0
Forwarded Octets   : 5713798            RPF Mismatches     : 0
```

```
Spt threshold      : 0 kbps            ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-1#
```

Leaf Node (Egress LER)

In the PIM context, the same tunnel interface must be created as the head-end node. An explicit reference to the head-end system address, using the **sender** *systemIP_head-end_node* parameter is needed.

```
*A:PE-7# configure
    router
        pim
            tunnel-interface rsvp-p2mp LSP-p2mp-1 sender 192.0.2.1
```

The **show router pim tunnel-interface** command provides the admin state of the tunnel interface and an association to an internal local ifindex (73728 in this example, by coincidence the same ifindex as the one on the head-end node PE-1).

```
*A:PE-7# show router pim tunnel-interface

===============================================================================
PIM Interfaces ipv4
===============================================================================
Interface                       Originator Address   Adm  Opr  Transport Type
-------------------------------------------------------------------------------
mpls-if-73728                   N/A                  Up   Up   Tx-IPMSI
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-7#
```

The main goal on the leaf nodes is to get traffic off the P2MP LSP/tunnel interface. This is done using a multicast information policy (**multicast-info-policy**). Inside this MC policy, a range of multicast group addresses must be defined under a bundle context *(bundle1*) in order to see traffic (**channel**). Also inside the bundle context, the P2MP LSP is presented by the tunnel interface (**primary-tunnel-interface**). This is configured as follows:

```
*A:PE-7# configure
    mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle1" create
                primary-tunnel-interface rsvp-p2mp LSP-p2mp-1 sender 192.0.2.1
                channel "227.1.1.1" "227.1.1.1" create
                exit
            exit
            bundle "default" create
            exit
        exit
```

Note: The **channel** command must be seen as a range command with a start-mc-group-address and an end-mc-group-address. In this example, only one MC group address, 227.1.1.1 is seen.

The configured multicast information policy must be applied to the base router instance.

```
*A:PE-7# configure router multicast-info-policy "p2mp-pol"
```

On the leaf nodes PE-7 and PE-6, multicast clients are connected. IGMP is enabled on those multicast clients with a static <S,G> join to redirect multicast traffic downstream to the multicast client. This is configured as follows:

```
*A:PE-7# configure
    router
        igmp
            interface "int-PE-7-MC-client1"
                static
                    group 227.1.1.1
                        source 192.168.9.2
                    exit
                exit
            exit
```

The **show router igmp group** provides the configured <S,G> entry and outgoing interface “int-PE-7-MC-client1”.

```
*A:PE-7# show router igmp group 227.1.1.1
===============================================================================
IGMP Interface Groups
===============================================================================

(192.168.9.2,227.1.1.1)                                   UpTime: 0d 00:09:02
    Fwd List  : int-PE-7-MC-client1
-------------------------------------------------------------------------------
Entries : 1
===============================================================================
IGMP Host Groups
===============================================================================
No Matching Entries
===============================================================================
IGMP SAP Groups
===============================================================================
No Matching Entries
===============================================================================
*A:PE-7#
```

Now, users can verify if multicast traffic is sent to the multicast client using the **show router pim group** *group-address* **detail** command

```
*A:PE-7# show router pim group 227.1.1.1 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 227.1.1.1
Source Address     : 192.168.9.2
RP Address         : 0
Advt Router        :
Flags              :                    Type               : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:01:39        Resolved By        : unresolved

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:21
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-7-MC-client1

Curr Fwding Rate   : 8352.8 kbps
Forwarded Packets  : 226349             Discarded Packets  : 0
Forwarded Octets   : 10412054           RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 1
===============================================================================
*A:PE-7#
```

OAM Tools

P2P LSP Operation, Administration, and Maintenance (OAM) commands (**oam lsp-ping** and **oam lsp-trace**) are extended for P2MP LSP. The user can instruct the head-end node to generate an P2MP LSP ping or a P2MP LSP trace by entering the command **oam p2mp-lsp-ping** or **oam p2mp-lsp-trace**. The P2MP OAM extensions are defined in *draft-ietf-mpls-p2mp-lsp-ping*.

For P2MP LSP ping, the echo request is sent on the active P2MP instance and replicated in the data path over all branches of the P2MP LSP instance. By default, all egress LER nodes which are leaves of the P2MP LSP instance will reply. Echo reply messages can be reduced by configuring the **s2l-dest-address** (a maximum of up to five egress nodes in a single run of the OAM command). Replies are sent by IP.

```
*A:PE-1# oam p2mp-lsp-ping
  - p2mp-lsp-ping {<lsp-name> [p2mp-instance <instance-name> [s2l-dest-address
                  <ipv4-  address> [... up to 5]]] [ttl <label-ttl>]}
  - p2mp-lsp-ping {ldp <p2mp-identifier> [vpn-recursive-fec] [sender-addr
                  <ipv4-address>] [leaf-addr <ipv4-address> [... up to 5]]}
  - p2mp-lsp-ping {ldp-ssm source <ip-address> group <ip-address> [router
                  <router-instance>|service-name <service-name>][sender-addr
                  <ipv4-address>] [leaf-addr <ipv4-address> [... up to 5]]}
  - options common to all p2mp-lsp-ping cases:  [fc <fc-name> [profile {in|out}]]
                  [size <octets>][timeout <timeout>] [detail]

 <lsp-name>           : [64 chars max]
 <instance-name>      : [32 chars max]
 <ipv4-address>       : a.b.c.d
 <in|out>             : in|out - Default: out
 <fc-name>            : be|l2|af|l1|h2|ef|h1|nc - Default: be
 <octets>             : [1..9198] - Default: 1
 <label-ttl>          : [1..255] - Default: 255
 <timeout>            : [1..120] seconds - Default: 10
 <detail>             : keyword - displays detailed information
 <p2mp-identifier>    : [1..4294967295]
 <ldp-ssm>         : keyword  - Label Distribution Protocol, Source-Specific Multicast
 <ip-address>         : ipv4-address   - a.b.c.d
                        ipv6-address   - x:x:x:x:x:x:x:x   (eight 16-bit pieces)
                                         x:x:x:x:x:x:d.d.d.d
                                         x - [0..FFFF]H
                                         d - [0..255]D
 <router-instance>    : <router-name>|<vprn-svc-id>
                        router-name    - "Base"  Default - Base
                        vprn-svc-id     - [1..2147483647]
 <service-name>       : [64 chars max]
 <vpn-recursive-fec>  : keyword - add a VPN Recursive FEC element to the launched
                        packet (useful for pinging a VPN BGP inter-AS Option B leaf)
```

```
*A:PE-1# oam p2mp-lsp-ping "LSP-p2mp-1" detail
P2MP LSP LSP-p2mp-1: 92 bytes MPLS payload

===============================================================================
S2L Information
===============================================================================
From             RTT                  Return Code
-------------------------------------------------------------------------------
192.0.2.6        =1.12ms              EgressRtr(3)
192.0.2.7        =1.13ms              EgressRtr(3)
===============================================================================

Total S2L configured/up/responded = 2/2/2,
          round-trip min/avg/max  = 1.12 / 1.12 / 1.13 ms

Responses based on return code:
        EgressRtr(3)=2

*A:PE-1#
```

Return codes are based on RFC 4379. Value 3 means the replying router is an egress for the FEC at stack depth.

P2MP LSP trace allows the user to trace the path of a single S2L path of a P2MP LSP from head-end node to leaf node. Using the downstream mapping TLV, each node along the S2L path can fill in the appropriate flags: B or E flag. The B-flag is set when the responding node is a branch LSR and the E-flag is set when the responding node is an egress LER.

```
*A:PE-1# oam p2mp-lsp-trace
  - p2mp-lsp-trace <lsp-name> p2mp-instance <instance-name> s2l-dest-address
    <ip-address> [fc <fc-name> [profile {in|out}]] [size <octets>] [max-fail
    <no-response-count>] [probe-count <probes-per-hop>] [min-ttl <min-label-ttl>]
    [max-ttl <max-label-ttl>] [timeout <timeout>]  [interval <interval>] [detail]

 <lsp-name>           : [64 chars max]
 <instance-name>      : [32 chars max]
 <ip-address>         : ipv4 address    a.b.c.d
 <fc-name>            : be|l2|af|l1|h2|ef|h1|nc - Default: be
 <in|out>             : in|out - Default: out
 <octets>             : [1..9198] - Default: 1
 <no-response-count>  : [1..10] - Default: 5
 <probes-per-hop>     : [1..10] - Default: 1
 <min-label-ttl>      : [1..255] - Default: 1
 <max-label-ttl>      : [1..255] - Default: 30
 <timeout>            : [1..60] seconds - Default: 3
 <detail>             : keyword - displays detailed information
 <interval>           : [1..10] seconds - Default: 1


*A:PE-1# oam p2mp-lsp-trace "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l-dest-
address 192.0.2.7 detail
P2MP LSP LSP-p2mp-1: 132 bytes MPLS payload
P2MP Instance p-LSP-p2mp-1, S2L Egress 192.0.2.7

 1  192.0.2.3  rtt=0.435 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.35.2 ifaddr=192.168.35.2 iftype=ipv4Numbered MRU=1564
     label=524287 proto=4(RSVP-TE) B/E flags:0/0
  2  192.0.2.5  rtt=0.838 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.57.2 ifaddr=192.168.57.2 iftype=ipv4Numbered MRU=1564
     label=524287 proto=4(RSVP-TE) B/E flags:0/0
  3  192.0.2.7  rtt=1.77 ms rc=3(EgressRtr)
*A:PE-1#
```

Return codes are based on RFC 4279. Value 8 means that the label is switched at stack depth. This is the case for a transit LSR doing MPLS label swapping. No B or E flag is set.

Additional Topics

<*,G> IGMP join instead of <S,G> IGMP join

In the Head-end Node (Ingress LER) PE-1 and Leaf Node (Egress LER) steps, a source specific IGMP join (<S,G> join) was used at the head-end node and leaf nodes. Another possibility is to use a source unknown or starg IGMP join (<*,G> join). When doing the latter, a rendezvous point (RP) must be defined in the PIM network. The RP allows multicast data flows between sources and receivers to meet at a predefined network location (in this example, the loopback address of node PE-1). It must be seen as an intermediate device to establish a multicast flow.

The RP can be defined in a dynamic way (BSR protocol) or a static way. In this example, the static way is chosen meaning that on all involved PIM nodes, the RP address will be statically configured. The following configuration is needed on head-end and leaf nodes.

```
*A:PE-1/PE-6/PE-7# configure
    router
        pim
            rp
                static
                    address 192.0.2.1
                        group-prefix 227.1.1.1/32
                    exit
                exit
```

The **group-prefix** is a mandatory keyword. It references a group address or group address range for which this rendezvous point will be used.

```
*A:PE-1/PE-6/PE-7# show router pim rp

===============================================================================
PIM RP Set ipv4
===============================================================================
Group Address                                                Hold Expiry
  RP Address                                   Type     Prio Time Time
-------------------------------------------------------------------------------
227.1.1.1/32
  192.0.2.1                                    Static   1    N/A  N/A
-------------------------------------------------------------------------------
Group Prefixes : 1
===============================================================================
*A:PE-1#
```

As previously mentioned, the configuration of the <*,G> IGMP join is done on the head-end node (PE-1) and leaf nodes (PE-6 and PE-7)

```
*A:PE-1# configure
```

```
    router
        igmp
            tunnel-interface rsvp-p2mp "LSP-p2mp-1"
                no shutdown
                static
                    group 227.1.1.1
                        starg
                    exit
                exit
            exit
```

```
*A:PE-6# configure
    router
        igmp
            interface "int-PE-6-MC-client2"
                static
                    group 227.1.1.1
                        starg
                    exit
                exit
            exit
```

```
*A:PE-7# configure
    router
        igmp
            interface "int-PE-7-MC-client1"
                static
                    group 227.1.1.1
                        starg
                    exit
                exit
            exit
```

The same preceding **show** command can be used to verify the multicast traffic on head-end node and leaf nodes, **show router igmp group 227.1.1.1** and **show router pim group 227.1.1.1 detail**.

```
*A:PE-7# show router igmp group 227.1.1.1

===============================================================================
IGMP Interface Groups
===============================================================================

(*,227.1.1.1)                                             UpTime: 0d 00:06:58
    Fwd List  : int-PE-7-MC-client1
-------------------------------------------------------------------------------
Entries : 1
===============================================================================
IGMP Host Groups
===============================================================================
No Matching Entries
===============================================================================
IGMP SAP Groups
===============================================================================
No Matching Entries
===============================================================================
*A:PE-7#
```

```
*A:PE-7# show router pim group 227.1.1.1 detail

===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 227.1.1.1
Source Address     : *
RP Address         : 192.0.2.1
Advt Router        :
Flags              :                    Type               : (*,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : remote
Keepalive Timer    : Not Running
Up Time            : 0d 00:05:34        Resolved By        : unresolved

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:25
Up JP Rpt          : Not Joined StarG   Up JP Rpt Override : 0d 00:00:00

Rpf Neighbor       :
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-7-MC-client1

Curr Fwding Rate   : 0.0 kbps
Forwarded Packets  : 31                 Discarded Packets  : 0
Forwarded Octets   : 1426               RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
===============================================================================
PIM Source Group ipv4
===============================================================================
Group Address      : 227.1.1.1
Source Address     : 192.168.9.2
RP Address         : 192.0.2.1
Advt Router        :
Flags              : spt                Type               : (S,G)
Mode               : sparse
MRIB Next Hop      :
MRIB Src Flags     : remote
Keepalive Timer Exp: 0d 00:01:27
Up Time            : 0d 00:05:34        Resolved By        : unresolved

Up JP State        : Joined             Up JP Expiry       : 0d 00:00:25
Up JP Rpt          : Not Pruned         Up JP Rpt Override : 0d 00:00:00

Register State     : No Info
Reg From Anycast RP: No

Rpf Neighbor       :
Incoming Intf      : mpls-if-73728
Outgoing Intf List : int-PE-7-MC-client1

Curr Fwding Rate   : 8352.8 kbps
Forwarded Packets  : 3539800            Discarded Packets  : 0
Forwarded Octets   : 162830800          RPF Mismatches     : 0
Spt threshold      : 0 kbps             ECMP opt threshold : 7
Admin bandwidth    : 1 kbps
-------------------------------------------------------------------------------
Groups : 2
```

```
===============================================================================
*A:PE-7#
```

Influence IGP Metric

Suppose that the IGP metric is increased on all links pointing to/from PE-2 and on the link between PE-5 and PE-7.

```
*A:PE-1# configure router ospf area 0 interface "int-PE-1-PE-2" metric 10000
```

```
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-1" metric 10000
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-3" metric 10000
*A:PE-2# configure router ospf area 0 interface "int-PE-2-PE-4" metric 10000
```

```
*A:PE-3# configure router ospf area 0 interface "int-PE-3-PE-2" metric 10000
```

```
*A:PE-4# configure router ospf area 0 interface "int-PE-4-PE-2" metric 10000
```

```
*A:PE-5# configure router ospf area 0 interface "int-PE-5-PE-7" metric 10000
```

```
*A:PE-7# configure router ospf area 0 interface "int-PE-7-PE-5" metric 10000
```

The existing P2MP LSP *LSP-p2mp-1* will not take into account these new constraints. The two S2L paths (one *loose* toward PE-6 and another one *loose* toward PE-7) are calculated using the default OSPF metric. To trigger MPLS to re-compute the S2L paths, configure a p2mp-resignal-timer on the head-end node inside the global MPLS context. Each time this timer expires (in the example, every 60 minutes), MPLS will trigger CSPF to re-compute the whole set of S2L paths of all active P2MP instances. MPLS performs a global Make-Before-Break (MBB) and moves each S2L sub-LSP in the instance into its new path using a new P2MP LSP ID if the global MBB is successful. **show router mpls status** gives an indication when the P2MP resignal timer will expire and which types of LSPs are set up on the node.

```
*A:PE-1# configure router mpls p2mp-resignal-timer 60
```

```
*A:PE-1# show router mpls status
```

```
===============================================================================
MPLS Status
===============================================================================
Admin Status              : Up          Oper Status               : Up
Oper Down Reason          : n/a
FRR Object                : Enabled     Resignal Timer            : Disabled
Hold Timer                : 1 seconds   Next Resignal             : N/A
Srlg Frr                  : Disabled    Srlg Frr Strict           : Disabled
Admin Group Frr           : Disabled
Dynamic Bypass            : Enabled     User Srlg Database        : Disabled
```

```
BypassResignalTimer       : Disabled    BypassNextResignal        : N/A
LeastFill Min Thd         : 5 percent   LeastFill Reopti Thd      : 10 percent
Local TTL Prop            : Enabled     Transit TTL Prop          : Enabled
AB Sample Multiplier      : 1           AB Adjust Multiplier      : 288
Exp Backoff Retry         : Disabled    CSPF On Loose Hop         : Disabled
Lsp Init RetryTimeout     : 30 seconds  MBB Pref Current Hops     : Disabled
Logger Event Bundling     : Disabled
RetryIgpOverload          : Disabled

P2mp Resignal Timer       : 60 minutes  P2mp Next Resignal        : 41 minutes
Sec FastRetryTimer        : Disabled    Static LSP FR Timer       : 30 seconds
P2P Max Bypass Association: 1000
P2PActPathFastRetry       : Disabled    P2MP S2L Fast Retry       : Disabled
In Maintenance Mode       : No
MplsTp                    : Disabled
Next Available Lsp Index  : 2
Entropy Label RSVP-TE     : Enabled     Entropy Label SR-TE       : Enabled
PCE Report RSVP-TE        : Disabled    PCE Report SR-TE          : Disabled
---snip---
===============================================================================
```

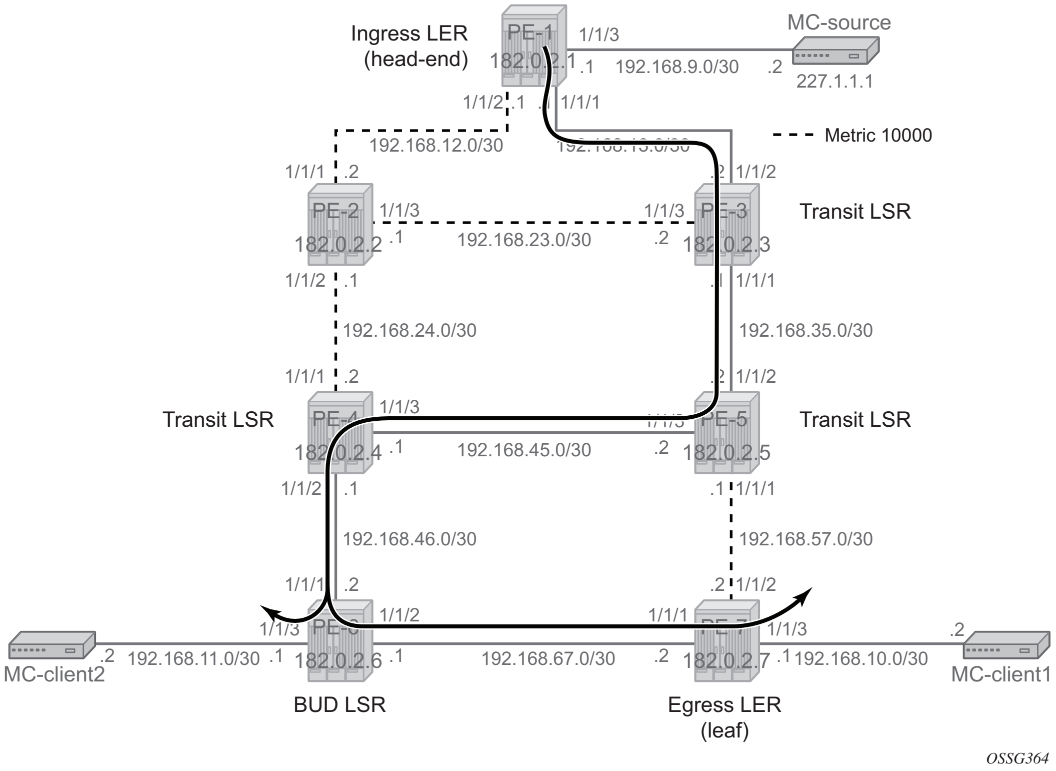
As an alternative, the user can also perform a manual resignal of a P2MP instance on the head-end node using the following tools command.

```
*A:PE-1# tools perform router mpls resignal p2mp-lsp "LSP-p2mp-1" p2mp-instance"p-LSP-
p2mp-1"
```

```
*A:PE-1# tools perform router mpls resignal p2mp-delay 0
```

Figure 295 shows the resignaled S2L paths. Node PE-6 is now a bud LSR node (instead of egress LER before).

Figure 295 **P2MP LSP p-to-mp-1 with Metric Change**



The resignaled S2L paths to PE-7 can be verified with the following command:

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l loose
to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-1 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                             # - Detour In Use
    b - Bandwidth Protected                          n - Node Protected
    S - Strict                                       L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name          : LSP-p2mp-1
S2L LSP ID        : 19970
P2MP ID           : 0                      S2L Grp Id           : 1
Admin State       : Up                     Oper State           : Up
S2L State:        : Active                                      :
S2L Name          : loose
To                : 192.0.2.7
S2L Admin         : Up                     S2L Oper             : Up
OutInterface      : 1/1/1                  Out Label            : 524283
S2L Up Time       : 0d 01:02:21            S2L Dn Time          : 0d 00:00:00
```

```
RetryAttempt     : 0                   NextRetryIn          : 0 sec
S2L Trans        : 2                   CSPF Queries         : 2
Failure Code     : noError             Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.13.1 (192.0.2.1) @                  Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                  Record Label        : 524283
 -> 192.168.35.2 (192.0.2.5) @                  Record Label        : 524283
 -> 192.168.45.1 (192.0.2.4) @                  Record Label        : 524283
 -> 192.168.46.2 (192.0.2.6) @                  Record Label        : 524284
 -> 192.168.67.2 (192.0.2.7)                    Record Label        : 524284
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.45.1(S)
 -> 192.168.46.2(S)
 -> 192.168.67.2(S)
LastResignal     : n/a
===============================================================================
*A:PE-1#
```

The resignaled S2L paths to PE-6 can be verified as follows:

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l loose
to 192.0.2.6 detail

===============================================================================
MPLS LSP LSP-p2mp-1 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                         # - Detour In Use
    b - Bandwidth Protected                      n - Node Protected
    S - Strict                                   L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-1
S2L LSP ID       : 19970
P2MP ID          : 0                   S2L Grp Id           : 2
Admin State      : Up                  Oper State           : Up
S2L State:       : Active                                   :
S2L Name         : loose
To               : 192.0.2.6
S2L Admin        : Up                  S2L Oper             : Up
OutInterface     : 1/1/1               Out Label            : 524283
S2L Up Time      : 0d 01:04:48         S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                   NextRetryIn          : 0 sec
S2L Trans        : 2                   CSPF Queries         : 2
Failure Code     : noError             Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.13.1 (192.0.2.1) @                  Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                  Record Label        : 524283
 -> 192.168.35.2 (192.0.2.5) @                   Record Label        : 524283
 -> 192.168.45.1 (192.0.2.4) @                   Record Label        : 524283
 -> 192.168.46.2 (192.0.2.6)                     Record Label        : 524284
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.45.1(S)
 -> 192.168.46.2(S)
LastResignal     : n/a
===============================================================================
*A:PE-1#
```

An **oam p2mp-lsp-trace** command toward PE-7 will now set the E flag on PE-6 because PE-6 acts also as an egress LER node.

```
*A:PE-1# oam p2mp-lsp-trace "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1"
s2l-dest-address 192.0.2.7 detail
P2MP LSP LSP-p2mp-1: 132 bytes MPLS payload
P2MP Instance p-LSP-p2mp-1, S2L Egress 192.0.2.7

  1  192.0.2.3  rtt=0.395 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.35.2 ifaddr=192.168.35.2 iftype=ipv4Numbered MRU=1564
     label=524283 proto=4(RSVP-TE) B/E flags:0/0
  2  192.0.2.5  rtt=0.611 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.45.1 ifaddr=192.168.45.1 iftype=ipv4Numbered MRU=1564
     label=524283 proto=4(RSVP-TE) B/E flags:0/0
  3  192.0.2.4  rtt=0.947 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.46.2 ifaddr=192.168.46.2 iftype=ipv4Numbered MRU=1564
     label=524284 proto=4(RSVP-TE) B/E flags:0/0
  4  192.0.2.6  rtt=1.26 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.67.2 ifaddr=192.168.67.2 iftype=ipv4Numbered MRU=1564
     label=524284 proto=4(RSVP-TE) B/E flags:0/1
  5  192.0.2.7  rtt=1.53 ms rc=3(EgressRtr)

*A:PE-1#
```

In the next step, the S2L path toward PE-7 is changed from **loose** to a **strict** direct MPLS path (*strict-to-PE-7)*. In that way, OSPF is not calculating anymore the shortest path to the leaf node.

```
*A:PE-1# configure
    router
        mpls
            path "path-strict-to-PE-7"
                hop 10 192.168.13.2 strict
                hop 20 192.168.35.2 strict
                hop 30 192.168.57.2 strict
                no shutdown
            exit
```

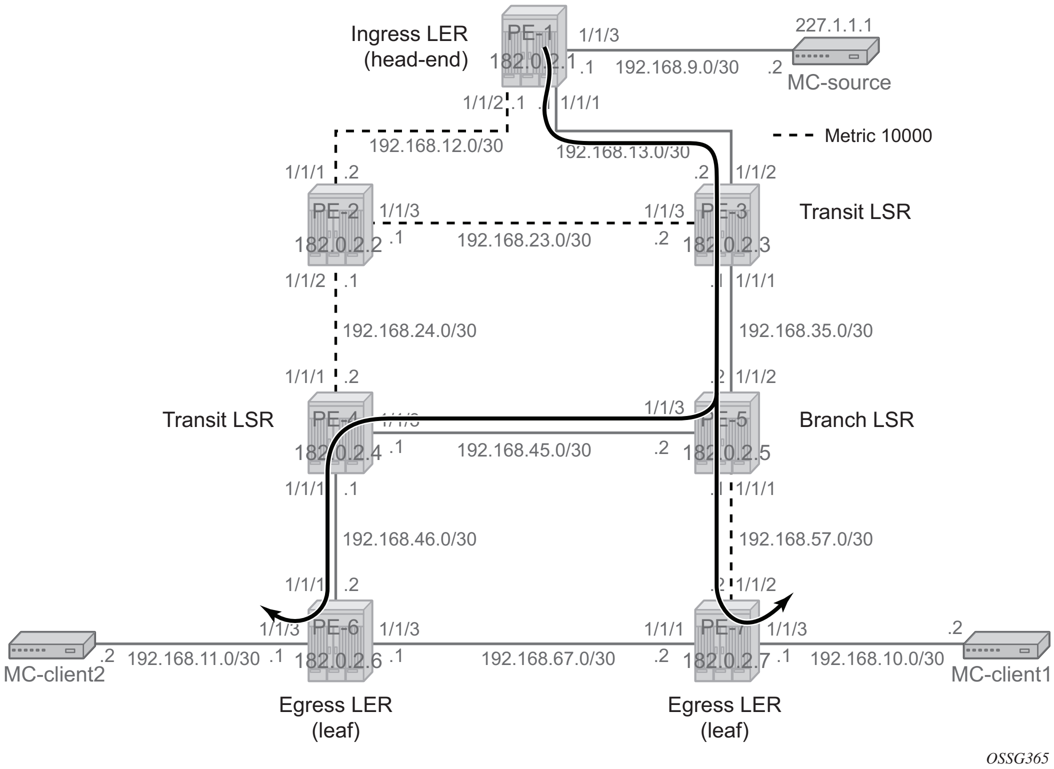
Before applying this new S2L path to the existing P2MP LSP (*LSP-p2mp-1*), the existing S2L path toward PE-7 must be removed.

```
*A:PE-1# configure
```

```
router
    mpls
        lsp "LSP-p2mp-1"
            primary-p2mp-instance "p-LSP-p2mp-1"
                s2l-path "loose" to 192.0.2.7 shutdown
                no s2l-path "loose" to 192.0.2.7
                s2l-path"path-strict-to-PE-7" to 192.0.2.7
                exit
            exit
```

As a consequence of this, only the S2L Grp Id has changed while S2L LSP ID remains the same as before. Figure 296 shows the P2MP LSP LSP-p2mp-1 with strict S2L path to leaf PE-7.

Figure 296 **P2MP LSP LSP-p2mp-1 with Strict S2L Path toward PE-7**



S2L paths can be verified according to Figure 296. PE-5 is now a branch LSR node (instead of a transit LSR before).

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l path-
strict-to-PE-7 to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-1 S2L path-strict-to-PE-7 (Detail)
===============================================================================
Legend :
    @ - Detour Available                          # - Detour In Use
    b - Bandwidth Protected                       n - Node Protected
    S - Strict                                    L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-1
S2l LSP ID       : 19970
P2MP ID          : 0                    S2L Grp Id          : 3
Admin State      : Up                   Oper State          : Up
S2L State:       : Active                                   :
S2L Name         : path-strict-to-PE-7
To               : 192.0.2.7
S2L Admin        : Up                   S2L Oper            : Up
OutInterface     : 1/1/1                Out Label           : 524283
S2L Up Time      : 0d 00:05:26          S2L Dn Time         : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn         : 0 sec
S2L Trans        : 1                    CSPF Queries        : 1
Failure Code     : noError              Failure Node        : n/a
Inter-area  : False
ExplicitHops:
    192.168.13.2(S)    -> 192.168.35.2(S)    -> 192.168.57.2(S)
Actual Hops :
    192.168.13.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                   Record Label        : 524283
 -> 192.168.35.2 (192.0.2.5) @                   Record Label        : 524283
 -> 192.168.57.2 (192.0.2.7)                     Record Label        : 524287
ComputedHops:
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.57.2(S)
LastResignal: n/a
===============================================================================
*A:PE-1#
```

An **oam p2mp-lsp-trace** command toward PE-7 will now set the B flag on PE-5 because PE-5 became a branch LSR now.

```
*A:PE-1# oam p2mp-lsp-trace "LSP-p2mp-1" p2mp-instance "p-LSP-p2mp-1" s2l-dest-
address  192.0.2.7 detail
P2MP LSP LSP-p2mp-1: 132 bytes MPLS payload
P2MP Instance p-LSP-p2mp-1, S2L Egress 192.0.2.7

 1  192.0.2.3  rtt=0.380 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.35.2 ifaddr=192.168.35.2 iftype=ipv4Numbered MRU=1564
     label=524283 proto=4(RSVP-TE) B/E flags:0/0
  2  192.0.2.5  rtt=0.724 ms rc=8(DSRtrMatchLabel)
     DS 1: ipaddr=192.168.57.2 ifaddr=192.168.57.2 iftype=ipv4Numbered MRU=1564
     label=524287 proto=4(RSVP-TE) B/E flags:1/0
  3  192.0.2.7  rtt=1.31 ms rc=3(EgressRtr)
```

Intelligent Re-merge

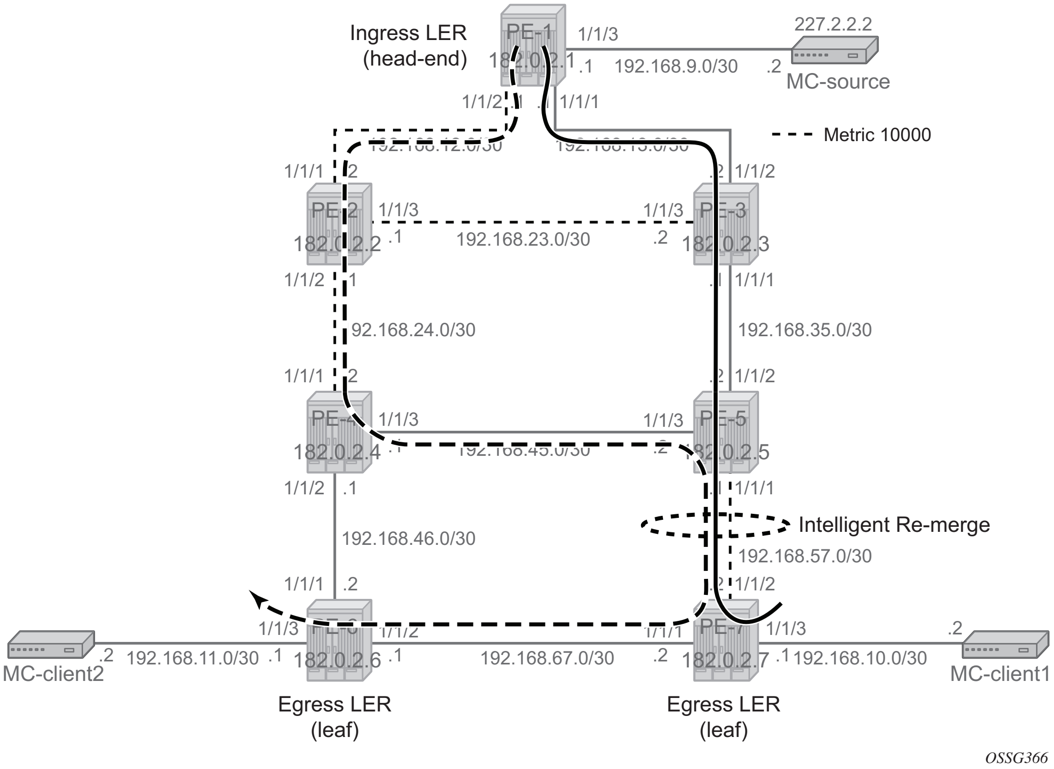
Intelligent re-merge protects users from receiving duplicate multicast traffic during convergence. It also protects against duplicate traffic in case of badly designed S2L paths. Three cases are described for which intelligent re-merge is implemented.

Case 1

When the paths of two different S2Ls of the same P2MP LSP instance have ingress label maps (ILMs) on different ports but go out on the same Next-Hop label Forwarding Entry (NHLFE).

Figure 298 shows P2MP LSP *LSP-p2mp-2* with two incoming S2L paths at PE-5.

Figure 297 **Intelligent Remerge, Case 1**



On the head-end node (PE-1), P2MP LSP *LSP-p2mp-2* is created with two strict direct MPLS paths *(strict-to-PE-7'*and *strict-to-PE-6*), as follows. Intelligent re-merge is performed at node PE-5.

```
*A:PE-1# configure
    router
        mpls
```

```
            path "path-strict-to-PE-7"
                hop 10 192.168.13.2 strict
                hop 20 192.168.35.2 strict
                hop 30 192.168.57.2 strict
                no shutdown
            exit
            path "path-strict-to-PE-6"
                hop 10 192.168.12.2 strict
                hop 20 192.168.24.2 strict
                hop 30 192.168.45.2 strict
                hop 40 192.168.57.2 strict
                hop 50 192.168.67.1 strict
                no shutdown
            exit
            lsp "LSP-p2mp-2" p2mp-lsp
                primary-p2mp-instance "p-LSP-p2mp-2"
                    s2l-path "path-strict-to-PE-7" to 192.0.2.7
                    exit
                    s2l-path "path-strict-to-PE-6" to 192.0.2.6
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-2" p2mp-instance "p-LSP-p2mp-2"

===============================================================================
MPLS P2MP Instance (Originating)
===============================================================================
-------------------------------------------------------------------------------
Type : Originating
-------------------------------------------------------------------------------
LSP Name    : LSP-p2mp-2
P2MP ID         : 0                         LSP Tunnel ID        : 2
Adm State       : Up                        Oper State           : Up

P2MPInstance    : p-LSP-p2mp-2
                                            P2MP-Inst-type       : Primary
P2MP Inst Id    : 2                         P2MP Lsp Id          : 7168
Inst Admin      : Up                        Inst Oper            : Up
Inst Up Time    : 0d 00:00:03               Inst Dn Time         : 0d 00:00:00
Hop Limit       : 255                       Adaptive             : Enabled
Record Route    : Record                    Record Label         : Record
Include Grps    :                           Exclude Grps         :
None                                           None
Bandwidth       : No Reservation            Oper Bw              : 0 Mbps
S2L-Name        : path-strict-to-PE-7
                                            To                   : 192.0.2.7
S2L Admin       : Up                        S2L Oper             : Up
S2L-Name        : path-strict-to-PE-6
                                            To                   : 192.0.2.6
S2L Admin       : Up                        S2L Oper             : Up
-------------------------------------------------------------------------------
P2MP instances : 1
===============================================================================
*A:PE-1#
```

To verify that node PE-5 is not sending duplicate multicast traffic downstream toward PE-7 while it receives two incoming multicast streams, a new tunnel interface and a new static <S,G> IGMP join will be configured on head-end node (PE-1) and leaf nodes (PE-6 and PE-7). Also on the leaf nodes, an extension to the existing multicast information policy is needed. This is configured as follows:

```
*A:PE-1# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-2"
```

```
*A:PE-1# configure
    router
        igmp
            tunnel-interface rsvp-p2mp "LSP-p2mp-2"
                static
                    group 227.2.2.2
                        source 192.168.9.2
                    exit
                exit
            exit
```

```
*A:PE-6# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-2" sender 192.0.2.1
```

```
*A:PE-6# configure
    router
        igmp
            interface "int-PE-6-MC-client2"
                static
                    group 227.2.2.2
                        source 192.168.9.2
                    exit
                exit
            exit
```

```
*A:PE-6# configure
    mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle2" create
                primary-tunnel-interface rsvp-p2mp "LSP-p2mp-2" sender 192.0.2.1
                    channel 227.2.2.2 create
                    exit
                exit
            exit
```

```
*A:PE-7# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-2" sender 192.0.2.1
```

```
*A:PE-7# configure
    router
        igmp
            interface "int-PE-7-MC-client1"
                static
                    group 227.2.2.2
                        source 192.168.9.2
                    exit
                exit
            exit
```

```
*A:PE-7# configure
    mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle2" create
                primary-tunnel-interface rsvp-p2mp "LSP-p2mp-2" sender 192.0.2.1
                    channel 227.2.2.2 create
                    exit
                exit
            exit
```

For verification of incoming/outgoing multicast traffic at node PE-5, the **monitor** command is used.

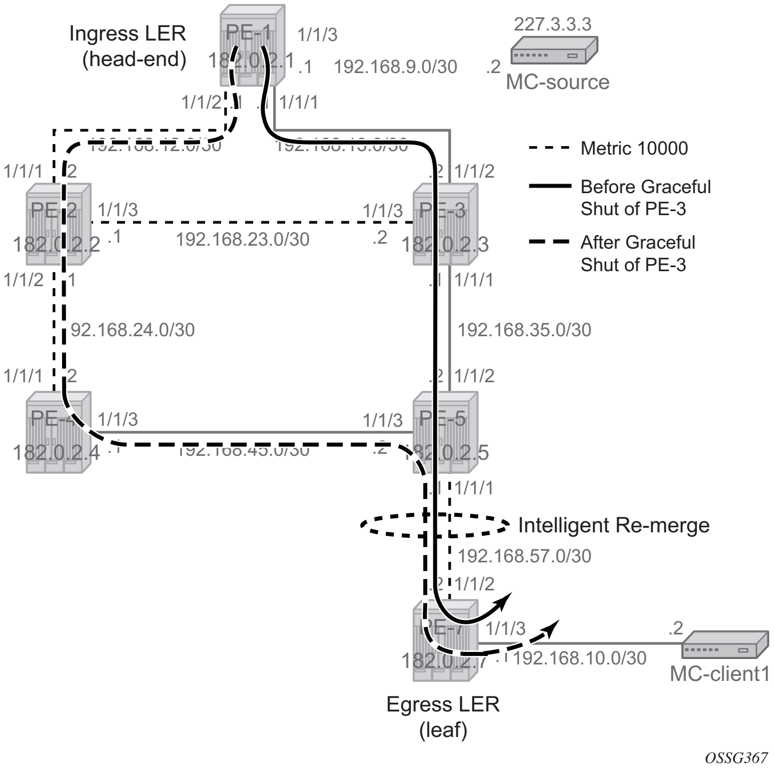
```
*A:PE-5# monitor port 1/1/1 1/1/2 1/1/3 rate interval 3 repeat 100

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------

---snip---
-------------------------------------------------------------------------------
At time t = 18 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                                21                1058955
Packets                                                0                    697
Errors                                                 0                      0
Bits                                                 168                8471640
Utilization (% of port capacity)                   ~0.00                   0.08

Port 1/1/2
-------------------------------------------------------------------------------
Octets                                           1059461                     50
Packets                                              697                      0
Errors                                                 0                      0
Bits                                             8475688                    400
Utilization (% of port capacity)                    0.08                  ~0.00

Port 1/1/3
-------------------------------------------------------------------------------
Octets                                           1059021                     21
Packets                                              697                      0
Errors                                                 0                      0
Bits                                             8472168                    168
Utilization (% of port capacity)                    0.08                  ~0.00
---snip---
```

Two incoming multicast streams are seen at PE-5 node (**port 1/1/2** and **port 1/1/3**) and only one outgoing multicast stream (**port 1/1/1**) is sent. No traffic duplication is seen.

Case 2

Figure 298 shows two paths of the same S2L that have ILMs on different incoming ports and go out on the same NHLFE. This is the case when we perform make-before-break (MBB) on an S2L path due to graceful shutdown or global revertive. This is only a temporary situation because the original path will be torn down.

Figure 298   **Intelligent Re-merge, Case 2**



For this test, only one multicast client will be looked at (the one connected to leaf node PE-7). On nodes PE-4 and PE-7, the port to PE-6 will be shut down to isolate PE-6. On the head-end node PE-1, a new P2MP LSP *LSP-p2mp-3* will be created with one loose MPLS path "loose" and keyword **cspf use-te-metric** to ensure that CSPF will use the TE metric instead of the IGP metric. In this case, the TE metric has the same value on all MPLS interfaces and the path has the hops PE-1, PE-3, PE-5, and PE-7. Also in this case, intelligent re-merge is performed at node PE-5.

```
*A:PE-1# configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-p2mp-3" p2mp-lsp
                cspf use-te-metric
                primary-p2mp-instance "p-LSP-p2mp-3"
```

```
                    s2l-path "loose" to 192.0.2.7
                    exit
                exit
                no shutdown
            exit
            no shutdown
```

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-3" p2mp-instance "p-LSP-p2mp-3" s2l loose
to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-3 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                          # - Detour In Use
    b - Bandwidth Protected                       n - Node Protected
    S - Strict                                    L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-3
S2L LSP ID       : 55810
P2MP ID          : 0                    S2L Grp Id           : 2
Admin State      : Up                   Oper State           : Up
S2L State:       : Active                                    :
S2L Name         : loose
To               : 192.0.2.7
S2L Admin        : Up                   S2L Oper             : Up
OutInterface     : 1/1/2                Out Label            : 524284
S2L Up Time      : 0d 00:04:56          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 2                    CSPF Queries         : 2
Failure Code     : noError              Failure Node         : n/a
Inter-area  : False
ExplicitHops:
    No Hops Specified
Actual Hops :
    192.168.13.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3)                     Record Label        : 524284
 -> 192.168.35.2 (192.0.2.5)                     Record Label        : 524281
 -> 192.168.57.2 (192.0.2.7)                     Record Label        : 524283
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.57.2(S)
LastResignal: n/a
===============================================================================
*A:PE-1#
```

In a normal situation, the P2MP LSP would follow the nodes PE-1, PE-3, PE-5, and PE-7. This can be verified with multicast traffic. Therefore, a new tunnel interface and a new static <S,G> IGMP join will be configured on head-end node PE-1 and leaf node PE-7. On the leaf node, an extension to the existing multicast information policy is needed. This is configured as follows:

```
*A:PE-1# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-3
```

```
*A:PE-1# configure router igmp
            tunnel-interface rsvp-p2mp "LSP-p2mp-3"
                static
                    group 227.3.3.3
                        source 192.168.9.2
                    exit
                exit
            exit
```

```
*A:PE-7# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-3" sender 192.0.2.1
```

```
*A:PE-7# configure router igmp
            interface "int-PE-7-MC-client1"
                static
                    group 227.3.3.3
                        source 192.168.9.2
                    exit
                exit
            exit
```

```
*A:PE-7# configure mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle3" create
                primary-tunnel-interface rsvp-p2mp LSP-p2mp-3 sender 192.0.2.1
                channel 227.3.3.3 create
                exit
            exit
        exit
```

The traffic on PE-5 is monitored. Under normal circumstances, the ingress port is 1/1/2 and the egress port is 1/1/1.

```
*A:PE-5# monitor port 1/1/1 1/1/2 1/1/3 rate interval 3 repeat 999

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------

---snip---
-------------------------------------------------------------------------------
At time t = 21 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                                21                1059461
Packets                                                0                    697
Errors                                                 0                      0
Bits                                                 168                8475688
Utilization (% of port capacity)                   ~0.00                   0.08

Port 1/1/2
```

```
-------------------------------------------------------------------------------
Octets                                           1059042                     21
Packets                                              697                      0
Errors                                                 0                      0
Bits                                             8472336                    168
Utilization (% of port capacity)                    0.08                  ~0.00

Port 1/1/3
-------------------------------------------------------------------------------
Octets                                                50                     21
Packets                                                0                      0
Errors                                                 0                      0
Bits                                                 400                    168
Utilization (% of port capacity)                   ~0.00                  ~0.00

-------------------------------------------------------------------------------
```

An RSVP graceful shutdown is performed on node PE-3, as follows:

```
*A:PE-3# configure router rsvp graceful-shutdown
```

Global revertive is triggered on head-end node PE-1. A new MPLS path will be calculated (see the dashed line in Figure 298). For a few seconds or even less than a second, the old path and new path are active (two incoming multicast streams on node PE-5). Node PE-5 is doing intelligent re-merge, not sending duplicate multicast traffic downstream toward PE-7:

```
*A:PE-5# monitor port 1/1/1 1/1/2 1/1/3 rate interval 3 repeat 100

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---

-------------------------------------------------------------------------------
At time t = 30 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                                82                1059685
Packets                                                1                    698
Errors                                                 0                      0
Bits                                                 656                8477480
Utilization (% of port capacity)                   ~0.00                   0.08

Port 1/1/2
-------------------------------------------------------------------------------
Octets                                           1059711                     21
Packets                                              699                      0
Errors                                                 0                      0
Bits                                             8477688                    168
Utilization (% of port capacity)                    0.08                  ~0.00

Port 1/1/3
```

```
-------------------------------------------------------------------------------
Octets                                               259534                 283
Packets                                                 172                   2
Errors                                                    0                   0
Bits                                                2076272                2264
Utilization (% of port capacity)                       0.02               ~0.00

===============================================================================
```

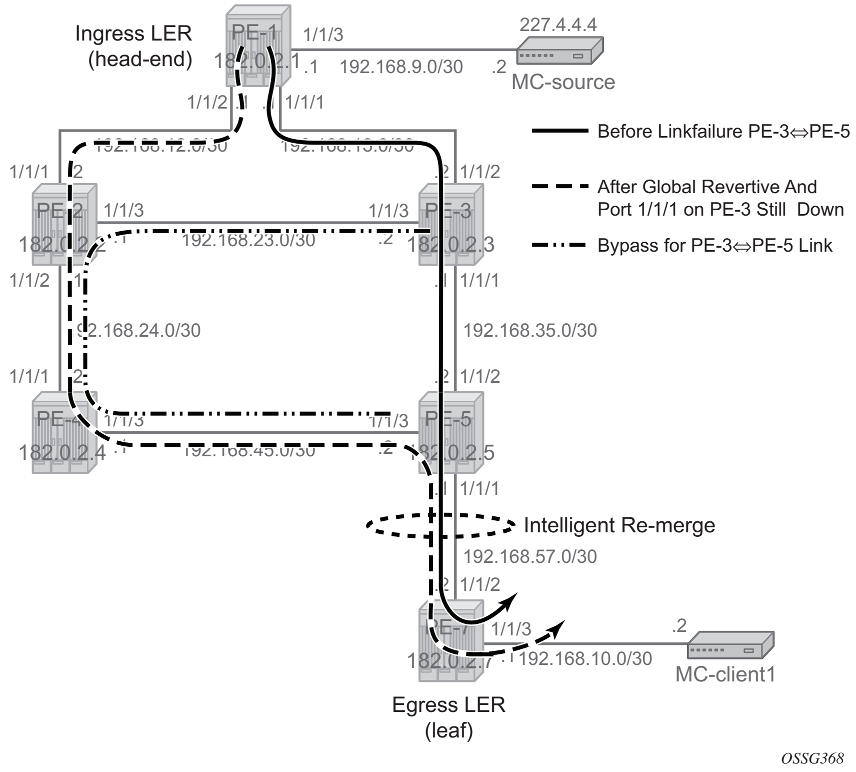
The granularity of the monitoring command is 3 seconds. The graceful shutdown takes less than 3 seconds. However, we can clearly see that the number of outgoing packets on port 1/1/1 equals the number of incoming packets on port 1/1/2. A number of these incoming packets also arrived on port 1/1/3, but no duplicate packets were sent on the outgoing port. No traffic duplication is seen.

Case 3

When a bypass is active on the S2L path and the new global revertive path of the same S2L arrives on the same incoming interface as the original path (interface flapped) at the FRR merge point node, the node will do intelligent re-merge. The implementation recognizes this specific case and will signal a different label from the original S2L path coming on that same interface.

Figure 299 shows the initial S2L path before the link failure (solid line), the bypass path to protect the link between PE-3 and PE-5, and the new global revertive path of the S2L (dotted line). Both the bypass path and the global revertive path share the links between PE-2 and PE-4 and between PE-4 and PE-5.

Figure 299 **Intelligent Re-merge, Case 3**



For this test, all the non-default OSPF metrics are removed from the interfaces. Only one MC-client will be looked at (the one connected to leaf node PE-7). On nodes PE-4 and PE-7, the port toward PE-6 will be shut down to isolate PE-6. On the head-end node PE-1, a new P2MP LSP “LSP-p2mp-4” will be created with one loose MPLS path “loose” and FRR enabled. Also in this case, intelligent re-merge is performed at node PE-5.

```
*A:PE-1# configure
    router
        mpls
            path "loose"
                no shutdown
            exit
            lsp "LSP-p2mp-4" p2mp-lsp
                cspf
                fast-reroute facility
                    no node-protect
                exit
                primary-p2mp-instance "p-LSP-p2mp-4"
                    s2l-path "loose" to 192.0.2.7
                    exit
                exit
                no shutdown
            exit
            no shutdown

*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-4" p2mp-instance "p-LSP-p2mp-4" s2l loose
to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-4 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                             # - Detour In Use
    b - Bandwidth Protected                          n - Node Protected
    S - Strict                                       L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-4
S2L LSP ID       : 54784
P2MP ID          : 0                    S2L Grp Id           : 1
Admin State      : Up                   Oper State           : Up
S2L State:       : Active                                    :
S2L Name         : loose
To               : 192.0.2.7
S2L Admin        : Up                   S2L Oper             : Up
OutInterface     : 1/1/1                Out Label            : 524283
S2L Up Time      : 0d 00:01:00          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 1                    CSPF Queries         : 1
Failure Code     : noError              Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.13.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @                   Record Label        : 524283
 -> 192.168.35.2 (192.0.2.5)                     Record Label        : 524285
 -> 192.168.57.2 (192.0.2.7)                     Record Label        : 524287
ComputedHops     :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.57.2(S)
LastResignal     : n/a
===============================================================================
*A:PE-1#
```

In the normal situation, the P2MP LSP follows the nodes PE-1, PE-3, PE-5, and PE-7. This can be verified with multicast traffic. Therefore, a new tunnel interface and a new static <S,G> IGMP join will be configured on head-end node PE-1 and leaf node PE-7. On the leaf node, an extension to the existing multicast information policy is needed. This is configured as follows:

```
*A:PE-1# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-4"


*A:PE-1# configure router igmp
            tunnel-interface rsvp-p2mp LSP-p2mp-4
                static
```

```
                    group 227.4.4.4
                        source 192.168.9.2
                    exit
                exit
            exit

*A:PE-7# configure router pim tunnel-interface rsvp-p2mp "LSP-p2mp-4" sender 192.0.2.1

*A:PE-7# configure router igmp
            interface "int-PE-7-MC-client1"
                static
                    group 227.4.4.4
                        source 192.168.9.2
                    exit
                exit
            exit

*A:PE-7# configure mcast-management
        multicast-info-policy "p2mp-pol" create
            bundle "bundle4" create
                primary-tunnel-interface rsvp-p2mp LSP-p2mp-4 sender 192.0.2.1
                channel 227.4.4.4 create
                exit
            exit
        exit
```

When the initial path is taken, the incoming traffic arrives at port 1/1/2 on PE-5 and is forwarded to port 1/1/1 to PE-7, as follows.

```
*A:PE-5# monitor port 1/1/1 1/1/2 1/1/3 rate interval 3 repeat 999

===============================================================================
Monitor statistics for Ports
===============================================================================
                                                   Input                 Output
-------------------------------------------------------------------------------
---snip---
-------------------------------------------------------------------------------
At time t = 3 sec (Mode: Rate)
-------------------------------------------------------------------------------
Port 1/1/1
-------------------------------------------------------------------------------
Octets                                                67                1056556
Packets                                                1                    696
Errors                                                 0                      0
Bits                                                 536                8452448
Utilization (% of port capacity)                   ~0.00                   0.08

Port 1/1/2
-------------------------------------------------------------------------------
Octets                                           1056421                    219
Packets                                              695                      1
Errors                                                 0                      0
Bits                                             8451368                   1752
Utilization (% of port capacity)                    0.08                  ~0.00
```

```
Port 1/1/3
--------------------------------------------------------------------------------
Octets                                                185                     67
Packets                                                 1                      1
Errors                                                  0                      0
Bits                                                 1480                    536
Utilization (% of port capacity)                    ~0.00                  ~0.00
```

Now a link failure on the interface from PE-3 to PE-5 is emulated as follows:

```
*A:PE-3# configure port 1/1/1 shutdown
```

As a consequence of this, traffic will be flowing over the bypass link (see Figure 299 and note the '#' symbol in the following **show** command, as well as the failure code).

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-4" p2mp-instance "p-LSP-p2mp-4" s2l loose
to 192.0.2.7 detail

================================================================================
MPLS LSP LSP-p2mp-4 S2L loose (Detail)
================================================================================
Legend :
    @ - Detour Available                           # - Detour In Use
    b - Bandwidth Protected                        n - Node Protected
    S - Strict                                     L - Loose
    A - ABR
    s - Soft Preemption
================================================================================
LSP Name          : LSP-p2mp-4
S2L LSP ID        : 54784
P2MP ID           : 0                    S2L Grp Id          : 1
Admin State       : Up                   Oper State          : Up
S2L State:        : Active                                   :
S2L Name          : loose
To                : 192.0.2.7
S2L Admin         : Up                   S2L Oper            : Up
OutInterface      : 1/1/1                Out Label           : 524283
S2L Up Time       : 0d 00:17:13          S2L Dn Time         : 0d 00:00:00
RetryAttempt      : 0                    NextRetryIn         : 0 sec
S2L Trans         : 1                    CSPF Queries        : 1
Failure Code      : tunnelLocallyRepaire Failure Node        : 192.0.2.3
                    d
Inter-area        : False
ExplicitHops      :
    No Hops Specified
Actual Hops       :
    192.168.13.1 (192.0.2.1) @                    Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3) @ #                  Record Label        : 524283
 -> 192.168.35.2 (192.0.2.5)                      Record Label        : 524285
 -> 192.168.57.2 (192.0.2.7)                      Record Label        : 524287
ComputedHops      :
    192.168.13.1(S)
 -> 192.168.13.2(S)
 -> 192.168.35.2(S)
 -> 192.168.57.2(S)
LastResignal      : n/a
In Prog MBB :
 MBB Type        : GlobalRevert         NextRetryIn          : 9 sec
 Started At      : 09/10/2018 13:08:10  RetryAttempt         : 0
 FailureCode     : noError              Failure Node         : n/a
===============================================================================
*A:PE-1#
```

In the meantime, PE-3 will trigger a global revertive action (sending PathErr message) toward the head-end node PE-1.

```
*A:PE-1# show router mpls p2mp-lsp "LSP-p2mp-4" p2mp-instance "p-LSP-p2mp-4" s2l loose
to 192.0.2.7 detail

===============================================================================
MPLS LSP LSP-p2mp-4 S2L loose (Detail)
===============================================================================
Legend :
    @ - Detour Available                          # - Detour In Use
    b - Bandwidth Protected                       n - Node Protected
    S - Strict                                    L - Loose
    A - ABR
    s - Soft Preemption
===============================================================================
LSP Name         : LSP-p2mp-4
S2L LSP ID       : 54784
P2MP ID          : 0                    S2L Grp Id           : 2
Admin State      : Up                   Oper State           : Up
S2L State:       : Active                                    :
S2L Name         : loose
To               : 192.0.2.7
S2L Admin        : Up                   S2L Oper             : Up
OutInterface     : 1/1/2                Out Label            : 524283
S2L Up Time      : 0d 00:17:33          S2L Dn Time          : 0d 00:00:00
RetryAttempt     : 0                    NextRetryIn          : 0 sec
S2L Trans        : 2                    CSPF Queries         : 2
Failure Code     : noError              Failure Node         : n/a
Inter-area       : False
ExplicitHops     :
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 524283
 -> 192.168.24.2 (192.0.2.4)                     Record Label        : 524284
 -> 192.168.45.2 (192.0.2.5)                     Record Label        : 524283
 -> 192.168.57.2 (192.0.2.7)                     Record Label        : 524287
ComputedHops     :
    192.168.12.1(S)
 -> 192.168.12.2(S)
 -> 192.168.24.2(S)
 -> 192.168.45.2(S)
 -> 192.168.57.2(S)
LastResignal     : n/a
Last MBB    :
 MBB Type        : GlobalRevert         MBB State            : Success
 Ended At        : 09/10/2018 13:08:4
===============================================================================
*A:PE-1#
```

For a short time, PE-5 will receive two incoming MC streams (both arriving on port 1/1/3). One from the bypass path (PE-3 => PE-2 => PE-4 => PE-5) and one from the new MPLS path (PE-1 => PE-2 => PE-4 => PE-5 => PE-7). Port 1/1/1 on PE-5 performs intelligent remerge, so only one MC stream is sent downstream toward leaf node PE-7.

Conclusion

From a configuration point of view, a P2MP LSP is only configured on the head-end node of that P2MP LSP; no explicit configuration is needed on the transit LSRs, branch LSRs, bud LSRs, and egress LERs/leaf nodes.

Because the PIM protocol is only needed on the head-end node and the leaf nodes, we can work in a PIM-free core network. Although convergence is not covered in this chapter, failures in the core will be resolved by MPLS (in case of FRR, traffic loss for less than 50ms is expected). This is a major improvement compared to PIM convergence.

Seamless MPLS: Isolated IGP/LDP Domains and Labeled BGP

This chapter provides information about Seamless MPLS: Isolated IGP/LDP domains and Labeled BGP.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS routers and was initially written for SR OS Release 13.0.R7. The CLI in this edition is based on SR OS Release 15.0.R1.

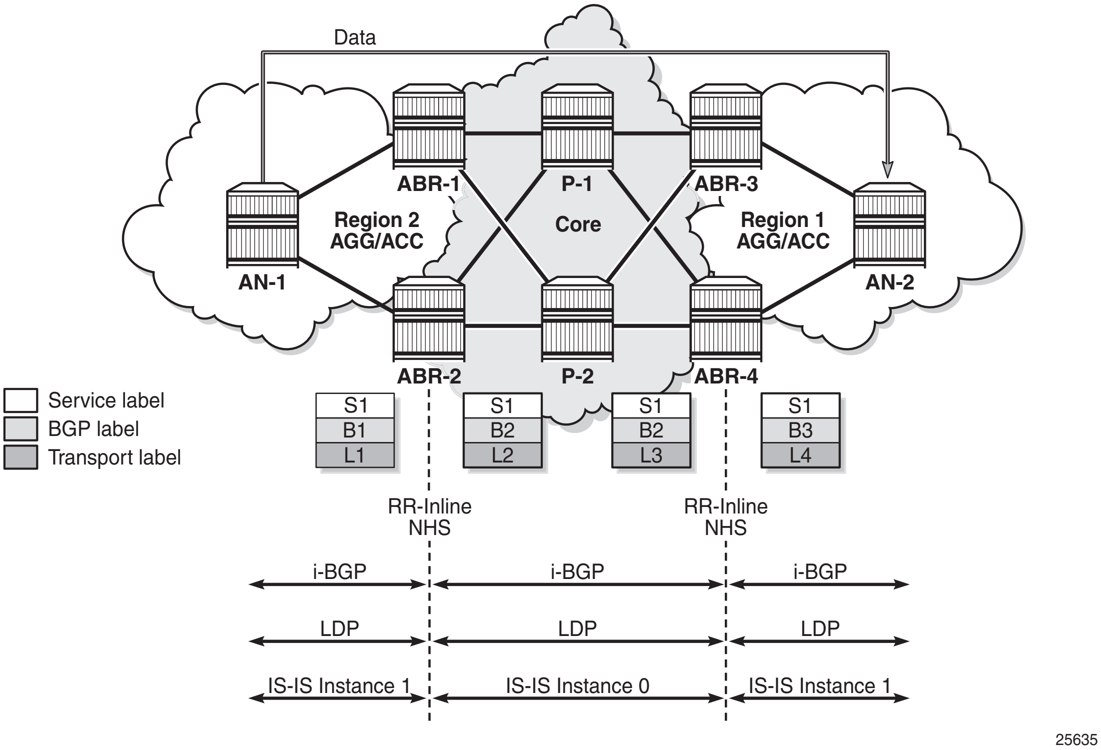
Overview

Seamless Multi-Protocol Label Switching (MPLS) is a network architecture that extends MPLS networks to integrate access and aggregation networks into a single MPLS domain, to solve the scaling problems in flat MPLS-based deployments. The Seamless MPLS transport concept described in this chapter partitions the core, aggregation, and access networks into isolated IGP/LDP domains. Seamless MPLS does not define any new protocols or technologies and is based on existing and well-known ones. Seamless MPLS provides end-to-end service-independent transport, separating the service and transport plane. Therefore, it removes the need for service-specific configurations in network transport nodes. Service provisioning is restricted only at the points of the network where it is required.

When BGP is used to distribute a route, it can also distribute an MPLS label that is mapped to that route. The label mapping information is appended to the BGP update message that is used to distribute the route. This is described in RFC 3701, *Carrying Label Information in BGPv4*.

Figure 300 shows a network with a core area and regional areas. Figure 300 also shows the control plane used in this Seamless MPLS implementation. For simplification, the control plane is displayed from right to left and the data plane from left to right. In this example, LDP will be used as the underlying transport inside each IGP domain. Alternatively, RSVP-TE could be used.

Figure 300 **Seamless MPLS - Network Topology, Control and Data Plane**



In typical Seamless MPLS solutions, multiple ABRs are in place that result in some specific BGP configurations to send/receive multiple paths, such as the add-path feature. Due to this, ANs and ABRs will have several next hops for the same prefix, allowing the use of redundancy mechanisms such as BGP Prefix Independent Convergence (PIC) edge, also known as BGP Fast ReRoute (FRR). These mechanisms are beyond the scope of this chapter.

AN routers in a regional area learn the reachability of AN routers in other regional areas through BGP labeled routes redistributed by the local ABRs (RFC 3107).

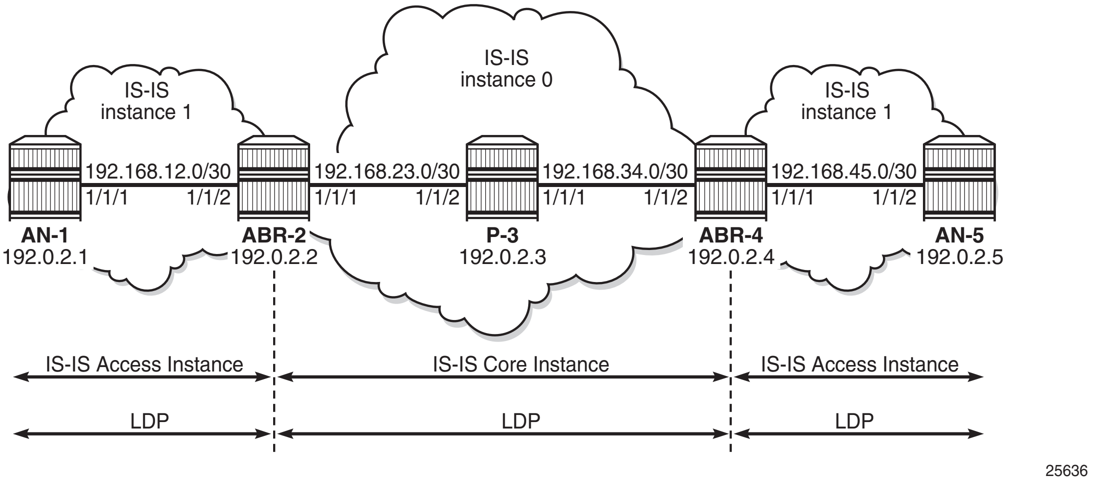
The label stack contains three labels for packets sent in a VPN service between the access nodes:

- The ANs push a service label to the packets sent in the VPN service. The service label remains unchanged end-to-end between ANs. The service label is popped by the remote AN and is the inner label of the label stack.
- The BGP label is the middle label of the label stack and should be regarded as a transport label. The transport label stack is increased to two labels: BGP and LDP transport labels. The BGP label is pushed by the iLER AN and is swapped at the BGP next hop, which can be one of the two local ABRs. Both ABRs are configured with next-hop-self. The BGP label is also swapped by the remote ABR.
- The iLER AN pushes an LDP transport label to the packets sent to the remote AN to reach the BGP next hop. At the local ABR, the LDP transport label is popped and a new LDP transport label is pushed to reach the BGP next hop (remote ABR). The LDP transport label is swapped in every label switching router (LSR) and popped by the ABR nearest to the remote AN. That ABR pops the LDP transport label, swaps the BGP label, and pushes an LDP transport label to reach the remote eLER AN.

Configuration

Figure 301 shows the example topology that is used in this chapter. An Epipe and VPRN will be established between the access nodes AN-1 and AN-5. In the regional areas, and in the core area, IS-IS L2 capability is used.

Figure 301 **Seamless MPLS - IGP/LDP domains**



Initial Configuration

All nodes have the following initial configuration:

- Cards, media dependent adapters (MDAs), ports
  - Router interfaces:

```
*A:ABR-2# configure
    router
        interface "int-ABR-2-AN-1"
            address 192.168.12.2/30
            port 1/1/2
        exit
        interface "int-ABR-2-P-3"
            address 192.168.23.1/30
            port 1/1/1
        exit
        interface "system"
            address 192.0.2.2/32
        exit
```

- IS-IS (alternatively, OSPF could be used). Core area and regional areas run an isolated IS-IS instance. ABRs run two IS-IS instances: instance 0 belongs to the core and instance 1 belongs to the access network.
  - **Core instance**. All ABRs and Ps have level 2 (L2) capability, as follows:

```
*A:ABR-2# configure
    router
        isis 0
            level-capability level-2
            area-id 49.0001
            interface "system"
            exit
            interface "int-ABR-2-P-3"
                interface-type point-to-point
            exit
            no shutdown
        exit
```

  - **Access instance**. All ABRs and ANs have L2 capability, as follows:

```
*A:ABR-2# configure
    router
        isis 1
            level-capability level-2
            interface "system"
            exit
            interface "int-ABR-2-AN-1"
                interface-type point-to-point
            exit
            no shutdown
        exit
```

- LDP

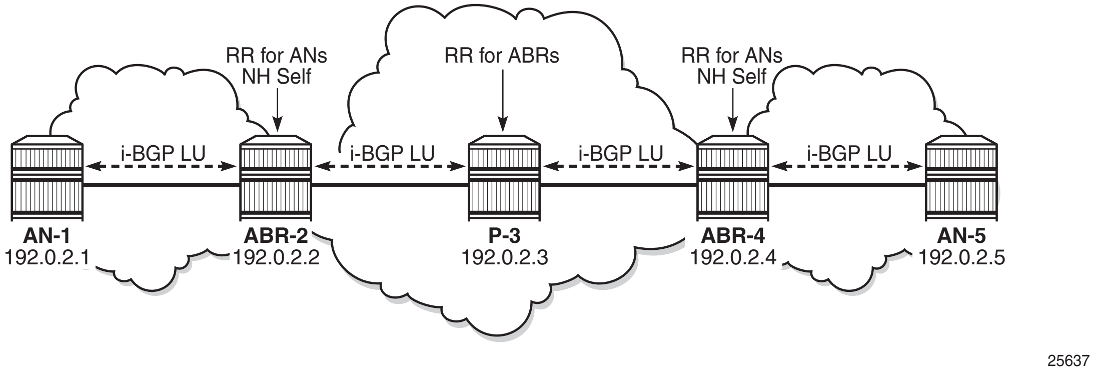
Link LDP is enabled on all router interfaces on all nodes, as follows:

```
*A:ABR-2# configure
    router
        ldp
            interface-parameters
                interface "int-ABR-2-AN-1"
                exit
                interface "int-ABR-2-P-3"
                exit
            exit
        exit
```

Configure BGP

BGP is configured on all ABRs and all ANs. P-3 acts as a core Route Reflector (RR). To allow for separation of core/access IGP domains, the ABRs become RRs inline and implement next-hop-self on labeled IPv4 BGP prefixes. Figure 302 shows the exchange of iBGP Labeled Unicast (LU) routes.

Figure 302 **Seamless MPLS - BGP**



BGP configuration on ABRs

There are two BGP groups on the ABRs: one group toward the core RR and another group toward the AN, as follows:

```
*A:ABR-2# configure
    router
        autonomous-system 64496
        bgp
            group "core"
                family vpn-ipv4 label-ipv4
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.3
                    description "coreRR_P-3"
                    next-hop-self
                exit
            exit
            no shutdown
```

Advertise-inactive must be enabled on the BGP group toward the core. The /32 system IP addresses, learned in labeled BGP, will also be learned in IS-IS. Because IS-IS has a lower preference compared to iBGP, the IS-IS routes will be installed in the routing table. BGP default behavior only advertises those prefixes that were elected by RTM and used. The VPNv4 address family is also included, along with labeled IPv4, to allow setting up L3 VPN services, as shown in next sections. The next-hop attribute of VPNv4 prefixes remains unchanged.

```
*A:ABR-2# configure
    router
        bgp
            group "ANs_Label_IPv4+VPNv4"
                family vpn-ipv4 label-ipv4
                cluster 2.2.2.2
                peer-as 64496
                neighbor 192.0.2.1
                    description "AN-1"
                    next-hop-self
                exit
            exit
            no shutdown
```

BGP configuration on the core RR

```
*A:P-3# configure
    router
        autonomous-system 64496
        bgp
            group "core"
                family vpn-ipv4 label-ipv4
                cluster 3.3.3.3
                peer-as 64496
                advertise-inactive
                neighbor 192.0.2.2
                    description "ABR-2"
                exit
                neighbor 192.0.2.4
                    description "ABR-4"
                exit
            exit
            no shutdown
```

BGP configuration on ANs toward ABRs

```
*A:AN-1# configure
    router
        autonomous-system 64496
        bgp
            group "ABRs_Label_IPv4+VPNv4"
                family vpn-ipv4 label-ipv4
                peer-as 64496
                neighbor 192.0.2.2
                exit
            exit
            no shutdown
        exit
```

Configuring address family **label-ipv4** implies that all IPv4 prefixes advertised will be sent to the remote BGP peer as an RFC 3107 formatted label. The next-hop-self command only applies to labeled IPv4 prefixes, not to VPN-IPv4.

The BGP sessions can be shown with the following command:

```
*A:P-3# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
-------------------------------------------------------------------------------
192.0.2.2
ABR-2
Def. Instance  64496        3    0 00h00m26s 0/0/0 (VpnIPv4)
                            3    0           0/0/0 (Lbl-IPv4)
192.0.2.4
ABR-4
Def. Instance  64496        4    0 00h00m20s 0/0/0 (VpnIPv4)
                            3    0           0/0/0 (Lbl-IPv4)

-------------------------------------------------------------------------------
*A:P-3#


*A:AN-1# show router bgp summary all

===============================================================================
BGP Summary
===============================================================================
Legend : D - Dynamic Neighbor
===============================================================================
Neighbor
Description
                   AS PktRcvd InQ  Up/Down   State|Rcv/Act/Sent (Addr Family)
                      PktSent OutQ
```

```
-------------------------------------------------------------------------------
192.0.2.2
Def. Instance  64496        4    0 00h00m31s 0/0/0 (VpnIPv4)
                            4    0           0/0/0 (Lbl-IPv4)
-------------------------------------------------------------------------------
*A:AN-1#
```

Export Policies for BGP

A policy is required on the ANs to advertise the system IP address in labeled BGP toward the ABRs. The same policy is required on the ABRs to advertise their system IP address in labeled BGP toward the core and the AN.

Policy configuration on ANs and ABRs

```
configure
    router
        policy-options
            begin
            prefix-list "system"
                prefix 192.0.2.1/32 exact
            exit
            policy-statement "export-system"
                entry 10
                    from
                        protocol direct
                        prefix-list "system"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

This export policy must be applied in the BGP context on AN-1: either in the general settings or per group or per neighbor, as follows:

```
*A:AN-1# configure
    router
        bgp
            group "ABRs_Label_IPv4+VPNv4"
                export "export-system"
            exit
        exit
```

The same export policy is applied in the group "core" on ABR-2, as follows:

```
*A:ABR-2# configure
```

```
    router
        bgp
            group "core"
                export "export-system"
            exit
        exit
```

A similar export policy is defined to export prefix 192.0.2.5 from AN-5 to ABR-4 and from ABR-4 to the RR in the core network, P-3.

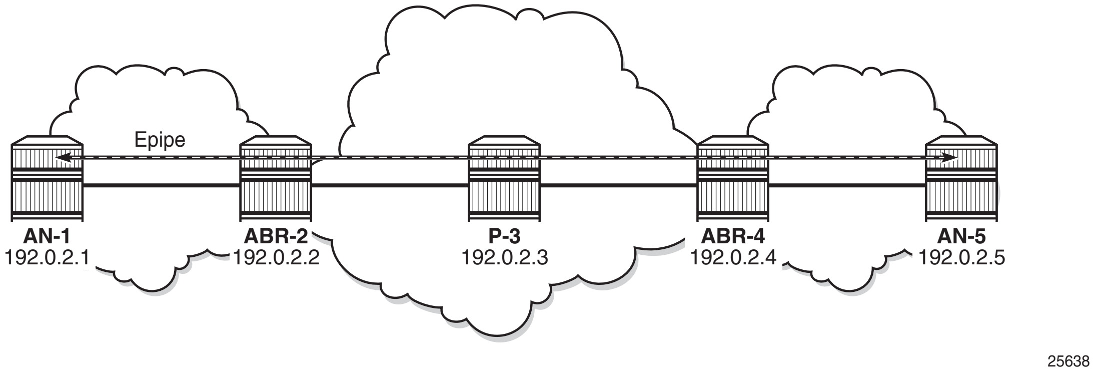
The prefix of the remote AN is added to the routing table in AN-1 and services can be configured in the ANs. No service configuration is required in the transit nodes.

```
*A:AN-1# show router route-table

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Local   Local     00h06m58s  0
       system                                                       0
192.0.2.2/32                                  Remote  ISIS(1)   00h06m12s  18
       192.168.12.2                                                 10
192.0.2.5/32                                  Remote  BGP       00h00m40s  170
       192.0.2.2 (tunneled)                                         0
192.168.12.0/30                               Local   Local     00h06m58s  0
       int-AN-1-ABR-2                                               0
-------------------------------------------------------------------------------
No. of Routes: 4
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:AN-1#
```

Configure SDP and Epipe

An end-to-end Epipe service is established between AN-1 and AN-5, as shown in Figure 303.

Figure 303 **End-to-End Epipe Service**



The SDP is configured on AN-1 and AN-5, as follows:

```
*A:AN-1# configure service
        sdp 15 mpls create
            far-end 192.0.2.5
            bgp-tunnel
            no shutdown
        exit
```

```
*A:AN-5# configure service
        sdp 51 mpls create
            far-end 192.0.2.1
            bgp-tunnel
            no shutdown
        exit
```

Epipe 1 is configured on AN-1 and AN-5, as follows:

```
*A:AN-1# configure service
        epipe 1 customer 1 create
            sap 1/1/3:1 create
            exit
            spoke-sdp 15:1 create
                no shutdown
            exit
            no shutdown
        exit
```

```
*A:AN-5# configure service
        epipe 1 customer 1 create
            sap 1/1/3:1 create
            exit
            spoke-sdp 51:1 create
                no shutdown
            exit
            no shutdown
        exit
```

The state of the SDP and of the Epipe service can be verified on AN-1, as follows:

```
*A:AN-1# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
15     0       1552    192.0.2.5        Up   Up          MPLS    B     TLDP
----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:AN-1#

*A:AN-1# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
---snip---

Admin State       : Up                  Oper State        : Up
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/3:1                              q-tag        1518    1518    Up   Up
sdp:15:1 S(192.0.2.5)                    Spok         0       1552    Up   Up
===============================================================================
*A:AN-1#
```

The state of the SDP and of the Epipe service can be verified on AN-5, as follows:

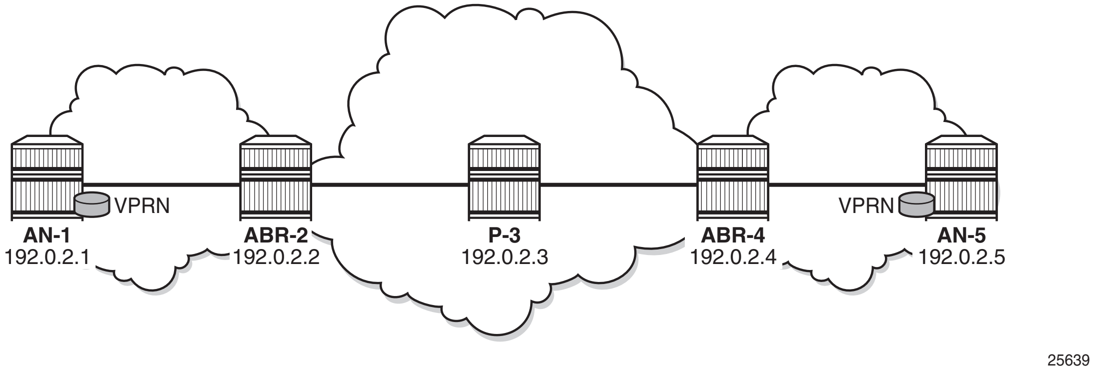
```
*A:AN-5# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
51     0       1552    192.0.2.1        Up   Up          MPLS    B     TLDP
----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:AN-5#
```

```
*A:AN-5# show service id 1 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 1                   Vpn Id            : 0
Service Type      : Epipe
---snip---

Admin State       : Up                  Oper State        : Up
---snip---

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/1/3:1                              q-tag        1518    1518    Up   Up
sdp:51:1 S(192.0.2.1)                    Spok         0       1552    Up   Up
===============================================================================
*A:AN-5#
```

Configure VPRN

An L3 VPN service is established on AN-1 and AN-5, as shown in Figure 304.

Figure 304 **L3 VPN service**



The VPRN service is configured on AN-1 and AN-5, as follows. For simplicity, no CEs are attached to the ANs and only one loopback is created for verification.

```
*A:AN-1# configure
    service
        vprn 2 customer 1 create
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:2
            interface "loopback" create
                address 192.0.1.1/32
                loopback
            exit
            no shutdown
        exit
```

```
*A:AN-5#  configure
    service
        vprn 2 customer 1 create
            route-distinguisher 64496:2
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:2
            interface "loopback" create
                address 192.0.1.5/32
                loopback
            exit
            no shutdown
        exit
```

The routing table for VPRN 2 contains the local and the remote loopback addresses. On AN-1, this can be verified as follows:

```
*A:AN-1# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
192.0.1.1/32                                  Local   Local     00h04m28s  0
       loopback                                                     0
192.0.1.5/32                                  Remote  BGP VPN   00h03m05s  170
       192.0.2.5 (tunneled:BGP)                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:AN-1#
```

On AN-5, this can be verified as follows:

```
*A:AN-5# show router 2 route-table

===============================================================================
Route Table (Service: 2)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
-------------------------------------------------------------------------------
```

```
192.0.1.1/32                                  Remote  BGP VPN   00h00m49s  170
       192.0.2.1 (tunneled:BGP)                                     0
192.0.1.5/32                                  Local   Local     00h01m38s  0
       loopback                                                     0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
*A:AN-5#
```

Ping messages can be sent from the loopback address in VPRN 2 on AN-1 to the remote loopback address in VPRN 2 on AN-5, as follows:

```
*A:AN-1# ping router 2 192.0.1.5
PING 192.0.1.5 56 data bytes
64 bytes from 192.0.1.5: icmp_seq=1 ttl=64 time=2.22ms.
64 bytes from 192.0.1.5: icmp_seq=2 ttl=64 time=2.11ms.
64 bytes from 192.0.1.5: icmp_seq=3 ttl=64 time=3.00ms.
64 bytes from 192.0.1.5: icmp_seq=4 ttl=64 time=2.14ms.
64 bytes from 192.0.1.5: icmp_seq=5 ttl=64 time=2.88ms.
---- 192.0.1.5 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 2.11ms, avg = 2.47ms, max = 3.00ms, stddev = 0.387ms
*A:AN-1#
```

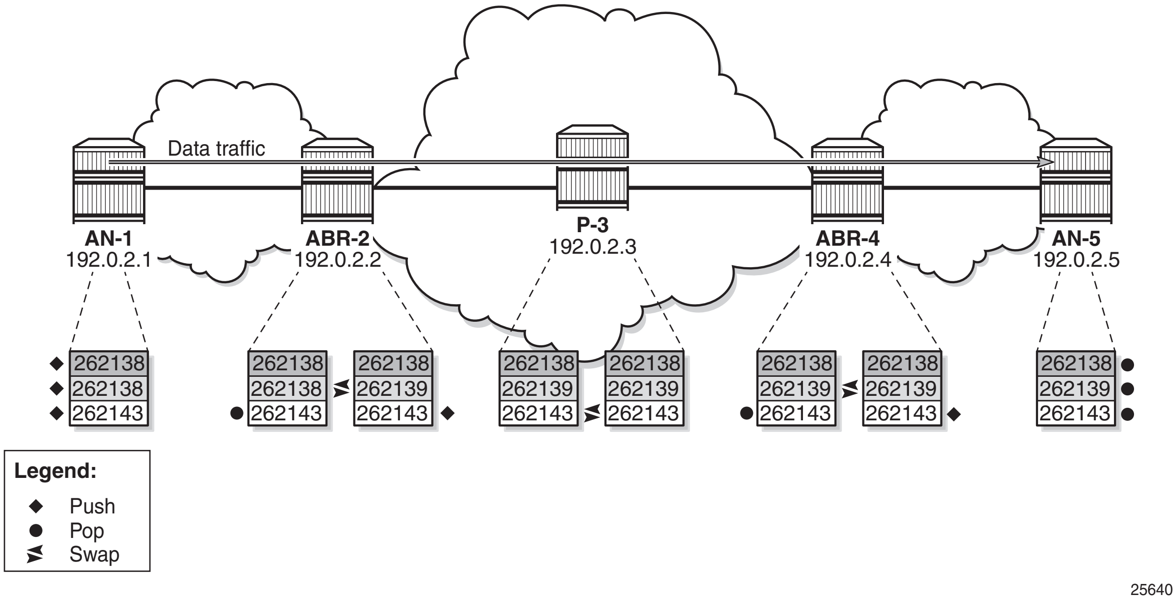
In a similar way, ping messages are sent from the loopback address in VPRN 2 on AN-5 to the loopback address in VPRN 2 on AN-1, as follows:

```
*A:AN-5# ping router 2 192.0.1.1
PING 192.0.1.1 56 data bytes
64 bytes from 192.0.1.1: icmp_seq=1 ttl=64 time=2.10ms.
64 bytes from 192.0.1.1: icmp_seq=2 ttl=64 time=1.94ms.
64 bytes from 192.0.1.1: icmp_seq=3 ttl=64 time=1.91ms.
64 bytes from 192.0.1.1: icmp_seq=4 ttl=64 time=2.11ms.
64 bytes from 192.0.1.1: icmp_seq=5 ttl=64 time=3.10ms.
---- 192.0.1.1 PING Statistics ----
5 packets transmitted, 5 packets received, 0.00% packet loss
round-trip min = 1.91ms, avg = 2.23ms, max = 3.10ms, stddev = 0.441ms
*A:AN-5#
```

Data Plane Overview

Figure 305 shows the label stacks used for traffic from AN-1 to AN-5. As an example, an Epipe service is used.

Figure 305 **Label Stacks for Traffic from AN-1 to AN-5**



1. The service label added for Epipe 1 on AN-1 for egress traffic to AN-5 is 262138. Ingress traffic on AN-1 has service label 262138. This can be shown as follows:

```
*A:AN-1# show service id 1 labels

===============================================================================
Martini Service Labels
===============================================================================
Svc Id     Sdp Binding        Type  I.Lbl                E.Lbl
-------------------------------------------------------------------------------
1          15:1               Spok  262138               262138
-------------------------------------------------------------------------------
Number of Bound SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:AN-1#
```

This service label remains unchanged end-to-end.

On AN-1, the (middle) BGP label for traffic with destination AN-5 is 262138, as follows:

```
*A:AN-1# show router bgp routes 192.0.2.5 label-ipv4

===============================================================================
 BGP Router ID:192.0.2.1        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       100         None
      192.0.2.2                                          None        262138
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:AN-1#
```

The next hop is ABR-2. AN-1 pushes the LDP label 262143 to reach ABR-2, as follows:

```
*A:AN-1# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.1)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Push              --      262143
192.168.12.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 1
===============================================================================
*A:AN-1#
```

2. At ABR-2, the service label 262138 remains unchanged. The LDP label 262143 is popped, as follows:

```
*A:ABR-2# show router ldp bindings active prefixes prefix 192.0.2.2/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.2/32                                Pop             262143      --
  --                                          --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 1
===============================================================================
*A:ABR-2#
```

On ABR-2, the BGP next hop is ABR-4 for prefix 192.0.2.5, as follows:

```
*A:ABR-2# show router bgp routes 192.0.2.5 label-ipv4

===============================================================================
 BGP Router ID:192.0.2.2        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete

===============================================================================
BGP IPv4 Routes
===============================================================================
Flag  Network                                            LocalPref   MED
      Nexthop (Router)                                   Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
u*>i  192.0.2.5/32                                       100         None
      192.0.2.4                                          None        262139
      No As-Path
-------------------------------------------------------------------------------
Routes : 1
===============================================================================
*A:ABR-2#
```

On ABR-2, the BGP label 262138 is swapped with 262139 for BGP next hop ABR-4, as follows:

```
*A:ABR-2# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.1                     262139         262139         Internal
192.0.2.4                     262139         262138         Internal
-------------------------------------------------------------------------------
Total Labels allocated:   2
===============================================================================
*A:ABR-2#
```

ABR-2 pushes a new LDP label (262140) to reach the BGP next hop (ABR-4), as follows:

```
*A:ABR-2# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.2)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push              --      262140
192.168.23.2                                1/1/1

192.0.2.4/32                                Swap            262140    262140
192.168.23.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:ABR-2#
```

3. At LSR P-3, only an LDP label swap occurs. P-3 swaps LDP label 262140 with 262143, as follows:

```
*A:P-3# show router ldp bindings active prefixes prefix 192.0.2.4/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.4/32                                Push              --      262143
192.168.34.2                                1/1/1

192.0.2.4/32                                Swap            262140    262143
192.168.34.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:P-3#
```

4. At ABR-4, LDP label 262143 is popped and BGP label 262139 is swapped to the same label 262139, as follows:

```
*A:ABR-4# show router bgp inter-as-label
```

```
===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.2                     262139         262138         Internal
192.0.2.5                     262139         262139         Internal
-------------------------------------------------------------------------------
Total Labels allocated:   2
===============================================================================
*A:ABR-4#
```

ABR-4 pushes a new LDP label 262143 to reach AN-5, as follows:

```
*A:ABR-4# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Push              --      262143
192.168.45.2                                1/1/1

192.0.2.5/32                                Swap            262140    262143
192.168.45.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:ABR-4#
```

5. Finally, at AN-5, all labels in the stack are popped. The LDP label 262143 is popped as follows:

```
*A:AN-5# show router ldp bindings active prefixes prefix 192.0.2.5/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.5)
             (IPv6 LSR ID ::)
===============================================================================
---snip---
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.5/32                                Pop             262143      --
  --                                          --
```

```
-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 1
===============================================================================
*A:AN-5#
```

The BGP label 262139 is popped.

```
*A:AN-5# show router bgp inter-as-label

===============================================================================
BGP Inter-AS labels
Flags: B - entry has backup, P - entry is promoted
===============================================================================
NextHop                       Received       Advertised     Label
                              Label          Label          Origin
-------------------------------------------------------------------------------
192.0.2.5                     0              262139         Edge
-------------------------------------------------------------------------------
Total Labels allocated:   1
===============================================================================
*A:AN-5#
```

The ingress service label 262138 is popped, as follows:

```
*A:AN-5# show service id 1 labels
===============================================================================
Martini Service Labels
===============================================================================
Svc Id     Sdp Binding        Type  I.Lbl                 E.Lbl
-------------------------------------------------------------------------------
1          51:1               Spok  262138                262138
-------------------------------------------------------------------------------
Number of Bound SDPs : 1
-------------------------------------------------------------------------------
===============================================================================
*A:AN-5#
```

OAM

The following Operations, Administration, and Maintenance (OAM) commands can be launched to validate reachability between regions using BGP labeled IPv4 routes.

```
*A:AN-1# oam lsp-ping bgp-label prefix 192.0.2.5/32
LSP-PING 192.0.2.5/32: 80 bytes MPLS payload
Seq=1, send from intf int-AN-1-ABR-2, reply from 192.0.2.5
       udp-data-len=32 ttl=255 rtt=2.11ms rc=3 (EgressRtr)

---- LSP 192.0.2.5/32 PING Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 2.11ms, avg = 2.11ms, max = 2.11ms, stddev = 0.000ms


*A:AN-5# oam lsp-ping bgp-label prefix 192.0.2.1/32
LSP-PING 192.0.2.1/32: 80 bytes MPLS payload
```

```
Seq=1, send from intf int-AN-5-ABR-4, reply from 192.0.2.1
       udp-data-len=32 ttl=255 rtt=2.21ms rc=3 (EgressRtr)

---- LSP 192.0.2.1/32 PING Statistics ----
1 packets sent, 1 packets received, 0.00% packet loss
round-trip min = 2.21ms, avg = 2.21ms, max = 2.21ms, stddev = 0.000ms
```

In a similar way, LSP trace can validate the reachability of the remote AN, as follows:

```
*A:AN-1# oam lsp-trace bgp-label prefix 192.0.2.5/32
lsp-trace to 192.0.2.5/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.2  rtt=0.666ms rc=8(DSRtrMatchLabel) rsc=1
2  192.0.2.4  rtt=1.78ms rc=8(DSRtrMatchLabel)
3  192.0.2.5  rtt=2.09ms rc=3(EgressRtr) rsc=1
```

```
*A:AN-5# oam lsp-trace bgp-label prefix 192.0.2.1/32
lsp-trace to 192.0.2.1/32: 0 hops min, 0 hops max, 104 byte packets
1  192.0.2.4  rtt=0.631ms rc=8(DSRtrMatchLabel) rsc=1
2  192.0.2.2  rtt=1.48ms rc=8(DSRtrMatchLabel)
3  192.0.2.1  rtt=2.06ms rc=3(EgressRtr) rsc=1
```

Conclusion

Seamless MPLS helps to solve the scalability problems of large networks. Seamless MPLS partitions the core, aggregation, and access networks into isolated IGP/LDP domains, which helps to maintain IGP databases small and controlled. Label BGP allows the establishment of hierarchical LSPs for end-to-end service set up.

Segment Routing – Traffic Engineered Tunnels

This chapter provides information about Segment Routing – Traffic Engineered Tunnels.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter was initially written for SR OS Release 14.0.R7, but the CLI in the current edition corresponds to SR OS Release 15.0.R2.

Overview

Segment Routing (SR) is described in the chapter Segment Routing with IS-IS Control Plane, where the advertisement of node prefix segment identifiers (SIDs) cause the automatic creation of ECMP-aware shortest path MPLS tunnels on each SR-aware router. Each node prefix SID is a globally unique value and becomes an MPLS label in the MPLS data plane. The label is advertised and learned by each SR-capable router using control plane extensions to the IS-IS and OSPF protocols.

It is also possible to create source-routed traffic-engineered end-to-end segment routing paths, where routing constraints such as strict or loose hops can be used to determine a data path to be taken through a network.

These are known as Segment Routing Traffic Engineered (SR-TE) Label Switched Paths (LSPs) and use the same command line construct as that used in configuring RSVP-TE LSPs. However, SR-TE LSPs differ in that there is no mid-point state; each intermediate and tail-end router is unaware of the presence of the LSP because there is no signaling protocol used to create the path. The path can be computed locally by the ingress PE or by offloading the path computation to an external controller.

If a packet is forwarded through the SR tunnel, each router along the path will read the top label and forward the packet according to the SR tunnel table entry for that label.

This chapter describes the configuration of SR-TE LSPs with locally-computed source-routed paths and how they can be used in the data plane of Layer 2 and Layer 3 services. In the cases described, an SR-TE LSP containing a number of strict or loose hops is created at the head-end router and used to construct an LSP by translating the IP addresses configured in the MPLS path to an SID. This results in an MPLS path with state at the head end only, comprising a stack of SIDs, where each SID is an MPLS label.

In this chapter, OSPF is used to advertise the SIDs and a set of extensions to OSPF have been defined, which require additional configuration on each network router.

The LSP is instantiated—the state is operationally "up"—and a tunnel table entry is created that is owned by the **sr-te** protocol. Any data packet that is resolved to use the resulting tunnel has the label stack imposed at the head-end router and is forwarded out of the appropriate next-hop interface. This interface is determined by the topmost label in the stack.

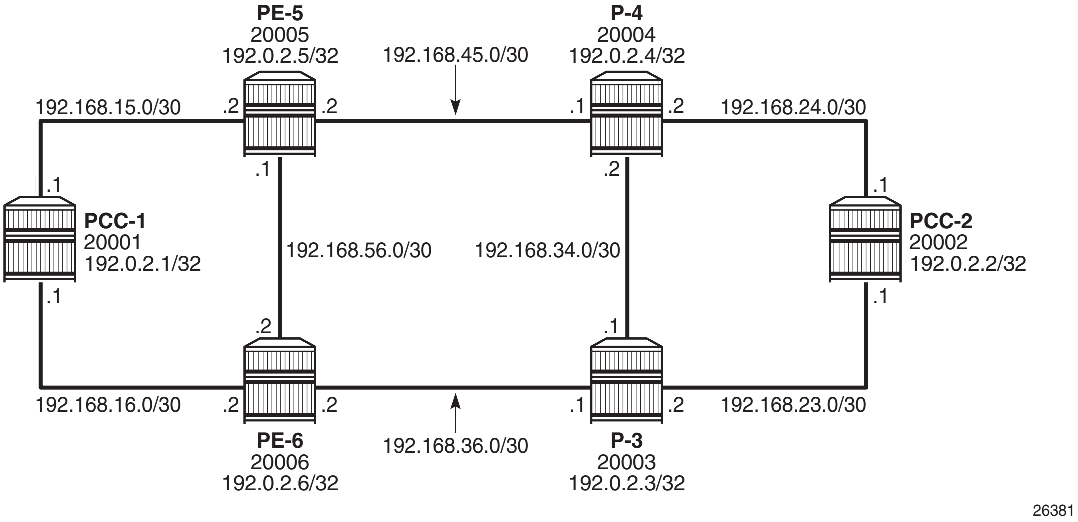
If the label is a node SID, the outgoing interface is determined by the IGP—the shortest path to the router that the node SID represents.

If the label is a local adjacency SID, the outgoing interface is the local interface for which this SID is generated by the IGP.

The segments referenced can be a prefix segment, such as a node segment or an adjacency segment, which represents a specific adjacency between two nodes. The SIDs are used as MPLS labels.

In the following configuration examples, the LSP path is created at the head-end router, and computed by translating a list of hops containing IP addresses into a list of SIDs, by examining the OSPF TE database. The head-end router is referred to as a Path Computation Client (PCC). Figure 306 shows the example topology used, and a pair of bidirectional connected SR-TE LSPs between PCC-1 and PCC-2 will be configured to illustrate SR-TE LSPs. All interfaces between PCC-1 and its neighbors have the OSPF metric set to 1000. Similarly, for PCC-2, the OSPF metric is also set to 1000 between itself and its neighbors. The OSPF metric on router interfaces between the core routers P-3, P-4, PE-5, and PE-6 are set to 100.

Figure 306 **Segment Routing Network Schematic**



Configuration

MPLS Label Range

The MPLS label range must be configured. This represents the Segment Routing Global Block (SRGB) from which node SIDs are allocated. The choice of SRGB in this example is the same as that chosen in the chapter Segment Routing with IS-IS Control Plane, where the label block is the same for each router. The SRGB is a contiguous range within the dynamic range 18432 to 524287, as in the following output:

```
*A:PCC-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           504846      505856
    Seg-Route   0           0           -           0           505856
============================================================================
```

In this example, a range of 1000 labels is chosen. For operational simplicity, Nokia recommends that the same label range is chosen for each router. However, this is not an explicit requirement.

A label range of 20000 to 20999 for SR is configured with the following command:

```
*A:PCC-1# configure router mpls-labels
            sr-labels start 20000 end 20999
```

When the SRGB label range has been configured, the MPLS label range looks like:

```
*A:PCC-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           504846      505856
    Seg-Route   20000       20999       -           0           1000
============================================================================
```

Global OSPF Configuration

The first step is to configure OSPF on each router, as shown in Figure 1. All router interfaces are members of a single backbone area: area 0.0.0.0.

The configuration for PCC-1 to enable OSPF is:

```
configure
    router
        ospf
            area 0.0.0.0
                interface "system"
                    no shutdown
                exit
                interface "int-PCC-1-PE-5"
                    interface-type point-to-point
                    metric 1000
                    no shutdown
                exit
                interface "int-PCC-1-PE-6"
                    interface-type point-to-point
                    metric 1000
                    no shutdown
                exit
            exit
            no shutdown
```

The configuration for all other nodes is the same, apart from the IP addresses. The IP addresses can be derived from Figure 1.

For each router to be segment-routing capable, additional configuration within the OSPF context is required, as in the following output for PCC-1:

```
configure
    router
        ospf
            traffic-engineering
            advertise-router-capability area
            area 0.0.0.0
                interface "system"
                    node-sid label 20001
                    no shutdown
                exit
            exit
            segment-routing
                prefix-sid-range global
                no shutdown
            exit
            no shutdown
```

The router capability is enabled using the **advertise-router-capability area** command, which defines the flooding scope of the opaque LSA used for this purpose as area. Traffic engineering is also enabled.

Also, MPLS must be enabled on each interface within the **configure router mpls** context, and RSVP must be placed in the **no shutdown** state using **configure router rsvp no shutdown** to ensure that OSPF opaque LSAs are generated.

A node SID is manually configured as a label, equivalent to the absolute node SID value. It is possible to configure the **node-sid** as an index. Indexing is explained in the chapter Segment Routing with IS-IS Control Plane.

Finally, segment routing is enabled, along with the **prefix-sid-range** command that states that the node prefix SID values of all routers within the network will be within the range of the global block.

The value of the **prefix-sid-range** must be the same for all routers; in this case, the range is always 1000.

The following output taken from PCC-1 shows the prefix SIDs configured on the routers in the network and advertised using OSPF. This will be identical for all routers in the network.

```
*A:PCC-1# show router ospf prefix-sids

========================================================================
Rtr Base OSPFv2 Instance 0 Prefix-Sids
========================================================================
Prefix                             Area            RtType      SID
```

```
                                   Adv-Rtr                    Flags
------------------------------------------------------------------------
192.0.2.1/32                       0.0.0.0         INTRA-AREA  1
                                   192.0.2.1                   NnP
192.0.2.2/32                       0.0.0.0         INTRA-AREA  2
                                   192.0.2.2                   NnP
192.0.2.3/32                       0.0.0.0         INTRA-AREA  3
                                   192.0.2.3                   NnP
192.0.2.4/32                       0.0.0.0         INTRA-AREA  4
                                   192.0.2.4                   NnP
192.0.2.5/32                       0.0.0.0         INTRA-AREA  5
                                   192.0.2.5                   NnP
192.0.2.6/32                       0.0.0.0         INTRA-AREA  6
                                   192.0.2.6                   NnP
------------------------------------------------------------------------
No. of Prefix/SIDs: 6
SID Flags : N = Node-SID
           nP = no penultimate hop POP
            M = Mapping server
            E = Explicit-Null
            V = Prefix-SID carries a value
            L = value/index has local significance
            I = Inter Area flag
            A = Attached flag
            B = Backup flag
========================================================================
```

The prefix SID for each node is displayed as an index; for example, 1. The absolute value of the Node SID is obtained by adding the (label_base) + (advertised SID index) = node prefix SID. The base label value for each router is chosen to be 20000, so the node prefix SID for PCC-1, for example, is 20000 + 1 = 20001.

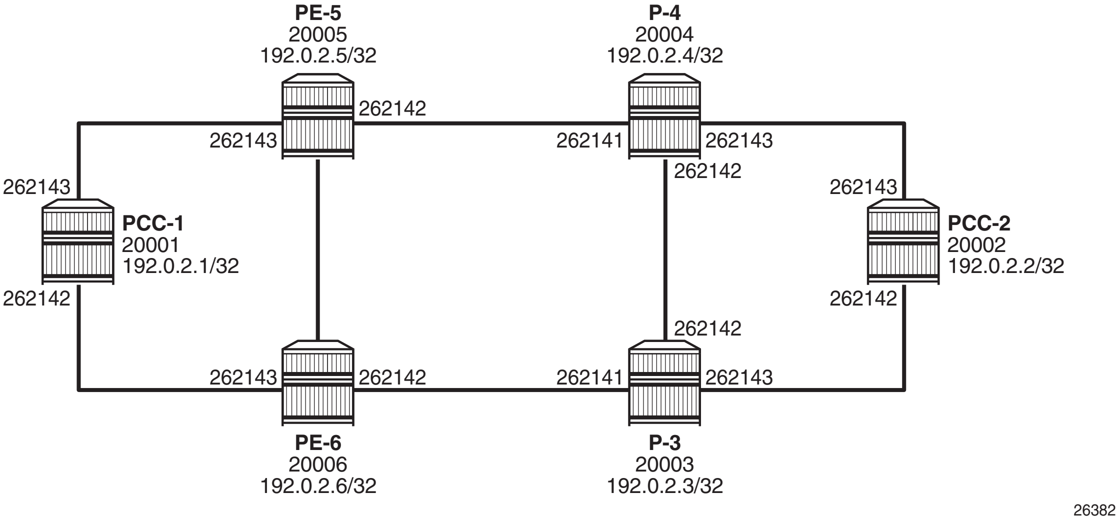
Adjacency SIDs are generated by OSPF for each interface link, and are advertised within the extended link opaque LSA using the adjacency SID sub-TLV. The following output shows the extended link opaque LSAs of PCC-1. There are two network links, so there are two LSAs, with link state IDs of 8.0.0.3 and 8.0.0.4.

```
*A:PCC-1# show router ospf opaque-database adv-router 192.0.2.1 detail

===============================================================================
Rtr Base OSPFv2 Instance 0 Opaque Link State Database (type: All) (detail)
===============================================================================
---snip---
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0              Adv Router Id    : 192.0.2.1
Link State Id    : 8.0.0.3              LSA Type         : Area Opaque
Sequence No      : 0x80000001           Checksum         : 0xb8f2
Age              : 48                   Length           : 48
Options          :  E
Advertisement    : Extended Link
    TLV Extended link (1) Len 24  :
        link Type=P2P (1)  Id=192.0.2.5 Data=192.168.15.1
        Sub-TLV Adj-SID (2) len 7 :
            Flags=Value Local (0x60)
            MT-ID=0 Weight=0 SID/Index/Label=262143
-------------------------------------------------------------------------------
Opaque LSA
-------------------------------------------------------------------------------
Area Id          : 0.0.0.0              Adv Router Id    : 192.0.2.1
Link State Id    : 8.0.0.4              LSA Type         : Area Opaque
Sequence No      : 0x80000001           Checksum         : 0xb0f8
Age              : 48                   Length           : 48
Options          :  E
Advertisement    : Extended Link
    TLV Extended link (1) Len 24  :
        link Type=P2P (1)  Id=192.0.2.6 Data=192.168.16.1
        Sub-TLV Adj-SID (2) len 7 :
            Flags=Value Local (0x60)
            MT-ID=0 Weight=0 SID/Index/Label=262142
===============================================================================
```

The adjacency SID for interface on PCC-1 toward PE-5 is 262143, and the adjacency SID for the interface toward PE-6 is 262142.

A full collection of SIDs for the whole network is shown in Figure 307.

Figure 307 **Node & Adjacency SIDs**



Segment Routing TE-LSPs

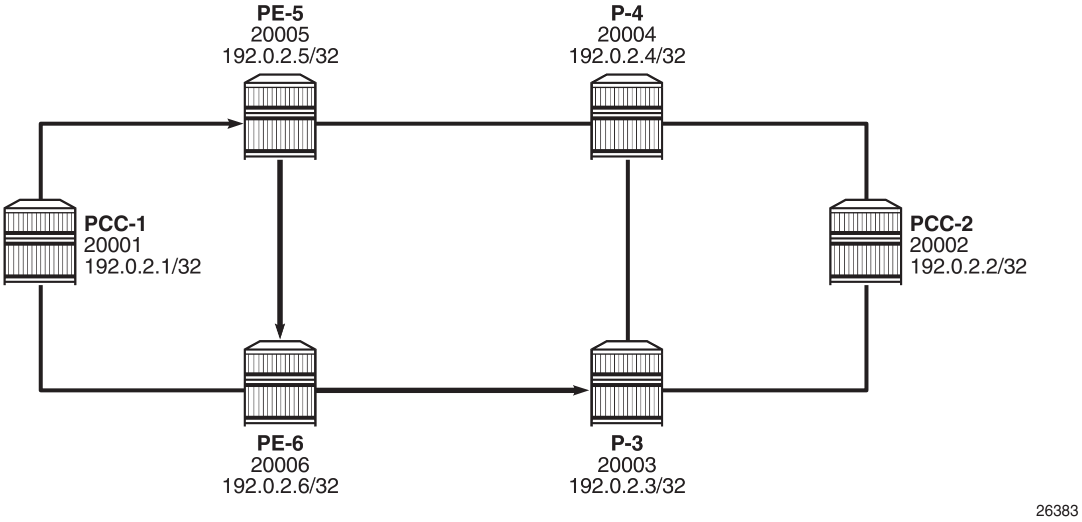
This section describes SR-TE LSPs that are configured on the head-end router (the PCC). The path taken through the network is computed locally by the PCC. To influence the path taken, a series of strict and/or loose hops are configured in an MPLS path.

Note: SR-TE LSPs configured with a loose path that contains no hops is effectively a shortest path tunnel to the destination node. The destination address is resolved to the node SID of the tail-end router.

PCC-initiated and Computed LSP – Strict Path

Consider an SR-TE LSP configured on PCC-1, with tail end at PCC-2. Assume there is a requirement for the LSP to avoid the link from PE-5 to P-4 during normal working, so a strict path from PCC-1 via PE-5 to PE-6, and then on to P-3 is required before being forwarded to PCC-2. This is shown in Figure 308.

Figure 308 **PCC Computed Strict Path between PCC-1 and PCC-2**



To meet these requirements, an MPLS path is configured containing the following strict hops, using the system addresses to identify the hops. The following output shows the configuration for the MPLS path required on PCC-1. This uses the identical CLI construct as an MPLS path used in configuring an RSVP-TE LSP.

```
configure
    router
        mpls
            path "PCC-controlled-strict"
                hop 1 192.0.2.5 strict
                hop 2 192.0.2.6 strict
                hop 3 192.0.2.3 strict
                no shutdown
```

The SR-TE LSP is configured on PCC-1 as per the following output:

```
configure
    router
        mpls
            lsp "PCC-1-PCC-2-PCC-strict" sr-te
                to 192.0.2.2
                primary "PCC-controlled-strict"
                exit
                no shutdown
            exit
```

Again, the same CLI construct as an RSVP-TE LSP is used, except for the **sr-te** keyword at creation time. If the **sr-te** keyword is not used, the LSP is signaled as an RSVP-TE LSP. The LSP configuration references the previously-created MPLS path as the primary path.

When placed in a **no shutdown** state, the LSP path status is as in the following output:

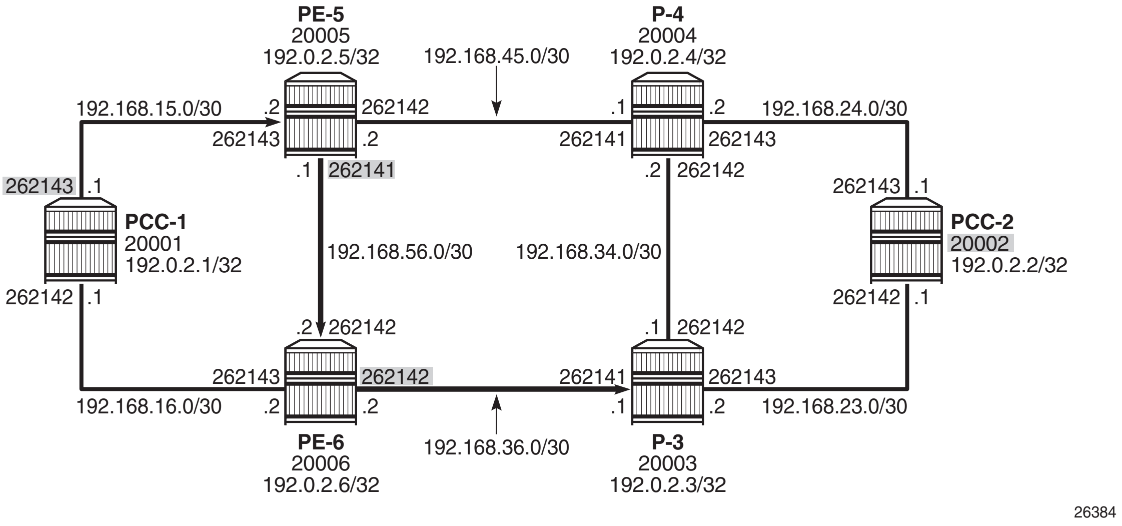
```
*A:PCC-1# show router mpls sr-te-lsp "PCC-1-PCC-2-PCC-strict" path detail

===============================================================================
MPLS SR-TE LSP PCC-1-PCC-2-PCC-strict Path  (Detail)
===============================================================================
Legend :
    S      - Strict                      L      - Loose
    A-SID  - Adjacency SID               N-SID  - Node SID
    +      - Inherited
===============================================================================
-------------------------------------------------------------------------------
SR-TE LSP PCC-1-PCC-2-PCC-strict Path PCC-controlled-strict
-------------------------------------------------------------------------------
LSP Name         : PCC-1-PCC-2-PCC-strict
Path LSP ID      : 46592
From             : 192.0.2.1            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : PCC-controlled-      Path Type            : Primary
                   strict
Path Admin       : Up                   Path Oper            : Up
Path Up Time     : 0d 00:00:10          Path Down Time       : 0d 00:00:00
Retry Limit      : 0                    Retry Timer          : 30 sec
Retry Attempt    : 1                    Next Retry In        : 0 sec

CSPF             : Disabled             Oper CSPF            : Disabled

Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Setup Priority   : 7                    Oper Setup Priority  : 7
Hold Priority    : 0                    Oper Hold Priority   : 0
Inter-area       : N/A

PCE Updt ID      : 0                    PCE Updt State       : None
PCE Upd Fail Code: noError

PCE Report       : Disabled+            Oper PCE Report      : Disabled
PCE Control      : Disabled             Oper PCE Control     : Disabled
PCE Compute      : Disabled             Oper PCE Compute     : Disabled
```

```
Include Groups    :                       Oper Include Groups  :
None                                             None
Exclude Groups    :                       Oper Exclude Groups  :
None                                             None

IGP/TE Metric     : 16777215              Oper Metric          : 16777215
Oper MTU          : 1548                  Path Trans           : 1
Failure Code      : noError
Failure Node      : n/a
Explicit Hops     :
    192.0.2.5(S)       -> 192.0.2.6(S)       -> 192.0.2.3(S)
Actual Hops       :
    192.168.15.2 (192.0.2.5)(A-SID)               Record Label        : 262143
 -> 192.168.56.2 (192.0.2.6)(A-SID)               Record Label        : 262141
 -> 192.168.36.1 (192.0.2.3)(A-SID)               Record Label        : 262142
 -> 192.0.2.2 (192.0.2.2)(N-SID)                  Record Label        : 20002
===============================================================================
```

The Actual Hops output shows the address of the upstream router facing the configured strict hop (in brackets) referenced in the MPLS path, plus a loose hop for the destination hop of 192.0.2.2.

The interface addresses are translated into SIDs to be used as MPLS labels, by the head-end PCC router, PCC-1, by examining the OSPF TE database. Each strict hop is always translated into an adjacency SID (A-SID), and a loose hop is always translated into a node SID (N-SID). This is shown in Figure 309.

Figure 309 **PCC Computed LSP Hop-To-Label Translation**



When the LSP is connected, the Tunnel Table Manager (TTM) adds an entry for the SR-TE LSP. This LSP is available for the provisioning of services that use the TTM. The following output shows the tunnel table for PCC-1, which includes the shortest-path tunnels to all other routers in the network, plus the entry for the provisioned SR-TE LSP. The default preference for an SR-TE LSP in the tunnel table is 8.

```
*A:PCC-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      sr-te     MPLS  655362    8        192.168.15.2   16777215
192.0.2.2/32      ospf (0)  MPLS  524291    10       192.168.15.2   2100
192.0.2.3/32      ospf (0)  MPLS  524294    10       192.168.16.2   1100
192.0.2.4/32      ospf (0)  MPLS  524292    10       192.168.15.2   1100
192.0.2.5/32      ospf (0)  MPLS  524293    10       192.168.15.2   1000
192.0.2.6/32      ospf (0)  MPLS  524295    10       192.168.16.2   1000
192.168.15.2/32   ospf (0)  MPLS  524289    10       192.168.15.2   0
192.168.16.2/32   ospf (0)  MPLS  524290    10       192.168.16.2   0
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
```

The value of the metric is set to 16777215 (infinity – 1), because there is no CSPF and the head-end router is unaware of the full topology between head- and tail-end router.

PCC-initiated and Computed LSP – Loose Path

Consider an LSP configured on PCC-2, with the tail end at PCC-1. There is a requirement for traffic on the LSP to pass through PE-6 before reaching PCC-1, so a loose path of PCC-2 to PE-6 before being forwarded to PCC-1 is required.

Figure 310 **SR-TE LSP with Loose Path**

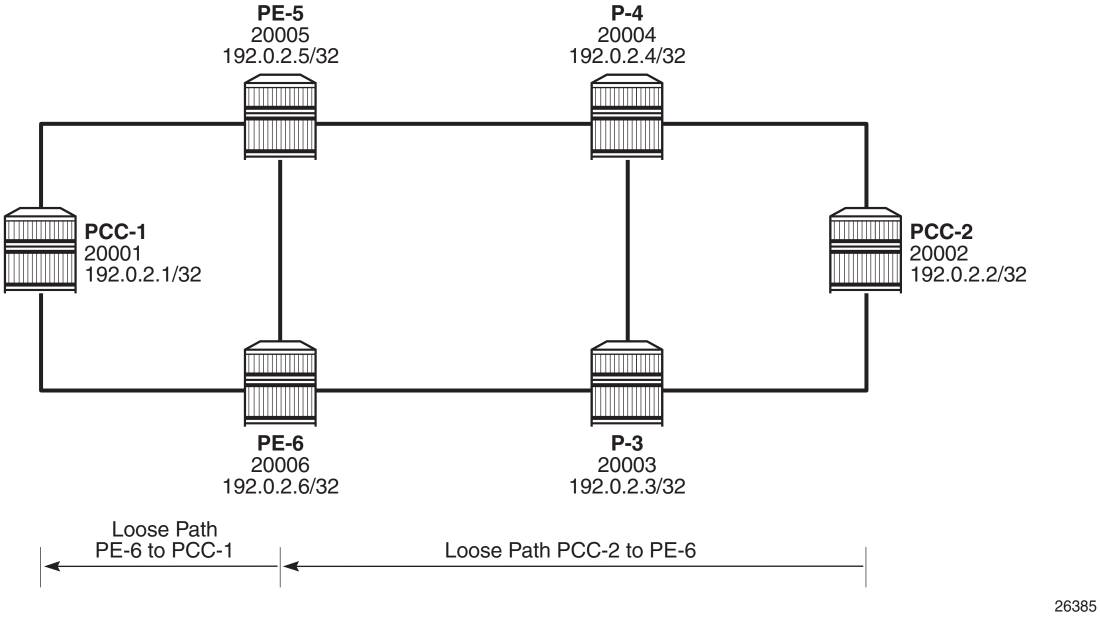


Figure 310 shows the concept of the loose path. The following output shows the MPLS path containing a loose hop configured on PCC-2:

```
configure
    router
        mpls
            path "PCC-controlled-loose"
                hop 1 192.0.2.6 loose
                no shutdown
            exit
```

The SR-TE LSP configuration is shown in the following output, which references the previously created MPLS path as the primary path:

```
configure
    router
        mpls
            lsp "PCC-2-PCC-1-PCC-loose" sr-te
                to 192.0.2.1
                primary "PCC-controlled-loose"
                exit
                no shutdown
            exit
```

When placed in a **no shutdown** state, the LSP path status becomes operationally up, as in the following output:

```
*A:PCC-2# show router mpls sr-te-lsp "PCC-2-PCC-1-PCC-loose" path detail
```

```
===============================================================================
MPLS SR-TE LSP PCC-2-PCC-1-PCC-loose Path  (Detail)
===============================================================================
Legend :
    S      - Strict                      L      - Loose
    A-SID  - Adjacency SID               N-SID  - Node SID
    +      - Inherited
===============================================================================
-------------------------------------------------------------------------------
SR-TE LSP PCC-2-PCC-1-PCC-loose Path PCC-controlled-loose
-------------------------------------------------------------------------------
LSP Name         : PCC-2-PCC-1-PCC-loose
Path LSP ID      : 27648
From             : 192.0.2.2            To                   : 192.0.2.1
Admin State      : Up                   Oper State           : Up
Path Name        : PCC-controlled-loose Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Path Up Time     : 0d 00:00:10          Path Down Time       : 0d 00:00:00
Retry Limit      : 0                    Retry Timer          : 30 sec
Retry Attempt    : 1                    Next Retry In        : 0 sec

CSPF             : Disabled             Oper CSPF            : Disabled

Bandwidth        : No Reservation       Oper Bandwidth       : 0 Mbps
Hop Limit        : 255                  Oper HopLimit        : 255
Setup Priority   : 7                    Oper Setup Priority  : 7
Hold Priority    : 0                    Oper Hold Priority   : 0
Inter-area       : N/A

PCE Updt ID      : 0                    PCE Updt State       : None
PCE Upd Fail Code: noError

PCE Report       : Disabled+            Oper PCE Report      : Disabled
PCE Control      : Disabled             Oper PCE Control     : Disabled
PCE Compute      : Disabled             Oper PCE Compute     : Disabled

Include Groups   :                      Oper Include Groups  :
None                                           None
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None

IGP/TE Metric    : 16777215             Oper Metric          : 16777215
Oper MTU         : 1556                 Path Trans           : 1
Failure Code     : noError
Failure Node     : n/a
Explicit Hops    :
    192.0.2.6(L)
Actual Hops      :
    192.0.2.6 (192.0.2.6)(N-SID)                 Record Label        : 20006
 -> 192.0.2.1 (192.0.2.1)(N-SID)                 Record Label        : 20001
===============================================================================
```

The Actual Hops in the MPLS path are the configured loose hop plus a hop for the destination of 192.0.2.1.

Again, the configured hop addresses are translated into labels by the head-end PCC router, PCC-2, by examining the OSPF TE database. The hop-to-label translation always translates a loose hop to a node SID (N-SID).

The LSP is installed by the TTM into the tunnel table, alongside OSPF advertised shortest path tunnels, for use by the TTM users.

```
*A:PCC-2# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      sr-te     MPLS  655362    8        192.0.2.6      16777215
192.0.2.1/32      ospf (0)  MPLS  524291    10       192.168.23.2   2100
192.0.2.3/32      ospf (0)  MPLS  524292    10       192.168.23.2   1000
192.0.2.4/32      ospf (0)  MPLS  524293    10       192.168.24.2   1000
192.0.2.5/32      ospf (0)  MPLS  524294    10       192.168.24.2   1100
192.0.2.6/32      ospf (0)  MPLS  524295    10       192.168.23.2   1100
192.168.23.2/32   ospf (0)  MPLS  524289    10       192.168.23.2   0
192.168.24.2/32   ospf (0)  MPLS  524290    10       192.168.24.2   0
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PCC-2#
```

Service Provisioning – VPRN

SR-TE tunnels are another MPLS tunnel type, and can be used in the context of **auto-bind-tunnel** for resolving BGP next hops for IPv4 routes within a VPRN.

Figure 311 **VPRN Service Schematic**

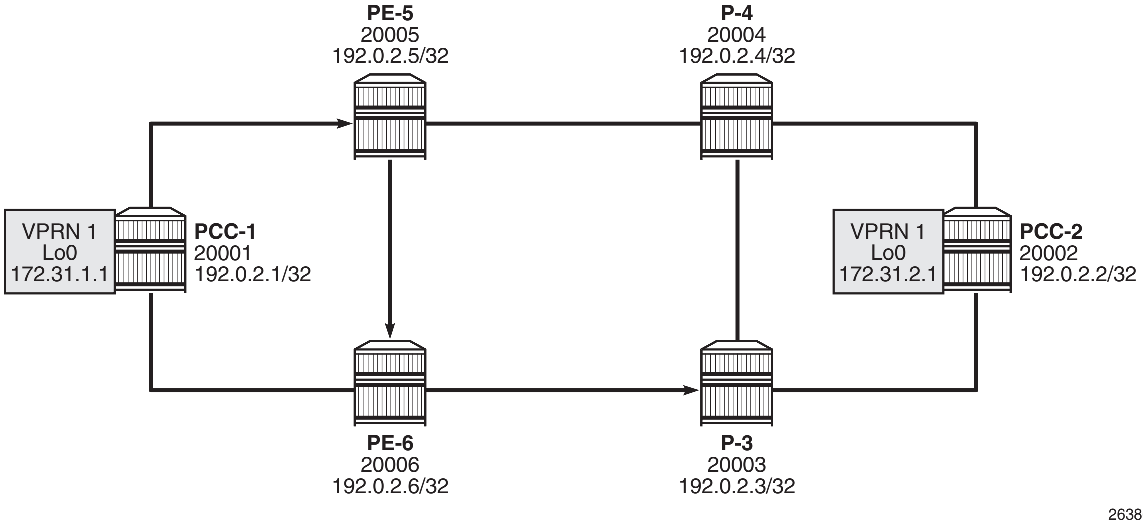


Figure 311 shows a VPRN service, configured on PCC-1 and PCC-2. The following output shows the VPRN 1 configuration on PCC-1, which includes a local interface using a /32 loopback address, which will be used to verify that routing is working correctly.

```
configure
    service
        vprn 1 customer 1 create
            autonomous-system 65545
            route-distinguisher 65545:1
            auto-bind-tunnel
                resolution-filter
                    sr-te
                exit
                resolution filter
            exit
            vrf-target target:65545:1
            interface "loopback" create
                address 172.31.1.1/32
                loopback
            exit
            no shutdown
```

Note: The **auto-bind-tunnel** command has the **resolution-filter** option set to **sr-te**, so that any BGP routes received will have the next-hop resolved to an SR-TE LSP. The VPRN configuration on PCC-2 also uses **auto-bind-tunnel sr-te**.

```
configure
    service
        vprn 1 customer 1 create
```

```
            autonomous-system 65545
            route-distinguisher 65545:1
            auto-bind-tunnel
                resolution-filter
                    sr-te
                exit
                resolution filter
            exit
            vrf-target target:65545:1
            interface "loopback" create
                address 172.31.2.1/32
                loopback
            exit
            no shutdown
```

Examination of the VPRN route table shows that the route prefix representing the IP address of the loopback address configured in VPRN 1 is shown, and is resolved via the SR-TE tunnel.

```
*A:PCC-1# show router 1 route-table

===============================================================================
Route Table (Service: 1)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                  Metric
-------------------------------------------------------------------------------
172.31.1.1/32                                 Local   Local     13d04h48m  0
       loopback                                                     0
172.31.2.1/32                                 Remote  BGP VPN   04h20m26s  170
       192.0.2.2 (tunneled:SR-TE:655363)                          0
-------------------------------------------------------------------------------
No. of Routes: 2
Flags: n = Number of times nexthop is repeated
       B = BGP backup route available
       L = LFA nexthop available
       S = Sticky ECMP requested
===============================================================================
```

Connectivity is verified by sending a ping from the loopback interface within VPRN 1 on PCC-1 to the loopback address within VPRN 1 on PCC-2, as shown in the following output:

```
*A:PCC-1# ping router 1 172.31.2.1 source 172.31.1.1
PING 172.31.2.1 56 data bytes
64 bytes from 172.31.2.1: icmp_seq=1 ttl=64 time=4.36ms.
64 bytes from 172.31.2.1: icmp_seq=2 ttl=64 time=4.47ms.
64 bytes from 172.31.2.1: icmp_seq=3 ttl=64 time=5.30ms.
64 bytes from 172.31.2.1: icmp_seq=4 ttl=64 time=5.22ms.
64 bytes from 172.31.2.1: icmp_seq=5 ttl=64 time=5.29ms.
```

For completeness, a ping is sent in the opposite direction, between the PCC-2 VPRN 1 interface to PCC-1 VPRN 1, as per the following output:

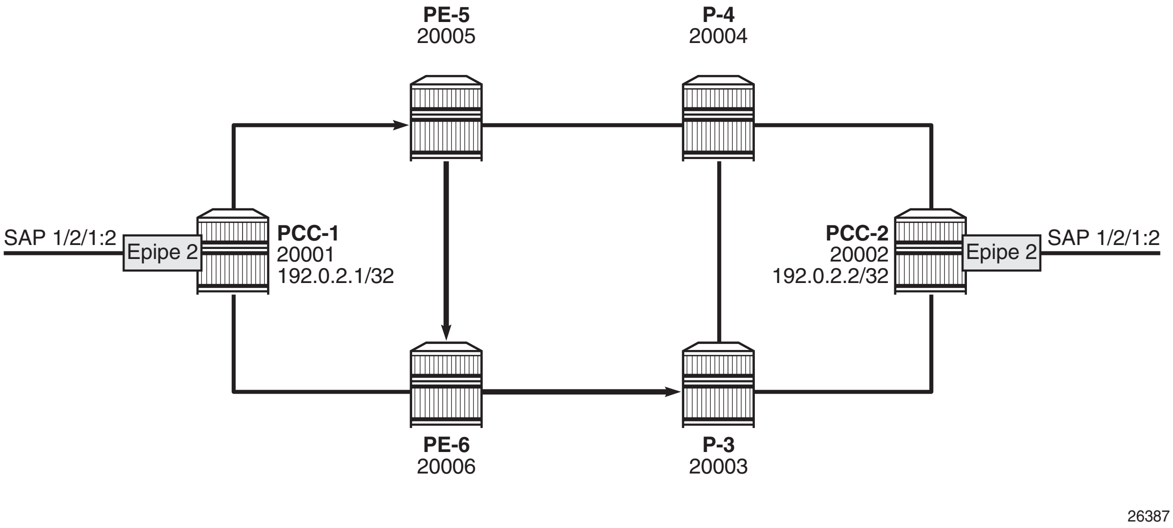
```
*A:PCC-2#  ping router 1 172.31.1.1 source 172.31.2.1
```

```
PING 172.31.1.1 56 data bytes
64 bytes from 172.31.1.1: icmp_seq=1 ttl=64 time=5.36ms.
64 bytes from 172.31.1.1: icmp_seq=2 ttl=64 time=4.71ms.
64 bytes from 172.31.1.1: icmp_seq=3 ttl=64 time=6.59ms.
64 bytes from 172.31.1.1: icmp_seq=4 ttl=64 time=5.74ms.
64 bytes from 172.31.1.1: icmp_seq=5 ttl=64 time=5.46ms.
```

Layer 2 Service Provisioning – SR-TE

SR-TE tunnels can also be bound as a transport tunnel within SDPs. To illustrate this, consider the following example of a simple Epipe connected between PCC-1 and PCC-2, as shown in Figure 312.

Figure 312 **Epipe Service Schematic**



Create an SDP on PCC-1, with far end on PCC-2, and bind it to the previously created SR-TE LSP:

```
configure
    service
        sdp 12 mpls create
            far-end 192.0.2.2
            sr-te-lsp "PCC-1-PCC-2-PCC-strict"
            no shutdown
        exit
```

Create an Epipe on PCC-1:

```
configure
    service
        epipe 2 customer 1 create
```

```
            sap 1/2/1:2 create
            exit
            spoke-sdp 12:2 create
            exit
            no shutdown
        exit
```

Similarly, for PCC-2, create an MPLS SDP and explicitly bind the SR-TE LSP, as follows:

```
configure
    service
        sdp 21 mpls create
            far-end 192.0.2.1
            sr-te-lsp "PCC-2-PCC-1-PCC-loose"
            no shutdown
        exit
```

Configure Epipe 2 on PCC-2, referencing the SDP as a spoke-SDP:

```
configure
    service
        epipe 2 customer 1 create
            sap 1/2/1:2 create
            exit
            spoke-sdp 21:2 create
            exit
            no shutdown
        exit
```

Service Verification

The state of SDP 12 on PCC-1 is shown in the following output:

```
*A:PCC-1# show service sdp

=============================================================================
Services: Service Destination Points
=============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
-----------------------------------------------------------------------------
12     0       1544    192.0.2.2        Up   Up          MPLS     T    TLDP
-----------------------------------------------------------------------------
Number of SDPs : 1
-----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
=============================================================================
```

The output shows the LSP type as an SR-TE LSP - "T".

On PCC-1, the following output shows the base state of the Epipe service entities:

```
*A:PCC-1# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                   Creation Origin   : manual
Last Status Change: 04/21/2017 09:55:44
Last Mgmt Change  : 04/21/2017 09:55:30
Test Service      : No
Admin State       : Up                  Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                   SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/2/1:2                              q-tag        1518    1518    Up   Up
sdp:12:2 S(192.0.2.2)                    Spok         0       1544    Up   Up
===============================================================================
*A:PCC-1#
```

Similarly, on PCC-2, the status of SDP 21 is as follows:

```
*A:PCC-2# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
21     0       1552    192.0.2.1        Up   Up          MPLS    T     TLDP
 ----------------------------------------------------------------------------
Number of SDPs : 1
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
===============================================================================
*A:PCC-2#
```

The state of the Epipe service on PCC-2 is shown in the following output:

```
*A:PCC-2# show service id 2 base

===============================================================================
Service Basic Information
===============================================================================
Service Id        : 2                   Vpn Id            : 0
Service Type      : Epipe
Name              : (Not Specified)
Description       : (Not Specified)
Customer Id       : 1                    Creation Origin   : manual
Last Status Change: 04/21/2017 09:53:20
Last Mgmt Change  : 04/21/2017 09:53:12
Test Service      : No
Admin State       : Up                   Oper State        : Up
MTU               : 1514
Vc Switching      : False
SAP Count         : 1                    SDP Bind Count    : 1
Per Svc Hashing   : Disabled
Vxlan Src Tep Ip  : N/A
Force QTag Fwd    : Disabled

-------------------------------------------------------------------------------
Service Access & Destination Points
-------------------------------------------------------------------------------
Identifier                               Type         AdmMTU  OprMTU  Adm  Opr
-------------------------------------------------------------------------------
sap:1/2/1:2                              q-tag        1518    1518    Up   Up
sdp:21:2 S(192.0.2.1)                    Spok         0       1552    Up   Up
===============================================================================
*A:PCC-2#
```

Conclusion

Segment routing LSPs extend the use of MPLS labels into traffic engineering applications. This chapter provides the configuration for router instantiated and controlled SR-TE LSPs along with some examples of the application in a VPRN and Epipe. The chapter also shows the associated commands and outputs that can be used for verifying and troubleshooting.

Segment Routing with IS-IS Control Plane

This chapter provides information about Segment Routing (SR) with Intermediate System to Intermediate System (IS-IS) control plane.

Topics in this chapter include:

Applicability

Segment routing is supported in SR OS release 13.0, and later. This chapter was initially written for SR OS release 13.0.R3, but the CLI in the current edition corresponds to SR OS release 16.0.R3.

Overview

Segment Routing (SR) is a technology for IP/Multi-Protocol Label Switching (MPLS) networks that enables source routing. With source routing, operators can specify a forwarding path, from ingress to egress, that is independent of the shortest path determined by the Interior Gateway Protocol (IGP).

The main benefit of segment routing compared to other source routing protocols (such as ReSource reserVation Protocol with Traffic Engineering (RSVP-TE)) is that, from a control plane perspective, no signaling protocol is required. Segment routing provides a path or tunnel, encoded as a sequential list of sub-paths or segments that are advertised within the segment routing domain, using extensions to well-known link state routing protocols, such as IS-IS or Open Shortest Path First (OSPF).

Implementation

A segment routing tunnel can contain a single segment that represents the destination node, or it can contain a list of segments that the tunnel must traverse. The tunnel can be established over an IPv4/IPv6 MPLS or IPv6 data plane, encoded as a stack of MPLS labels or as a number of IPv6 addresses contained in an IPv6 extension header.

Network elements are modeled as segments. For each segment, IGP advertises an identifier referred to as a segment ID (SID).

The two segment types are:

- **Prefix segment** — Globally unique and allocated from a Segment Routing Global Block (SRGB), typically multi-hop and signaled by the IGP. It is the Equal Cost Multi-Path ECMP-aware shortest path IGP route to a related prefix. A typical example of a prefix segment is a node SID. Within the SR OS implementation, the node SID is either the system address or another interface address in the Global Routing Table (GRT) of type loopback. Node SIDs are advertised in IS-IS using a prefix SID sub-TLV (Type Length Value).
- **Adjacency segment** — Locally unique and allocated from the (local) dynamic label space, so that other routers in the SR domain can use the same label space. Adjacency segments are signaled by the IGP. Within the SR OS implementation, adjacency SIDs are automatically assigned and advertised when the SR context within the IGP instance is set in no shutdown. Adjacency SIDs are advertised in IS-IS using an adjacency SID sub-TLV.

To make prefix segments globally unique within the segment routing domain, an indexing mechanism is required, because production networks consist of multiple vendors and multiple products. As a result, it is often difficult to agree on a common SRGB for the prefix SIDs.

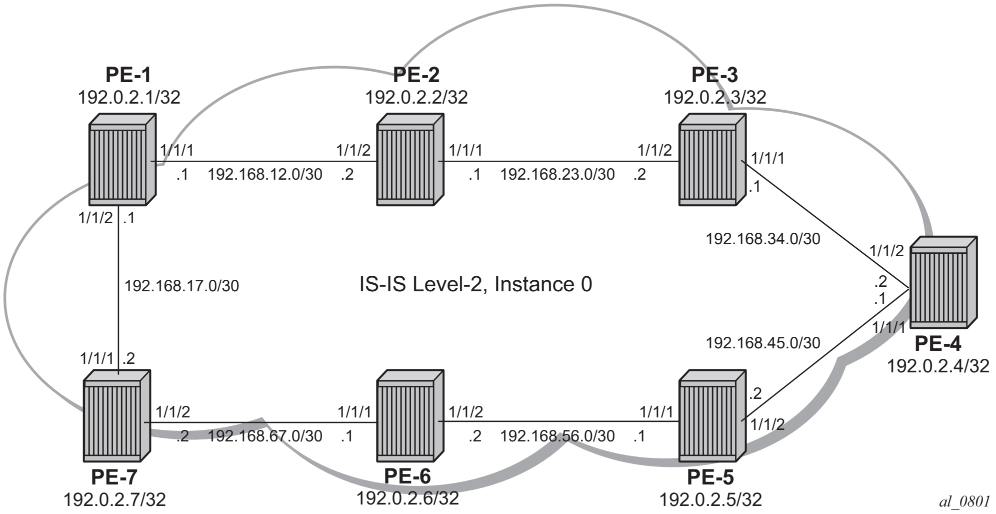
All routers within the SR domain are expected to configure and advertise the same Prefix SID index range for an IGP instance. The label value used by each router to represent a prefix can be local to that router by the use of an offset label, referred to as a start label:

Local label (for a prefix) = (local) start label + {Prefix SID index}

Within the SR OS implementation, prefix Loop-Free Alternate (LFA) is supported for segment routing to improve the Fast ReRoute (FRR) coverage. Remote LFA (RLFA) is also supported. With RLFA, segment routing shortest path tunnels are used as a virtual LFA or repair tunnel toward the PQ node.

The following example uses IS-IS as an IGP protocol, with an MPLS data plane and services enabled using LFA and RLFA. Figure 313 shows the example topology with seven PEs.

Figure 313 **Example Topology**



Configuration

Step 1. Configure router interfaces and IS-IS according to Figure 313.

- The system and IP interface addresses are configured according to Figure 313.
- IS-IS level 2 is selected as the IGP to distribute routing information between all PEs. All IS-IS interfaces are of type point-to-point to avoid running the Designated Router/Backup Designated Router (DR/BDR) election process.

Step 2. Configure segment routing.

Before enabling segment routing on a router, define a dedicated SRGB. This SRGB is required on each individual router part of the SR domain and is used to allocate the Prefix SIDs.

By default, an SRGB is not instantiated and, when configured by the operator, it is taken from the system dynamic label range. By default, the following label ranges are available:

```
*A:PE-1# show router mpls-labels label-range

===============================================================================
Label Ranges
```

```
===========================================================================
Label Type      Start Label End Label   Aging       Available   Total
---------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           505856      505856
    Seg-Route   0           0           -           0           0
===========================================================================
*A:PE-1#
```

For simplicity, the same SRGB is used in this example for all SR domain routers. Within the command, a start value and end value define the size of the SRGB. The following command configures an SRGB of 100 MPLS labels, starting from label 20000:

```
*A:PE-1# configure router mpls-labels sr-labels start 20000 end 20099
```

```
*A:PE-1# show router mpls-labels label-range

===========================================================================
Label Ranges
===========================================================================
Label Type      Start Label End Label   Aging       Available   Total
---------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           505756      505856
    Seg-Route   20000       20099       -           0           100
===========================================================================
```

This command is repeated for all other nodes. The allocated MPLS labels are only for the prefix SIDs. The adjacency SIDs, which are only locally unique, are taken from the dynamic range; in this example, between 18432 and 524287.

1. Enable router capability in the IGP instance.

   It is mandatory to enable the router-capability parameter inside the IS-IS instance, to advertise SR support among the IS-IS adjacencies. By configuring this command within the IGP instance, the SR capability sub-TLV is propagated and is used to indicate the index range and the start label. The SR algorithm sub-TLV is also used to advertise the algorithm used for path calculations. Only Shortest Path First (SPF) (value 0) is defined. This is configured as follows:

```
<all-nodes-within-SR-domain># configure router isis advertise-router-capability area
```

   The flooding parameter is a mandatory parameter in this CLI command. The keyword **area** or **as** indicates that the router capabilities label switched path (LSP) should be advertised throughout the same level or throughout the whole Autonomous System (AS). In the preceding example, all routers belong to the same level, so the **area** argument is sufficient. When the SR context within the IGP instance is set in no shutdown, both IS-IS sub-TLVs are flooded.

2. Define the Prefix SID index range.

   The SR OS implementation for SR provides two mutually exclusive modes of operation to define the Prefix SID index range: global mode and per-instance mode. Per-instance mode is useful in a seamless MPLS environment when multiple IGP instances are used. The main difference between the modes is the way that the start label and index range are calculated.

   A comparison of the modes is shown in following table:

Table 27 **Mode Comparison**

Global	Per Instance
Applicable for all IGP instances on that node	Applicable for one dedicated IGP instance
Start label is first label of SRGB	Start label is configurable (but part of SRGB range); use of non-overlapping sub-ranges of SRGB
Prefix SID index range is “size” of SRGB	Prefix SID index-range is configurable
If SRGB needs to change, shut down SR and delete prefix-SID-ranges in all IGP instances	If prefix SID index and/or label range needs to change, shut down SR in that specific IGP instance
SW checks whether any allocated SID index/label goes out of range. SW checks also for overlaps of the resulting net label value range across IGP instances.	

   For simplicity, global mode is used for this example, as follows:

```
<all-nodes-within-SR-domain># configure router isis segment-routing prefix-sid-range
global
```

3. Assign a prefix SID index or label to the prefix representing a node.

   To be able to set up SR shortest path tunnels to all routers of the SR domain, each router needs to be uniquely defined within the SR domain. Therefore, the system address or other loopback interface in the GRT will be assigned an **ipv4-node-sid index** or **label** value that is unique within the SR domain. The prefix SID index is assigned as follows:

```
*A:PE-1# configure router isis interface "system" ipv4-node-sid index 1
*A:PE-2# configure router isis interface "system" ipv4-node-sid index 2
*A:PE-3# configure router isis interface "system" ipv4-node-sid index 3
*A:PE-4# configure router isis interface "system" ipv4-node-sid index 4
*A:PE-5# configure router isis interface "system" ipv4-node-sid index 5
*A:PE-6# configure router isis interface "system" ipv4-node-sid index 6
*A:PE-7# configure router isis interface "system" ipv4-node-sid index 7
```

Because the SRGB is the same on all nodes, each node in the network can be reached using the same MPLS label. For example, the node SID for PE-5 on all nodes has a start label (first label of the SRGB (= 20000) + ipv4-node-sid index on node PE-5 (= 5)) of 20005.

When there is one consistent SRGB for the SR domain, the SR OS CLI allows the use of absolute MPLS label values instead of index values. For example, on PE-1, an operator can use an explicit MPLS label value, as follows:

```
*A:PE-1# configure router isis interface "system" ipv4-node-sid label 20001
```

Internally, this explicit value is translated into an index value (index-value 1) before advertising it toward its neighbors, taking into account the prefix SID index-range mode (global or per-instance) and the SRGB.

4. Enable SR context within the IGP instance, as follows:

```
<all-nodes-within-SR-domain># configure router isis segment-routing no shutdown
```

After enabling the SR context within an IGP instance, the SR capability sub-TLV, and the SR algorithm sub-TLV between all routers within the SR domain, are flooded. The following show command displays the SR related router capability information on PE-1:

```
*A:PE-1# show router isis capabilities level 2

===============================================================================
Rtr Base ISIS Instance 0 Capabilities
===============================================================================

Displaying Level 2 capabilities
-------------------------------------------------------------------------------
LSP ID    : PE-1.00-00
  Router Cap : 192.0.2.1, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

LSP ID    : PE-2.00-00
  Router Cap : 192.0.2.2, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

LSP ID    : PE-3.00-00
  Router Cap : 192.0.2.3, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

LSP ID    : PE-4.00-00
  Router Cap : 192.0.2.4, D:0, S:0
```

```
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

LSP ID    : PE-5.00-00
  Router Cap : 192.0.2.5, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

LSP ID    : PE-6.00-00
  Router Cap : 192.0.2.6, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF
LSP ID    : PE-7.00-00
  Router Cap : 192.0.2.7, D:0, S:0
    TE Node Cap : B E M  P
    SR Cap: IPv4 MPLS-IPv6
       SRGB Base:20000, Range:100
    SR Alg: metric based SPF

Level (2) Capability Count : 7
===============================================================================
*A:PE-1#
```

A similar output occurs for each router in the SR domain.

After enabling the SR context within the IGP instance, the assigned index for each locally configured prefix SID is advertised. After the advertisement of prefix SIDs, MPLS data plane Ingress Label Mapping (ILM) is programmed with a pop operation. In this context, a show command can be used to display the prefix SIDs, in order, within the SR domain. As an example, on PE-1, this becomes:

```
*A:PE-1# show router isis prefix-sids

===============================================================================
Rtr Base ISIS Instance 0 Prefix/SID Table
===============================================================================
Prefix                            SID        Lvl/Typ    SRMS   AdvRtr
                                                         MT     Flags
-------------------------------------------------------------------------------
192.0.2.1/32                      1          2/Int.      N     PE-1
                                                            0       NnP
192.0.2.2/32                      2          2/Int.      N     PE-2
                                                            0       NnP
192.0.2.3/32                      3          2/Int.      N     PE-3
                                                            0       NnP
192.0.2.4/32                      4          2/Int.      N     PE-4
                                                            0       NnP
192.0.2.5/32                      5          2/Int.      N     PE-5
                                                            0       NnP
192.0.2.6/32                      6          2/Int.      N     PE-6
                                                            0       NnP
```

```
192.0.2.7/32                      7          2/Int.      N     PE-7
                                                            0       NnP
-------------------------------------------------------------------------------
No. of Prefix/SIDs: 7 (7 unique)
-------------------------------------------------------------------------------
SRMS : Y/N  = prefix SID advertised by SR Mapping Server (Y) or not (N)
       S    = SRMS prefix SID is selected to be programmed
Flags: R    = Re-advertisement
       N    = Node-SID
       nP   = no penultimate hop POP
       E    = Explicit-Null
       V    = Prefix-SID carries a value
       L    = value/index has local significance
===============================================================================
*A:PE-1#
```

By default, the SR OS implementation sets the node SID (or **N**–flag) and no Penultimate hop PoP (or **nP**–flag) inside the prefix SID TLV. Another useful flag that can be set is the re-advertisement (or R-flag). The R-flag is set when a prefix SID is propagated between levels or areas, or redistribution is in place (from another protocol).

Prefix SID information can also be viewed within the IGP database attached to (extended) IP prefix reachability TLVs. For example, on PE-1, as follows:

```
*A:PE-1# show router isis database level 2 PE-1.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================

Displaying Level 2 database
-------------------------------------------------------------------------------
LSP ID    : PE-1.00-00                                  Level     : L2
Sequence  : 0x5                    Checksum  : 0x830c   Lifetime  : 1111
Version   : 1                      Pkt Type  : 20       Pkt Ver   : 1
Attributes: L1L2                   Max Area  : 3        Alloc Len : 1492
SYS ID    : 1920.0000.2001         SysID Len : 6        Used Len  : 248

TLVs :
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-1
  Router ID   :
    Router ID   : 192.0.2.1
---snip---

  TE IP Reach   :
---snip---

    Default Metric  : 0
    Control Info:   S, prefLen 32
    Prefix   : 192.0.2.1
    Sub TLV   :
      Prefix-SID Index:1, Algo:0, Flags:NnP
```

```
Level (2) LSP Count : 1
---snip---
```

After enabling the SR context within the IGP instance, adjacency SIDs are also automatically assigned and advertised for each formed adjacency over an IP interface. From a data plane perspective, one local adjacency SID consumes one ILM entry, programming a pop operation.

Similar to prefix SIDs, adjacency SID information can be viewed within the IGP database attached to IS neighbor TLVs, as follows:

```
*A:PE-1# show router isis database level 2 PE-1.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================
Displaying Level 2 database
-------------------------------------------------------------------------------
LSP ID    : PE-1.00-00                                 Level     : L2
Sequence  : 0x5                    Checksum  : 0x830c  Lifetime  : 1111
Version   : 1                      Pkt Type  : 20      Pkt Ver   : 1
Attributes: L1L2                   Max Area  : 3       Alloc Len : 1492
SYS ID    : 1920.0000.2001         SysID Len : 6       Used Len  : 248

TLVs :
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-1
  Router ID   :
    Router ID   : 192.0.2.1
---snip---

 TE IS Nbrs   :
    Nbr   : PE-2.00
    Default Metric  : 10
    Sub TLV Len     : 19
    IF Addr   : 192.168.12.1
    Nbr IP    : 192.168.12.2
    Adj-SID: Flags:v4VL Weight:0 Label:524287
  TE IS Nbrs   :
    Nbr   : PE-7.00
    Default Metric  : 10
    Sub TLV Len     : 19
    IF Addr   : 192.168.17.1
    Nbr IP    : 192.168.17.2
    Adj-SID: Flags:v4VL Weight:0 Label:524286
---snip---
```

By default, the SR OS implementation sets the value (**V**–flag), meaning that the adjacency SID carries a value (as opposed to an index). Also, the local **L**-flag is set by default, meaning that the adjacency SID has only local significance. The **v4**-flag set to 0 means that the adjacency SID references to an adjacency with outgoing IPv4 encapsulation.

Another way to display adjacency SID information is using the **show router isis adjacency detail** command.

```
*A:PE-1# show router isis adjacency "int-PE-1-PE-2" detail

===============================================================================
Rtr Base ISIS Instance 0 Adjacency (detail)
===============================================================================
SystemID    : PE-2                             SNPA        : 4a:c5:01:01:00:02
Interface   : int-PE-1-PE-2                    Up Time     : 0d 00:07:08
State       : Up                               Priority    : 0
Nbr Sys Typ : L2                               L. Circ Typ : L2
Hold Time   : 22                               Max Hold    : 27
Adj Level   : L2                               MT Enabled  : No
Topology    : Unicast

IPv6 Neighbor     : ::
IPv4 Neighbor     : 192.168.12.2
IPv4 Adj SID      : Label 524287
Restart Support   : Disabled
Restart Status    : Not currently being helped
Restart Supressed : Disabled
Number of Restarts: 0
Last Restart at   : Never

===============================================================================



*A:PE-1# show router isis adjacency "int-PE-1-PE-7" detail

===============================================================================
Rtr Base ISIS Instance 0 Adjacency (detail)
===============================================================================
SystemID    : PE-7                             SNPA        : 4a:a4:01:01:00:01
Interface   : int-PE-1-PE-7                    Up Time     : 0d 00:06:39
State       : Up                               Priority    : 0
Nbr Sys Typ : L2                               L. Circ Typ : L2
Hold Time   : 23                               Max Hold    : 27
Adj Level   : L2                               MT Enabled  : No
Topology    : Unicast

IPv6 Neighbor     : ::
IPv4 Neighbor     : 192.168.17.2
IPv4 Adj SID      : Label 524286
Restart Support   : Disabled
Restart Status    : Not currently being helped
Restart Supressed : Disabled
Number of Restarts: 0
Last Restart at   : Never

===============================================================================
*A:PE-1#
```

Finally, when enabling the SR context within the IGP instance, the SR module resolves received prefixes with prefix SID sub-TLVs present. As a result, MPLS data plane resources are consumed. The ILM is programmed with a swap operation and the label-to-next-hop-label-forwarding-entry (LTN) with a push operation, both pointing to the primary and/or LFA next-hop label forwarding entry (NHLFE). Also, an SR tunnel is added in the Tunnel Table Manager (TTM). As a result, an SR shortest path tunnel is set up to each other router that is part of the SR domain. Now, SR shortest path tunnels can be used for all users of TTM.

Example 1: VPRN service with LFA and RLFA enabled

In the network topology of Figure 313, no LDP and RSVP-TE signaling protocols are enabled. Each router of the SR domain has a full mesh of SR shortest path tunnels to the other routers, and no LDP and RSVP-TE LSPs are present. For example, on PE-1, the TTM looks as follows:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.2/32      isis (0)  MPLS  524291    11       192.168.12.2   10
192.0.2.3/32      isis (0)  MPLS  524295    11       192.168.12.2   20
192.0.2.4/32      isis (0)  MPLS  524293    11       192.168.12.2   30
192.0.2.5/32      isis (0)  MPLS  524292    11       192.168.17.2   30
192.0.2.6/32      isis (0)  MPLS  524296    11       192.168.17.2   20
192.0.2.7/32      isis (0)  MPLS  524294    11       192.168.17.2   10
192.168.12.2/32   isis (0)  MPLS  524289    11       192.168.12.2   0
192.168.17.2/32   isis (0)  MPLS  524290    11       192.168.17.2   0
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The objective is to configure a VPRN between PE-1 and PE-7, using SR shortest path tunnels as transport tunnel. The configuration is as follows:

```
*A:PE-1# configure
    service
        vprn 100 name "VPRN 100" customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:10001
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:100
            interface "loopback" create
                address 192.0.1.1/32
                loopback
            exit
            no shutdown

*A:PE-7# configure
    service
        vprn 100 name "VPRN 100" customer 1 create
            autonomous-system 64496
            route-distinguisher 64496:10007
            auto-bind-tunnel
                resolution any
            exit
            vrf-target target:64496:100
            interface "loopback" create
                address 192.0.1.7/32
                loopback
            exit
            no shutdown
```

Within the VPRN service configuration, a loopback interface is created on both PEs to verify the transport mechanism. Tunnel information displaying the MPLS label value is retrieved using the **show router fp-tunnel-table <slot number>** command, as follows:

```
*A:PE-1# show router fp-tunnel-table 1 192.0.2.7/32

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.7/32                                 SR-ISIS-0             -
  20007
  192.168.17.2                                                 1/1/2
-------------------------------------------------------------------------------
Total Entries : 1
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

This means that, when traffic arrives on PE-1, the MPLS label 20007 is pushed to reach destination PE-7. Because, in this example, the prefix SID index range global mode is used, the value 20007 comes from the start label on PE-7 (first label of the SRGB, which is 20000, plus the configured index value of node SID PE-7 (7)), so 20007.

Enabling prefix LFA within the IS-IS context on PE-1 will enable LFA/FRR protection. Next-hop LFA protection is present for node PE-4, node PE-5, and the link between PE-4 and PE-5, as follows:

```
*A:PE-1# configure router isis loopfree-alternate
```

```
*A:PE-1# show router isis lfa-coverage
```

```
===============================================================================
Router Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
---snip---
IPV4 Unicast     L2      2/6(33%)       3/11(27%)           0/0(0%)
---snip---

*A:PE-1# show router route-table alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
---snip---
192.0.2.4/32                                  Remote  ISIS      00h27m26s  18
       192.168.12.2                                                 30
       192.168.17.2 (LFA)                                           40
192.0.2.5/32                                  Remote  ISIS      00h23m34s  18
       192.168.17.2                                                 30
       192.168.12.2 (LFA)                                           40
---snip---
192.168.45.0/30                               Remote  ISIS      00h27m27s  18
       192.168.12.2                                                 40
       192.168.17.2 (LFA)                                           50
---snip---
-------------------------------------------------------------------------------
No. of Routes: <...>
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested


*A:PE-1# show router fp-tunnel-table 1
===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                  Protocol            Tunnel-ID
  Lbl
    NextHop                                                      Intf/Tunnel
-------------------------------------------------------------------------------
---snip---
192.0.2.4/32                                 SR-ISIS-0             -
  20004
  192.168.12.2                                                 1/1/1
  20004
  192.168.17.2(B)                                              1/1/2
192.0.2.5/32                                 SR-ISIS-0             -
  20005
  192.168.17.2                                              1/1/2
  20005
  192.168.12.2(B)                                           1/1/1
---snip---

*A:PE-1# show router tunnel-table detail

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
---snip---
-------------------------------------------------------------------------------
Destination      : 192.0.2.4/32
NextHop          : 192.168.12.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m52s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524293               Preference       : 11
Tunnel Label     : 20004                Tunnel Metric    : 30
Tunnel MTU       : 1560                 Max Label Stack  : 1
-------------------------------------------------------------------------------
Destination      : 192.0.2.5/32
NextHop          : 192.168.17.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m52s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524292               Preference       : 11
Tunnel Label     : 20005                Tunnel Metric    : 30
Tunnel MTU       : 1560                 Max Label Stack  : 1
-------------------------------------------------------------------------------
---snip---
Number of tunnel-table entries with LFA : 2
```

When a failure occurs on the primary SR path (only applicable for prefix PE-4/PE-5 and the link between PE-4 and PE-5), the traffic takes the LFA backup SR path to the destination using the same MPLS label value.

To extend the LFA/FRR coverage, for example, to find an LFA protection for node PE-7, which is one of the VPRN service endpoints, RLFA can be enabled. RLFA creates a virtual LFA by using a repair tunnel to carry packets to a point in the network from where they will not be looped back to the source, but forwarded (SPF-based) toward the destination prefix.

The RLFA implementation uses the PQ algorithm. The node where RLFA is configured (PE-1 in this example) computes an extended P-space and a Q-space. The intersection of both spaces is called the PQ-node. This PQ node is the destination node of the repair tunnel using an SR shortest path tunnel. To compute both spaces, SPF is used.

In this example, IS-IS is used as the IGP, using a default metric value of 10 for all links. With the assumption that the link between PE-1 and PE-7 is broken, the calculation of both the extended P-space and the Q-space at PE-1 is as follows:

- extended P-space — An SPF computed from node PE-1 and rooted at PE-2. It is used to calculate the set of routers that are reachable without any path transiting the protected link between PE-1 and PE-7. The following nodes belong to the extended P-space: PE-2, PE-3, PE-4, and PE-5.
- Q-space — A reverse SPF computed from PE-1 and rooted from PE-7 (acting as destination proxy). It is used to calculate the set of routers that can reach PE-7 without transiting the protected link between PE-1 and PE-7. The nodes PE-4, PE-5, and PE-6 belong to the Q-space.

Possible PQ-nodes are PE-4 or PE-5, because they are in the intersection of both spaces.

RLFA is configured as follows:

```
*A:PE-1# configure router isis loopfree-alternate remote-lfa
```

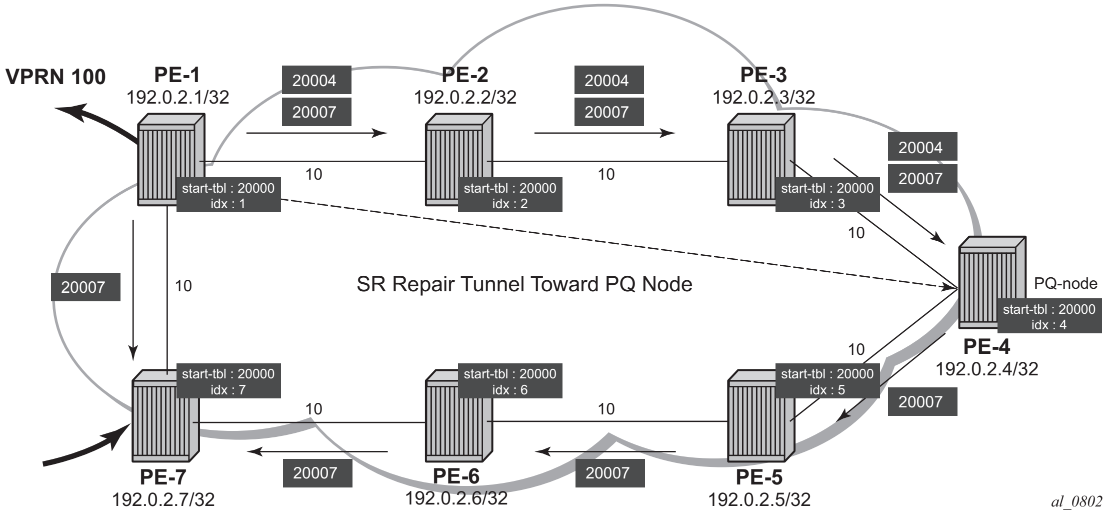
The nodes PE-2, PE-3, PE-6, and PE-7 now have RLFA protection, whereas PE-4 and PE-5 have LFA protection.

```
*A:PE-1# show router fp-tunnel-table 1

===============================================================================
Tunnel Table Display

Legend:
B - FRR Backup
===============================================================================
Destination                                   Protocol            Tunnel-ID
  Lbl
    NextHop                                                       Intf/Tunnel
-------------------------------------------------------------------------------
192.0.2.2/32                                  SR-ISIS-0             -
  20002
  192.168.12.2                                                    1/1/1
  20002/20005
  192.168.17.2(B)                                                 1/1/2
192.0.2.3/32                                  SR-ISIS-0             -
  20003
  192.168.12.2                                                    1/1/1
  20003/20005
  192.168.17.2(B)                                                 1/1/2
192.0.2.4/32                                  SR-ISIS-0             -
  20004
  192.168.12.2                                                    1/1/1
  20004
  192.168.17.2(B)                                                 1/1/2
192.0.2.5/32                                  SR-ISIS-0             -
  20005
  192.168.17.2                                                    1/1/2
  20005
  192.168.12.2(B)                                                 1/1/1
192.0.2.6/32                                  SR-ISIS-0             -
  20006
  192.168.17.2                                                    1/1/2
  20006/20004
  192.168.12.2(B)                                              1/1/1
192.0.2.7/32                                 SR-ISIS-0            -
  20007
  192.168.17.2                                                 1/1/2
  20007/20004
  192.168.12.2(B)                                              1/1/1
192.168.12.2/32                              SR                   524289
  3
  192.168.12.2                                                 1/1/1
  20002/20005
  192.168.17.2(B)                                              1/1/2
192.168.17.2/32                              SR                   524290
  3
  192.168.17.2                                                 1/1/2
  20007/20004
  192.168.12.2(B)                                              1/1/1
-------------------------------------------------------------------------------
Total Entries : 8
-------------------------------------------------------------------------------
===============================================================================
*A:PE-1#
```

The main difference between normal prefix LFA and RLFA is that for RLFA a two-MPLS label stack is pushed by the head-end node (PE-1). The top label is the SR-label to reach the PQ node (for example, 20004 for PE-4) and the bottom label is the SR-label to reach the destination node (for example, 20007 for PE-7). The notation inside the show command is bottom-label/top-label.

Figure 314 illustrates the RLFA traffic path protecting the link between PE-1 and PE-7:

Figure 314 **RLFA Traffic Path During Protection**



Inside the TTM, a tunnel-flag, **has-lfa**, is set for all destination nodes that have LFA protection available. The last two tunnels are adjacency tunnels and have in addition the flag **is-adjacency-tunnel**.

```
*A:PE-1# show router tunnel-table detail

===============================================================================
Tunnel Table (Router: Base)
===============================================================================
Destination      : 192.0.2.2/32
NextHop          : 192.168.12.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m34s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524291               Preference       : 11
Tunnel Label     : 20002                Tunnel Metric    : 10
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.0.2.3/32
NextHop          : 192.168.12.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m34s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524295               Preference       : 11
Tunnel Label     : 20003                Tunnel Metric    : 20
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.0.2.4/32
NextHop          : 192.168.12.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h09m52s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524292               Preference       : 11
Tunnel Label     : 20004                Tunnel Metric    : 30
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.0.2.5/32
NextHop          : 192.168.17.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h09m52s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524295               Preference       : 11
Tunnel Label     : 20005                Tunnel Metric    : 30
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.0.2.6/32
NextHop          : 192.168.17.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m34s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524296               Preference       : 11
Tunnel Label     : 20006                Tunnel Metric    : 20
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
```

```
Destination      : 192.0.2.7/32
NextHop          : 192.168.17.2
Tunnel Flags     : has-lfa entropy-label-capable
Age              : 00h00m34s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524294               Preference       : 11
Tunnel MTU       : 1556
Tunnel Label     : 20007                Tunnel Metric    : 10
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.168.12.2/32
NextHop          : 192.168.12.2
Tunnel Flags     : has-lfa is-adjacency-tunnel
Age              : 00h09m51s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524289               Preference       : 11
Tunnel Label     : 3                    Tunnel Metric    : 0
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Destination      : 192.168.17.2/32
NextHop          : 192.168.17.2
Tunnel Flags     : has-lfa is-adjacency-tunnel
Age              : 00h09m52s
CBF Classes      : (Not Specified)
Owner            : isis (0)             Encap            : MPLS
Tunnel ID        : 524290               Preference       : 11
Tunnel Label     : 3                    Tunnel Metric    : 0
Tunnel MTU       : 1556                 Max Label Stack  : 2
-------------------------------------------------------------------------------
Number of tunnel-table entries          : 8
Number of tunnel-table entries with LFA : 8
===============================================================================
*A:PE-1#
```

Verification of the loopback address configured within the VPRN service context on PE-7 (using loopback address 192.0.1.7/32) shows that an SR shortest path tunnel is used as the transport mechanism:

```
*A:PE-1# show router 100 route-table 192.0.1.7/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.7/32
  Protocol              : BGP_VPN
  Age                   : 00h00m57s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.7
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.7 (SR-ISIS:0 tunnel)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

Example 2: TTM preference with VPRN service

The following example is a variant on the previous example. The difference in this example is that, in addition to SR, LDP and RSVP-TE are also enabled between PE-1 and PE-7. A single RSVP LSP is configured originating at PE-1 and terminating at PE-7.

The objective of this example is to show the difference in protocol preference within TTM and how to influence the default behavior. This can be useful in case of migration scenarios from a non-SR environment toward a hybrid environment having LDP/RSVP and SR enabled.

In the following example, LFA/RLFA is no longer configured on the PE-1 node.

Translated into configuration commands, this becomes:

```
*A:PE-1# configure router isis no loopfree-alternate


*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-7"
            exit
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-7"
                to 192.0.2.7
                primary "dyn"
                exit
                no shutdown
            exit
            no shutdown
        exit
        rsvp no shutdown
        ldp
            interface-parameters
                interface "int-PE-1-PE-7" dual-stack
                    ipv4
                        no shutdown
                    exit
                    no shutdown
                exit
            exit
        exit


*A:PE-7# configure
    router
```

```
mpls
    interface "int-PE-7-PE-1"
    exit
    no shutdown
exit
rsvp no shutdown
ldp
    interface-parameters
        interface "int-PE-7-PE-1" dual-stack
            ipv4
                no shutdown
            exit
            no shutdown
        exit
    exit
exit
```

By enabling LDP and RSVP between PE-1 and PE-7, the TTM on both nodes changed. With the VPRN service between PE-1 and PE-7 of example 1, only those two specific service endpoints are displayed:

```
*A:PE-1# show router tunnel-table 192.0.2.7

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.7/32      rsvp      MPLS  1         7        192.168.17.2   10
192.0.2.7/32      ldp       MPLS  65537     9        192.168.17.2   10
192.0.2.7/32      isis (0)  MPLS  524294    11       192.168.17.2   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

```
*A:PE-7# show router tunnel-table 192.0.2.1

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        192.168.17.1   10
192.0.2.1/32      isis (0)  MPLS  524294    11       192.168.17.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-7#
```

On node PE-1, an RSVP LSP, an LDP LSP, and an SR shortest path tunnel (using IS-IS) are present. Because the VPRN service has **auto-bind-tunnel resolution any** enabled, the protocol type with the highest TTM preference (meaning the lowest absolute preference value in TTM) is taken; in this case, the RSVP LSP. This can be verified for the configured loopback address within the VPRN service context, as follows:

```
*A:PE-1# show router 100 route-table 192.0.1.7/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.7/32
  Protocol              : BGP_VPN
  Age                   : 00h00m48s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.7
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.7 (RSVP tunnel:1)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

On node PE-7, only an LDP LSP and an SR shortest path tunnel (using IS-IS) are present. Because the VPRN service has **auto-bind-tunnel resolution any** enabled, the protocol type with highest TTM preference (meaning the lowest absolute preference value in TTM) is taken; in this case, the LDP LSP. This can be verified for the configured loopback address within the VPRN service context, as follows:

```
*A:PE-7# show router 100 route-table 192.0.1.1/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.1/32
  Protocol              : BGP_VPN
  Age                   : 00h01m03s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.1
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.1 (LDP tunnel)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-7#
```

Some configuration changes are possible to change this default behavior:

- It is possible to change the **auto-bind-tunnel resolution any** command into **auto-bind-tunnel resolution filter**. Because this is a service-specific parameter, the operator has the choice to only configure this on one specific service endpoint. From a migration point of view, a smooth and easy SR migration is possible, not affecting any other deployed services on this node.
- It is possible to change the SR tunnel-table protocol preference on a node. From a migration point of view, this affects all services initiating on this node.

Using the current example, PE-1 implements the auto-bind-tunnel change (option 1), while PE-7 implements the TTM preference change (option 2).

First, a resolution-filter CLI context within VPRN service 100 on node PE-1 must be created. The example uses a **resolution-filter** context, which uses a filter to only allow SR shortest path tunnels (IS-IS based) and is configured as follows:

```
*A:PE-1# configure service vprn 100 auto-bind-tunnel resolution-filter sr-isis
```

Then, change the **auto-bind-tunnel resolution any** command into **resolution filter** on PE-1, as follows:

```
*A:PE-1# configure service vprn 100 auto-bind-tunnel resolution filter
```

As a result, the RSVP LSP is no longer used. Instead, the SR shortest path tunnel is used for the traffic from PE-1 to PE-7:

```
*A:PE-1# show router 100 route-table 192.0.1.7/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.7/32
  Protocol              : BGP_VPN
  Age                   : 00h00m12s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.7
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.7 (SR-ISIS:0 tunnel)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
```

```
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

The VPRN service on node PE-7 is still using the LDP LSP as transport mechanism to reach node PE-1 at this point. Because the previous CLI change is only done within the VPRN service context 100 on PE-1, only the direction from PE-1 to PE-7 is affected.

Another way to influence the default TTM preference is shown as follows on the PE-7 node. Using the default behavior, the LDP LSP is used, because of the preference value of 9. If the SR tunnel table preference value is lowered to a value smaller than LDP, for instance 4, the SR shortest path tunnels originating on this node will always have preference compared to LDP LSP. Translated into a configuration command, this becomes:

```
*A:PE-7# configure router isis segment-routing tunnel-table-pref 4


*A:PE-7# show router tunnel-table 192.0.2.1

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
   Color
-------------------------------------------------------------------------------
192.0.2.1/32      isis (0)  MPLS  524294    4        192.168.17.1   10
192.0.2.1/32      ldp       MPLS  65537     9        192.168.17.1   10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-7#
```

As a result, the LDP LSP is no longer used and the SR shortest path tunnel is the preferred transport tunnel:

```
*A:PE-7# show router 100 route-table 192.0.1.1/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.1/32
  Protocol              : BGP_VPN
  Age                   : 00h00m25s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.1
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.1 (SR-ISIS:0 tunnel)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-7#
```

At this point, within the VPRN service, the SR shortest path tunnels are used bidirectionally between PE-1 and PE-7.

If, for example, an operator configures explicit SDP binding within the same VPRN service on both endpoints, the explicit SDPs will always have preference. In this example, manual SDPs are configured on nodes PE-1 and PE-7, both using LDP, as follows:

```
*A:PE-1# configure
    service
        sdp 17 mpls create
            far-end 192.0.2.7
            ldp
            no shutdown
        exit
        vprn 100
            spoke-sdp 17 create
            exit
        exit


*A:PE-7# configure
    service
        sdp 71 mpls create
            far-end 192.0.2.1
            ldp
            no shutdown
        exit
        vprn 100
            spoke-sdp 71 create
            exit
        exit
```

As a result, SR shortest path tunnels are no longer used, but rather LDP-based SDPs are used instead:

```
*A:PE-1# show router 100 route-table 192.0.1.7/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix               : 192.0.1.7/32
  Protocol                : BGP_VPN
  Age                     : 00h00m34s
  Preference              : 170
  Indirect Next-Hop       : 192.0.2.7
    Label                 : 524285
    QoS                   : Priority=n/c, FC=n/c
    Source-Class          : 0
    Dest-Class            : 0
    ECMP-Weight           : N/A
    Resolving Next-Hop  : 192.0.2.7 (SDP tunnel:17)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-1#
```

```
*A:PE-7# show router 100 route-table 192.0.1.1/32 extensive

===============================================================================
Route Table (Service: 100)
===============================================================================
Dest Prefix             : 192.0.1.1/32
  Protocol              : BGP_VPN
  Age                   : 00h00m35s
  Preference            : 170
  Indirect Next-Hop     : 192.0.2.1
    Label               : 524285
    QoS                 : Priority=n/c, FC=n/c
    Source-Class        : 0
    Dest-Class          : 0
    ECMP-Weight         : N/A
    Resolving Next-Hop  : 192.0.2.1 (SDP tunnel:71)
      Label             : 524285
      Metric            : 10
      ECMP-Weight       : N/A
-------------------------------------------------------------------------------
No. of Destinations: 1
===============================================================================
*A:PE-7#
```

Conclusion

Segment Routing is a technique using extensions of the existing link state protocols, and using existing MPLS or IPv6 infrastructure as the data plane. It is a source routing technique similar to RSVP-TE, but without the need to run an extra signaling protocol. SR also avoids other scaling restrictions of associated RSVP-TE, such as midpoint state. SR is simple to control and operate because the intelligence and state are part of the packet, not held by the network. Other benefits are that SR can be introduced in an incremental way using different migration scenarios to assure a smooth transition.

Shared Risk Link Groups for RSVP-Based LSP

This chapter provides information about Shared Risk Link Groups for RSVP-Based LSPs.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This feature is applicable to SR OS. See the release notes for a full and up to date list of supported hardware.

This chapter was initially written for SR OS release 7.0.R5, but the CLI in the current edition corresponds to SR OS release 16.0.R3. There are no prerequisites.

Overview

Introduction

Shared Risk Link Group (SRLG) is a feature which allows the user to establish a backup secondary label switched path (LSP) path or a fast-reroute (FRR) LSP path which is disjoint from the path of the primary LSP. Links which are members of the same SRLG represent resources which share the same risk. For example, fiber links sharing the same conduit or multiple wavelengths sharing the same fiber.

A typical application of the SRLG feature is to provide an automatic placement of secondary backup LSPs or FRR bypass/detour LSPs that minimizes the probability of fate sharing with the path of the primary LSP.

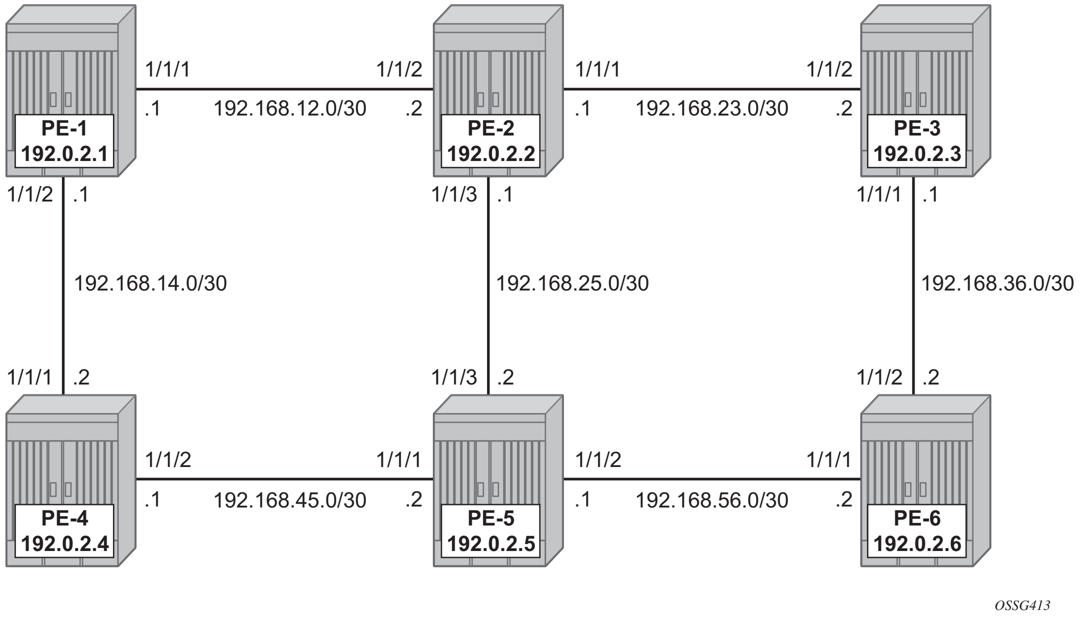
SRLG groups are used to determine which links belong to the same SRLG. The mechanism is similar to Multi-Protocol Label Switching (MPLS) admin groups. To advertise SRLG, the information is part of the IGP TE parameters in an opaque link state advertisement (LSA). The SRLG is advertised in a new Shared Risk Link Group TLV (type 138) in IS-IS (RFC 4205, *Intermediate System to Intermediate System (IS-IS) Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)*). It is advertised in a new SRLG sub-TLV (type 16) of the existing Link TLV in OSPF (RFC 4203, *OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)*.

For FRR, a choice can be made on what to do when no FRR tunnel can be found with the SRLG constraints. No FRR tunnel might be signaled or a FRR tunnel might be signaled not taking the SRLG constraints into account.

SRLG

Figure 315 shows the example topology for this chapter.

Figure 315 **Example Topology**



A single IGP area (IS-IS in this case) with traffic engineering (TE) enabled is required for the SRLG feature to work properly.

When OSPF is used as the IGP, the functionality is similar.

Configuration

Configuring the IP/MPLS network

IS-IS, MPLS, and RSVP are configured on all interfaces. TE is enabled in IS-IS. Optionally, admin groups “green” and “red” are configured on all nodes. The “green” links are the following: the link between PE-1 and PE-2, the link between PE-2 and PE-3, and the link between PE-3 and PE-6. The “red” links are: the link between PE-1 and PE-4, the link between PE-4 and PE-5, and the link between PE-5 and PE-6. The remaining link is the link between PE-2 and PE-5, which does not belong to an admin group. For more information about admin groups, see chapter RSVP Point-to-Point LSPs.

In addition, ECMP is set to 2, instead of the default value 1 in order to highlight the application of SRLG in the final example: SRLG database.

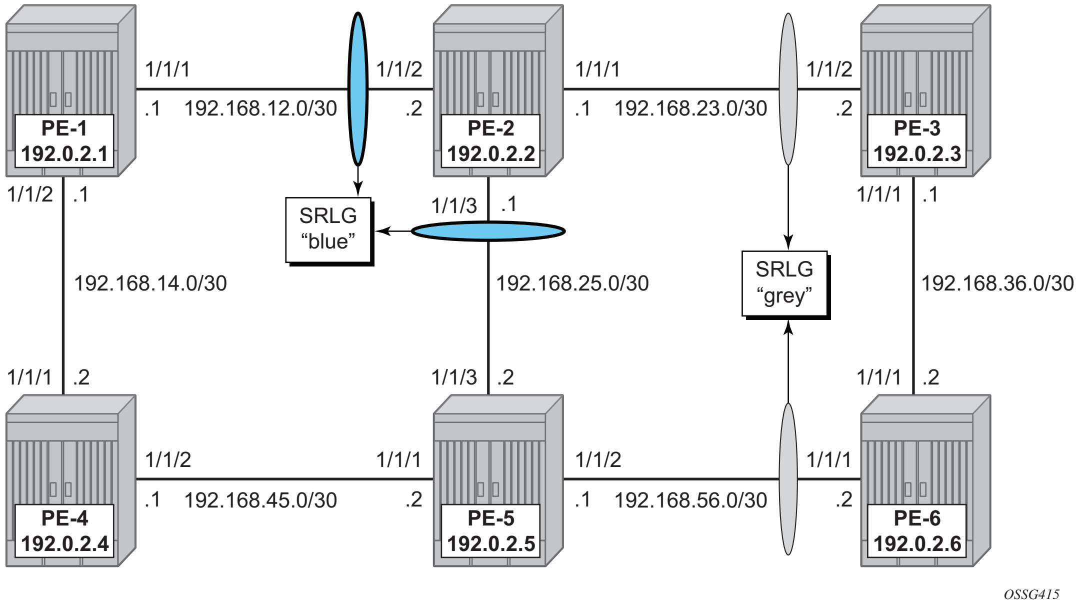
```
*A:PE-1# configure router ecmp 2
```

Define SRLG groups

Define the SRLG groups, and link them to the related MPLS interfaces.

Two SRLG groups are defined, named blue and gray, as shown in Figure 316.

Figure 316 **SRLG Topology**



The configuration of the blue SRLG group is only mandatory on PE-1, PE-2, and PE-5, while the gray SRLG group is only mandatory on PE-2, PE-3, PE-5, and PE-6. However, it is good practice to configure both SRLG groups on all nodes, for example as follows on PE-1.

```
*A:PE-1# configure router if-attribute srlg-group blue value 1
*A:PE-1# configure router if-attribute srlg-group gray value 2
```

The IP/MPLS interfaces need to be linked to the related SRLG group, which is a uni-directional indicator, applying only to the egress direction; therefore, it needs to be configured on both sides of the IP/MPLS interface. For example on PE-1, the interface to PE-2 is part of **srlg-group "blue"**. An interface can be part of multiple SRLG groups similar to the admin-group functionality.

```
*A:PE-1>config>router>mpls# info
----------------------------------------------
            interface "system"
                no shutdown
            exit
            interface "int-PE-1-PE-2"
                admin-group "green"
                no shutdown
            exit
            interface "int-PE-1-PE-4"
                admin-group "red"
                no shutdown
            exit
            no shutdown

*A:PE-1#
```

```
configure
    router
        mpls
            interface "int-PE-1-PE-2"
                srlg-group blue
            exit
```

The same must be done on PE-2, PE-3, PE-5, and PE-6. Afterward, verify the MPLS configuration for example on PE-2, where the SRLG groups are linked to the interfaces. Admin groups are configured in parallel to indicate that both can be configured and will work independently.

```
*A:PE-2>config>router>mpls# info
----------------------------------------------
        interface "system"
            no shutdown
        exit
        interface "int-PE-2-PE-1"
            admin-group "green"
            srlg-group "blue"
            no shutdown
        exit
        interface "int-PE-2-PE-3"
            admin-group "green"
            srlg-group "gray"
            no shutdown
        exit
        interface "int-PE-2-PE-5"
            srlg-group "blue"
            no shutdown
        exit
        no shutdown
```

The SRLG configuration can be verified using the following show commands.

The following shows all SRLG groups on the node:

```
*A:PE-2# show router if-attribute srlg-group

======================================================================
Interface Srlg Groups
======================================================================
Group Name                       Group Value      Penalty Weight
----------------------------------------------------------------------
blue                             1                0
gray                             2                0
----------------------------------------------------------------------
No. of Groups: 2
======================================================================
*A:PE-2#
```

In the following list of MPLS interfaces, admin groups and SRLG groups are indicated.

```
A:PE-2# show router mpls interface
```

```
===============================================================================
MPLS Interfaces
===============================================================================
Interface                           Port-id          Adm    Opr    TE-metric
-------------------------------------------------------------------------------
system                              system           Up     Up     None
  Admin Groups                      None
  SRLG Groups                       None
int-PE-2-PE-1                       1/1/2            Up     Up     None
  Admin Groups                      green
  SRLG Groups                       blue
int-PE-2-PE-3                       1/1/1            Up     Up     None
  Admin Groups                      green
  SRLG Groups                       gray
int-PE-2-PE-5                       1/1/3            Up     Up     None
  Admin Groups                      None
  SRLG Groups                       blue
-------------------------------------------------------------------------------
Interfaces : 4
```

To verify the SRLG groups in the IGP TE database, the following command can be used. The output can be extensive, but searching on the SRLG group name will lead to the correct interfaces.

The following output shows the link-state advertisements of PE-2 on PE-1 in this case. The SRLG information is linked to the IP interfaces in a dedicated TE-TLV.

```
*A:PE-1# show router isis database PE-2.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================

Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : PE-2.00-00                                Level     : L1
Sequence  : 0x6                    Checksum  : 0x3cc1   Lifetime  : 1168
Version   : 1                      Pkt Type  : 18       Pkt Ver   : 1
Attributes: L1                     Max Area  : 3        Alloc Len : 508
SYS ID    : 1920.0000.2002         SysID Len : 6        Used Len  : 508

TLVs :

---snip---

TE SRLGs     :
    SRLGs : PE-1.00
    Lcl Addr  : 192.168.12.2
    Rem Addr  : 192.168.12.1
    Num SRLGs       : 1
        1

---snip---

  TE SRLGs     :
    SRLGs : PE-3.00
    Lcl Addr  : 192.168.23.1
    Rem Addr  : 192.168.23.2
    Num SRLGs       : 1
        2

---snip---

  TE SRLGs     :
    SRLGs : PE-5.00
    Lcl Addr  : 192.168.25.1
    Rem Addr  : 192.168.25.2
    Num SRLGs       : 1
        1

---snip---
```

On-Line Verification

An on-line verification can be done by a **tools perform** command. This will trigger a Constrained Shortest Path First (CSPF) call to the Interior Gateway Protocol (IGP) TE database, and the result will be an Explicit Route Object (ERO) object which can potentially be used to set up a CSPF-based LSP.

The following shows the command syntax.

```
*A:PE-1# tools perform router mpls cspf
  - cspf to <ip-addr> [from <ip-addr>] [bandwidth <bandwidth>]
            [include-bitmap <bitmap>] [exclude-bitmap <bitmap>] [hop-limit <limit>]
            [exclude-address <excl-addr> [<excl-addr>...(up to 8 max)]]
            [use-te-metric] [strict-srlg] [srlg-group <grp-id>...(up to 8 max)]
            [exclude-node <excl-node-id> [<excl-node-id>..(up to 8 max)]]
            [skip-interface <interface-name>] [ds-class-type <class-type>]
            [cspf-reqtype <req-type>] [least-fill-min-thd <thd>]
            [setup-priority <val>] [hold-priority <val>]

 <ip-addr>            : a.b.c.d
 <bandwidth>          : [1..100000] in Mbps
 <bitmap>             : [0..4294967295] - accepted in decimal, hex(0x) or binary(0b)
 <limit>              : [2..255]
 <excl-addr>          : a.b.c.d (outbound interface)
 <use-te-metric>      : keyword
 <strict-srlg>        : keyword
 <grp-id>             : [0..4294967295]
 <excl-node-id>       : [a.b.c.d] (outbound interface)
 <interface-name>     : [max 32 chars]
 <class-type>         : [0..7]
 <req-type>           : all|random|least-fill : keywords
 <thd>                : [1..100]
 <priority>           : [0..7]
```

Where the relevant parameters are:

- **to** — Defines the far-end address of the LSP. This is the system-address of the destination LER
- **srlg-group** — Specifies which SRLG groups should be avoided while building the path to the destination (ERO object)
- **strict-srlg** — Indicates whether the SRLG group is a strict requirement or not. When this parameter is given, only paths without traversing the SRLG will be displayed.

Example:

On PE-1, a CSPF calculation is made with PE-3 as destination, without any SRLG restrictions, as follows:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(no Diffserv) class: 0 , setup Priority 7,
    Hold Priority 0 TE Class: 7

CSPF Path
To        : 192.0.2.3
Path 1    : (cost 20)
    Src:   192.0.2.1   (= Rtr)
   Egr:  192.168.12.1      -> Ingr:  192.168.12.2       Rtr:  192.0.2.2     (met 10)
   Egr:  192.168.23.1      -> Ingr:  192.168.23.2       Rtr:  192.0.2.3     (met 10)
    Dst:   192.0.2.3   (= Rtr)

*A:PE-1#
```

With a restriction on **srlg-group “blue”** (grp-id =1), the CSPF calculation is as follows:

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 srlg-group 1
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(no Diffserv) class: 0 , setup Priority 7,
    Hold Priority 0 TE Class: 7

CSPF Path
To        : 192.0.2.3
Path 1    : (cost 40)
    Src:   192.0.2.1   (= Rtr)
   Egr:  192.168.14.1       -> Ingr:  192.168.14.2      Rtr:  192.0.2.4     (met 10)
   Egr:  192.168.45.1       -> Ingr:  192.168.45.2      Rtr:  192.0.2.5     (met 10)
   Egr:  192.168.56.1       -> Ingr:  192.168.56.2      Rtr:  192.0.2.6     (met 10)
      1 SRLGs:  2
   Egr:  192.168.36.2       -> Ingr:  192.168.36.1      Rtr:  192.0.2.3     (met 10)
    Dst:   192.0.2.3   (= Rtr)

*A:PE-1#
```

The path will be through PE-4, PE-5, and PE-6.

When a strict restriction is requested on **srlg-group “gray**”, no valid CSPF path toward the destination can be found.

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 srlg-group 2 strict-srlg
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(no Diffserv) class: 0 , setup Priority 7,
    Hold Priority 0 TE Class: 7

MINOR: CLI No CSPF path to "192.0.2.3" with specified constraints.
*A:PE-1#
```

Removing the **strict** restriction results in a successful return of CSPF, indicating that the CSPF path is not SRLG disjoint.

```
*A:PE-1# tools perform router mpls cspf to 192.0.2.3 srlg-group 2
Req CSPF for all ECMP paths
    from: this node to: 192.0.2.3 w/(no Diffserv) class: 0 , setup Priority 7,
    Hold Priority 0 TE Class: 7

CSPF Path
To        : 192.0.2.3 (NOT SRLG DISJOINT)
Path 1    : (cost 20)
    Src:   192.0.2.1   (= Rtr)
   Egr:  192.168.12.1       -> Ingr:  192.168.12.2       Rtr:  192.0.2.2    (met 10)
      1 SRLGs:  1
   Egr:  192.168.23.1       -> Ingr:  192.168.23.2       Rtr:  192.0.2.3    (met 10)
      1 SRLGs:  2
    Dst:   192.0.2.3   (= Rtr)

*A:PE-1#
```

The best practice for debugging is to enable debug-tracing on the CSPF process, with following command:

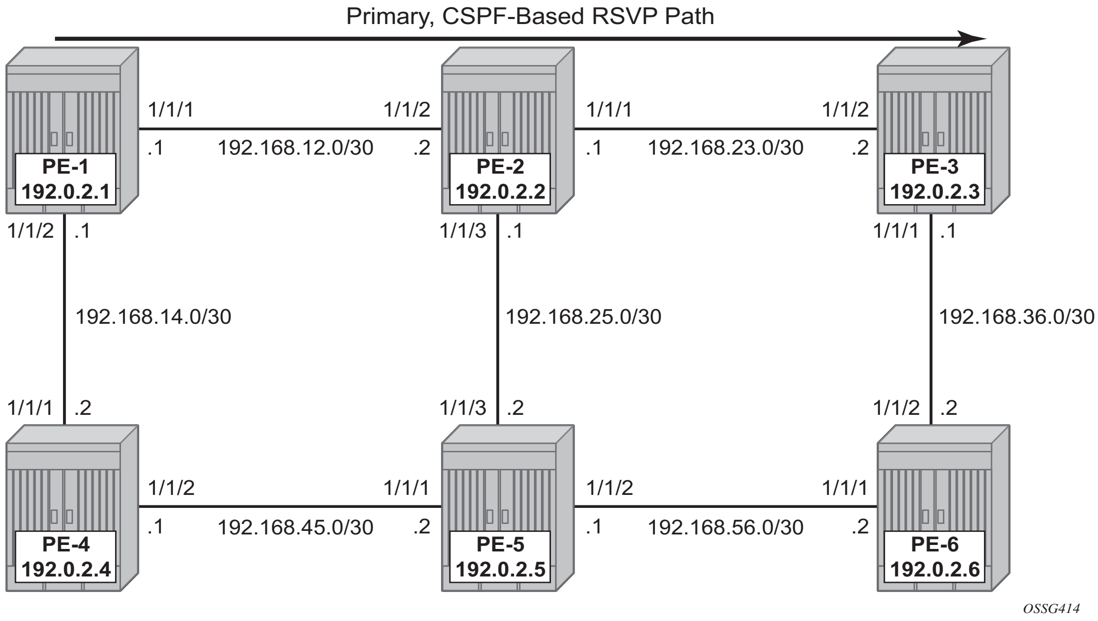
```
*A:PE-1# debug router isis cspf
```

SRLG for FRR

The fast-reroute mechanism used here is facility link protection (**fast-reroute facility no node-protect**). The SRLG feature is independent of the FRR type and works for all combinations (facility versus one-to-one, link versus node protection).

Configure an LSP from PE-1 to PE-3, and enable CSPF.

Figure 317 **Path Primary RSVP_TE LSP**



The configuration of the LSP”LSP-PE-1-PE-3_FRR_facility-link” is based on an empty path, with FRR facility link protection enabled.

```
*A:PE-1# configure
    router
        mpls
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-3_FRR_facility-link"
                to 192.0.2.3
                cspf
                fast-reroute facility
                    no node-protect
                exit
                primary "dyn"
                exit
                no shutdown
            exit
```

To verify the primary path, **oam lsp-trace** command can be used, checking the intermediate nodes.

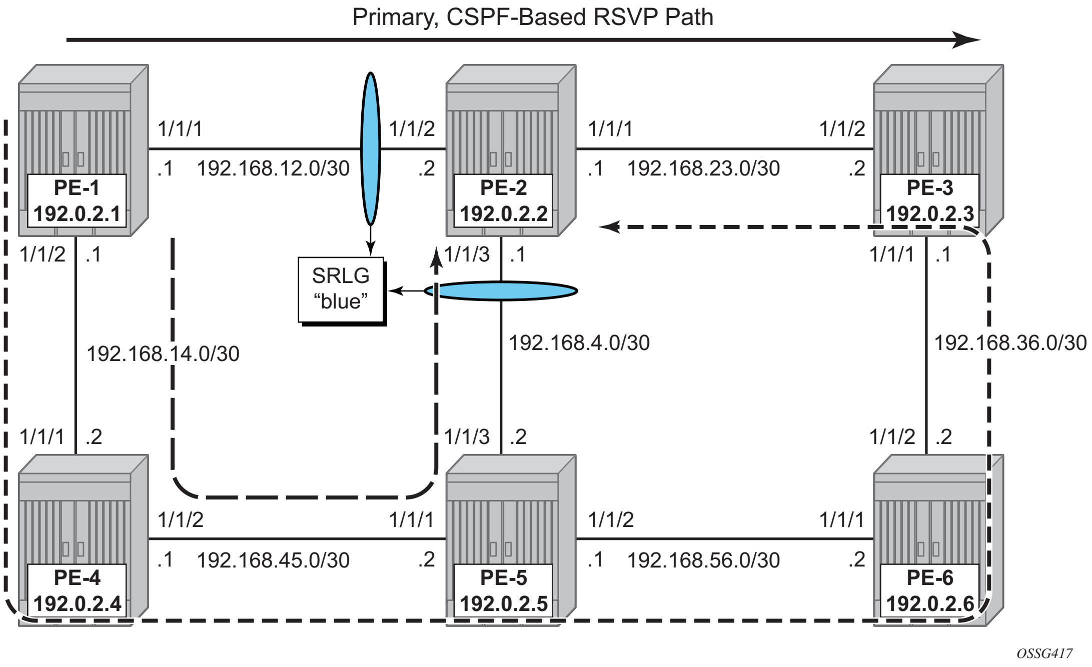
```
*A:PE-1# oam lsp-trace "LSP-PE-1-PE-3_FRR_facility-link" detail
lsp-trace to LSP-PE-1-PE-3_FRR_facility-link: 0 hops min, 0 hops max,
116 byte packets
1  192.0.2.2  rtt=1.41ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.23.2 ifaddr=192.168.23.2 iftype=ipv4Numbered MRU=1564
           label[1]=524287 protocol=4(RSVP-TE)
2  192.0.2.3  rtt=1.13ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

To verify if the bypass tunnels are up and running, an indication (@)can be found in the detail output of **show router mpls ls <x> path detail** as seen in the following output.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-3_FRR_facility-link" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-3_FRR_facility-link Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-3_FRR_facility-link Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-1-PE-3_FRR_facility-link
From             : 192.0.2.1            To                   : 192.0.2.3
Admin State      : Up                   Oper State           : Up
Path Name        : dyn
Path LSP ID      : 12800                Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 524287

---snip---

Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1) @                   Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2) @                   Record Label        : 524287
 -> 192.168.23.2 (192.0.2.3)                     Record Label        : 524287
Computed Hops    :
    192.168.12.1(S)
 -> 192.168.12.2(S)
 -> 192.168.23.2(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 20
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-1#
```

Two links are protected: one bypass tunnel originates in PE-1 protecting the link between PE-1 and PE-2. Another bypass tunnel originates in PE-2 protecting the link between PE-2 and PE-3. The focus is on the bypass tunnel originating in PE-1. When SRLG is enabled, the bypass tunnel originating in PE-1 will have different hops. The expected paths followed by the bypass tunnels originating in PE-1 with and without SRLG are shown in Figure 318.

Figure 318 FRR Bypass Tunnels Originating in PE-1 With and Without SRLG



To verify the bypass data path on the point of local repair (PLR) PE-1, the following command can be used.

```
*A:PE-1# show router mpls bypass-tunnel detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.12.2-61441
-------------------------------------------------------------------------------
To              : 192.168.25.1        State               : Up
Out I/F         : 1/1/2               Out Label           : 524287
Up Time         : 0d 00:11:31         Active Time         : n/a
Reserved BW     : 0 Kbps              Protected LSP Count : 1
Type            : Dynamic             Bypass Path Cost    : 30
Setup Priority  : 7                   Hold Priority       : 0
Class Type      : 0
Exclude Node    : None                Inter-Area          : False
Computed Hops   :
    192.168.14.1(S)                   Egress Admin Groups :
                                      red
 -> 192.168.14.2(S)                   Egress Admin Groups :
                                      red
 -> 192.168.45.2(S)                   Egress Admin Groups : None
 -> 192.168.25.1(S)                   Egress Admin Groups : None
Actual Hops     :
    192.168.14.1 (192.0.2.1)          Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)          Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)          Record Label        : 524286
 -> 192.168.25.1 (192.0.2.2)         Record Label        : 524286
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

===============================================================================
*A:PE-1#
```

The SRLG restriction is not taken into account at this moment at PLR PE-1. The actual hops are PE-4, PE-5, and PE-2 visualized by the path with the long dashes in Figure 318.

To take the SRLG restrictions into account, the following additional configuration is needed for MPLS on PE-1.

```
*A:PE-1>config>router>mpls# srlg-
srlg-database  srlg-frr

*A:PE-1>config>router>mpls# srlg-frr
  - no srlg-frr
  - srlg-frr [strict]

 <strict>             : keyword


*A:PE-1# configure router mpls srlg-frr strict
```

The option **strict** should only be used if the logical topology allows this. In other words, one must be sure that an alternative path is possible which avoids SRLG-groups.

Note: Enabling or disabling SRLG for FRR is a system-wide configuration and requires the MPLS routing instance to be manually set to shutdown and then to no shutdown to activate the change. This may cause service outage. Nokia recommends that the operator incorporates the SRLG into the initial network design and implementation to minimize the traffic loss.

```
*A:PE-1# configure router rsvp shutdown
*A:PE-1# configure router rsvp no shutdown
```

The bypass tunnel originating in PLR PE-1 can be verified with a previously used command.

```
*A:PE-1# show router mpls bypass-tunnel detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.168.12.2-61442
-------------------------------------------------------------------------------
To             : 192.168.23.1        State                : Up
```

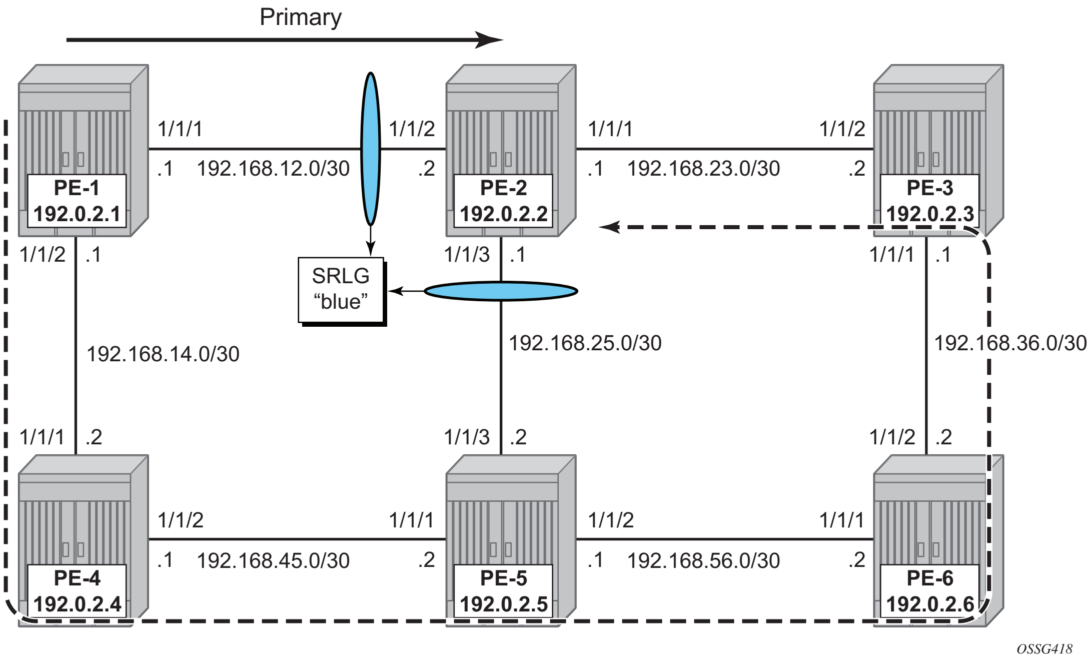
```
Out I/F        : 1/1/2              Out Label           : 524287
Up Time        : 0d 00:00:42        Active Time         : n/a
Reserved BW    : 0 Kbps             Protected LSP Count : 1
Type           : Dynamic            Bypass Path Cost    : 50
Setup Priority : 7                  Hold Priority       : 0
Class Type     : 0
Exclude Node   : None               Inter-Area          : False
Computed Hops  :
    192.168.14.1(S)                 Egress Admin Groups :
                                    red
 -> 192.168.14.2(S)                 Egress Admin Groups :
                                    red
 -> 192.168.45.2(S)                 Egress Admin Groups :
                                    red
 -> 192.168.56.2(S)                 Egress Admin Groups :
                                    green
 -> 192.168.36.1(S)                 Egress Admin Groups :
                                    green
 -> 192.168.23.1(S)                 Egress Admin Groups : None
Actual Hops    :
    192.168.14.1 (192.0.2.1)        Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)        Record Label        : 524287
 -> 192.168.45.2 (192.0.2.5)        Record Label        : 524286
 -> 192.168.56.2 (192.0.2.6)        Record Label        : 524286
 -> 192.168.36.1 (192.0.2.3)        Record Label        : 524285
 -> 192.168.23.1 (192.0.2.2)        Record Label        : 524286
Last Resignal  :
Attempted At   : n/a                Resignal Reason     : n/a
Resignal Status: n/a                Reason              : n/a

===============================================================================
*A:PE-1#
```

This path taking the SRLG constraints into account is represented by the line with the short dashes in Figure 318.

SRLG for Standby Path

Where SRLG groups can be constraints for bypass tunnels, they can also be a constraint to set up a secondary path. Figure 319shows that the secondary path is expected to follow the dashed line instead of passing over the direct link between PE-5 and PE-2.

Figure 319 **SRLG for Secondary Path**



An LSP is configured with a primary and a secondary path, which have no hops defined. The configuration of the LSP will need a specific indication at the level of the secondary path to enable the restriction on the srlg-groups.

```
*A:PE-1# configure
    router
        mpls
            path "prim"
                no shutdown
            exit
            path "secon"
                no shutdown
            exit
            lsp "LSP-PE-1-PE-2-srlg"
                to 192.0.2.2
                cspf
                primary "prim"
                exit
                secondary "secon"
                    standby
                    srlg
                exit
                no shutdown
            exit
```

Where both paths are empty paths, the ERO object creation solely relies on CPSF without any specific hop.

To verify the data path, the detailed output of the **show router mpls** lsp <..> path command can be used, as well as the **lsp-trace** OAM command. This output shows both ERO objects of the primary and secondary path.

```
*A:PE-1# show router mpls lsp "LSP-PE-1-PE-2-srlg" path detail

===============================================================================
MPLS LSP LSP-PE-1-PE-2-srlg Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-2-srlg Path prim
-------------------------------------------------------------------------------
---snip---

ExplicitHops:
    No Hops Specified
Actual Hops      :
    192.168.12.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.12.2 (192.0.2.2)                     Record Label        : 524285
ComputedHops:
    192.168.12.1(S)
 -> 192.168.12.2(S)
ResigEligib*: False
LastResignal: n/a                                CSPF Metric : 10
-------------------------------------------------------------------------------
LSP LSP-PE-1-PE-2-srlg Path secon
-------------------------------------------------------------------------------
---snip---

ExplicitHops:
    No Hops Specified
Actual Hops      :
    192.168.14.1 (192.0.2.1)                     Record Label        : N/A
 -> 192.168.14.2 (192.0.2.4)                     Record Label        : 524286
 -> 192.168.45.2 (192.0.2.5)                     Record Label        : 524285
 -> 192.168.56.2 (192.0.2.6)                     Record Label        : 524285
 -> 192.168.36.1 (192.0.2.3)                     Record Label        : 524284
 -> 192.168.23.1 (192.0.2.2)                     Record Label        : 524284
ComputedHops:
    192.168.14.1(S)
 -> 192.168.14.2(S)
 -> 192.168.45.2(S)
 -> 192.168.56.2(S)
 -> 192.168.36.1(S)
 -> 192.168.23.1(S)
Srlg        : Enabled
SrlgDisjoint: True
ResigEligib*: False
LastResignal: n/a                                CSPF Metric : 50
===============================================================================
*A:PE-1#
```

The **lsp-trace** command can be used for secondary path as well. The intermediate LSRs and the MPLS labels used can be clearly seen.

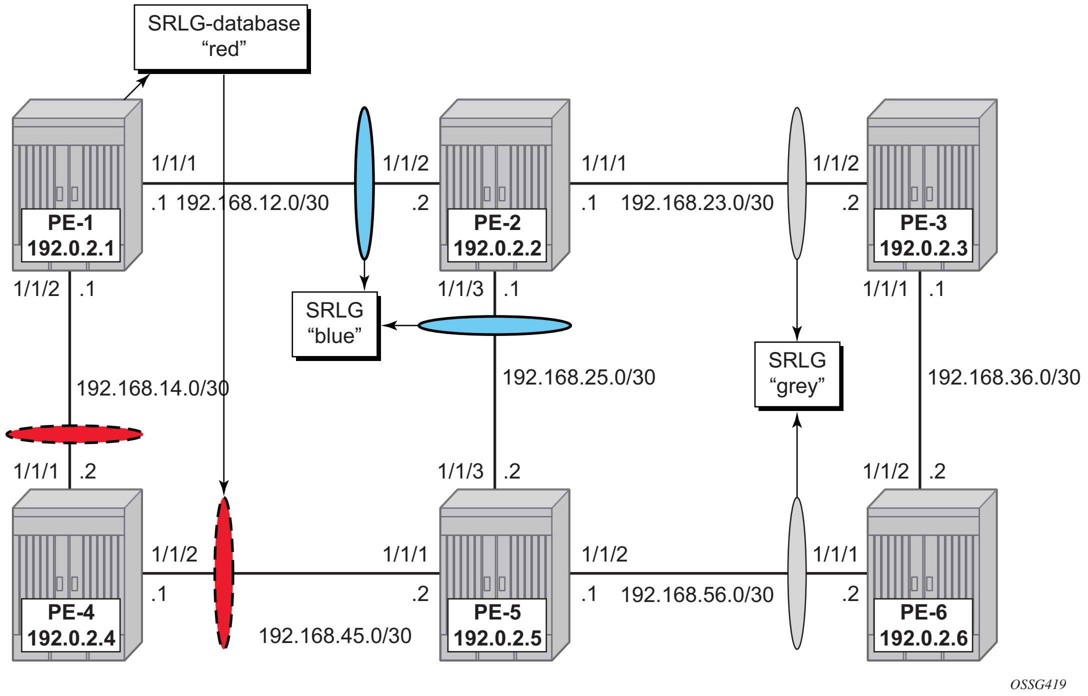
```
*A:PE-1# oam lsp-trace "LSP-PE-1-PE-2-srlg" path "secon" detail
lsp-trace to LSP-PE-1-PE-2-srlg: 0 hops min, 0 hops max, 116 byte packets
1  192.0.2.4  rtt=0.639ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.45.2 ifaddr=192.168.45.2 iftype=ipv4Numbered MRU=1564
           label[1]=524285 protocol=4(RSVP-TE)
2  192.0.2.5  rtt=1.20ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.56.2 ifaddr=192.168.56.2 iftype=ipv4Numbered MRU=1564
           label[1]=524285 protocol=4(RSVP-TE)
3  192.0.2.6  rtt=2.18ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.36.1 ifaddr=192.168.36.1 iftype=ipv4Numbered MRU=1564
           label[1]=524284 protocol=4(RSVP-TE)
4  192.0.2.3  rtt=1.92ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=192.168.23.1 ifaddr=192.168.23.1 iftype=ipv4Numbered MRU=1564
           label[1]=524284 protocol=4(RSVP-TE)
5  192.0.2.2  rtt=3.94ms rc=3(EgressRtr) rsc=1
*A:PE-1#
```

SRLG Database

In case not all IP/MPLS routers in the area support SRLG, a static SRLG database can be created on the systems which will be used as an additional constraint when performing the CSPF calculation to define the path.

Figure 320 shows an example where an additional SRLG group “red” is defined on PE-1, with information related to the interface between PE-4 and PE-5.

Figure 320 SRLG Database Example



```
*A:PE-1# configure router if-attribute srlg-group "red" value 3
```

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-4"
                srlg-group "red"
            exit
            srlg-database
                router-id 192.0.2.4
                    interface 192.168.45.1 srlg-group "red"
                    no shutdown
                exit
                router-id 192.0.2.5
                    interface 192.168.45.2 srlg-group "red"
                    no shutdown
                exit
            exit
```

This information is local to PE-1 and will only have effect on CSPF calculations on PE-1, not on the other nodes.

When a CSPF calculation is done for a path from PE-1 to PE-5, the result will be two equal-cost paths, because ECMP equals 2. When adding the **srlg-group "red"** as a restriction, only a single path will be found, passing PE-2.

Conclusion

Interpreting the SRLG information into the TE database makes it possible to protect an LSP even when multiple IP/MPLS interfaces fail as a result of an underlying transmission failure. Transmission failures can occur quite often because not all transmission links are one to one protected.

SRLG groups in MPLS provide a very dynamic and simple way to assure LSP FRR path protection on every PLR throughout the followed LSP path. The SRLG groups are also taken into account when defining the ERO for secondary paths, at least if the configured secondary path is empty.

For interoperability reasons, the SRLG-database is available, because systems can link interfaces to an SRLG with interconnecting systems that do not support the SRLG feature; so they cannot advertise the SRLG information through the IGP.

The creation and maintenance of an SRLG database requires operational effort and systems that do not support SRLG will never take any SRLG information into account during CSPF calculation for the creation of FRR bypass or detour tunnels.

Static Point-to-Point LSPs

This chapter provides information about static point-to-point label switched paths (LSPs).

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS and was originally written for release 7.0.R5. The output in the current edition corresponds to release 16.0.R3. There are no pre-requisites or conditions on the hardware for this configuration.

Overview

Due to the connectionless nature of the network layer protocol IP, packets travel through the network on a hop-by-hop basis with routing decisions made at each node. As a result, hyperaggregation of data on certain links may occur and it may impact the provider's ability to provide guaranteed service levels across the network end-to-end. To address these shortcomings, multiprotocol label switching (MPLS) was developed. The technology provides the capability to establish connection-oriented paths, called label switched paths (LSPs), over a connectionless (IP) network. The LSP offers a mechanism to engineer network traffic independently from the underlying network routing protocol (mostly IP) to improve the network resiliency and recovery options and to permit delivery of new services that are not readily supported by conventional IP routing techniques (Layer 2 IP Virtual Private Networks (VPNs)). These benefits are essential for today's communication network explaining the wide deployment base of the MPLS technology.

RFC 3031, *Multiprotocol Label Switching Architecture,* specifies the MPLS architecture while this document describes the configuration and troubleshooting of static point-to-point LSPs on SR OS. Point-to-point LSPs can also be dynamically established using a label signaling protocol, such as label distribution protocol (LDP) or resource reservation protocol (RSVP). See chapters LDP Point-to-Point LSPs and RSVP Point-to-Point LSPs.

Packet Forwarding

As a packet of a connectionless network layer protocol travels from one router to the next, each router in the network makes an independent forwarding decision by performing the following basic tasks: first analyzing the packet header, then referencing the local routing table to find the longest match based on the destination address in the IP header, and finally sending out the packet on the selected interface. In other terms, the first function partitions the entire set of possible packets into a set of forwarding equivalence classes (FECs). All packets associated to a particular FEC will be forwarded along the same logical path to the same destination. The second function maps each FEC to a next hop destination router. Each router along the path performs these actions.

On the other hand, in MPLS, the assignment of a particular packet to a particular FEC is done just once, as the packet enters the network. In turn, the FEC is mapped to an LSP, which is established prior to any data flowing. An MPLS label, representing the FEC to which the packet is assigned, is attached to the packet (push operation) and once labeled, the packet is forwarded to the next hop router along that LSP path. At subsequent hops, there is no further analysis of the network layer header of the packet. Instead, the label is used as an index into a table which specifies the next hop and a new label. The old label is replaced with the new label (swap operation), and the packet is forwarded to its next hop. At the MPLS network egress, the label is removed from the packet (pop operation). If this router is the final destination (based on the remaining packet), the packet is handed to the receiving application (such as a virtual private LAN service (VPLS) domain). If this router is not the final destination of the packet, the packet will be sent into a new MPLS tunnel or forwarded by conventional IP forwarding toward the layer 3 destination.

Terminology

Figure 321 **Generic MPLS Network, MPLS Label Operations**

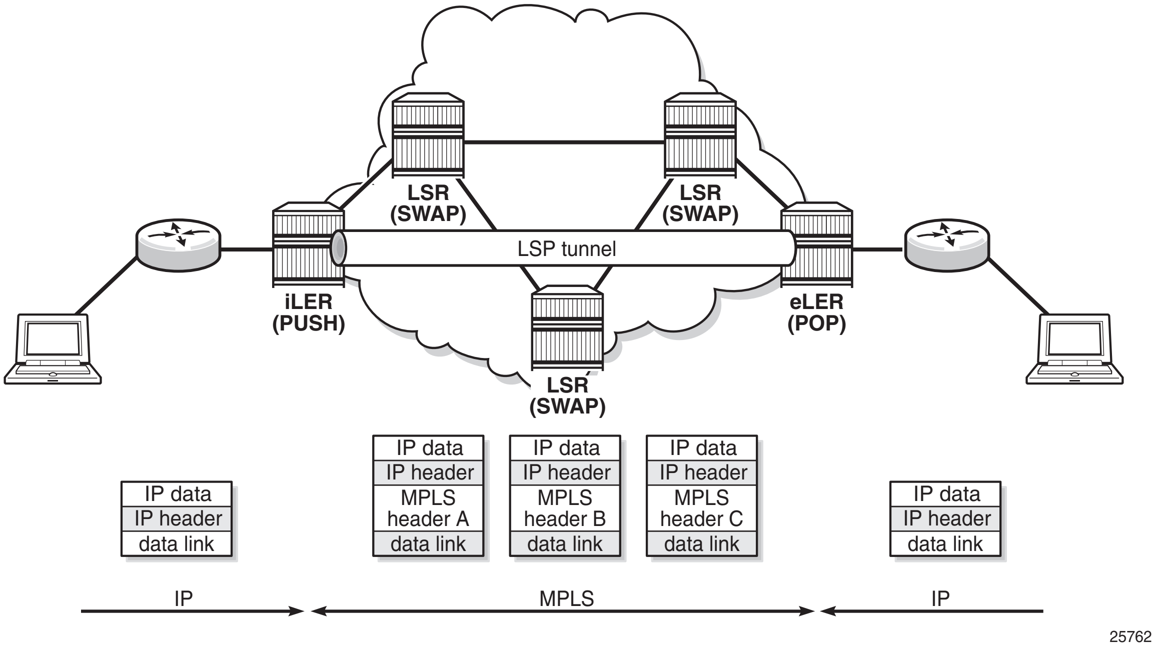


Figure 321 shows a general network topology clarifying the MPLS-related terms. A Label Edge Router (LER) is a device at the edge of an MPLS network, with at least one interface outside the MPLS domain. A router is usually defined as an LER based on its position relative to a particular LSP. The MPLS router at the head-end of an LSP is called the ingress label edge router (iLER). The MPLS router at the tail-end of an LSP is called the egress label edge router (eLER). The iLER receives unlabeled packets from outside the MPLS domain, then applies MPLS labels to the packets, and forwards the labeled packets into the MPLS domain. The eLER receives labeled packets from the MPLS domain, then removes the labels, and forwards unlabeled packets outside the MPLS domain. The last LSR before the eLER can be configured with an implicit-null label (numeric value 3). This LSR will pop the outer label and send MPLS packets without an outer label to the eLER. This is known as Penultimate Hop Popping (PHP). A Label Switching Router (LSR) is a device internal to an MPLS network, with all interfaces inside the MPLS domain. These devices switch labeled packets inside the MPLS domain. In the core of the network, LSRs ignore the network layer (IP) header of the packet and simply forward the packet using the MPLS label swapping mechanism.

A single LSP is unidirectional. In common practice, because the bidirectional nature of most traffic flows is implied, the term LSP often is used to define the pair of LSPs that enable the bidirectional flow. For ease of terminology and discussion however, the LSP in this chapter is referred to as a single entity.

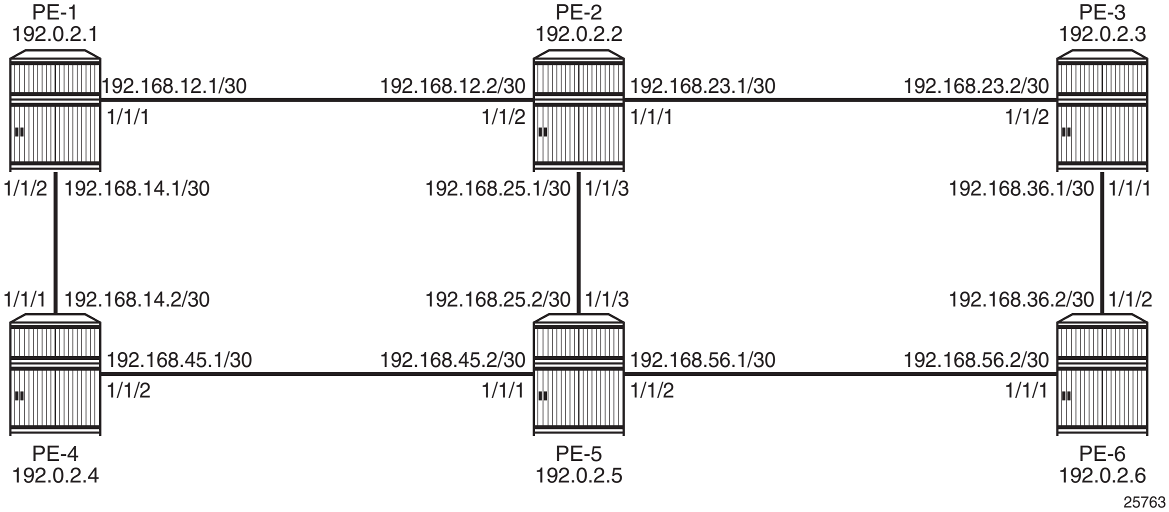
LSP Establishment

Prior to packet forwarding, the LSP must be established. In order to do so, labels need to be distributed for the path. For static LSPs, the label distribution is done manually by the network administrator. Although a high control level of the labels in use is achieved, the LSP cannot enjoy the resilience and recovery functionality the dynamic label signaling protocols can offer.

Example Topology

The example topology is displayed in Figure 322. The setup consists of six 7750 SR nodes located in a single autonomous system.

Figure 322 **MPLS Example Topology**



Configuration

For static LPSs, there is no need for an IGP.

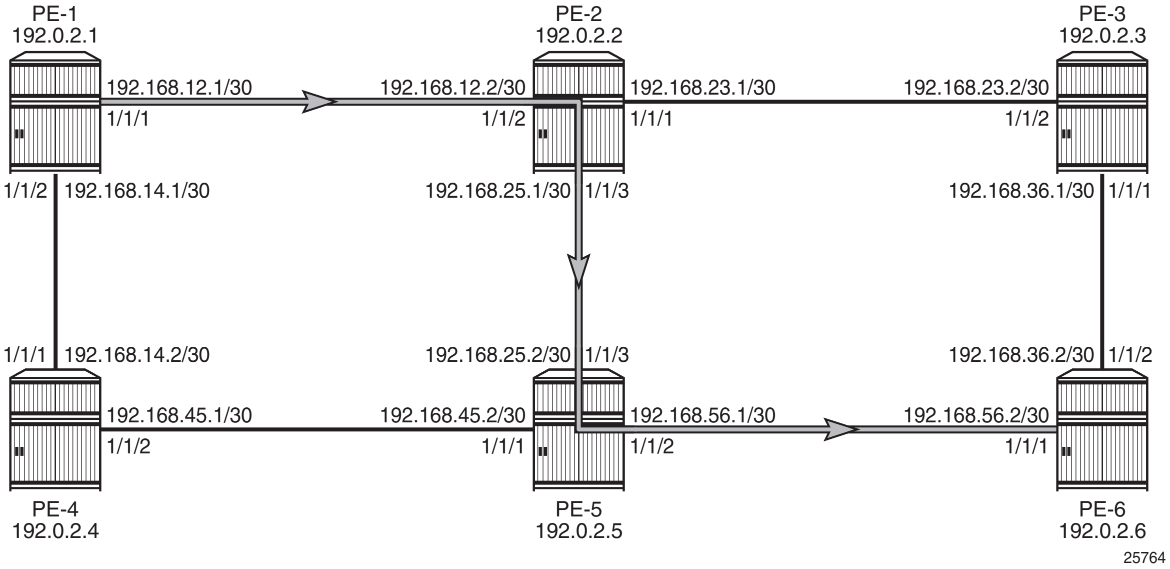
For LSPs that are set up manually, the first step is to enable MPLS on all network interfaces that will be used to carry LSPs. MPLS is automatically enabled on the system IP addresses.

For manually configured LSPs, any interface used by the static LSP must be added into the MPLS protocol instance, even though RSVP is not actually used to signal labels. For router PE-1, this results in the following configuration:

```
*A:PE-1# configure
    router
        mpls
            interface "int-PE-1-PE-2"
            exit
            interface "int-PE-1-PE-4"
            exit
            no shutdown
```

As an example, a static LSP will be created starting from PE-1, running over PE-2 and PE-5, then terminating on PE-6 as shown in Figure 323.

Figure 323 **Static LSP Running over PE-1, PE-2, PE-5, PE-6**



Verify the acceptable label range for use with static configurations for each node; as follows:

```
*A:PE-1# show router mpls-labels label-range

=============================================================================
Label Ranges
=============================================================================
Label Type      Start Label End Label   Aging       Available   Total
-----------------------------------------------------------------------------
Static          32          18431       -           18400       18400
Dynamic         18432       524287      0           505856      505856
    Seg-Route   0           0           -           0           0
=============================================================================
*A:PE-1#
```

The label range for static LSPs extends from the value 32 to 18431. To ensure the labels have not yet been allocated to another configuration, use the command:

```
*A:PE-1# show router mpls-labels label 32 18431 in-use

=================================================================
MPLS Labels from 32 to 18431 (In-use)
=================================================================
Label               Label Type          Label Owner
-----------------------------------------------------------------
-----------------------------------------------------------------
In-use labels (Owner: All) in specified range   : 0
In-use labels in entire range                   : 0
=================================================================
*A:PE-1#
```

This command shows the number of incoming labels in use. At the iLER, the number of labels in use will remain 0 even after the static LSP has been configured where the iLER has a push operation for a label. The reason is that the labels shown are relevant to the labels that the router is generating, as for label swap or pop operations. There is no information shown about labels that other routers are advertising. For the push operation, any label can be used, even if it is not within the label range of the router pushing the label. For the originating router PE-1, the label 100 will be used for the push operation on the interface toward PE-2.

Static LSPs are configured within the MPLS configuration context, but do not rely on dynamic label signaling.

The configuration of the MPLS static LSP head-end PE-1 contains:

- The system IP address of the destination router PE-6 (to).
- A push operation of the label 100.
- The interface address facing the current node of the next hop along the static path, which is PE-2 (nexthop).

```
*A:PE-1# configure
    router
        mpls
            static-lsp "LSP-PE-1-PE-6-static"
```

```
                to 192.0.2.6
                push 100 nexthop 192.168.12.2
                no shutdown
            exit
```

The transit LSRs (PE-2 and PE-5) perform swap operations and forward the packet to the manually defined next-hop. On the LSR under the context of the interface on which the incoming LSP arrives, the correct label is selected (label-map) and in this context a swap operation with a new label and the new next hop (nexthop) is entered.

```
*A:PE-2# configure
    router
        mpls
            interface "int-PE-2-PE-1"
                label-map 100
                    swap 150 nexthop 192.168.25.2
                    no shutdown
                exit
                no shutdown
            exit


*A:PE-5# configure
    router
        mpls
            interface "int-PE-5-PE-2"
                label-map 150
                    swap 200 nexthop 192.168.56.2
                    no shutdown
                exit
```

The terminating router PE-6 performs a pop operation and forwards the now unlabeled packets external to the MPLS domain.

```
*A:PE-6# configure
    router
        mpls
            interface "int-PE-6-PE-5"
                label-map 200
                    pop
                    no shutdown
                exit
```

To verify the operational status of the static LSP configuration, the **show router mpls static-lsp** command is used on the iLER. A static LSP is considered to be operationally up when its next-hop is reachable. Since there is no check whether the end-to-end LSP path is up (the LSP connectivity to the eLER is never verified), it can be that the static LSP path is broken while the iLER displays an operational enabled LSP.

```
*A:PE-1# show router mpls static-lsp

===============================================================================
MPLS Static LSPs (Originating)
```

```
================================================================================
LSP Name     To              Next Hop        Out Label Up/Down Time   Adm  Opr
 ID           Metric          Oper Metric     Out Port
--------------------------------------------------------------------------------
LSP-PE-1-PE-6-static
             192.0.2.6       192.168.12.2    100       0d 00:04:31    Up   Up
 1           N/A             N/A             1/1/1
--------------------------------------------------------------------------------
LSPs : 1
================================================================================
*A:PE-1#
```

On the LSR, the **transit** keyword is added to the command. On PE-2:

```
*A:PE-2# show router mpls static-lsp transit

================================================================================
MPLS Static LSPs (Transit)
================================================================================
In Label    In Port     Out Label   Out Port    Next Hop            Adm   Opr
--------------------------------------------------------------------------------
100         1/1/2       150         1/1/3       192.168.25.2        Up    Up
--------------------------------------------------------------------------------
LSPs : 1
================================================================================
*A:PE-2#
```

On LSR PE-5:

```
*A:PE-5# show router mpls static-lsp transit

================================================================================
MPLS Static LSPs (Transit)
================================================================================
In Label    In Port     Out Label   Out Port    Next Hop            Adm   Opr
--------------------------------------------------------------------------------
150         1/1/3       200         1/1/2       192.168.56.2        Up    Up
--------------------------------------------------------------------------------
LSPs : 1
================================================================================
*A:PE-5#
```

On the terminating router (eLER), the keyword **terminate** is added, as follows:

```
*A:PE-6# show router mpls static-lsp terminate

================================================================================
MPLS Static LSPs (Terminate)
================================================================================
In Label    In Port     Out Label   Out Port    Next Hop            Adm   Opr
--------------------------------------------------------------------------------
200         1/1/1       n/a         n/a         n/a                 Up    Up
--------------------------------------------------------------------------------
LSPs : 1
================================================================================
*A:PE-6#
```

To track the label action associated with the static LSP configuration, the **show router mpls interface label-map** command can be used on all LSRs and eLERs (not on the iLER).

```
*A:PE-2# show router mpls interface label-map

===============================================================================
MPLS Interfaces (Label-Map)
===============================================================================
In Label  In I/F     Out Label Out I/F    Next Hop          Type      Adm  Opr
-------------------------------------------------------------------------------
100       1/1/2      150       1/1/3      192.168.25.2      Static    Up   Up
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-2#
```

```
*A:PE-6# show router mpls interface label-map

===============================================================================
MPLS Interfaces (Label-Map)
===============================================================================
In Label  In I/F     Out Label Out I/F    Next Hop          Type      Adm  Opr
-------------------------------------------------------------------------------
200       1/1/1      n/a       n/a        n/a               Static    Up   Up
-------------------------------------------------------------------------------
Interfaces : 1
===============================================================================
*A:PE-6#
```

The **show router mpls status** command is used to verify each of the LSP types, the number of configured LSPs and whether they originate on, transit through or terminate on the router.

```
*A:PE-1# show router mpls status

===============================================================================
MPLS Status
===============================================================================
Admin Status              : Up          Oper Status               : Up
Oper Down Reason          : n/a
FRR Object                : Enabled     Resignal Timer            : Disabled
Hold Timer                : 1 seconds   Next Resignal             : N/A
Srlg Frr                  : Disabled    Srlg Frr Strict           : Disabled
Admin Group Frr           : Disabled
Dynamic Bypass            : Enabled     User Srlg Database        : Disabled
BypassResignalTimer       : Disabled    BypassNextResignal        : N/A
LeastFill Min Thd         : 5 percent   LeastFill Reopti Thd      : 10 percent
Local TTL Prop            : Enabled     Transit TTL Prop          : Enabled
AB Sample Multiplier      : 1           AB Adjust Multiplier      : 288
Exp Backoff Retry         : Disabled    CSPF On Loose Hop         : Disabled
Lsp Init RetryTimeout     : 30 seconds  MBB Pref Current Hops     : Disabled
Logger Event Bundling     : Disabled
RetryIgpOverload          : Disabled

P2mp Resignal Timer       : Disabled    P2mp Next Resignal        : N/A
```

```
Sec FastRetryTimer        : Disabled    Static LSP FR Timer       : 30 seconds
P2P Max Bypass Association: 1000
P2PActPathFastRetry       : Disabled    P2MP S2L Fast Retry       : Disabled
In Maintenance Mode       : No
MplsTp                    : Disabled
Next Available Lsp Index  : 1
Entropy Label RSVP-TE     : Enabled     Entropy Label SR-TE       : Enabled
PCE Report RSVP-TE        : Disabled    PCE Report SR-TE          : Disabled

===============================================================================
MPLS LSP Count
===============================================================================
                          Originate         Transit           Terminate
-------------------------------------------------------------------------------
Static LSPs               1                 0                 0
Dynamic LSPs              0                 0                 0
Detour LSPs               0                 0                 0
P2MP S2Ls                 0                 0                 0
MPLS-TP LSPs              0                 0                 0
Mesh-P2P LSPs             0                 N/A               N/A
One Hop-P2P LSPs          0                 N/A               N/A
SR-TE LSPs                0                 N/A               N/A
Mesh-P2P SR-TE LSPs       0                 N/A               N/A
One Hop-P2P SR-TE LSPs    0                 N/A               N/A
===============================================================================
*A:PE-1#
```

Penultimate Hop Popping (PHP) can be used with static LSPs. This is achieved by configuring the penultimate LER to swap the incoming label to implicit-null instead of a specific label value (the label-map must be shut down to add the **swap** command).

```
*A:PE-5# configure
    router
        mpls
            interface "int-PE-5-PE-2"
                label-map 150
                    shutdown
                    swap implicit-null-label nexthop 192.168.56.2
                    no shutdown
                exit
```

The previous configuration will cause PE-5 to pop the top label from the incoming labeled frame received from PE-2 and send it to PE-6 without adding another outer label. The result can be seen from the following command (label 3 is never actually pushed onto a frame).

```
*A:PE-5# show router mpls static-lsp transit

===============================================================================
MPLS Static LSPs (Transit)
===============================================================================
In Label    In Port     Out Label   Out Port    Next Hop            Adm   Opr
-------------------------------------------------------------------------------
150         1/1/3       3           1/1/2       192.168.56.2        Up    Up
-------------------------------------------------------------------------------
LSPs : 1
```

```
===============================================================================
*A:PE-5#
```

If the traffic arriving at PE-5 was IP with a single label, then it would arrive at PE-6 as unlabeled IP traffic.

If the static LSP spans a single hop (PE-1 to PE-2), the ingress LER would push the implicit-null instead of pushing a label.

```
*A:PE-1# configure
    router
        mpls
            static-lsp "LSP-PE-1-PE-2-static"
                to 192.0.2.2
                push implicit-null-label nexthop 192.168.12.2
                no shutdown
            exit
```

In this case, no MPLS action (swap or pop) is required for this LSP on PE-2.

Conclusion

MPLS provides the capability to establish connection-oriented paths over a connectionless network. The static LSP offers a mechanism to engineer network traffic In this chapter, the configuration of static LSPs is given together with the associated show output which can be used to verify and troubleshoot.

Tunneling of ICMP Reply Packets over MPLS LSPs

This chapter provides information about tunneling of ICMP reply packets over MPLS LSPs.

Topics in this chapter include:

- Applicability
- Overview
- Configuration
- Conclusion

Applicability

This chapter is applicable to SR OS routers and was initially written for SR OS Release 13.0.R7. The CLI in this edition corresponds to release 15.0.R1. Internet Control Message Protocol (ICMP) tunneling over Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) is supported in SR OS release 12.0.R4 or later.

Overview

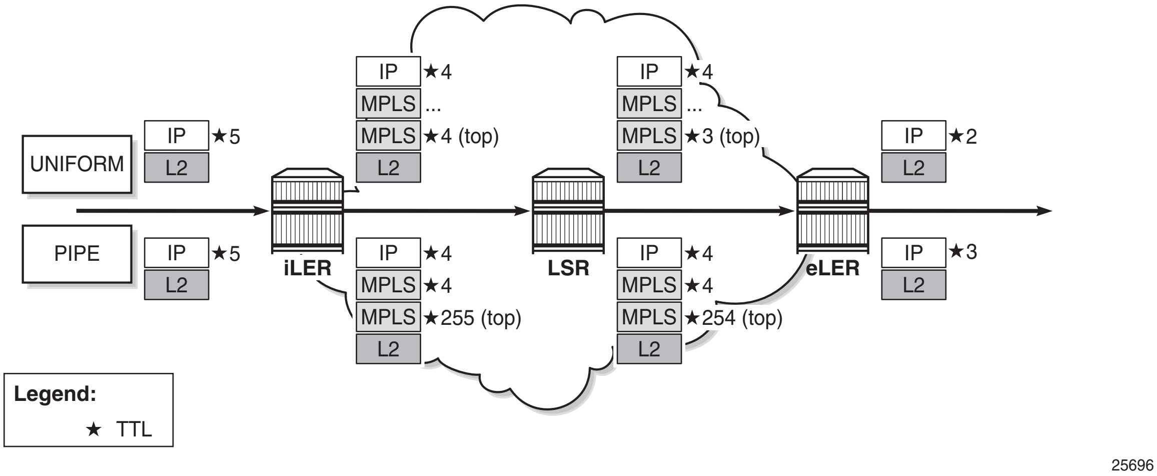
In IP forwarding, Time-To-Live (TTL) is a well-known mechanism to mitigate the damage in case of a loop. The TTL value in the IP header is decremented by one at each hop and the packet is discarded when the TTL equals 0. TTL is also used in traceroute, where the first batch of echo requests are sent with TTL equal to 1, the second batch of echo requests is sent with TTL equal to 2, and so on. Any intermediate node where the TTL expires (is decremented to 0) sends an ICMP reply of type "Time exceeded" (type 11) to the sender. From the replies, the sequence of hops can be determined.

If ICMP messages are sent in an MPLS tunnel, in pipe mode, the hops in the tunnels are invisible and the TTL is only decremented by the Label Edge Routers (LERs), not by the intermediate Label Switching Routers (LSRs). However, there are two modes for TTL handling, according to *RFC 3443: Time To Live Processing in MPLS Networks*:

1. Uniform mode: the MPLS network is visible from the outside. MPLS nodes use the TTL in the same way as any other IP node.
2. Pipe mode: the MPLS network is invisible from the outside. MPLS use of TTL is independent from IP TTL use. The network appears like a pipe between ingress Label Edge Router (iLER) and egress Label Edge Router (eLER).

Both TTL uses are shown in Figure 324:

Figure 324 **Use of TTL: Uniform versus Pipe**



Independent of the mode, the iLER decrements the TTL in the IP header by one. The iLER adds service and transport MPLS headers.

Note: In an L2 Virtual Private Network (VPN), the TTL in the IP header is kept intact.

- In uniform mode, the iLER sets the TTL of every MPLS header to match the TTL in the IP header and every LSR decrements the MPLS TTLs. The IP header remains unchanged as long as the packet is in the MPLS tunnel. The eLER pops the MPLS labels and decrements the minimum TTL of the headers (which is the TTL in both MPLS headers) by one. This TTL is used in the IP header.
- In pipe mode, the iLER sets the TTL of the top MPLS header to 255 and every LSR decrements that TTL by one. The eLER pops the MPLS labels and decrements the minimum TTL of the headers (which is the IP TTL) by one. This TTL is used in the IP header. There can be uncounted hops in pipe mode, because the LSRs are not counted.

The LERs can be in uniform mode and the LSRs in pipe mode, and vice versa.

The default use of TTL in SR OS is as follows:

- Uniform mode for LSP shortcuts: ReSource reserVation Protocol (RSVP) shortcuts, Label Distribution Protocol (LDP) shortcuts, and Border Gateway Protocol (BGP) shortcuts.
- Pipe mode for L2 and L3 VPN services, BGP labeled routes, IPv6 Provider Edge (6PE) router, and IPv6 on VPN to PE router (6VPE).

However, the use of TTL can be changed by configuration.

Figure 325 shows the use of TTL for an L2 VPN service in pipe mode. The TTL in the IP header is preserved. There is no processing of the IP header for an L2 service. The TTL in the pushed MPLS headers is 255 and the TTL in the top MPLS header is decremented by one in the LSRs. The eLER pops the MPLS labels.

Figure 325 **Use of TTL in an L2 VPN Service in Pipe Mode**

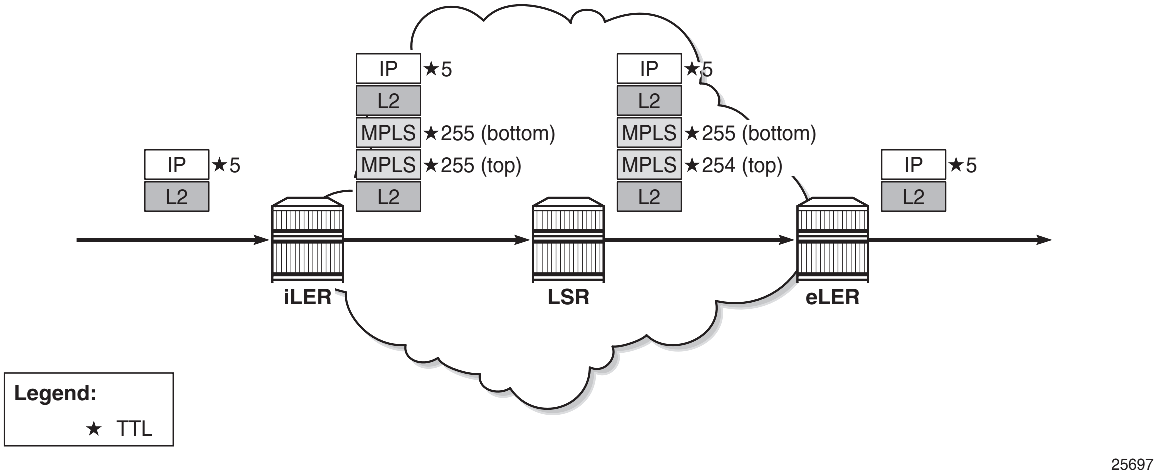
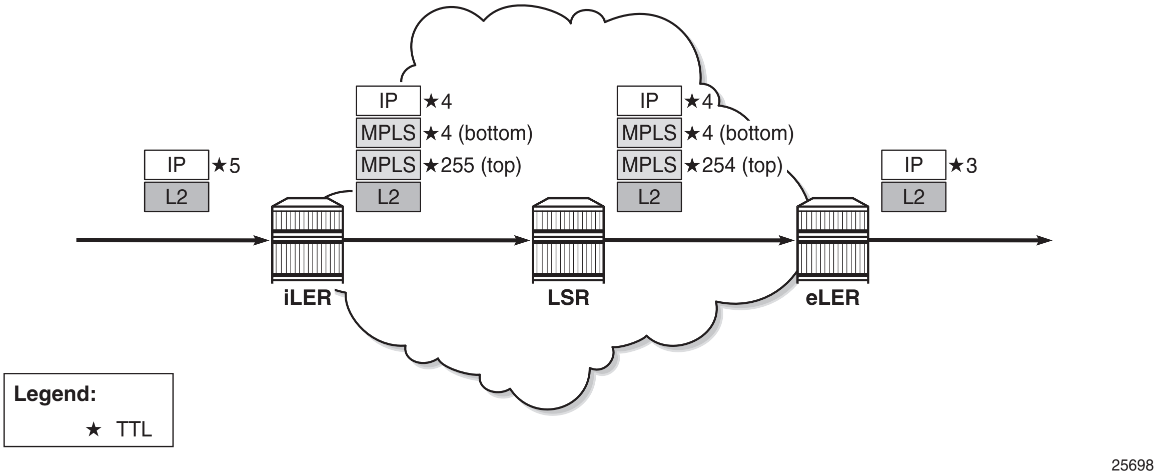


Figure 326 shows the use of TTL for an L3 VPN service in pipe mode. The TTL in the IP header is decremented by the iLER and the eLER, but not by the LSRs. In pipe mode, the bottom MPLS header inherits the IP TTL after it has been decremented by the iLER. The transport MPLS header gets TTL 255 and this TTL is decremented by one at each LSR. The eLER takes the minimum of the TTL of the MPLS headers and the IP TTL and decrements that by one. This will match the IP TTL in the forwarded packet. The MPLS labels are popped. There are uncounted hops, because the LSRs are invisible in pipe mode.

Figure 326 **Use of TTL in an L3 VPN Service in Pipe Mode**

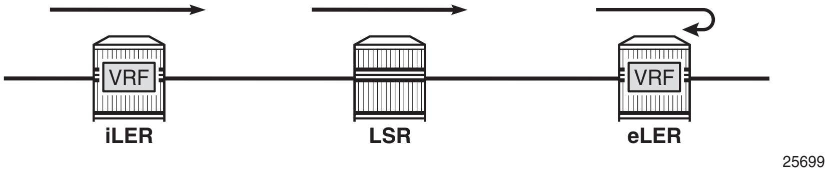


Tunneling of ICMP reply packets over MPLS LSPs provides the ability for a network operator or customer to trace the MPLS network hops in the path, for Virtual Private Routed Network (VPRN), 6PE/6VPE, and BGP labeled routes.

ICMP Tunneling over an MPLS LSP

Figure 327 shows the actions performed in iLER, LSR, and eLER, when tunneling ICMP messages over an MPLS LSP:

Figure 327 **Tunneling of ICMP Reply Packets over an MPLS LSP**



- In the iLER, uniform mode is required within the VPRN service. The IP TTL is propagated in the MPLS TTL, both for in-transit and Control Processing Module (CPM) generated IP packets. In this example, it is assumed that a UDP traceroute message is forwarded with source IP address S1.

- In all LSRs, ICMP tunneling is enabled globally on the system, according to RFC 3032, *MPLS Label Stack Encoding*. When the MPLS TTL expires in an LSR, the LSR generates an ICMP reply with code "Time exceeded" and destination IP address S1. However, the CPM sends this ICMP reply packet in the forward direction of the MPLS LSP tunnel that the packet arrived on. The ICMP reply packet is sent to the eLER, not to the iLER.
- The eLER performs a lookup for the IP address S1 and sends the ICMP reply to S1 toward the iLER.
  - The lookup of the IP address S1 is in the Global Routing Table (GRT) for BGP shortcut, 6PE, and BGP labeled route prefixes.
  - The lookup of IP address S1 is in the Virtual Routing and Forwarding (VRF) table for VPRN and 6VPE prefixes.

TTL Propagation

The TTL propagation can be configured in LERs and LSRs.

TTL Propagation at the iLER

Different commands are used for TTL propagation in a VPRN versus BGP labeled routes. Pipe mode is enabled by default in either case.

TTL Propagation in iLER for VPRN

The TTL propagation of VPN-IPv4 or VPN-IPv6 packets in a VPRN service can be enabled globally as follows:

```
*A:PE-3# configure router ttl-propagate vprn-local
  - vprn-local <ttl-prop-type>

 <ttl-prop-type>      : none|all|vc-only - Default: vc-only


*A:PE-3# configure router ttl-propagate vprn-transit
  - vprn-transit <ttl-prop-type>

 <ttl-prop-type>      : none|all|vc-only - Default: vc-only
```

There are three options for the propagation of TTL in the iLER of a VPRN:

None	No IP TTL propagation to any MPLS header in the stack: the transport and the VC MPLS headers will have a value of 255. This is needed for proper operation of traceroute in an inter-AS option B VPRN.
All	Uniform mode: IP TTL is propagated to all MPLS headers in the stack.
VC-only (default)	Pipe mode: IP TTL is propagated to the VC header, but not to the transport headers in the stack.

For more information about inter-AS option B, see chapter *Rosen MVPN Inter-AS Option B*.

In inter-AS option B, a traceroute for a VPN IP prefix issued from a Customer Edge (CE) router results in both ingress Autonomous System Boundary Router (ASBR) and egress ASBR not responding. The traceroute also misses a couple of hops if the target CE node is two or more hops away from the egress PE. The reason is that the VC label TTL inherits the decremented IP TTL at the ingress PE, but is decremented twice in the MPLS network whereas the IP TTL is only decremented at the ingress and the egress PE nodes. The option "none" for **ttl-propagate** makes the ASBRs transparent to the traceroute behavior and corrects the uncounted hop issue.

The global configuration can be overruled within each VPRN, as follows

```
*A:PE-3# configure service vprn 1 ttl-propagate local
  - local <ttl-prop-type>

 <ttl-prop-type>      : none|all|vc-only|inherit - Default: inherit


*A:PE-3# configure service vprn 1 ttl-propagate transit
  - transit <ttl-prop-type>

 <ttl-prop-type>      : none|all|vc-only|inherit - Default: inherit
```

TTL Propagation in iLER for BGP Labeled Route

IPv4 and IPv6 packets are forwarded using BGP labeled routes in the GRT, as described in *RFC 3107, Carrying Label Information in BGP-4.* This also applies to 6PE. TTL propagation for RFC 3107 label routes can be configured as follows:

```
*A:PE-3# configure router ttl-propagate label-route-local
  - label-route-local <ttl-prop-type>

 <ttl-prop-type>      : none|all - Default: none


*A:PE-3# configure router ttl-propagate label-route-transit
  - label-route-transit <ttl-prop-type>
```

```
<ttl-prop-type>      : none|all - Default: none
```

There are two options for TTL propagation in the iLER for BGP labeled routes:

None (default) Pipe mode: No TTL is propagated from the IP header to the MPLS headers in the transport MPLS stack. However, the IP TTL is propagated to the bottom header: the virtual circuit (VC) header

All Uniform mode: TTL is propagated to all headers in the transport MPLS stack.

If the BGP peer advertises the implicit-null label value for the BGP labeled route (in the case of a third-party implementation), the TTL propagation follows the configuration of the RSVP/LDP LSP shortcut that the BGP labeled route resolves to. This is not controlled by the preceding commands.

TTL Propagation at the LSR

In a VPRN service, there is no TTL propagation to be configured in the LSRs.

TTL Propagation in LSR for BGP Labeled Route

The IP TTL and VC TTL are not decremented by the LSRs. The TTL that is decremented is the minimum of the RSVP/LDP transport TTL and the BGP TTL.

Step 1. The LSR determines the TTL using the following function:

TTL = MIN {incoming transport label stack TTL, incoming swapped/stitched label TTL}

Step 2. The LSR decrements the TTL by one and writes it to the outgoing swapped/stitched BGP label.

This is always performed when an LSR is swapping or stitching a label at any stack depth.

The control plane indicates to the data plane whether a BGP labeled route is stitched or an LDP FEC is being stitched. The same node can perform stitching for one BGP labeled route and swapping for another one. See chapter LDP FEC to BGP Label Route Stitching for more information.

Step 3. The LSR can propagate the decremented TTL to the outgoing transport label stack (if any) that is pushed on top of the BGP swapped/stitched label. This is configured as follows:

```
*A:PE-2# configure router ttl-propagate lsr-label-route
  - lsr-label-route <ttl-prop-type>
```

```
<ttl-prop-type>      : none|all - Default: none
```

There are two options for TTL propagation in the LSR:

None (default)	No TTL propagation of the decremented TTL to the MPLS transport label stack.
All	TTL propagation of the decremented TTL to all LDP/RSVP transport labels.

It is safe to not propagate the TTL to the transport label stack for an ASBR/Area Border Router (ABR)/data path Route Reflector (RR)/BGP-LDP stitching node. Not propagating the TTL provides isolation of the network domains downstream of the LSRs. Operations, Administration, and Maintenance (OAM) packets, such as traceroute and ping, sent in the context of a BGP labeled route or VPRN will not expire in LSR nodes within these domains.

A node performing pseudowire (PW) switching terminates the transport label stack in pipe mode; the node ignores the TTL of the incoming transport label stack and propagates the TTL of the VC label. The TTL of the new pushed transport label stack is always 255.

Some Considerations on TTL Propagation in LSR for BGP Labeled Routes

- When an LSR stitches an LDP label to a BGP label, the decremented TTL of the stitched label can be propagated to the LDP/RSVP transport labels with the preceding configuration.
- When an LSR stitches a BGP label to an LDP label, the decremented TTL of the stitched label is automatically propagated to the RSVP label if the outgoing LDP LSP is tunneled over RSVP.
- When the LSR pops a BGP label and forwards the packet using an IGP route (IGP route is preferred over BGP labeled route), the LSR pushes an LDP label on the packet and the TTL behavior is the same as when an LSR stitches a BGP label to an LDP label.
- In a Carrier Supporting Carrier (CSC) VPRN, the ingress CSC CE swaps an iBGP label for an eBGP label and the ingress CSC PE swaps the incoming eBGP label for a VPN-IPv4 label. The reverse operation is performed by the egress CSC PE and the egress CSC CE. In all cases, the decremented TTL of the swapped label is propagated to the LDP/RSVP transport labels.
- SR OS does not support ASBR or data path RR functionality for labeled IPv6 routes in the global routing instance (6PE).

TTL Propagation at the eLER

For packets received with a BGP labeled route and searched for in the GRT, the TTL of the forwarded IP packet is set to MIN{MPLS_TTL-1, IP_TTL-1}, where MPLS_TTL refers to the TTL in the outermost label in the popped stack. This is the same behavior as for LSP shortcuts.

For packets received in the context of VPRN, the TTL of the forwarded IP packet is set to MIN{MPLS_TTL-1, VC_TTL-1, IP_TTL-1}, where MPLS_TTL refers to the TTL in the outermost label in the popped stack and VC_TTL refers to the TTL in the VC label in the popped stack.

Some Considerations on TTL Propagation at the eLER

- When a packet is received in one VPRN instance and is redirected using policy-based routing to be forwarded in another VPRN instance, the TTL is governed by the configuration of the outgoing VPRN instance.
- When a packet is received in a VPRN context but is searched for in the GRT (GRT leaking configured), the behavior of the TTL propagation is governed by:
  - the BGP labeled route configuration when the matching route is an RFC 3107 label route or a 6PE route
  - the LSP shortcut configuration when the matching route is an RSVP or LDP shortcut (default uniform mode)

    For shortcuts, uniform mode is default. Pipe mode can be configured as follows:

```
*A:PE-6# configure router ldp no shortcut-transit-ttl-propagate
*A:PE-6# configure router ldp no shortcut-local-ttl-propagate

*A:PE-6# configure router mpls no shortcut-transit-ttl-propagate
*A:PE-6# configure router mpls no shortcut-local-ttl-propagate
```

Enabling ICMP Tunneling on LSRs

For all scenarios (VPRN and BGP labeled routes), ICMP tunneling needs to be enabled on all LSRs, as follows:

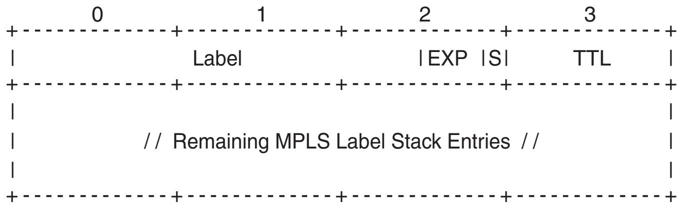
```
*A:PE-2# configure router icmp-tunneling
```

The LSR will generate the ICMP reply packet of type 11 - "Time exceeded", with source IP address set to a local address of the LSR node and appending the IP header and leading octets of the original datagram. The LSR does not perform a lookup for the destination IP address of the ICMP reply packet, which is the source IP address of the sender of the label TTL expiry packet. The CPM injects the ICMP reply packet in the forward direction toward the eLER. The TTL of pushed labels is 255.

There is no need to enable ICMP tunneling on the eLER. The eLER performs a user packet lookup in the data path in the VRF table or GRT and forwards the ICMP reply packet to the destination. If the eLER does not have a route to the destination, the packet is dropped.

RFC 4950: ICMP Extensions for Multiprotocol Label Switching defines an extension object (MPLS label stack object) that permits LSRs to include label stack information to ICMP messages; see Figure 328:

Figure 328 **MPLS Label Stack Object**



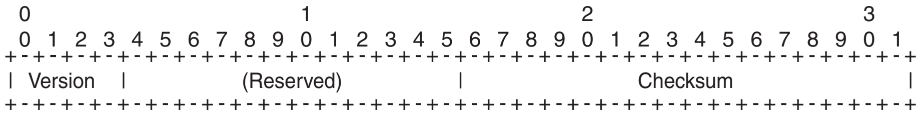
```
        0             1             2             3
+-------------+-------------+-------------+-------------+
|                 Label                  |EXP |S|   TTL   |
+-------------+-------------+-------------+-------------+
|                                                       |
|         // Remaining MPLS Label Stack Entries //      |
|                                                       |
+-------------+-------------+-------------+-------------+
```

25700

The MPLS label stack object is applicable for ICMPv4 and ICMPv6. The MPLS label stack contains the MPLS shim header: label, experimental bits for Type of Service (ToS), S-bit indicating the bottom of the stack, and TTL. The object can be appended to the ICMP Time Exceeded and ICMP Destination Unreachable messages. The LSR that sends the ICMP reply message will not change the MPLS label stack.

RFC 4884 Extended ICMP to Support Multi-Part Messages defines the ICMP extension header; see Figure 329:

Figure 329 **ICMP Extension Header**



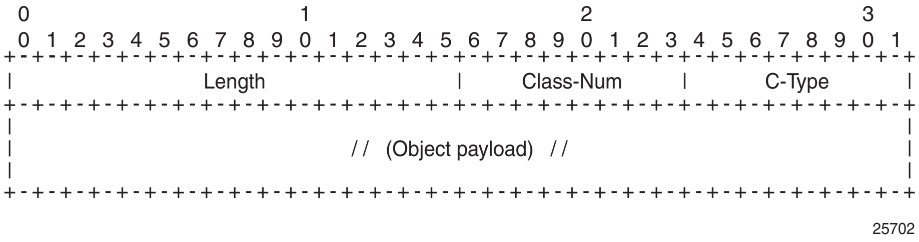
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Version |        (Reserved)       |            Checksum          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

25701

The version of the ICMP extension header is 2. The twelve reserved bits must be set to 0.

An extension object contains 32-bit words, representing an object header and payload, as defined in RFC 4884; see Figure 330:

Figure 330 **ICMP Extension Object: Object Header and Payload**



The length of the object is the length of the header (4 octets) plus the length of the object payload: 4 octets per LSR. The class number identifies the object class; in this case, object class 1 for MPLS label stack class. The C-type defines the object subtype; in this case, the subtype is 1 for an incoming MPLS label stack.

Backward compatibility is guaranteed between the ICMP message with extension header, the ICMP messages without extension header, and the ICMP message with a non-compliant extension header.

Effect of ICMP Tunneling on OAM

ICMP tunneling over MPLS LSPs affects the behavior of some CPM originated OAM packets that are forwarded within a VPRN context.

- ICMP ping and UDP traceroute are sent according to the TTL propagation configured in the VPRN context.
- VPRN-ping and VPRN-trace are not affected.

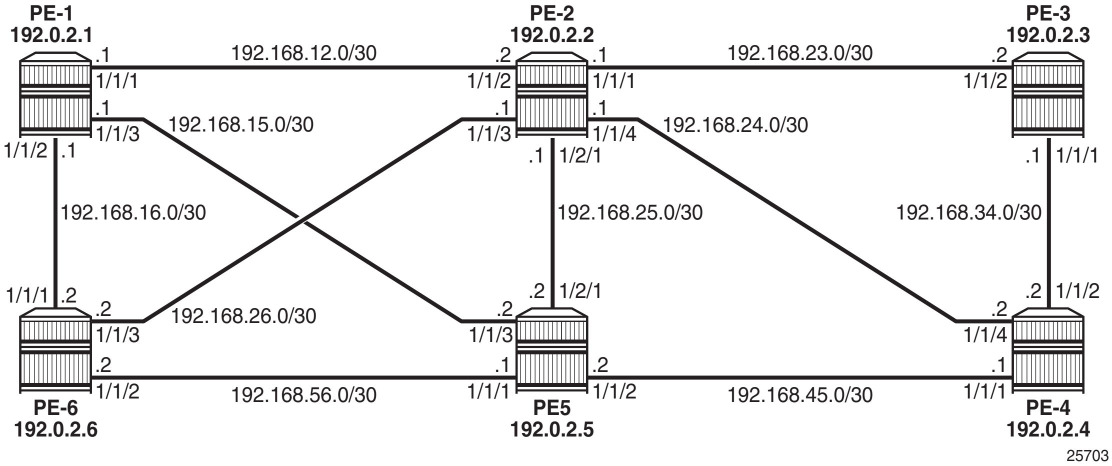
OAM packets forwarded over a BGP labeled route follow the TTL configuration of the iLER.

ICMP tunneling behavior at an LSR only applies to UDP traceroute packets. Other OAM packets expiring at the LSR, such as ICMP ping, VPRN ping, VPRN trace, LSP ping, and LSP trace, follow their specific procedures or are silently dropped.

Configuration

Figure 331 shows the example configuration, which has six 7750 SRs:

Figure 331 **Example Configuration**



Initial Configuration

The nodes have the following initial configuration:

- Cards, MDAs, ports
- Router interfaces. For PE-3:

```
*A:PE-3# configure
    router
        interface "int-PE-3-PE-2"
            address 192.168.23.2/30
            port 1/1/2
        exit
        interface "int-PE-3-PE-4"
            address 192.168.34.1/30
            port 1/1/1
        exit
        interface "system"
            address 192.0.2.3/32
        exit
```

- IGP: OSPF (alternatively, any IGP could have been used). For PE-3:

```
*A:PE-3# configure
    router
        ospf
```

```
            area 0.0.0.0
                interface "system"
                exit
                interface "int-PE-3-PE-2"
                    interface-type point-to-point
                exit
                interface "int-PE-3-PE-4"
                    interface-type point-to-point
                exit
            exit
            no shutdown
        exit
```

- Link LDP. For PE-3:

```
*A:PE-3# configure
    router
        ldp
            interface-parameters
                interface "int-PE-3-PE-2"
                exit
                interface "int-PE-3-PE-4"
                exit
            exit
```

Configure VPRN

A VPRN service is configured on PE-3 and PE-6. The routes are exchanged via BGP. BGP is configured on all nodes with PE-2 as route reflector (RR).

```
*A:PE-3# configure
    router
        autonomous-system 64496
        bgp
            group "internal"
                family vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.2
                exit
            exit
            no shutdown
```

The configuration on RR PE-2 is as follows:

```
*A:PE-2# configure
    router
        autonomous-system 64496
        bgp
            cluster 1.1.1.1
            group "internal"
                family vpn-ipv4
                peer-as 64496
                neighbor 192.0.2.1
                exit
                neighbor 192.0.2.3
                exit
                neighbor 192.0.2.4
                exit
                neighbor 192.0.2.5
                exit
                neighbor 192.0.2.6
                exit
            exit
            no shutdown
```

Import and export policies are configured on PE-3 and PE-6, as follows:

```
*A:PE-3# configure
    router
        policy-options
            begin
            community "VPN1" members "target:64496:1"
            policy-statement "VPN1-export"
                entry 10
                    from
                        protocol direct
                    exit
                    to
                        protocol bgp-vpn
                    exit
                    action accept
                        community add "VPN1"
                    exit
                exit
            exit
            policy-statement "VPN1-import"
                entry 10
                    from
                        protocol bgp-vpn
                        community "VPN1"
                    exit
                    action accept
                    exit
                exit
            exit
            commit
        exit
```

VPRN 1 is configured on PE-3 and PE-6, as follows:

```
*A:PE-3# configure
    service
        vprn 1 customer 1 create
            vrf-import "VPN1-import"
            vrf-export "VPN1-export"
            route-distinguisher 64496:13
            auto-bind-tunnel
                resolution-filter
                    ldp
                exit
                resolution filter
            exit
            interface "loopback1" create
                address 192.0.1.3/32
                loopback
            exit
            no shutdown
        exit
```

The configuration on PE-6 is similar, with loopback address 192.0.1.6/32.

Default TTL Handling in VPRN

The default configuration for TTL propagation on the iLER corresponds to pipe mode, as follows:

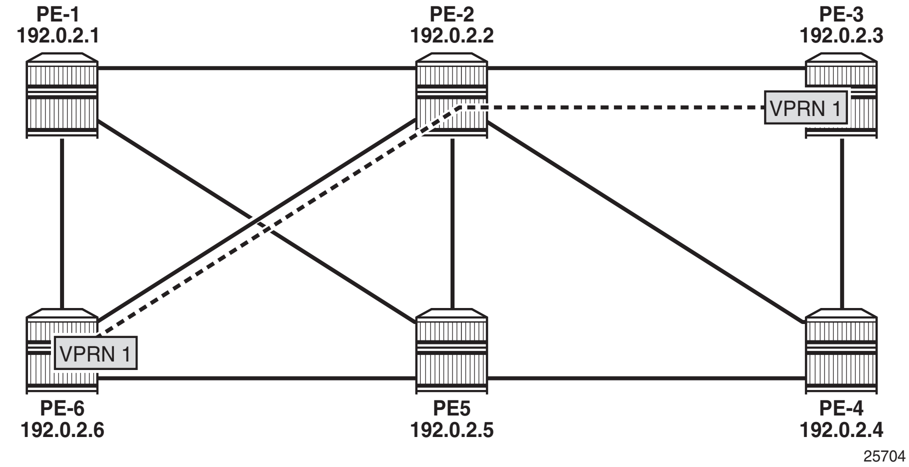
```
*A:PE-3# configure router ttl-propagate
*A:PE-3>config>router>ttl-propagate# info detail
----------------------------------------------
            label-route-local none
            label-route-transit none
            lsr-label-route none
            vprn-local vc-only
            vprn-transit vc-only
----------------------------------------------
```

No ICMP tunneling is enabled in the LSRs, which implies that no ICMP "Time exceeded" messages will be tunneled by the LSR to the eLER. A traceroute message sent in VPRN 1 from PE-3 to the loopback address in VPRN 1 on PE-6 shows that the loopback address is the next hop. There are no intermediate hops detected.

```
*A:PE-3# traceroute router 1 192.0.1.6
traceroute to 192.0.1.6, 30 hops max, 40 byte packets
  1  192.0.1.6 (192.0.1.6)    1.84 ms  1.76 ms  1.75 ms
*A:PE-3#
```

Figure 332 shows the tunnel from iLER PE-3 to eLER PE-6:

Figure 332 **Tunnel from iLER PE-3 to eLER PE-6 via LSR PE-2**



For comparison, a traceroute in the base router toward the system address of PE-6 shows PE-2 as intermediate hop, as follows:

```
*A:PE-3# traceroute 192.0.2.6
traceroute to 192.0.2.6, 30 hops max, 40 byte packets
  1  192.168.23.1 (192.168.23.1)    0.714 ms  0.687 ms  0.695 ms
  2  192.0.2.6 (192.0.2.6)    1.73 ms  1.63 ms  1.61 ms
*A:PE-3#
```

Uniform Mode in iLER and ICMP Tunneling in LSR

In the iLER PE-3, uniform mode is enabled for local VPRNs, as follows:

```
*A:PE-3# configure service vprn 1 ttl-propagate local all
```

This is a specific configuration for VPRN 1 that overrules the global configuration. By default, it is set to inherit the global configuration. By default, the global configuration is pipe mode.

This TTL propagation is only configured on PE-3, not on PE-6. This implies that traceroute messages from the VPRN in PE-3 will have TTL propagation to all MPLS labels (uniform mode), while traceroute messages from the VPRN in PE-6 will have pipe mode.

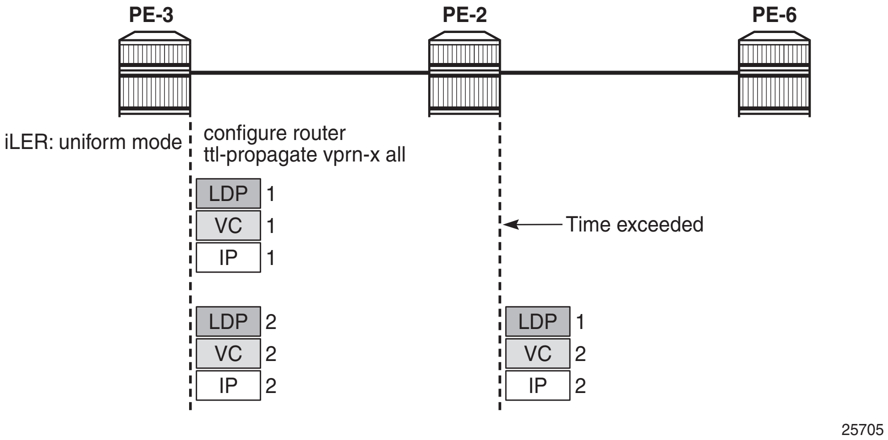
In the LSR PE-2, ICMP tunneling is enabled, as follows:

```
*A:PE-2# configure router icmp-tunneling
```

A UDP traceroute is sent from VPRN 1 on PE-3 to the loopback address in VPRN 1 on PE-6. A message with TTL 1 is sent first. The IP TTL and VC TTL are not decremented by the LSR PE-2. Only the LDP TTL is decremented, so this message times out on the LSR PE-2, which tunnels the ICMP Time Exceeded reply with destination address 192.0.1.3 toward eLER PE-6. The eLER looks up the prefix 192.0.1.3 in the VRF table and sends the ICMP Time Exceeded reply toward VPRN 1 in PE-3. Three UDP traceroute messages with TTL 1 are sent. Then, UDP traceroute messages with TTL 2 are sent. They reach the destination PE-6 before time-out.

Figure 333 shows the TTLs in the UDP traceroute messages:

Figure 333 **UDP Traceroute in VPRN with iLER in Uniform Mode**



In the following output, there is only an MPLS label stack object when TTL expires in the LSR, because ICMP tunneling is occurring. The MPLS label stack object is not used by the eLER. The LSR where ICMP tunneling occurs adds an MPLS label stack object to the ICMP reply message. The MPLS label stack object contains information about the MPLS labels (VC label and LDP transport label) in the stack: MPLS labels, experimental bits for ToS, and TTL. S indicates the bottom of the label stack. In the detailed output of the traceroute command, the MPLS label stack information is shown for the echo requests that timed out in the LSR:

```
*A:PE-3# traceroute router 1 192.0.1.6 detail
traceroute to 192.0.1.6, 30 hops max, 40 byte packets
  1   1  192.168.26.1  (192.168.26.1)  1.66 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262138, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   1, S = 1
  1   2  192.168.26.1  (192.168.26.1)  1.32 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262138, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   1, S = 1
  1   3  192.168.26.1  (192.168.26.1)  1.57 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262138, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   1, S = 1
  2   1  192.0.1.6  (192.0.1.6)  1.89 ms
  2   2  192.0.1.6  (192.0.1.6)  1.54 ms
  2   3  192.0.1.6  (192.0.1.6)  1.86 ms

*A:PE-3#
```

The top label or transport label 262138 is the LDP label pushed by PE-3:

```
*A:PE-3# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.3)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262138
192.168.23.1                                1/1/2

192.0.2.6/32                                Swap            262138    262138
192.168.23.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-3#
```

The bottom label or service label 262137 is the BGP label, which remains the same end-to-end:

```
*A:PE-3# show router bgp routes vpn-ipv4

===============================================================================
 BGP Router ID:192.0.2.3        AS:64496       Local AS:64496
===============================================================================
 Legend -
 Status codes  : u - used, s - suppressed, h - history, d - decayed, * - valid
                 l - leaked, x - stale, > - best, b - backup, p - purge
 Origin codes  : i - IGP, e - EGP, ? - incomplete
```

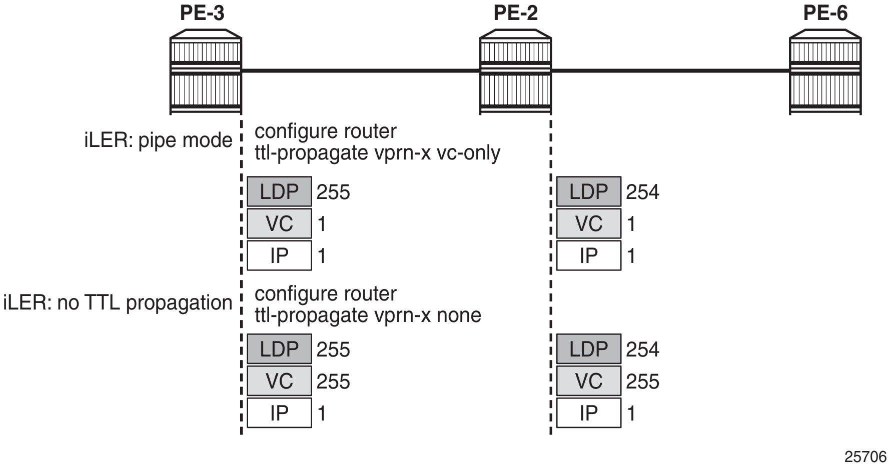
```
===============================================================================
BGP VPN-IPv4 Routes
===============================================================================
Flag  Network                                                  LocalPref   MED
      Nexthop (Router)                                         Path-Id     Label
      As-Path
-------------------------------------------------------------------------------
i     64496:13:192.0.1.3/32                                    100         None
      192.0.2.3                                                None        262137
      No As-Path
u*>i  64496:16:192.0.1.6/32                                    100         None
      192.0.2.6                                                None        262137
      No As-Path
-------------------------------------------------------------------------------
Routes : 2
===============================================================================
*A:PE-3#
```

When the iLER is configured in pipe mode (VC-only) or if there is no TTL propagation to any MPLS label (none), the output of the traceroute detail command does not contain the MPLS label stack object information. In pipe mode, the IP TTL is propagated to the VC header, but not to the LDP header. When the TTL propagation is none, the IP TTL is not propagated to VC or LDP. The LSRs are invisible and there will be missing hops.

```
*A:PE-3# configure service vprn 1 ttl-propagate local vc-only
*A:PE-3# traceroute router 1 192.0.1.6 detail
traceroute to 192.0.1.6, 30 hops max, 40 byte packets
  1   1  192.0.1.6  (192.0.1.6)  1.80 ms
  1   2  192.0.1.6  (192.0.1.6)  1.85 ms
  1   3  192.0.1.6  (192.0.1.6)  2.16 ms

*A:PE-3#
```

The reason is that the TTL of the LDP header is 255 at the iLER and it is decremented by one in every LSR. The UDP traceroute message will not time out on the LSR PE-2. The TTL of the VC header is not decremented in the LSR. When the traceroute message does not time out in PE-2, the hop PE-2 is invisible. In a similar way, the traceroute messages will not time out in the LSR when no TTLs are propagated to any MPLS header. Figure 334 shows the TTLs in both cases:

Figure 334 **UDP Traceroute in VPRN without TTL Propagation to LDP**



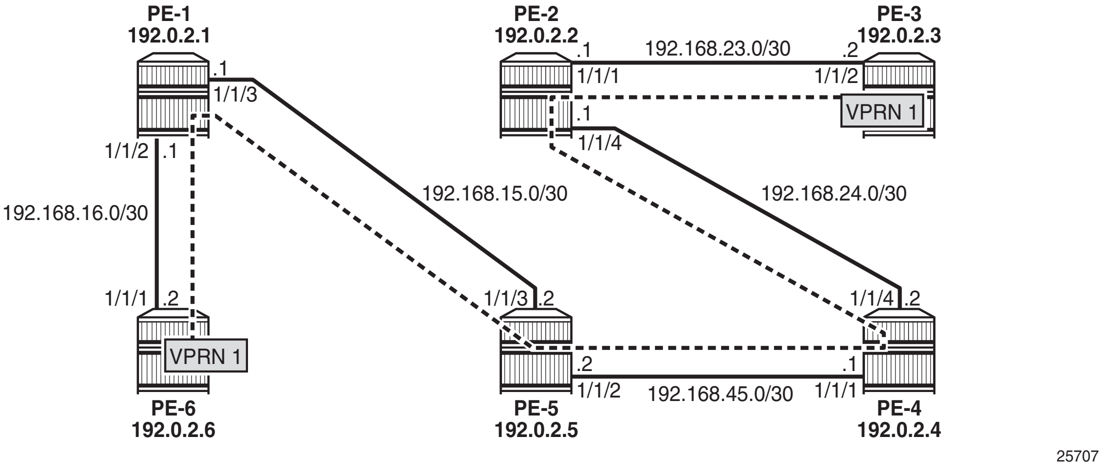
The TTL propagation is restored to uniform mode in the iLER as follows:

```
*A:PE-3# configure service vprn 1 ttl-propagate local all
```

Uniform Mode in iLER and ICMP Tunneling in Multiple LSRs

After shutting down some ports in the nodes, the tunnel from iLER PE-3 to PE-6 has four intermediate hops (LSRs) instead of one, as shown in Figure 335:

Figure 335 **Tunnel from iLER PE-3 to eLER PE-6 with Multiple LSRs**



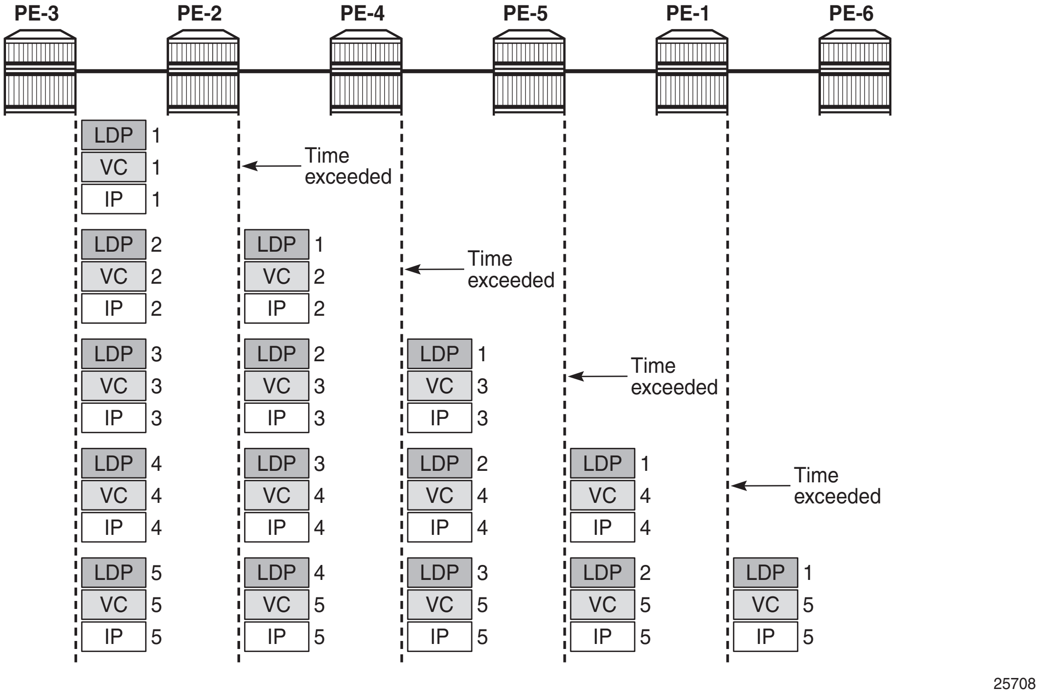
```
*A:PE-2# configure port 1/1/[2..3] shutdown
*A:PE-2# configure port 1/2/1 shutdown
*A:PE-3# configure port 1/1/1 shutdown
*A:PE-5# configure port 1/1/1 shutdown
```

TTL propagation on the iLER PE-3 is in uniform mode. ICMP tunneling needs to be enabled on all LSRs, as follows:

```
*A:PE-1# configure router icmp-tunneling
```

UDP traceroute messages are sent from VPRN 1 on iLER PE-3 to VPRN 1 on PE-6, as shown in Figure 336:

Figure 336 **UDP Traceroute with iLER in Uniform Mode**



The detailed output of the traceroute command shows the MPLS label stack object information as added by the LSR where the ICMP Time Exceeded message was tunneled. For brevity, only the first of the three messages from each intermediate node is shown.

```
*A:PE-3# traceroute router 1 192.0.1.6 detail
traceroute to 192.0.1.6, 30 hops max, 40 byte packets
  1   1  192.168.24.1  (192.168.24.1)  3.90 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262135, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   1, S = 1
  ---snip---
  2   1  192.168.45.1  (192.168.45.1)  2.81 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262137, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   2, S = 1
  ---snip---
  3   1  192.168.15.2  (192.168.15.2)  2.80 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262138, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   3, S = 1
  ---snip---
  4   1  192.168.16.1  (192.168.16.1)  2.97 ms
         returned MPLS Label Stack Object
             entry  1:  MPLS Label =  262138, Exp = 7, TTL =   1, S = 0
             entry  2:  MPLS Label =  262137, Exp = 7, TTL =   4, S = 1
  ---snip---
  5   1  192.0.1.6  (192.0.1.6)  2.90 ms
  ---snip---

*A:PE-3#
```

The TTL for the bottom MPLS header (BGP) is not decremented in each hop; only the TTL for the transport MPLS header (LDP) is decremented. The bottom label or BGP label of 262137 is not changed end-to-end. The LDP transport label for the different nodes is as follows.

For iLER PE-3, the LDP transport label for traffic toward PE-6 is 262135:

```
*A:PE-3# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
---snip---
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op            IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262135
192.168.23.1                                1/1/2

192.0.2.6/32                                Swap          262134    262135
192.168.23.1                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-3#
```

For LSR PE-2, the LDP transport label toward PE-6 is 262137:

```
*A:PE-2# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
---snip---
LDP IPv4 Prefix Bindings (Active)
```

```
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262137
192.168.24.2                                1/1/4

192.0.2.6/32                                Swap            262135    262137
192.168.24.2                                1/1/4

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-2#
```

For LSR PE-4, the LDP transport label toward PE-6 is 262138:

```
*A:PE-4# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
---snip---
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262138
192.168.45.2                                1/1/1

192.0.2.6/32                                Swap            262137    262138
192.168.45.2                                1/1/1

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-4#
```

For LSR PE-5, the LDP transport label toward PE-6 is 262138:

```
*A:PE-5# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
---snip---
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262138
192.168.15.1                                1/1/3

192.0.2.6/32                                Swap            262138    262138
192.168.15.1                                1/1/3

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-5#
```

For LSR PE-1, the LDP transport label toward PE-6 is 262143, but this label will not be present in the traceroute detailed output, because this message cannot time out on an LSR where ICMP tunneling takes place:

```
*A:PE-1# show router ldp bindings active prefixes prefix 192.0.2.6/32

===============================================================================
---snip---
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.6/32                                Push              --      262143
192.168.16.2                                1/1/2

192.0.2.6/32                                Swap            262138    262143
192.168.16.2                                1/1/2

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 2
===============================================================================
*A:PE-1#
```

Conclusion

Tunneling of ICMP reply messages over MPLS LSPs provides the ability to trace the hops in an MPLS tunnel. This mechanism applies to VPRN, 6PE/6VPE, and BGP labeled routes.

ICMP tunneling at an LSR applies to UDP traceroute packets that time out at the LSR. The ICMP Time Exceeded message is tunneled by the LSR toward the eLER and the eLER routes the packet to the sender of the traceroute message.

Unnumbered Interfaces in RSVP-TE and LDP

This chapter provides information about Unnumbered Interfaces in RSVP-TE and LDP.

Topics in this chapter include:

Applicability

This chapter is applicable to SR OS routers and was initially written for SR OS Release 13.0.R7. The CLI in this edition corresponds to SR OS Release 15.0.R1. SR OS supports unnumbered interfaces in RSVP-TE and LDP in SR OS Release 11.0.R1 or later.

Overview

Unnumbered interfaces enable IP processing on a point-to-point (P2P) interface without an explicit IP address. Unnumbered interfaces are supported in ReSource reserVation Protocol with Traffic Engineering (RSVP-TE) and Label Distribution Protocol (LDP).

An unnumbered interface is uniquely identified in the network by the tuple (router-id,ifIndex), where the interface index (ifIndex) is unique on the router. The two endpoints of an unnumbered link exchange the ifIndex that they assigned to the link.

The (router-id,ifIndex) tuple is used by the following:

- Intermediate System to Intermediate System (IS-IS) or Open Shortest Path First (OSPF) to advertise link information
- RSVP to signal Label Switched Paths (LSPs) over this unnumbered interface

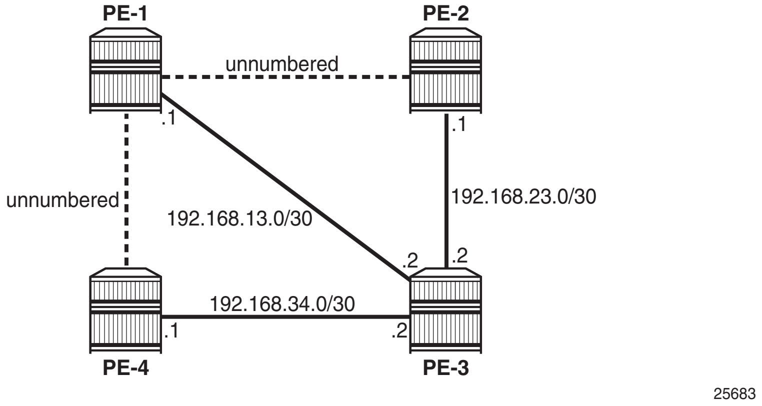
- LDP to establish hello adjacencies and resolve Forwarding Equivalence Classes (FECs)
- Operations, Administration, and Maintenance (OAM) to send or respond to an Multi-Protocol Label Switching (MPLS) echo request over an unnumbered interface

The unnumbered interface can "borrow" the IP address of another interface on the node.

The borrowed IP address is used exclusively as the source address for IP packets that are originated from the unnumbered interface. The borrowed IP address defaults to the system loopback interface address, but can be changed manually. The borrowed IP address corresponds to the router ID in the tuple representing the unnumbered interface.

The configuration used in this chapter is shown in Figure 337. There are two unnumbered links: one between PE-1 and PE-2, and another between PE-1 and PE-4. The remaining links are numbered.

Figure 337 **Example Topology for Unnumbered Interfaces in RSVP and LDP**



Configure unnumbered interfaces as follows:

```
configure router interface <itf-name> unnumbered [<ip-int-name|ip-address>]
```

To configure an unnumbered link with the system address as the borrowed IP address, no address needs to be configured:

```
*A:PE-4# configure router
        interface "int-PE-4-PE-1"
            port 1/1/1
            unnumbered
        exit
```

An unnumbered interface has to be a P2P link.

Unnumbered Interfaces in IS-IS

Unnumbered interfaces are identified in IS-IS by a combination of the system ID and an extended local circuit ID, as described in *RFC 5307 IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching*.

Enable debugging on the unnumbered interface on PE-4 (int-PE-4-PE-1) as follows:

```
*A:PE-4# debug router isis packet "int-PE-4-PE-1" detail
```

The following IS-IS hello Protocol Data Unit (PDU) is received from PE-1. The I/F Address is the borrowed IP address; in this case, the system address of PE-1, because the interface is unnumbered.

```
76 2017/03/29 12:50:42.96 UTC MINOR: DEBUG #2001 Base ISIS PKT
"ISIS PKT:
RX ISIS PDU ifId 2 len 52:
  DMAC          : 09:00:2b:00:00:05
  Proto Disc    : 131
  Header Len    : 20
  Version PID   : 1
  ID Length     : 0
  Version       : 1
  Reserved      : 0
  Max Area Addr : 3
  PDU Type      : (11) Point-2-Point IS-IS Hello Pdu
  Circuit Type  : L1
  Source Id     : 19 20 00 00 20 01
  Hold Time     : 27
  Packet length : 52
  Circuit Id    : 0
  Area Addresses:
    Area Address : (3) 49.0001
  Supp Protocols:
    Protocols     : IPv4
  I/F Addresses :
    I/F Address   : 192.0.2.1
  3Way Adjacency   :
    State          : UP
    Ext ckt ID     : 4
    NbrSysID       : 19 20 00 00 20 04
    Nbr ext ckt ID : 2
```

The three-way adjacency contains the neighbor extended local circuit ID (Nbr ext ckt ID: 2). This ID is the local interface index of the unnumbered interface (int-PE-4-PE-1), which can be verified as follows:

```
*A:PE-4# show router interface "int-PE-4-PE-1" detail | match "If Index"
If Index         : 2                    Virt. If Index    : 2
```

```
Last Oper Chg    : 03/29/2017 12:40:28  Global If Index   : 1
*A:PE-4#
```

On PE-1, the interface toward PE-4 has a different index, as follows:

```
*A:PE-1# show router interface "int-PE-1-PE-4" detail | match "If Index"
If Index         : 4                    Virt. If Index    : 4
Last Oper Chg    : 03/29/2017 12:40:14  Global If Index   : 3
*A:PE-1#
```

For numbered interfaces, such as int-PE-4-PE-3, the I/F Address is the interface address; in that case, 192.168.34.1, for messages received from PE-3 instead of the router ID.

When a Shared Risk Link Group (SRLG) is configured in combination with IS-IS and unnumbered interfaces, the least significant bit in the flags field of the SRLG Type-Length Value (TLV) indicates that the interface is unnumbered (0) or numbered (1).

Unnumbered Interfaces in OSPF

For unnumbered interfaces in OSPF, link local and remote identifiers are defined in *RFC 4203 OSPF Extensions in Support of Generalized Multi-Protocol Label Switching*. The OSPF link state advertisement (LSA) is defined in *RFC 2328, OSPF version 2*.

For numbered interfaces, the link data is the IP interface address; for unnumbered interfaces, the link data is the interface index value. The value starts from 1 in the format 0.0.0.1. SR OS recognizes an unnumbered interface when the first byte in the link data has a value of 0; SR OS then treats the link data as an interface index instead of an IP address.

Unnumbered Interfaces in RSVP-TE

Unnumbered IP interfaces can be used as Traffic Engineering (TE) links for the signaling of RSVP P2P LSPs and point-to-multipoint (P2MP) LSPs.

Fast Reroute (FRR) facility backup over unnumbered interfaces is supported, whereas FRR one-to-one will only use numbered interfaces in the detour path.

The unnumbered IP address is advertised by IS-IS or OSPF, and Constrained Shortest Path First (CSPF) can include them in the computation of a path.

Unnumbered interfaces of the remote router can be specified in the Explicit Route Object (ERO), and in the Record Route Object (RRO), by a combination of router ID (borrowed IP address) and interface ID, as defined in *RFC 3477 Signaling Unnumbered Links in RSVP-TE*.

The choice of the data interface is indicated in the path message by including the interface identifier of the data channel. In the path message, the IP address equals the local borrowed IP address; in the resv message, the IP address is the remote borrowed IP address. As well as the borrowed IP address, there is also a Logical Interface Handle (LIH). This interface identification is defined in *RFC 3473 Generalized Multi-Protocol Label Switching Signaling Resource Reservation Protocol Traffic Engineering Extensions*.

To see the path and resv messages on PE-4, enter the following debug commands:

```
*A:PE-4# debug router rsvp packet path detail
*A:PE-4# debug router rsvp packet resv detail
```

The path message contains the RSVPHop object with the local interface identifier of the data channel, as follows:

```
960 2017/03/28 11:50:21.83 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: PATH Msg
Send PATH From:192.0.2.4, To:192.0.2.2
          TTL:255, Checksum:0x3636, Flags:0x0
Session    - EndPt:192.0.2.2, TunnId:1, ExtTunnId:192.0.2.4
SessAttr   - Name:LSP-PE-4-PE-2::dyn
             SetupPri:7, HoldPri:0, Flags:0x17
RSVPHop    - Ctype:3, Addr:192.1.2.4, LIH:2
             RouterId :192.0.2.4, InterfaceId :2
TimeValue  - RefreshPeriod:30
SendTempl  - Sender:192.0.2.4, LspId:31266
SendTSpec  - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
             MPU:20, MTU:1564
LabelReq   - IfType:General, L3ProtID:2048
RRO        - Unnumbered: RouterId 192.0.2.4 InterfaceID 2, Flags:0x0
ERO        - Unnumbered RouterId 192.0.2.1, LinkId 4, Strict
             Unnumbered RouterId 192.0.2.2, LinkId 2, Strict
FRRObj     - SetupPri:7, HoldPri:0, HopLimit:16, BW:0.000 bps, Flags:0x2
             ExcAny:0x0, IncAny:0x0, IncAll:0x0
"
```

The ERO and RRO objects are also shown. The unnumbered interfaces are defined by the combination of the router ID and the interface ID.

The resv message also contains the RSVPHop object, but the address is now the remote address of PE-1 instead of the local address of PE-4, as follows:

```
962 2017/03/28 11:50:29.72 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESV Msg
Recv RESV From:192.0.2.1, To:192.0.2.4
          TTL:255, Checksum:0x4e21, Flags:0x0
```

```
Session     - EndPt:192.0.2.2, TunnId:1, ExtTunnId:192.0.2.4
RSVPHop     - Ctype:3, Addr:192.0.2.1, LIH:2
              RouterId :192.0.2.1, InterfaceId :4
TimeValue   - RefreshPeriod:30
Style       - SE
FlowSpec    - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
              MPU:20, MTU:1560, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.4, LspId:31266, Label:262142
RRO         - Unnumbered: RouterId 192.0.2.1 InterfaceID 4, Flags:0x1
              Label:262142, Flags:0x1
              Unnumbered: RouterId 192.0.2.2 InterfaceID 2, Flags:0x0
              Label:262142, Flags:0x1
"
```

To see the patherr and resverr messages on PE-4, enter the following debug command:

```
*A:PE-4# debug router rsvp packet patherr detail
*A:PE-4# debug router rsvp packet resverr detail
```

The patherr message contains the following ErrorSpec object, as defined by RFC 3473. In this case, the error is caused by disabling TE on ingress Label Egress Router (iLER) PE-4. No route can be found to the destination because there is no lookup in the TE database. The LSP will not come up, even if CSPF is disabled on the LSP.

```
1079 2017/03/28 12:05:55.73 UTC MINOR: DEBUG #2001 Base RSVP
"RSVP: RESVERR Msg
Send RESVERR From:192.1.2.4, To:192.0.2.1
              TTL:255, Checksum:0xf8ae, Flags:0x0
Session     - EndPt:192.0.2.2, TunnId:1, ExtTunnId:192.0.2.4
RSVPHop     - Ctype:3, Addr:192.1.2.4, LIH:2
              RouterId :192.0.2.4, InterfaceId :2
ErrorSpec   - Ctype:3, ErrNode:192.1.2.4, Flags:0x0, ErrCode:3, ErrValue:0
              RouterId :192.0.2.4, InterfaceId :2
Style       - SE
FlowSpec    - Ctype:QOS, CDR:0.000 bps, PBS:0.000 bps, PDR:infinity
              MPU:20, MTU:1560, RSpecRate:0, RSpecSlack:0
FilterSpec - Sender:192.0.2.4, LspId:31266
"
```

Considerations for Unnumbered Interfaces in RSVP-TE

Consider the following for unnumbered interfaces in RSVP-TE:

- With RSVP, TE must be enabled in IS-IS or OSPF. The router ID of the router that advertised an unnumbered interface index is obtained from the TE database. Therefore, if TE is disabled in IS-IS or OSPF, a non-CSPF LSP with the next hop for this path over an unnumbered interface will not come up. The router ID of the neighbor that has the next hop of the path message cannot be searched for.
  - The operational state of the LSP path will remain down with reason "noRouteToDestination".
  - If a path message is received at the LSR in which TE is disabled and the next hop for the LSP path is over an unnumbered interface, a patherr message is sent back to the iLER with error code 24: Routing problem; Error value 5: "No route available toward destination".
- "Only FRR facility protection is supported; FRR one-to-one protection only works for numbered interfaces.
- There is no FRR facility protection if the point of local repair (PLR) is the iLER and the bypass tunnel egress interface is unnumbered.
- Bi-directional Forwarding Detection (BFD) cannot be enabled on an unnumbered router interface. Therefore, RSVP FRR procedures will not be triggered via a BFD session timeout, but only by physical failures and local interface down events.
- Unnumbered interfaces cannot be configured as hops in a path. This is true for RSVP-TE LSPs, as well as for static LSPs.
- RSVP hello and hello related capabilities, such as graceful restart helper, are not supported.
- SRLG is supported, but the user SRLG DB (user-srlg-db) feature at the iLER is not supported. Unnumbered interfaces cannot be added to the SRLG DB. When the user SRLG DB feature is enabled on the iLER, all unnumbered interfaces are considered as having no SRLG membership.

Unnumbered Interfaces in LDP

LDP can establish hello adjacencies and can resolve unicast and multicast FECs over unnumbered interfaces.

For link LDP, hello adjacencies are brought up using hello packets with source IP address set to the borrowed IP address and a destination IP address set to 224.0.0.2. The borrowed IP address is the system address, by default. Hello packets with the same source IP address are accepted when received over parallel unnumbered interfaces from the same peer LSR ID. The corresponding hello adjacencies are associated with a single LDP session.

The transport address for the TCP connection, which is encoded in the hello packet, will always be set to the LSR ID of the node. The user can configure the **local-lsr-id** option on the interface and change the value of the LSR ID to either the local interface or some other interface name: loopback or not, numbered or not. The transport address for the LDP session will be updated with the new LSR ID.

For targeted LDP, the source and destination addresses of targeted hello packets are the LDP LSR IDs of the nodes. The user can configure the **local-lsr-id** option on the targeted session. The transport address for the LDP session and the source IP address of targeted hello messages will be updated to the new LSR ID value.

LDP will advertise/withdraw unnumbered interfaces using the address/address-withdraw messages. The borrowed IP address of the interface is used.

A FEC can be resolved to an unnumbered interface in the same way as it is resolved to a numbered interface. The outgoing interface and the next hop are searched for in the Routing Table Manager (RTM). The next hop consists of the router ID and link identifier of the interface to the peer LSR. All LDP FEC types are supported. LDP FEC Equal Cost Multi-Path (ECMP) over a mix of unnumbered and numbered interfaces is supported.

RFC 5036 LDP Specification describes the address list TLV that is used in the LDP address message, and the LDP address withdrawal message. For unnumbered interfaces, the borrowed IP address is used, which is typically the system address of the sender node.

On PE-1, enable debugging for LDP packets from peer 192.0.2.2 as follows:

```
*A:PE-1# debug router ldp peer 192.0.2.2 packet init detail
*A:PE-1# debug router ldp peer 192.0.2.2 packet label detail
```

The following LDP address packets are shown at PE-1:

```
278 2017/03/29 07:59:20.11 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address packet (msgId 6) to 192.0.2.2:0
Protocol version = 1
Address Family = 1  Number of addresses = 2
Address 1 = 192.0.2.1
Address 2 = 192.168.13.1
"


277 2017/03/29 07:59:20.09 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Address packet (msgId 6) from 192.0.2.2:0
Protocol version = 1
Address Family = 1  Number of addresses = 2
Address 1 = 192.0.2.2
Address 2 = 192.168.23.1
"
```

The received LDP address packet contains two addresses: the system IP address 192.0.2.2 for the unnumbered interfaces and the interface address 192.168.23.1 for a numbered interface on that node.

Considerations for Unnumbered Interfaces in LDP

All LDP features are supported on unnumbered interfaces, except for the following:

- BFD cannot be enabled on an unnumbered router interface. Therefore, LDP FRR procedures will not be triggered via a BFD session timeout, but only by physical failures and local interface down events.
- Unnumbered interfaces cannot be added into LDP global and peer prefix policies.

Unnumbered Interfaces in OAM

The following applies to unnumbered interfaces in RSVP-TE or LDP.

The downstream mapping object is a TLV that can be included in an echo request, as described in *RFC 4379 Detecting Multi-Protocol Label Switched Data Plane Failures*.

Only one downstream mapping object may appear in an echo request. The presence of a downstream mapping object is a request that a downstream mapping object be included in the echo reply.

For unnumbered interfaces, the address type is 2 (IPv4 unnumbered), the downstream IP address is the borrowed IP address of the downstream LSR, and the downstream interface address is the index assigned by the upstream LSR to the interface.

The downstream detailed mapping object is a TLV that can be included in an echo request, as described in *RFC 6424 Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels*.

The following output shows the detailed LSP trace for an RSVP LSP from PE-4 to PE-2. Two unnumbered interfaces are used: the first between PE-4 and PE-1 and the second between PE-1 and PE-2. The interface type (iftype) is ipv4Unnumbered.

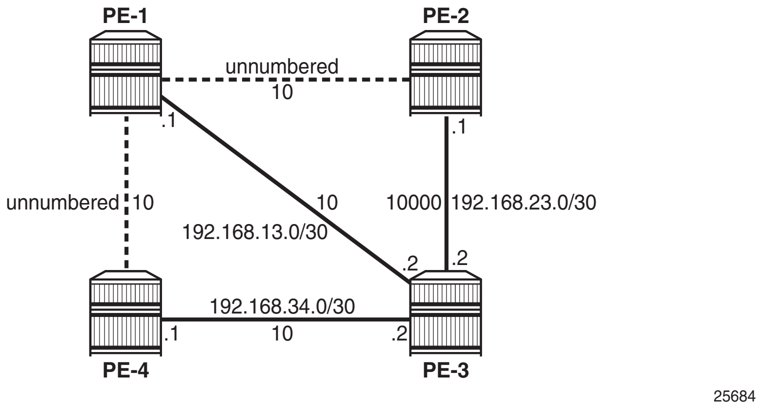
```
*A:PE-4# oam lsp-trace "LSP-PE-4-PE-2" detail
lsp-trace to LSP-PE-4-PE-2: 0 hops min, 0 hops max, 116 byte packets
1  192.0.2.1  rtt=0.602ms rc=8(DSRtrMatchLabel) rsc=1
     DS 1: ipaddr=0.0.0.0 ifaddr=0 iftype=ipv4Unnumbered MRU=1564
```

```
          label[1]=262143 protocol=4(RSVP-TE)
2  192.0.2.2  rtt=1.14ms rc=3(EgressRtr) rsc=1
*A:PE-4#
```

Configuration

The following configuration example is for unnumbered interfaces in RSVP and LDP; see Figure 338. The nodes are 7750 SRs.

Figure 338 **Configuration Example for Unnumbered Interfaces in RSVP and LDP**



All interfaces have a TE metric of 10, while the link between PE-2 and PE-3 has a TE metric of 10000. As such, the preferred path from PE-4 to PE-2 will be over the unnumbered interfaces between PE-4 and PE-1 and between PE-1 and PE-2.

Unnumbered Interfaces

Router interfaces are configured on all nodes, numbered and unnumbered. Initially, the unnumbered interfaces are configured with default settings. The following interfaces are configured on PE-1:

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            port 1/1/1
            unnumbered
        exit
        interface "int-PE-1-PE-3"
```

```
            address 192.168.13.1/30
            port 1/1/3
        exit
        interface "int-PE-1-PE-4"
            port 1/1/2
            unnumbered
        exit
        interface "system"
            address 192.0.2.1/32
        exit
```

There are two unnumbered interfaces: int-PE-1-PE-2 and int-PE-1-PE-4. There is no borrowed IP address configured for the unnumbered interfaces. This implies that the borrowed IP address will be the system address.

For the unnumbered interfaces, the borrowed IP address is indicated between square brackets in the following output:

```
*A:PE-1# show router interface

===============================================================================
Interface Table (Router: Base)
===============================================================================
Interface-Name                   Adm         Opr(v4/v6)  Mode    Port/SapId
   IP-Address                                                    PfxState
-------------------------------------------------------------------------------
int-PE-1-PE-2                    Up          Up/Down     Network 1/1/1
   Unnumbered If[system]                                         n/a
int-PE-1-PE-3                    Up          Up/Down     Network 1/1/3
   192.168.13.1/30                                               n/a
int-PE-1-PE-4                    Up          Up/Down     Network 1/1/2
   Unnumbered If[system]                                         n/a
system                           Up          Up/Down     Network system
   192.0.2.1/32                                                  n/a
-------------------------------------------------------------------------------
Interfaces : 4
===============================================================================
*A:PE-1#
```

Each interface, numbered or unnumbered, gets an interface index. This interface index can be retrieved as follows:

```
*A:PE-1# show router interface "int-PE-1-PE-2" detail | match "If Index"
If Index         : 2                   Virt. If Index    : 2
Last Oper Chg    : 03/29/2017 08:30:28  Global If Index   : 1
*A:PE-1# show router interface "int-PE-1-PE-3" detail | match "If Index"
If Index         : 3                   Virt. If Index    : 3
Last Oper Chg    : 03/29/2017 08:30:28  Global If Index   : 2
*A:PE-1# show router interface "int-PE-1-PE-4" detail | match "If Index"
If Index         : 4                   Virt. If Index    : 4
Last Oper Chg    : 03/29/2017 08:30:28  Global If Index   : 3
*A:PE-1# show router interface "system" detail | match "If Index"
If Index         : 1                   Virt. If Index    : 1
Last Oper Chg    : 03/29/2017 08:30:28  Global If Index   : 256
*A:PE-1#
```

The unnumbered interface toward PE-2 has ifIndex 2, and the unnumbered interface toward PE-4 has ifIndex 4.

BFD cannot be enabled on unnumbered interfaces. When BFD is configured on an unnumbered interface, the following error is raised:

```
*A:PE-4# configure router
*A:PE-4>config>router# interface "int-PE-4-PE-1"
*A:PE-4>config>router>if# bfd 100 receive 100 multiplier 3
INFO: BFD #1007 Bfd not allowed
```

An Interior Gateway Protocol (IGP) needs to be configured; in this case, IS-IS is chosen. OSPF could have been used equally well. TE must be enabled for unnumbered interfaces used in RSVP, even when CSPF is disabled. The IS-IS configuration on PE-1 is as follows:

```
*A:PE-1# configure
    router
        isis
            level-capability level-1
            area-id 49.0001
            traffic-engineering
            interface "system"
            exit
            interface "int-PE-1-PE-2"
                interface-type point-to-point
            exit
            interface "int-PE-1-PE-3"
                interface-type point-to-point
            exit
            interface "int-PE-1-PE-4"
                interface-type point-to-point
            exit
            no shutdown
```

An unnumbered interface has to be a P2P link.

The TE database contains the router ID and the ifIndex for unnumbered interfaces, as follows:

```
*A:PE-1# show router isis database PE-2.00-00 detail

===============================================================================
Rtr Base ISIS Instance 0 Database (detail)
===============================================================================
Displaying Level 1 database
-------------------------------------------------------------------------------
LSP ID    : PE-2.00-00                                 Level     : L1
Sequence  : 0x2                 Checksum  : 0x5ba3     Lifetime  : 1047
Version   : 1                   Pkt Type  : 18         Pkt Ver   : 1
Attributes: L1                  Max Area  : 3          Alloc Len : 180
SYS ID    : 1920.0000.2002      SysID Len : 6          Used Len  : 180

TLVs :
  Area Addresses:
    Area Address : (3) 49.0001
  Supp Protocols:
    Protocols     : IPv4
  IS-Hostname   : PE-2
  Router ID   :
    Router ID   : 192.0.2.2
  I/F Addresses :
    I/F Address   : 192.168.23.1
    I/F Address   : 192.0.2.2
  TE IS Nbrs   :
    Nbr   : PE-1.00
    Default Metric  : 10
    Sub TLV Len     : 10
    LclId    : 2
    RmtId    : 2
  TE IS Nbrs   :
    Nbr   : PE-3.00
    Default Metric  : 10
    Sub TLV Len     : 12
    IF Addr   : 192.168.23.1
    Nbr IP    : 192.168.23.2
---snip---
```

PE-2 has an unnumbered interface toward PE-1 (Nbr: PE-1.00), with local interface index 2 (LclId: 2) and remote interface index 2 (RmtId: 2). For the numbered interface toward PE-3, the local and remote interface IP addresses are shown (IF Addr and Nbr IP), not the interface index.

Unnumbered Interfaces in RSVP

MPLS and RSVP need to be enabled on the interfaces on the nodes. TE metrics are configured on the MPLS interfaces. For node PE-4, the configuration is as follows:

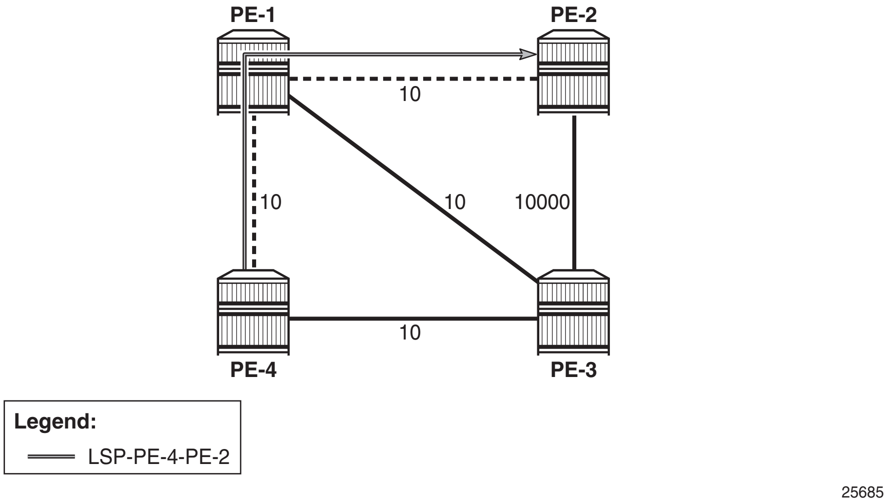
```
*A:PE-4# configure
    router
        mpls
            interface "system"
            exit
            interface "int-PE-4-PE-1"
                te-metric 10
            exit
            interface "int-PE-4-PE-3"
                te-metric 10
            exit
        exit
        rsvp
            no shutdown
        exit
```

An LSP is configured from PE-4 to PE-2 with CSPF enabled and using the TE metrics, not the IGP metrics. Unnumbered interfaces cannot be configured as hops in a path. A dynamic path "dyn", without any hops, is configured to be used in an LSP from PE-4 to PE-2, as follows:

```
*A:PE-4# configure
    router
        mpls
            path "dyn"
                no shutdown
            exit
            lsp "LSP-PE-4-PE-2"
                to 192.0.2.2
                cspf use-te-metric
                primary "dyn"
                exit
                no shutdown
            exit
            no shutdown
        exit
```

The LSP from PE-4 to PE-2 will have TE metric 20 when the next hop is PE-1, and TE metric 30 or 10010 when the next hop is PE-3. Figure 339 shows LSP-PE-4-PE-2, which uses only unnumbered interfaces.

Figure 339 **LSP-PE-4-PE-2 on Unnumbered Interfaces**



The following tunnel table shows a next hop 0.0.0.1, which implies that it is an unnumbered interface. The only unnumbered interface at PE-4 is int-PE-4-PE-1. The metric in this tunnel table is 16777215 because the IGP metric is not used.

```
*A:PE-4# show router tunnel-table
```

```
===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      rsvp      MPLS  1         7        0.0.0.1        16777215
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-4#
```

The actual and computed hops can be verified, as well as the CSPF metric (TE metric), as follows:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail
```

```
===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55808
From             : 192.0.2.4            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262143
---snip---
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.0.2.4, If Index : 2                      Record Label        : N/A
 -> 192.0.2.1, If Index : 4                      Record Label        : 262143
 -> 192.0.2.2, If Index : 2                      Record Label        : 262143
Computed Hops    :
    192.0.2.4, If Index : 2(S)
 -> 192.0.2.1, If Index : 4(S)
 -> 192.0.2.2, If Index : 2(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 20
===============================================================================
*A:PE-4#
```

The computed hops are strict hops, as indicated by "(S)". Because the interfaces are unnumbered, the system address and the ifIndex are displayed. The CSPF metric is 20.

Configuring the Borrowed IP Address

The borrowed IP address does not need to be the system address, but the address must exist on the node. When the unnumbered interface is configured with a borrowed IP address that does not exist on the node, the interface goes down. This can be verified by assigning a non-existent address to the unnumbered interface int-PE-1-PE-2, as follows:

```
*A:PE-1# configure
    router
        interface "int-PE-1-PE-2"
            port 1/1/1
            unnumbered 192.1.2.1
        exit
```

The operational state of this interface goes down, which can be verified as follows:

```
*A:PE-1# show router interface

===============================================================================
Interface Table (Router: Base)
===============================================================================
Interface-Name                   Adm         Opr(v4/v6)  Mode    Port/SapId
   IP-Address                                                    PfxState
-------------------------------------------------------------------------------
int-PE-1-PE-2                    Up          Down/Down   Network 1/1/1
   Unnumbered If[192.1.2.1]                                      n/a
int-PE-1-PE-3                    Up          Up/Down     Network 1/1/3
   192.168.13.1/30                                               n/a
int-PE-1-PE-4                    Up          Up/Down     Network 1/1/2
   Unnumbered If[system]                                         n/a
system                           Up          Up/Down     Network system
   192.0.2.1/32                                                  n/a
-------------------------------------------------------------------------------
Interfaces : 4
===============================================================================
*A:PE-1#
```

The borrowed IP address is indicated between square brackets. The interface is down because the IP address is not known on PE-1. The down reason code (noIfAddress) can be retrieved as follows:

```
*A:PE-1# show router interface "int-PE-1-PE-2" detail | match "Down Reason Code"
Down Reason Code : noIfAddress
```

The IP address can be configured as a loopback address on PE-1 and assigned to all unnumbered interfaces, as follows:

```
*A:PE-1# configure
    router
        interface "loopback1"
            address 192.1.2.1/32
            loopback
        exit
        interface "int-PE-1-PE-4"
            port 1/1/2
            unnumbered 192.1.2.1
        exit
```

When the borrowed IP address is known on node PE-1, the unnumbered interface is operationally up, which can be verified as follows:

```
*A:PE-1# show router interface

===============================================================================
Interface Table (Router: Base)
===============================================================================
Interface-Name                   Adm         Opr(v4/v6)  Mode     Port/SapId
   IP-Address                                                     PfxState
-------------------------------------------------------------------------------
int-PE-1-PE-2                    Up          Up/Down     Network 1/1/1
   Unnumbered If[192.1.2.1]                                       n/a
int-PE-1-PE-3                    Up          Up/Down     Network 1/1/3
   192.168.13.1/30                                                n/a
int-PE-1-PE-4                    Up          Up/Down     Network 1/1/2
   Unnumbered If[192.1.2.1]                                       n/a
loopback1                        Up          Up/Down     Network loopback
   192.1.2.1/32                                                   n/a
system                           Up          Up/Down     Network system
   192.0.2.1/32                                                   n/a
-------------------------------------------------------------------------------
Interfaces : 5
===============================================================================
*A:PE-1#
```

In a similar way, the borrowed IP address on PE-2 is configured as 192.1.2.2 and on PE-4 as 192.1.2.4.

TE Required for Unnumbered Interfaces in RSVP

For unnumbered interfaces, the IGP needs to look up the router ID in the TE database. Therefore, TE must be enabled even if CSPF is disabled. TE is disabled in IS-IS and CSPF is disabled in the LSP on PE-4, as follows:

```
*A:PE-4# configure router isis no traffic-engineering
*A:PE-4# configure router mpls lsp "LSP-PE-4-PE-2" no cspf
```

LSP-PE-4-PE-2 is operationally down with failure code "noRouteToDestination", which can be verified as follows:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-4- PE-2 Path  (Detail)
```

```
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55818
From             : 192.0.2.4           To                   : 192.0.2.2
Admin State      : Up                  Oper State           : Down
Path Name        : dyn                 Path Type            : Primary
Path Admin       : Up                  Path Oper            : Down
---snip---
CSPF             : Disabled            Oper CSPF            : N/A
---snip---
Failure Code     : noRouteToDestination
Failure Node     : 192.0.2.4
---snip---
```

The configuration is restored by enabling TE in IS-IS and CSPF in the LSP context, as follows:

```
*A:PE-4# configure router isis traffic-engineering
*A:PE-4# configure router mpls lsp "LSP-PE-4-PE-2" cspf use-te-metric
```

FRR Facility

FRR facility is enabled on the LSP as follows:

```
*A:PE-4# configure
    router
        mpls
            lsp "LSP-PE-4-PE-2"
                fast-reroute facility
                exit
            exit
```

The following LSP path detail output shows where an FRR detour is available (@) and in which node a bypass tunnel originates. The letter "n" indicates that a node is protected, as in hop 192.0.2.4. When there is a detour available, but there is no "n", link protection is available, as in hop 192.0.2.1:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail
```

```
===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available                  # - Detour In Use
    b - Bandwidth Protected               n - Node Protected
    s - Soft Preemption
    S - Strict                            L - Loose
    A - ABR                               + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55822
From             : 192.0.2.4            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262141
---snip---
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.0.2.4, If Index : 2 @ n                  Record Label        : N/A
 -> 192.0.2.1, If Index : 4 @                    Record Label        : 262141
 -> 192.0.2.2, If Index : 2                      Record Label        : 262141
Computed Hops    :
    192.0.2.4, If Index : 2(S)
 -> 192.0.2.1, If Index : 4(S)
 -> 192.0.2.2, If Index : 2(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 20
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-4#
```

Information about the bypass tunnel originating in PE-4 can be retrieved as follows:

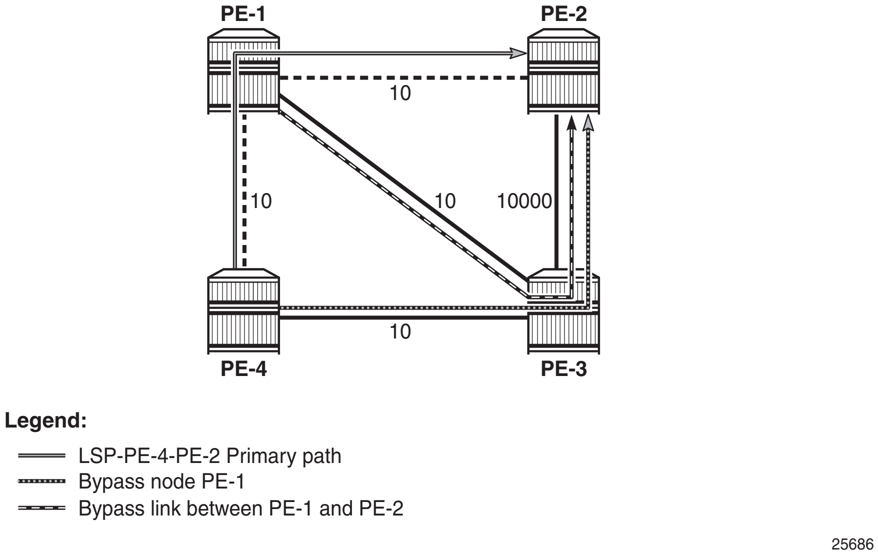
```
*A:PE-4# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-node192.0.2.1-61441
-------------------------------------------------------------------------------
To             : 192.168.23.1        State               : Up
Out I/F        : 1/1/2               Out Label           : 262142
Up Time        : 0d 00:02:49         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : Dynamic             Bypass Path Cost    : 10010
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.34.2(S)                  Egress Admin Groups : None
 -> 192.168.34.1(S)                  Egress Admin Groups : None
 -> 192.168.23.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.34.2 (192.0.2.4)         Record Label        : N/A
 -> 192.168.34.1 (192.0.2.3)         Record Label        : 262142
 -> 192.168.23.1 (192.0.2.2)         Record Label        : 262139
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-4-PE-2::dyn
From           : 192.0.2.4           To                  : 192.0.2.2
Avoid Node/Hop : 192.0.2.1           Downstream Label    : 262141
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-4#
```

This bypass tunnel, via PE-3 to PE-2, offers node protection for node PE-1. There are no unnumbered interfaces in this path. In a similar way, information about the bypass tunnel to protect the link between PE-1 and PE-2 can be retrieved in PE-1, as follows:

```
*A:PE-1# show router mpls bypass-tunnel protected-lsp detail

===============================================================================
MPLS Bypass Tunnels (Detail)
===============================================================================
-------------------------------------------------------------------------------
bypass-link192.0.2.2-61441
-------------------------------------------------------------------------------
To             : 192.168.23.1        State               : Up
Out I/F        : 1/1/3               Out Label           : 262143
Up Time        : 0d 00:02:51         Active Time         : n/a
Reserved BW    : 0 Kbps              Protected LSP Count : 1
Type           : Dynamic             Bypass Path Cost    : 10010
Setup Priority : 7                   Hold Priority       : 0
Class Type     : 0
Exclude Node   : None                Inter-Area          : False
Computed Hops  :
    192.168.13.1(S)                  Egress Admin Groups : None
 -> 192.168.13.2(S)                  Egress Admin Groups : None
 -> 192.168.23.1(S)                  Egress Admin Groups : None
Actual Hops    :
    192.168.13.1 (192.0.2.1)         Record Label        : N/A
 -> 192.168.13.2 (192.0.2.3)         Record Label        : 262143
 -> 192.168.23.1 (192.0.2.2)         Record Label        : 262140
Last Resignal  :
Attempted At   : n/a                 Resignal Reason     : n/a
Resignal Status: n/a                 Reason              : n/a

Protected LSPs -
LSP Name       : LSP-PE-4-PE-2::dyn
From           : 192.0.2.4           To                  : 192.0.2.2
Avoid Node/Hop : 192.0.2.2           Downstream Label    : 262141
Bandwidth      : 0 Kbps

===============================================================================
*A:PE-1#
```

Figure 340 shows the LSP and the two bypass tunnels: one in PE-4, offering node protection for node PE-1, and another in PE-1, bypassing the link between PE-1 and PE-2.

Figure 340 **LSP and FRR Facility Bypass Tunnels**



For each bypass tunnel, an additional RSVP session is set up. The following output shows that, in PE-4, two LSPs are signaled: the primary LSP and the bypass tunnel for node PE-1.

```
*A:PE-4# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55822 LSP-PE-4-PE-2::dyn           Up
192.0.2.4       192.168.23.1    61441  2     bypass-node192.0.2.1-61441   Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-4#
```

Similarly, PE-1 has an RSVP session for the primary LSP, but also for the bypass tunnel for the link toward PE-2, as follows:

```
*A:PE-1# show router rsvp session

===============================================================================
RSVP Sessions
```

```
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55822 LSP-PE-4-PE-2::dyn           Up
192.0.2.1       192.168.23.1    61441  2     bypass-link192.0.2.2-61441   Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-1#
```

PE-3 is only used by the bypass tunnels, as follows:

```
*A:PE-3# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.1       192.168.23.1    61441  2     bypass-link192.0.2.2-61441   Up
192.0.2.4       192.168.23.1    61441  2     bypass-node192.0.2.1-61441   Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-3#
```

PE-2 terminates the LSP and the bypass tunnels, as follows:

```
*A:PE-2# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55822 LSP-PE-4-PE-2::dyn           Up
192.0.2.1       192.168.23.1    61441  2     bypass-link192.0.2.2-61441   Up
192.0.2.4       192.168.23.1    61441  2     bypass-node192.0.2.1-61441   Up
-------------------------------------------------------------------------------
Sessions : 3
===============================================================================
*A:PE-2#
```

FRR One-to-One Only Supported on Numbered Interfaces

When FRR one-to-one is enabled on the LSP, the LSP will not use unnumbered interfaces.

FRR is reconfigured on the LSP as follows:

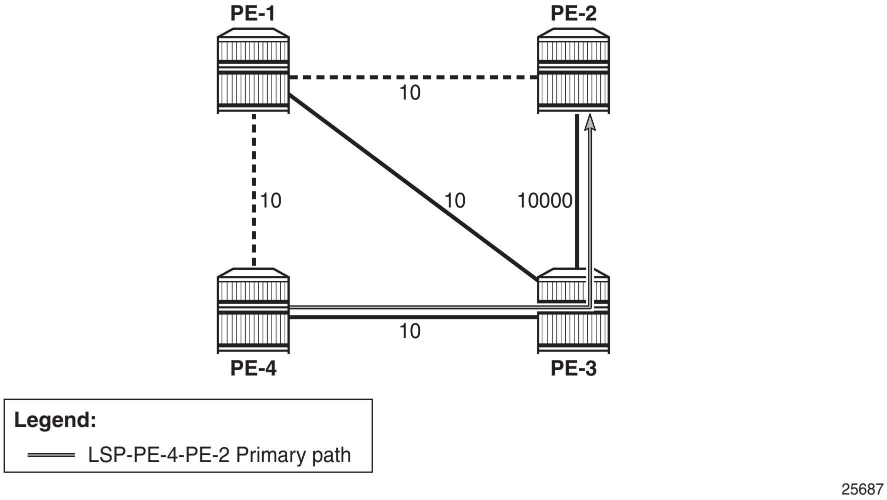
```
*A:PE-4# configure
    router
        mpls
            lsp "LSP-PE-4-PE-2"
                no fast-reroute
                fast-reroute one-to-one
                exit
            exit
```

The LSP will only come up if it can use numbered interfaces end-to-end. In this case, the LSP will take the path via PE-3 with CSPF metric 10010, as follows:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55828
From             : 192.0.2.4            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/2                Out Label            : 262143
---snip---
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.168.34.2 (192.0.2.4)                     Record Label        : N/A
 -> 192.168.34.1 (192.0.2.3)                     Record Label        : 262143
 -> 192.168.23.1 (192.0.2.2)                     Record Label        : 262142
Computed Hops    :
    192.168.34.2(S)
 -> 192.168.34.1(S)
 -> 192.168.23.1(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 10010
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-4#
```

Figure 341 shows the LSP in case of FRR one-to-one. Only numbered interfaces are used. Unfortunately, there is no bypass tunnel possible with only numbered interfaces; therefore, there is no protection.

Figure 341 **FRR One-to-One Only Supported on Numbered Interfaces**



If there is no path available with only numbered interfaces, the LSP will remain operationally down with failure code "noCspfRouteToDestination". This can be verified by shutting down port 1/1/2 toward PE-3, as follows:

```
*A:PE-4# configure port 1/1/2 shutdown
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail
===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55830
From             : 192.0.2.4            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Down
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Down
Out Interface    : n/a                  Out Label            : n/a
---snip---
Failure Code     : noCspfRouteToDestination
Failure Node     : 192.0.2.4
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    No Hops Specified
Computed Hops    :
    No Hops Specified
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : N/A
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-4#
```

The port is enabled again and the LSP configuration is restored to FRR facility, as follows:

```
*A:PE-4# configure port 1/1/2 no shutdown
*A:PE-4# configure
    router
        mpls
            lsp "LSP-PE-4-PE-2"
                no fast-reroute
                fast-reroute facility
                exit
            exit
```

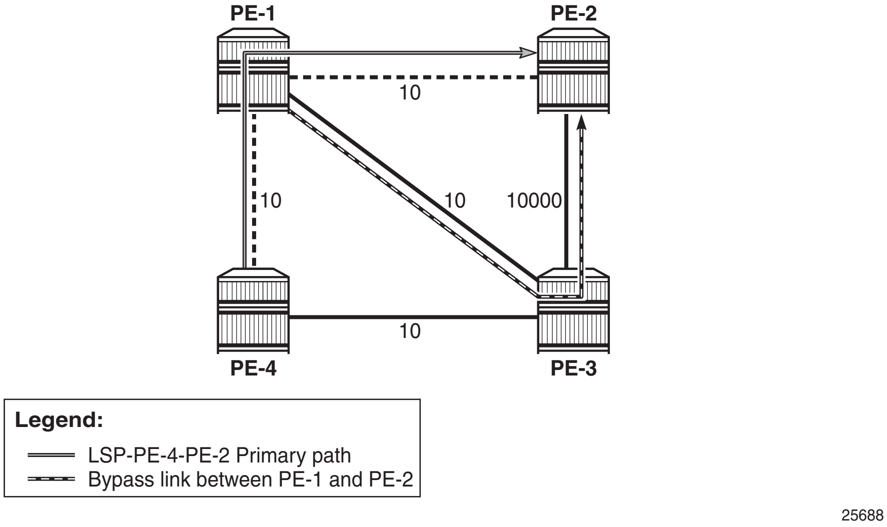
FRR Bypass Not Possible on iLER on Unnumbered Interfaces

FRR facility is not supported on the iLER PE-4 if the bypass is over an unnumbered interface. This restriction only applies to the iLER, not to the LSRs. The interface toward PE-3 is reconfigured as unnumbered, as follows:

```
*A:PE-3# configure router interface "int-PE-3-PE-4" no address
*A:PE-3# configure router interface "int-PE-3-PE-4" unnumbered
*A:PE-3# configure router mpls interface "int-PE-3-PE-4" te-metric 10
*A:PE-4# configure router interface "int-PE-4-PE-3" no address
*A:PE-4# configure router interface "int-PE-4-PE-3" unnumbered
*A:PE-4# configure router mpls interface "int-PE-4-PE-3" te-metric 10
```

When an interface changes from numbered to unnumbered or vice versa, it is no longer known in the MPLS context. Therefore, the interface needs to be added in the MPLS context again. When the interface toward PE-3 is numbered, there is a bypass tunnel in PE-4 to protect node PE-1, but this bypass tunnel cannot be established on an unnumbered interface. The only remaining protection for the LSP is the bypass tunnel originating in PE-1 to protect the link between PE-1 and PE-2, as shown in Figure 342.

Figure 342 **FRR on iLER: No Bypass on Unnumbered Interfaces**



The following output shows that there is only a detour available in PE-1:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55836
From             : 192.0.2.4           To                   : 192.0.2.2
Admin State      : Up                  Oper State           : Up
Path Name        : dyn                 Path Type            : Primary
Path Admin       : Up                  Path Oper            : Up
Out Interface    : 1/1/1               Out Label            : 262143
---snip---
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.0.2.4, If Index : 2                    Record Label        : N/A
 -> 192.0.2.1, If Index : 4 @                  Record Label        : 262143
 -> 192.0.2.2, If Index : 2                    Record Label        : 262143
Computed Hops    :
    192.0.2.4, If Index : 2(S)
 -> 192.0.2.1, If Index : 4(S)
 -> 192.0.2.2, If Index : 2(S)
Resignal Eligible: False
Last Resignal    : n/a                  CSPF Metric          : 20
===============================================================================
* indicates that the corresponding row element may have been truncated.
*A:PE-4#
```

In iLER PE-4, there is only the LSP tunnel, no bypass tunnel, as follows:

```
*A:PE-4# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                         State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55836 LSP-PE-4-PE-2::dyn           Up
-------------------------------------------------------------------------------
Sessions : 1
===============================================================================
*A:PE-4#
```

The original configuration is restored with numbered interfaces between PE-3 and PE-4, as follows:

```
*A:PE-3# configure router interface "int-PE-3-PE-4" no unnumbered
*A:PE-3# configure router interface "int-PE-3-PE-4" address 192.168.34.1/30
*A:PE-3# configure router mpls interface "int-PE-3-PE-4" te-metric 10
*A:PE-4# configure router interface "int-PE-4-PE-3" no unnumbered
*A:PE-4# configure router interface "int-PE-4-PE-3" address 192.168.34.2/30
*A:PE-4# configure router mpls interface "int-PE-4-PE-3" te-metric 10
```

Admin Groups for Unnumbered Interfaces in RSVP

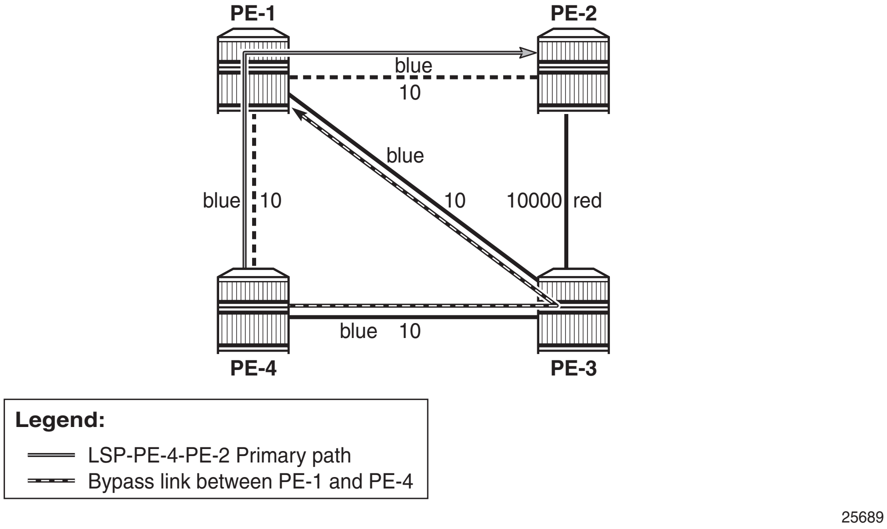
Administrative groups (link-coloring) can be used to calculate a path with the restriction to only include, or exclude, links of a particular admin group (color). Paths can be disjointed from each other, without the need for an explicit hops list. For unnumbered interfaces, an explicit hop list is not an option, but admin groups are.

Two admin groups are configured on all nodes, as follows:

```
*A:PE-4# configure router if-attribute admin-group "red" value 0
*A:PE-4# configure router if-attribute admin-group "blue" value 1
```

Admin group "blue" is assigned to all MPLS interfaces, except for the link between PE-2 and PE-3; see Figure 343.

Figure 343 **FRR Facility and Admin Groups**



The admin groups are assigned to the interfaces in the MPLS context, as follows:

```
*A:PE-2# configure
    router
        mpls
            interface "int-PE-2-PE-1"
                admin-group "blue"
            exit
            interface "int-PE-2-PE-3"
                admin-group "red"
            exit
```

To ensure that FRR bypass tunnels will only use links belonging to the same admin group, the following is configured on all nodes. It is required on all PLRs.

```
*A:PE-4# configure router mpls admin-group-frr
```

In the LSP context, the admin group "blue" is included. The option **propagate-admin-group** implies the tunnels must use links belonging to the admin group "blue". This is configured for the LSP tunnel, and for the FRR bypass tunnels, as follows:

```
*A:PE-4# configure
    router
        mpls
            lsp "LSP-PE-4-PE-2"
                include "blue"
                propagate-admin-group
                fast-reroute facility
                    propagate-admin-group
                exit
            exit
        exit
```

This configuration implies that the link that does not belong to admin group "blue" is excluded, and cannot be used by the LSP nor by a bypass tunnel. Therefore, there will be no bypass tunnel to protect node PE-1 and no bypass tunnel originating in PE-1 protecting the link to PE-2. There will be a bypass tunnel originating in PE-4 to protect the link between PE-4 and PE-1, as shown in Figure 343. The following output shows that a detour is available for link protection in PE-4:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name         : LSP-PE-4-PE-2
Path LSP ID      : 55842
From             : 192.0.2.4            To                   : 192.0.2.2
Admin State      : Up                   Oper State           : Up
Path Name        : dyn                  Path Type            : Primary
Path Admin       : Up                   Path Oper            : Up
Out Interface    : 1/1/1                Out Label            : 262140
---snip---
Include Groups   :                      Oper Include Groups  :
blue                                           blue
Exclude Groups   :                      Oper Exclude Groups  :
None                                           None
---snip---
Explicit Hops    :
    No Hops Specified
Actual Hops      :
    192.0.2.4, If Index : 2 @                    Record Label        : N/A
 -> 192.0.2.1, If Index : 4                      Record Label        : 262140
 -> 192.0.2.2, If Index : 2                      Record Label        : 262138
Computed Hops    :
    192.0.2.4, If Index : 2(S)
 -> 192.0.2.1, If Index : 4(S)
 -> 192.0.2.2, If Index : 2(S)
---snip---
```

The following output shows two RSVP sessions in PE-4: one for the LSP and one for the bypass tunnel to protect the link between PE-4 and PE-1.

```
*A:PE-4# show router rsvp session

===============================================================================
RSVP Sessions
```

```
===============================================================================
From            To              Tunnel LSP   Name                          State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55842 LSP-PE-4-PE-2::dyn            Up
192.0.2.4       192.168.13.1    61597  8     bypass-link192.0.2.1-61597    Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-4#
```

The configuration is restored as follows:

```
*A:PE-4# configure router mpls no admin-group-frr
*A:PE-4# configure
    router
        mpls
            lsp "LSP-PE-4-PE-2"
                no include "blue"
                no propagate-admin-group
                fast-reroute no propagate-admin-group
            exit
```

SRLGs for Unnumbered Interfaces in RSVP

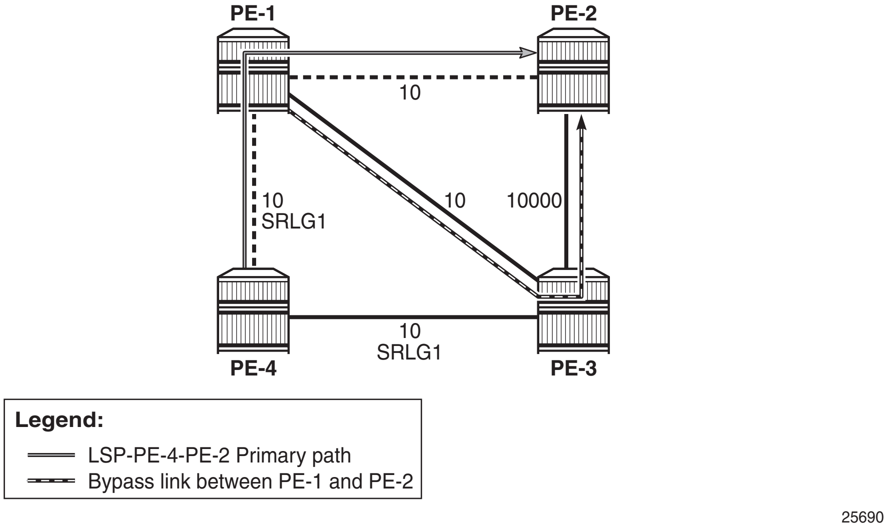
SRLGs allow operators to create automatic secondary LSPs or FRR tunnels that are disjointed from the protected primary tunnel. See chapter Shared Risk Link Groups for RSVP-Based LSP for more information.

One SRLG group is configured on all nodes, as follows:

```
*A:PE-4# configure router if-attribute srlg-group "SRLG1" value 1
```

SRLG "SRLG1" is assigned to the interface between PE-4 and PE-1, and to the interface between PE-4 and PE-3, as shown in Figure 344.

Figure 344 **SRLG-FRR Strict: No Bypass on PE-4**



The SRLG is assigned to the interfaces in the MPLS context, as follows:

```
*A:PE-4# configure
    router
        mpls
            interface "int-PE-4-PE-1"
                srlg-group "SRLG1"
            exit
            interface "int-PE-4-PE-3"
                srlg-group "SRLG1"
            exit
        exit
```

The configuration on PE-1 and PE-3 is similar.

When SRLG for FRR is enabled in strict mode, CSPF will not establish any detour LSP if there is no path that meets the SRLG constraint. This configuration implies that there is no bypass tunnel in PE-4. The following enables SRLG for FRR in strict mode on all nodes:

```
*A:PE-4# configure router mpls srlg-frr strict
```

Note: Enabling or disabling SRLG for FRR is a system-wide configuration that requires the MPLS routing instance to be manually set to shutdown, then to no shutdown, to activate the change. This can be service affecting. Nokia recommends that the operator include the SRLG in the initial network design and implementation to minimize the traffic loss.

The following output shows that there is only a detour available in PE-1:

```
*A:PE-4# show router mpls lsp "LSP-PE-4-PE-2" path detail

===============================================================================
MPLS LSP LSP-PE-4-PE-2 Path  (Detail)
===============================================================================
Legend :
    @ - Detour Available              # - Detour In Use
    b - Bandwidth Protected           n - Node Protected
    s - Soft Preemption
    S - Strict                        L - Loose
    A - ABR                           + - Inherited
===============================================================================
-------------------------------------------------------------------------------
LSP LSP-PE-4-PE-2 Path dyn
-------------------------------------------------------------------------------
LSP Name          : LSP-PE-4-PE-2
Path LSP ID       : 55850
From              : 192.0.2.4           To                    : 192.0.2.2
Admin State       : Up                  Oper State            : Up
Path Name         : dyn                 Path Type             : Primary
Path Admin        : Up                  Path Oper             : Up
Out Interface     : 1/1/1               Out Label             : 262143
---snip---
Explicit Hops     :
    No Hops Specified
Actual Hops       :
    192.0.2.4, If Index : 2                      Record Label        : N/A
 -> 192.0.2.1, If Index : 4 @                    Record Label        : 262143
 -> 192.0.2.2, If Index : 2                      Record Label        : 262143
Computed Hops     :
    192.0.2.4, If Index : 2(S)
 -> 192.0.2.1, If Index : 4(S)
 -> 192.0.2.2, If Index : 2(S)
---snip---
```

The following output shows that PE-1 has two RSVP sessions: one for the LSP and one for the bypass tunnel to protect the link between PE-1 and PE-2.

```
*A:PE-1# show router rsvp session

===============================================================================
RSVP Sessions
===============================================================================
From            To              Tunnel LSP   Name                          State
                                ID     ID
-------------------------------------------------------------------------------
192.0.2.4       192.0.2.2       1      55850 LSP-PE-4-PE-2::dyn            Up
192.0.2.1       192.168.23.1    61492  16    bypass-link192.0.2.2-61492    Up
-------------------------------------------------------------------------------
Sessions : 2
===============================================================================
*A:PE-1#
```

This was the last example for unnumbered interfaces in RSVP. MPLS and RSVP are disabled in all nodes as follows:

```
*A:PE-1# configure router mpls shutdown
*A:PE-1# configure router rsvp shutdown
```

Unnumbered Interfaces in LDP

Link LDP is configured on PE-4, as follows:

```
*A:PE-4# configure
    router
        ldp
            interface-parameters
                interface "int-PE-4-PE-1"
                exit
                interface "int-PE-4-PE-3"
                exit
            exit
```

The configuration of link LDP on the other nodes is similar. Link LDP sessions are established, which can be verified as follows:

```
*A:PE-4# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.1:0         Link      Established   11        12        0d 00:00:13
192.0.2.3:0         Link      Established   11        12        0d 00:00:13
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
```

The peer LDP ID is the LSR ID, which is the system address, by default. The IP address configured on the unnumbered interface (such as 192.1.2.1) is not used. The following tunnel table shows a distinction between numbered and unnumbered interfaces in the next hop:

```
*A:PE-4# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.1/32      ldp       MPLS  65537     9        0.0.0.1        10
192.0.2.2/32      ldp       MPLS  65538     9        0.0.0.1        20
192.0.2.3/32      ldp       MPLS  65539     9        192.168.34.1   10
```

```
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-4#
```

For destination 192.0.2.1 or 192.0.2.2, the unnumbered interface toward PE-1 is taken. The next hop is represented by 0.0.0.1. When a node has several unnumbered interfaces, the corresponding next hop values are different, as follows, for PE-1:

```
*A:PE-1# show router tunnel-table

===============================================================================
IPv4 Tunnel Table (Router: Base)
===============================================================================
Destination       Owner     Encap TunnelId  Pref     Nexthop        Metric
-------------------------------------------------------------------------------
192.0.2.2/32      ldp       MPLS  65537     9        0.0.0.1        10
192.0.2.3/32      ldp       MPLS  65538     9        192.168.13.2   10
192.0.2.4/32      ldp       MPLS  65539     9        0.0.0.3        10
-------------------------------------------------------------------------------
Flags: B = BGP backup route available
       E = inactive best-external BGP route
===============================================================================
*A:PE-1#
```

The LDP active prefix bindings only contain system addresses, no other loopback prefixes, as follows:

```
*A:PE-4# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262142
Unnumbered                                  1/1/1

192.0.2.1/32                                Swap            262142    262142
```

```
Unnumbered                                  1/1/1

192.0.2.2/32                                Push              --      262143
Unnumbered                                  1/1/1

192.0.2.2/32                                Swap            262141    262143
Unnumbered                                  1/1/1

192.0.2.3/32                                Push              --      262143
192.168.34.1                                1/1/2

192.0.2.3/32                                Swap            262140    262143
192.168.34.1                                1/1/2

192.0.2.4/32                                Pop             262143      --
  --                                          --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 7
===============================================================================
```

For prefixes 192.0.2.1 and 192.0.2.2, the egress next hop is unnumbered. The egress interface for both is 1/1/1. There is no ifIndex. Local addresses are advertised in LDP address messages, such as:

```
6 2017/03/31 09:39:16.85 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Send Address packet (msgId 5) to 192.0.2.1:0
Protocol version = 1
Address Family = 1  Number of addresses = 3
Address 1 = 192.0.2.4
Address 2 = 192.1.2.4
Address 3 = 192.168.34.2
"

5 2017/03/31 09:39:16.78 UTC MINOR: DEBUG #2001 Base LDP
"LDP: LDP
Recv Address packet (msgId 5) from 192.0.2.1:0
Protocol version = 1
Address Family = 1  Number of addresses = 3
Address 1 = 192.0.2.1
Address 2 = 192.1.2.1
Address 3 = 192.168.13.1
"
```

The list of advertised local addresses includes the loopback addresses "loopback1": 192.1.2.1 and 192.1.2.4. These loopback addresses did not occur in the preceding lists of LDP sessions or LDP bindings, but they occur in the LDP session local addresses, as follows:

```
*A:PE-4# show router ldp session local-addresses

===============================================================================
LDP Session Local-Addresses
===============================================================================
-------------------------------------------------------------------------------
```

```
Session with Peer 192.0.2.1:0,
             Local 192.0.2.4:0
-------------------------------------------------------------------------------
IPv4 Sent Addresses:
                192.0.2.4       192.1.2.4       192.168.34.2
IPv6 Sent Addresses:
                None
IPv4 Recv Addresses:
                192.0.2.1       192.1.2.1       192.168.13.1
IPv6 Recv Addresses:
                None
-------------------------------------------------------------------------------
---snip---
```

If there were only unnumbered addresses and no additional loopback addresses, only the system address and other loopback addresses would be sent or received. The interface addresses in the list of local addresses are from numbered interfaces.

Configuring the Local LSR ID

To use the loopback address "loopback1" in the LDP sessions, the local LSR ID is configured as follows:

```
A:PE-4# configure
    router
         ldp
            interface-parameters
                interface "int-PE-4-PE-1"
                    ipv4
                        local-lsr-id interface-name "loopback1"
                        transport-address interface
                        no shutdown
                    exit
                exit
            exit
```

The transport address is the system address, by default, but here it is changed to the address of "loopback1", which is 192.1.2.4. The configuration is similar on PE-1. On PE-2, the system addresses are kept and no additional configuration is required.

LDP hello messages are sent from the transport address to 224.0.0.2 to establish hello adjacencies. The transport address is 192.1.2.4 for the unnumbered interface toward PE-1, and 192.0.2.4 (system address) for the numbered interface toward PE-3. LDP hello adjacencies are verified as follows:

```
*A:PE-4# show router ldp discovery ipv4

===============================================================================
LDP IPv4 Hello Adjacencies
===============================================================================
Interface Name                   Local Addr                               State
```

```
AdjType                          Peer Addr
-------------------------------------------------------------------------------
int-PE-4-PE-1                    192.1.2.4:0                            Estab
link                             192.1.2.1:0

int-PE-4-PE-3                    192.0.2.4:0                            Estab
link                             192.0.2.3:0

-------------------------------------------------------------------------------
No. of IPv4 Hello Adjacencies: 2
===============================================================================
```

The LDP hello adjacencies are established, but the LDP session on the unnumbered interface is non-existent, as follows:

```
*A:PE-4# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.3:0         Link      Established   85        87        0d 00:03:29
192.1.2.1:0         Link      Nonexistent   19        20        0d 00:01:11
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
```

The LDP session is non-existent because the prefix 192.1.2.1/32 is not in the routing table and the LDP session is to be established between 192.1.2.4 and 192.1.2.1. The following export policy is configured and added in the IS-IS context on the nodes:

```
*A:PE-4# configure
    router
        policy-options
            begin
            policy-statement "export_ISIS"
                entry 10
                    from
                        protocol direct
                    exit
                    action accept
                    exit
                exit
                default-action drop
                exit
            exit
            commit
        exit
    exit

*A:PE-4# configure router isis export "export_ISIS"
```

The loopback addresses are now exported in IS-IS. When the loopback addresses are in the routing table, the LDP session is established, as follows:

```
*A:PE-4# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.3:0         Link      Established   146       146       0d 00:06:08
192.1.2.1:0         Link      Established   71        72        0d 00:03:50
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
```

The local LSR ID can also be configured for targeted LDP sessions, as follows:

```
*A:PE-4# configure
    router
        ldp
            targeted-session
                peer 192.1.2.1
                    local-lsr-id "loopback1"
                    no shutdown
                exit
            exit
```

The configuration on PE-1 is similar. The LDP adjacency type is now both link and targeted for peer 192.1.2.1, as follows:

```
*A:PE-4# show router ldp session ipv4

==============================================================================
LDP IPv4 Sessions
==============================================================================
Peer LDP Id         Adj Type  State         Msg Sent  Msg Recv  Up Time
------------------------------------------------------------------------------
192.0.2.3:0         Link      Established   257       257       0d 00:11:07
192.1.2.1:0         Both      Established   185       186       0d 00:08:49
------------------------------------------------------------------------------
No. of IPv4 Sessions: 2
==============================================================================
```

Even though the LDP sessions are established, there is no LDP prefix binding for the loopback address, as follows:

```
*A:PE-4# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262142
Unnumbered                                  1/1/1

192.0.2.1/32                                Swap            262142    262142
Unnumbered                                  1/1/1

192.0.2.2/32                                Push              --      262143
Unnumbered                                  1/1/1

192.0.2.2/32                                Swap            262141    262143
Unnumbered                                  1/1/1

192.0.2.3/32                                Push              --      262143
192.168.34.1                                1/1/2

192.0.2.3/32                                Swap            262140    262143
192.168.34.1                                1/1/2

192.0.2.4/32                                Pop             262143      --
  --                                          --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 7
===============================================================================
```

There is no label mapping for prefix 192.1.2.1/32. SDPs are created toward all other nodes, as follows:

```
*A:PE-4# configure
    service
        sdp 411 mpls create
            far-end 192.0.2.1
            ldp
            no shutdown
        exit
        sdp 412 mpls create
            far-end 192.1.2.1
            ldp
            no shutdown
        exit
        sdp 421 mpls create
            far-end 192.0.2.2
            ldp
            no shutdown
        exit
        sdp 431 mpls create
            far-end 192.0.2.3
            ldp
```

```
                no shutdown
            exit
```

The following output shows that, between PE-4 and PE-1, there are two LDP SDPs: one using the system address and another using the loopback address 192.1.2.x. The configuration on the other nodes is similar. All SDPs that have a system address as the far end are operationally up, whereas the SDP toward 192.1.2.1 is down:

```
*A:PE-4# show service sdp

============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
411    0       1556    192.0.2.1        Up   Up          MPLS    L     TLDP
412    0       0       192.1.2.1        Up   Down        MPLS    L     TLDP
421    0       1556    192.0.2.2        Up   Up          MPLS    L     TLDP
431    0       1556    192.0.2.3        Up   Up          MPLS    L     TLDP
----------------------------------------------------------------------------
Number of SDPs : 4
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-4#
```

The SDP toward 192.1.2.1 is down because the transport tunnel is down, as follows:

```
*A:PE-4# show service sdp detail

===============================================================================
Services: Service Destination Points Details
===============================================================================
---snip---
-------------------------------------------------------------------------------
 Sdp Id 412  -192.1.2.1
-------------------------------------------------------------------------------
Description           : (Not Specified)
SDP Id                : 412                    SDP Source         : manual
Admin Path MTU        : 0                      Oper Path MTU      : 0
Delivery              : MPLS
Far End               : 192.1.2.1
Tunnel Far End        : 192.1.2.1              LSP Types          : LDP

Admin State           : Up                     Oper State         : Down
Signaling             : TLDP                   Metric             : 0
---snip---
Flags                 : TranspTunnDown
---snip---
```

The solution is to manually add an LDP prefix binding, as described in the following section.

Configuring the FEC Originate

The labels to be used for a manually created LDP prefix binding must be chosen from the range for static labels: from 32 to 18431. This range can be retrieved as follows:

```
*A:PE-1# show router mpls-labels label-range

============================================================================
Label Ranges
============================================================================
Label Type      Start Label End Label   Aging       Available   Total
----------------------------------------------------------------------------
Static          32          18431       -           18399       18400
Dynamic         18432       524287      0           505850      505856
    Seg-Route   0           0           -           0           505856
============================================================================
*A:PE-1#
```

To manually add an LDP prefix binding for the loopback prefixes, configure the following:

On PE-4:

```
*A:PE-4# configure
    router
        ldp
            fec-originate 192.1.2.1/32 next-hop 192.0.2.4
                          interface "int-PE-4-PE-1" swap-label 101
            fec-originate 192.1.2.4/32 pop advertised-label 104
        exit
```

On PE-1:

```
*A:PE-1# configure
    router
        ldp
            fec-originate 192.1.2.1/32 pop advertised-label 101
            fec-originate 192.1.2.4/32 next-hop 192.0.2.1
                          interface "int-PE-1-PE-4" swap-label 104
        exit
```

This configuration for unnumbered interfaces includes the interface name, such as int-PE-4-PE-1. This parameter is optional for numbered interfaces.

If the label is chosen from the dynamic range instead of the static range, an error is raised for the pop operation, as follows:

```
*A:PE-1# configure router ldp fec-originate 192.1.2.4/32 pop advertised-label 100001
                                                                             ^
Error: Invalid parameter. Label value not in allowed range
```

For interoperability, no error is raised for the swap operation.

As a result, three active LDP bindings are added: one pop operation for the local loopback prefix, and a swap and a push operation for the remote loopback prefix, as follows:

```
*A:PE-4# show router ldp bindings active prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
        (S) - Static           (M) - Multi-homed Secondary Support
        (B) - BGP Next Hop     (BU) - Alternate Next-hop for Fast Re-Route
        (I) - SR-ISIS Next Hop (O) - SR-OSPF Next Hop
        (C) - FEC resolved with class-based-forwarding
===============================================================================
LDP IPv4 Prefix Bindings (Active)
===============================================================================
Prefix                                      Op              IngLbl    EgrLbl
EgrNextHop                                  EgrIf/LspId
-------------------------------------------------------------------------------
192.0.2.1/32                                Push              --      262142
Unnumbered                                  1/1/1

192.0.2.1/32                                Swap            262142    262142
Unnumbered                                  1/1/1

192.0.2.2/32                                Push              --      262143
Unnumbered                                  1/1/1

192.0.2.2/32                                Swap            262141    262143
Unnumbered                                  1/1/1

192.0.2.3/32                                Push              --      262143
192.168.34.1                                1/1/2

192.0.2.3/32                                Swap            262140    262143
192.168.34.1                                1/1/2

192.0.2.4/32                                Pop             262143      --
  --                                          --

192.1.2.1/32                                Push              --      101
Unnumbered                                  1/1/1

192.1.2.1/32                                Swap            262139    101
Unnumbered                                  1/1/1

192.1.2.4/32(S)                             Pop             104         --
  --                                          --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Active Bindings: 10
```

```
===============================================================================
```

The following output shows that the SDPs are all operationally up, including the one toward the loopback address:

```
*A:PE-4# show service sdp
============================================================================
Services: Service Destination Points
============================================================================
SdpId  AdmMTU  OprMTU  Far End          Adm  Opr         Del     LSP   Sig
----------------------------------------------------------------------------
411    0       1556    192.0.2.1        Up   Up          MPLS    L     TLDP
412    0       1556    192.1.2.1        Up   Up          MPLS    L     TLDP
421    0       1556    192.0.2.2        Up   Up          MPLS    L     TLDP
431    0       1556    192.0.2.3        Up   Up          MPLS    L     TLDP
----------------------------------------------------------------------------
Number of SDPs : 4
----------------------------------------------------------------------------
Legend: R = RSVP, L = LDP, B = BGP, M = MPLS-TP, n/a = Not Applicable
        I = SR-ISIS, O = SR-OSPF, T = SR-TE, F = FPE
============================================================================
*A:PE-4#
```

LDP FRR Loop-Free Alternate on Unnumbered Interfaces

LDP FRR Loop-Free Alternate (LFA) is supported on unnumbered interfaces and on numbered interfaces. For information about LDP FRR LFA, see chapter MPLS LDP FRR using ISIS as IGP. LDP FRR LFA can be configured as follows:

```
*A:PE-4# configure router isis loopfree-alternate
*A:PE-4# configure router ldp fast-reroute
*A:PE-4# configure router ip-fast-reroute
```

Enabling FRR LFA is a local decision. In this configuration, it is configured on all nodes. The LFA coverage can be retrieved as follows:

```
*A:PE-4# show router isis lfa-coverage

===============================================================================
Rtr Base ISIS Instance 0 LFA Coverage
===============================================================================
Topology         Level   Node           IPv4                IPv6
-------------------------------------------------------------------------------
IPV4 Unicast     L1      3/3(100%)      7/7(100%)           0/0(0%)
IPV6 Unicast     L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L1      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Unicast     L2      0/0(0%)        7/7(100%)           0/0(0%)
IPV6 Unicast     L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV4 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
IPV6 Multicast   L2      0/0(0%)        0/0(0%)             0/0(0%)
===============================================================================
*A:PE-4#
```

There is protection for the three other nodes and for all remote prefixes in the routing table, which can be verified as follows:

```
*A:PE-4# show router route-table alternative

===============================================================================
Route Table (Router: Base)
===============================================================================
Dest Prefix[Flags]                            Type    Proto     Age        Pref
      Next Hop[Interface Name]                                    Metric
      Alt-NextHop                                                Alt-
                                                                Metric
-------------------------------------------------------------------------------
192.0.2.1/32                                  Remote  ISIS      00h57m41s  15
       int-PE-4-PE-1                                                10
       192.168.34.1 (LFA)                                           20
192.0.2.2/32                                  Remote  ISIS      00h53m52s  15
       int-PE-4-PE-1                                                20
       192.168.34.1 (LFA)                                           30
192.0.2.3/32                                  Remote  ISIS      00h34m39s  15
       192.168.34.1                                                 10
       int-PE-4-PE-1 (LFA)                                          20
192.0.2.4/32                                  Local   Local     00h58m17s  0
       system                                                       0
192.1.2.1/32                                  Remote  ISIS      00h19m17s  15
       int-PE-4-PE-1                                                10
       192.168.34.1 (LFA)                                           20
192.1.2.2/32                                  Remote  ISIS      00h19m17s  15
       int-PE-4-PE-1                                                20
       192.168.34.1 (LFA)                                           30
192.1.2.4/32                                  Local   Local     00h53m39s  0
       loopback1                                                    0
192.168.13.0/30                               Remote  ISIS      00h34m39s  15
       int-PE-4-PE-1                                                20
       192.168.34.1 (LFA)                                           30
192.168.23.0/30                               Remote  ISIS      00h34m39s  15
       192.168.34.1                                               10010
       int-PE-4-PE-1 (LFA)                                        10020
192.168.34.0/30                               Local   Local     00h34m40s  0
       int-PE-4-PE-3                                                0
-------------------------------------------------------------------------------
No. of Routes: 10
Flags: n = Number of times nexthop is repeated
       Backup = BGP backup route
       LFA = Loop-Free Alternate nexthop
       S = Sticky ECMP requested
===============================================================================
*A:PE-4#
```

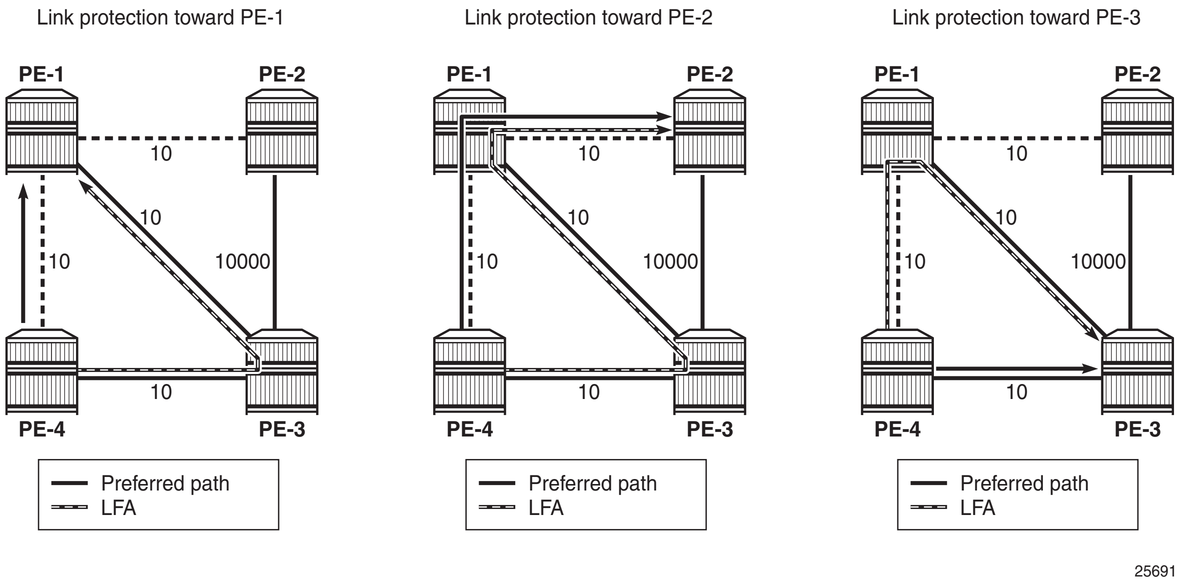
For unnumbered interfaces, the interface name is shown (int-PE-4-PE-1); for numbered interfaces, the next hop IP address is shown (192.168.34.1). The LFA type is link protection for the three nodes, as follows:

```
*A:PE-4# show router isis topology lfa detail
```

```
===============================================================================
Rtr Base ISIS Instance 0 Topology Table
===============================================================================
-------------------------------------------------------------------------------
IS-IS IP paths (MT-ID 0),   Level 1
-------------------------------------------------------------------------------
Node       : PE-1.00                              Metric      : 10
Interface  : int-PE-4-PE-1                        SNPA        : none
Nexthop    : PE-1

LFA intf   : int-PE-4-PE-3                        LFA Metric  : 20
LFA nh     : PE-3                                 LFA type    : linkProtection

Node       : PE-2.00                              Metric      : 20
Interface  : int-PE-4-PE-1                        SNPA        : none
Nexthop    : PE-1

LFA intf   : int-PE-4-PE-3                        LFA Metric  : 30
LFA nh     : PE-3                                 LFA type    : linkProtection

Node       : PE-3.00                              Metric      : 10
Interface  : int-PE-4-PE-3                        SNPA        : none
Nexthop    : PE-3

LFA intf   : int-PE-4-PE-1                        LFA Metric  : 20
LFA nh     : PE-1                                 LFA type    : linkProtection

===============================================================================
*A:PE-4#
```

The LFA protection is shown in Figure 345.

Figure 345   LDP FRR LFA Link Protection on PE-4


Link protection toward PE-1
Link protection toward PE-2
Link protection toward PE-3
PE-1
PE-2
PE-3
PE-4
10
10
10
10000
10
Preferred path
LFA
25691

The LDP bindings for FRR LFA indicate alternate ("BU") in the list, as follows:

```
*A:PE-4# show router ldp bindings prefixes ipv4

===============================================================================
LDP Bindings (IPv4 LSR ID 192.0.2.4)
             (IPv6 LSR ID ::)
===============================================================================
Label Status:
        U - Label In Use, N - Label Not In Use, W - Label Withdrawn
        WP - Label Withdraw Pending, BU - Alternate For Fast Re-Route
        e - Label ELC
FEC Flags:
        LF - Lower FEC, UF - Upper FEC, BA - ASBR Backup FEC
===============================================================================
LDP IPv4 Prefix Bindings
===============================================================================
Prefix                                      IngLbl                    EgrLbl
Peer                                        EgrIntf/LspId
EgrNextHop
-------------------------------------------------------------------------------
192.0.2.1/32                                262142U                   262141BU
192.0.2.3:0                                 1/1/2
192.168.34.1

192.0.2.1/32                                  --                      262142
192.1.2.1:0                                 1/1/1
Unnumbered

192.0.2.2/32                                262141U                   262142BU
192.0.2.3:0                                 1/1/2
192.168.34.1

192.0.2.2/32                                262141N                   262143
192.1.2.1:0                                 1/1/1
Unnumbered

192.0.2.3/32                                  --                      262143
192.0.2.3:0                                 1/1/2
192.168.34.1

192.0.2.3/32                                262140U                   262141BU
192.1.2.1:0                                 1/1/1
Unnumbered

192.0.2.4/32                                262143U                     --
192.0.2.3:0                                   --
  --

192.0.2.4/32                                262143U                     --
192.1.2.1:0                                   --
  --

192.1.2.1/32                                262139U                   262139BU
192.0.2.3:0                                 1/1/2
192.168.34.1

192.1.2.1/32                                  --                      101
192.1.2.1:0                                 1/1/1
```

```
Unnumbered

192.1.2.4/32                                104U                        --
192.0.2.3:0                                   --
  --

192.1.2.4/32                                104U                        --
192.1.2.1:0                                   --
  --

-------------------------------------------------------------------------------
No. of IPv4 Prefix Bindings: 12
===============================================================================
```

Conclusion

Unnumbered interfaces were initially supported for SONET/SDH/ATM/FR, and later also on Ethernet access ports. IS-IS adjacencies and OSPF neighbors can be established on unnumbered interfaces. This chapter showed that unnumbered interfaces can be added to RSVP or LDP. Most features that are supported on numbered interfaces are also supported on unnumbered interfaces.

Customer Document and Product Support

Customer documentation

Customer Documentation Welcome Page

Technical Support

Product Support Portal

Documentation feedback

Customer Documentation Feedback