



Nokia Service Router Linux
7215 Interconnect System
7220 Interconnect Router
7250 Interconnect Router
7730 Service Interconnect Router
Release 24.10

VPN Services Guide

3HE 20959 AAAA TQZZA
Edition: 01
November 2024

Nokia is committed to diversity and inclusion. We are continuously reviewing our customer documentation and consulting with standards bodies to ensure that terminology is inclusive and aligned with the industry. Our future customer documentation will be updated accordingly.

This document includes Nokia proprietary and confidential information, which may not be distributed or disclosed to any third parties without the prior written consent of Nokia.

This document is intended for use by Nokia's customers ("You"/"Your") in connection with a product purchased or licensed from any company within Nokia Group of Companies. Use this document as agreed. You agree to notify Nokia of any errors you may find in this document; however, should you elect to use this document for any purpose(s) for which it is not intended, You understand and warrant that any determinations You may make or actions You may take will be based upon Your independent judgment and analysis of the content of this document.

Nokia reserves the right to make changes to this document without notice. At all times, the controlling version is the one available on Nokia's site.

No part of this document may be modified.

NO WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY OF AVAILABILITY, ACCURACY, RELIABILITY, TITLE, NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, IS MADE IN RELATION TO THE CONTENT OF THIS DOCUMENT. IN NO EVENT WILL NOKIA BE LIABLE FOR ANY DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, DIRECT, INDIRECT, INCIDENTAL OR CONSEQUENTIAL OR ANY LOSSES, SUCH AS BUT NOT LIMITED TO LOSS OF PROFIT, REVENUE, BUSINESS INTERRUPTION, BUSINESS OPPORTUNITY OR DATA THAT MAY ARISE FROM THE USE OF THIS DOCUMENT OR THE INFORMATION IN IT, EVEN IN THE CASE OF ERRORS IN OR OMISSIONS FROM THIS DOCUMENT OR ITS CONTENT.

Copyright and trademark: Nokia is a registered trademark of Nokia Corporation. Other product names mentioned in this document may be trademarks of their respective owners.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

© 2024 Nokia.

Table of contents

1	About this guide.....	8
1.1	Precautionary and information messages.....	8
1.2	Conventions.....	8
2	What's new.....	10
3	Overview.....	11
3.1	About EVPN.....	11
3.2	About the service infrastructure for EVPN.....	11
3.3	About EVPN for Layer 2.....	12
3.4	About EVPN for Layer 3.....	13
3.5	About EVPN for multicast.....	14
3.6	About IP-VPN.....	15
4	EVPN services.....	17
4.1	Layer 2 services infrastructure.....	17
4.1.1	MAC-VRF network instance.....	17
4.1.1.1	MAC selection.....	18
4.1.1.2	MAC duplication detection and actions.....	18
4.1.1.3	Bridge table configuration.....	19
4.1.2	Interface extensions for Layer 2 services.....	19
4.1.2.1	Traffic classification and ingress/egress mapping actions.....	19
4.1.2.2	Routed and bridged subinterfaces.....	20
4.1.3	IRB interfaces.....	20
4.1.3.1	Using ACLs with IRB interfaces and Layer 2 subinterfaces.....	21
4.1.4	Layer 2 services configuration.....	22
4.1.4.1	MAC-VRF network instance configuration example.....	22
4.1.4.2	Bridged subinterface configuration example.....	23
4.1.4.3	IRB interface configuration example.....	24
4.1.5	Displaying bridge table information.....	25
4.1.6	Deleting entries from the bridge table.....	28
4.1.7	Storm control on bridged subinterfaces.....	28
4.1.7.1	Configuring storm control for an interface.....	29
4.1.7.2	Displaying storm control information.....	30

4.1.8	L2CP transparency.....	31
4.1.8.1	Configuring L2CP transparency.....	33
4.1.8.2	Displaying L2CP transparency information.....	34
4.1.8.3	Displaying L2CP transparency statistics.....	36
4.1.9	Split horizon groups on MAC-VRF.....	38
4.1.9.1	Configuring split horizon groups.....	39
4.1.9.2	Split horizon group configuration example.....	39
4.1.9.3	Platform specifications for split horizon groups.....	40
4.1.10	Pseudowires.....	41
4.1.10.1	Pseudowires in MAC-VRF network instances.....	41
4.1.10.2	EVPN-MPLS and VPLS seamless integration (RFC 8560).....	46
4.1.10.3	Pseudowires in VPWS network instances.....	48
4.1.11	Server aggregation configuration example.....	50
4.1.11.1	Configuration for server aggregation example.....	52
4.2	VXLAN v4.....	53
4.2.1	VXLAN configuration.....	54
4.2.1.1	Source and destination VTEP addresses.....	57
4.2.1.2	Ingress or egress VNI.....	57
4.2.1.3	VLAN tagging for VXLAN.....	57
4.2.1.4	Network instance and interface MTU.....	57
4.2.1.5	Fragmentation for VXLAN traffic.....	58
4.2.2	VXLAN and ECMP.....	58
4.2.3	VXLAN ACLs.....	58
4.2.4	QoS for VXLAN tunnels.....	59
4.2.4.1	Configuring a VXLAN classifier policy.....	60
4.2.5	VXLAN statistics collection.....	60
4.2.5.1	Enabling VXLAN statistics collection.....	60
4.2.5.2	Displaying VXLAN statistics.....	61
4.2.5.3	Clearing VXLAN statistics.....	62
4.3	MPLS transport for EVPN services.....	62
4.4	EVPN for Layer 2 ELAN services.....	63
4.4.1	Basic configuration for EVPN for Layer 2 ELAN services.....	63
4.4.1.1	EVPN Layer 2 basic routes.....	65
4.4.1.2	Creation of EVPN destinations based on received EVPN routes.....	66
4.4.1.3	EVPN route selection.....	69
4.4.1.4	BGP next hop configuration for EVPN routes.....	69

4.4.2	MAC duplication detection for Layer 2 loop prevention in EVPN.....	70
4.4.3	EVPN Layer 2 multihoming.....	70
4.4.3.1	EVPN Layer 2 multihoming procedures.....	71
4.4.3.2	All-active multihoming.....	72
4.4.3.3	Single-active multihoming.....	72
4.4.3.4	Reload-delay timer.....	74
4.4.3.5	EVPN multihoming configuration example.....	76
4.4.4	Layer 2 proxy-ARP/ND.....	79
4.4.4.1	Dynamic learning for proxy-ARP/ND.....	80
4.4.4.2	Static proxy-ARP entries.....	82
4.4.4.3	EVPN learning for proxy-ARP/ND.....	83
4.4.4.4	Configuring proxy-ARP/ND traffic flooding options.....	84
4.4.4.5	Proxy-ARP/ND duplicate IP detection.....	86
4.4.4.6	Proxy-ARP/ND table.....	89
4.4.4.7	EVPN ARP/ND extended community flags.....	91
4.4.4.8	Displaying proxy-ARP/ND information.....	95
4.4.4.9	Displaying proxy-ARP/ND statistics.....	97
4.4.5	EVPN VLAN-aware bundle services.....	98
4.4.5.1	Configuring the Ethernet Tag ID for VLAN-aware bundle services.....	99
4.4.5.2	Displaying the Ethernet Tag ID for VLAN-aware bundle services.....	100
4.5	EVPN for Layer 2 ELINE services.....	100
4.5.1	Configuring EVPN-VPWS services.....	101
4.5.2	Operational state in VPWS network instances.....	102
4.5.3	Layer 2 attribute mismatch.....	103
4.5.4	Flow-aware transport label for EVPN-VPWS.....	104
4.5.5	Interface/LAG link loss forwarding.....	104
4.6	EVPN for Layer 3.....	106
4.6.1	EVPN Layer 3 basic configuration.....	106
4.6.1.1	Asymmetric IRB.....	107
4.6.1.2	Symmetric IRB interface-less ip-vrf-to-ip-vrf model.....	112
4.6.2	Anycast gateways.....	118
4.6.3	EVPN Layer 3 multihoming and anycast gateways.....	120
4.6.4	EVPN Layer 3 host route mobility.....	122
4.6.5	EVPN-VXLAN IFL interoperability with EVPN IFF.....	125
4.6.6	VIP discovery for redundant servers.....	125
4.6.6.1	Configuring VIP discovery.....	127

4.6.6.2	Displaying VIP discovery information.....	128
4.6.7	Layer 3 proxy-ARP/ND.....	129
4.6.7.1	Layer 3 proxy-ARP.....	130
4.6.7.2	Layer 3 proxy-ND.....	132
4.6.8	EVPN IP aliasing.....	132
4.6.8.1	Configuring Layer 3 Ethernet Segments for IP aliasing.....	137
4.6.9	Unequal ECMP for EVPN IP prefix routes.....	142
4.6.9.1	Advertising the EVPN link-bandwidth extended community.....	144
4.6.9.2	Enabling weighted ECMP.....	145
4.6.9.3	Configuring weighted ECMP for received PE-CE BGP routes.....	146
4.6.9.4	Displaying ECMP weights.....	146
4.6.10	Resilient hashing for EVPN IFL.....	147
4.6.10.1	Resilient hashing and unequal ECMP for EVPN IFL.....	148
4.7	Dynamic multicast ID management.....	150
4.7.1	Displaying MCID information.....	151
4.8	EVPN for multicast.....	152
4.8.1	IGMP/MLD snooping.....	152
4.8.2	EVPN multicast for Layer 2.....	153
4.8.2.1	EVPN multihoming and multicast synchronization.....	154
4.8.2.2	Configuring EVPN multicast for Layer 2.....	156
4.8.2.3	Displaying EVPN multicast for Layer 2 information.....	159
4.9	BGP and routing policy extensions for EVPN.....	166
4.9.1	BGP extensions for EVPN.....	166
4.9.2	Best path selection for EVPN IFL routes.....	167
4.9.3	Routing policy extensions for EVPN.....	170
4.10	Route internal tags (tag-sets in routing policies).....	170
4.10.1	Order of precedence when multiple applications write internal tags.....	172
4.10.2	Configuring route internal tags.....	174
5	IP-VPN services.....	178
5.1	Configuring IP-VPN services.....	178
6	Multi-instance IP-VRF.....	181
7	Next-hop-self route reflector and inter-AS option B.....	184
7.1	Displaying next-hop-self route reflector and inter-AS option B information.....	193

8	Combined ECMP for BGP PE-CE and EVPN IFL/IP-VPN.....	197
8.1	Combined ECMP example.....	198
9	EVPN configuration examples.....	205
9.1	All-active redundant connectivity example.....	205
9.1.1	Configuration for all-active connectivity example.....	206
9.2	Hierarchical active-active connectivity example.....	209
9.2.1	Configuration for hierarchical active-active connectivity example.....	211
9.3	EVPN multihoming as standalone solution for MC-LAG.....	217
9.3.1	Configuration for EVPN multihoming as standalone MC-LAG.....	219

1 About this guide

This document describes basic configuration for Layer 2 and Layer 3 Virtual Private Network service functionality on the Nokia Service Router Linux (SR Linux). It presents examples to configure and implement various protocols and services.

This document is intended for network technicians, administrators, operators, service providers, and others who need to understand how the router is configured.

**Note:**

This manual covers the current release and may also contain some content to be released in later maintenance loads. See the *SR Linux Release Notes* for information about features supported in each load.

Configuration and command outputs shown in this guide are examples only; actual displays may differ depending on supported functionality and user configuration.

1.1 Precautionary and information messages

The following are information symbols used in the documentation.



DANGER: Danger warns that the described activity or situation may result in serious personal injury or death. An electric shock hazard could exist. Before you begin work on this equipment, be aware of hazards involving electrical circuitry, be familiar with networking environments, and implement accident prevention procedures.



WARNING: Warning indicates that the described activity or situation may, or will, cause equipment damage, serious performance problems, or loss of data.



Caution: Caution indicates that the described activity or situation may reduce your component or system performance.



Note: Note provides additional operational information.



Tip: Tip provides suggestions for use or best practices.

1.2 Conventions

Nokia SR Linux documentation uses the following command conventions.

- **Bold** type indicates a command that the user must enter.
- Input and output examples are displayed in `Courier` text.
- An open right-angle bracket indicates a progression of menu choices or simple command sequence (often selected from a user interface). Example: **start** > **connect to**.

- A vertical bar (|) indicates a mutually exclusive argument.
- Square brackets ([]) indicate optional elements.
- Braces ({ }) indicate a required choice. When braces are contained within square brackets, they indicate a required choice within an optional element.
- *Italic* type indicates a variable.

Generic IP addresses are used in examples. Replace these with the appropriate IP addresses used in the system.

2 What's new

Topic	Location
Tunneling EFM-OAM, E-LMI, and ESMC L2CP frames on 7730 SXR platforms	L2CP transparency
EVPN-VPWS services on VPWS network instances	EVPN for Layer 2 ELINE services Configuring EVPN-VPWS services Operational state in VPWS network instances
Pseudowires for Layer 2 services (MAC-VRF and VPWS)	Pseudowires
Interface/LAG link loss forwarding (LLF)	Interface/LAG link loss forwarding
Flow-aware transport (FAT) label for EVPN-VPWS	Flow-aware transport label for EVPN-VPWS
Storm control enhancements on bridged subinterfaces	Storm control on bridged subinterfaces
Multi-instance IP-VRF	Multi-instance IP-VRF
Split horizon groups (SHGs) on MAC-VRF	Split horizon groups on MAC-VRF
Configuring IP-VPN services	IP-VPN services

3 Overview

This chapter contains the following topics:

- [About EVPN](#)
- [About the service infrastructure for EVPN](#)
- [About EVPN for Layer 2](#)
- [About EVPN for Layer 3](#)
- [About EVPN for multicast](#)
- [About IP-VPN](#)

3.1 About EVPN

Ethernet Virtual Private Network (EVPN) is a technology that bridges Layer 2 and routes Layer 3 VPN traffic across a shared provider environment. When extending Broadcast Domains (BDs), EVPN instances configured on Provider Edge (PE) routers act as virtual bridges, facilitating traffic between Customer Edge (CE) devices at different locations. Conversely, when connecting devices in different IP subnets, EVPN instances function as virtual routers, enabling inter-subnet forwarding for devices of the same tenant.

Essentially, PE routers exchange reachability information, encapsulate Layer 2 or Layer 3 traffic from CE devices, and forward it across the provider or data center (DC) network. EVPN is a standard technology in multi-tenant DCs (RFC 8365) and MPLS/Segment Routing networks (RFC 7432).

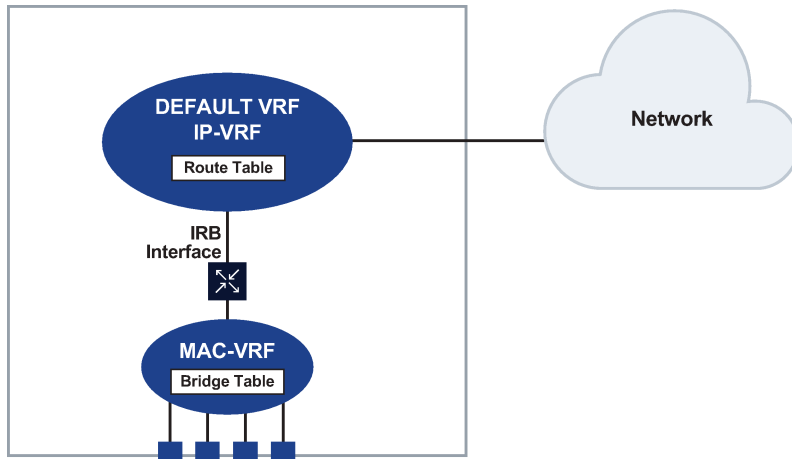
The SR Linux EVPN solution supports EVPN multi-tenant DCs using VXLAN as the data plane, and in service provider networks, using MPLS as the data plane.

3.2 About the service infrastructure for EVPN

The SR Linux infrastructure to support EVPN services uses network-instances of type MAC-VRF and IP-VRF, along with IRB interfaces. The MAC-VRF network-instance is associated with a network-instance of type `default` or `ip-vrf` via an Integrated Routing and Bridging (IRB) interface.

The following figure shows the relationship between an IRB interface and MAC-VRF, and IP-VRF network-instance types.

Figure 1: MAC-VRF, IRB interface, and IP-VRF

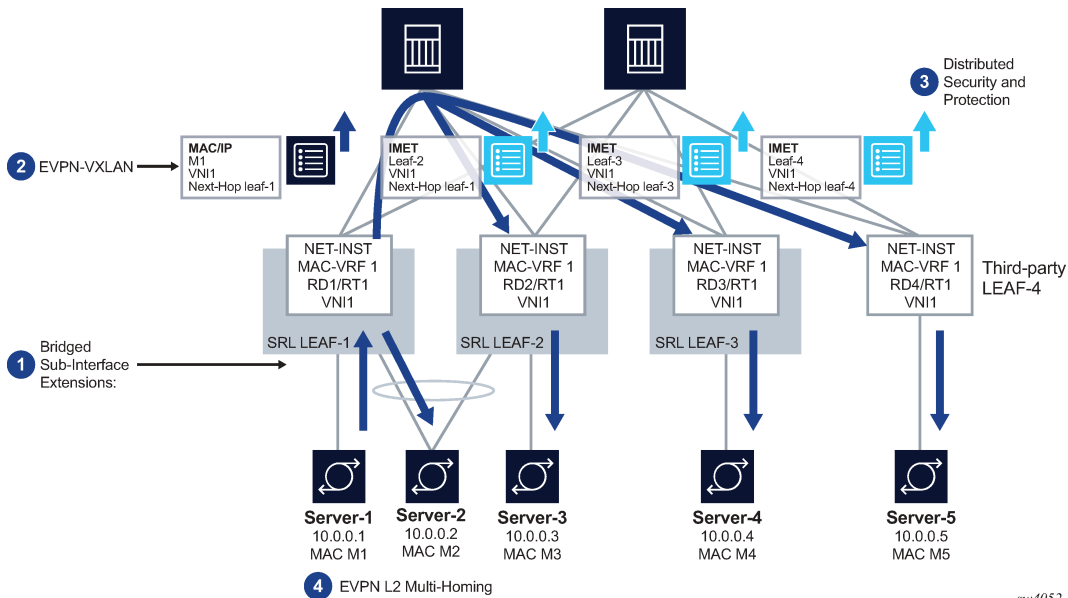


See [Layer 2 services infrastructure](#) for information about MAC-VRFs, IP-VRFs, and IRB interfaces in SR Linux.

3.3 About EVPN for Layer 2

The primary usage for EVPN for Layer 2 is the extension of a BD in overlay multi-tenant DCs or service provider networks. An example of this type of topology in DCs is shown in the following figure:

Figure 2: BD extension in overlay DCs



SR Linux features that support this topology fall into the following categories:

- bridged subinterfaces, including:

- default subinterfaces, configured with the parameter **vlan-id optional**, which capture untagged and non-explicitly configured VLAN-tagged frames on tagged subinterfaces
- transparency of inner qtags not being used for service classification
- EVPN control and data plane extensions, as described in RFC 8365 for VXLAN tunnels and in RFC 7432 for MPLS tunnels:
 - EVPN routes type MAC/IP and IMET (Inclusive Multicast Ethernet Tag) routes
 - VXLANv4 model for MAC-VRFs
 - MPLS model for MAC-VRFs
- distributed security and protection, including:
 - an extension to the MAC duplication mechanism that can be applied to MACs received from EVPN
 - protection of static MACs
- EVPN Layer 2 multihoming, including:
 - the Ethernet Segment (ES) model definition for all-active and single-active multihoming
 - interface-level reload-delay timers to avoid service impact when links recover
 - load-balancing and redundancy using aliasing, as in RFC 7432

SR Linux also supports static and TLDP-signaled pseudowires, which can be used on MAC-VRF and VPWS instances.

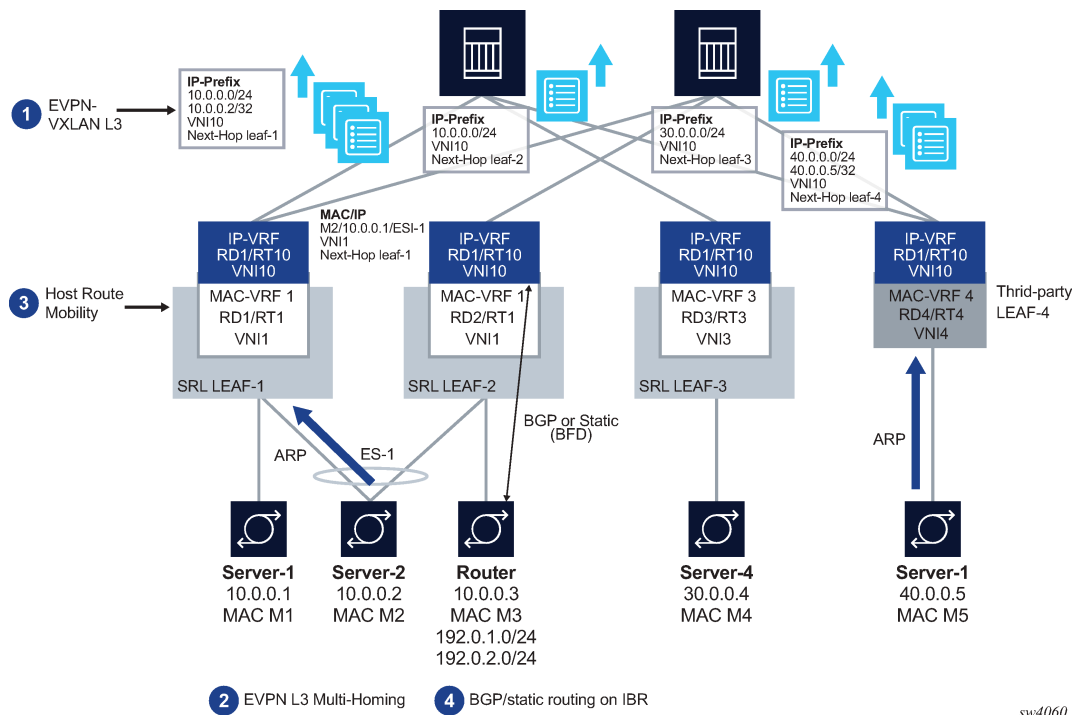
See [EVPN for Layer 2 ELAN services](#) for information about the components of EVPN Layer 2 in SR Linux.

3.4 About EVPN for Layer 3

The primary usage for EVPN for VXLAN tunnels (Layer 3) is inter-subnet-forwarding for unicast traffic within the same tenant infrastructure. This type of topology is shown in the following figure:

The primary usage for EVPN for Layer 3 is inter-subnet-forwarding for unicast traffic within the same tenant infrastructure, which can use a shared provider network or a data center. An example of this type of topology in a data center is shown in the following figure:

Figure 3: Inter-subnet forwarding with EVPN Layer 3



SR Linux features that support this topology fall into the following categories:

- EVPN Layer 3 control plane (mostly EVPN IP prefix routes or RT5s) and data plane, as described in RFC 9136
- EVPN Layer 3 multihoming on MAC-VRFs with IRB interfaces that use anycast gateway IP and MAC addresses in all leaves attached to the same BD
- host route mobility procedures to allow fast mobility of hosts between leaf nodes attached to the same BD

Other supported features include:

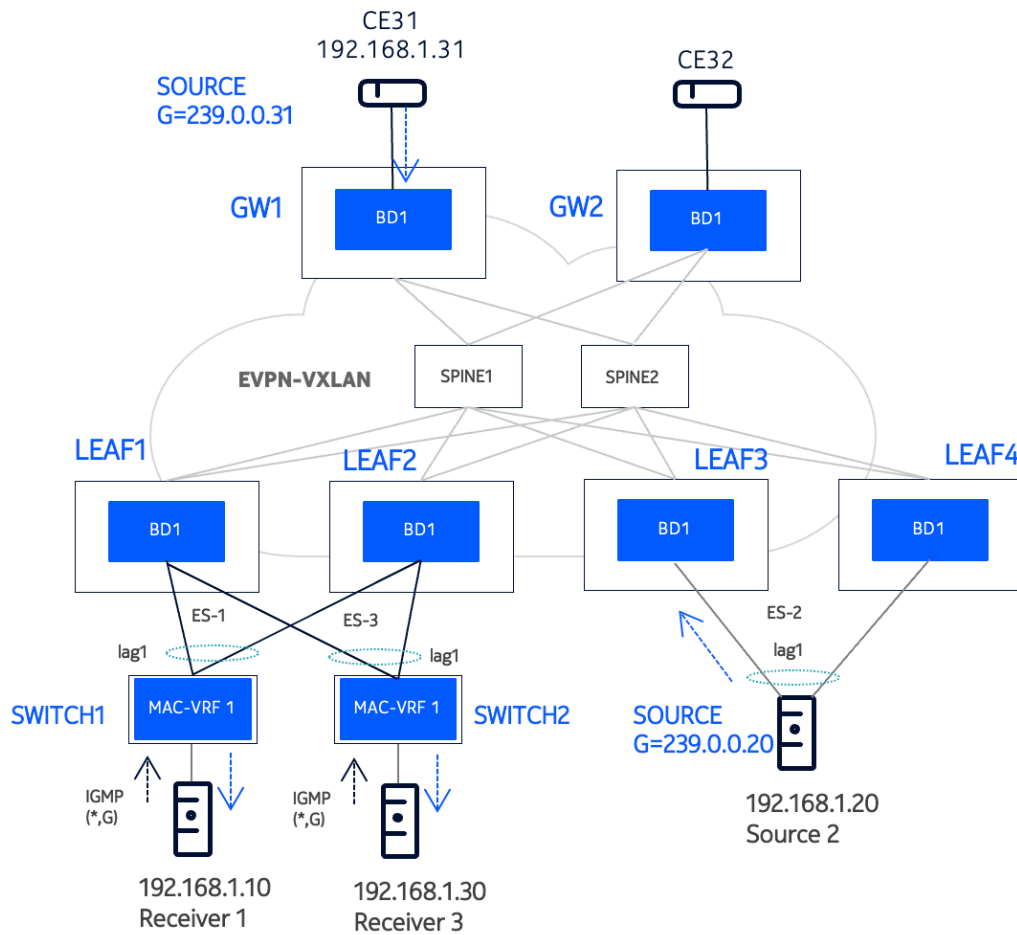
- interface-less (IFL) model interoperability with unnumbered interface-ful (IFF) model
- ECMP over EVPN, including unequal ECMP, IP aliasing, and combined ECMP
- support for interface-level OAM (ping) in anycast deployments

[EVPN for Layer 3](#) describes the components of EVPN Layer 3 in SR Linux.

3.5 About EVPN for multicast

The primary usage for EVPN for multicast is to forward IP multicast traffic from sources to receivers in an efficient way. An example of this type of service in data centers is shown in the following figure:

Figure 4: EVPN for multicast



SR Linux features that support this topology fall into the following categories:

- IGMP/MLD snooping on MAC-VRFs
- EVPN IGMP/MLD proxy support in MAC-VRFs, as described in RFC 9251
- EVPN multihoming support on MAC-VRFs where IGMP/MLD snooping is enabled, and synchronization of the multicast states on all the PEs attached to the same Ethernet Segment, as described in RFC 9251

[EVPN for multicast](#) describes the components of EVPN multicast in SR Linux.

3.6 About IP-VPN

IP-VPN services use a combination of MP-BGP and MPLS to distribute IPv4/v6 routing information and provide Layer 3 VPN services.

Each IP-VPN consists of a set of customer end points connected to one or more PE routers. Each associated PE router maintains a separate IP forwarding table for each IP-VPN instance. The PE routers exchange the routing information configured or learned from all customer sites via MP-BGP peering. Each route exchanged via the MP-BGP protocol includes a Route Distinguisher (RD), which identifies the IP-VPN association and handles any potential IP address overlap.

Multi-Protocol BGP (MP-BGP) is used to exchange the routes of a particular VPN among the PE routers that are attached to that VPN.

4 EVPN services

This chapter describes the components of EVPN services on SR Linux. It contains the following topics:

- [Layer 2 services infrastructure](#)
- [VXLAN v4](#)
- [MPLS transport for EVPN services](#)
- [EVPN for Layer 2 ELAN services](#)
- [EVPN for Layer 2 ELINE services](#)
- [EVPN for Layer 3](#)
- [Dynamic multicast ID management](#)
- [EVPN for multicast](#)
- [BGP and routing policy extensions for EVPN](#)
- [Route internal tags \(tag-sets in routing policies\)](#)

4.1 Layer 2 services infrastructure

This chapter describes the components of Layer 2 services on SR Linux. It contains the following topics:

- [MAC-VRF network instance](#)
- [Interface extensions for Layer 2 services](#)
- [IRB interfaces](#)
- [Layer 2 services configuration](#)
- [Displaying bridge table information](#)
- [Deleting entries from the bridge table](#)
- [Storm control on bridged subinterfaces](#)
- [L2CP transparency](#)
- [Pseudowires](#)
- [Split horizon groups on MAC-VRF](#)

4.1.1 MAC-VRF network instance

The network-instance type MAC-VRF functions as a broadcast domain. Each MAC-VRF network instance builds a bridge table composed of MAC addresses that can be learned via the data path on network instance interfaces or via static configuration. You can configure the size of the bridge table for each MAC-VRF network instance, as well as the aging for dynamically learned MAC addresses and other parameters related to the bridge table.

The MAC-VRF network-instance type features a MAC duplication mechanism that monitors MAC address moves across network-instance interfaces and across interfaces.

4.1.1.1 MAC selection

Each MAC-VRF network instance builds a bridge table to forward Layer 2 frames based on a MAC address lookup. SR Linux selects the MAC addresses to be sent for installation to the line card (XDP), based on the following priority:

1. local application MACs (for example, IRB interface MACs)
2. local static MACs
3. EVPN static MACs
4. local duplicate MACs
5. learned and EVPN-learned MACs

4.1.1.2 MAC duplication detection and actions

MAC duplication is the mechanism used by SR Linux for loop prevention. MAC duplication monitors MAC addresses that move between subinterfaces. It consists of detection, actions, and process restart.

4.1.1.2.1 MAC duplication detection

Detection of duplicate MAC addresses is necessary when extending broadcast domains to multiple leaf nodes. SR Linux supports a MAC duplication mechanism that monitors MAC address moves across network-instance interfaces and across interfaces and EVPN.

A MAC address is considered a *duplicate* when its number of detected moves is greater than a configured threshold within a configured time frame where the moves are observed. When the threshold is exceeded, the system holds on to the prior local destination of the MAC and executes an action.

4.1.1.2.2 MAC duplication actions

The action taken on detecting one or more MAC addresses as duplicate on a subinterface can be configured for the MAC-VRF network instance or for the subinterface. The following are the configurable actions:

- **oper-down**

When one or more duplicate MAC addresses are detected on the subinterface, the subinterface is brought operationally down.

- **blackhole**

On detection of a duplicate mac on the subinterface, the mac is blackholed.

- **stop learning**

On detection of a duplicate mac on the subinterface, the MAC address is no longer relearned on this or any subinterface. This is the default action for a MAC-VRF network instance.

- **use-network-instance-action**

(Only available for subinterfaces) Use the action specified for the MAC-VRF network instance. This is the default action for a subinterface.

4.1.1.2.3 MAC duplication process restarts

When at least one duplicate MAC address is detected, the duplicate MAC addresses are visible in the state datastore and can be displayed with the **info from state mac-duplication duplicate-entries** CLI command. See [Displaying bridge table information](#).

4.1.1.2.4 Configurable hold-down time

The **info from state mac-duplication duplicate-entries** command also displays the hold-down time for each duplicate MAC address. After the hold-down time expires for all of the duplicate MAC addresses for the subinterface, the oper-down or stop-learning action is cleared, and the subinterface is brought operationally up or starts learning again.

The hold-down time is configurable, ranging from 2 and 60 minutes. You can optionally specify **indefinite** for the hold-down time, which prevents the oper-down or stop-learning action from being cleared after a duplicate MAC address is detected; in this case, you can manually clear the oper-down or stop-learning action by changing the **mac-duplication** configuration or using the **tools network-instance bridge-table mac-duplication** command.

4.1.1.3 Bridge table configuration

The bridge table, its MAC address limit, and maximum number of entries can be configured on a per mac-vrf or per-subinterface basis.

When the size of the bridge table exceeds its maximum number of entries, the MAC addresses are removed in reverse order of the priority listed in [MAC selection](#).

You can also configure aging for dynamically learned MAC addresses and other parameters related to the bridge table.

4.1.2 Interface extensions for Layer 2 services

To accommodate the Layer 2 services infrastructure, SR Linux interfaces support the following features:

- traffic classification and ingress or egress mapping actions
- subinterfaces of types routed and bridged

4.1.2.1 Traffic classification and ingress/egress mapping actions

On MAC-VRF network instances, traffic can be classified based on VLAN tagging. Interfaces where VLAN tagging is set to **false** or **true** can be used with MAC-VRF network instances.

A default subinterface can be specified, which captures untagged and non-explicitly configured VLAN-tagged frames in tagged subinterfaces.

Within a tagged interface, a default subinterface (where the **vlan-id** value is set to **optional**) and an untagged subinterface can be configured. This type of configuration behaves as follows:

- the **vlan-id optional** subinterface captures untagged and non-explicitly configured VLAN-tagged frames
- the untagged subinterface captures untagged and packets with tag 0 as the outermost tag

When **vlan-id optional** and untagged subinterfaces are configured on the same tagged interface, packets for unconfigured VLANs go to the **vlan-id optional** subinterface and tag 0 or untagged packets go to the untagged subinterface.

Classification is based on the following:

- all traffic, for interfaces where VLAN tagging is set to false, regardless of existing VLAN tags
- single outermost tag for tagged interfaces where VLAN tagging is set to true. By default, only Ethertype 0x8100 is considered for tags; other Ethertypes are treated as payload. The default Ethertype used for classification can be modified by the command **tpid [TPID_0X8100|TPID_0X88A8|TPID_0X9100|TPID_0X9200|TPID_ANY]**.

The following ingress and egress VLAN mapping actions are supported by default:

- at ingress, pop the single outermost tag (tagged interfaces) or perform no user-visible action (untagged interfaces)
- at egress, push a specified tag at the top of the stack (tagged interfaces) or perform no user-visible action (untagged interfaces)
- if the **vlan-id** value is set to **optional** or the subinterface uses an **untagged** configuration, no tag is popped at ingress or pushed at egress

There is one exception: on a subinterface that uses an **untagged** configuration, if a received packet has tag 0 as its outermost tag, the subinterface pops tag 0.

For more information on subinterface VLAN configuration, Dot1q tagged bridged subinterfaces, and QinQ bridged subinterfaces, see the *Interfaces* guide.

4.1.2.2 Routed and bridged subinterfaces

SR Linux subinterfaces can be specified as type routed or bridged:

- routed subinterfaces can be assigned to a network instance of type mgmt, default, or ip-vrf
- bridged subinterfaces can be assigned to a network instance of type mac-vrf

Routed subinterfaces allow for configuration of IPv4 and IPv6 settings, and bridged subinterfaces allow for configuration of bridge table and VLAN ingress or egress mapping.

Bridged subinterfaces do not have MTU checks other than the interface-level MTU (port MTU) or the value set with the **l2-mtu** command. The IP MTU is only configurable on routed subinterfaces.

4.1.3 IRB interfaces

IRB interfaces enable inter-subnet forwarding. Network instances of type mac-vrf are associated with a network instance of type ip-vrf via an IRB interface. See [Figure 1: MAC-VRF, IRB interface, and IP-VRF](#) for an illustration of the relationship between MAC-VRF and IP-VRF network instances.

On SR Linux, IRB interfaces are named `irbN`, where N is an integer in the range of 0 to 255. Up to 4095 subinterfaces can be defined under an IRB interface. An IP-VRF network instance can have multiple IRB subinterfaces, while a MAC-VRF network instance can reference only one IRB subinterface.

IRB subinterfaces are type routed. They cannot be configured as type bridged.

IRB subinterfaces operate in the same way as other routed subinterfaces, including support for the following:

- IPv4 and IPv6 access control lists (ACLs)
- static routes and BGP (IPv4 and IPv6 families)
- IP MTU (with the same range of valid values as Ethernet subinterfaces)
- all settings in the `subinterface/ipv4` and `subinterface/ipv6` containers; for IPv6, the IRB subinterface also gets an IPv6 link local address
- BFD
- subinterface statistics

IRB subinterfaces also support QoS across the following platforms:

Table 1: IRB QoS support on various platforms

Platform	IRB QoS support
7220 IXR-Dx	Support for DSCP-based QoS (input and output classifiers and rewrite rules).
7250 IXR 6/10/6e/10/X1b/X3b	No support.
7730 SXR	No support.
7215 IXS-A1	No support.
7220 IXR-Hx	No support.

IRB interfaces do not support sFlow, VLAN tagging, or interface statistics.

4.1.3.1 Using ACLs with IRB interfaces and Layer 2 subinterfaces

The following items apply when using access control lists (ACLs) with an IRB interface or Layer 2 subinterface:

- Input ACLs associated with Layer 2 subinterfaces match all the traffic entering the subinterface, including Layer 2 switched traffic or Layer 3 traffic forwarded to the IRB.
- Input ACLs associated with IRB subinterfaces only match Layer 3 traffic, that is, traffic with a MAC destination address (DA) matching the IRB MAC address.
- The same ACL can be attached to a Layer 2 subinterface and an IRB subinterface in the same service. In this case, there are two ACL instances: one for the IRB with higher priority and another one for the bridged traffic. Routed traffic matches the higher priority instance entries of the ACL.
- The same ACL can be attached to IRB subinterfaces and Layer 2 subinterfaces if both subinterface types belong to different services.

- On 7220 IXR-D1, D2, and D3 systems, egress ACLs, unlike ingress ACLs, cannot match both routed and switched traffic when a Layer 2 IP ACL is attached.
- Received traffic on a MAC-VRF is automatically discarded if the MAC source address (SA) matches the IRB MAC address, unless the MAC is an anycast gateway MAC.
- Packet capture filters show the Layer 2 subinterface for switched traffic and the IRB interface for routed traffic.

4.1.4 Layer 2 services configuration

The examples in this section show how to configure a MAC-VRF network instance, bridged interface, and IRB interface.

4.1.4.1 MAC-VRF network instance configuration example

The following example configures a MAC-VRF network instance and settings for the bridge table. The bridge table is set to a maximum of 500 entries. Learned MAC addresses are aged out of the bridge table after 600 seconds.

MAC duplication detection is configured so that a MAC address is considered a duplicate when its number of detected moves across network-instance interfaces is greater than three over a 5-minute interval. In this example, the MAC address is blackholed. After the hold-down time of 3 minutes, the MAC address is flushed from the bridge table, and the monitoring process for the MAC address is restarted.

The example includes configuration for a static MAC address in the bridge table.

The MAC-VRF network instance is associated with a bridged interface and an IRB interface.

```
--{ candidate shared default }--[ ]--
# info network-instance mac-vrf-1
  network-instance mac-vrf-1 {
    type mac-vrf
    admin-state enable
    description "Sample mac-vrf network instance"
    interface ethernet-1/1.1 {
    }
    interface irb1.1 {
    }
    bridge-table {
      mac-learning {
        admin-state enable
        aging {
          admin-state enable
          age-time 600
        }
      }
    }
    mac-duplication {
      admin-state enable
      monitoring-window 5
      num-moves 3
      hold-down-time 3
      action blackhole
    }
    mac-limit {
      maximum-entries 500
    }
    static-mac {
```

```

        mac 00:00:5E:00:53:FF {
        }
    }
}

```

4.1.4.2 Bridged subinterface configuration example

The following example configures the bridged subinterface that is associated with the MAC-VRF in the previous example.

```

--{ candidate shared default }--[ ]--
# info interface ethernet-1/1
  interface ethernet-1/1 {
    admin-state enable
    subinterface 1 {
      type bridged
      admin-state enable
      vlan {
        encap {
          single-tagged {
            vlan-id 10
          }
        }
      }
    }
  }
}

```

The **vlan-id** value can be configured as a specific valid number or with the keyword **optional**, which means any frame that does not hit the **vlan-id** configured in other subinterfaces of the same interface is classified in this subinterface.

In the following example, the **vlan encap untagged** setting is enabled for subinterface 1. This setting allows untagged frames to be captured on tagged interfaces.

For subinterface 2, the **vlan encap single-tagged vlan-id optional** setting allows non-configured VLAN IDs and untagged traffic to be classified to this subinterface.

With the **vlan encap untagged** setting on one subinterface, and the **vlan encap single-tagged vlan-id optional** setting on the other subinterface, traffic enters the appropriate subinterface; that is, traffic for unconfigured VLANs goes to subinterface 2, and tag0 or untagged traffic goes to subinterface 1.

```

--{ candidate shared default }--[ ]--
# info interface ethernet-1/2
  interface ethernet-1/2 {
    vlan-tagging true
    subinterface 1 {
      type bridged
      vlan {
        encap {
          untagged {
          }
        }
      }
    }
    subinterface 2 {
      type bridged
      vlan {
        encap {

```

```

        single-tagged {
            vlan-id optional
        }
    }
}

```

SR Linux also supports QinQ (or double-tagged) subinterfaces. The following example adds a bridged subinterface that uses inner and outer tags to classify the traffic (in addition to the existing subinterfaces).



Note: QinQ subinterfaces are supported on 7250 IXR and 7730 SXR platforms.

```

--{ * candidate shared default }--[ ]--
A:srl1# info interface ethernet-1/2
interface ethernet-1/2 {
    vlan-tagging true
    subinterface 1 {
        type bridged
        vlan {
            encap {
                untagged {
                }
            }
        }
    }
    subinterface 2 {
        type bridged
        vlan {
            encap {
                single-tagged {
                    vlan-id optional
                }
            }
        }
    }
    subinterface 3 {
        type bridged
        vlan {
            encap {
                double-tagged {
                    inner-vlan-id 10
                    outer-vlan-id 30
                }
            }
        }
    }
}

```

For more information on subinterface VLAN configuration, see the *Interfaces* guide.

4.1.4.3 IRB interface configuration example

The following example configures an IRB interface. The IRB interface is operationally up when its admin-state is enabled, and its IRB subinterfaces are operationally up when associated with MAC-VRF and IP-

VRF network instances. At least one IPv4 or IPv6 address must be configured for the IRB subinterface to be operationally up.

```
--{ candidate shared default }--[ ]--
# info interface irb1
  interface irb1 {
    description IRB_Interface
    admin-state enable
    subinterface 1 {
      admin-state enable
      ipv4 {
        admin-state enable
        address 192.168.1.1/24 {
        }
      }
    }
  }
}
```

4.1.5 Displaying bridge table information

Procedure

You can display information from the bridge table of a MAC-VRF network instance using **show** commands and the **info from state** command.

Example: Display summary information for MAC-VRF network instances

To display a summary of the bridge table contents for the MAC-VRF network instances configured on the system:

```
# show network-instance * bridge-table mac-table summary
-----
Network-Instance Bridge table summary
-----
Name                               : mac-vrf-1
Irb mac                             : 1 Total 1 Active
Static macs                         : 19 Total 19 Active
Duplicate macs                      : 10 Total 10 Active
Learnt macs                         : 15 Total 14 Active
Total macs                          : 45 Total 44 Active
Maximum-Entries                    : 200
Warning Threshold Percentage        : 95% (190)
Clear Warning                       : 90% (180)
-----
Name                               : mac-vrf-2
Irb mac                             : 1 Total 1 Active
Static macs                         : 1 Total 1 Active
Duplicate macs                      : 10 Total 10 Active
Learnt macs                         : 15 Total 14 Active
Total macs                          : 27 Total 26 Active
Maximum-Entries                    : 200
Warning Threshold Percentage        : 95% (190)
Clear Warning                       : 90% (180)
-----
Total Irb macs                     : 2 Total 2 Active
Total Static macs                  : 29 Total 29 Active
Total Duplicate macs               : 20 Total 20 Active
Total Learnt macs                  : 30 Total 29 Active
```

```
Total Macs          : 81 Total 80 Active
-----
```

Example: Display bridge table for a MAC-VRF network instance

To list the contents of the bridge table for a MAC-VRF network instance:

```
# show network-instance mac-vrf-1 bridge-table mac-table all
-----
Mac-table of network instance mac-vrf-1
-----
+-----+-----+-----+-----+-----+-----+-----+
| Mac           | Destination | Dest-Index| Type   | Active | Aging | Last-update |
+=====+=====+=====+=====+=====+=====+=====+
| 00:00:00:00:00:01 | ethernet-1/1.1 | 65      | Learnt | True  | 256  | 2020-2-3T3:37:26 |
| 00:00:00:00:00:02 | irbl.1       | 0       | Irb    | True  | N/A  | 2019-2-1T3:37:26 |
| 00:00:00:00:00:03 | blackhole    | 0       | Duplicate | True  | N/A  | 2019-2-1T3:37:26 |
| 00:00:00:00:00:04 | ethernet-1/1.2 | 66      | Learnt | True  | 256  | 2019-2-1T3:37:26 |
+-----+-----+-----+-----+-----+-----+-----+
Total Irb macs      : 2 Total 2 Active
Total Static macs   : 0 Total 0 Active
Total Duplicate macs : 1 Total 1 Active
Total Learnt macs   : 3 Total 3 Active
Total Macs          : 6 Total 6 Active
-----
```

Example: Display information about a MAC address

To display information about a specific MAC address in the bridge table:

```
# show network-instance * bridge-table mac-table mac 00:00:00:00:00:01
-----
Mac-table of network instance mac-vrf-1
-----
Mac           : 00:00:00:00:00:01
Destination   : ethernet-1/1.1
Destination Index : 65
Type          : Learnt
Programming status : Success | Failed | Pending
Aging         : 250 seconds
Last update    : 2019-12-13T23:37:26.000
Duplicate Detect Time : N/A
Hold-down-time-remaining: N/A
-----
```

Example: Display duplicate MAC addresses

To display the duplicate MAC address entries in the bridge table:

```
# show network-instance * bridge-table mac-duplication duplicate-entries
-----
Mac-duplication in network instance mac-vrf-1
-----
Admin-state      : Enabled
Monitoring window : 3 minutes
Number of moves allowed : 5
Hold-down-time   : 10 seconds
Action           : Stop Learning
-----
Duplicate entries in network instance mac-vrf-1
-----
+-----+-----+-----+-----+-----+-----+-----+
-----
```

Duplicate mac	Destination	Dest-Index	Detect-Time	Hold down time remaining
00:00:00:00:00:01	ethernet-1/1.1	65	2019-12-13T23:37:26.000	6
00:00:00:00:00:02	ethernet-1/1.1	65	2019-12-13T23:37:26.000	6
00:00:00:00:00:03	ethernet-1/1.1	65	2019-12-13T23:37:26.000	6

Total Duplicate macs : 3 Total 3 Active				

Example: Use info from state command to display MAC address entries

You can display the duplicate, learned, or static MAC address entries in the bridge table using **info from state** commands. For example, the following command displays the duplicate MAC entries:

```
# info from state network-instance * bridge-table mac-duplication duplicate-entries mac *
| as
| table
+-----+-----+-----+-----+-----+
| Network-instance | Duplicate-mac | Dest-idx | Detect-time | Hold-down-
| time |
| |
| |
+-----+-----+-----+-----+-----+
=====+
| red | 00:00:00:00:00:01 | 1 | 2019-12-13T23:37:26.000 |
| 10 |
| red | 00:00:00:00:00:02 | 2 | 2019-12-13T23:37:26.000 |
| 20 |
| red | 00:00:00:00:00:03 | 3 | 2019-12-13T23:37:26.000 |
| 90 |
| blue | 00:00:00:00:00:04 | 4 | 2019-12-13T23:37:26.000 |
| 90 |
| blue | 00:00:00:00:00:05 | 0 | 2019-12-13T23:37:26.000 |
| 90 |
+-----+-----+-----+-----+-----+
|-----+
|-----+
```

Example: Display learned MAC addresses

The following command displays the learned MAC entries in the table:

```
# info from state network-instance * bridge-table mac-learning learnt-entries mac * | as
| table
+-----+-----+-----+-----+-----+
| Network-instance | Learnt-mac | Dest | Aging | Last-update |
+-----+-----+-----+-----+-----+
| red | 00:00:00:00:00:01 | ethernet-1/1.1 | 300 | 2019-12-13T23:37:26.000 |
| red | 00:00:00:00:00:02 | ethernet-1/1.1 | 212 | 2019-12-13T23:37:26.000 |
| red | 00:00:00:00:00:03 | ethernet-1/1.2 | 10 | 2019-12-13T23:37:26.000 |
| blue | 00:00:00:00:00:04 | ethernet-1/1.3 | 10 | 2019-12-13T23:37:26.000 |
| blue | 00:00:00:00:00:05 | ethernet-1/1.4 | 20 | 2019-12-13T23:37:26.000 |
+-----+-----+-----+-----+-----+
```

Example: Display static MAC entries

The following command displays the static MAC entries in the table:

```
# info from state network-instance * bridge-table static-mac mac * | as table
+-----+-----+-----+
| Network-instance | Static-mac | Dest |
```

```

+-----+-----+-----+
| red      | 00:00:00:00:00:01 | ethernet-1/1.1 |
| red      | 00:00:00:00:00:02 |          irb1.1 |
| red      | 00:00:00:00:00:03 |          blackhole |
| blue     | 00:00:00:00:00:04 | ethernet-1/1.3 |
| blue     | 00:00:00:00:00:05 | ethernet-1/1.4 |
+-----+-----+-----+

```

4.1.6 Deleting entries from the bridge table

Procedure

The SR Linux includes commands for deleting duplicate or learned MAC entries from the bridge table. For a MAC-VRF or subinterface, you can delete all MAC entries, MAC entries with a blackhole destination, or a specific MAC entry.

Example: Clear blackhole MAC entries

The following example clears MAC entries in the bridge table for a MAC-VRF network instance that have a blackhole destination:

```

--{ candidate shared default }--[ ]--
# tools network-instance mac-vrf-1 bridge-table mac-duplication delete-blackhole-macs

```

Example: Delete a learned MAC address

The following example deletes a specified learned MAC address from the bridge table for a MAC-VRF network instance:

```

--{ candidate shared default }--[ ]--
# tools network-instance mac-vrf-1 bridge-table mac-learning delete-mac 00:00:00:00:00:04

```

Example: Clear duplicate MAC entries

The following example clears all duplicate MAC entries in the bridge table for a subinterface:

```

--{ candidate shared default }--[ ]--
# tools interface ethernet-1/1.1 bridge-table mac-duplication delete-all-macs

```

4.1.7 Storm control on bridged subinterfaces



Note: Storm control is supported on 7220 IXR-D1/D2/D3/D4/D5 and 7215 IXS-A1 platforms.

Storm control allows you to rate limit broadcast, unknown unicast, and multicast (BUM) traffic on bridged subinterfaces. This feature measures the rate for the individual types of BUM traffic, and performs rate limiting based on thresholds configured for each traffic type.

Storm control is configured as an interface-level policer. You set thresholds for each type of BUM traffic. You can configure the threshold as a percentage of interface bandwidth or as a rate in kb/s.

If kbps is specified as the unit type, the rate can be configured as a multiple of 64; for example, 64, 128, 192, and so on. If the rate is configured as any value in the 1 to 127 kb/s range, the effective rate is 64 kb/s.

s, which is shown as the operational-rate in **info from state** output. Similarly, if the rate is configured as any value in the 128 to 191 kb/s range, the operational rate is 128 kb/s, and so on.

The default threshold is 100% of the interface bandwidth for the three traffic types. Specifying a threshold of 0 for a traffic type blocks all traffic of that type on the interface. When a storm control policer is set to 0, the first packet of a flow affected by the policer is forwarded, but subsequent packets are dropped.

Storm control policers can be configured for Ethernet interfaces only. The Ethernet interfaces can be members of a LAG, but storm control must be configured on each interface individually; aggregate rate limiting for the LAG is not supported.

Statistics are not collected for the storm control policers. However, it is possible to check discarded packets at the subinterface level. For example, if storm control for broadcast traffic is configured on interface ethernet-1/1 at a rate of 50%, and broadcast traffic on the ethernet-1/1.1 subinterface exceeds 50% of the port capacity, statistics for the ethernet-1/1.1 subinterface show ingress drops.

Storm control on 7220 IXR-D1/D2/D3/D4/D5 platforms

On 7220 IXR-D1/D2/D3/D4/D5 platforms, storm control supports an action for interfaces when the traffic exceeds any of the configured policers. This action is configured by the command **rising-threshold-action** and the possible actions are:

- **disable-interface**

This action brings down the interface if traffic exceeds the rate in any of the policers. The interface oper-down-reason shows storm-control-action. In order to bring the interface back up, the **admin-state** must be toggled (first disable, then enable).

- **trigger-event**

This action creates an event when a configured rate is exceeded. The event indicates that at least one of the configured storm control rates has been exceeded, but it does not indicate which one.

- **none**

This is the default action.

Storm control on 7215 IXS-A1 platforms

The implementation of storm control on 7215 IXS-A1 platforms differs compared to other platforms.

The configured storm control rates are Layer 1 rates, as opposed to Layer 2 rates. The configured storm control rates include frame preamble and inter-frame gap (IFG) bytes.

For example, on the 7215 IXS-A1, 64-byte frames and a broadcast rate set to 1 000 000 kb/s corresponds to 761 904 kb/s of actual Layer 2 rate: $(1\ 000\ 000 * 64) / (64+20) = 761\ 904$ kb/s.

If storm control is enabled, the in-discarded-packets statistics on the interface are increased when packets are dropped due to the storm control policer. However, the subinterface statistics are not modified.

The interface storm control policers are never applied to TCP SYN packets whether their destination MAC is broadcast, multicast, or unknown unicast.

4.1.7.1 Configuring storm control for an interface

Procedure

To configure storm control, set thresholds for each type of BUM traffic, either as a percentage of interface bandwidth or as a rate in kbps.

Example: Configure storm control for an interface

The following example configures storm control for an interface on 7220 IXR-D1/D2/D3/D4/D5 and 7215 IXS-A1 platforms. The unit type is specified as kbps. Individual thresholds are set for each traffic type.

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1
  interface ethernet-1/1 {
    admin-state enable
    ethernet {
      storm-control {
        units kbps
        broadcast-rate 128
        multicast-rate 192
        unknown-unicast-rate 225
      }
    }
  }
}
```

Example: Configure storm control with an action

The following example configures storm control with an action on 7220 IXR-D1/D2/D3/D4/D5 platforms. When the amount of broadcast traffic in the interface exceeds 40% of the interface capacity, the interface is disabled.

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1
  interface ethernet-1/1 {
    admin-state enable
    ethernet {
      storm-control {
        units percentage
        broadcast-rate 40
        rising-threshold-action disable-interface
      }
    }
  }
}
```

4.1.7.2 Displaying storm control information

Procedure

Display the configured rates for an interface using **info from state** and **show interface** commands.

Example: Display configured rates using info from state command

```
--{ * candidate shared default }--[ ]--
# info from state interface ethernet-1/1 ethernet storm-control
  interface ethernet-1/1 {
    ethernet {
      storm-control {
        units kbps
        broadcast-rate 128
        multicast-rate 172
        unknown-unicast-rate 225
        operational broadcast-rate 128
        operational multicast-rate 192
        operational unknown-unicast-rate 192
      }
    }
  }
}
```

```

    }
  }
}

```

Example: Display configured rates using show interface command

```

--{ * candidate shared default }--[ ]--
# show interface ethernet-1/1 detail
=====
Interface: ethernet-1/1
-----
Description      : dut2
Oper state       : up
Down reason      : N/A
Last change      : 9h21m16s ago, 1 flaps since last clear
Speed            : 100G
Flow control     : Rx is disabled, Tx is disabled
Loopback mode    : false
MTU              : 9232
VLAN tagging     : true
Queues           : 8 output queues supported, 0 used since the last clear
MAC address      : 00:00:5e:00:53:00
Last stats clear: never
...
-----
Storm control for ethernet-1/1
-----
Broadcast Rate (kbps)           : 128
Multicast Rate (kbps)           : 172
Unknown Unicast (kbps)         : 225
Operational Broadcast Rate (kbps) : 128
Operational Multicast Rate (kbps) : 192
Operational Unknown Unicast Rate (kbps) : 192

```

4.1.8 L2CP transparency

Layer 2 Control Protocol (L2CP) transparency refers to the ability to control whether L2CP frames are tunneled across a carrier Ethernet network. Tunneling in this context means injecting the L2CP frames into the regular Layer 2 datapath, as opposed to trapping them to the CPU or discarding them.

By default, SR Linux handles L2CP frames according to the default transparency action listed in the following table.

Table 2: Default L2CP PDU actions for tagged and untagged frames

Layer 2 control protocol	L2CP destination address	Protocol identifier	Default transparency action (7220 IXR-Dx)	Default transparency action (7730 SXR)
LLDP	01-80-c2-00-00-0e Only the above MAC address is identified as LLDP; other LLDP destination MAC addresses are not.	Ethertype: 0x88cc	Trap to CPU only if untagged, drop otherwise	Tunnel if tagged, trap to CPU if untagged

Layer 2 control protocol	L2CP destination address	Protocol identifier	Default transparency action (7220 IXR-Dx)	Default transparency action (7730 SXR)
LACP	01-80-c2-00-00-02	Ethertype: 0x8809 Subtype: 0x01	Trap to CPU only if untagged, drop otherwise	Tunnel if tagged, trap to CPU if untagged
xSTP (STP/RSTP/MSTP)	01-80-c2-00-00-00	Ethertype: any	Drop	Tunnel if tagged, drop if untagged
PAUSE	01-80-c2-00-00-01	Ethertype: 0x8808	Drop	Drop
dot1x	01-80-c2-00-00-03	Ethertype: 0x888e	Trap to CPU only if untagged, drop otherwise	Tunnel if tagged, trap to CPU if untagged
PTP	01-80-c2-00-00-0e	Ethertype: 0x88f7	Drop	Trap to CPU only if untagged, drop otherwise (sync platforms) Tunnel if tagged, drop if untagged (non-sync platforms)
E-LMI	01-80-c2-00-00-07	Ethertype: 0x88ee	Drop	Tunnel if tagged, drop if untagged
EFM-OAM	01-80-c2-00-00-02	Ethertype: 0x8809 Subtype: 0x03		Tunnel if tagged, drop if untagged
LACP Marker-Protocol/Link OAM 802.3ah/ESMC	01-80-c2-00-00-02	Ethertype: 0x8809 Subtype: 0x0a	Drop	Tunnel if tagged, trap to CPU if untagged (sync platforms) Tunnel if tagged, drop if untagged (non-sync platforms)

The L2CP transparency feature allows you to configure SR Linux to tunnel the following types of L2CP frames instead of performing the default transparency action, or you can configure SR Linux to tunnel all L2CP frames.

- LLDP
- LACP
- xSTP (STP/RSTP/MSTP)
- dot1x
- PTP
- EFM-OAM (7730 SXR only)
- E-LMI (7730 SXR only)
- ESMC (7730 SXR only)

When tunneling is enabled, the following take place:

- Tagged and untagged L2CP frames are classified into a bridged subinterface based on their dot1q tag VLAN ID. If the L2CP frames are classified for forwarding into a MAC-VRF, they are flooded based on their MAC DA. If no bridged subinterface matches the incoming untagged or VLAN-ID tagged frames, they are discarded.
- L2CP transparency is supported on Ethernet interfaces only, with the processing and tunneling possible on bridged subinterfaces of a MAC-VRF. The system does not apply L2CP transparency rules to VXLAN tunnel interfaces, however it does apply them to Layer 3 subinterfaces in the default network instance.
- When an L2CP PDU is tunneled, the source MAC is learned. When the PDU is dropped, the source MAC is not learned.

On 7220 IXR Dx platforms, to tunnel L2CP frames end-to-end between two EVPN-VXLAN leaf nodes, you must enable L2CP transparency on the ingress Layer 3 subinterfaces of the default network instance on the egress leaf node. Otherwise, even if L2CP transparency for a protocol is configured on the bridged subinterfaces of the ingress leaf, the system discards the L2CP frames at the ingress VXLAN tunnel of the egress leaf.

L2CP transparency is supported in configurations where tunneling for LLDP frames is used in the overlay (between PE and CE devices), and LLDP is used in the underlay (between spine and leaf nodes).

L2CP transparency is not supported on interfaces configured in breakout mode. For LAG interfaces, you must configure L2CP transparency on all member interfaces.

4.1.8.1 Configuring L2CP transparency

Procedure

You can enable tunneling for all L2CP protocol types listed in [Table 2: Default L2CP PDU actions for tagged and untagged frames](#) or you can enable tunneling individually for the following L2CP protocol types: LLDP, LACP xSTP (STP/RSTP/MSTP), dot1x, and PTP.

Example: Enable tunneling for all L2CP protocols

```
--{ candidate shared default }--[ ]--
# info interface ethernet-1/1 ethernet l2cp-transparency
  interface ethernet-1/1 {
    ethernet {
      l2cp-transparency {
        tunnel-all-l2cp true
      }
    }
  }
}
```

When **tunnel-all-l2cp true** is configured, all L2CP frames coming into the interface, identified by MAC DA = 01:80:c2:00:00:0x or MAC DA = 01:80:c2:00:00:2x where x is any hex value, are tunneled.

The **tunnel-all-l2cp** command is not supported on interfaces where LLDP or LACP are enabled.

Example: Disable tunneling for all L2CP protocols

```
--{ candidate shared default }--[ ]--
# info interface ethernet-1/1 ethernet l2cp-transparency
  interface ethernet-1/1 {
```

```

    ethernet {
      l2cp-transparency {
        tunnel-all-l2cp false
      }
    }
  }
}

```

When **tunnel-all-l2cp false** is configured, all L2CP frames not specifically enabled for tunneling are discarded.

Example: Enable tunneling for individual protocol types

You can enable tunneling individually for LLDP, LACP xSTP (STP/RSTP/MSTP), dot1x, and PTP L2CP protocol types. These L2CP frames are identified by the destination MAC address, Ethertype and sometimes slow-protocol subtype; see [Table 2: Default L2CP PDU actions for tagged and untagged frames](#).

The following example enables tunneling for LLDP frames:

```

# info interface ethernet-1/1 ethernet l2cp-transparency
interface ethernet-1/1 {
  ethernet {
    l2cp-transparency {
      lldp {
        tunnel true
      }
    }
  }
}

```

By default, all the **tunnel** commands are set to **false**, so all L2CP protocols are discarded, except for LLDP and LACP, for which a copy is trapped to the CPU (if untagged) and another copy is discarded.



Note: Only LLDP frames with MAC 01-80-c2-00-00-0e are tunneled. LLDP frames with MAC DA = 01-80-c2-00-00-00 or 01-80-c2-00-00-03 are not tunneled or trapped to the CPU.

Tunneling for LLDP frames is not supported on an interface where LLDP is enabled.

Tunneling for LACP frames is not supported on an interface where `aggregate-id` is configured or the interface is a member of a LAG where LACP is enabled.

4.1.8.2 Displaying L2CP transparency information

Procedure

Each of the L2CP protocols that can be tunneled is associated with an L2CP transparency rule (`oper-rule`) that indicates the state of the actual rule installed in the datapath. This rule can be any of the following:

- `trap-to-cpu-untagged`

This rule is used for LLDP and LACP frames if they are not configured to be tunneled.

On 7220 Dx platforms, untagged frames are trapped to the CPU; that is, a copy is trapped and sent to the CPM, while another copy is discarded. Tagged frames are discarded and not trapped. On 7730 SXR platforms, tagged frames are tunneled, with the exception of PAUSE frames, which are dropped.

- `drop-tagged-and-untagged`

This rule is used for xSTP, dot1x, and PTP frames if they are not configured for tunneling.

- `tunnel-tagged-and-untagged`
This rule is used for any of the L2CP protocols when they are configured to be tunneled.
- `tunnel-tagged-drop-untagged`
This rule is used on 7730 SXR platforms for some of the L2CP protocols when they are configured to not be tunneled.
- `tunnel-tagged-trap-to-cpu-untagged`
This rule is used on 7730 SXR platforms for some of the L2CP protocols when they are not configured to be tunneled.

You can display the L2CP transparency configuration and `oper-rule` for each L2CP protocol that can be tunneled.

Example: Display tunneling configuration and `oper-rule` for an interface on 7220 IXR Dx platforms

Use the **info from state** command for an interface to display its tunnelling configuration and `oper-rule` for each protocol. For example:

```
--{ candidate shared default }--[ ]--
# info from state interface ethernet-1/1 ethernet l2cp-transparency
interface ethernet-1/1 {
  ethernet {
    l2cp-transparency {
      tunnel-all-l2cp false
      lldp {
        tunnel false
        oper-rule trap-to-cpu-untagged
      }
      lacp {
        tunnel false
        oper-rule trap-to-cpu-untagged
      }
      xstp {
        tunnel false
        oper-rule drop-tagged-and-untagged
      }
      dot1x {
        tunnel false
        oper-rule drop-tagged-and-untagged
      }
      ptp {
        tunnel false
        oper-rule drop-tagged-and-untagged
      }
    }
  }
}
```

Example: Display the `oper-rule` for each protocol type on 7220 IXR Dx platforms

Use the **show interface detail** command to display the L2CP transparency rule (`oper-rule`) for each protocol type; for example:

```
--{ candidate shared default }--[ ]--
# show interface ethernet-1/1 detail
=====
Interface: ethernet-1/1
```

```

-----
Description      : <None>
Oper state       : up
Down reason      : N/A
Last change      : 2d22h27m45s ago, 1 flaps since last clear
Speed            : 10G
Flow control     : Rx is disabled, Tx is disabled
MTU              : 9232
VLAN tagging     : false
Queues           : 8 output queues supported, 2 used since the last clear
MAC address      : 00:01:01:FF:00:01
Last stats clear: never
Breakout mode    : false
-----
L2cp transparency rules for ethernet-1/1
-----
Lldp              : trap-to-cpu-untagged
Lacp              : trap-to-cpu-untagged
xStp              : drop-tagged-and-untagged
Dot1x             : drop-tagged-and-untagged
Ptp               : tunnel-tagged-and-untagged
Non-specified L2cp : drop-tagged-and-untagged
-----

```

4.1.8.3 Displaying L2CP transparency statistics

Procedure

To display L2CP transparency statistics, use the **info from state** command in candidate or running mode, or the **info** command in state mode.

Statistics are displayed for the following:

- total number of L2CP packets of any L2CP protocol on all interfaces, as well as the total number of discarded and tunneled L2CP packets
- total number of ingress tunneled and trapped-to-CPU packets per L2CP protocol at the system level. Individual statistics are displayed for LLDP, LACP, xSTP, dot1x, and PTP protocols if `tunnel-all-l2cp false` is configured

Example: Display L2CP transparency statistics for 7220 IXR Dx platforms

```

--{ candidate shared default }--[ ]--
# info from state system l2cp-transparency l2cp-statistics
system {
  l2cp-transparency {
    l2cp-statistics {
      total-in-packets 0
      total-in-discarded-packets 0
      total-in-tunneled-packets 0
      total-in-trap-to-cpu-packets 0
      last-clear "a day ago"
      lldp {
        in-tunneled-packets 0
        in-trap-to-cpu-packets 0
      }
      lacp {
        in-tunneled-packets 0
        in-trap-to-cpu-packets 0
      }
      xstp {

```

```

        in-tunneled-packets 0
        in-trap-to-cpu-packets 0
    }
    dot1x {
        in-tunneled-packets 0
        in-trap-to-cpu-packets 0
    }
    ptp {
        in-tunneled-packets 0
        in-trap-to-cpu-packets 0
    }
}
}
}
}

```

Example: Display L2CP transparency statistics for 7730 SXR platforms

```

--{ candidate shared default }--[ ]--
# info from state system l2cp-transparency l2cp-statistics--{ candidate shared default }--
[ ]--
system {
    l2cp-transparency {
        l2cp-statistics {
            total-in-packets 4558
            total-in-discarded-packets 0
            total-in-tunneled-packets 0
            total-in-trap-to-cpu-packets 4558
            lldp {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 4558
            }
            lacp {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            xstp {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            dot1x {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            ptp {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            esmc {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            elmi {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
            efm-oam {
                in-tunneled-packets 0
                in-trap-to-cpu-packets 0
            }
        }
    }
}
}
}
}

```

**Note:**

- Discarded L2CP frames are counted as `in-discarded-packets` in interface-level statistics. They are not counted in subinterface statistics.
- Tunneled L2CP frames are counted in interface and subinterface statistics for outgoing interfaces in the appropriate counters.
- If **`tunnel-all-l2cp true`** is configured, all L2CP frames (including LLDP, LACP, xSTP, dot1x, and PTP) are counted only in `l2cp-statistics/total-in-tunneled-packets` regardless of protocol-specific configuration. In this case, the protocol-specific statistics are displayed as 0.

4.1.8.3.1 Clearing L2CP transparency statistics

Procedure

You can use a **tools** command to clear the total counters and the per-protocol counters for L2CP transparency statistics.

Example: Clear statistics counters for all L2CP protocols

```
--{ running }--[ ]--
# tools system l2cp-transparency l2cp-total-statistics clear
```

Example: Clear statistics counters for a specific L2CP protocol type

```
--{ running }--[ ]--
# tools system l2cp-transparency ptp clear
```

4.1.9 Split horizon groups on MAC-VRF



Note: Split horizon groups are supported on 7730 SXR, 7220 IXR-Dx, and 7215 IXS-A1 platforms. See [Platform specifications for split horizon groups](#) for more information.

By default, in a MAC-VRF network instance, all EVPN destinations are considered part of the same split horizon group (SHG) and every interface and pseudowire is part of a different SHG. With that default configuration, traffic from a subinterface or pseudowire can be forwarded to other subinterfaces or pseudowires and EVPN destinations of the same MAC-VRF. However, traffic from EVPN destinations can only be forwarded to subinterfaces and pseudowires, but not to other EVPN destinations.

The explicit configuration of SHGs in a MAC-VRF network instance modifies this default behavior. SHGs can be configured and associated with bridged subinterfaces, BGP-EVPN instances, and pseudowires.

SHGs are typically used to prevent client-to-client communication while allowing client-to-server communication. SHGs are also used for the seamless integration of pseudowires and EVPN (RFC 8560). See [EVPN-MPLS and VPLS seamless integration \(RFC 8560\)](#) and [MPLS transport for EVPN services](#).

With explicit SHGs, traffic (unicast and BUM) cannot be forwarded between objects that are part of the same SHG.

For control plane traffic, incoming traffic on object A cannot be reinjected by the CPM on object B, if both object A and B are configured with the same SHG. An example of this behavior is proxy-ARP and two Layer 2 subinterfaces configured with the same SHG and attached to the same MAC-VRF. An ARP

request received on one of the two subinterfaces cannot be reinjected into the other subinterface if the proxy-ARP lookup fails to find a successful entry.

4.1.9.1 Configuring split horizon groups

Procedure

Before you can reference them in any object, you must first manually create SHGs using the following command: **network-instance bridge-table split-horizon-group *name***

When created, SHGs are only used in the context of the MAC-VRF.

You can reference SHGs in the following contexts:

- network instance interfaces

network-instance interface bridge-table split-horizon-group *name*

- BGP-EVPN MPLS

network-instance protocols bgp-evpn bgp-instance mpls bridge-table split-horizon-group *name*

- connection points

network-instance connection-point bridge-table split-horizon-group *name*

4.1.9.2 Split horizon group configuration example

The following example illustrates the use of SHGs.

Interfaces **ethernet-1/1.2** and **ethernet-1/1.3** cannot forward traffic between each other, but can send traffic to the other interfaces and EVPN destinations.

Interfaces **ethernet-1/2.1** and **ethernet-1/2.2** cannot forward traffic between each other, but can send traffic to the other interfaces and EVPN destinations.

Interface **ethernet-1/2.3** can forward traffic normally to any other interface or EVPN destination.

Example: Configuring split horizon groups

```
--{ * candidate shared default }--[ network-instance * ]--
A:leaf1# info
  network-instance MAC-VRF-2 {
    interface ethernet-1/1.2 {
      bridge-table {
        split-horizon-group SHG-1
      }
    }
    interface ethernet-1/1.3 {
      bridge-table {
        split-horizon-group SHG-1
      }
    }
    interface ethernet-1/2.1 {
      bridge-table {
        split-horizon-group SHG-2
      }
    }
    interface ethernet-1/2.2 {
      bridge-table {
        split-horizon-group SHG-2
      }
    }
  }
```

```

    }
  }
  interface ethernet-1/2.3 {
  }
  vxlan-interface vxlan0.1 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        encapsulation-type vxlan
        vxlan-interface vxlan0.1
        evi 2
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
  bridge-table {
    split-horizon-group SHG-1 {
    }
    split-horizon-group SHG-2 {
    }
  }
}

```

4.1.9.3 Platform specifications for split horizon groups

SHGs are supported on 7730 SXR, 7220 IXR-Dx, and 7215 IXS-A1 platforms.

The following limitations apply depending on the platform.

7220 IXR-D1/D2/D2L/D3/D3L platform support

Six SHGs per MAC-VRF are supported on 7220 IXR-D1/D2/D2L/D3/D3L platforms. When used with EVPN in the same MAC-VRF, five SHGs can be configured.

7220 IXR-D4/D5 platform support

Two SHGs per MAC-VRF are supported on 7220 IXR-D4/D5 platforms.

7215 IXS-A1 platform support

For expected SHG behavior, all bridged subinterfaces in a MAC-VRF are assigned to the SHG. A bridged subinterface in an SHG drops (at egress) all packets coming from another bridged subinterface, whether or not the origin is a subinterface in the same SHG.

SHGs on 7215 IXR-A1 platforms have the following capabilities:

- In MAC-VRFs without IRB, if all subinterfaces are applied to an SHG there is no connectivity among them. If not all subinterfaces are associated with the SHG, traffic from SHG subinterfaces to non-SHG subinterfaces is forwarded, but traffic from non-SHG subinterfaces to SHG subinterfaces is dropped.
- In MAC-VRFs with IRB, all subinterfaces of the MAC-VRF must be configured on the SHG. Traffic among subinterfaces is dropped, but traffic to and from IRB is forwarded correctly.

4.1.10 Pseudowires



Note: Pseudowires are currently supported exclusively on the 7730 SXR platform.

Static and Targeted Label Distribution Protocol (T-LDP) based pseudowires (PWs) are supported on SR Linux. Pseudowires can be used to establish Layer 2 point-to-point services between two PEs (RFC 4448), stretch broadcast domains in an MPLS network, or set up Virtual Private LAN Services (VPLS), as per RFC 4762. Pseudowires can also be seamlessly integrated with EVPN-MPLS (RFC 8560). All these use cases are described in the following sections.

4.1.10.1 Pseudowires in MAC-VRF network instances

SR Linux supports pseudowires in MAC-VRF network instances. The configuration model requires a **connection-point**, under which a single pseudowire (static or TLDP based) can be configured. Multiple connection-points (therefore multiple pseudowires) are supported per MAC-VRF. The following considerations about the use of pseudowires apply:

- Pseudowires are supported in the same MAC-VRF where bridged subinterfaces and EVPN-MPLS are configured.
- Pseudowires use MPLS tunnels for transport, where the binding between the pseudowire and the transport tunnel is completed by configuring of the **pw-tunnel** under the pseudowire. The **pw-tunnel** uses the MPLS tunnels in the tunnel table, unless the type is filtered by the **allowed-tunnel-types** command.
- The forwarding rules for pseudowires depend on the association of **split-horizon-groups** to those pseudowires and the rest of the objects in the network instance. For example, to emulate an RFC 4762 compliant VPLS service, all the **connection-points** in the MAC-VRF must be associated with the same **split-horizon-group**.
- Pseudowires support all bridge table parameters that are also applicable to bridged subinterfaces, including **mac-duplication**, **mac-limit**, **mac-learning** and other parameters.
- The generation of TLDP MAC withdraw messages to flush MAC addresses on the remote nodes is controlled by the **bridge-table tldp-mac-flush send-flush-on-failure** command. When enabled, a fault on a subinterface, another pseudowire, or EVPN will trigger a flush-all-from-me message. A flush-all-from-me message is a TLDP-based MAC Withdraw message, by which a PE requests a remote PE to remove all MAC addresses associated with the PE transmitting the message.

Pseudowire configuration characteristics

Pseudowires have an administrative and operational state. Pseudowires can also be operationally down due to the following **oper-down-reason** flags:

- `admin-disabled`
- `network-instance-oper-down`
- `no-ingress-vc-label`
- `no-egress-vc-label`
- `network-instance-mtu-mismatch`
- `pseudowire-remote-system-fault-status-bits`

- `transport-tunnel-oper-down`
- `no-destination-id`
- `evpn-route-conflict`



Note: The `evpn-route-conflict` flag is used for EVPN-VPLS seamless integration to handle conflicts when there is a pseudowire and an EVPN Inclusive Multicast Ethernet Tag (IMET) route to the same far end IP address (RFC 8560).

Virtual circuit type configuration, with **ethernet** or **vlan** options determines the **vc-type** signaling for the pseudowire. The VLAN tag value that is pushed on the frames depends on the VLAN configuration on the ingress subinterface and the type does not push any extra VLAN tag on pseudowires.

Control word configuration determines whether control word support is signaled in the control plane. If the peer signals control word support too, the pseudowire is established, and the control word is pushed on all frames. If there is a mismatch between the local router and the peer, the pseudowire will remain operationally down. The router will show a control-word-mismatch oper - down reason at the **ldp service-fecs fec128** level. A change in the control word configuration that causes a mismatch will trigger the TLDP withdraw and re-advertisement with the control word set. It does not cause a label release locally, but the pseudowire will bounce.

Flow label configuration determines whether the Flow Aware Transport (FAT) label is used on pseudowires. FAT labels allow for granular load balancing without deep packet inspection, improving the efficiency of traffic flow management. The capability to use the FAT label is signaled in TLDP. Contrary to the control word, discrepancies in the signaling of the FAT label capability between two TLDP peers do not bring down the pseudowire, but they determine whether the flow label is sent on the wire.

SR Linux supports the signaling and processing of the following pseudowire status bits:

- `pseudowire-forwarding`
- `pseudowire-not-forwarding`
- `local-attachment-circuit-ingress-fault`
- `local-attachment-circuit-egress-fault`
- `provider-service-network-ingress-fault`
- `provider-service-network-egress-fault`
- `pseudowire-forwarding-standby`
- `pseudowire-request-switchover`

Pseudowires do not have a specific Layer 2 MTU value that can be configured, and the data path layer does not enforce MTU or MRU values other than **subinterface mpls-mtu** in the default network instance. However, TLDP signals the Layer 2 MTU as follows:

The default value signaled for a pseudowire by TLDP is taken from the **oper-mac-vrf-mtu**. This default value can be overridden by the value configured with the command **advertise-l2-mtu**. In case there is a mismatch with the MTU signaled by the peer, the pseudowire will go operationally down with the reason `network-instance-mtu-mismatch`, unless **ignore-mtu-mismatch true** is configured.

The **oper-mac-vrf-mtu**, signaled by TLDP, is computed as the minimum value among all Layer 2 MTU of all the subinterfaces in the MAC-VRF (irrespective of their operational state), with a default value of the **system mtu default-l2-mtu** in case there are no subinterfaces in the MAC-VRF.

4.1.10.1.1 Configuring pseudowires in MAC-VRF network instances

The following examples show a MAC-VRF configuration with a subinterface and three pseudowires.

Example: Assigning a label block

A label block must be assigned to the network instances so that labels can be allocated for each pseudowire.

```
--{ candidate shared default }--[ system mpls services network-instance ]--
A:srl-pe4# info
    dynamic-label-block range-5-services
```

Example: Configuring pseudowire tunnels to each PE

A **pw-tunnel** to each remote PE is then configured. The **pw-tunnel** requires the configuration of a **remote-system** IP address and the selection of the tunnel types that are used at the transport level, via the **allowed-tunnel-types** command. If there are multiple tunnels (of the allowed type) that can resolve the remote system IP address, the best tunnel in the tunnel table is selected (based on preference and metrics). The configuration of a **pw-tunnel** automatically creates a TLDP session to the remote system IP address for the setup of the pseudowire.

```
--{ candidate shared default }--[ tunnel pseudowire-tunnel ]--
A:srl-pe4# info
    tunnel to-pe1 {
        remote-system 10.0.0.1
        allowed-tunnel-types [
            ldp
        ]
    }
    tunnel to-pe2 {
        remote-system 10.0.0.2
        allowed-tunnel-types [
            bgp
            ldp
            sr-isis
        ]
    }
    tunnel to-pe3 {
        remote-system 10.0.0.3
        allowed-tunnel-types [
            ldp
        ]
    }
--{ candidate shared default }--[ tunnel pseudowire-tunnel ]--
A:srl-pe4# info from state tunnel to-pe*
    tunnel to-pe1 {
        remote-system 10.0.0.1
        operational-tunnel-type ldp
        operational-tunnel-id 65539
        allowed-tunnel-types [
            ldp
        ]
    }
    tunnel to-pe2 {
        remote-system 10.0.0.2
        operational-tunnel-type ldp
        operational-tunnel-id 65541
        allowed-tunnel-types [
            bgp
```

```

        ldp
        sr-isis
    ]
}
tunnel to-pe3 {
    remote-system 10.0.0.3
    operational-tunnel-type ldp
    operational-tunnel-id 65543
    allowed-tunnel-types [
        ldp
    ]
}

```

Example: Configuring pseudowires in the MAC-VRF

The pseudowires are then configured in the MAC-VRF. Each pseudowire is configured under a different **connection-point**. A **connection-point** is a logical structure that can contain multiple objects, but only one is active. Only one pseudowire can be configured per **connection-point**.

The association of **connection-point** and **split-horizon-group** determines how forwarding is carried out between objects in the MAC-VRF. The following example illustrates a typical RFC 4762 VPLS configuration, where all the pseudowires of the MAC-VRF belong to the same "mesh" or split-horizon-group. This configuration ensures that traffic from a pseudowire can never be forwarded to another pseudowire in the MAC-VRF and only to the subinterface. Should you want traffic to be forwarded between two pseudowires (typical RFC 4762 H-VPLS architectures), those two pseudowires must be part of different split horizon groups.

Bridge table properties are supported for pseudowires, such as **discard-unknown-src-mac**, **mac-limit** settings, **mac-learning** settings, or **mac-duplication**. These properties are configured and operated under the connection point to which the pseudowire belongs.

```

--{ candidate shared default }--[ network-instance mac-vrf-5 ]--
A:srl-pe4# info
  type mac-vrf
  interface ethernet-1/1.5 {
  }
  bridge-table {
    split-horizon-group shg-1 {
    }
  }
  connection-point A {
    bridge-table {
      split-horizon-group shg-1
    }
    pseudowire pw-to-3 {
      admin-state enable
      pw-tunnel to-pe3
      signaling {
        virtual-circuit-identifier 5
        tldp {
        }
      }
    }
  }
  connection-point B {
    bridge-table {
      split-horizon-group shg-1
    }
    pseudowire pw-to-1 {
      admin-state enable
      pw-tunnel to-pe1
    }
  }

```

```

        signaling {
            virtual-circuit-identifier 5
            tldp {
            }
        }
    }
}
connection-point C {
    bridge-table {
        split-horizon-group shg-1
    }
    pseudowire pw-to-2 {
        admin-state enable
        pw-tunnel to-pe2
        signaling {
            virtual-circuit-identifier 5
            tldp {
            }
        }
    }
}
}

```

Example: Displaying pseudowire state information

The following shows the state of one of the pseudowires configured. The output shows the operational values and labels.

```

--{ candidate shared default }--[ network-instance mac-vrf-5 connection-point A pseudowire
pw-to-3 ]--
A:srl-pe4# info from state
admin-state enable
oper-state up
index 1
destination-index 373975667326
pw-tunnel to-pe3
control-word false
flow-label false
flow-label-oper-state down
signaling {
    virtual-circuit-identifier 5
    tldp {
        virtual-circuit-type ethernet
        ignore-mtu-mismatch false
    }
}
local {
    operational-ingress-vc-label 1007
}
remote {
    operational-egress-vc-label 1009
}

```

Example: Displaying fec-128 bindings state information

The following shows the state of the corresponding fec-128 bindings.

```

--{ candidate shared default }--[ network-instance mac-vrf-5 connection-point A pseudowire
pw-to-3 ]--
A:srl-pe4# info from state /network-instance default protocols ldp ipv4 bindings service-
fec128 ethernet virtual-circuit-identifier 5 peer-lsr-id 10.0.0.3
binding-oper-state up
advertised {

```

```

label 1007
l2-mtu 1500
control-word false
flow-aware-transport-label-transmit-capability false
flow-aware-transport-label-receive-capability false
pw-status true
withdraw-reason none
label-status [
  in-use-pop
]
signaling-status [
  pseudowire-forwarding
]
}
received {
label 1009
l2-mtu 1500
control-word false
flow-aware-transport-label-transmit-capability false
flow-aware-transport-label-receive-capability false
pw-status true
label-status [
  in-use-push
]
signaling-status [
  pseudowire-forwarding
]
}
}

```

4.1.10.2 EVPN-MPLS and VPLS seamless integration (RFC 8560)

MAC-VRF network instances support seamless integration of pseudowires, interfaces, and EVPN-MPLS destinations by configuring manual **split-horizon-group** objects.

If a **split-horizon-group shg-1** is configured in a MAC-VRF and **interface ethernet-1/1.1**, **connection-point conn-point-1**, and **bgp-evpn bgp-instance mpls** are associated with **shg1**, the following would be true:

- Any traffic among the three objects is discarded.
- Any MAC addresses learned on **ethernet-1/1.1** are not advertised by EVPN.
- Any MAC addresses learned on **connection-point conn-point-1** are not advertised by EVPN.
- If the pseudowire that belongs to **connection-point conn-point-1** and an EVPN multicast destination have the same far-end IP address, the pseudowire will be **oper-state down** with oper-down-reason evpn-conflict.

This behavior allows the seamless integration of VPLS and EVPN-MPLS networks (RFC 4762). In a group of PEs attached to the same MAC-VRF broadcast domain, some PEs can be configured with only pseudowires and subinterfaces in the MAC-VRF (referred to as VPLS PEs), and other PEs can be hybrid, supporting pseudowires to the VPLS PEs and EVPN destinations to other hybrid PEs. On the hybrid PEs, the pseudowires and EVPN destinations can be placed into the same "mesh" by configuring the same **split-horizon-group** under the connection points and EVPN. The following example shows this configuration.

```

--{ candidate shared default }--[ network-instance mac-vrf-5 ]--
A:srl-pe4# info
  type mac-vrf
  interface ethernet-1/1.5 {

```

```

}
protocols {
  bgp-evpn {
    bgp-instance 1 {
      encapsulation-type mpls
      evi 5
      mpls {
        bridge-table {
          split-horizon-group shg-1
        }
        next-hop-resolution {
          allowed-tunnel-types [
            bgp
            ldp
            sr-isis
          ]
        }
      }
    }
  }
  bgp-vpn {
    bgp-instance 1 {
    }
  }
}
bridge-table {
  split-horizon-group shg-1 {
  }
}
connection-point A {
  bridge-table {
    split-horizon-group shg-1
  }
  pseudowire pw-to-3 {
    admin-state enable
    pw-tunnel to-pe3
    signaling {
      virtual-circuit-identifier 5
      tldp {
      }
    }
  }
}
connection-point B {
  bridge-table {
    split-horizon-group shg-1
  }
  pseudowire pw-to-1 {
    admin-state enable
    pw-tunnel to-pe1
    signaling {
      virtual-circuit-identifier 5
      tldp {
      }
    }
  }
}
connection-point C {
  bridge-table {
    split-horizon-group shg-1
  }
  pseudowire pw-to-2 {
    admin-state enable
    pw-tunnel to-pe2
  }
}

```

```

        signaling {
            virtual-circuit-identifier 5
            tldp {
            }
        }
    }
}

```

By configuring all connection points and BGP-EVPN with the same **split-horizon-group**, traffic from a pseudowire is not forwarded to EVPN, and vice versa. Also, MACs learned over one of those pseudowires will not be advertised in EVPN to avoid duplication. A pseudowire is brought down if its remote system IP matches a received IMET route's next hop, to avoid BUM duplication between hybrid PEs.

4.1.10.3 Pseudowires in VPWS network instances

Virtual Private Wire Service (VPWS) is a type of network instance used for Layer 2 point-to-point connectivity. VPWS network instances support two connection points, where each connection point contains a single object.

The supported combinations in the same VPWS network instance are:

- EVPN-MPLS and a subinterface
- Two subinterfaces
- A pseudowire and a subinterface

A single pseudowire and a single bridged subinterface are supported in the same VPWS network instance as follows:

- The pseudowire can be added to a VPWS **connection-point**, where a bridge subinterface exists in the other **connection-point**.
- The configuration and state model for the pseudowire in a VPWS network instance follows the model described in [Pseudowires in MAC-VRF network instances](#) except for the **bridge-table** context, which is not supported when the **connection-point** is added to a VPWS network instance.

4.1.10.3.1 Configuring pseudowires in VPWS network instances

The process for creating a pseudowire in a VPWS network instance is the same as creating a pseudowire in a MAC-VRF:

1. Configure the **pw-tunnel** with the same support as for pseudowires in MAC-VRF.
2. At the network instance **type vpws** level, configure the **connection-point** (without bridge table aspects).
3. Configure the pseudowire (with signaling static or TLDP) under the configured **connection-point**.

The following example shows the configuration of a pseudowire in a VPWS network instance.

```

--{ * candidate shared default }--[ network-instance VPWS-1 ]--
A:srl# info
  type vpws
  description "VPWS-1 pw and subinterface example"
  interface ethernet-1/5.1 {
    connection-point A
  }
  connection-point A {

```



```

}
connection-point B {
  pseudowire pw-to-4 {
    admin-state enable
    pw-tunnel to-pe4
    signaling {
      virtual-circuit-identifier 5
      tldp {
      }
    }
  }
}
}
}

```

The pseudowires configured in VPWS network instances support the same characteristics as in the case of MAC-VRFs, with the following differences:

- The evpn-route-conflict **oper-down-reason** does not apply to pseudowires in VPWS services.
- The reception of **pw-status** bits signaling an error does not bring down the pseudowire. Therefore, some of the **pseudowire oper-down-reason** options are not applicable to pseudowires in VPWS services.
- The Layer 2 MTU is handled the same way as in MAC-VRFs, with the only difference being that **oper-vpws-mtu** is used instead of **oper-mac-vrf-mtu**.
- A mismatch in the signaled **mtu**, **vc-type**, **flow-label**, or **control-word** is handled similarly to the MAC-VRFs.
- MAC-flush behavior is not applicable to pseudowires in VPWS services. The same applies to other interactions with MAC-VRF bridge table specific features.

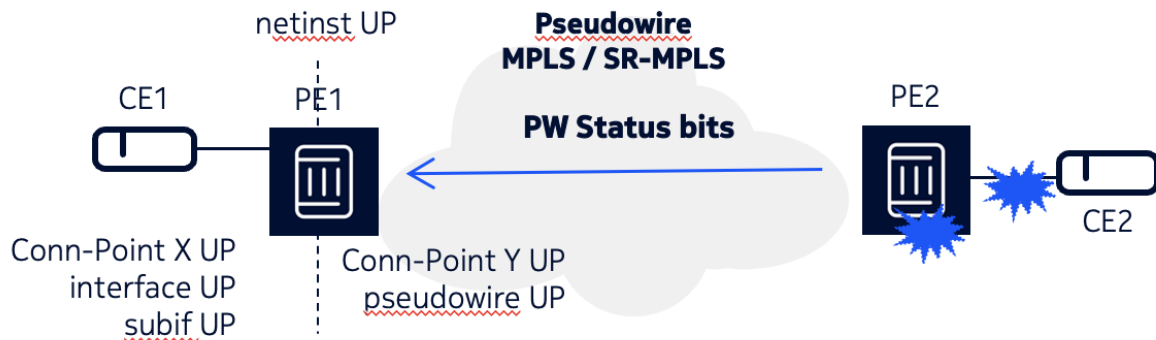
In VPWS network instances, the following objects have their own operational state:

- **network-instance**
- **bgp-evpn bgp-instance**
- **connection-point**
- **sub-interface**

The **network-instance oper-state** is always up as long as the network instance **admin-state enable** is configured. The **connection-point oper-state** is up as long as at least one object in the **connection-point** is operationally up. This object can be a subinterface **oper-state up** or a pseudowire.

The following figure shows a case where PE1 is attached to a VPWS service with a subinterface and a pseudowire.

Figure 5: VPWS service structure



If the interface and/or subinterface CE2 goes down, the following occurs:

- PE2 signals the corresponding **pw-status** bits local-attachment-circuit-ingress/egress-fault.
- Upon reception of the **pw-status** bits from PE2, PE1 does not bring down the pseudowire or connection point. This is different from the case of MAC-VRFs.
- If the VPWS network instance goes **oper-state down**, PE2 releases the label allocated to the VPWS network instance and signals the withdrawal of the label to PE1. This will bring down the pseudowire.

Link Loss Forwarding in VPWS network instances

Link Loss Forwarding is supported in cases where the VPWS network instance contains a pseudowire and a subinterface.

The Link Loss Forwarding feature is enabled by the **interface ethernet link-loss-forwarding true** command on the interface connected to the CE over which the fault signaling is propagated. In addition, the command **standby-signaling[=power-off|lacp]** must be configured on the same interface so that `ch_mgr` can trigger the relevant standby signaling to the CE.

In the case where VPWS contains a pseudowire and a subinterface, the reception of **pw-status** bits notifying the remote fault triggers the link loss forwarding procedures to propagate the fault.

In the case where **interface ethernet link-loss-forwarding true** is configured on PE1 for the VPWS in the preceding figure (CE1 subinterface on connection point X, pseudowire-1 on connection point Y), the following occurs:

- If PE2 signals a remote fault, PE1 brings down the interface to CE1 due to link loss forwarding, and connection point X goes operationally down.
- Subsequently, the pseudowire receives an event that connection point X is down, but `pw_mgr` will not react to this (it will not signal any fault in the **pw-status** bits) because connection point X is down due to link loss forwarding.

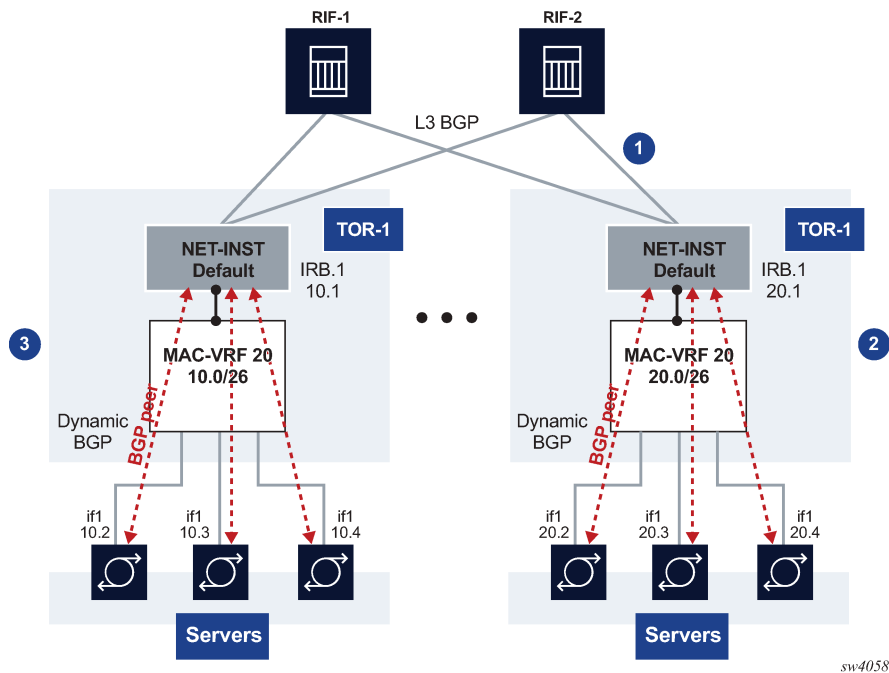
4.1.11 Server aggregation configuration example

Figure 6: Server aggregation example shows an example of using MAC-VRF network instances to aggregate servers into the same subnet.

In this example, Leaf-1 and Leaf-2 are configured with MAC-VRF instances that aggregate a group of servers. These servers are assigned IP addresses in the same subnet and are connected to the leaf default network instance by a single IRB subinterface. The servers use a PE-CE BGP session with the IRB IP address to exchange reachability.

Using the MAC-VRF with an IRB subinterface saves routed subinterfaces on the default network instance; only one routed subinterface is needed, as opposed to one per server.

Figure 6: Server aggregation example



In this example:

1. TORs peer (eBGP) to two or four RIFs.
2. MAC-VRF 20 is defined in TORs with an IRB interface with IPv4/IPv6 addresses. DHCP relay is supported on IRB subinterfaces.
3. The following Layer 2 features are implemented or loop and MAC duplication protection:
 - MAC duplication with oper-down or blackhole actions configured on the bridged subinterfaces
 - storm control for BUM traffic on bridged subinterfaces

This example uses the following features:

- MAC-VRF with bridge subinterfaces and IRB subinterfaces to the default network instance
- PE-CE BGP sessions for IPv4 and IPv6 address families
- MAC duplication with oper-down or blackhole actions configured on the bridged subinterfaces
- storm control for BUM traffic on bridged subinterfaces

4.1.11.1 Configuration for server aggregation example

The following shows the configuration of Leaf-1 in [Figure 6: Server aggregation example](#) and its BGP session via IRB to server 1. Similar configuration is used for other servers and other TORs.

```
--{ [FACTORY] + candidate shared default }--[ interface * ]--
A:Leaf-1# info
  interface ethernet-1/1 {
    description tor1-server1
    vlan-tagging true
    subinterface 1 {
      type bridged
      vlan {
        encap {
          single-tagged {
            vlan-id 100
          }
        }
      }
    }
  }

// Configure an IRB interface and sub-interface that will connect
the MAC-VRF to the existing default network-instance.

--{ [FACTORY] + candidate shared default }--[ interface irb* ]--
A:Leaf-1# info
  interface irb1 {
    subinterface 1 {
      ipv4 {
        admin-state enable
        address 10.0.0.2/24 {
        }
      }
      ipv6 {
        admin-state enable
        address 2001:db8::2/64 {
        }
      }
    }
  }

// Configure the network-instance type mac-vrf
and associate the bridged and irb interfaces to it.

--{ [FACTORY] + candidate shared default }--[ network-instance MAC-VRF-1 ]--
A:Leaf-1# info
  type mac-vrf
  admin-state enable
  interface ethernet-1/1.1 {
  }
  interface irb1.1 {
  }

// Associate the same IRB interface to the network-
instance default and configure the BGP IPv4 and IPv6 neighbors to DUT1 and DUT3.

--{ [FACTORY] + candidate shared default }--[ network-instance default ]--
A:Leaf-1# info
  type default
  admin-state enable
  router-id 2.2.2.2
```

```
interface irb1.1 {
}
protocols {
  bgp {
    admin-state enable
    afi-safi ipv4-unicast {
      admin-state enable
    }
    autonomous-system 64502
    router-id 10.0.0.2
    ebgp-default-policy {
      import-reject-all false
    }
    failure-detection {
      enable-bfd true
      fast-failover true
    }
  }
  group leaf {
    admin-state enable
    export-policy pass-all
    afi-safi ipv4-unicast {
      admin-state enable
    }
    afi-safi ipv6-unicast {
      admin-state enable
    }
    local-as as-number 64502 {
    }
    timers {
      minimum-advertisement-interval 1
    }
  }
  afi-safi ipv4-unicast {
    admin-state enable
  }
  afi-safi ipv6-unicast {
    admin-state enable
  }
  neighbor 10.0.0.1 {
    peer-as 64501
    peer-group leaf
    transport {
      local-address 10.0.0.2
    }
  }
  neighbor 2001:db8::1 {
    peer-as 64501
    peer-group leaf
    transport {
      local-address 2001:db8::2
    }
  }
}
}
```

4.2 VXLAN v4

This chapter describes the implementation for VXLAN tunnels that use IPv4 in the underlay.

- [VXLAN configuration](#)

- [VXLAN and ECMP](#)
- [VXLAN ACLs](#)
- [QoS for VXLAN tunnels](#)
- [VXLAN statistics collection](#)

4.2.1 VXLAN configuration

VXLAN on SR Linux uses a tunnel model where VXLAN interfaces are bound to network instances, in the same way that subinterfaces are bound to network instances. Up to one VXLAN interface per network instance is supported.

Configuration of VXLAN on SR Linux is tied to EVPN. VXLAN configuration consists of the following steps:

1. Configure a tunnel interface and VXLAN interface.

A tunnel interface for VXLAN is configured as `vxlان<N>`, where N can be 0 to 255.

A VXLAN interface is configured under a tunnel interface. At a minimum, a VXLAN interface must have an index, type, and the ingress VXLAN network identifier (VNI).

- The index can be a number in the range 0 to 4294967295.
- The type can be bridged or routed and indicates whether the vxlan-interface can be linked to a MAC-VRF (bridged) or IP-VRF (routed).
- The system looks for the ingress VNI in incoming VXLAN packets to classify them to this VXLAN interface and its network instance.

Configuration of an explicit VXLAN interface egress source IP is not permitted, given that the data path supports one source tunnel IP address for all VXLAN interfaces. The source IP used in the VXLAN interfaces is the IPv4 address of sub-interface `system0.0` in the default network instance.

2. Associate the VXLAN interface to a network instance.

A VXLAN interface can only be associated with one network instance, and a network instance can have only one VXLAN interface.

3. Associate the VXLAN interface to a BGP-EVPN instance.

The VXLAN interface must be linked to a BGP-EVPN instance so that VXLAN destinations can be dynamically discovered and used to forward traffic.

The following configuration example shows these steps:

```
tunnel-interface vxlan1 {
  // (Step 1) Creation of the tunnel-interface and vxlan-interface
  vxlan-interface 1 {
    type bridged
    ingress {
      vni 1
    }
    egress {
      source-ip use-system-ipv4-address
    }
  }
}
network-instance blue {
  type mac-vrf
  admin-state enable
  description "network instance blue"
```

```

interface ethernet-1/2.1 {
}
vxlان-interface vxlan1.1 {
// (Step 2) Association of the vxlan-interface to the network-instance
}
protocols {
    bgp-evpn {
        bgp-instance 1 {
            admin-state enable
            vxlan-interface vxlan1.1
// (Step 3) Association of the vxlan-interface to the bgp-evpn instance
            evi 1
        }
    }
    bgp-vpn {
        bgp-instance 1 {
            route-distinguisher {
                route-distinguisher 1.1.1.1:1
            }
            route-target {
                export-rt target:1234:1
                import-rt target:1234:1
            }
        }
    }
}
}
}
}
}

```

When EVPN routes are received with VXLAN encapsulation, the SR Linux creates VXLAN Termination Endpoints (VTEPs) from the EVPN route next-hops, and each VTEP is allocated an index number (per source and destination tunnel IP addresses).

When a VTEP is created in the vxlan-tunnel table and a non-zero index allocated, a tunnel-table entry is also created for the tunnel in the tunnel-table.

If the next hop is not resolved to any route in the network-instance default route-table, the index in the vxlan-tunnel table shows as 0 for the VTEP, and no tunnel-table entry would be created in the tunnel-table for that VTEP.

The following example shows the created vxlan-tunnel entries and tunnel-table entries on reception of IMET routes from three different PEs.

```

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state tunnel vxlan-tunnel vtep *
tunnel {
    vxlan-tunnel {
        vtep 10.22.22.2 {
            index 677716962894
// index allocated per source and destination tunnel IP addresses.
// Index of 0 would mean that 10.22.22.2 is not resolved in the route-table
// and no tunnel-table entry is created.
            last-change "17 hours ago"
        }
        vtep 10.33.33.3 {
            index 677716962900
            last-change "17 hours ago"
        }
        vtep 10.44.44.4 {
            index 677716962897
            last-change "17 hours ago"
        }
    }
}
}
}

```

```

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance default tunnel-table ipv4
  network-instance default {
    tunnel-table {
      ipv4 {
        tunnel 10.22.22.2/32 type vxlan owner vxlan_mgr id 1 {
// tunnel table entry for VTEP 10.22.22.2, created
after the vxlan-tunnel vtep 10.22.22.2
          next-hop-group 677716962900 // NHG-ID allocated by fib_mgr
          metric 0
          preference 0
          last-app-update "17 hours ago"
          vxlan {
            destination-address 10.22.22.2
            source-address 10.11.11.1
            time-to-live 255
          }
        }
        tunnel 10.33.33.3/32 type vxlan owner vxlan_mgr id 3 {
          next-hop-group 677716962900
          metric 0
          preference 0
          last-app-update "17 hours ago"
          vxlan {
            destination-address 10.33.33.3
            source-address 10.11.11.1
            time-to-live 255
          }
        }
        tunnel 10.44.44.4/32 type vxlan owner vxlan_mgr id 2 {
          next-hop-group 677716962897
          metric 0
          preference 0
          last-app-update "17 hours ago"
          vxlan {
            destination-address 10.44.44.4
            source-address 10.11.11.1
            time-to-live 255
          }
        }
      }
    }
  }
}

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance default route-table next-hop-group 677716962900
  network-instance default {
    route-table {
      next-hop-group 677716962900 {
        next-hop 0 {
          next-hop 677716962900
// NH ID allocated by fib_mgr for the NHG-ID
          active true
        }
      }
    }
  }
}

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance default route-table next-hop 677716962900
  network-instance default {
    route-table {
      next-hop 677716962900 {

```



```
// resolution of the NH ID
    type direct
    ip-address 10.1.2.2
    subinterface ethernet-1/1.1
  }
}
```

4.2.1.1 Source and destination VTEP addresses

In the network egress direction, the vxlan-interface/egress/source IP leaf determines the loopback interface that the system uses to source VXLAN packets (outer IP SA). The source IP used in the VXLAN interfaces is the IPv4 address of subinterface `system0.0` in the default network instance.

The egress VTEP (outer IP DA) is determined by EVPN and must be of the same family IPv4 as the configured source IP.

Only unicast VXLAN tunnels are supported (outer IP DA is always unicast), and ingress replication is used to deliver BUM frames to the remote VTEPs in the current release.

In the network ingress direction, the IP DA matches one of the local loopback IP addresses in the default network instance to move the packet to the VNI lookup stage (for loopback interfaces only, not other interfaces in the default network instance, such as IRB subinterfaces). The loopback IP address does not need to match the configured source IP address in the VXLAN interface.

The system can terminate any VXLAN packet with an outer destination IP matching a local loopback address, with no set restriction on the number of IPs.

4.2.1.2 Ingress or egress VNI

The configured ingress VNI determines the value used by the ingress lookup to find the network instance for a further MAC lookup. The egress VNI is specified by EVPN. For a MAC-VRF, only one egress VNI is supported. The system ignores the value of the I flag on reception. According to RFC 7348, the I flag must be set to 1. However, the system accepts VXLAN packets with I flag set to 0; the I flag is set to 1 on transmission.

4.2.1.3 VLAN tagging for VXLAN

Outer VLAN tagging is supported (one VLAN tag only), assuming that the egress subinterface in the default network instance uses VLAN tagging.

Inner VLAN tagging is transparent, and no specific handling is needed at network ingress for Layer 2 network instances. Inner VLAN tagging is not supported for VXLAN-originated traffic or VXLAN-terminated traffic in IP-VRF network instances that are BGP-EVPN IFL enabled.

4.2.1.4 Network instance and interface MTU

No specific MTU checks are done in network instances configured with VXLAN. Make the default network-instance interface MTU large enough to allow room for the VXLAN overhead. If the size of the egress VXLAN packets exceeds the IP MTU of the egress subinterface in the default network instance, the packets are still forwarded. No statistics are collected, other than those for forwarded packets.

IP MTU checks are used only for the overlay domain; that is, for interfaces doing inner packet modifications. IP MTU checks are not done for VXLAN-encapsulated packets on egress subinterfaces of the default network instance (which are in the underlay domain).

4.2.1.5 Fragmentation for VXLAN traffic

Fragmentation for VXLAN traffic is handled as follows:

- The Don't Fragment (DF) flag is set in the VXLAN outer IP header.
- The TTL of the VXLAN outer IP header is always 255.
- No reassembly is supported for VXLAN packets.

4.2.2 VXLAN and ECMP

Unicast traffic forwarded to VXLAN destinations can be load-balanced on network (underlay) ECMP links or overlay aliasing destinations.

Network LAGs, that is, LAG subinterfaces in the default network instance, are not supported when VXLAN is enabled on the platform. LAG access subinterfaces on MAC-VRFs are supported.

VXLAN-originated packets support double spraying based on overlay ECMP (or aliasing) and underlay ECMP on the default network instance.

Load-balancing is supported for the following:

- encapsulated Layer 2 unicast frames (coming from a subinterface within the same broadcast domain)
- Layer 3 frames coming from an IRB subinterface

For BUM frames, load balancing operates as follows:

- BUM supports spraying in access LAGs based on a hash. That is, BUM flows received from a VXLAN or a Layer 2 subinterface of a MAC-VRF are sprayed across egress access LAG links.
- BUM does not support spraying in underlay VXLAN next-hops. That is, BUM flows received from VXLAN or a Layer 2 subinterface of a MAC-VRF are sent to a single underlay subinterface.
- BUM VXLAN packets are sent to a single member of the next-hop group (NHG) associated with a specific VXLAN multicast destination. The chosen member is based on a hash of the NHG-ID of the VXLAN destination and the number of links in the NHG.

4.2.3 VXLAN ACLs

You can configure system filter ACLs to drop incoming VXLAN packets for reasons such as the following:

- the source IP is not recognized
- the destination IP is not an address to be used for termination
- the default destination UDP port is not being used

SR Linux supports logging VXLAN of packets dropped by ACL policies.

A system filter ACL is an IPv4 or IPv6 ACL that is evaluated before tunnel termination has occurred and before interface ACLs have been applied. A system filter can match and drop unauthorized VXLAN tunnel packets before they are decapsulated. When a system filter ACL is created, its rules are evaluated against

all transit and terminating IPv4 or IPv6 traffic that is arriving on any subinterface of the router, regardless of where that traffic entered in terms of network instance, subinterface, and so on.

The system filter matches the outer header of tunneled packets; they do not filter the payload of VXLAN tunnels. If the system-filter does not drop the VXLAN-terminated packets, only egress IRB ACLs can match the inner packets. System filters can be applied only at ingress, not egress.

See the *SR Linux Configuration Basics Guide* for information about configuring system filter ACLs.

4.2.4 QoS for VXLAN tunnels

When the SR Linux receives a terminating VXLAN packet on a subinterface, it classifies the packet to one of eight forwarding classes and one of three drop probabilities (low, medium, or high). The classification is based on the following considerations:

- The outer IP header DSCP is not considered.
- If the payload packet is non-IP, the classified FC is fc0 and the classified drop probability is lowest.
- If the payload packet is IP, and there is a classifier policy referenced by the **qos classifiers vxlan-default** command, that policy is used to determine the FC and drop probability from the header fields of the payload packet.
- If the payload packet is IP, and there is no classifier policy referenced by the **qos classifiers vxlan-default** command, the default DSCP classifier policy is used to determine the FC and drop probability from the header fields of the payload packet.

When the SR Linux adds VXLAN encapsulation to a packet and forwards it out a subinterface, the inner header IP DSCP value is not modified if the payload packet is IP, even if the egress routed subinterface has a DSCP rewrite rule policy bound to it that matches the packet FC and drop probability. The outer header IP DSCP is set to a static value or copied from the inner header IP DSCP. However, this static or copied value is modified by the DSCP rewrite rule policy that is bound to the egress routed subinterface, if the rule policy exists.

Example:

You can specify a classifier policy that applies to ingress packets received from any remote VXLAN VTEP. The policy applies to payload packets after VXLAN decapsulation has been performed.

The following example specifies a VXLAN classifier policy:

```
--{ * candidate shared default }--[ ]--
# info qos
  qos {
    classifiers {
      vxlan-default p1 {
        traffic-class 1 {
          forwarding-class fc0
        }
      }
    }
  }
}
```

See the *SR Linux Quality of Service Guide* for information on configuring QoS.

4.2.4.1 Configuring a VXLAN classifier policy

Procedure

You can specify a classifier policy that applies to ingress packets received from any remote VXLAN VTEP. The policy applies to payload packets after VXLAN decapsulation has been performed.

Example

The following example specifies a VXLAN classifier policy:

```
--{ * candidate shared default }--[ ]--
# info qos
  qos {
    classifiers {
      vxlan-default p1 {
        traffic-class 1 {
          forwarding-class fc0
        }
      }
    }
  }
}
```

4.2.5 VXLAN statistics collection

You can configure SR Linux to collect statistics for VXLAN tunnels. By default, statistics collection for VXLAN tunnels is disabled.

On 7220 IXR-D2/D3 devices, system resources allocated to statistics collection is distributed among Layer 3 or IRB subinterface statistics (if configured), Layer 2 subinterface statistics, and VXLAN statistics.

When Layer 3 or IRB subinterfaces are configured, enabling VXLAN statistics collection reduces the amount of resources available for collecting Layer 2 subinterface statistics:

- If VXLAN statistics collection is disabled (the default), resources are available for collecting statistics for all supported Layer 2 subinterfaces.
- If VXLAN statistics collection is enabled, resources are shared between Layer 2 subinterfaces and VXLAN tunnels, reducing the Layer 2 subinterface scale.

If the maximum scale is required for Layer 2 subinterfaces, VXLAN statistics collection must be disabled. Disabling VXLAN statistics collection deallocates the resources for VXLAN statistics collection and allocates them to Layer 2 subinterface statistics collection. The VXLAN statistics are not cleared, however; to clear them, use the appropriate **tools** command. See [Clearing VXLAN statistics](#).

- The amount of resources available for collecting Layer 3 or IRB subinterface statistics does not change if VXLAN statistics collection is enabled or disabled.

4.2.5.1 Enabling VXLAN statistics collection

Procedure

To enable VXLAN statistics collection, use the following command. Enabling VXLAN statistics collection disables and re-enables statistics collection on the subinterfaces. During this transition, statistics are not counted for the subinterfaces.

After enabling VXLAN statistics collection, if you attempt to configure more than the maximum number of Layer 2 subinterfaces supported with VXLAN statistics, an error message is displayed. If more than the maximum number of Layer 2 subinterfaces with VXLAN statistics are already configured in the system, VXLAN statistics collection cannot be enabled.

Example

```
--{ candidate shared default }--[ ]--
# info tunnel vxlan-tunnel statistics
  tunnel {
    vxlan-tunnel {
      statistics {
        admin-state enable
      }
    }
  }
}
```

4.2.5.2 Displaying VXLAN statistics

Procedure

To display statistics for all VXLAN tunnels or for a specified VTEP, use the **info from state** command in candidate or running mode, or the **info** command in state mode.

Example: Display statistics for all VXLAN tunnels

```
--{ running }--[ ]--
# info from state vxlan-tunnel statistics
  vxlan-tunnel {
    statistics {
      in-octets 7296882
      in-packets 83012
      in-discarded-packets 5
      out-octets 7297496
      out-packets 83007
      last-clear 2021-01-29T21:58:40.919Z
    }
  }
}
```

Example: Display statistics for a specified VTEP

```
--{ running }--[ ]--
# info from state vxlan-tunnel vtep 10.22.22.2
  vxlan-tunnel {
    vtep {
      address 10.22.22.2
      index 677716962894
      last-change 2021-01-29T21:52:34.151Z
      statistics {
        in-octets 7296882
        in-packets 83012
        in-discarded-packets 5
        out-octets 7297496
        out-packets 83007
        last-clear 2021-01-29T21:58:40.919Z
      }
    }
  }
}
```

4.2.5.3 Clearing VXLAN statistics

Procedure

You can clear the statistics for all VXLAN tunnels or for a specific VTEP.

Example: Clear statistics for all VXLAN tunnels

```
--{ running }--[ ]--
# tools vxlan-tunnel statistics clear
```

Example: Clear statistics for a specific VTEP

```
--{ running }--[ ]--
# tools vxlan-tunnel vtep 10.22.22.2 clear
```

4.3 MPLS transport for EVPN services

This chapter describes the infrastructure in SR Linux required to support EVPN for MPLS services. SR Linux support for MPLS itself and a detailed description of the MPLS label manager are described in the *SR Linux MPLS Guide*.

SR Linux features an MPLS label manager process that shares the MPLS label space among client applications that require MPLS labels; these applications include not only the applications managing the different MPLS supported protocols (LDP, SR-ISIS, and so on) but also EVPN manager and network-instance manager.

For an application such as LDP or a BGP-EVPN instance to become operational, it must be configured with a reference to a predefined range of labels, called a label block. A label block configuration includes a start-label value and an end-label value. When the label block is made available to an application, the application can use any label in the range between the start-label and end-label, inclusive of those values. The MPLS label manager ensures there is no configuration overlap between the labels of two different label blocks.

EVPN/IP-VPN services and pseudowires require label blocks allocated by label manager, as follows:

- A label block to be used by EVPN/IP-VPN and dynamic pseudowires is requested by network-instance manager based on the configuration of the **system mpls services network-instance dynamic-label-block <name>** statement. Out of this label block, a label per IP-VRF is used.
- Static pseudowires can make use of a separate static label block configured by the **system mpls services network-instance static-label-block <name>** statement.
- On 7250 IXR-X1b/X3b platforms, an additional label block must be configured to support multicast labels in MAC-VRF network-instances. This additional label block is configured with the **system mpls services network-instance dynamic-evpn-inclusive-multicast-label-block <name>** statement. This additional label block is not required on other platforms, where the multicast labels can be drawn from the dynamic label block allocated to network-instances.
- EVPN manager also requires configuration of a dynamic label block to be used for EVPN applications that are not network-instance specific; for example, EVPN multihoming ESI labels. This block is configured with the **system mpls services evpn dynamic-label-block <name>** statement.

- If the router is configured for inter-AS option B or next-hop-self route reflector functions, when acting as ASBR/ABR, a new label block is required so that service-label swap operations can be programmed. These labels are drawn from the label block configured with the **network-instance default protocols bgp bgp-label bgp-vpn dynamic-label-block <name>** statement, and are used for both EVPN and IP-VPN services.

If the system is not able to allocate a label, for example for a MAC-VRF where EVPN MPLS is enabled (either for unicast or multicast), the `bgp-instance` is kept down (there is no automatic retrying) with `oper-down-reason no-mpls-label`. In this case, ensure labels are made available, then re-add the `bgp-instance` so that the system can make another attempt to allocate a label.

4.4 EVPN for Layer 2 ELAN services

This chapter describes the components of EVPN for Layer 2 ELAN services on SR Linux.

4.4.1 Basic configuration for EVPN for Layer 2 ELAN services

Basic configuration of EVPN for Layer 2 ELAN services for SR Linux consists of the following:

- **VXLAN interface**

Contains the ingress VNI of the incoming VXLAN packets associated with the `vxlan-interface`. This is required only when EVPN uses VXLAN tunnels; it is not required when EVPN uses MPLS tunnels.

- **MAC-VRF network instance**

Where the VXLAN interface (if present) and the BGP-EVPN instance are attached. If VXLAN is used, only one VXLAN interface can be attached to a MAC-VRF network instance.

- **BGP-EVPN**

Enabled in the same MAC-VRF with a minimum configuration of the EVI and the VXLAN interface or the MPLS resolution associated with it.

The BGP instance under BGP-EVPN has an `encapsulation-type` leaf, which is VXLAN by default, but can be set as MPLS for EVPN MPLS services.

For EVPN, this configuration determines whether the BGP encapsulation extended community is advertised with value VXLAN or MPLS and the value encoded in the label fields of the advertised NLRIs is a VNI or MPLS label.

For EVPN-MPLS services, the `bgp-evpn bgp-instance 1 mpls` container supports the following additional configuration and state:

- `next-hop-resolution allowed-tunnel-types` is a list of the MPLS transport tunnel types that can be used to resolve the imported EVPN routes in the BGP-EVPN instance. At least one type is needed so that the EVPN routes can be resolved. If an EVPN route with next hop X is received, and multiple tunnel types (included in the allowed list) can resolve X, the best tunnel in the tunnel table is used.
- `control-word` determines whether the control-word is used when encapsulating Ethernet frames into EVPN-MPLS packets, and whether the control-word is expected when receiving EVPN-MPLS packets in the context of the MAC-VRF.

If the `route-distinguisher` or `route-targets/policies` are not configured, the required values are automatically derived from the configured EVI as follows:

- The route-distinguisher is derived as <ip-address:evi>, where the ip-address is the IPv4 address of the default network instance subinterface system0.0.
- The route-target is derived as <asn:evi>, where the asn is the autonomous system configured in the default network instance.

The following example shows a basic EVPN-VXLAN Layer 2 configuration consisting of a VXLAN interface, MAC-VRF network instance, and BGP-EVPN configuration:

```
--{ candidate shared default }--[ ]--
# info
...
  tunnel-interface vxlan1 {
    vxlan-interface 1 {
      type bridged
      ingress {
        vni 10
      }
      egress {
        source-ip use-system-ipv4-address
      }
    }
  }

// In the network-instance:
A:dut2# network-instance blue
--{ candidate shared default }--[ network-instance blue ]--
# info
  type mac-vrf
  admin-state enable
  description "Blue network instance"
  interface ethernet-1/2.1 {
  }
  vxlan-interface vxlan1.1 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.1
        evi 10
      }
    }
    bgp-vpn {
      bgp-instance 1 {
        // rd and rt are auto-derived from evi if this context is not configured
        export-policy pol-def-1
        import-policy pol-def-1
        route-distinguisher {
          route-distinguisher 64490:200
        }
        route-target {
          export-rt target:64490:200
          import-rt target:64490:100
        }
      }
    }
  }
}
```


The following example shows a basic EVPN-MPLS Layer 2 configuration, and the previous configuration of the label-blocks required:

```
--{ + candidate shared default }--[ network-instance MAC-VRF-2 ]--
A:leaf3# info
type mac-vrf
description "MAC-VRF-2 EVPN-MPLS"
interface ethernet-1/1.2 {
}
interface ethernet-1/2.2 {
}
interface irb0.1 {
}
protocols {
  bgp-evpn {
    bgp-instance 1 {
      admin-state enable
      encapsulation-type mpls
      evi 2
      ecmp 4
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [
            ldp
            sr-isis
          ]
        }
      }
    }
  }
  bgp-vpn {
    bgp-instance 1 {
    }
  }
}

// label blocks required for EVPN and network-instances

--{ + candidate shared default }--[ network-instance MAC-VRF-2 ]--
A:leaf3# /system mpls services
--{ + candidate shared default }--[ system mpls services ]--
A:leaf3# info
evpn {
  dynamic-label-block range-4-evpn
}
network-instance {
  dynamic-label-block range-5-services
}
```



Note: On 7250 IXR platforms, an additional label block for EVPN multicast labels must be configured.

4.4.1.1 EVPN Layer 2 basic routes

EVPN Layer 2 without multihoming includes the implementation of the BGP-EVPN address family and support for the following route types:

- EVPN MAC/IP route (or type 2, RT2)
- EVPN Inclusive Multicast Ethernet Tag route (IMET or type 3, RT3)

The MAC/IP route is used to convey the MAC and IP information of hosts connected to subinterfaces in the MAC-VRF. The IMET route is advertised as soon as BGP-EVPN is enabled in the MAC-VRF; it has the following purpose:

- auto-discovery of the remote VXLAN VTEPs or MPLS TEPs (termination endpoints) attached to the same EVI or BD
- creation of a VXLAN or MPLS default flooding list in the MAC-VRF so that BUM frames are replicated

Advertisement of the MAC/IP and IMET routes is configured on a per-MAC-VRF basis. The following example shows the default setting **advertise true**, which advertises MAC/IP and IMET routes. The example is for EVPN-VXLAN services, but is also valid for EVPN-MPLS.



Note: Changing the setting of the **advertise** parameter and committing the change internally flaps the BGP instance.

```
--{ candidate shared default }--[ network-instance blue protocols bgp-evpn bgp-
instance 1 ]--
# info detail
  admin-state enable
  vxlan-interface vxlan1.1
  evi 1
  ecmp 1
  default-admin-tag 0
  routes {
    next-hop use-system-ipv4-address
    mac-ip {
      advertise true
    }
    inclusive-mcast {
      advertise true
    }
  }
}
```

4.4.1.2 Creation of EVPN destinations based on received EVPN routes

Upon receiving EVPN routes in the context of a MAC-VRF, the router creates EVPN destinations, referred to as EVPN-VXLAN destinations or EVPN-MPLS destinations, depending on whether the BGP-EVPN instance is configured for VXLAN or MPLS encapsulation.

EVPN-VXLAN destinations

Creation of VXLAN destinations of types unicast, unicast Ethernet segment (ES), and multicast for each VXLAN interface is driven by the reception of EVPN routes.

The created unicast, unicast ES, and multicast VXLAN destinations are visible in state. Each destination is allocated a system-wide unique destination index and is an internal next-hop group ID (NHG-ID).

The destination indexes for the VXLAN destinations are shown in the following example for destination 10.22.22.4, vni 1:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info from state tunnel-interface vxlan1 vxlan-interface 1 bridge-table unicast- destinations
destination * vni *
  tunnel-interface vxlan1 {
    vxlan-interface 1 {
      bridge-table {
```

```

        unicast-destinations {
            destination 10.44.44.4 vni 1 {
                destination-index 677716962904 // destination index
                statistics {
                }
                mac-table {
                    mac 00:00:00:01:01:04 {
                        type evpn-static
                        last-update "16 hours ago"
                    }
                }
            }
        }
    }
}
--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance blue bridge-table mac-table mac 00:00:00:01:01:04
network-instance blue {
    bridge-table {
        mac-table {
            mac 00:00:00:01:01:04 {
                destination-type vxlan
                destination-index 677716962904 // destination index
                type evpn-static
                last-update "16 hours ago"
                destination "vxlan-interface:vxlan1.1 vtep:10.44.44.4 vni:1"
            }
        }
    }
}
}

```

The following is an example of dynamically created multicast destinations for a VXLAN interface:

```

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state tunnel-interface vxlan1 vxlan-interface 1 bridge-table multicast-
destinations
tunnel-interface vxlan1 {
    vxlan-interface 1 {
        bridge-table {
            multicast-destinations {
                destination 40.1.1.2 vni 1 {
                    multicast-forwarding BUM
                    destination-index 46428593833
                }
                destination 40.1.1.3 vni 1 {
                    multicast-forwarding BUM
                    destination-index 46428593835
                }
                destination 40.1.1.4 vni 1 {
                    multicast-forwarding BUM
                    destination-index 46428593829
                }
            }
        }
    }
}
}

```

EVPN-MPLS destinations

For EVPN-MPLS services, the state for the destinations can be displayed under the MAC-VRF, as shown in the following example:

```
--{ + candidate shared default }--[ network-instance MAC-VRF-2 protocols bgp-evpn bgp-instance
1 mpls ]--
# info from state
control-word false
bridge-table {
  ingress-multicast-mpls-label 1001
  ingress-unicast-mpls-label 1000
  unicast-destinations {
    destination 100.0.0.4 evi-label 1000 tunnel-id 303822694264 {
      destination-index 303822694290
      statistics {
        active-entries 1
        total-entries 1
        failed-entries 0
        mac-type evpn {
          active-entries 0
          total-entries 0
          failed-entries 0
        }
        mac-type evpn-static {
          active-entries 1
          total-entries 1
          failed-entries 0
        }
      }
    }
    mac-table {
      mac 1A:01:0B:FF:00:24 {
        type evpn-static
        last-update "2024-06-27T07:55:05.000Z (9 hours ago)"
      }
    }
  }
}
statistics {
  active-entries 1
  total-entries 1
  failed-entries 0
  mac-type evpn {
    active-entries 0
    total-entries 0
    failed-entries 0
  }
  mac-type evpn-static {
    active-entries 1
    total-entries 1
    failed-entries 0
  }
}
multicast-destinations {
  multicast-limit {
    maximum-entries 1024
    current-usage 1
  }
  destination 100.0.0.4 evi-label 5000 tunnel-id 303822694264 {
    multicast-forwarding BUM
    destination-index 303822694291
  }
}
}
```

```

next-hop-resolution {
  allowed-tunnel-types [
    ldp
    sr-isis
  ]
}

```

4.4.1.3 EVPN route selection

When a MAC is received from multiple sources, the route is selected based on the priority listed in [MAC selection](#). Learned and EVPN-learned routes have equal priority; the latest received route is selected.

When multiple EVPN-learned MAC/IP routes arrive for the same MAC but with a different key (for example, two routes for MAC M1 with different route-distinguishers), a selection is made based on the following priority:

1. EVPN MACs with higher SEQ number
2. EVPN MACs with lower IP next-hop
3. EVPN MACs with lower Ethernet Tag
4. EVPN MACs with lower RD

4.4.1.4 BGP next hop configuration for EVPN routes

You can configure the BGP next hop to be used for the EVPN routes advertised for a network instance. This next hop is by default the IPv4 address configured in interface `system 0.0` of the default network instance. However, the next-hop address can be changed to any IPv4 address.

The system does not check that the configured IP address exists in the default network instance. Any valid IP address can be used as next hop of the EVPN routes advertised for the network instance, irrespective of its existence in any subinterface of the system. However, the receiver leaf nodes create their unicast, multicast and ES destinations to this advertised next-hop, so it is important that the configured next-hop is a valid IPv4 address.

When the system or loopback interface configured for the BGP next-hop is administratively disabled, EVPN still advertises the routes, as long as a valid IP address is available for the next-hop. However, in EVPN-VXLAN instances received VXLAN traffic on that interface is dropped.

The following example configures a BGP next hop to be used for the EVPN routes advertised for a network instance.

```

--{ candidate shared default }--[ network-instance MAC-VRF-1 protocols bgp-evpn bgp-instance
1 ]--
# info
  routes {
    bridge-table {
      next-hop 1.1.1.1
    }
  }
}

```

4.4.2 MAC duplication detection for Layer 2 loop prevention in EVPN

MAC loop prevention in EVPN broadcast domains is based on the SR Linux MAC duplication feature (see [MAC duplication detection and actions](#)), but considers MACs that are learned via EVPN as well. The feature detects MAC duplication for MACs moving among bridge subinterfaces of the same MAC-VRF, as well as MACs moving between bridge subinterfaces and EVPN in the same MAC-VRF, but not for MACs moving from a VTEP to a different VTEP, via EVPN, in the same MAC-VRF.

Also, when a MAC is declared as duplicate, and the **blackhole** configuration option is added to the interface, then not only incoming frames on bridged subinterfaces are discarded if their MAC SA or DA match the blackhole MAC, but also frames encapsulated in EVPN-VXLAN/MPLS packets are discarded if their source MAC or destination MAC match the blackhole MAC in the mac-table.

When a MAC exceeds the allowed number of moves, the MAC is moved to a type duplicate irrespective of the type of move: EVPN-to-local, local-to-local, local-to-EVPN. The EVPN application receives an update that advertises the MAC with a higher sequence number, which may trigger the duplication in other nodes. The duplicate MAC can be overwritten by a higher priority type, or flushed by the **tools** command (see [Deleting entries from the bridge table](#)).

4.4.3 EVPN Layer 2 multihoming

SR Linux supports single-active multihoming and all-active multihoming, as defined in RFC 7432. The EVPN multihoming implementation uses the following SR Linux features:

- **system network instance**

A system network instance container hosts the configuration and state of EVPN for multihoming.

- **BGP network instance**

The ES model uses a BGP instance from where the RD/RT and export/import policies are taken to advertise and process the multihoming ES routes. Only one BGP instance is allowed, and all the ESes are configured under this BGP instance. The RD/RTs cannot be configured when the BGP instance is associated with the system network instance; however, the operational RD/RTs are still shown in state.

- **ESs**

An ES has an admin-state (disabled by default) setting that must be toggled to change any of the parameters that affect the EVPN control plane. In particular, the ESes support the following:

- general and per-ES boot and activation timers
- manual 10-byte ESI configuration
- all-active and single-active multihoming modes
- DF Election algorithm type Default (modulo based) or type Preference
- configuration of ES and AD per-ES routes next-hop, and ES route originating-IP per ES
- an AD per ES route is advertised per MAC-VRF, where the route carries the network-instance RD and RT
- association with an interface that can be of type Ethernet or LAG. When associated with a LAG, the LAG can be static or LACP-based. In case of LACP, the same **system-id**, **system-priority**, and **port-key** settings must be configured on all the nodes attached to the same ES.

- **aliasing load balancing**

This hashing operation for aliasing load balancing uses the hash fields that the platform uses for ECMP or LAG load balancing.

- **reload-delay timer**

The reload-delay timer configures an interface to be shut down for a period of time following a node reboot or an IMM reset to avoid black-holing traffic.

4.4.3.1 EVPN Layer 2 multihoming procedures

EVPN relies on three different procedures to handle multihoming: DF election, split-horizon, and aliasing. DF election is relevant to single-active and all-active multihoming; split-horizon and aliasing are relevant only to all-active multihoming.

- **Designated Forwarder (DF) Election**

The DF is the leaf that forwards BUM traffic in the ES. Only one DF can exist per ES at a time, and it is elected based on the exchange of ES routes (type 4) and the subsequent DF Election Algorithm (DF Election Alg).

In single-active multihoming, the non-DF leafs bring down the subinterface associated with the ES.

In all-active multihoming, the non-DF leafs do not forward BUM traffic received from remote EVPN PEs.

- **split-horizon**

This is the mechanism by which BUM traffic received from a peer ES PE is filtered so that it is not looped back to the CE that first transmitted the frame. This mechanism is different depending on whether the transport is VXLAN or MPLS. For VXLAN, local bias is applied, as described in RFC 8365. For EVPN-MPLS packets, the split-horizon filtering is based on ESI-label based filtering, as described in RFC 7432.

- **aliasing**

This is the procedure by which PEs that are not attached to the ES can process non-zero ESI MAC/IP routes and AD routes and create ES destinations to which per-flow ECMP can be applied.

To support multihoming, EVPN supports two additional route types:

- **ES routes (type 4)**

Used for ES discovery on all the leafs attached to the ES and DF Election.

ES routes use an ES-import route target extended community, its value derived from the ESI, so that its distribution is limited to only the leafs that are attached to the ES.

The ES route is advertised with the DF Election extended community, which indicates the intent to use a specific DF Election Alg and capabilities.

On reception of the remote ES routes, each PE builds a DF candidate list based on the originator IP of the ES routes. Then, based on the agreed DF Election Alg, each PE elects one of the candidates as DF for each MAC-VRF where the ES is defined.

- **AD route (type 1)**

Advertised to the leafs attached to an ES. There are two versions of AD routes:

- **AD per-ES route**

Used to advertise the multihoming mode (single-active or all-active) and the ESI label, which is not advertised or processed in case of EVPN-VXLAN, but it is advertised and used in the case of EVPN-MPLS. Its withdrawal enables the mass withdrawal procedures in the remote PEs.

– AD per-EVI route

Used to advertise the availability of an ES in an EVI and its VNI/label. It is needed by the remote leafs for the aliasing procedures.

Both versions of AD routes can influence the DF Election. Their withdrawal from a leaf results in removing that leaf from consideration for DF Election for the associated EVI, as long as **ac-df exclude** is configured. The AC-DF capability can be set to **exclude** only if the DF election algorithm type is set to **preference**.

4.4.3.2 All-active multihoming

4.4.3.2.1 EVPN-VXLAN local bias for all-active multihoming

Local bias for all-active multihoming in EVPN-VXLAN services is based on the following behavior at the ingress and egress leafs:

- At the ingress leaf, any BUM traffic received on an all-active multihoming LAG subinterface associated with an EVPN-VXLAN MAC-VRF is flooded to all local subinterfaces, irrespective of their DF or NDF status, and VXLAN tunnels.
- At the egress leaf, any BUM traffic received on a VXLAN subinterface associated with an EVPN-VXLAN MAC-VRF is flooded to single-homed subinterfaces and multihomed subinterfaces whose ES is not shared with the owner of the source VTEP if the leaf is DF for the ES.

In SR Linux, the local bias filtering entries on the egress leaf are added or removed based on the ES routes, and they are not modified by the removal of AD-per-EVI/ES routes. This configuration may cause blackholes in the multihomed CE for BUM traffic if the local subinterfaces are administratively disabled.

4.4.3.2.2 EVPN-MPLS split-horizon filtering for all-active multihoming

For Ethernet segments attached to EVPN-MPLS MAC-VRFs, EVPN manager requests a label block from which it draws the ES labels. A single label is allocated per ES on the router, and that label is advertised in the EVPN auto-discovery-per-ES route. The peer PEs attached to the same ES as the local router push the received ESI label at the bottom of the stack when sending BUM encapsulated packets to the advertising PE. The ESI label identifies BUM packets that are sourced from a given ES, and therefore cannot be forwarded to the same ES from where they came.

4.4.3.3 Single-active multihoming

[Figure 7: EVPN Layer 2 single-homing configuration](#) shows a single-active ES attached to two leaf nodes. In this configuration, the ES in single-active mode can be configured to do the following:

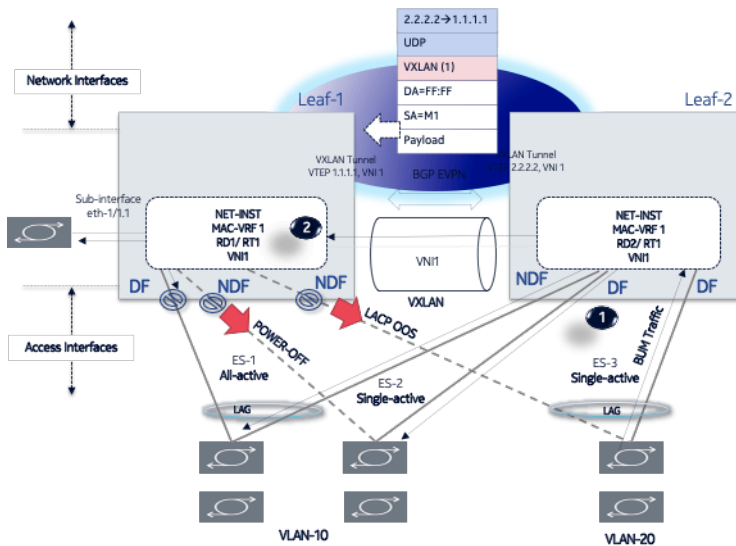
- associate with an Ethernet interface or a LAG interface, as all-active ESes
- coexist with all-active ESes on the same node, as well as in the same MAC-VRF service
- signal the non-DF state to the CE by using LACP out-of-sync signaling or power off

Optionally, the ES can be configured not to signal to the CE. When the LACP sync flag or power off is used to signal the non-DF state to the CE or server, all of the subinterfaces are active on the same

node; that is, load balancing is not per-service, but rather per-port. This mode of operation is known as EVPN multihoming port-active mode.

- connect to a CE that uses a single LAG to connect to the ES or separate LAG/ports per leaf in the ES

Figure 7: EVPN Layer 2 single-homing configuration



All peers in the ES must be configured with the same multihoming mode; if the nodes are not configured consistently, the oper-multi-homing-mode in state is single-active. From a hardware resource perspective, no local-bias-table entries are populated for ESes in single-active mode.

The following features work in conjunction with single-active mode:

- [Preference-based DF election/non-revertive option](#) configures ES peers to elect a DF based on a preference value, as well as an option to prevent traffic from reverting back to a former DF node.
- [Attachment Circuit-influenced DF Election \(AC-DF\)](#) allows the DF election candidate list for a network instance to be modified based on the presence of the AD per-EVI and per-ES routes.
- [Standby LACP-based or power-off signaling](#) configures how the node's non-DF state is signaled to the multihomed CE.

4.4.3.3.1 Preference-based DF election/non-revertive option

Preference-based DF election is defined in *draft-ietf-bess-evpn-pref-df* and specifies a way for the ES peers to elect a DF based on a preference value where highest preference-value wins. The *draft-ietf-bess-evpn-pref-df* document also defines a non-revertive mode, so that on recovery of a former DF node, traffic does not revert to the node. This mode is desirable in most cases to avoid double impact on the traffic (failure and recovery).

The configuration requires the command **df-election/algorithm/type preference** and the corresponding **df-election/algorithm/preference-alg/preference-value**. Optionally, you can set non-revertive mode to **true**. See [EVPN multihoming configuration example](#).

Nokia recommends that all of the peers in the ES be configured with the same algorithm type. However, if that is not the case, all the peers fall back to the default algorithm or operational type.

4.4.3.3.2 Attachment Circuit-influenced DF Election (AC-DF)

Attachment Circuit-influenced DF Election (AC-DF) refers to the ability to modify the DF election candidate list for a network instance based on the presence of the AD per-EVI and per-ES routes. When AC-DF is enabled with the **ac-df include** command, a node cannot become DF if it has the ES subinterface in administratively disabled state. The AC-DF capability is defined in RFC 8584, and it is by default enabled.

Nokia recommends disabling the AC-DF capability, **ac-df exclude** command, when single-active multihoming is used and standby-signaling (**lACP power-off** command) signals the non-DF state to the multihomed CE or server. In this case, the same node must be DF for all the contained subinterfaces. Administratively disabling one subinterface does not cause a DF switchover for the network instance if **ac-df exclude** is configured.

The AC-DF capability is configured with the command **df-election algorithm preference-alg capabilities ac-df; include** is the default. See [EVPN multihoming configuration example](#).

4.4.3.3.3 Standby LACP-based or power-off signaling

Standby LACP-based or power-off signaling is used for cases where the AC-DF capability is excluded, and the DF election port-active mode is configured.

When single-active multihoming is used and all subinterfaces on the node for the ES must be in DF or non-DF state, the multihomed CE does not send traffic to the non-DF node. SR Linux supports two ways of signaling the non-DF state to the multihomed CE: LACP standby or power-off.

Signaling the non-DF state is configured at the interface level, using the command **interface ethernet standby-signaling**, and must also be enabled for a specific ES using the **ethernet-segment df-election interface-standby-signaling-on-non-df** command. See [EVPN multihoming configuration example](#).

The LACP signaling method is only available on LAG interfaces with LACP enabled. When the node is in the non-DF state, it uses an LACP out-of-sync notification (the sync bit is clear in the LACP PDUs) to signal the non-DF state to the CE. The CE then brings down LACP, and the system does not jump to the collecting-distributing state and neither does the peer (because of the out_of_sync state). After the port is out of standby mode, LACP needs to be re-established, after which the forwarding ports need to be programmed.

The power-off signaling is available on Ethernet and LAG interfaces. When the node is in non-DF state, the interface goes operationally down, and the lasers on the Ethernet interfaces (all members in case of a LAG) are turned off. This process brings the CE interface down and avoids any traffic on the link. The interfaces state show `oper - state down` and `oper - down - reason standby-signaling`.

4.4.3.4 Reload-delay timer

After the system boots, the reload-delay timer keeps an interface shut down with the laser off for a configured amount of time until connectivity with the rest of network is established. When applied to an access multihomed interface, typically an ES interface), this delay can prevent black-holing traffic coming from the multihomed server or CE.

In EVPN multihoming scenarios, if one leaf in the ES peer group is rebooting, coming up after an upgrade or a failure, it is important for the ES interface not to become active until after the node is ready to forward traffic to the core. If the ES interface comes up too quickly and the node has not programmed its forwarding tables yet, traffic from the server is black-holed. To prevent this from happening, you can configure a

reload-delay timer on the ES interface so that the interface does not become active until after network connectivity is established.

When a reload-delay timer is configured, the interface port is shut down and the laser is turned off from the time that the system determines the interface state following a reboot or reload of the XDP process until the number of seconds specified in the reload-delay timer elapse.

The reload-delay timer is only supported on Ethernet interfaces that are not enabled with breakout mode. For a multihomed LAG interface, configure the reload-delay timer on all the interface members. The reload-delay timer can be from 1 to 86 400 seconds. There is no default value; if not configured for an interface, there is no reload-delay timer.

Only configure ES interfaces with a non-zero reload-delay timer. Do not configure a reload-delay timer on single-homed interfaces and network interfaces (used to forward EVPN traffic).

The following example sets the reload-delay timer for an interface to 20 seconds. The timer starts following a system reboot or when the IMM is reconnected, and the system determines the interface state. During the timer period, the interface is deactivated and the port laser is inactive.

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1
  interface ethernet-1/1 {
    admin-state enable
    ethernet {
      reload-delay 20
    }
  }
}
```

When the reload-delay timer is running, the port-oper-down-reason for the port is shown as interface-reload-timer-active. The reload-delay-expires state indicates the amount of time remaining until the port becomes active. For example:

```
--{ * candidate shared default }--[ ]--
# info from state interface ethernet-1/1
  interface ethernet-1/1 {
    description eth_seg_1
    admin-state enable
    mtu 9232
    loopback-mode false
    ifindex 671742
    oper-state down
    oper-down-reason interface-reload-time-active
    last-change "51 seconds ago"
    linecard 1
    forwarding-complex 0
    vlan-tagging true
    ...
    ethernet {
      auto-negotiate false
      lacp-port-priority 32768
      port-speed 100G
      hw-mac-address 00:01:01:FF:00:15
      reload-delay 20
      reload-delay-expires "18 seconds from now"
      flow-control {
        receive false
        transmit false
      }
    }
  }
}
```

4.4.3.5 EVPN multihoming configuration example

The following is an example of a single-active multihoming configuration, including standby signaling power-off, AC-DF capability, and preference-based DF algorithm.

The following configures power-off signaling and the reload delay timer for an interface:

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/21 ethernet
  standby-signaling power-off // needed to signal non-DF state to the CE
  reload-delay 100 // upon reboot, this is required to avoid attracting traffic
  from the multi-homed CE until the node is ready to forward.
  // The time accounts for the time it takes all network protocols to be up
  and forwarding entries ready.
```

The following configures DF election settings for the ES, including preference-based DF election and a preference value for the DF election alg. The **ac-df** setting is set to **exclude**, which disables the AC-DF capability. The **non-revertive** option is enabled, which prevents traffic from reverting back to a former DF node when the node reconnects to the network.

```
--{ * candidate shared default }--[ system network-instance protocols evpn ethernet-
segments bgp-instance 1 ethernet-segment eth_seg_1 ]--
# info
  admin-state enable
  esi 00:01:00:00:00:00:00:00:00:00
  interface ethernet-1/21
  multi-homing-mode single-active
  df-election {
    interface-standby-signaling-on-non-df { // presence container that enables
the standby-signaling for the ES
    }
    algorithm {
      type preference // enables the use of preference based DF election
      preference-alg {
        preference-value 100 // changes the default 32767 to a
specific value
      }
      capabilities {
        ac-df exclude // turns off the default ac-df capability
        non-revertive true // enables the non-revertive mode
      }
    }
  }
}
```

The following shows the state, and consequently the configuration, of an ES for single-active multihoming and indicates the default settings for the algorithm or operational type. Nokia recommends configuring all of the peers in the ES with the same algorithm type. However, if that is not the case, all the peers fall back to the default algorithm.

```
--{ * candidate shared default }--[ system network-instance protocols evpn ethernet-
segments bgp-instance 1 ethernet-segment eth_seg_1 ]--
# info
  admin-state enable
  oper-state up
  esi 00:01:00:00:00:00:00:00:00:00
  interface ethernet-1/21
  multi-homing-mode single-active
  oper-multi-homing-mode single-active // oper mode may be different if not all
the ES peers are configured in the same way
```

```

df-election {
  interface-standby-signaling-on-non-df {
  }
  algorithm {
    type preference
    oper-type preference // if at least one peer in the ES is in type
default, all the peers will fall back to default
    preference-alg {
      preference-value 100
      capabilities {
        ac-df exclude
        non-revertive true
      }
    }
  }
}
routes {
  next-hop use-system-ipv4-address
  ethernet-segment {
    originating-ip use-system-ipv4-address
  }
}
association {
  network-instance blue {
    bgp-instance 1 {
      designated-forwarder-role-last-change "2 seconds ago"
      designated-forwarder-activation-start-time "2 seconds ago"
      designated-forwarder-activation-time 3
      computed-designated-forwarder-candidates {
        designated-forwarder-candidate 40.1.1.1 {
          add-time "2 seconds ago"
          designated-forwarder true
        }
        designated-forwarder-candidate 40.1.1.2 {
          add-time "2 minutes ago"
          designated-forwarder false
        }
      }
    }
  }
}
}
}
}
}

```

To display information about the ES, use the **show system network-instance ethernet-segments** command. For example:

```

--{ [FACTORY] + candidate shared default }--[ ]--
# show system network-instance ethernet-segments eth_seg_1
-----
eth_seg_1 is up, single-active
ESI      : 00:01:00:00:00:00:00:00:00:00
Alg      : preference
Peers    : 40.1.1.2
Interface: ethernet-1/21
Network-instances:
  blue
    Candidates : 40.1.1.1 (DF), 40.1.1.2
    Interface  : ethernet-1/21.1
-----
Summary
 1 Ethernet Segments Up
 0 Ethernet Segments Down
-----

```

The **detail** option displays more information about the ES. For example:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# show system network-instance ethernet-segments eth_seg_1 detail
=====
Ethernet Segment
=====
Name                : eth_seg_1
40.1.1.1 (DF)
Admin State         : enable                Oper State           : up
ESI                 : 00:01:00:00:00:00:00:00:00:00
Multi-homing        : single-active         Oper Multi-homing    : single-active
Interface           : ethernet-1/21
ES Activation Timer  : None
DF Election         : preference            Oper DF Election     : preference

Last Change         : 2021-04-06T08:49:44.017Z
=====
MAC-VRF    Actv Timer Rem   DF
eth_seg_1  0                Yes
-----
DF Candidates
-----
Network-instance  ES Peers
blue              40.1.1.1 (DF)
blue              40.1.1.2
=====
```

On the DF node, the **info from state** command displays the following:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info from state interface ethernet-1/21 | grep oper
oper-state up
oper-state not-present
oper-state up

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance blue interface ethernet-1/21.1
network-instance blue {
  interface ethernet-1/21.1 {
    oper-state up
    oper-mac-learning up
    index 6
    multicast-forwarding BUM
  }
}
```

On the non-DF node, the **info from state** command displays the following:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info from state interface ethernet-1/21 | grep oper
oper-state down
oper-down-reason standby-signaling
oper-state not-present
oper-state down
oper-down-reason port-down

--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance blue interface ethernet-1/21.1
network-instance blue {
  interface ethernet-1/21.1 {
    oper-state down
  }
}
```

```

        oper-down-reason subif-down
        oper-mac-learning up
        index 7
        multicast-forwarding none
    }
}

```

When the ES is used for EVPN-MPLS segments, the state also shows the allocated ESI label; for example:

```

--{ + candidate shared default }--[ system network-instance protocols evpn ethernet-segments
  bgp-instance 1 ]--
# info from state ethernet-segment ES-ce31 esi-label
esi-label 500

```

4.4.4 Layer 2 proxy-ARP/ND



Note: Layer 2 proxy-ARP/ND are supported on 7220 IXR-Dx and 7730 SXR platforms.

Proxy-Address Resolution Protocol (ARP) and proxy-neighbor discovery (ND) are Layer 2 functions supported on MAC-VRF network instances, specified in RFC 9161. Proxy-ARP/ND enables the learning of IP-to-MAC bindings so that leaf nodes can reply to ARP-requests or neighbor solicitations (NS) without having to flood those requests in the BD. The proxy-ARP/ND function supports ARP/ND flooding suppression and security protection against ARP/ND spoofing attacks in large BDs.

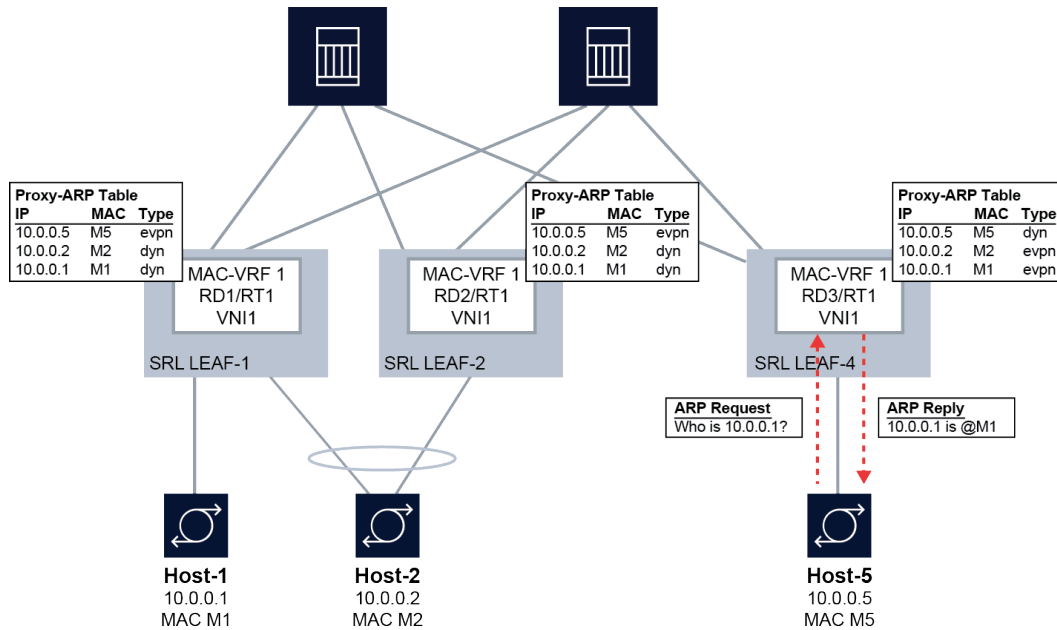
When proxy-ARP/ND is enabled for a MAC-VRF, a table is created that contains entries related to proxy-ARP/ND for the BD. Entries in the proxy-ARP/ND table can be of the following types:

- **dynamic**
Dynamic IP-MAC entries are learned by snooping ARP and ND messages; requires enabling proxy-ARP/ND and enabling dynamic learning. See [Dynamic learning for proxy-ARP/ND](#).
- **static**
Static IP-MAC entries are manually provisioned in the proxy-ARP/ND table; requires enabling proxy-ARP/ND and configuring static entries. See [Static proxy-ARP entries](#).
- **EVPN-learned**
EVPN-learned IP-MAC entries are learned from information received in EVPN MAC/IP (type 2 or RT2) routes from remote PE devices; requires enabling proxy-ARP/ND and configuring **bgp-evpn** on the MAC-VRF. See [EVPN learning for proxy-ARP/ND](#).
- **duplicate**
Duplicate entries are identified as duplicates by the IP duplication detection procedure. You can configure the criteria that determines whether an entry is a duplicate and optionally inject anti-spoofing MACs in case of duplication. See [Proxy-ARP/ND duplicate IP detection](#).

The proxy-ARP/ND table for the MAC-VRF has a default size of 250 entries. You can modify the table size. See [Proxy-ARP/ND table](#).

The following figure shows how proxy-ARP/ND functions in a BD with an example showing proxy-ARP.

Figure 8: Layer 2 proxy-ARP example



In this example, the SR Linux leaf nodes snoop and learn the local hosts and add dynamic IP-MAC entries for them in the proxy-ARP table for the MAC-VRF. The IP-MAC bindings for the local hosts are advertised in EVPN MAC/IP (type 2 or RT2) routes and installed as EVPN-learned IP-MAC entries in the proxy-ARP table on the remote SR Linux leaf nodes.

For example, SRL LEAF-1 dynamically learns the IP-MAC binding for Host-1 (10.0.0.1-M1) and adds it to the proxy-ARP table as a dynamically learned entry. The IP-MAC binding for Host-1 is advertised in an EVPN MAC/IP route. The remote SRL LEAF-4 installs the IP-MAC binding for Host-1 in its proxy-ARP table as an EVPN-learned entry. When Host-5 sends an ARP-request for 10.0.0.1, SRL LEAF-4 looks it up in the proxy-ARP table and replies with M1. Because Leaf-4 has the 10.0.0.1-M1 binding in its proxy-ARP table, it does not need to flood the request in the BD.

If the lookup is unsuccessful, the ARP-request is re-injected into the datapath and flooded in the BD. Alternatively, this flooding can be suppressed. See [Configuring proxy-ARP/ND traffic flooding options](#).

See RFC 9161 for details about proxy-ARP/ND in EVPN deployments.

4.4.4.1 Dynamic learning for proxy-ARP/ND

When dynamic learning for proxy-ARP/ND is enabled, all frames coming into bridged subinterfaces of the MAC-VRF that have Ethertype 0x0806 (for proxy-ARP) or ICMPv6 ND (for proxy-ND) are sent to the CPM for learning. This includes ARP-request and ARP-reply (including gratuitous ARP) messages and neighbor solicitation (NS) and neighbor advertisement (NA) messages (for ND). Dynamic entries in the proxy-ARP/ND table are created from both message types. For proxy-ND, dynamic entries are created only from NA messages. Learning an entry is done irrespective of the MAC DA of the ARP or ND packet (unicast or broadcast).

The dynamically learned information in the table is based on the ARP/ND payload MAC SA and IP DA, and not the frame outer MAC SA, although they normally match. A valid MAC SA must be present for the entry to be learned.

In addition to learning the dynamic entry from the ARP-request, ARP-reply, or NA, the system re-injects and forwards messages as follows:

- For received ARP-request or NS messages, the system looks up the requested IP address, and if the lookup is successful, it sends an ARP-reply or NA with the MAC→IP information. If the lookup is not successful, the ARP-request/NS is re-injected and flooded based on the flood list and the configured flooding option, with source squelching. See [Configuring proxy-ARP/ND traffic flooding options](#).

The re-injected ARP-request or NS keeps all existing non-service-delimiting tags of the original frame. Unicast ARP-requests are replied to if there is an entry in the proxy table. If the lookup is not successful, the frame is forwarded to the MAC DA.

- For ARP-reply/NA messages, the MAC DA (unicast) is looked up in the MAC table. In case of a hit, the frame is re-injected and unicasted based on the MAC table information. If there is no hit, the frame is re-injected and flooded based on the flood list information (with source squelching). The re-injected ARP-reply/NA keeps all existing non-service-delimiting tags of the original frame.
- For ARP/ND frames that are sent to the CPM, the ARP reply/NA is always unicasted to the subinterface on which the ARP-request/NS arrived, even if the MAC itself has not yet been learned in the MAC table.

If a MAC ACL is bound to the ingress of a bridged subinterface (see the *SR Linux ACL and Policy-Based Routing Guide*) and proxy-ARP/ND is enabled in the network instance, the rules in the MAC ACL are applied before extraction to the CPM. Consequently, if the MAC ACL drops an ARP packet, it never reaches the CPM.

Disabling dynamic learning for proxy-ARP/ND causes the dynamically learned entries to be flushed from the proxy-ARP/ND table. You can also set an age timer (default disabled) for dynamic entries, after which they are flushed from the proxy table.

In addition, you can set a timer to refresh dynamic entries. At a configured interval (default never) the system generates ARP-requests/NS with the intent to refresh the proxy entry; if no response is received, another one is attempted.

4.4.4.1.1 Configuring dynamic learning for proxy-ARP/ND

Procedure

To configure dynamic learning for entries in the proxy-ARP table, enable proxy-ARP and dynamic learning for the MAC-VRF. The same commands exist under proxy-nd.

Example

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-arp {
        admin-state enable
        dynamic-learning {
          admin-state enable
          age-time 600
          send-refresh 200
        }
      }
    }
  }
}
```

This example also configures the **age-time** and **send-refresh** timers. By default, both timers are disabled.

- The **age-time** specifies in seconds the aging timer for each proxy-ARP/ND entry. When the aging expires, the entry is flushed. The age is reset when a new ARP/GARP/NA for the same IP-MAC binding is received.
- The **send-refresh** timer sends ARP-request/NS messages at the configured time so that the owner of the IP address can reply and therefore refresh its IP-MAC (proxy-ARP/ND) and MAC (FDB) entries.

4.4.4.2 Static proxy-ARP entries

You can configure static entries in the proxy-ARP/ND table. Static entries have higher priority than snooped dynamic and EVPN entries in the table.

A static proxy-ARP/ND entry requires a static or dynamic MAC entry in the MAC table to become active. The static entries are advertised in MAC/IP routes, with the MAC Mobility extended community following the information associated with the MAC entry (static bit or sequence number).

The system does not validate the MAC addresses configured in static entries.

When **proxy-arp** or **proxy-nd** is disabled, the configured static entries remain in the proxy table, unlike dynamically learned and EVPN-learned entries, which are flushed from the table.

4.4.4.2.1 Configuring static proxy-ARP entries

Procedure

To configure static proxy-ARP, add entries to the proxy-ARP table.

Example

The following example enables proxy-ARP and configures a static entry in the proxy-ARP table.



Note: The system does not validate MAC addresses specified in static entries.

The same commands exist under `proxy-nd`, although the configured IP address is an IPv6 address.

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp
network-instance MAC-VRF-1 {
  bridge-table {
    proxy-arp {
      admin-state enable
      static-entries {
        neighbor 101.1.1.1 {
          link-layer-address 00:00:64:01:01:01
        }
      }
    }
  }
}
```

4.4.4.3 EVPN learning for proxy-ARP/ND

When proxy-ARP/ND is configured in an EVPN, the PE devices dynamically learn IP-MAC bindings for their local hosts and advertise them in EVPN MAC/IP routes to remote PE devices. The remote PE devices add these IP-MAC bindings to the proxy-ARP/ND table as EVPN-learned entries.

When a host sends an ARP-request or NS for a host on the remote side of the EVPN, if the PE device has the EVPN-learned IP-MAC binding in its proxy-ARP/ND table, it sends back an ARP-reply or NA with the remote host's MAC. If the IP-MAC binding does not exist in the PE device's proxy-ARP/ND table, the ARP-request or NS is flooded in the BD (ARP-request/NS flooding can be optionally disabled). See [Figure 8: Layer 2 proxy-ARP example](#) for an illustration of this process.

4.4.4.3.1 Configuring EVPN learning for proxy-ARP/ND

Procedure

To configure EVPN learning for proxy-ARP/ND, enable proxy-ARP/ND for the MAC-VRF and configure **bgp-evpn** to advertise the IP-MAC bindings in EVPN MAC/IP Advertisement routes and learn new IP-MAC bindings from the imported MAC/IP Advertisement routes.

For deployments that use a multihoming solution, **advertise-arp-nd-only-with-mac-table-entry** must be set to **true** to avoid issues with ESIs. If this setting is not configured, race conditions may result in MAC/IP routes advertised with ESI = 0 and later with non-zero ESIs.

Example: Advertise local MACs and local MAC-IP pairs

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-arp {
        admin-state enable
      }
    }
  }
}
```

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bgp-evpn bgp-instance 1 routes bridge-table
  network-instance MAC-VRF-1 {
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          routes {
            bridge-table {
              mac-ip {
                advertise true
              }
            }
          }
        }
      }
    }
  }
}
```

Example: Advertise local MAC-IP pairs only with a local MAC in the MAC table

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp
network-instance MAC-VRF-1 {
  bridge-table {
    proxy-arp {
      admin-state enable
    }
  }
}
```

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bgp-evpn bgp-instance 1 routes bridge-table
network-instance MAC-VRF-1 {
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        routes {
          bridge-table {
            mac-ip {
              advertise-arp-nd-only-with-mac-table-entry true
            }
          }
        }
      }
    }
  }
}
```



Note: EVPN MAC/IP routes advertised for Layer 2 and Layer 3 differ in that Layer 3 ARP entries are always advertised in non-zero IP MAC/IP routes, while Layer 2 proxy-ARP/ND entries are advertised only when there is a local MAC in the MAC table, because the proxy-ARP/ND entry must be active.

Enabling **advertise-arp-nd-only-with-mac-table-entry** is recommended for Layer 2 proxy-ARP/ND multihoming deployments, because the MAC delete process is handled separately from the proxy-ARP/ND delete process. If the MAC delete process happens first, the proxy-ARP/ND entry is advertised with ES = 0 right away, which flushes the MAC on the ES peer.

4.4.4.4 Configuring proxy-ARP/ND traffic flooding options

Procedure

If a lookup in the proxy-ARP/ND table is unsuccessful, by default the system re-injects the ARP-request into the data path and floods it in the BD.

For proxy-ARP, you can configure the following options for how ARP frames are flooded into the EVPN network. The default for both of these options is **true**.

- **unknown-arp-req**
This option configures whether unknown broadcast ARP-requests are flooded to EVPN destinations. Unknown in this context means the lookup in the proxy-ARP/ND table was unsuccessful. Non-broadcast ARP-requests are not affected by this option.
- **gratuitous-arp**

This option configures whether Gratuitous ARP (GARP) requests or replies are flooded to EVPN destinations. GARPs are ARP messages where the sender's IP address matches the target's IP address. Normally the MAC DA is a broadcast address.

For proxy-ND, you can configure the following options for how NS and NA frames are flooded into the EVPN network. The default for these options is **true**.

- **unknown-neighbor-advertise-host**
This option configures whether to flood NA replies for type host into the EVPN network.
- **unknown-neighbor-advertise-router**
This option configures whether to flood NA replies for type router into the EVPN network.
- **unknown-neighbor-solicitation**
This option configures whether to flood NS messages (with source squelching) into the EVPN network

Example: Disable EVPN flooding for unknown broadcast ARP-requests and GARPs

The following example disables flooding to EVPN destinations for both unknown broadcast ARP-requests and GARPs:

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp evpn flood
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-arp {
        evpn {
          flood {
            unknown-arp-req false
            gratuitous-arp false
          }
        }
      }
    }
  }
}
```

Example: Disable EVPN flooding for Neighbor Advertisements

The following example disables flooding to EVPN destinations for NA replies for type router and type host:

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-nd evpn flood
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-nd {
        evpn {
          flood {
            unknown-neighbor-advertise-router false
            unknown-neighbor-advertise-host false
          }
        }
      }
    }
  }
}
```

4.4.4.5 Proxy-ARP/ND duplicate IP detection

Proxy-ARP/ND duplicate IP detection is a security mechanism described in RFC 9161 to detect ARP/ND-spoofing attacks. In an ARP/ND-spoofing attack, an attacker sends false ARP/ND messages into a BD, with the goal of associating the attacker's IP address with a target host and directing traffic for that host to the attacker.

The proxy-ARP/ND duplicate IP detection feature monitors changes to active entries in the proxy-ARP/ND table. When an IP move occurs, for example, IP1→MAC1 is replaced by IP1→MAC2 in the table, a **monitoring-window** timer is started for a default 3 minutes. If a specified number of IP moves (default 5) is detected before the **monitoring-window** timer expires, the IP is considered to be a duplicate.

When the system detects an IP move in the proxy table, for example, IP1→MAC1 changing to IP1→MAC2, it places the IP1→MAC2 proxy table entry in pending-confirmation state for a maximum of 30 seconds. During the pending-confirmation period, the ARP/ND entry is inactive, and a confirm-message is unicast to MAC1. If no reply from MAC1 is received during the pending-confirmation period, the IP1→MAC2 entry is confirmed as legitimate and becomes active. If a reply from MAC1 is received, then MAC2 is sent a confirm-message. If MAC2 replies, an additional confirm-message is sent to MAC1. If both MAC1 and MAC2 keep replying to the confirm-messages, it triggers the duplicate IP detection procedure for IP1, because the number of IP moves exceeds the maximum allowed during the monitoring window.

When an IP is detected as a duplicate, the proxy table cannot be updated with new dynamic or EVPN-learned entries for the same IP, although you can configure a static entry for the IP. The duplicate IP is subject to this restriction until a configured **hold-down-time** expires, of which the default is 9 minutes, after which the entry for the IP is removed from the proxy table, and the monitoring process for the IP is restarted.

Anti-Spoofing MAC

You can configure an Anti-Spoofing MAC (AS-MAC). If an AS-MAC is configured, the system associates the duplicate IP with the AS-MAC in the proxy-ARP/ND table. A GARP or unsolicited-NA message with IP1→AS-MAC is sent to the local CEs, and an EVPN MAC/IP route with IP1→AS-MAC is sent to the remote PEs. This process updates the ARP/ND caches on the CEs in the BD so that CEs in the BD use the AS-MAC as MAC DA when sending traffic to IP1.

If you configure the `static-blackhole true` option, the AS-MAC is installed in the MAC table as a blackhole MAC, which discards incoming frames with a MAC source or destination matching the AS-MAC.

4.4.4.5.1 Configuring proxy-ARP/ND duplicate IP detection

Procedure

To configure proxy-ARP/ND duplicate IP detection, set the following options:

- **monitoring-window**
This is the number of minutes the system monitors a proxy-ARP/ND table entry following an IP move (default 3 minutes).
- **num-moves**
This is the maximum number of IP moves a proxy-ARP/ND table entry can have during the **monitoring-window** before the IP is considered duplicate (default 5 IP moves).
- **hold-down-time**

This is the number of minutes from the time an IP is declared duplicate to the time the IP is removed from the proxy-ARP/ND table (default 9 minutes).

- **anti-spoof-mac**
If configured, this replaces the MAC of the duplicate IP in the proxy-ARP/ND table. The AS-MAC is advertised in EVPN to remote PEs.
- **static-blackhole**
If this option is set to `true`, this installs the AS-MAC in the MAC table as a blackhole MAC, causing incoming frames with a MAC source or destination matching the AS-MAC to be discarded.

Example: Configure proxy-ARP duplicate IP detection

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp ip-duplication
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-arp {
        ip-duplication {
          monitoring-window 5
          num-moves 7
          hold-down-time 10
          anti-spoof-mac 00:CA:FE:CA:FE:08
          static-blackhole true
        }
      }
    }
  }
}
```

Example: Configure proxy-ND duplicate IP detection

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-nd ip-duplication
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-nd {
        ip-duplication {
          monitoring-window 5
          num-moves 7
          hold-down-time 10
          anti-spoof-mac 00:CA:FE:CA:FE:08
          static-blackhole true
        }
      }
    }
  }
}
```

4.4.4.5.2 Configuring processing for proxy-ARP/ND probe packets

Procedure

ARP Request and NS packets that have 0.0.0.0 or :: as the sender's IP address are used as probes by the IPv4/IPv6 duplicate address detection (DAD) function. Proxy-ARP/ND identifies these probe packets and can either process them locally, or forward them into the MAC-VRF (flooding them into the MAC-VRF) so that they reach all the hosts in the BD. You can configure whether the probe packets are processed locally or flooded to the MAC-VRF.

Example: Configure flooding for proxy-ARP probe packets

The following example configures local processing for ARP probe messages. When set to `false`, ARP probe messages are flooded to the remote nodes if unknown-ARP-requests are configured to be flooded.

By default, local processing is set to `true`, which causes ARP probe messages used by the hosts for DAD to be processed, replied if a proxy-ARP entry is hit, or reinjected into the data path.

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1
  network-instance MAC-VRF-1 {
    type mac-vrf
    bridge-table {
      proxy-arp {
        process-arp-probes false
      }
    }
  }
}
```

Example: Configure flooding for Neighbor Solicitation DAD messages

The following example configures local processing for NS DAD messages. When set to `false`, NS DAD messages are flooded to the remote nodes if unknown-neighbor-solicitation is configured, so that unknown NS messages are flooded.

By default, local processing is set to `true`, which causes NS DAD messages used by the hosts for DAD to be processed, replied if a proxy-ND entry is hit, or reinjected into the data path.

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-nd {
        process-dad-neighbor-solicitations false
      }
    }
  }
}
```

4.4.4.5.3 Displaying proxy-ARP/ND duplicate IP detection information

Procedure

You can use a **show** command to display information about duplicate IPs detected in the proxy-ARP table. A similar **show** command exists for proxy-ND.

Example

To display duplicate IP information from the proxy-ARP table for all MAC-VRFs:

```
--{ running }--[ ]--
# show network-instance * bridge-table proxy-arp ip-duplication duplicate-entries
-----
IP-duplication in network instance mac-vrf-1
-----
Monitoring window      : 3 minutes
Number of moves allowed : 5
Hold-down-time         : 10 seconds
```



```

Anti-Spoof-MAC      : 00:DE:AD:00:00:01 (Static-blackhole)
-----
Duplicate entries in network instance mac-vrf-1
-----
+-----+-----+-----+-----+
| Neighbor | MAC Address | Detect Time | Hold down time |
|          |             |             | remaining      |
+-----+-----+-----+-----+
| 10.10.10.1 | 00:DE:AD:00:00:01 | 2021-12-11T12:48:24.000Z | 10 |
| 10.10.10.2 | 00:DE:AD:00:00:01 | 2021-12-11T12:48:24.000Z | 9  |
| 10.10.10.3 | 00:DE:AD:00:00:01 | 2021-12-11T12:48:24.000Z | 10 |
+-----+-----+-----+-----+
IP-duplication in network instance mac-vrf-2
...
Total Duplicate IPs      : 4 Total 4 Active
-----

```

4.4.4.6 Proxy-ARP/ND table

When you enable proxy-ARP/ND for a MAC-VRF, this creates a table containing proxy-ARP/ND entries learned dynamically by snooping ARP/ND messages, configured statically, or learned from EVPN MAC/IP routes from remote PE nodes. By default, this table can contain up to 250 entries. You can configure the size of this table to be from 1 to 8192 entries.

The system generates a log event when the size of the table reaches 90% of the maximum size and when the table reaches 95% of the maximum size. When the table reaches 100% of its maximum size, entries for an IP can be replaced, that is, a different MAC can be learned and added for the IP, but no new IP entries can be added to the table, regardless of the type (dynamic, static, or EVPN-learned).

4.4.4.6.1 Configuring the proxy-ARP/ND table size

Procedure

By default, the proxy-ARP or proxy-ND table can contain up to 250 entries of all types (dynamic, static, EVPN, duplicate). You can increase or decrease the maximum size of the table. If you configure the table size to be lower than the number of entries it currently contains, the system stops and restarts the proxy-ARP/ND application, causing the non-static entries to be flushed from the table.

Example: Configure the proxy-ARP table size

The following example configures the size of the proxy-ARP table for a MAC-VRF:

```

--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp table-size
network-instance MAC-VRF-1 {
  bridge-table {
    proxy-arp {
      table-size 125
    }
  }
}

```

Example: Configure the proxy-ND table size

The following example configures the size of the proxy-ND table for a MAC-VRF:

```
--{ candidate shared default }--[ ]--
# info network-instance MAC-VRF-1 bridge-table proxy-arp table-size
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-nd {
        table-size 125
      }
    }
  }
}
```

4.4.4.6.2 Clearing entries from the proxy-ARP/ND table

Procedure

You can use **tools** commands to clear dynamic and duplicate entries from the proxy-ARP table. You can clear all dynamic and duplicate entries or you can clear specific entries.

Example: Clear dynamic entries from the proxy-ARP table

To clear dynamic entries from the proxy-ARP table:

```
--{ running }--[ ]--
# tools network-instance MAC-VRF-1 bridge-table proxy-arp dynamic delete-all
# tools network-instance MAC-VRF-1 bridge-table proxy-arp dynamic entry 10.10.10.1 delete-ip
```

Example: Clear duplicate entries from the proxy-ARP table

To clear duplicate entries from the proxy-ARP table:

```
--{ running }--[ ]--
# tools network-instance MAC-VRF-1 bridge-table proxy-arp duplicate delete-all
# tools network-instance MAC-VRF-1 bridge-table proxy-arp duplicate entry 10.10.10.1 delete-ip
```

Example: Clear dynamic entries from the proxy-ND table

To clear dynamic entries from the proxy-ND table:

```
--{ running }--[ ]--
# tools network-instance MAC-VRF-1 bridge-table proxy-nd dynamic delete-all
# tools network-instance MAC-VRF-1 bridge-table proxy-nd dynamic entry 10.10.10.1 delete-ip
```

Example: Clear duplicate entries from the proxy-ND table

To clear duplicate entries from the proxy-ND table:

```
--{ running }--[ ]--
# tools network-instance MAC-VRF-1 bridge-table proxy-nd duplicate delete-all
# tools network-instance MAC-VRF-1 bridge-table proxy-nd duplicate entry 10.10.10.1 delete-ip
```

4.4.4.7 EVPN ARP/ND extended community flags

RFC 9047 defines an extended community (EC) for EVPN MAC/IP advertisement routes that carry flags related to ARP or ND. This EC includes flags that can indicate to a remote PE router whether an ARP/ND entry learned via EVPN belongs to a host or a router, or if the address is an anycast address. Information from the EC flags can allow remote PE routers configured for proxy ARP/ND to reply to local ARP or ND requests with correct information.

On SR Linux, the RFC 9047 EC is not carried in EVPN MAC/IP advertisement routes by default. You can configure SR Linux to advertise the RFC 9047 EC along with the MAC/IP routes advertised for local static, dynamic, and duplicate proxy-ARP/ND entries. When you enable SR Linux to advertise the RFC 9047 EC, you can also configure how the flags are advertised in EVPN MAC/IP routes and how the flags are set in replies to NS messages for EVPN entries.

This feature is not supported along with IRB interfaces.

The feature must be configured consistently on all PEs attached to the same BD, and in particular attached to the same ES and BD.

At the BGP level, only one EVPN ARP/ND EC is expected to be received along with an EVPN MAC/IP advertisement route. If more than one EVPN ARP/ND EC is received, the router considers only the first one on the list for processing purposes on the PE.

At the EVPN level, the flags are ignored if they come in a MAC/IP route without IP.

RFC 9047 extended community flags

RFC 9047 defines the following flags for the EVPN ARP/ND extended community. The flags field is the third octet of the EC.

- **router flag (R flag) (bit 7 in the flags field)**
If set, the R flag indicates the IPv6→MAC pair in the MAC/IP advertisement route belongs to an IPv6 router; if not set, the IPv6→MAC pair belongs to a host.
- **override flag (O flag) (bit 6 in the flags field)**
The O flag is normally set to 1 by the egress router when advertising IPv6→MAC pairs, set to 0 when IPv6 anycast is enabled for the BD or interface, and ignored when the received MAC/IP advertisement route belongs to an IPv4→MAC pair.



Note: Anycast IPv6 addresses are not supported in SR Linux proxy-ND tables.

The ingress PE installs the ND entry with the received O flag and always uses this O flag value when replying to an NS message for the IPv6 address.

The O flag is conveyed in EVPN advertisements and advertised based on the received NA messages, and is installed in the proxy-ND entries. However, other than transferring the information between ND and EVPN, the O flag does not change any of the selection of ND entries in proxy-ND.

- **immutable flag (I flag) (bit 4 in the flags field)**
If set, the I flag indicates the IP→MAC pair in the MAC/IP advertisement route is a configured ARP/ND entry; the IP address in the EVPN MAC/IP advertisement route can only be bound together with the MAC address specified in the same route.

The I flag applies to IPv6 and IPv4 entries and is set in MAC/IP advertisements for static proxy-ARP/ND entries.

4.4.4.7.1 Enabling advertisement of the RFC 9047 extended community

When you enable advertisement of the RFC 9047 extended community, the EC is advertised along with the MAC/IP routes advertised for local static, dynamic, and duplicate proxy-ARP/ND entries. Enabling this feature also allows you to control how the EC is processed and how ARP/ND entries are selected based on the I flag.

The following example enables SR Linux to advertise the RFC 9047 EC:

```
--{ candidate shared default }--[ ]--
# info network-instance m1 protocols bgp-evpn bgp-instance 1 routes bridge-table
  network-instance m1 {
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          routes {
            bridge-table {
              mac-ip {
                advertise-arp-nd-extended-community true
              }
            }
          }
        }
      }
    }
  }
}
```

By default, the `advertise-arp-nd-extended-community` command is set to `false`. When this command is set to `true`, the R, I, and O flags are advertised and processed as described in the following sections:

- [Advertising and processing for the R flag](#)
- [Advertising and processing for the I flag](#)
- [Advertising and processing for the O flag](#)

Advertising and processing for the R flag

After you enable SR Linux to advertise the RFC 9047 EC, you can control the advertisement of proxy-ND EVPN entries and the reply to NS messages received for those EVPN entries by configuring the `advertise-neighbor-type` command. For example:

```
--{ candidate shared default }--[ ]--
# info network-instance m1 bridge-table
  network-instance m1 {
    bridge-table {
      proxy-nd {
        evpn {
          advertise-neighbor-type router-host
        }
      }
    }
  }
}
```

The `advertise-neighbor-type` command has three options: `host`, `router`, and `router-host`. The effect this command has on how SR Linux processes the R flag for static, dynamic, EVPN, and duplicate proxy-ND entries is described in [Table 3: Effect the advertise-neighbor-type setting has on proxy ND entries](#).

Note the following:

- System-generated unsolicited NA messages for duplicate entries carry R = 0 or R = 1, and are flooded to EVPN if allowed by the configured flooding options.
- The R flag in the RFC 9047 EC is ignored for proxy-ARP entries.

Table 3: Effect the advertise-neighbor-type setting has on proxy ND entries

For this proxy-ND entry type	The advertise-neighbor-type setting has this effect
Static proxy-ND entries	<p>Static proxy-ND entries are added with an R = 1 flag if configured with <code>neighbor <ipv6-address> type router</code>.</p> <p>NS messages received for the entry are replied to with the configured R flag.</p> <p>The entries are advertised in EVPN if the command <code>advertise-neighbor-type</code> is configured.</p> <p>If the configuration for an entry is changed from <code>host</code> to <code>router</code>, or from <code>router</code> to <code>host</code>, an unsolicited NA message is triggered containing the new flag.</p> <p>This unsolicited NA message is or isn't flooded into EVPN based on the configured flooding options.</p>
Dynamic proxy-ND entries	<p>When <code>proxy-nd/dynamic-learning/admin-state enable</code> is configured, the router learns dynamic entries from NA messages; the corresponding R flag is also learned and added to the proxy-ND table.</p> <p>Subsequent NS messages received for the entry are replied to with the corresponding R flag.</p> <p>The learned dynamic entries are advertised in EVPN depending configured <code>advertise-neighbor-type</code> setting:</p> <ul style="list-style-type: none"> • <code>router</code> advertises dynamic and static entries for which R = 1 • <code>host</code> advertises dynamic and static entries for which R = 0 • <code>router-host</code> advertises dynamic and static entries for which R = 0 or R = 1 <p>The <code>router-host</code> option is available only if <code>advertise-arp-nd-extended-community true</code> is configured. The R flag of the entry is then propagated in the R bit of the extended community for the route.</p> <ul style="list-style-type: none"> • the default value is <code>advertise-neighbor-type router</code>
EVPN proxy-ND entries	<p>The setting for <code>advertise-neighbor-type</code> controls the proxy-ND dynamic/static/duplicate entries that are advertised in EVPN:</p> <ul style="list-style-type: none"> • <code>router</code> only advertises in EVPN dynamic entries learned with R = 1, or static entries configured as <code>router</code> • <code>host</code> only advertises in EVPN dynamic entries learned with R = 0, or static entries configured as <code>host</code>

For this proxy-ND entry type	The advertise-neighbor-type setting has this effect
	<ul style="list-style-type: none"> router-host advertises in EVPN dynamic and static entries irrespective of the router flag, and with the correct R flag in each case. For an EVPN entry, the command adds the Router flag depending on the R flag coming in the RFC 9047 EC. irrespective of the configured option, the absence of the RFC 9047 EC in a received MAC/IP route creates a proxy-ND entry of type R = 1 (router) <p>The setting for advertise-neighbor-type must be consistent on all nodes for the same service.</p>
Duplicate proxy-ND entries	<p>Duplicate proxy-ND entries are treated as host or router entries as far as EVPN advertisement and response to NS messages is concerned.</p> <p>If advertise-neighbor-type host is configured, the duplicate entry is treated as host.</p> <p>If advertise-neighbor-type router or router-host is configured, the duplicate entry is treated as router.</p>

Advertising and processing for the I flag

When the `advertise-arp-nd-extended-community true` command is configured, SR Linux advertises the I flag as follows:

- any static proxy-ARP/ND entry is advertised with I = 1
- duplicate entries with AS-MAC are advertised with I = 1 (in addition to O = 1 and R = 0 or 1, based on configuration)
- setting of the I flag is independent of the Static bit associated with the FDB entry; the I flag is only used with proxy-ARP/ND advertisements

On reception, the I flag is processed as follows:

When SR Linux receives an EVPN MAC/IP advertisement route containing an IP→MAC and I = 1, it installs the IP→MAC entry in the proxy-ARP/ND table as an immutable binding, overriding any existing non-immutable binding for the same IP→MAC. MAC mobility does not consider the I flag.

For multiple routes to the same IP, proxy-ARP/ND selection operates according the following rules:

- local immutable ARP/ND entries (static) are preferred first
- EVPN immutable ARP/ND entries are preferred next
- regular existing ARP/ND selection applies third

If SR Linux receives multiple EVPN MAC/IP advertisement routes with the I flag set to 1 with the same IP but a different MAC address, a route is selected using these selection rules.

If SR Linux originates an EVPN MAC/IP advertisement route containing an IP→MAC and I = 1, it also originates the route with the MAC mobility EC sticky/static flag set if the MAC is static. In this case, the MAC→IP binding is immutable, and it cannot move. If an update for the same immutable and static IP→MAC is received from a different PE, one of the two routes is selected.

There are no changes for generation of confirmation messages or GARP or unsolicited-NA messages when a new entry is learned.

Advertising and processing for the O flag

When the `advertise-arp-nd-extended-community true` command is configured, SR Linux advertises the O flag as follows:

- For dynamic entries, the setting for the O flag is taken from what was learned and added to the proxy-ND table.
- For static and duplicate entries, O = 1.

On reception, the O flag is processed as follows:

- The O flag is stored in the proxy-ND table, and is used when replying to a received NS message.
- For link-local neighbor advertisements, generated for solicitations to the local chassis MAC, NA messages with O = 0 are not learned.
- If an EVPN route is received without the RFC 9047 EC, the entry is installed with the default value of O = 1.
- For solicited or unsolicited NA messages, the O flag is O = 1 or O = 0, depending on the flag in the proxy-ND entry that was previously learned.

4.4.4.8 Displaying proxy-ARP/ND information

Procedure

You can use **show** commands to display information about the contents of the proxy-ARP table. Similar **show** commands exist for proxy-ND.

Example: Display the entire proxy-ARP table

To display all entries in the proxy-ARP table for a MAC-VRF:

```
--{ candidate shared default }--[ ]--
# show network-instance MAC-VRF-1 bridge-table proxy-arp all
-----
Proxy-ARP table of network instance MAC-VRF-1
-----
+-----+-----+-----+-----+-----+-----+
| Neighbor | MAC Address | Type | State | Aging | Last Update |
+-----+-----+-----+-----+-----+-----+
| 10.10.10.1 | 00:CA:FE:CA:FE:01 | dynamic | active | 300 | 2021-12-11T12:48:24.000Z |
| 10.10.10.2 | 00:CA:FE:CA:FE:02 | evpn | active | N/A | 2021-12-11T12:48:24.000Z |
| 10.10.10.3 | 00:CA:FE:CA:FE:03 | duplicate | active | N/A | 2021-12-11T12:48:24.000Z |
+-----+-----+-----+-----+-----+-----+
Total Static Neighbors : 0 Total 0 Active
Total Dynamic Neighbors : 1 Total 1 Active
Total Evpn Neighbors : 1 Total 1 Active
Total Duplicate Neighbors : 1 Total 1 Active
Total Neighbors : 3 Total 3 Active
-----
```

The output of this command indicates the total number of entries of each type, as well as the number that are active (replies are sent for received ARP-requests). For an entry to be considered active in the proxy-ARP table, it must have a corresponding entry in the MAC table of the type listed in the following table:

Table 4: MAC table entry types required for proxy-ARP table entry types to be active

Proxy-ARP table entry type	MAC table entry type
Dynamic	Learned
Dynamic	Static
Dynamic	EVPN
Static	Learned
Static	Static
EVPN	EVPN (irrespective of the ESI)
EVPN	Static or dynamic matching the EVPN ESI
Duplicate	–

If a proxy-ARP table entry has a corresponding entry in the MAC table of a type not listed in this table, then the proxy-ARP table entry is considered inactive, so no replies are sent for received ARP-requests. If the MAC-VRF is not active, its proxy-ARP table entries are not active as well.

Example: Display a specific entry in the proxy-ARP table

```
--{ candidate shared default }--[ ]--
# show network-instance MAC-VRF-1 bridge-table proxy-arp neighbor 10.10.10.1
-----
Proxy-ARP table of network instance MAC-VRF-1
-----
Neighbor           : 10.10.10.1
MAC Address        : 00:CA:FE:CA:FE:01
Type               : dynamic
Programming Status : Success
Aging              : 300
Last Update        : 2021-12-11T12:48:24.000Z
Duplicate Detect time : N/A
Hold down time remaining: N/A
-----
```

Example: Display a summary of the proxy-ARP table

```
--{ candidate shared default }--[ ]--
# show network-instance * bridge-table proxy-arp summary
-----
Network Instance Proxy-ARP Table Summary
-----
Network Instance: MAC-VRF-1
Total Static Neighbors      :    0 Total    0 Active
Total Dynamic Neighbors    :    1 Total    1 Active
Total Evpn Neighbors       :    1 Total    1 Active
Total Duplicate Neighbors  :    1 Total    1 Active
Total Neighbors            :    3 Total    3 Active
Maximum Entries           : 250
Warning Threshold: 95% (237)
Clear Warning              : 90% (225)
-----
```



```

Network Instance: MAC-VRF-2
Total Static Neighbors      :    1 Total    1 Active
Total Dynamic Neighbors    :    1 Total    1 Active
Total Evpn Neighbors       :    0 Total    0 Active
Total Duplicate Neighbors  :    0 Total    0 Active
Total Neighbors            :    2 Total    2 Active
Maximum Entries           : 250
Warning Threshold: 95% (237)
Clear Warning              : 90% (225)
-----

```

```

-----
Total Static Neighbors      :    1 Total    1 Active
Total Dynamic Neighbors    :    2 Total    2 Active
Total Evpn Neighbors       :    1 Total    1 Active
Total Duplicate Neighbors  :    1 Total    1 Active
Total Neighbors            :    5 Total    5 Active
-----

```

4.4.4.9 Displaying proxy-ARP/ND statistics

Procedure

To display proxy-ARP statistics, use the **info from state** command in candidate or running mode or the **info** command in state mode. You can display system-wide statistics or statistics for a specific MAC-VRF network instance.

Example: Display proxy-ARP statistics

The following example displays proxy-ARP statistics for a MAC-VRF network instance:

```

--{ candidate shared default }--[ ]--
# info from state network-instance MAC-VRF-1 bridge-table proxy-arp statistics
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-arp {
        statistics {
          total-entries 0
          active-entries 0
          in-active-entries 0
          pending-entries 0
          neighbor-origin {
            origin
            total-entries 0
            active-entries 0
            in-active-entries 0
            pending-entries 0
          }
        }
      }
    }
  }
}

```

Example: Display proxy-ND statistics

The following example displays proxy-ND statistics for a MAC-VRF network instance:

```

--{ candidate shared default }--[ ]--
# info from state network-instance MAC-VRF-1 bridge-table proxy-nd statistics
  network-instance MAC-VRF-1 {
    bridge-table {
      proxy-nd {

```

```

statistics {
  total-entries 0
  active-entries 0
  in-active-entries 0
  pending-entries 0
  neighbor-origin {
    origin
    total-entries 0
    active-entries 0
    in-active-entries 0
    pending-entries 0
  }
}
}
}
}

```

4.4.5 EVPN VLAN-aware bundle services

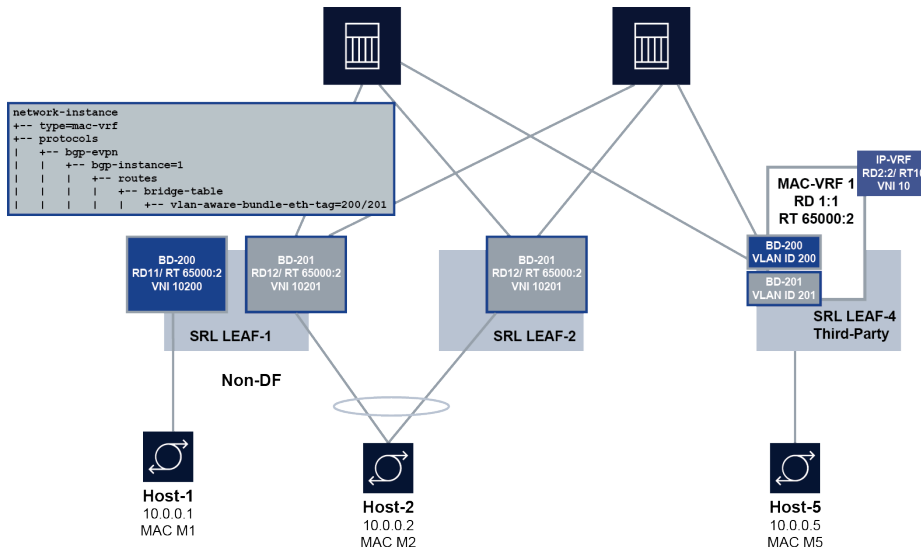
This chapter describes the components of EVPN VLAN-aware bundle services.

- [Configuring the Ethernet Tag ID for VLAN-aware bundle services](#)
- [Displaying the Ethernet Tag ID for VLAN-aware bundle services](#)

RFC 7432 defines VLAN-aware bundle services as those EVPN instances that consist of multiple broadcast domains. On SR Linux, a broadcast domain instance on a leaf node is identified by a MAC-VRF, and a MAC-VRF can contain only one broadcast domain. However, SR Linux can be attached to VLAN-aware bundle broadcast domains along with other third-party routers.

The following figure shows a configuration that features interoperability between SR Linux systems and a third-party device configured with VLAN-aware bundle services.

Figure 9: EVPN interoperability with VLAN-aware bundle services



In this configuration, Leaf-1 and Leaf-2 are SR Linux systems, and Leaf-4 is a third-party device that supports VLAN-aware bundle services. On Leaf-4, MAC-VRF1 is configured with two broadcast domains

(BDs), BD-200 and BD-201. BD-200 and BD-201 are configured with the corresponding VLAN ID value encoded as Ethernet Tag ID in the EVPN routes and the VNI.

To allow Leaf-1 and Leaf-2 to interoperate with Leaf-4, the SR Linux devices are configured to advertise a non-zero Ethernet Tag ID in the EVPN routes, and process the Ethernet Tag ID in EVPN routes received for the MAC-VRF. When this Ethernet Tag ID is configured, all MAC-VRFs of the same bundle have the same import/export route target (RT), but different route distinguisher (RD), Ethernet Tag ID, and VNI/label.

The Ethernet Tag ID can be set to a value in the range 0 to 16777215 (24-bit value). When the Ethernet Tag ID is set to a non-zero value, MAC/IP, IMET, and AD per-EVI routes for the MAC-VRF are advertised encoding the Ethernet Tag ID, configured with the **vlan-aware-bundle-eth-tag** parameter, into the `ethernet-tag-id` field of the routes. To interoperate in a VLAN-aware bundle broadcast domain where there are multiple vendors and multihomed CEs, all the vendors must advertise and process the AD per-EVI routes as per RFC 8584, where an AD per-EVI route is advertised per broadcast domain for the ES. SR Linux is fully compliant with RFC 8584.

For received routes, BGP processes the routes as usual, and imports them based on the import route-target. The `ethernet-tag-id` is part of the route-key, so BGP may keep multiple routes with same RD/RT/prefix but different `ethernet-tag-id`. For the routes imported in a specific MAC-VRF, only those routes whose `ethernet-tag-id` matches the locally configured Ethernet Tag ID are processed.

The feature is supported for the following:

- MAC-VRF network instances
- MAC-VRF network instances with IRB interfaces
- MAC-VRF with multihoming, for Layer 2 and Layer 3

4.4.5.1 Configuring the Ethernet Tag ID for VLAN-aware bundle services

Procedure

To configure VLAN-aware bundle services for SR Linux, you specify a value for the **vlan-aware-bundle-eth-tag** parameter.

When the **vlan-aware-bundle-eth-tag** is set to a non-zero value, routes are accepted only when the incoming `ethernet-tag-id` matches the configured value, for all route types imported in the MAC-VRF.

When the **vlan-aware-bundle-eth-tag** is set to zero (the default value):

- all received routes with `ethernet-tag-id = 0` are accepted (irrespective of type)
- for received routes with a non-zero `ethernet-tag-id` value, IMET and MAC/IP routes for the VXLAN instance are always accepted
- AD per-EVI routes with a non-zero `ethernet-tag-id` are rejected

Example

The following example configures a value for the **vlan-aware-bundle-eth-tag** parameter. All the EVPN routes advertised for the MAC-VRF contain this value in the `ethernet-tag-id` field.

```
--{ * candidate shared default }--[ ]--
# info network-instance mac_vrf_1 protocols bgp-evpn bgp-instance 1 routes
  network-instance mac_vrf_1 {
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          routes {
```

```

    }
  }
}
bridge-table {
  vlan-aware-bundle-eth-tag 1
}

```

4.4.5.2 Displaying the Ethernet Tag ID for VLAN-aware bundle services

Procedure

Use the **show network-instance** command to display the value configured for the **vlan-aware-bundle-eth-tag** parameter.

Example

```

--{ * candidate shared default }--[ ]--
# show network-instance mac_vrf_1 protocols bgp-evpn bgp-instance 1
=====
Net Instance   : mac_vrf_1
  bgp Instance 1 is enabled and up
-----
VXLAN-Interface : vxlan1.1
evi              : 1
ecmp             : 2
default-admin-tag : 0
oper-down-reason : N/A
EVPN Routes
  Next hop                : 10.20.1.3/32 (network-instance "default" system0.0
                           IPv4 address)
VLAN Aware Bundle Ethernet tag : 1
MAC/IP Routes            : enabled
IMET Routes              : enabled, originating-ip 10.20.1.3/32
=====

```

4.5 EVPN for Layer 2 ELINE services



Note: These services are currently supported exclusively on the 7730 SXR platform.

SR Linux supports EVPN-VPWS services under VPWS network instances where the network instance can only use a single subinterface and BGP-EVPN.

Two connection point objects and at least one bridged subinterface must be configured. Each subinterface configured must be associated with a different connection point.

When one subinterface and EVPN network are configured, they must be associated with different connection points. The **local-attachment-circuit** and **remote-attachment-circuit** parameters must be associated with the same connection point.

The **vpws-attachment-circuits** container hosts the specific EVPN-VPWS configuration, including the configuration and state for local and remote Ethernet tags. The **vpws-attachment-circuits** can only be configured using the **encapsulation-type mpls** command.

EVPN multihoming and VPWS network instances (single and all-active) are supported on SR Linux. The configuration of the Ethernet Segments (ESs), supported features, and association with interfaces follows what is described in [EVPN Layer 2 multihoming](#).

The main differences between multihoming in MAC-VRF and VPWS network instances are:

- VPWS services a backup PE in single-active mode, which is the next best PE after the elected DF. Remote nodes forward VPWS traffic to the primary PE (which signals flag P=1), and immediately switch over to the backup PE (which signals flag B=1) upon a failure on the primary PE.
- For MAC-VRF services, the primary/backup concept is not implemented. Upon a failure on the primary PE, the backup PE does not wait for the activation timer to activate the interface.
- VPWS services do not use the ESI label, because traffic from a PE can never be forwarded back to the ES peer PE.
- All-active multihoming for VPWS services does not run DF election, however the router still shows only four candidates in the ES state.

Entropy labels are supported on EVPN-VPWS network instances, but configuration per network instance is not supported. The entropy label is automatically added if supported by the LSP. If the EVPN destination in the VPWS service uses multiple transport LSPs, and all support entropy labels, only the inner most level uses the entropy label.

The use of **control-word** is supported.

To resolve EVPN-MPLS routes to tunnels in the tunnel-table, the network-instance must be configured with:

- **encapsulation-type mpls**
- **mpls next-hop-resolution allowed-tunnel-types[X]**, where X is at least one of the tunnel types supported by SR Linux

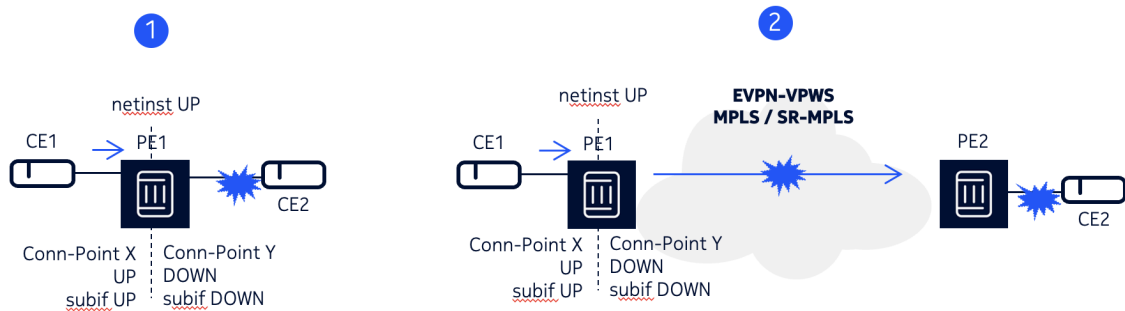
When EVPN routes are resolved to tunnels that support (underlay) ECMP, EVPN packets are sprayed over multiple paths for the tunnel.

4.5.1 Configuring EVPN-VPWS services

The following example configures a single-homing EVPN-VPWS service:

```
--{ * candidate shared default }--[ ]--
A:srl2# info network-instance VPWS-4
  network-instance VPWS-4 {
    type vpws
    description "VPWS-4 EVPN-MPLS single-homing"
    interface ethernet-1/1.4 {
      connection-point A
    }
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          admin-state enable
          encapsulation-type mpls
          evi 4
          ecmp 4
          vpws-attachment-circuits {
            local {
              local-attachment-circuit ac-4 {
                ethernet-tag 4
                connection-point B
              }
            }
          }
        }
      }
    }
  }
}
```


Figure 10: VPWS operational state



In case 1, PE1 is attached to a VPWS service with two subinterfaces. If the interface and/or subinterface CE2 goes down:

- Connection point Y goes operationally down.
- Connection point X and subinterface CE1 remain operationally up and the network instance remains operationally up.

In case 2, PE1 is attached to an EVPN-VPWS service with a subinterface and an EVPN destination. If the interface and/or subinterface CE2 goes down:

- EVPN destination on PE is removed and connection point Y goes operationally down.
- **bgp-evpn bgp-instance** remains operationally up.
- Connection point X and subinterface CE1 remain operationally up and the network instance remains operationally up. This changes when **link-loss-forwarding true** is enabled. For more information, see [Interface/LAG link loss forwarding](#).

In a third example (not shown in the figure), the two connection points can go operationally down but the network instance remains operationally up.

4.5.3 Layer 2 attribute mismatch

A Layer 2 attribute mismatch occurs when an AD per EVI route is imported into a VPWS service and the signaled layer 2 attributes (Layer 2 MTU, control word, or flow label) do not match the local settings.

In general, a Layer 2 attribute mismatch causes the `evpn_mgr` to discard the AD per EVI route and remove the EVPN destination from the PE that advertised the route.

The received Layer 2 MTU must match the **oper-`vpws-mtu`** so that the EVPN destination can be created. If the received MTU value is zero, the value is ignored and the destination is created if the rest of the attributes are valid.

A mismatch on the C flag (the flag that indicates the capability to use **control-word**) removes the EVPN destination. Even if the received C flag matches the local setting, if packets are received without the **control-word** and the **control-word** is configured, the data path discards the packets. Packets are also dropped if the **control-word** is received but **control-word false** is configured.

A mismatch on the F flag (flow label) also removes the EVPN destination.

4.5.4 Flow-aware transport label for EVPN-VPWS

Flow-aware transport (FAT) labels (also known as hash labels or flow labels) add entropy to divisible flows and create ECMP load balancing in a network. A FAT label can be used by EVPN services as per *draft-ietf-bess-rfc7432bis*.

SR Linux supports FAT labels for EVPN-VPWS services. When configured, FAT label support is always signaled as a new flag in the Layer 2 attributes extended community (F flag).

When **flow-label** is configured to be true, the EVPN AD per EVI route for the VPWS service is signaled with the F flag set and the local node expects the peer route to be received with F=1. If there is a mismatch in the F flag value between peers, the EVPN destination is removed.

Example: FAT label configuration

The following example shows a FAT label configuration:

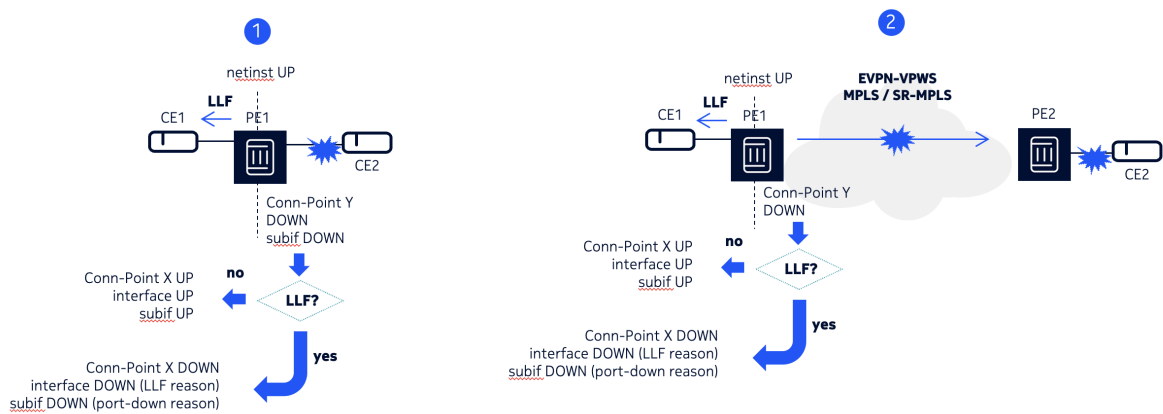
```
--{ * candidate shared default }--[ network-instance mgmt ]--
A:srl2# info
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        mpls {
          flow-label true
        }
      }
    }
  }
}
```

4.5.5 Interface/LAG link loss forwarding

Interface/LAG link loss forwarding (LLF) covers the propagation of a fault on an EVPN-VPWS service. The removal of all the AD per EVI routes from a remote PE triggers a fault propagation event on the local PE to the local CE. That fault propagation can use power-off or LACP out-of-synch signaling.

As illustrated in the following figure, LLF works in VPWS network instances irrespective of the connection points configured with two subinterfaces, one subinterface, and BGP-EVPN (or a pseudowire, as described [Pseudowires in VPWS network instances](#)).

Figure 11: LLF in VPWS network instances



In the cases 1 and 2 in the preceding figure, upon a failure that causes connection point Y to go operationally down, the system checks the configuration of the interface on connection point X. If **interface ethernet link-loss-forwarding true** is configured in the interface of connection point X, the **interface ethernet standby-signaling** action is triggered. Otherwise, connection point X and its subinterface remain operationally up.

In case 2, where BGP-EVPN is configured, when the subinterface connected to CE1 is operationally down due to active LLF, the EVPN AD per EVI route is still advertised by PE1.

LLF does not introduce any **oper-down-reason** at connection point or subinterface levels. LLF introduces **port-oper-per-down-reason link-loss-forwarding** at the interface level.

Referring to the preceding figure, suppose connection point Y goes operationally down and LLF is configured to be true. In this case the:

- connection point X goes operationally down with the **oper-down-reason** being **associations-oper-down**.
- subinterface to CE1 goes operationally down with the **oper-down-reason** being **port-down**.
- interface goes operationally down with the **oper-down-reason** being **link-loss-forwarding**.

LLF is compatible with ESs in all-active or single-active mode.

Configuring link loss forwarding

LLF is enabled by the **interface ethernet link-loss-forwarding true** command on the interface connected to the CE over which the fault signaling is propagated. In addition, the command **standby-signaling power-off lacp** must be configured on the same interface so that `ch_mgr` can trigger the relevant standby signaling to the CE.

```
--{ * candidate shared default }--[ interface ethernet-1/1 ]--
A:srl# info
  ethernet {
    standby-signaling power-off
    link-loss-forwarding true
  }
```

4.6 EVPN for Layer 3



Note: The components of EVPN Layer 3 are supported 7720 IXR-Dx, 7250 IXR, and 7730 SXR platforms.



Note: EVPN-VXLAN is currently supported exclusively on 7220 IXR-Dx platforms.



Note: EVPN-MPLS is supported on 7250 IXR and 7730 SXR platforms.

This chapter describes the components of EVPN Layer 3 on SR Linux.

- [EVPN Layer 3 basic configuration](#)
- [Anycast gateways](#)
- [EVPN Layer 3 multihoming and anycast gateways](#)
- [EVPN Layer 3 host route mobility](#)
- [EVPN-VXLAN IFL interoperability with EVPN IFF](#)
- [VIP discovery for redundant servers](#)
- [Layer 3 proxy-ARP/ND](#)
- [EVPN IP aliasing](#)
- [Unequal ECMP for EVPN IP prefix routes](#)
- [Combined ECMP for BGP PE-CE and EVPN IFL/IP-VPN](#)

4.6.1 EVPN Layer 3 basic configuration

The basic EVPN Layer 3 configuration model builds on the model for EVPN routes described in [EVPN for Layer 2 ELAN services](#), extending it with an additional route type to support inter-subnet forwarding: EVPN IP prefix route (or type 5, RT5).

The EVPN IP prefix route conveys IP prefixes of any length and family (IPv4 or IPv6) that need to be installed in the ip-vrfs of remote leaf nodes. The EVPN Layer 3 configuration model has two modes of operation:

- **asymmetric IRB**

This is a basic mode of operation EVPN Layer 3 using IRB interfaces. The term asymmetric refers to how there are more lookups performed at the ingress leaf than at the egress leaf.

While the asymmetric model allows inter-subnet-forwarding in EVPN-VXLAN networks in a very simple way, it requires the instantiation of all the MAC-VRFs of all the tenant subnets on all the leafs attached to the tenant. Because all the MAC-VRFs of the tenant are instantiated, FDB and ARP entries are consumed for all the hosts in all the leafs of the tenant.

These scale implications may make the symmetric model a better choice for data center deployment.

- **symmetric IRB**

The term symmetric refers to how MAC and IP lookups are needed at ingress, and MAC and IP lookups are performed at egress. As opposed to asymmetric, symmetric IRB implies the same number of lookups at ingress and egress.

SR Linux support for symmetric IRB includes the prefix routing model using RT5s as in RFC 9136, including the interface-less IP-VRF-to-IP-VRF model (EVPN IFL model).

Compared to the asymmetric model, the symmetric model scales better because hosts' ARP and FDB entries are installed only on the directly attached leaves and not on all the leaves of the tenant.

The following sections show asymmetric and symmetric IFL forwarding configurations.

4.6.1.1 Asymmetric IRB



Note: EVPN-VXLAN is currently supported exclusively on 7220 IXR-Dx platforms.

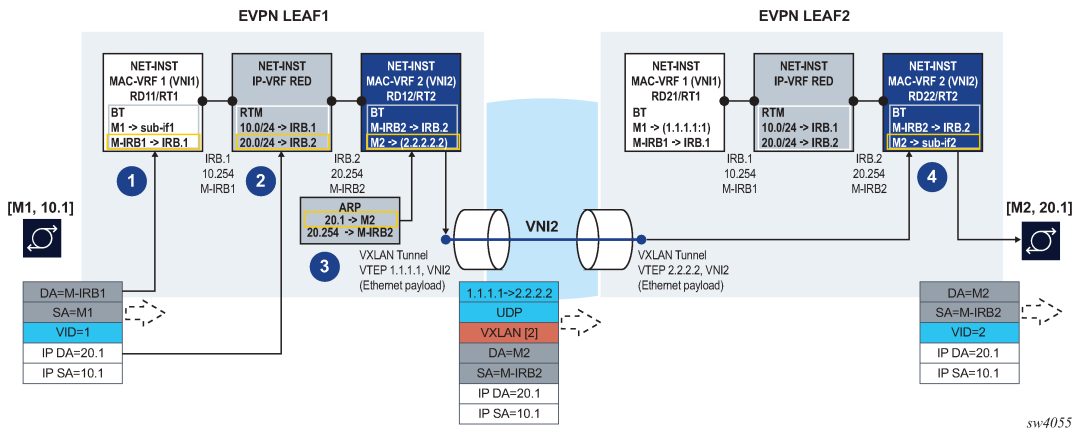


Note: EVPN-MPLS is supported on 7250 IXR and 7730 SXR platforms.

The asymmetric IRB model is the basic Layer 3 forwarding model when the IP-VRF (or default network-instance) interfaces are all IRB-based. The asymmetric model assumes that all the subnets of a tenant are local in the IP-VRF/default route table, so there is no need to advertise EVPN RT5 routes.

The following figure shows the asymmetric IRB model for EVPN-VXLAN services. This model is also applicable to EVPN-MPLS services, where VXLAN encapsulation is replaced with MPLS encapsulation for Ethernet frames.

Figure 12: EVPN-VXLAN Layer 3 asymmetric forwarding



In this example, the host with IP address 10.1 (abbreviation of 10.0.0.1) sends a unicast packet with destination 20.1 (a host in a different subnet and remote leaf). Because the IP DA is in a remote subnet, the MAC DA is resolved to the local default gateway MAC, M-IRB1. The frame is classified for MAC lookup on MAC-VRF 1, and the result is IRB.1, which indicates that an IP DA lookup is required in IP-VRF red.

An IP DA longest-prefix match in the route table yields IRB.2, a local IRB interface, so an ARP and MAC DA lookup are required in the corresponding IRB interface and MAC-VRF bridge table.

The ARP lookup yields M2 on MAC-VRF 2, and the M2 lookup yields the EVPN destination [VTEP:VNI]=[2.2.2.2:2]. For EVPN-MPLS, this destination would be [TEP:label]. The routed packet is

encapsulated with the corresponding inner MAC header and VXLAN or MPLS encapsulation before being sent to the wire.

In the asymmetric IRB model, if the ingress leaf routes the traffic via the IRB to a local subnet, and the destination MAC is aliased to multiple leaf nodes in the same ES destination, SR Linux can do load balancing on a per-flow basis.

EVPN Leaf 1 in [Figure 12: EVPN-VXLAN Layer 3 asymmetric forwarding](#) has the following configuration:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info
interface ethernet-1/2 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    type bridged
    admin-state enable
    vlan {
      encap {
        single-tagged {
          vlan-id 1
        }
      }
    }
  }
}
interface irb0 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 10.0.0.254/24 {
        anycast-gw true
      }
    }
    anycast-gw {
    }
  }
  subinterface 2 {
    ipv4 {
      admin-state enable
      address 20.0.0.254/24 {
        anycast-gw true
      }
    }
    anycast-gw {
    }
  }
}
network-instance ip-vrf-red {
  type ip-vrf
  interface irb0.1 {
  }
  interface irb0.2 {
  }
}
network-instance mac-vrf-1 {
  type mac-vrf
  admin-state enable
  interface ethernet-1/2.1 {
  }
  interface irb0.1 {
  }
  vxlan-interface vxlan1.1 {
```

```

    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.1
                evi 1
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
network-instance mac-vrf-2 {
    type mac-vrf
    admin-state enable
    interface irb0.2 {
    }
    vxlan-interface vxlan1.2 {
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.2
                evi 2
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
tunnel-interface vxlan1 {
    vxlan-interface 1 {
        type bridged
        ingress {
            vni 1
        }
    }
    vxlan-interface 2 {
        type bridged
        ingress {
            vni 2
        }
    }
}
}

```

EVPN Leaf 2 in [Figure 12: EVPN-VXLAN Layer 3 asymmetric forwarding](#) has the following configuration:

```

--{ [FACTORY] + candidate shared default }--[ ]--
A:LEAF2# info
interface ethernet-1/12 {
    admin-state enable
    vlan-tagging true
    subinterface 1 {
        type bridged
        admin-state enable
        vlan {
            encap {

```

```
        single-tagged {
            vlan-id 2
        }
    }
}
interface irb0 {
    subinterface 1 {
        ipv4 {
            admin-state enable
            address 10.0.0.254/24 {
                anycast-gw true
            }
        }
        anycast-gw {
        }
    }
    subinterface 2 {
        ipv4 {
            admin-state enable
            address 20.0.0.254/24 {
                anycast-gw true
            }
        }
        anycast-gw {
        }
    }
}
network-instance ip-vrf-red {
    type ip-vrf
    interface irb0.1 {
    }
    interface irb0.2 {
    }
}
network-instance mac-vrf-1 {
    type mac-vrf
    admin-state enable
    interface irb0.1 {
    }
    vxlan-interface vxlan1.1 {
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.1
                evi 1
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
network-instance mac-vrf-2 {
    type mac-vrf
    admin-state enable
    interface irb0.2 {
    }
    vxlan-interface vxlan1.2 {
    }
}
```

```

    protocols {
      bgp-evpn {
        bgp-instance 1 {
          admin-state enable
          vxlan-interface vxlan1.2
          evi 2
        }
      }
      bgp-vpn {
        bgp-instance 1 {
        }
      }
    }
  }
  tunnel-interface vxlan1 {
    vxlan-interface 1 {
      type bridged
      ingress {
        vni 1
      }
    }
    vxlan-interface 2 {
      type bridged
      ingress {
        vni 2
      }
    }
  }
}

```

When EVPN-MPLS is used instead of EVPN-VXLAN, the MAC-VRF network instances replace their VXLAN configuration with the configuration for MPLS resolution. For example, `mac-vrf-1` on Leaf 1 would have the following configuration:

```

--{ [FACTORY] + candidate shared default }--[ ]--
# info
<snip>
  network-instance mac-vrf-1 {
    type mac-vrf
    admin-state enable
    interface ethernet-1/2.1 {
    }
    interface irb0.1 {
    }
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          admin-state enable
          evi 1
          mpls {
            next-hop-resolution {
              allowed-tunnel-types [
                ldp
                sr-isis
              ]
            }
          }
        }
      }
      bgp-vpn {
      }
    }
  }
<snip>

```

4.6.1.2 Symmetric IRB interface-less ip-vrf-to-ip-vrf model



Note: EVPN-VXLAN is currently supported exclusively on 7220 IXR-Dx platforms.



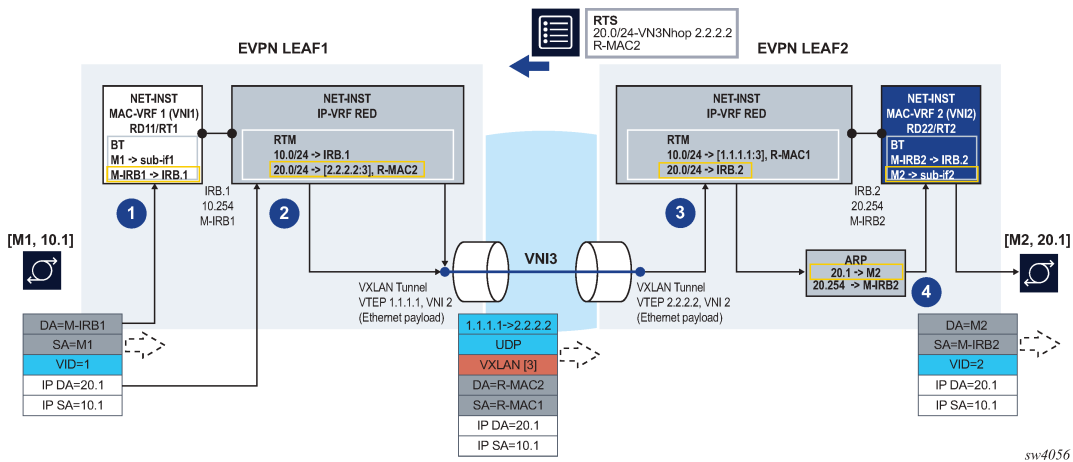
Note: EVPN-MPLS is supported on 7250 IXR and 7730 SXR platforms.

SR Linux support for symmetric IRB is based on the prefix routing model using RT5s, and implements the EVPN interface-less (EVPN IFL) IP-VRF-to-IP-VRF model.

In the EVPN IFL model, all interface and local routes (static, ARP-ND, BGP, and so on) are automatically advertised in RT5s without the need for any export policy. Interface host and broadcast addresses are not advertised. On the ingress PE, RT5s are installed in the route table as indirect with owner BGP-EVPN.

The following figure shows the forwarding for symmetric IRB for EVPN-VXLAN services.

Figure 13: EVPN-VXLAN Layer 3 symmetric forwarding



As in the asymmetric model, the frame is classified for bridge-table lookup on the MAC-VRF and processed for routing in IP-VRF red.

In contrast to the asymmetric model, a longest prefix match does not yield a local subnet, but a remote subnet reachable via [VTEP:VNI]=[2.2.2.2:3] for EVPN-VXLAN and [TEP:label] for EVPN-MPLS, and inner MAC DA R-MAC2. SR Linux supports the EVPN interface-less (EVPN IFL) model, so that information is found in the IP-VRF route-table directly; a route lookup on the IP-VRF red route-table yields a VXLAN tunnel and VNI for EVPN-VXLAN and an MPLS tunnel and service label for EVPN-MPLS.

In the case of EVPN-VXLAN, as shown in [Figure 13: EVPN-VXLAN Layer 3 symmetric forwarding](#):

- Packets are encapsulated with an inner Ethernet header and the VXLAN tunnel encapsulation.
- The inner Ethernet header uses the system-mac as MAC SA, and the MAC advertised along with the received RT5 as MAC DA. No VLAN tag is transmitted or received in this inner Ethernet header.

At the egress PE, the packet is classified for an IP lookup on the IP-VRF red (the inner Ethernet header is ignored).

The inner and outer IP headers are updated as follows:

- inner IP header TTL is decremented

- outer IP header TTL is set to 255
- outer DSCP value is marked as described in [QoS for VXLAN tunnels](#)
- no IP MTU check is performed before or after encapsulation

Because SR Linux supports EVPN IFL, the IP lookup in the IP-VRF red route-table yields a local IRB interface.

Subsequent ARP and MAC lookups provide the information to send the routed frame to subinterface 2.

EVPN Leaf 1 in [Figure 13: EVPN-VXLAN Layer 3 symmetric forwarding](#) has the following configuration:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info
interface ethernet-1/2 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    type bridged
    admin-state enable
    vlan {
      encaps {
        single-tagged {
          vlan-id 1
        }
      }
    }
  }
}
interface irb0 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 10.0.0.254/24 {
        anycast-gw true
      }
    }
    anycast-gw {
    }
  }
}
network-instance ip-vrf-red {
  type ip-vrf
  interface irb0.1 {
  }
  vxlan-interface vxlan1.3 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.3
        evi 3
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
network-instance mac-vrf-1 {
  type mac-vrf
  admin-state enable
}
```

```

interface ethernet-1/2.1 {
}
interface irb0.1 {
}
vxlan-interface vxlan1.1 {
}
protocols {
  bgp-evpn {
    bgp-instance 1 {
      admin-state enable
      vxlan-interface vxlan1.1
      evi 1
    }
  }
  bgp-vpn {
    bgp-instance 1 {
    }
  }
}
}
tunnel-interface vxlan1 {
  vxlan-interface 1 {
    type bridged
    ingress {
      vni 1
    }
  }
  vxlan-interface 3 {
    type routed
    ingress {
      vni 3
    }
  }
}
}

```

EVPN Leaf 2 in [Figure 13: EVPN-VXLAN Layer 3 symmetric forwarding](#) has the following configuration:

```

--{ [FACTORY] + candidate shared default }--[ ]--
# info
interface ethernet-1/12 {
  admin-state enable
  vlan-tagging true
  subinterface 2 {
    type bridged
    admin-state enable
    vlan {
      encaps {
        single-tagged {
          vlan-id 2
        }
      }
    }
  }
}
interface irb0 {
  subinterface 2 {
    ipv4 {
      admin-state enable
      address 20.0.0.254/24 {
        anycast-gw true
      }
    }
  }
  anycast-gw {

```

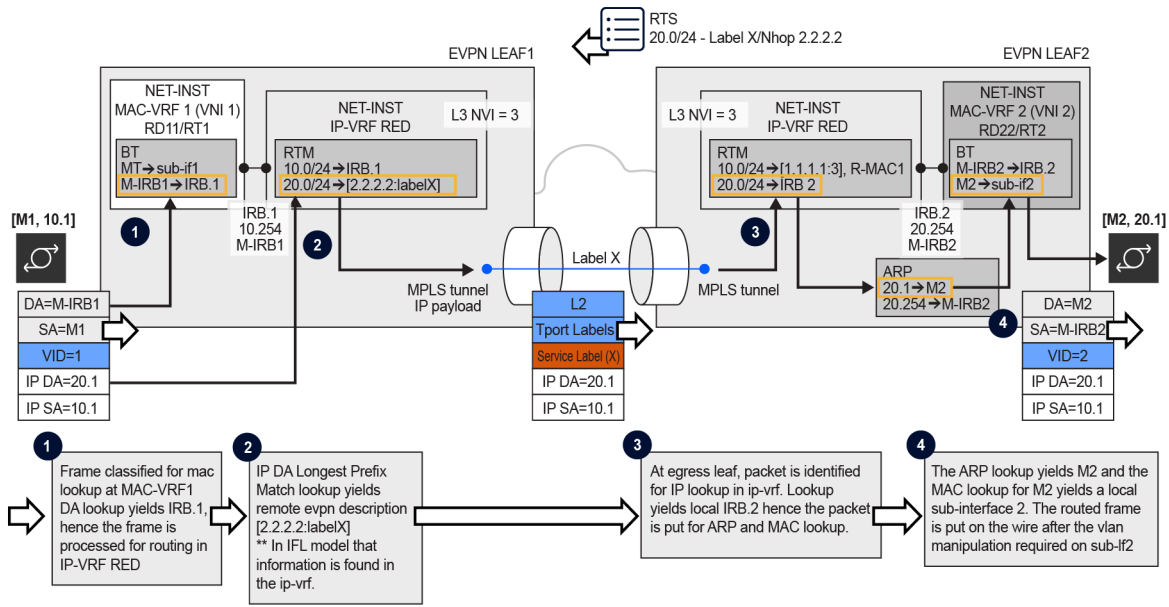
```

    }
  }
}
network-instance ip-vrf-red {
  type ip-vrf
  interface irb0.2 {
  }
  vxlan-interface vxlan1.3 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.3
        evi 3
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
network-instance mac-vrf-2 {
  type mac-vrf
  admin-state enable
  interface ethernet-1/12.2 {
  }
  interface irb0.2 {
  }
  vxlan-interface vxlan1.2 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.2
        evi 2
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
tunnel-interface vxlan1 {
  vxlan-interface 2 {
    type bridged
    ingress {
      vni 2
    }
  }
  vxlan-interface 3 {
    type routed
    ingress {
      vni 3
    }
  }
}
}

```

In the case of EVPN-MPLS, MPLS tunnels exist between Leaf 1 and Leaf 2, instead of VXLAN. The following figure shows the forwarding for symmetric IRB for EVPN-MPLS services.

Figure 14: EVPN-MPLS Layer 3 symmetric forwarding



hw4244

The following characteristics apply in EVPN-MPLS encapsulation:

- To resolve EVPN-MPLS routes to tunnels in the tunnel table, the network instance is configured with **encapsulation-type mpls** and at least one MPLS tunnel type with the **bgp-evpn bgp-instance mpls next-hop-resolution allowed-tunnel-types [tunnel-types]** command.
- Received EVPN-MPLS next hops are resolved to the best tunnels in the allowed list.
- In order for MPLS to support the encapsulation of IP packets directly, packets are not encapsulated with an inner Ethernet header (as opposed to VXLAN which always requires inner Ethernet encapsulation).
- At the egress PE (shown as Leaf 2 in the preceding figure) and based on an ILM lookup on Label X, the packet is classified for an IP lookup on the `ip-vrf red` network instance.
- Subsequent ARP and MAC lookups provide the information required to send the routed packet to subinterface 2.

EVPN Leaf 1 in Figure 14: EVPN-MPLS Layer 3 symmetric forwarding has the following configuration.

```
--{ [FACTORY] + candidate shared default }-- [ ]--
A:LEAF1# info
  interface ethernet-1/2 {
    admin-state enable
    vlan-tagging true
    subinterface 1 {
      type bridged
      admin-state enable
      vlan {
        encap {
          single-tagged {
            vlan-id 1
          }
        }
      }
    }
  }
}
```

```

}
interface irb0 {
  subinterface 1 {
    ipv4 {
      address 10.0.0.254/24 {
        anycast-gw true
      }
    }
    anycast-gw {
    }
  }
}
network-instance ip-vrf-red {
  type ip-vrf
  interface irb0.1 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        encapsulation-type mpls
        evi 3
        mpls {
          next-hop-resolution {
            allowed-tunnel-types [ldp sr-isis]
          }
        }
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
network-instance mac-vrf-1 {
  type mac-vrf
  admin-state enable
  interface ethernet-1/2.1 {
  }
  interface irb0.1 {
  }
}
}

```

EVPN Leaf 2 in [Figure 14: EVPN-MPLS Layer 3 symmetric forwarding](#) has the following configuration.

```

- { [FACTORY] + candidate shared default }--[ ]--
A:LEAF2# info
  interface ethernet-1/12 {
    admin-state enable
    vlan-tagging true
    subinterface 2 {
      type bridged
      admin-state enable
      vlan {
        encap {
          single-tagged {
            vlan-id 2
          }
        }
      }
    }
  }
}

```

```

interface irb0 {
  subinterface 2 {
    ipv4 {
      address 20.0.0.254/24 {
        anycast-gw true
      }
    }
    anycast-gw {
    }
  }
}
network-instance ip-vrf-red {
  type ip-vrf
  interface irb0.2 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        encapsulation-type mpls
        evi 3
        mpls {
          next-hop-resolution {
            allowed-tunnel-types [ldp sr-isis]
          }
        }
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
network-instance mac-vrf-2 {
  type mac-vrf
  admin-state enable
  interface ethernet-1/12.2 {
  }
  interface irb0.2 {
  }
}

```



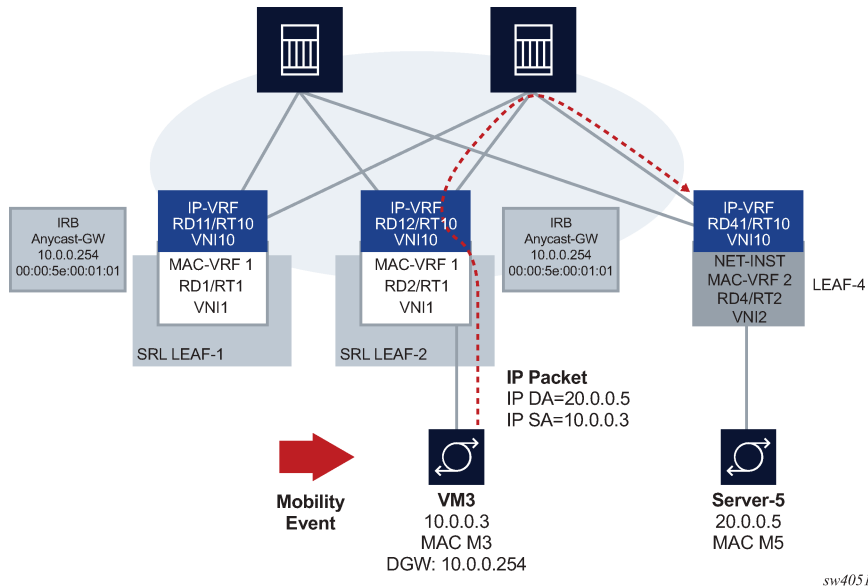
Note: Though it is not shown in this example, MAC-VRFs in Leaf 1 and Leaf 2 can also be EVPN-MPLS enabled.

4.6.2 Anycast gateways

Anycast gateways (anycast-GWs) are a common way to configure IRB subinterfaces in DC leaf nodes. Configuring anycast-GW IRB subinterfaces on all leaf nodes of the same BD avoids tromboning for upstream traffic from hosts moving between leaf nodes.

The following figure shows an example anycast-GW IRB configuration.

Figure 15: Anycast-GW IRB configuration



Anycast-GWs allow configuration of the same IP and MAC addresses on the IRB interfaces of all leaf switches attached to the same BD; for example, SRL LEAF-1 and SRL LEAF-2 in the preceding figure. This feature optimizes the south-north forwarding because a host's default gateway always belongs to the connected leaf, irrespective of the host moving between leaf switches; for example VM3 moving from SRL LEAF-1 to SRL LEAF-2 in the figure.

When an IRB subinterface is configured as an anycast-GW, it must have one IP address configured as **anycast-gw**. The subinterface may or may not have other non-anycast-GW IP addresses configured.

To simplify provisioning, an option to automatically derive the anycast-GW MAC is supported, as described in RFC 9135. The auto-derivation uses a virtual-router-id similar to MAC auto-derivation in RFC 5798 (VRRP). Anycast-GWs use a default virtual-router-id of 01, if not explicitly configured. Because only one anycast-gw-mac per IRB sub-interface is supported, the anycast-gw-mac for IPv4 and IPv6 is the same in the IRB sub-interface.

The following is an example configuration for an anycast-GW subinterface:

```
// Configuration Example of an anycast-gw IRB sub-interface

[interface irb1 subinterface 1 ]
A:leaf-1/2# info
  ipv4 {
    admin-state enable
    address 10.0.0.254/24 {
      primary true
      anycast-gw true
    }
  }
  anycast-gw {
    virtual-router-id 2
  }

// State Example of an anycast-gw IRB sub-interface

[interface irb1 subinterface 1 ]
A:leaf-1/2# info from state
```

```
ipv4 {
  address 10.0.0.254/24 {
    primary true
    anycast-gw true
  }
}
anycast-gw {
  virtual-router-id 2
  anycast-gw-mac 00:00:5e:00:01:02
  anycast-gw-mac-origin auto-derived
}
```

The **anycast-gw true** command designates the associated IP address as an anycast-GW address of the subinterface and associates the IP address with the **anycast-gw-mac** address in the same sub-interface. ARP requests or NSs received for the anycast-GW IP address are replied using the **anycast-gw-mac** address, as opposed to the regular system-derived MAC address. Similarly, CPM-originated GARPs or unsolicited NAs sent for the anycast-GW IP address use the **anycast-gw-mac** address as well. Packets routed to the IRB use the **anycast-gw-mac** as the SA in Ethernet header.

All IP addresses of the IRB subinterface and their associated MACs are advertised in MAC/IP routes with the static flag set. The non-anycast-GW IPs are advertised along with the interface hardware MAC address, and the anycast-GW IP addresses along with the anycast-gw-mac address.

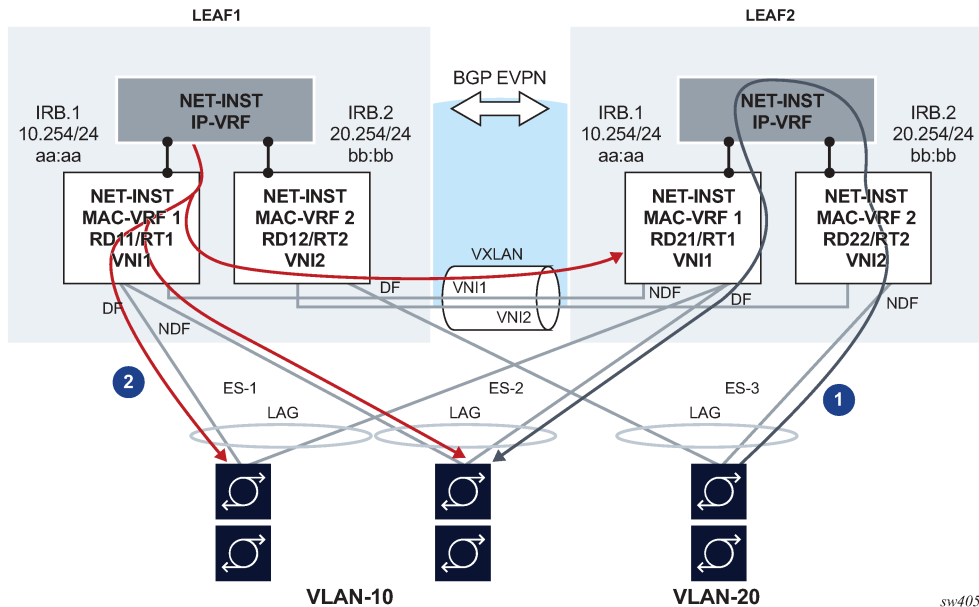
In addition, the **anycast-gw true** command makes the system skip the ARP/ND duplicate-address-detection procedures for the anycast-GW IP address.

4.6.3 EVPN Layer 3 multihoming and anycast gateways

In an EVPN Layer 3 scenario, all IRB interfaces facing the hosts must have the same IP address and MAC; that is, an anycast-GW configuration. This avoids inefficiencies for all-active multihoming or speeds up convergence for host mobility.

The use of anycast-GW along with all-active multihoming in an EVPN-VXLAN scenario is shown in the following figure:

Figure 16: EVPN-VXLAN Layer 3 model – multihoming and anycast gateway



In this example:

1. Routed unicast traffic is always routed to the directly connected leaf (no tromboning).
2. BUM traffic sent from the IRB interface is sent to all DF and NDF subinterfaces, similar to BUM entering a subinterface.

This applies to:

- system-generated unknown unicast or broadcast traffic
- ARP requests, GARPs, and Neighbor Discovery multicast messages
- unicast traffic with unknown MAC DA

This solution also applies when EVPN-MPLS is used on MAC-VRFs instead of VXLAN. Flooded traffic coming from the IRB interface to the MAC-VRF is flooded to local DF subinterfaces and to other leaf nodes attached to the same broadcast domain.

When a host connected to ES-3 sends a unicast flow to be routed in the IP-VRF, the flow must be routed in the leaf receiving the traffic, irrespective of the server hashing the flow to Leaf-1 or Leaf-2. To do this, the host is configured with only one default gateway, 20.254/24. When the host ARPs for it, it does not matter if the ARP request is sent to Leaf-1 or Leaf-2. Either leaf replies with the same anycast-GW MAC, and when receiving the traffic either leaf can route the packet.

This scenario is supported on IP-VRF network instances and the default network instance.



Note:

When IRB subinterfaces are attached to MAC-VRF network instances with all-active multihoming ESs, set the `arp timeout / neighbor-discovery stale-time` settings on the IRB subinterface to a value that is 30 seconds lower than the `age-time` configured in the MAC-VRF. This setting avoids transient packet loss situations triggered by the MAC address of an active ARP/ND entry being removed from the MAC table.



Note: To avoid false MAC duplicates in scenarios with IRB subinterfaces and all-active multihoming (if the ARP/ND entries are learned earlier than the corresponding MAC entries), the configuration of **bgp-evpn bgp-instance routes bridge-table mac-ip advertise-arp-nd-only-with-mac-table-entry true** is highly recommended.

4.6.4 EVPN Layer 3 host route mobility

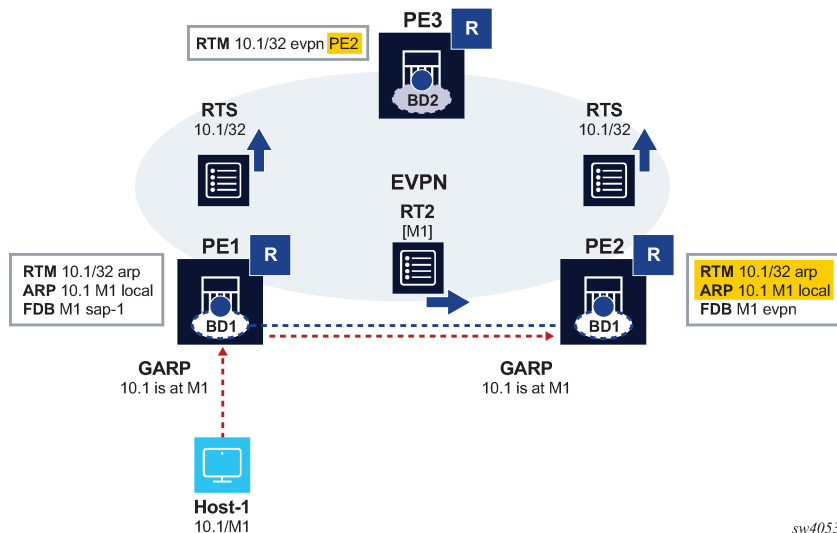
EVPN host route mobility refers to the procedures that allow the following:

- learning ARP/ND entries out of unsolicited messages from hosts
- generating host routes out of those ARP/ND entries
- refreshing the entries when mobility events occur within the same BD

EVPN host route mobility is part of basic EVPN Layer 3 functionality as defined in [RFC 9135](#).

The following figure EVPN host route mobility.

Figure 17: EVPN host route mobility



In [Figure 17: EVPN host route mobility](#) a host is attached to PE1, so all traffic from PE3 to that host must be forwarded directly to PE1. When the host moves to PE2, all the tables must be immediately updated so that PE3 sends the traffic to PE2. EVPN host route mobility works by doing the following:

1. snooping and learning ARP/ND entries for hosts upon receiving unsolicited GARP or NA messages
2. creating host routes out of those dynamic ARP/ND entries. These routes only exist in the control plane and are not installed in the forwarding plane to avoid FIB exhaustion with too many /32 or /128 host routes.
3. advertising the locally learned ARP/ND entries in MAC/IP routes so that ARP/ND caches can be synchronized across leaf nodes
4. advertising the host routes as IP prefix routes
5. triggering ARP/ND refresh messages for changes in the ARP/ND table or MAC table for a specific IP, which allows updating the tables without depending on the ARP/ND aging timers (which can be hours long)

The following configuration enables an anycast-GW IRB subinterface to support mobility procedures:

```
// Example of the configuration of host route mobility features

--{ candidate shared default }--[ interface irb1 subinterface 1 ]--
A:--# info
  ipv4 {
    admin-state enable
    address 10.0.0.254/24 {
      anycast-gw true
      primary
    }
    arp {
      learn-unsolicited true
      host-route {
        populate static
        populate dynamic
      }
      evpn {
        advertise static
        advertise dynamic
      }
    }
  }
  ipv6 {
    admin-state enable
    address 200::254/64 {
      anycast-gw true
      primary
    }
    neighbor-discovery {
      learn-unsolicited true
      host-route {
        populate static
        populate dynamic
      }
    }
    evpn {
      advertise static
      advertise dynamic
    }
  }
  anycast-gw {
  }
--{ candidate shared default }--[ interface irb1 subinterface 1 ]--
# info from state
  admin-state enable
  ip-mtu 1500
  name irb1.1
  ifindex 1082146818
  oper-state up
  last-change "a minute ago"
  ipv4 {
    allow-directed-broadcast false
    address 10.0.0.254/24 {
      anycast-gw true
      origin static
    }
  }
  arp {
    duplicate-address-detection true
    timeout 14400
    learn-unsolicited true
    host-route {
      populate static
    }
  }
}
```

```

                populate dynamic
            }
        }
    }
}
ipv6 {
    address 200::254/64 {
        anycast-gw true
        origin static
        status unknown
    }
    address fe80::201:1ff:feff:42/64 {
        origin link-layer
        status unknown
    }
    neighbor-discovery {
        duplicate-address-detection true
        reachable-time 30
        stale-time 14400
        learn-unsolicited both
        host-route {
            populate static
            populate dynamic
        }
    }
    router-advertisement {
        router-role {
            current-hop-limit 64
            managed-configuration-flag false
            other-configuration-flag false
            max-advertisement-interval 600
            min-advertisement-interval 200
            reachable-time 0
            retransmit-time 0
            router-lifetime 1800
        }
    }
}
anycast-gw {
    virtual-router-id 1
    anycast-gw-mac 00:00:5E:00:01:01
    anycast-gw-mac-origin vrid-auto-derived
}
...

```

In this configuration, when `learn-unsolicited` is set to `true`, the node processes all solicited and unsolicited ARP/ND flooded messages received on subinterfaces (no VXLAN/MPLS) and learns the corresponding ARP/ND entries as dynamic. By default, this setting is `false`, so only solicited entries are learned by default.

The advertisement of EVPN MAC/IP routes for the learned ARP entries must be enabled/disabled by configuration; it is disabled by default. In the example above, this is configured with the `advertise dynamic` and `advertise static` settings.

The creation of host routes in the IP-VRF route table out of the dynamic or static ARP entries can be enabled/disabled by configuration; it is disabled by default. In the example above, this is configured with the `host-route populate dynamic` and `host-route populate static` settings.

By default, the ARP/ND host routes are not installed in the data path. You can configure `host-route populate dynamic datapath-programming true` to change this default behavior so that the ARP/

ND host routes are installed in the data path. This configuration is required when the ARP/ND host routes are leaked. See "Network instance route leaking" in the *SR Linux Configuration Basics Guide*.

The dynamic ARP entries are refreshed without any extra configuration. The system sends ARP requests for the dynamic entries to make sure the hosts are still alive and connected.

4.6.5 EVPN-VXLAN IFL interoperability with EVPN IFF

By default, the SR Linux EVPN IFL (interface-less) model, described in [Symmetric IRB interface-less ip-vrf-to-ip-vrf model](#) for VXLAN, does not interoperate with the EVPN IFF (interface-ful) model, as supported on, for example, Nuage WBX devices. However, it is possible to configure the SR Linux EVPN IFL model to interoperate with the EVPN IFF model.

To do this, configure the **advertise-gateway-mac** command for the IP-VRF network instance. When this command is configured, the node advertises a MAC/IP route using the following:

- gateway-mac for the IP-VRF (that is, the system-mac)
- RD/RT, next-hop, and VNI of the IP-VRF where the command is configured
- null IP address, ESI or Ethernet Tag ID

Nuage WBX devices support two EVPN Layer 3 IPv6 modes: IFF unnumbered and IFF numbered, as described in RFC 9136. The SR Linux interoperability mode enabled by the **advertise-gateway-mac** command only works with Nuage WBX devices that use the EVPN IFF unnumbered model. This is because the EVPN IFL and EVPN IFF unnumbered models both use the same format in the IP prefix route, and they differ only in the additional MAC/IP route for the gateway-mac. The EVPN IFL and EVPN IFF numbered models have different IP prefix route formats, so they cannot interoperate.

The following example enables interoperability with the Nuage EVPN IFF unnumbered model:

```
--{ [FACTORY] + candidate shared default }--[ ]--
# info from state network-instance protocols bgp-vpn
  bgp-evpn {
    bgp-instance 1 {
      admin-state enable
      vxlan-interface vxlan1.2
      routes {
        route-table {
          mac-ip {
            advertise-gateway-mac true
          }
        }
      }
    }
  }
}
```

4.6.6 VIP discovery for redundant servers

In data centers, it is common to have clusters of servers sharing the same IP address and working in an active-standby mode, so that only one of the servers in the cluster is active at a time. This shared IP address is known as a virtual IP (VIP) address. SR Linux can discover which server in the cluster owns the VIP address.

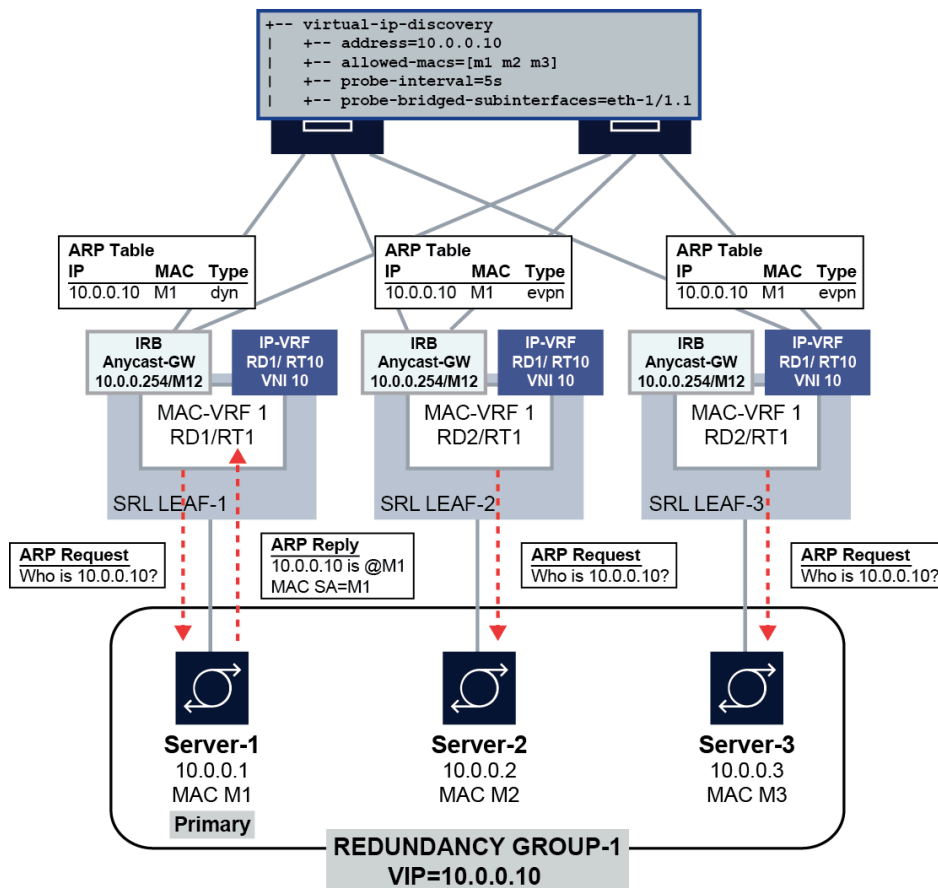
ARP requests (NS for IPv6) from a leaf node or gratuitous ARP (unsolicited NA for IPv6) from a server are used to discover which host owns the VIP. Among the servers sharing the VIP, only the active one

either sends a gratuitous ARP (or unsolicited NA) or replies to the ARP request (NS) from the leaf node. If a server does not send a gratuitous ARP, you can optionally configure SR Linux to send ARP requests periodically at a specified probe interval.

The leaf nodes create entries in their ARP tables that map the VIP to the MAC address of the active server. You can optionally configure a list of allowed MAC addresses so that the ARP/ND entry for the VIP is created only if the reply from the server comes from one of the MACs on the allowed list.

The following figure shows a configuration where three SR Linux devices in a MAC-VRF are connected to a redundant server group.

Figure 18: VIP discovery for the active server in a server group



In this configuration, the SR Linux devices (SRL LEAF-1/2/3) start sending ARP requests for the VIP 10.0.0.10 on the ethernet 1/1.1 subinterface every 5 seconds until the ARP entry for the VIP/VMAC is learned. The ARP entry is created only if the MAC address matches one of the allowed MACs: M1, M2, or M3.

Initially, Server-1 is the primary server in the server group, and it responds to the ARP request from Leaf-1, causing the server's MAC address M1 to be added to the ARP table on the SR Linux devices for VIP 10.0.0.10. If Server-2 (MAC M2) becomes the primary server, it sends out a gratuitous ARP message or responds to an ARP request from LEAF-2, which triggers an update of the ARP table on Leaf-2 and a MAC/IP route update with 10.0.0.10/M2. When Leaf-1 receives the update, it updates its ARP table immediately with the new entry.

VIP discovery is supported on single-homed and multihomed subinterfaces. For all-active multihoming, and based on the local-bias behavior, the ARP/ND probes are sent to all ES subinterfaces irrespective of their DF state. When the active server replies, the MAC/IP route is synchronized in the other ES peers based on the functionality described in [EVPN Layer 3 host route mobility](#).

4.6.6.1 Configuring VIP discovery

Procedure

To configure VIP discovery for an IRB interface, you specify the bridged subinterfaces that can be sent probe messages (ARP requests) for the VIP, optionally a probe interval that determines how often the probe messages are sent, and optionally a list of allowed MAC addresses that indicate the MAC addresses of the servers in the group.

Example: Configure VIP discovery for IPv4

```
--{ candidate shared default }--[ ]--
# info interface irb1 subinterface 1 ipv4 arp
interface irb1 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      arp {
        virtual-ipv4-discovery {
          address 10.10.10.10 {
            probe-bridged-subinterfaces [
              ethernet-1/1.1
            ]
            probe-interval 5
            allowed-macs [
              00:14:9C:78:E2:E1
              00:14:9C:78:E2:E2
              00:14:9C:78:E2:E3
            ]
          }
        }
      }
    }
  }
}
```

Example: Configure VIP discovery for IPv6

```
--{ candidate shared default }--[ ]--
# info interface irb1 subinterface 1 ipv6 neighbor-discovery
interface irb1 {
  subinterface 1 {
    ipv6 {
      admin-state enable
      neighbor-discovery {
        virtual-ipv6-discovery {
          address 2001::10:10:10:10 {
            probe-interval 5
            allowed-macs [
              00:14:9C:78:E2:E1
              00:14:9C:78:E2:E2
              00:14:9C:78:E2:E3
            ]
          }
        }
      }
      probe-bridged-subinterfaces [

```



```
arp {
  virtual-ipv4-discovery {
    address 10.10.10.10 {
      probe-bridged-subinterfaces [
        ethernet-1/1.1
      ]
      probe-interval 5
      allowed-macs [
        00:14:9C:78:E2:E1
        00:14:9C:78:E2:E2
        00:14:9C:78:E2:E3
      ]
      statistics {
        out-probe-packets 100
      }
      statistics {
        out-total-probe-packets 100
      }
    }
  }
}
```

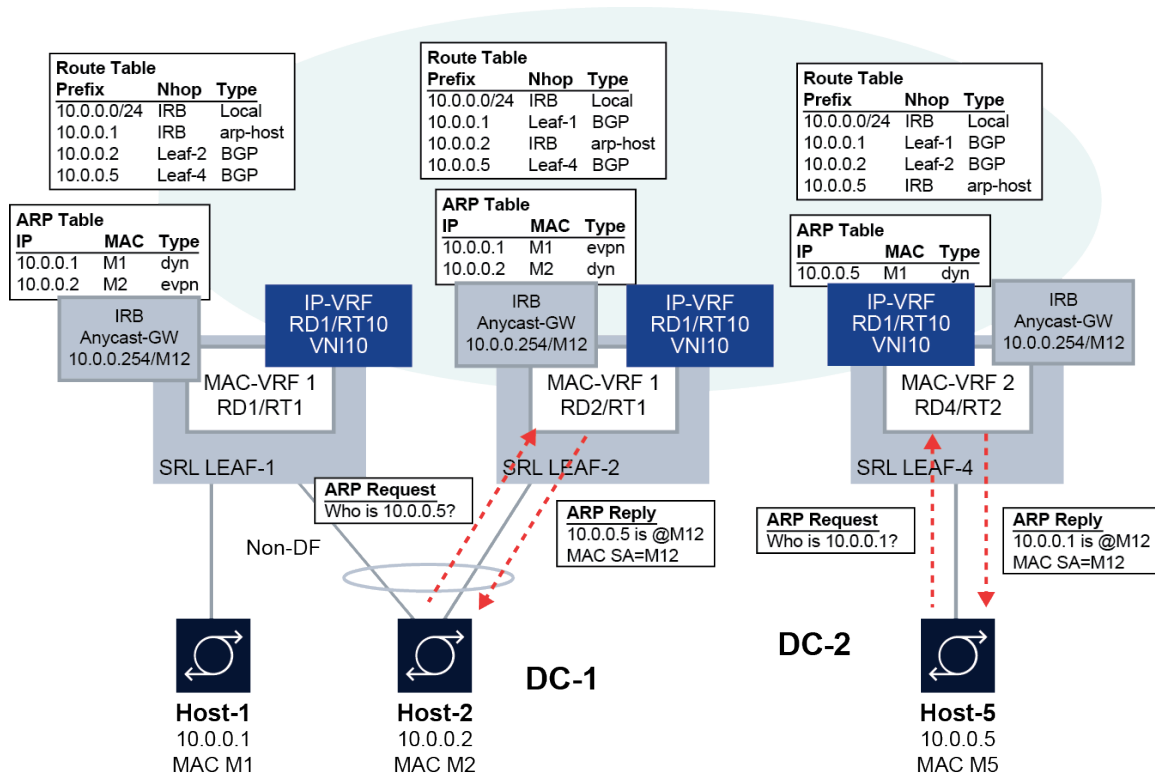
4.6.7 Layer 3 proxy-ARP/ND

Layer 3 proxy-ARP/ND configures the router to reply with its own MAC to ARP-requests and NS destined for any host. This feature is intended for the following use cases:

- Deployments that use a Layer 3 multihoming solution, but are unable to run EVPN, ESs or VXLAN/MPLS.
- Virtual subnet scenarios, where leaf nodes are attached to different broadcast domains (BDs), but the hosts connected to them are part of the same subnet. RFC 7814 describes this type of solution.

The following figure shows an example virtual subnet configuration that uses this feature.

Figure 19: Layer 3 Proxy ARP/ND



In this example, SR Linux Leaf-1/Leaf-2 and SR Linux-4 are attached to different BDs; however, the hosts connected to them are part of the same subnet. This type of configuration is typically used in scenarios where the BDs are kept deliberately small, while hosts of the same subnet can be part of more than one of the BDs. This is the case described in RFC 7814.

Layer 3 proxy-ARP/ND triggers the router to reply to any ARP-request or NS from hosts, so that the traffic is attracted to the IRB subinterface and forwarded by the router to a destination on the same subinterface from which the request came or a different subinterface or route.

Layer 3 proxy-ARP/ND is intended to be used along with the learn-unsolicited and host-route populate features, so that ARP/ND entries are generated and host routes created out of them and advertised by the routing protocol enabled in the routing network instance (default or IP-VRF).

This feature is supported on IRB and other Layer 3 subinterfaces. EVPN may or may not be enabled on the attached MAC-VRF. The subinterfaces can be attached to an IP-VRF or the default network instance.



Note: The virtual subnet expansion enabled by this feature only works for unicast IP traffic with TTL > 1 between hosts in different BDs.

4.6.7.1 Layer 3 proxy-ARP

When Layer 3 proxy-ARP is enabled, all ARP-requests are copied to the CPM and replied using the anycast-gw MAC, if configured; otherwise, the interface hw-mac-address MAC is used. This process is irrespective of the target IP address of the ARP-request being in the ARP table or in the route-table.

The system does not reply to gratuitous ARP (GARP) requests, where the ARP Sender Protocol Address and ARP Target Protocol Address are both set to the IP address of the cache entry to be updated.

When Layer 3 proxy-ARP is enabled, the router replies to any ARP-request unconditionally if received over a bridged subinterface. If the ARP-request is received over VXLAN/MPLS, the router replies only if the ARP-request is targeted to its own IP; if it is targeted to a non-local IP on the IRB, the ARP-request is flooded to local bridged subinterfaces, changing the source MAC and IP with the local ones.

If the original ARP-request came on a bridged subinterface, and the target IP has not been learned on the subinterface yet, the router sends an ARP-request with its own source information to get the target IP learned as soon as possible.

The enabled proxy-ARP behavior attracts all traffic from all the hosts attached to the IRB subinterface that sent an ARP-request for the destination IP address. As a result of this:

- If the received packets have an IP DA with an ARP entry on the same subinterface over which the ARP-request was received, the router forwards the packets to the next-hop indicated by the ARP entry.
- If the received packets have an IP DA with an ARP entry on a different subinterface, the router forwards the packets to the next-hop indicated by the ARP entry.
- If the received packets have an IP DA with a matching route in the route-table, the router forwards the packets based on the route-table entry.
- Otherwise, the attracted traffic is dropped.

The Layer 3 proxy-ARP feature does not require a route lookup before replying to the received ARP-requests. The received ARP-requests are only copied to CPM; they are not flooded to other objects in the MAC-VRF.

The ICMP redirects generated by XDP CPM are suppressed for IP packets targeting an IP address within the same subnet.

Received ARP-replies with a DA not equal to the local MAC, if any are received, are forwarded based on a mac-table lookup.



Note: The system verifies the source IP and replies regardless of the target IP (it floods to bridged subinterfaces if received on VXLAN/MPLS tunnels). The command `learn-unsolicited` is required to create an ARP entry.

4.6.7.1.1 Configuring Layer 3 proxy-ARP

Procedure

To configure Layer 3 proxy-ARP, enable it on a subinterface.

Example

```
--{ candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv4 arp proxy-arp
  interface ethernet-1/1 {
    subinterface 1 {
      ipv4 {
        admin-state enable
        arp {
          proxy-arp true
        }
      }
    }
  }
```

```
}

```

4.6.7.2 Layer 3 proxy-ND

When Layer 3 proxy-ND is enabled, it functions similarly to Layer 3 proxy-ARP. All NS messages sent to the Solicited-Node multicast address [ff02::1:ff][low-3-bytes-unicast-ip] are copied to CPM, and all of the NS messages are replied using the anycast-gw MAC, if configured (or interface hw-mac-address otherwise), without any route lookup.

Received NS messages are only copied to CPM and not flooded in the BD. In this way, the feature provides an implicit ND flood suppression.

The **learn-unsolicited** command is not needed to learn a neighbor from an NS, because it is basic ND behavior to create a neighbor for the SA of an NS if there is a reply and it is a valid neighbor.

4.6.7.2.1 Configuring Layer 3 proxy-ND

Procedure

To configure Layer 3 proxy-ND, enable it on a subinterface.

Example

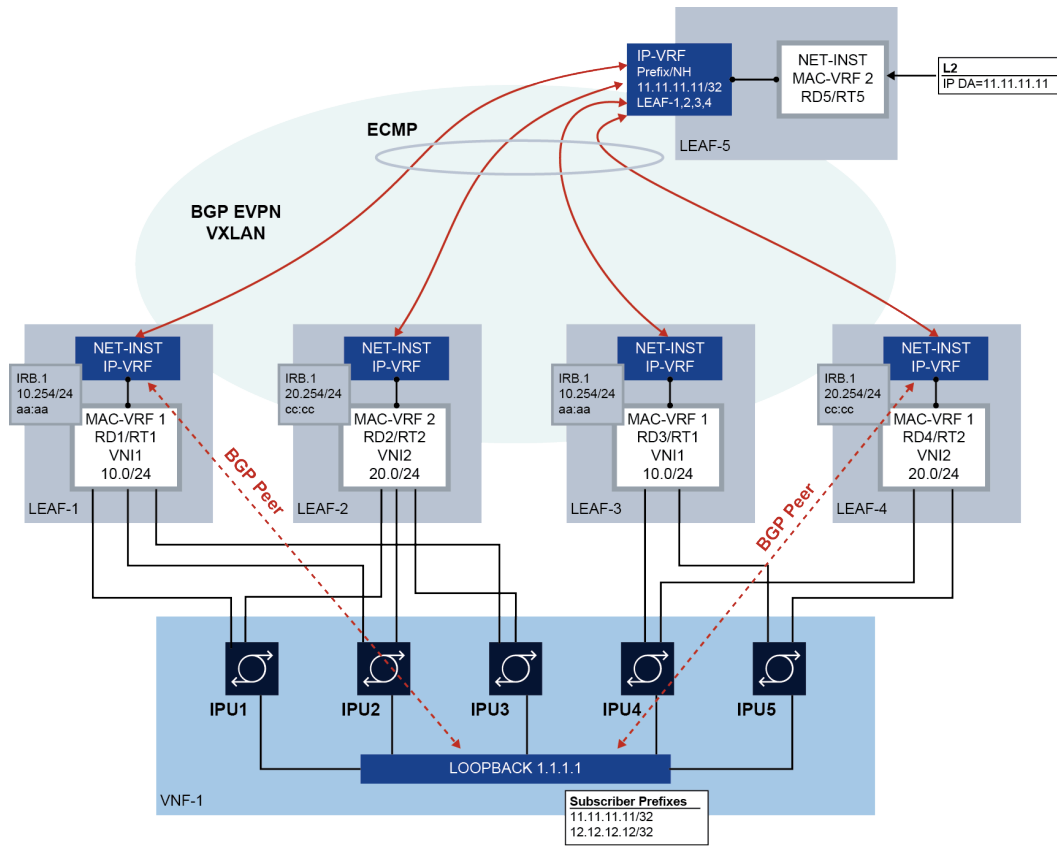
```
--{ candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv6 neighbor-discovery
  interface ethernet-1/1 {
    subinterface 1 {
      ipv6 {
        neighbor-discovery {
          proxy-nd true
        }
      }
    }
  }
}
```

4.6.8 EVPN IP aliasing

As defined in *draft-ietf-bess-evpn-ip-aliasing*, IP aliasing can be used with EVPN Layer 3 Ethernet Segments (ESs) to load-balance traffic to a host connected to multiple leaf nodes attached to the same IP-VRF, even if some of the leaves do not advertise reachability to that host.

SR Linux supports using IP aliasing for BGP peering where leaf switches are multihomed to a single virtual/container network function (referred to as VNF in this section) loopback address. The following figure shows this type of configuration.

Figure 20: BGP peering to VNF loopback



Note:

EVPN-VXLAN is used to describe IP aliasing in this section, however IP aliasing is also applicable to EVPN-MPLS services.

In this configuration, leaf nodes LEAF-1 through LEAF-5 are connected to an EVPN Layer 3 tenant domain, and four of these leaf nodes are multihomed to the same VNF. This VNF has five IPUs, similar to line cards, that are each dual-homed to two leaf nodes. Each IPU has two Layer 3 links that are connected to the leaf switches.

MAC-VRF 1 and MAC-VRF 2 are different Broadcast Domains (BDs) and have no Layer 2 EVPN connectivity between them.

The VNF is attached to a pair of subscriber prefixes, 11.11.11.11/32 and 12.12.12.12/32, that must be advertised to the leaf nodes, peering from the IP-VRF IRB interfaces, so the leaf nodes can distribute the reachability to the rest of the network.

In this example, the VNF can run PE-CE BGP from loopback interface 1.1.1.1/32, but supports a maximum of two BGP sessions.

In this configuration, IP aliasing on SR Linux allows the following:

- Full traffic spraying in both upstream and downstream directions, even though the number of BGP sessions is limited to two, and the VNF connects to four leaf nodes.

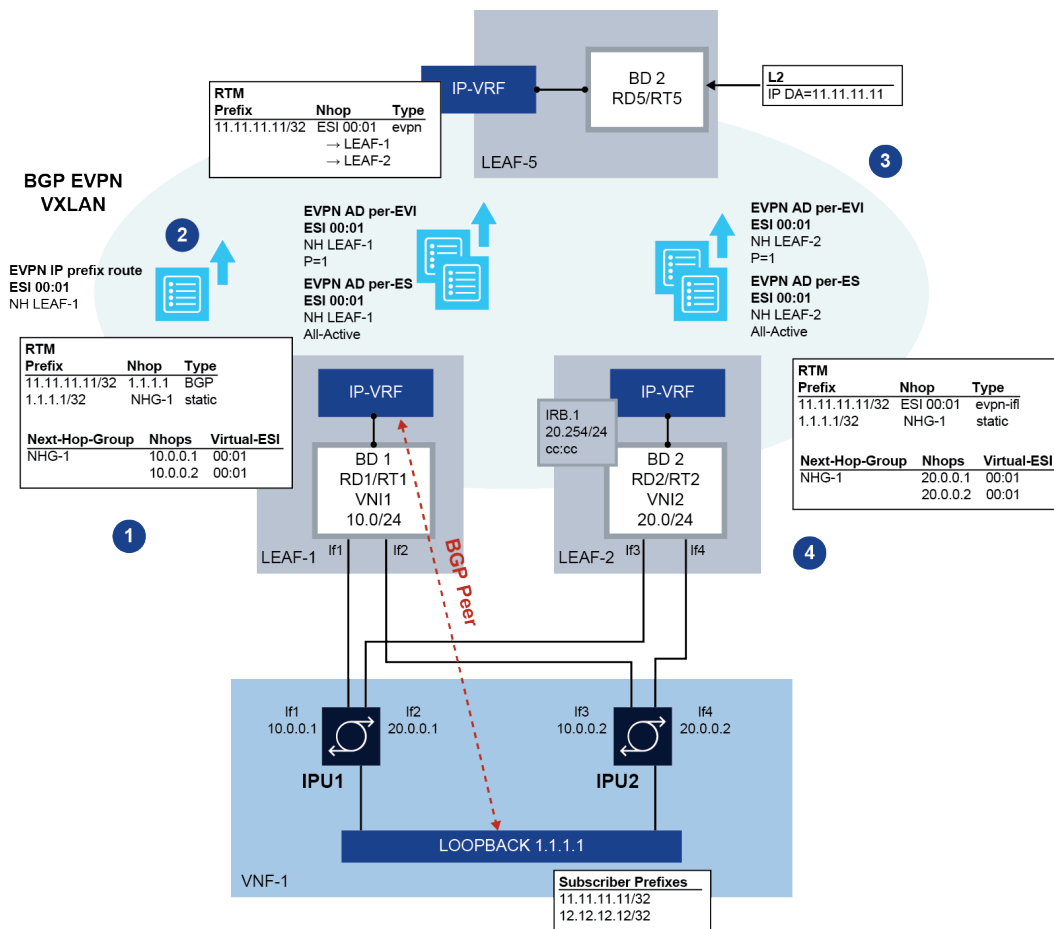
- No tromboning for downstream traffic routed to a leaf node that does not have a BGP session with the VNF. The local leaf must always route locally to the VNF, unless the VNF loopback becomes unreachable.

For example, when LEAF-5 sends traffic with destination 11.11.11.11 to LEAF-3, LEAF-3 needs to route locally to its local MAC-VRF 1, even if it has an RT5 in its route table pointing at LEAF-1 or LEAF-4.

EVPN IP aliasing on SR Linux

The following figure shows how IP aliasing on SR Linux works for BGP peering to a VNF loopback. To simplify the explanation, only two leaf nodes are shown; however, multiple leaf nodes can be multihomed to the same VNF loopback.

Figure 21: EVPN IP aliasing on SR Linux



In this configuration, BGP peering to the loopback on the VNF works as follows. The numbers represent the steps depicted in the figure.

- The VNF's loopback address (1.1.1.1) is associated with virtual Layer 3 ES ES-1 on both LEAF-1 and LEAF-2, working in all-active mode. Static routes to 1.1.1.1/32 are configured on both leaf switches. These static routes can run BFD to the VNF, if required, to speed up fault detection. After ES-1 is enabled, and as soon as 1.1.1.1 has a route (of any mask length) in the IP-VRF's route table, LEAF-1 and LEAF-2 advertise an EVPN ES route, as well as EVPN Auto Discovery (AD) per ES

and per EVI routes for ES-1. The AD routes are advertised based on the presence of a route for 1.1.1.1 in the route table. In this example, the route for 1.1.1.1/32 is statically configured in the route table, but any non-EVPN IP prefix route that provides reachability to 1.1.1.1 would trigger the advertisement of the AD routes.

If the ES ES-1 oper-state is up and a route for 1.1.1.1 is active in the route table, each leaf advertises:

- AD per-ES route for ES-1, indicating all-active mode
- AD per-EVI route for ES-1, indicating P = 1 (primary state is set), with route distinguisher and route target of the IP-VRF

The AD per-EVI routes contain two flags, P (primary) and B (backup), which provide an indication for the remote nodes to know whether they can send traffic to the owners of the routes. The remote nodes send traffic for the ES to all the nodes advertising themselves as primary (P = 1), and not to nodes signaling P = 0, B = 1, unless there are no primary nodes left in the ES.

For all-active multihoming mode, there is no DF election in the context of the ES.

2. Any prefixes that are resolved to the next-hop associated with ESI-1 are advertised as EVPN IP prefix routes with ESI 00:01 (representation of the ESI for ES-1). In the example in [Figure 21: EVPN IP aliasing on SR Linux](#), prefix 11.11.11.11/32 is advertised as an EVPN IP prefix route with ESI 00:01.
3. LEAF-5 identifies the EVPN IP prefix route with ESI 00:01 and installs the route in the IP-VRF route table with as many next-hops as received AD routes with ESI 00:01 and the primary flag set (P = 1). That is, the IP prefix route with ESI 00:01 is recursively resolved to the next hops that advertise the AD per-EVI (with P = 1) and per-ES routes for 00:01.
If the virtual Layer 3 ES is configured as single-active, a DF Election among all the leaf nodes attached to the same ES determines which leaf nodes advertise themselves as primary for the ES.
4. LEAF-2 identifies the EVPN IP prefix route with ESI 00:01 and next-hop LEAF-1 and resolves the route recursively to the next-hop associated with ES-1.
If the static route associated with ES-1 goes down, the leaf withdraws the associated Layer 3 AD per-EVI/ES routes to avoid attracting traffic from LEAF-5.

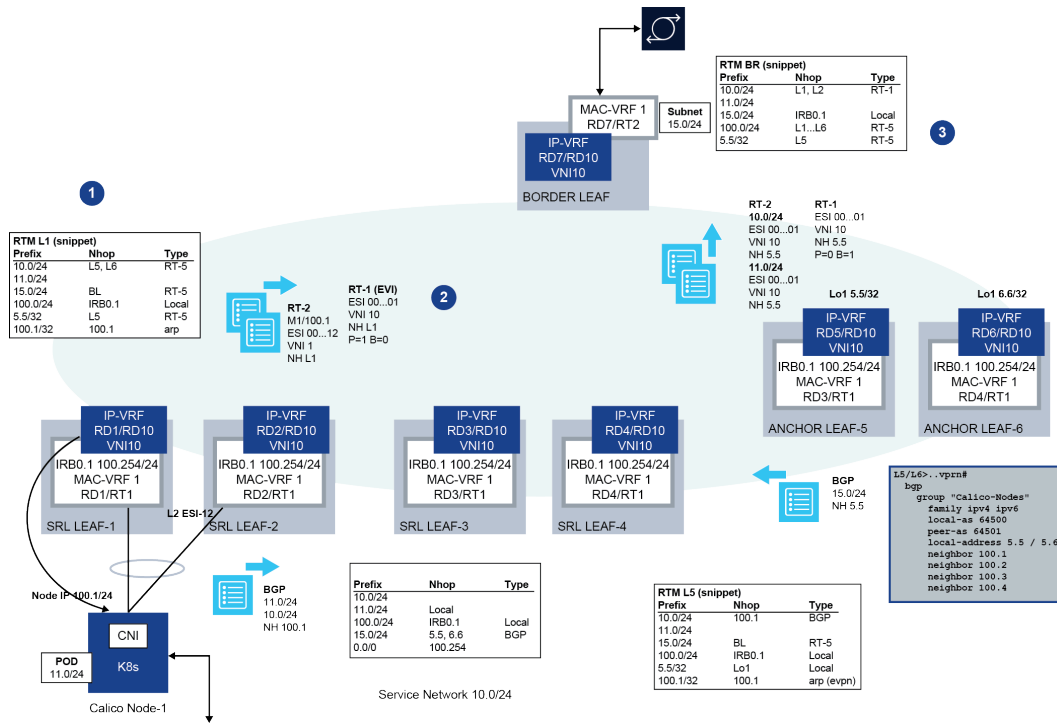
Centralized routing model and PE-CE routes resolved over EVPN-IFL

SR Linux supports PE-CE sessions over VXLAN/MPLS and PE-CE next-hop resolution via EVPN-IFL, which, along with Layer 3 ES, enables support for the centralized routing model, shown in [Figure 22: Centralized routing model](#).

This model uses BGP PE-CE peering on CEs to a pair of centralized anchor leaf nodes, as opposed to BGP PE-CE sessions to the directly connected leaf nodes. This is possible because BGP PE-CE route resolution via EVPN IFL routes is unblocked, which is necessary when the BGP session is not established via a local subinterface.

Having all the BGP PE-CE sessions centralized in a pair of anchor leaf nodes simplifies the provisioning of BGP not only on the rest of the leaf nodes, but also on the CEs connected to the leaf nodes.

Figure 22: Centralized routing model



In this example, the centralized routing model is used between all the Kubernetes (K8s) node IPs and the IP-VRF loopbacks on anchor leaf nodes; for example, between Node IP 100.1/24 (Node-1) and 5.5/32 in ANCHOR LEAF-5. Although only one Kubernetes node is depicted in the figure, multiple nodes may be attached to the leaf nodes throughout the data center. By using this model, the BGP configuration is simplified, especially on the Kubernetes nodes, where they can all use the same configuration with a peer to the anchor leaf nodes.

All the BGP PE-CE routes advertised by the CEs are re-advertised by the anchor leaf nodes using EVPN IP prefix routes. For example, prefixes 10.0/24 and 11.0/24 are advertised by Node-1 and re-advertised by LEAF-5 in IP prefix routes. IP Aliasing solves the issue of tromboning the traffic with destination 10.0 or 11.0 from the BORDER LEAF to LEAF-1 or LEAF-2 via anchor LEAF-5. IP Aliasing works as follows (where the numbers represent the steps depicted in [Figure 22: Centralized routing model](#)):

1. LEAF-1, LEAF-2, LEAF-5 and LEAF-6 are all configured with the virtual Layer 3 ES ES-1, which uses ESI 00..01, is associated with Node-1 IP 100.1, and is configured as single-active. Node-1 advertises BGP PE-CE routes for 10.0/24 and 11.0/24 to LEAF-5.
2. Because LEAF-1 and LEAF-2 have 100.1 in their route table as a local route and they are configured as primary nodes for the ES, they advertise AD per-EVI routes for ES-1 with P = 1. Anchor LEAF-5 and LEAF-6 are configured as non-primary, so even if they have 100.1 in their route table, they advertise their AD per-EVI routes with P = 0 and B = 1 as backup nodes for ES-1. When for example LEAF-5 receives the BGP PE-CE routes from Node-1, it re-advertises them in EVPN IP prefix routes with ESI 00..01. This is because the BGP PE-CE routes' next-hop matches the IP address in ES-1.
3. BORDER LEAF imports the EVPN IP prefix routes with ESI 00..01 and recursively resolves them to the next-hops of all the AD per-EVI routes advertised as primary (P = 1); that is, LEAF-1 and LEAF-2. When receiving traffic to 10.0/24 or 11.0/24, BORDER LEAF load balances the traffic to only LEAF-1 and LEAF-2, as opposed to the anchor leaf nodes.

4.6.8.1 Configuring Layer 3 Ethernet Segments for IP aliasing

To configure a Layer 3 Ethernet Segment for IP aliasing, you specify the Layer 3 next hop to be associated with an Ethernet Segment (ES). The AD per-ES/EVI routes for the ES are advertised when the specified next-hop address is active in the route table of the IP-VRF with the configured EVI.

The following configures two examples of Layer 3 Ethernet Segments, which are associated with an IPv4 and an IPv6 next-hop, respectively.

```
A:Leaf-1# info from running system network-instance protocols evpn ethernet-segments
system {
  network-instance {
    protocols {
      evpn {
        ethernet-segments {
          timers {
            boot-timer 180
            activation-timer 3
          }
          bgp-instance 1 {
            ethernet-segment es_v4_1 {
              type virtual
              admin-state enable
              esi 00:bc:de:00:00:00:00:01:00:04
              multi-homing-mode single-active
              next-hop 6.0.1.254 {
                evi 1 {
                }
              }
            }
            df-election {
              algorithm {
                type preference
                preference-alg {
                  preference-value 2
                }
                capabilities {
                  ac-df exclude
                  non-revertive false
                }
              }
            }
          }
          routes {
            next-hop use-system-ipv4-address
            ethernet-segment {
              originating-ip use-system-ipv4-address
            }
          }
        }
        ethernet-segment es_v6_1 {
          type virtual
          admin-state enable
          esi 00:bc:de:00:00:00:00:01:00:06
          multi-homing-mode single-active
          next-hop 6:0:1::254 {
            evi 1 {
            }
          }
        }
        df-election {
          algorithm {
            type preference
            preference-alg {
              preference-value 2
            }
          }
        }
      }
    }
  }
}
```


However, Leaf-2 in the following example is not configured as primary, so it advertises an AD per-EVI route with P = 0 and B = 1. Therefore, it does not attract traffic to the ES as long as there are primary nodes in the ES.

```
--{ +* candidate shared default }--[ system network-instance protocols evpn ethernet-segments
  bgp-instance 1 ethernet-segment L3-ES-1 ]--
A:Leaf-2# info
  type virtual
  admin-state enable
  esi 01:20:20:20:01:00:00:00:00:00
  multi-homing-mode single-active
  next-hop 20.20.20.1 {
    evi 2 {
    }
  }
  df-election {
    algorithm {
      type manual
      manual-alg {
      }
    }
  }
}
```

The following example shows the configuration of LEAF-1 and LEAF-2 in the [Figure 20: BGP peering to VNF loopback](#) figure. LEAF-5 is configured without any ES and with an IP-VRF that has BGP-EVPN enabled and **ecmp** greater than or equal to 2, so that it can load-balance traffic to the two leaf nodes attached to the ES.

```
# LEAF-1 and LEAF-2 configuration of the Ethernet Segment
--{ + candidate shared default }--[ system network-instance protocols evpn ]--
A:leaf1/leaf2# info
  ethernet-segments {
    bgp-instance 1 {
      ethernet-segment L3-ES-1 {
        type virtual
        admin-state enable
        esi 01:01:00:00:00:00:00:00:00:00
        multi-homing-mode all-active
        next-hop 1.1.1.1 {
          evi 2 {
          }
        }
      }
    }
  }
}

# LEAF-1 IP-VRF configuration, including BGP session to the VNF
--{ + candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf1# info
  type ip-vrf
  admin-state enable
  interface irb0.1 {
  }
  interface lo1.1 {
  }
  vxlan-interface vxlan1.2 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.2
      }
    }
  }
}
```

```

    evi 2
    ecmp 4
  }
}
bgp {
  autonomous-system 64501
  router-id 10.0.0.1
  group PE-CE {
    export-policy all
    import-policy all
    peer-as 64500
  }
  afi-safi ipv4-unicast {
    admin-state enable
  }
  neighbor 1.1.1.1 {
    peer-group PE-CE
    multihop {
      admin-state enable
      maximum-hops 10
    }
    transport {
      local-address 10.0.0.1
    }
  }
}
bgp-vpn {
  bgp-instance 1 {
    route-distinguisher {
      rd 1.1.1.1:2
    }
    route-target {
      export-rt target:64500:2
      import-rt target:64500:2
    }
  }
}
static-routes {
  route 1.1.1.1/32 {
    next-hop-group NHG-1
  }
}
next-hop-groups {
  group NHG-1 {
    nexthop 1 {
      ip-address 10.0.0.1
    }
    nexthop 2 {
      ip-address 10.0.0.2
    }
  }
}
}

# LEAF-2 IP-VRF configuration
--{ + candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf2# info
type ip-vrf
interface irb0.1 {
}
vxlan-interface vxlan1.2 {
}
protocols {
  bgp-evpn {
    bgp-instance 1 {

```

```

    vxlan-interface vxlan1.2
    evi 2
    ecmp 4
  }
}
bgp-vpn {
  bgp-instance 1 {
    route-distinguisher {
      rd 2.2.2.2:2
    }
    route-target {
      export-rt target:64500:2
      import-rt target:64500:2
    }
  }
}
static-routes {
  route 1.1.1.1/32 {
    next-hop-group NHG-1
  }
}
next-hop-groups {
  group NHG-1 {
    nexthop 1 {
      ip-address 20.0.0.1
    }
    nexthop 2 {
      ip-address 20.0.0.1
    }
  }
}
}
}
}

```

The following shows the route table for one of the subscriber prefixes (11.11.11.11/32) that are advertised from the VNF with a next-hop matching the Layer 3 ES next-hop. For LEAF-1, the route is learned via BGP PE-CE.

```

# LEAF-1 route table entry
--{ + candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf1# info from state route-table ipv4-unicast route 11.11.11.11/32 id 0 route-type bgp
route-owner bgp_mgr
route-table {
  ipv4-unicast {
    route 11.11.11.11/32 id 0 route-type bgp route-owner bgp_mgr {
      metric 0
      preference 170
      active true
      last-app-update 2022-10-19T10:26:51.020Z
      next-hop-group 423577430891
      next-hop-group-network-instance ip-vrf-2
      resilient-hash false
      fib-programming {
        last-successful-operation-type add
        last-successful-operation-timestamp 2022-10-19T10:26:51.021Z
        pending-operation-type none
        last-failed-operation-type none
      }
    }
  }
}
}
}
}

```

For LEAF-2, the route is learned via EVPN; however, the next-hop is resolved locally to the next-hop-group associated with the route for 1.1.1.1 (the ES next-hop IP address).

```
# LEAF-2 route table entry
--{ +* candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf2# info from state route-table ipv4-unicast route 11.11.11.11/32 id 0 route-type bgp-evpn
route-owner bgp_evpn_mgr
route-table {
  ipv4-unicast {
    route 11.11.11.11/32 id 0 route-type bgp-evpn route-owner bgp_evpn_mgr {
      metric 0
      preference 170
      active true
      last-app-update 2022-10-19T10:25:06.199Z
      next-hop-group 423577389148
      next-hop-group-network-instance ip-vrf-2
      resilient-hash false
      fib-programming {
        last-successful-operation-type add
        last-successful-operation-timestamp 2022-10-19T10:25:06.199Z
        pending-operation-type none
        last-failed-operation-type none
      }
    }
  }
}

--{ +* candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf2# info from state route-table next-hop-group 423577389148
route-table {
  next-hop-group 423577389148 {
    backup-next-hop-group 0
    next-hop 0 {
      next-hop 423577389136
      resolved true
    }
    next-hop 1 {
      next-hop 423577389137
      resolved true
    }
  }
}

--{ +* candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf2# show route-table next-hop 423577389136
-----
Next-hop route table of network instance ip-vrf-2
-----
Index : 423577389136
Next-hop : 20.20.20.1
Type : indirect
Subinterface : N/A
Resolving Route : 20.20.20.0/24 (local)
```

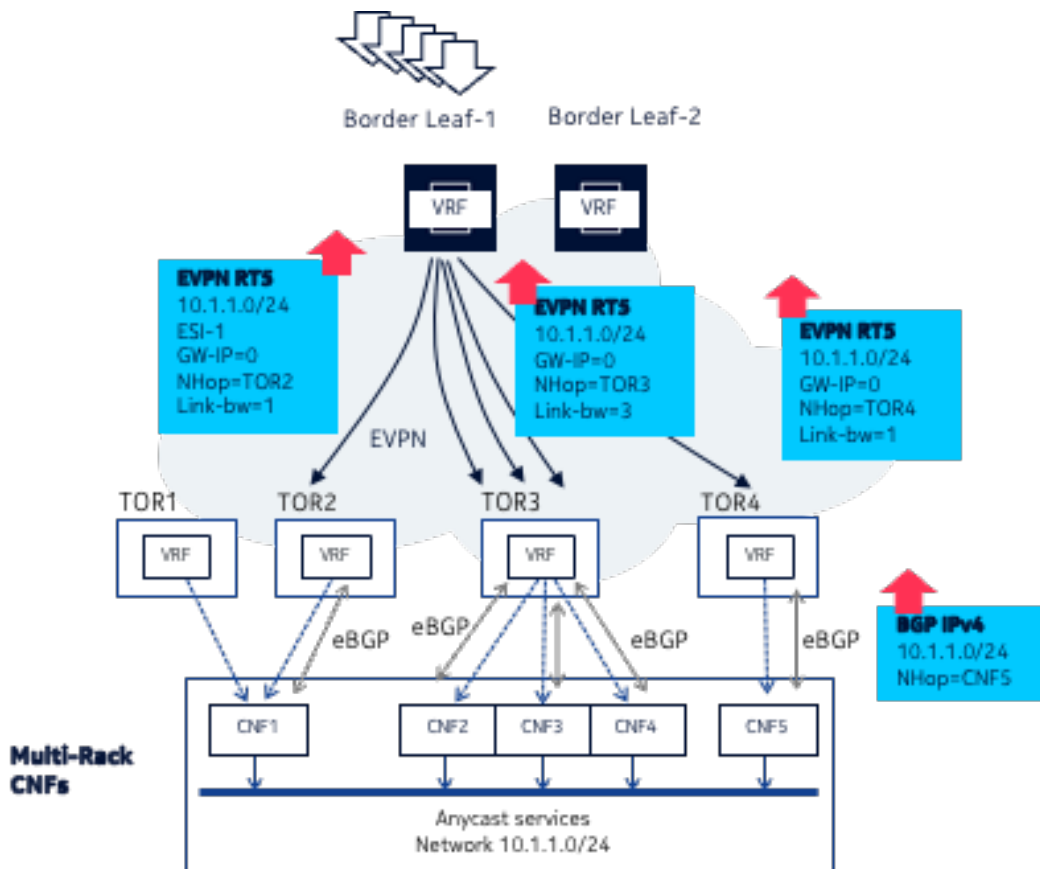
4.6.9 Unequal ECMP for EVPN IP prefix routes

SR Linux supports unequal Equal Cost Multi Path (ECMP) for EVPN IP prefix IFL (interface-less) routes for EVPN-VXLAN and EVPN-MPLS services. To do this, SR Linux makes use of the EVPN link-bandwidth extended community (EC) defined in [draft-ietf-bess-evpn-unequal-lb](#). This extended community indicates

the weight for a specific IP prefix; that is, the number of PE-CE multi-paths for an IP prefix that is re-advertised into an EVPN IP prefix route.

The following figure shows an example of weighted ECMP. Assuming each Container Network Function (CNF) advertises the anycast subnet 10.1.1.0/24 from a different next-hop, each TOR ends up with a different number of multi-paths in its PE-CE session. In the example below, TOR3 has three multi-paths for the anycast subnet, and the advertised EVPN IP prefix route includes an EVPN link-bandwidth extended community with a weight of 3. The rest of the TORs send a weight of 1. Note that TOR1 and TOR2 are multi-homed to the same CNF1, yet they send a weight of 1.

Figure 23: EVPN using weighted ECMP



On the border leafs / data center gateways, when this feature is enabled, if the EVPN IP prefix route has an Ethernet Segment Identifier (ESI) of 0, the PE sprays the flows to the EVPN IP prefix route based on the received weight; in this example, one-fifth of the flows are sent to TOR4, and three-fifths are sent to TOR3.

When the EVPN IP prefix route has a non-zero ESI, and there is a weight in the route:

- The EVPN link-bandwidth extended community received in the EVPN IP prefix route indicates the weight for the EVPN IP prefix route. The EVPN link-bandwidth extended community may also be received in the IP A-D per ES routes for each PE attached to the Ethernet segment (ES), but the system ignores it.
- The PE sprays the flows to the EVPN IP prefix route based on the received weight, dividing the flows to an ES among the number of PEs attached to the ES.

In the example above, one-fifth of the flows are sent to the aliased pair TOR1/TOR2 (either one is selected because TOR1 and TOR2 are assumed to send equal weights in the AD per ES routes).

- The system rounds up when the advertised weight for the ESI, divided by the number of PEs in the ES, is not an integer.

For example, if ES1 (TOR1/TOR2) advertises BW=3, and TOR4 advertises BW=1, then $3 \text{ (BW)} / 2 \text{ (PE in ES1)} = 1.5$. The system rounds up, and the remote nodes install weight=2 for TOR1, weight=2 for TOR2, and weight=1 for TOR4.

- If the weight received in a non-zero ESI IP prefix route exceeds 128, the system caps it at 128, then divides the weight into the number of PEs in the ES.
- If two EVPN IP prefix routes are received for the same prefix, same ESI, different route distinguishers (RDs), they should have the same weight. However, if they have different weights, the system selects the weight from the first EVPN IP prefix route.
- If the EVPN link-bandwidth extended community is missing from any of the PEs in an ECMP set, or the Value Units field of the extended community is inconsistent, the weight is ignored by the receiving PE, and regular ECMP forwarding is performed. The Value Units field can indicate "bandwidth" or a "generalized weight", with the latter being supported by SR Linux.

4.6.9.1 Advertising the EVPN link-bandwidth extended community

Procedure

To configure weighted ECMP, you enable the advertisement of the EVPN link-bandwidth extended community and specify the weight to be advertised in the extended community.

You can configure the following parameters for the advertised weight in the extended community:

- `weight` specifies the weight to be advertised in the EVPN link-bandwidth extended community for the advertised EVPN IP prefix routes for the service. If set to `dynamic` (the default value), the weight is dynamically set based on the number of BGP PE-CE paths for the IP prefix that is advertised in an EVPN IP prefix route. The dynamic weight only considers BGP PE-CE paths.

Alternatively, the weight can be set to a fixed integer value in the range 1 to 128.

- `maximum-dynamic-weight` specifies the maximum weight to be advertised in the EVPN link-bandwidth extended community for the advertised EVPN IP-Prefix routes for the service. If `weight dynamic` is configured, the actual advertised weight is the minimum of the number of BGP PE-CE paths for the prefix and the configured `maximum-dynamic-weight`.

Example

The following example enables advertisement of the EVPN link bandwidth extended community and specifies the weight to be included in the extended community for the advertised EVPN IP prefix routes.

```
--{ * candidate shared default }--[ ]--
# info network-instance VRF1 protocols bgp-evpn bgp-instance 1
  network-instance VRF1 {
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          routes {
            route-table {
              ip-prefix {
```



```
evpn-link-bandwidth {
  advertise {
    weight 60
  }
}
```

4.6.9.2 Enabling weighted ECMP

Procedure

When weighted ECMP is enabled, the system takes into account the EVPN link-bandwidth extended community when installing an ECMP set for an EVPN IP prefix route in the IP-VRF route table.

Flows to an IP prefix received with a weight and a zero ESI are sprayed according to the weight. If the EVPN IP prefix route received with a weight has a non-zero ESI, the weight is divided into the number of PEs attached to the ES, and rounded up if the result is not an integer.

The command also enables weighted ECMP for BGP CEs that are configured with a weight specified with the `link-bandwidth.add-next-hop-count-to-received-bgp-routes` setting.

Configuring `max-ecmp-hash-buckets-per-next-hop-group` preserves the datapath resources used for the weighted next-hops. The normalization algorithm also refers to this number of hash buckets. See [Displaying ECMP weights](#) for an example of how the weights are normalized the using the `max-ecmp-hash-buckets-per-next-hop-group` setting.

Example

The following example enables weighted ECMP and specifies the maximum number of ECMP hash buckets per next-hop-group. The weights for weighted ECMP are normalized based on this number of hash buckets.

```
--{ * candidate shared default }--[ ]--
# info network-instance VRF1 protocols bgp-evpn bgp-instance 1
network-instance VRF1 {
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        routes {
          route-table {
            ip-prefix {
              evpn-link-bandwidth {
                weighted-ecmp {
                  admin-state enable
                  max-ecmp-hash-buckets-per-next-hop-group 4
                }
              }
            }
          }
        }
      }
    }
  }
}
```

}

Example



Note: The range of the command `max-ecmp-hash-buckets-per-next-hop-group` differs depending on the platform. Validation of the configuration prevents the configuration of a value that the platform cannot support.

4.6.9.3 Configuring weighted ECMP for received PE-CE BGP routes

Procedure

When the system is enabled to process the link-bandwidth extended community, you can configure a weight to be internally added to the received PE-CE BGP routes for the purpose of EVPN unequal ECMP.

Example

```
--{ * candidate shared default }--[ ]--
# info network-instance VRF1 protocols bgp group g1 link-bandwidth
network-instance VRF1 {
  protocols {
    bgp {
      group g1 {
        link-bandwidth {
          add-next-hop-count-to-received-bgp-routes 100
        }
      }
    }
  }
}
```

4.6.9.4 Displaying ECMP weights

Procedure

The system normalizes the weights used in weighted ECMP using the `max-ecmp-hash-buckets-per-next-hop-group` setting and the advertised weights, according to the algorithm described in "Normalizing datapath weights" in the *SR Linux Routing Protocols Guide*.

You can display the weights for a next-hop group.

Example

In the following example, the same prefix 19.1.1.1/32 is received from two PEs with weights 40 and 60, respectively. The maximum supported ECMP paths is 8. The `max-ecmp-hash-buckets-per-next-hop-group` setting is 4. The system programs the next-hops with normalized weights.

To display the number of the next-hop-group for prefix 19.1.1.1/32:

```
--{ running }--[ ]--
# info from state network-instance VRF1 route-table ipv4-unicast route 19.1.1.1/32 id 0
route-type bgp-evpn route-owner bgp_evpn_mgr origin-network-instance VRF1
network-instance VRF1 {
  route-table {
    ipv4-unicast {
```

```

route 19.1.1.1/32 id 0 route-type bgp-evpn route-owner bgp_evpn_mgr
origin-network-instance VRF1 {
    leakable false
    metric 0
    preference 170
    active true
    last-app-update 2023-05-31T18:29:25.207Z
    next-hop-group 94413408183
    next-hop-group-network-instance VRF1
    resilient-hash false
    fib-programming {
        suppressed false
        last-successful-operation-type modify
        last-successful-operation-timestamp 2023-05-31T18:29:25.207Z
        pending-operation-type none
        last-failed-operation-type none
    }
}
}
}
}

```

To display the received weights for the individual next hops in the next-hop group:

```

--{ running }--[ ]--
# info from state network-instance VRF1 route-table next-hop-group 94413408183
network-instance VRF1 {
    route-table {
        next-hop-group 94413408183 {
            backup-next-hop-group 0
            fib-programming {
                last-successful-operation-type add
                last-successful-operation-timestamp 2023-05-31T18:29:25.207Z
                pending-operation-type none
                last-failed-operation-type none
            }
            next-hop 0 {
                next-hop 94413408157
                weight 40
                resolved true
            }
            next-hop 1 {
                next-hop 94413408152
                weight 60
                resolved true
            }
        }
    }
}
}

```

The system normalizes the weights using the algorithm described in "Normalizing datapath weights" in the *SR Linux Routing Protocols Guide*, taking into account the received weights, the maximum supported ECMP paths, and the `max-ecmp-hash-buckets-per-next-hop-group` setting. However, the normalized weights that are programmed in the data path cannot be displayed.

4.6.10 Resilient hashing for EVPN IFL

Resilient hashing is supported for EVPN IFL where an IFL route points to a next-hop group (NHG) that contains indirect next hops. The indirect next hops are then resolved by MPLS tunnels.

The resilient hashing feature has "sticky" flows even when flows are rehashed because of newly added or failed links. This feature prevents rerouting of flows that are unaffected by network link failures.

SR Linux supports resilient hashing for the following routes:

- regular ECMP with EVPN IFL routes
- unequal ECMP with EVPN IFL routes
- combined EVPN IFL and BGP PE-CE weighted ECMP

Resilient hashing is also supported for the addition or removal of next hops.

4.6.10.1 Resilient hashing and unequal ECMP for EVPN IFL

SR Linux supports combined resilient hashing and unequal ECMP for EVPN IP prefix IFL routes. Both resilient hashing and unequal ECMP can be used on the same prefix.

Configuring resilient hashing and unequal ECMP for EVPN IFL

The following is an example configuration for resilient hashing and unequal ECMP for EVPN IFL:

```
--{ * candidate shared default }--[ network-instance ip-vrf-2 ]--
A:leaf1# info
  type ip-vrf
  admin-state enable
  ip-load-balancing {
    resilient-hash-prefix 10.10.10.0/24 { // resilient hashing is configured for prefix
10.10.10.0/24
      hash-buckets-per-path 32
      max-paths 2
    }
  }
  interface irb0.1 {
  }
  interface lo1.1 {
  }
  vxlan-interface vxlan1.2 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.2
        evi 2
        ecmp 4
        routes {
          route-table {
            ip-prefix {
              evpn-link-bandwidth {
                advertise {
                  weight dynamic
                }
              }
              weighted-ecmp { // unequal ECMP is configured for EVPN IFL
                admin-state enable
                max-ecmp-hash-buckets-per-next-hop-group 128
              }
            }
          }
        }
      }
    }
  }
}
```

```

    }
  }
  bgp {
    autonomous-system 64501
    router-id 1.1.1.1
    afi-safi ipv4-unicast {
      admin-state enable
      multipath {
        max-paths-level-1 4
        max-paths-level-2 4
      }
    }
  }
  group PE-CE {
    export-policy all
    import-policy all
    peer-as 64500
    afi-safi ipv4-unicast {
      ipv4-unicast {
        link-bandwidth {
          add-next-hop-count-to-received-bgp-routes 1
        }
      }
    }
    timers {
      connect-retry 1
      hold-time 9
      keepalive-interval 3
      minimum-advertisement-interval 1
    }
  }
  neighbor 20.20.20.1 {
    peer-group PE-CE
    multihop {
      admin-state enable
      maximum-hops 10
    }
    transport {
      local-address 1.1.1.1
    }
  }
}
bgp-vpn {
  bgp-instance 1 {
    route-distinguisher {
      rd 1.1.1.1:2
    }
    route-target {
      export-rt target:64500:2
      import-rt target:64500:2
    }
  }
}
}

```

Resilient hashing is configured for the 10.10.10.0/24 prefix using the **resilient-hash-prefix** parameter. The **hash-buckets-per-path** parameter indicates the number of times each next hop is repeated in the fill pattern. The **max-paths** parameter indicates the maximum number of ECMP next hops per route associated with the **resilient-hash-prefix**. The product of the **hash-buckets-per-path** and **max-paths** parameters is the total number of buckets that are used when applying the normalized weights.

Unequal ECMP is configured for the same EVPN IFL prefix using the **weighted-ecmp** and **max-ecmp-hash-buckets-per-next-hop-group** parameters.

To configure both resilient hashing and unequal ECMP on the same prefix, the **hash-buckets-per-path** and **max-ecmp-hash-buckets-per-next-hop-group** parameters can be set to different values. The **hash-buckets-per-path** parameter will override **max-ecmp-hash-buckets-per-next-hop-group**.

Example

The resilient hashing configuration contains the following parameter values:

- **max-paths=16**
- **hash-buckets-per-path=16**

In this example, the unequal ECMP configuration contains the following parameter value: **max-ecmp-hash-buckets-per-next-hop-group=128**

The above values result in the router using 256 hash buckets for the prefix, with an ECMP member cap of 16.

4.7 Dynamic multicast ID management

In SR Linux, multicast IDs (MCIDs) are multicast groups used by systems in the data path. MCIDs are allocated by the SR Linux `mcid_mgr` application so that the XDP layer can program different multicast forwarding modes in network instances.

For example, on 7220 IXR platforms, the system allocates two MCIDs for each MAC-VRF network instance:

- One MCID is used for the default multicast or flood list.
- On EVPN-VXLAN MAC-VRFs, another MCID is dedicated to allow multi-homing correct forwarding of BUM traffic. Traffic coming from VXLAN that needs to be flooded automatically picks up the secondary multicast group that includes local single-homed subinterfaces plus Designated Forwarder (DF) subinterfaces, as opposed to the default flood list.
- Similarly, on EVPN-MPLS MAC-VRFs, another MCID is dedicated to BGP-EVPN instances with MPLS enabled.

The available MCIDs in SR Linux are divided among the following applications:

- `net_instance_mgr`, which requests an MCID per default flood list (allocated per MAC-VRF when the MAC-VRF is configured)
- `evpn_mgr`, which requests MCIDs per MAC-VRF (one MCID is used for EVPN multi-homing, allocated when VXLAN is configured on the MAC-VRF, and another one is used for BGP-EVPN MPLS instances)
- `l2_proxy_arp_nd_mgr`, which requests an MCID per MAC-VRF when the proxy-ARP/ND flood options `flood gratuitous-arp`, `unknown-arp-rq` are set to `false` to suppress flooding over EVPN destinations.
- Individual multicast modules, which request an MCID per MFIB entry created by IGMP/MLD/PIM/MVPN.



Note: The use of MCIDs and the maximum number of MCIDs vary depending on the platform.

Request-response model for MCIDs

To accommodate the SR Linux features that require MCIDs, the system allocates MCIDs dynamically, using a request-response model. When an application needs a new MCID, it makes a request to the

mcid_mgr application, which allocates the MCID and publishes a response with the allocated MCID to the requesting application. The mcid_mgr application also programs XDP with the MCID.

MCID exhaustion

The system has a finite number of MCIDs to allocate. If mcid_mgr runs out of MCIDs it can allocate, the following applications are brought down when a new object is configured:

- mac-vrf
- vxlan-interface
- bgp-evpn mpls instance

When this occurs, the oper - down - reason for the application is displayed as no-mcid. When the oper - state for an application is changed to down, then brought back up, if there are no MCIDs available, the application is kept down with oper - down - reason no-mcid. For example, a change in evi automatically toggles the oper - state of an EVPN instance. If there are no MCIDs available following the evi change, the EVPN instance may not come back up.

Upon MCID exhaustion, mcid_mgr keeps trying to allocate MCIDs for the requesting multicast applications. If the system is rebooted, it may result in some VXLAN interfaces / MAC-VRFs that were not allocated MCIDs earlier receiving them at the expense of the multicast routes.

For l2_proxy_arp_nd_mgr, the application is brought down only if the flood options that require a new MCID are configured.

4.7.1 Displaying MCID information

Procedure

You can use an **info from state** command to display information about MCID allocation and MCID usage per application.

The command displays the total number of MCIDs available in the system, the total number of MCIDs currently in use, and the total number of pending MCIDs that mcid_mgr has not yet allocated.

For each application that requires MCIDs (Layer 2 proxy ARP/ND, MAC VRFs, MAC VRF BGP-EVPN, and VXLAN interfaces) you can display the number of MCIDs in use by the application, as well as the number of MCIDs requested but not yet allocated for use by the application.

Example

The following command displays the current MCID usage:

```
--{ running }--[ ]--
# info from state system multicast multicast-ids statistics
  system {
    multicast {
      multicast-ids {
        statistics {
          maximum-ids 16380
          current-usage 3
          total-pending 0
          multicast-id-user-type mac-vrf {
            current-usage 2
            total-pending 0
          }
          multicast-id-user-type vxlan-interface {
```

```

        current-usage 1
        total-pending 0
    }
    multicast-id-user-type l2-proxy-arp-nd {
        current-usage 0
        total-pending 0
    }
    multicast-id-user-type mfib {
        current-usage 0
        total-pending 0
    }
    multicast-id-user-type mac-vrf-bgp-evpn {
        current-usage 0
        total-pending 0
    }
}
}
}
}
}

```

4.8 EVPN for multicast

This chapter describes EVPN for multicast on SR Linux.



Note: The components of EVPN for multicast are currently supported exclusively on 7220 IXR D2/D3/D2L/D3L/D4/D5 platforms.

4.8.1 IGMP/MLD snooping

IGMP snooping is enabled by configuring **igmp-snooping admin-state enable** under **network-instance.protocols**. The feature is supported on MAC-VRF network-instances with bridged subinterfaces and no IRBs.

The **igmp-snooping** model allows configuration of IGMP parameters at the MAC-VRF network-instance level and specific configuration for individual interfaces. The programming of the MFIB for the network-instance is enabled by the configuration of **network-instance.protocols.igmp-snooping**.

IGMP snooping support in SR Linux includes the following:

- IGMP versions 1, 2, and 3 are supported.
- IGMP snooping is supported on network-instances of type **mac-vrf**.
- The following features are supported:
 - Querier support, including timers and relevant parameters
 - Trace options
 - Mcast router support
 - State for group memberships, queriers, and mcast routers
 - Static membership groups configured under interfaces
 - Statistics

- As soon as IGMP snooping is enabled on the MAC-VRF, the node starts snooping IGMP messages on the interfaces added under `protocols.igmp-snooping`, and starts programming the MFIB and proxying IGMP reports.
The interfaces not configured under `protocols.igmp-snooping` but that exist in the network-instance are considered mrouter ports by the IGMP snooping application. The state model shows the learned queriers, multicast routers, and proxy-membership-groups.
- IGMP settings can be customized at the interface level. The state model shows the learned membership-groups on the interface, as well as the statistics.
- When originating reports from CPM, the system MAC is used.

MLD snooping is modeled after IGMP snooping (both are managed by IGMP manager), with MLD versions 1 and 2 supported. Similar to IGMP snooping, MLD snooping is enabled with the command **mld-snooping admin-state enable**.

IGMP/MLD snooping are supported on 7220 IXR-D series platforms.

4.8.2 EVPN multicast for Layer 2

This section describes EVPN multicast for Layer 2 on SR Linux.

IGMP/MLD proxies for EVPN (RFC 9251)

SR Linux supports IGMP/MLD proxies for EVPN networks, as described in RFC 9251.

When IGMP snooping and, or MLD snooping are enabled on a MAC-VRF, a multicast forwarding information base (MFIB) is created with entries for each (*,G) or (S,G) state that is snooped out of received IGMP/MLD reports. The result of that MFIB is advertised in EVPN Selective Multicast Ethernet Tag (SMET) routes, one route per (*,G)/(S,G) entry, as long as EVPN is enabled in the MAC-VRF.

The following summarizes what happens when both `igmp/mlld-snooping` and `bgp-evpn` are enabled:

- The router sends updates of the IMET routes for the MAC-VRF with the multicast flags EC signaling support for IGMP/MLD-proxy.
- The router collects information about the remote EVPN destinations that support iIGMP/MLD-proxy (based on the received IMET flags) and stops flooding IGMP reports to those destinations.
- The router also creates an EVPN proxy DB and sends the result in SMET routes.
- The router adds EVPN destinations as OIF (outgoing interfaces) in the MFIB based on the received SMET routes.

The SMET route (or EVPN route type 6) is described in RFC 9251 and conveys the information about the multicast source and the multicast group IP addresses (for IPv4 or IPv6) along with some flags. The flags indicate whether the proxy entry in the MFIB was created using IGMP/MLDv2 or IGMPv3; in case of IGMPv3/MLDv2, the `Exclude` flag (when set to 1) indicates if the source is excluded.

On reception of an SMET route, IGMP/MLD reports are generated and sent to mrouter subinterfaces based on the IGMP/MLD version that is locally valid.

Integration of EVPN proxy and non-proxy routers

SR Linux supports the integration of EVPN multicast proxy-enabled and non-proxy routers in the same EVPN broadcast domain. As described in RFC 9251, each router where IGMP/MLD snooping is enabled advertises an IMET route along with the Multicast Flags EC. This EC is used by the router to indicate that it has enabled IGMP/MLD snooping for that MAC-VRF.

Upon reception of the IMET route, a PE considers that the originating PE supports the EVPN proxy behavior if and only if, the Multicast Flags EC is present *and* the proxy bit value is 1, else the originating PE is not considered as an EVPN proxy PE. This information is used for backwards compatibility with non-supporting proxy or non-upgraded PEs. The following considerations apply:

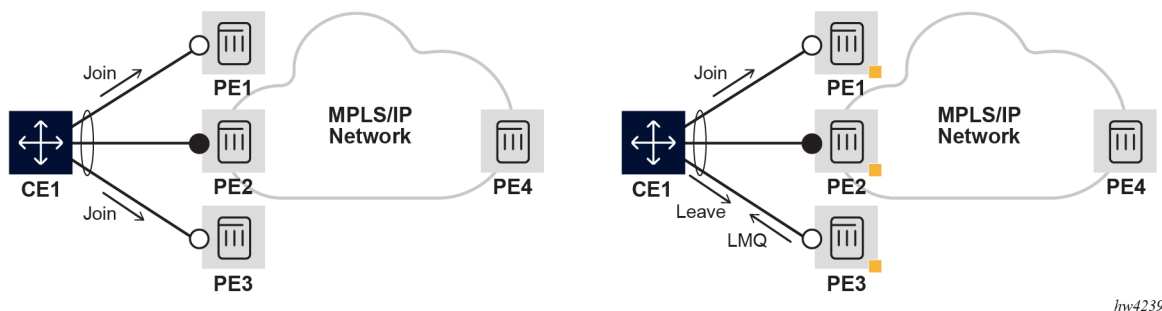
- On the egress PE (connected to receivers):
 - If an egress PE supports EVPN multicast proxy, but does not have any interest in a given (x, G) (with x meaning source S or *), it advertises its IGMP proxy capability using this EC, but does not advertise any SMET route for that (x, G).
- On the ingress PE (connected to sources):
 - When the PE receives an IMET advertisement with proxy support from the egress PE, it does not replicate the multicast traffic to that PE until a SMET route is received from it.
 - When the PE receives an IMET route with *no* proxy support from an egress PE, multicast traffic is replicated to that egress PE, even if the egress PE has no receivers for the stream. The egress PE is in this case added as an OIF for (*, *); that is, multicast is sent to it for any sources and groups.
 - Reports and queries are still forwarded to non-proxy PEs and processed on reception from non-proxy PEs.

To summarize, an EVPN multicast proxy PE never sends IGMP/MLD reports to EVPN destinations created from IMET routes with EVPN multicast proxy support, but it does send reports to EVPN destinations created from IMET routes without the proxy indication. This behavior allows the interworking between EVPN multicast proxy and non-proxy PEs.

4.8.2.1 EVPN multihoming and multicast synchronization

As described in RFC 9251, on EVPN multicast proxy-enabled PEs, multicast states must be synchronized across all the PEs attached to the same ES. This is illustrated in the following figure:

Figure 24: Multicast state synchronization for EVPN MH



The left side of the figure shows how it is possible for CE1 to send Join messages for a (S,G) to PE1 or PE3, while PE2 is the DF and therefore the only PE that can send multicast traffic to the CE. The multicast traffic would not be delivered to the CE1 unless the DF, PE2, creates a state for the (S,G).

The right side of the figure shows a similar issue for the Leave process. CE1 may send a Leave message to PE3, which replies back with a LMQ (Last Member Query). CE1 then may send a Join message to a different PE as a result of a different hashing; for example, PE1. Therefore, PE3 would think the CE has left the group and would remove the state.

These two issues are solved by synchronizing the multicast states in all the PEs attached to the same ES. This feature introduces two new EVPN route types described in RFC 9251 to synchronize the multicast states in EVPN MAC-VRFs:

- Route type 7, or Multicast Join Synch Route (synchronizes the join state in all the PEs of the ES).
- Route type 8, or Multicast Leave Synch Route (synchronizes the leave state in all the PEs of the ES).

Note that this feature is needed along with all-active and single-active multi-homing. In single-active multi-homing, the synchronization helps speed up convergence.

Operation of EVPN multicast and multihoming

EVPN multicast synchronization in multi-homed PEs does not require any specific configuration command. Configuration of **igmp/mld-snooping admin-state enable** and the existence of multicast state received in a bridged subinterface attached to an ES is the trigger for the router to advertise the EVPN multicast synchronization routes.

The following considerations apply:

- Reception of IGMP/MLD Join and Leave messages on a local ES triggers the advertisement of EVPN multicast Join/Leave synchronization routes respectively and the associated procedures in RFC 9251. These routes are advertised only if ES bridged subinterfaces with multicast state exist.

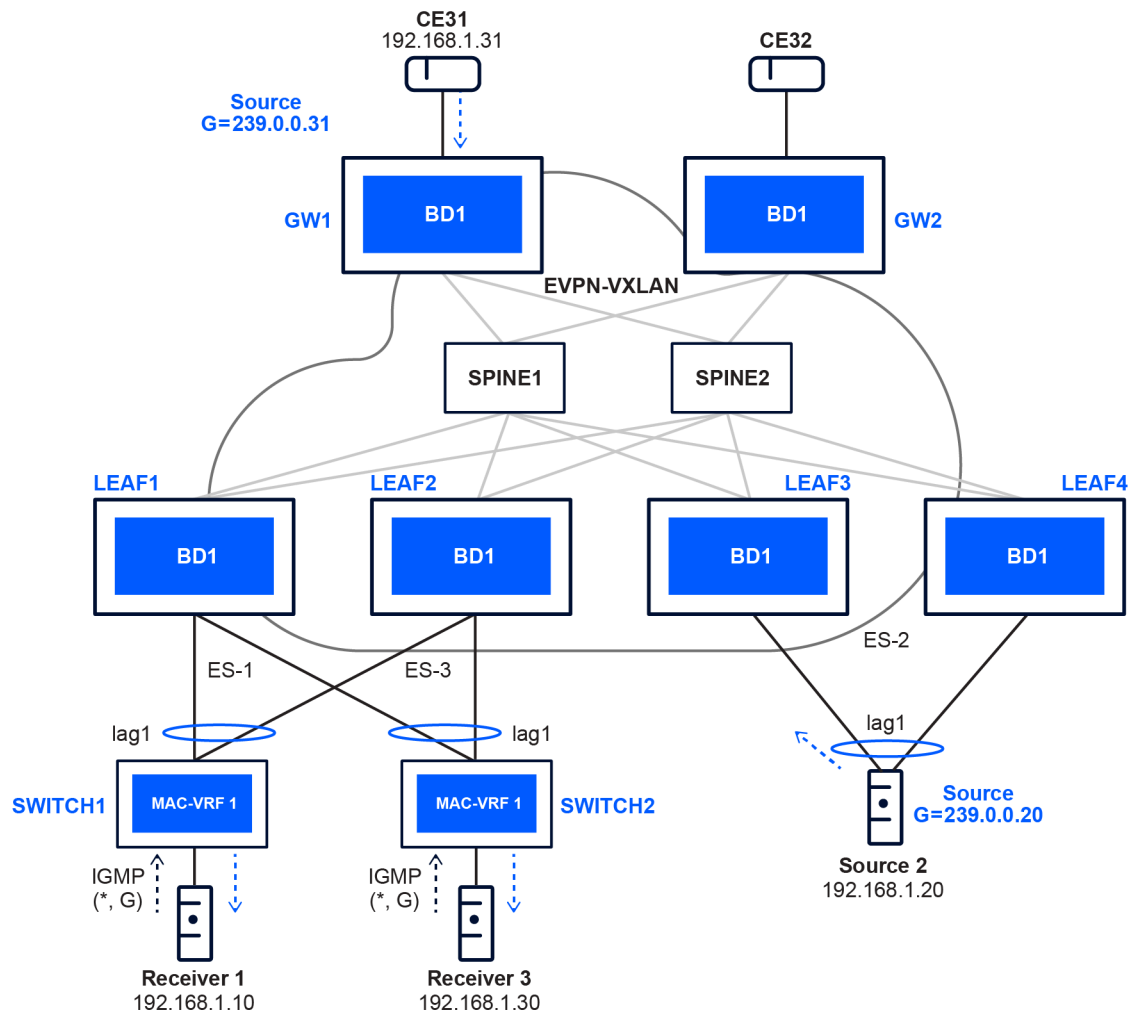
The routes are always advertised with the ESI-import route-target (which ensures the routes are only imported by the PEs attached to the ES) and the EVI route target EC (which guarantees the routes are only imported in the correct MAC-VRF).

- Reception of multicast Join/Leave synchronization routes trigger the synchronization of states and the associated procedures described in RFC 9251.
- About the Leave synchronization routes:
 - Leave synchronization routes synchronize the LMQ (Last Member Query) procedures in the ES peers.
 - The node triggering the route encodes (Last Member Query Interval * robust count) in the Maximum Response Time field of the NLRI, but applies an extra delta before expiring the max response time locally. This delta time is configurable with **system.network-instance.protocols.evpn.multicast.leave-sync-propagation <number>**, with a 0-to-300 second range and default value of 5 seconds.
 - This delta is a configured value, ideally configured the same on all the ES peers and long enough to account for the BGP propagation time between the ES peers.
For example, suppose there is a max response time of 5 seconds (50 deciseconds), advertised by A and a delay of 4 seconds in BGP propagation to B. In this case, the timer could already expire on A while B is still in LMQ time and can still receive Joins (which would recreate state in A after a Join synchronization route from B).
By adding an extra delta on A, it minimizes the potential churn of A removing and recreating the state.
 - The advertised maximum response time value in the multicast Leave synchronization route is the configured `robust-count * query-last-member-interval`, by default 2 seconds. The `robust-count` is taken from the `igmp` context (the value is defined by the querier), and the `query-last-member-interval` is taken from the ES interface `igmp-snooping` configuration.
 - Fast Leave is supported along with the EVPN Leave synchronization routes. In this case, the EVPN Leave synchronization route is advertised with a Max Response Time of 0 seconds.

4.8.2.2 Configuring EVPN multicast for Layer 2

To configure EVPN multicast for Layer 2, **igmp/mld-snooping** must be enabled on the MAC-VRF. The following example illustrates how EVPN multicast for Layer 2 is configured.

Figure 25: EVPN multicast for Layer 2



Inv4242

Where:

- GW1 and GW2 are two data center gateways (DC GWs).
- LEAF1 through LEAF4 are four SR Linux nodes. LEAF3 and LEAF4 are multi-homed to Client 2, a source that generates traffic for group 239.0.0.20.
- The four leaf nodes and the two DC GW nodes are attached to the same EVPN BD.

- SWITCH1 and SWITCH2 are two SR Linux systems that are not EVPN-enabled, but have IGMP snooping enables so they can generate IGMP reports on behalf of the clients. The two switches are dual-homed to LEAF1 and LEAF2, using all-active Ethernet Segments.

The configuration of the MAC-VRF in LEAF1 and its Ethernet Segments is as follows. The other leaf nodes are configured in a similar way.

```
--{ * candidate shared default }--[ ]--
# info network-instance "MAC-VRF 1"
  network-instance "MAC-VRF 1" {
    type mac-vrf
    interface ethernet-1/1.1 {
    }
    interface ethernet-1/2.1 {
    }
    vxlan-interface vxlan0.1 {
    }
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          admin-state enable
          vxlan-interface vxlan0.1
          evi 1
          ecmp 2
        }
      }
      bgp-vpn {
        bgp-instance 1 {
          route-distinguisher {
            rd 1:11
          }
          route-target {
            export-rt target:65011:1
            import-rt target:65011:1
          }
        }
      }
    }
    igmp-snooping { // this is the configuration required for EVPN multicast or evpn-
proxy in SRL
      admin-state enable
      trace-options {
        trace {
          packet {
            modifier egress-ingress-and-dropped
          }
        }
      }
      interface ethernet-1/1.1 {
        version 2
        send-queries true // queries are generated locally since queries are not
propagated in EVPN
      }
      interface ethernet-1/2.1 {
        version 2
        send-queries true
      }
    }
  }
}
```

```
--{ * candidate shared default }--[ ]--
# info system network-instance
```

```

system {
  network-instance {
    protocols {
      evpn {
        ethernet-segments {
          bgp-instance 1 {
            ethernet-segment ES-SROS-Client1 {
              admin-state enable
              esi 01:01:01:01:01:01:01:01:01
              multi-homing-mode all-active
              interface ethernet-1/1 {
            }
          }
          ethernet-segment ES-SROS-Client3 {
            admin-state enable
            esi 03:03:03:03:03:03:03:03:03
            multi-homing-mode all-active
            interface ethernet-1/2 {
          }
        }
      }
    }
  }
  bgp-vpn {
    bgp-instance 1 {
  }
}
}
}
}

```

Note that **send-queries true** must be configured on the interfaces of the MAC-VRF. This is because of the leaf routers not flooding the queries from the multicast routers. By enabling **send-queries** on all the interfaces, the EVPN BD becomes a virtual multicast router to the receivers. This is as described in RFC 9251.

Also note that trace-options are enabled to facilitate troubleshooting for the multicast messages.

Configuration of SWITCH1 and SWITCH2 is simpler:

```

--{ * candidate shared default }--[ ]--
# info network-instance "MAC-VRF 1"
network-instance "MAC-VRF 1" {
  type mac-vrf
  interface ethernet-1/1.1 {
  }
  interface lag1.1 {
  }
  protocols {
    igmp-snooping {
      admin-state enable
      trace-options {
        trace {
          packet {
            modifier egress-ingress-and-dropped
          }
        }
      }
    }
    interface ethernet-1/1.1 {
      static-membership-groups {
        group 239.0.0.20 {
          starg
        }
      }
    }
  }
}

```

```

                group 239.0.0.31 {
                    source 192.168.1.31 {
                        }
                    }
            }
    }
    interface lag1.1 {
        }
    }
}

```

4.8.2.3 Displaying EVPN multicast for Layer 2 information

The examples in this section display multicast information for the LEAF routers in the example configuration in [Figure 25: EVPN multicast for Layer 2](#) .

The following example shows the MFIB table in MAC-VRF 1:

```

--{ running }--[ ]--
# info from state network-instance "MAC-VRF 1" multicast-forwarding-information-base multicast-
route 0.0.0.0 group 239.0.0.20
network-instance "MAC-VRF 1" {
    multicast-forwarding-information-base {
        multicast-route 0.0.0.0 group 239.0.0.20 {
            last-update "2024-01-11T02:52:39.241Z (4 days ago)"
            outgoing-next-hop-group 203145816567 {
                forward true
            }
            outgoing-next-hop-group 203145816569 {
                forward true
            }
            outgoing-next-hop-group 203145816575 {
                forward true
            }
        }
    }
}

```

You can also use the **show network-instance multicast-forwarding-information-base** command to display the MFIB; for example:

```

--{ candidate shared default }--[ network-instance "MAC-VRF 1" ]--
A:leaf1# show multicast-forwarding-information-base
-----
IPv4 multicast FIB of network instance MAC-VRF 1
-----
+-----+-----+-----+-----+-----+-----+
| Source | Group | OIF list           | Forward | Line card | Last Update |
|        |       |                   |         | Replication |             |
|        |       |                   |         | Index     |             |
+-----+-----+-----+-----+-----+-----+
| 0.0.0.0 | 0.0.0.0 | ethernet-1/1.1    | True    | 1         | 5h52m8s ago |
| 0.0.0.0 | 239.0.0.20 | ethernet-1/2.1    | True    | 2         | 5h52m6s ago |
|         |         | vxlan:10.0.0.12:1 | True    |           |             |
| 0.0.0.0 | 239.0.0.31 | ethernet-1/1.1    | True    | 3         | 5h52m4s ago |
|         |         | ethernet-1/2.1    | True    |           |             |
|         |         | vxlan:10.0.0.12:1 | True    |           |             |
+-----+-----+-----+-----+-----+-----+

```

```

-----+-----+-----+-----+-----+-----+
IPv4 routes total : 3
IPv6 routes total : 0
-----+-----+-----+-----+-----+

```

The following command shows that all remote VTEPs support the EVPN multicast proxy features:

```

--{ candidate shared default }--[ ]--
A:leaf3# show network-instance "MAC-VRF 1" protocols igmp-snooping vxlan-destination
=====
Net-Inst "MAC-VRF 1" IGMP SNOOPING vxlan destinations
=====
-----+-----+-----+-----+-----+-----+-----+
| Vtep      | Vni | NHG Id      | Is MRouter | Is Evpn Proxy | Is SBD | Groups |
-----+-----+-----+-----+-----+-----+-----+
| 10.0.0.11 | 1   | 203145816567 | false      | true          | false  | 2      |
| 10.0.0.12 | 1   | 203145816569 | false      | true          | false  | 2      |
| 10.0.0.14 | 1   | 203145816565 | false      | true          | false  | 0      |
| 10.0.0.31 | 1   | 203145816575 | false      | true          | false  | 1      |
| 10.0.0.32 | 1   | 203145816573 | false      | true          | false  | 0      |
-----+-----+-----+-----+-----+-----+-----+
No. of destinations: 5
=====

```

The IGMP snooping state on LEAF1 is as follows:

```

--{ candidate shared default }--[ ]--
A:leaf1# info from state network-instance "MAC-VRF 1" protocols igmp-snooping
network-instance "MAC-VRF 1" {
  protocols {
    igmp-snooping {
      admin-state enable
      query-interval 125
      robust-count 2
      oper-state up
      transmitted-bgp-smet-routes 22
      proxy-membership-group-count 2
      proxy-evpn-membership-group-count 2
      trace-options {
        trace {
          packet {
            modifier egress-ingress-and-dropped
          }
        }
      }
      proxy-membership-groups { // --> proxy DB
        group 239.0.0.20 {
          filter-mode exclude
          up-time "2024-01-11T02:52:38.000Z (4 days ago)"
        }
        group 239.0.0.31 {
          filter-mode exclude
          up-time "2024-01-11T02:52:39.000Z (4 days ago)"
        }
      }
      proxy-evpn-membership-groups { // --> proxy-evpn DB used to propagate the reports with
the proper flags
        group 239.0.0.20 {
          filter-mode exclude
          up-time "2024-01-11T02:52:38.000Z (4 days ago)"

```



```

        v1-support false
        v2-support true
        v3-support false
    }
    group 239.0.0.31 {
        filter-mode exclude
        up-time "2024-01-11T02:52:39.000Z (4 days ago)"
        v1-support false
        v2-support true
        v3-support false
    }
}
interface ethernet-1/1.1 {
    router-alert-check true
    version 2
    maximum-number-groups 0
    maximum-number-sources 0
    maximum-number-group-sources 0
    query-interval 125
    query-last-member-interval 1
    query-response-interval 10
    robust-count 2
    fast-leave false
    mrouter-port false
    send-queries true
    membership-group-count 2
    is-mrouter-port false
    membership-groups {
        group 239.0.0.20 {
            group-type dynamic // --> type is dynamic because the report from SWITCH1
is hashed to LEAF1. On LEAF2, it shows as "group-type bgp-sync"
            filter-mode exclude
            expiry-time 245
            up-time "2024-01-11T02:52:38.000Z (4 days ago)"
            v1-host-timer 0
            v2-host-timer 245
            igmp-compatibility-mode 2
        }
        group 239.0.0.31 {
            group-type dynamic
            filter-mode exclude
            expiry-time 242
            up-time "2024-01-11T02:52:39.000Z (4 days ago)"
            v1-host-timer 0
            v2-host-timer 242
            igmp-compatibility-mode 2
        }
    }
}
statistics {
    received {
        general-queries 0
        group-queries 0
        group-source-queries 0
        v1-reports 0
        v2-reports 5860
        v3-reports 6
        leave-messages 3
        discarded-packets 2
        bgp-join-sync 4
        bgp-leave-sync 0
    }
    transmitted {
        general-queries 2
        group-queries 11
    }
}

```

```

        group-source-queries 0
        v1-reports 0
        v2-reports 0
        v3-reports 0
        leave-messages 0
        error-packets 0
        bgp-join-sync 6
        bgp-leave-sync 6
    }
    forwarded {
        general-queries 0
        group-queries 0
        group-source-queries 0
        v1-reports 0
        v2-reports 0
        v3-reports 0
        leave-messages 0
        unknown-type 0
        error-packets 0
    }
    error {
        bad-length 0
        unknown-type 0
        wrong-version 0
        missing-router-alert 0
        bad-encoding 0
        local-scope 0
        reached-maximum-number-groups 0
        reached-maximum-number-group-sources 0
        reached-maximum-number-sources 0
        out-of-memory-discarded-packets 0
        bad-igmp-checksum 0
        zero-source-ip-address 0
        send-query-configured-discarded-packets 0
        discarded-bgp-join-sync 0
        discarded-bgp-leave-sync 0
    }
    multicast-states {
        star-group-entries 2
        source-group-entries 0
    }
}
}
interface ethernet-1/2.1 {
    router-alert-check true
    version 2
    maximum-number-groups 0
    maximum-number-sources 0
    maximum-number-group-sources 0
    query-interval 125
    query-last-member-interval 1
    query-response-interval 10
    robust-count 2
    fast-leave false
    mrouter-port false
    send-queries true
    membership-group-count 2
    is-mrouter-port false
    membership-groups {
        group 239.0.0.20 {
            group-type dynamic
            filter-mode exclude
            expiry-time 246
            up-time "2024-01-11T02:54:43.000Z (4 days ago)"
        }
    }
}

```

```

        v1-host-timer 0
        v2-host-timer 246
        igmp-compatibility-mode 2
    }
    group 239.0.0.31 {
        group-type dynamic
        filter-mode exclude
        expiry-time 249
        up-time "2024-01-11T02:54:44.000Z (4 days ago)"
        v1-host-timer 0
        v2-host-timer 249
        igmp-compatibility-mode 2
    }
}
statistics {
    received {
        general-queries 0
        group-queries 0
        group-source-queries 0
        v1-reports 0
        v2-reports 5858
        v3-reports 0
        leave-messages 0
        discarded-packets 0
        bgp-join-sync 0
        bgp-leave-sync 0
    }
    transmitted {
        general-queries 2
        group-queries 0
        group-source-queries 0
        v1-reports 0
        v2-reports 0
        v3-reports 0
        leave-messages 0
        error-packets 0
        bgp-join-sync 2
        bgp-leave-sync 0
    }
    forwarded {
        general-queries 0
        group-queries 0
        group-source-queries 0
        v1-reports 0
        v2-reports 0
        v3-reports 0
        leave-messages 0
        unknown-type 0
        error-packets 0
    }
    error {
        bad-length 0
        unknown-type 0
        wrong-version 0
        missing-router-alert 0
        bad-encoding 0
        local-scope 0
        reached-maximum-number-groups 0
        reached-maximum-number-group-sources 0
        reached-maximum-number-sources 0
        out-of-memory-discarded-packets 0
        bad-igmp-checksum 0
        zero-source-ip-address 0
        send-query-configured-discarded-packets 0
    }
}

```

```

        discarded-bgp-join-sync 0
        discarded-bgp-leave-sync 0
    }
    multicast-states {
        star-group-entries 2
        source-group-entries 0
    }
}
vxlan-destination 10.0.0.12 vni 1 {
    index 203145159966
    is-mrouter-port false
    is-evpn-proxy true
    is-sbd false
    membership-group-count 2
    membership-groups {
        group 239.0.0.20 {
            group-type bgp-smet // --> created out of an SMET route
            filter-mode exclude
            expiry-time 2147483647
            up-time "2024-01-11T02:52:40.000Z (4 days ago)"
            v1-host-timer 0
            v2-host-timer 0
            igmp-compatibility-mode 2
        }
        group 239.0.0.31 {
            group-type bgp-smet
            filter-mode exclude
            expiry-time 2147483647
            up-time "2024-01-11T02:52:41.000Z (4 days ago)"
            v1-host-timer 0
            v2-host-timer 0
            igmp-compatibility-mode 2
        }
    }
    statistics {
        received-smet 10
        discarded-smet 0
    }
}
vxlan-destination 10.0.0.13 vni 1 {
    index 203145159969
    is-mrouter-port false
    is-evpn-proxy true
    is-sbd false
    membership-group-count 0
    statistics {
        received-smet 0
        discarded-smet 0
    }
}
vxlan-destination 10.0.0.14 vni 1 {
    index 203145159964
    is-mrouter-port false
    is-evpn-proxy true
    is-sbd false
    membership-group-count 0
    statistics {
        received-smet 0
        discarded-smet 0
    }
}
vxlan-destination 10.0.0.31 vni 1 {
    index 203145159974

```



```

GroupAddr : 239.0.0.20
2024-01-15T08:42:48.829737+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11805|D: VR 10240 IGMP
TX BGP MCAST_LEAVE_SYNC Route on network-instance "MAC-VRF 1": from local port 1/1/1
ADD (*,239.0.0.20) V2 maxRespTime 20
2024-01-15T08:42:49.812651+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11806|D: VR 10240 IGMP
TX packet on network-instance "MAC-VRF 1":
DstMac : 01:00:5e:00:00:14
Interface : ethernet-1/1.1
SrcIp : 10.0.0.11
DstIp : 239.0.0.20
Type : QUERY
MaxRespCode : 10 (0xa)
GroupAddr : 239.0.0.20
2024-01-15T08:42:55.825134+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11807|D: VR 10240 IGMP
TX BGP MCAST_LEAVE_SYNC Route on network-instance "MAC-VRF 1": from local port 1/1/1
DEL (*,239.0.0.20) maxRespTime 0
2024-01-15T08:42:55.825283+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11808|D: VR 10240 IGMP
TX BGP MCAST_JOIN_SYNC Route on network-instance "MAC-VRF 1": from local port 1/1/1
DEL (*,239.0.0.20)
2024-01-15T08:43:00.044934+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11809|D: VR 10240 IGMP
RX packet on network-instance "MAC-VRF 1":
SrcMac : 1a:9a:0b:ff:00:00
Interface : ethernet-1/1.1
SrcIp : 0.0.0.0
DstIp : 239.0.0.31
Type : V2 REPORT
GroupAddr : 239.0.0.31
2024-01-15T08:43:00.492810+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11810|D: VR 10240 IGMP
RX packet on network-instance "MAC-VRF 1":
SrcMac : 1a:61:0c:ff:00:00
Interface : ethernet-1/2.1
SrcIp : 0.0.0.0
DstIp : 239.0.0.20
Type : V2 REPORT
GroupAddr : 239.0.0.20
2024-01-15T08:43:03.196847+00:00 leaf1 local6|DEBU sr_igmp_mgr: igmp|2532|2733|11811|D: VR 10240 IGMP
RX packet on network-instance "MAC-VRF 1":
SrcMac : 1a:61:0c:ff:00:00
Interface : ethernet-1/2.1
SrcIp : 0.0.0.0
DstIp : 239.0.0.31
Type : V2 REPORT
GroupAddr : 239.0.0.31

```

4.9 BGP and routing policy extensions for EVPN

This chapter describes extensions added to SR Linux BGP and routing policy configuration to facilitate EVPN configuration.

- [BGP extensions for EVPN](#)
- [Routing policy extensions for EVPN](#)

4.9.1 BGP extensions for EVPN

SR Linux supports Multi-Protocol BGP with AFI/SAFI EVPN. The following BGP features are relevant to EVPN in a VXLAN network:

- The EVPN address family can use eBGP or iBGP.
- eBGP multihop is supported on SR Linux; local-AS override can be used for iBGP per session.
- A supported configuration or design with eBGP is eBGP with local-as override on the session to an iBGP RR.
- Rapid-update and rapid-withdrawal for EVPN family are supported and are always expected to be enabled, especially along with multihoming.



Note: Rapid-update is address-family specific, while rapid-withdrawal is generic for all address families.

- The BGP keep-all-routes option is supported for EVPN to avoid route-refresh messages attracting all EVPN routes when a policy changes or BGP-EVPN is enabled.
- The **receive-ipv6-next-hops** option does not apply to the EVPN address family.
- The **prefix-limit max-received-routes** and **threshold** options are supported for EVPN.
- The command **network-instance protocols bgp route-advertisement wait-for-fib-install** does not apply to EVPN.
- SR Linux BGP resolves BGP-EVPN routes' next-hops in the network-instance default route-table. If the next-hop is resolved, BGP can mark the route as $u^{*}>$ where u is used, $*$ is valid, and $>$ is best) and send the route to `evpn_mgr` if needed or reflected to other peers.
- BGP dynamic neighbors for IPv4 and IPv6 families are fully supported in SR Linux. EVPN-VXLAN routes can be resolved to the IPv4 addresses advertised by the dynamic neighbors.
- The command **default-received-encapsulation <mpls|vxlan>**, added to the BGP configuration in the default network-instance, indicates the encapsulation considered when the EVPN routes are received without a BGP encapsulation EC.

Most EVPN routes are usually received with a BGP encapsulation extended community (EC) that indicates the encapsulation and therefore how to interpret the value in the received Label fields of the NLRI. If no encapsulation is received, BGP validates the route as MPLS or VXLAN (or SRv6), depending on how this command is configured. The default is **vxlan**.

4.9.2 Best path selection for EVPN IFL routes

The SR Linux `fib_mgr` makes the route selection across owners for the same prefix in the network-instance route-table based on the following:

1. route-table PREFERENCE
2. METRIC
3. `source_id` (which is always zero for BGP-based routes)
4. `src_instance_id` (for leaked routes, the ID of the source network-instance)
5. `owner_id`, where `bgp` is preferred over `bgp-evpn`, and `bgp-evpn` is preferred over `bgp-ipvpn`.

For example, if the same prefix `10.0.0.0/16` is received on the same IP-VRF route table via BGP PE-CE (IPv4 family) and via EVPN-IFL, the route with lowest route table PREFERENCE wins. If both routes have the same PREFERENCE (for example, the default 170), the route with lowest METRIC wins (assuming `combined-ecmp` is not configured). The metric is a measurement of the distance to the next-hop, typically obtained from the IGP protocol that resolves the route.

As another example, if prefix 10.0.0.0/8 is received via BGP PE-CE and via EVPN IFL on the same IP-VRF, with same PREFERENCE and METRIC, the router typically selects the PE-CE route based on the lowest owner_id.

When an IP-VRF network-instance has multiple BGP paths to the same destination, the selection of the best path to install in the FIB is based on criteria the following table:

Table 5: Best path selection when IP-VRF has multiple BGP paths to a destination

Route A	Route B	Comparison steps to select the best route	ECMP supported?	
VPN-IP RDx	VPN-IP RDx	Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection	Yes, with maximum paths controlled by network-instance.protocols.bgp-ipvpn.bgp-instance.ecmp (of the IP-VRF network-instance)	bgp-ipvpn
VPN-IP RDx	VPN-IP RDy	Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection	Yes, with maximum paths controlled by network-instance.protocols.bgp-ipvpn.bgp-instance.ecmp (of the IP-VRF network-instance)	bgp-ipvpn
EVPN-IFL RDx	EVPN-IFL RDx	Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection	Yes, with maximum paths controlled by network-instance.protocols.bgp-evpn.bgp-instance.ecmp (of the IP-VRF network-instance)	bgp-evpn
EVPN-IFL RDx	EVPN-IFL RDy	Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection	Yes, with maximum paths controlled by network-instance.protocols.bgp-evpn.bgp-instance.ecmp (of the IP-VRF network-instance)	bgp-evpn
VPN-IP	PE-CE BGP	<p>When network-instance.protocols.bgp-ipvpn.combined-ecmp is present:</p> <ul style="list-style-type: none"> Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection <p>When network-instance.protocols.bgp-ipvpn.combined-ecmp is not present:</p> <ul style="list-style-type: none"> Tie break based on best route preference, then lowest route metric, then 	<p>Yes, only if network-instance.protocols.bgp-ipvpn.combined-ecmp is present. The maximum paths is the minimum of:</p> <ul style="list-style-type: none"> network-instance.protocols.bgp-afi-safi.multipath.maximum-paths (of the IP-VRF network-instance) network-instance.protocols.bgp-ipvpn.bgp-instance.ecmp (of the IP-VRF network-instance) 	bgp

Route A	Route B	Comparison steps to select the best route	ECMP supported?	
		lowest route owner (BGP < EVPN-IFL < VPN-IP)		
EVPN-IFL	PE-CE BGP	<p>When network-instance.protocols.bgp-vpn.combined-ecmp is present:</p> <ul style="list-style-type: none"> Full best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection <p>When network-instance.protocols.bgp-vpn.combined-ecmp is NOT present:</p> <ul style="list-style-type: none"> Tie break based on best route preference, then lowest route metric, then lowest route owner (BGP < EVPN-IFL < VPN-IP) 	<p>Yes, only if network-instance.protocols.bgp-vpn.combined-ecmp is present. The maximum paths is the minimum of:</p> <ul style="list-style-type: none"> network-instance.protocols.bgp.afi-safi.multipath.maximum-paths (of the IP-VRF network-instance) network-instance.protocols.bgp-evpn.bgp-instance.ecmp (of the IP-VRF network-instance) 	bgp
PE-CE BGP	PE-CE BGP	Full BGP best path selection controlled by IP VRF network-instance network-instance.protocols.bgp.best-path-selection	Yes, with maximum paths determined only by network-instance.protocols.bgp.afi-safi.multipath.maximum-paths (of the IP-VRF network-instance)	bgp
VPN-IP	VPN-IFL	Full BGP best path selection controlled by DEFAULT network-instance network-instance.protocols.bgp.best-path-selection	<p>Yes, only if network-instance.protocols.bgp-vpn.combined-ecmp is present. The maximum paths is the minimum of:</p> <ul style="list-style-type: none"> network-instance.protocols.bgp-ipvpn.bgp-instance.ecmp (of the IP-VRF network-instance) network-instance.protocols.bgp-evpn.bgp-instance.ecmp (of the IP-VRF network-instance) 	bgp-evpn

4.9.3 Routing policy extensions for EVPN

SR Linux includes support for applying routing policies to EVPN routes. You can specify the following match conditions in a policy statement:

- match routes based on EVPN route type
- match routes based on IP addresses and prefixes via prefix-sets for route types 2 and 5
- match routes based BGP encapsulation extended community
- match routes based on configured admin-tags

For information about configuring routing policies on SR Linux, see the *SR Linux Configuration Basics Guide*.

The following considerations apply to routing policies for EVPN:

- For IBGP neighbors, EVPN routes are imported and exported without explicit configuration of any policy at either the BGP or network-instance level.
- For EBGP neighbors, by default, routes are imported or exported based on the **ebgp-default-policy** configuration.
- When an explicit import/export route-target is configured in a network-instance bgp-instance, and an import/export policy is also configured on the same bgp-instance, the configured policy is used, and its route-target is added to the imported/exported route.
- When a network-instance policy and a peer policy are applied, they are executed as follows:
 - For export, the network-instance export policy is applied first, and the peer policy is applied afterwards (sequentially).
 - For import, peer import policy is applied first and network-instance import policy is applied afterwards (sequentially).
- Only one network-instance export and import policy is supported.

4.10 Route internal tags (tag-sets in routing policies)

Route tags are pieces of information or identifiers attached to individual routes. They can be either external (protocol) tags or internal tags:

- External (protocol) tags are sent on the wire and are typically part of a routing protocol; for example, IS-IS route tags and OSPF route tags.
- Internal tags are not sent on the wire, and they have meaning only within an individual system. Internal tags are assigned to routes by policies or specific commands, and are used by policies to match on a group of routes. Internal tags are typically used to facilitate the matching of multiple routes at the same time or propagate route attributes between different routing protocols.

This chapter describes SR Linux support for internal tags and how to configure them.

About internal tags

In SR Linux, route internal tags are data structures that are attached to IP routes in the FIB or EVPN-specific routes. Internal tags are useful for carrying protocol information such as communities, area-id, as-path, and so on, from one routing protocol context to another. For example, BGP can push its attribute

fields, such as local preference, into the FIB as an internal tag. On export, the internal tag associated with the route can be matched by export policies to map the tag into any other metric for a routing protocol.

Every FIB route can be associated with an array of tags. The number of array elements is limited to a maximum of 2 in the current release.

To define internal tags in SR Linux, you configure tag-sets at the `routing-policy` level, and refer to the tag-sets by name at other configuration levels. See [Configuring route internal tags](#).

SR Linux applications that support internal tags

The following SR Linux applications can write internal tags:

- **EVPN and Layer 2 proxy-ARP/ND**
EVPN and Layer 2 proxy-ARP/ND support writing internal tags for all MAC/IP advertisement routes generated out of the active MACs in the `mac-table` or IP-to-MAC bindings in the proxy-ARP/ND tables, respectively.
- **Layer 3 ARP/ND**
SR Linux supports writing internal tags into ARP/ND entries that generate an EVPN MAC/IP advertisement route and also into the ARP/ND entries that generate an `arp-nd` host route into the route-table.
- **Layer 3 routes**
In SR Linux, BGP can write internal tags into the route-table entries for a routed network-instance. In this case, all routes to be exported as prefix routes are assigned the configured internal tag.

The route-table includes a state leaf-list with the allocated `internal-tags` tag values that indicate whether a specific route entry is associated with an internal tag.

Routing policies and internal tags

An import or export policy can write an internal tag as a configured `action` or `default-action`. These actions are supported on BGP policies applied to the supported families; that is, `ipv4-unicast`, `ipv6-unicast`, and `evpn`.

When internal tags are used in BGP import policies, the policies can be applied at the default network-instance level (including global, per-group, or per-peer policies), MAC-VRF network-instance level, or IP-VRF network-instance level.

When internal tags are used in BGP export policies, the policies can be applied at the MAC-VRF or IP-VRF network-instance level, or IP-VRF/default network-instance level for inter-instance policies.

When an import policy matches an IPv4 BGP route, IPv6 BGP route, or EVPN IP prefix route (route type 5) and adds a tag-set, the corresponding route created in the route-table contains the internal tag and is exposed in the route-table entry `internal-tags`.

When the import policy matches an EVPN MAC/IP advertisement route (route type 2) and adds a tag-set as an action, the corresponding created MAC and/or ARP/ND entry is associated with the internal tag. Matching on any other EVPN route type and adding an internal tag has no effect, since the rest of the EVPN route types do not create entries into the MAC table, ARP/ND table, or route table.

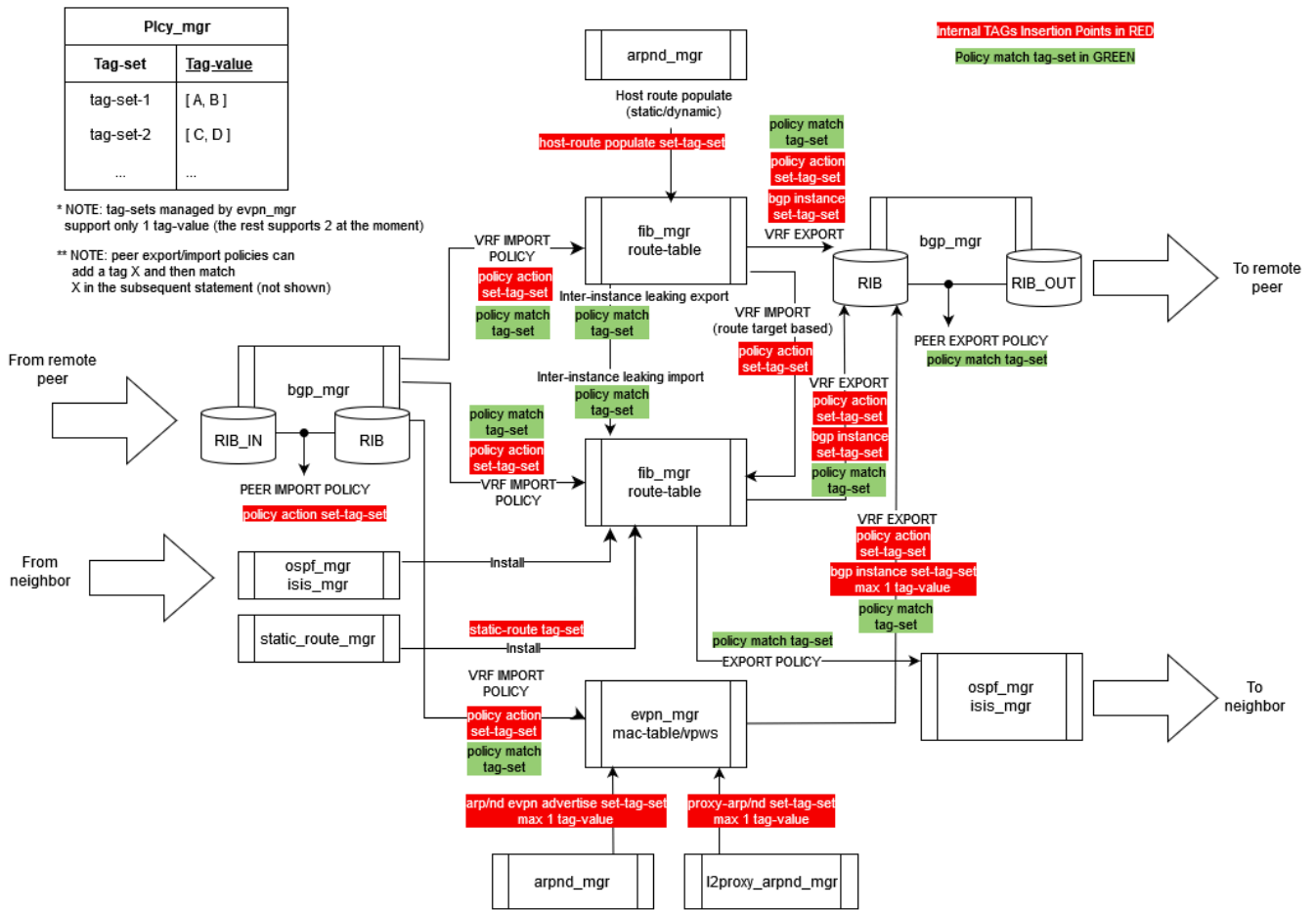
A MAC/IP VRF export policy can add a tag-set as an action in the following cases

- On a MAC-VRF, the `bgp-vpn.bgp-instance.export-policy` can match EVPN MAC/IP routes (with or without IP address) generated for dynamic or static entries in the `mac-table` or proxy-ARP/ND table; that is, basically for any entry that generates a MAC/IP advertisement route in the context of the MAC-VRF. In those policies, a tag-set can be added. The same tag-set can be matched on peer/group-level export policies in the default network-instance.

- On an IP-VRF, the `bgp-vpn.bgp-instance.export-policy` can match any route-table entry that generates an EVPN IP prefix route in the context of the IP-VRF. In these policies, a tag-set can be added as an action, and the same tag-set can be matched by a peer/group-level export policy in the default network-instance.
- On an IP-VRF or default network-instance, the `inter-instance-policies.export-policy` can match any route-table entry and make those entries available for leaking. The internal tag can be used by an `inter-instance-policies.import-policy` to match the routes that are actually leaked into the destination Layer 3 network-instance.

Insertion points for internal tags

The following diagram shows the SR Linux applications that can write an internal tag, insertion points for the internal tags, and the policy points where internal tag matching can take place.



4.10.1 Order of precedence when multiple applications write internal tags

As described in the previous section, multiple SR Linux applications can write internal tags for a route. For example, a newly created ARP-ND host 10.0.0.1/32 route can be associated with tag-set TAG-1, and at the same time an export policy applied on IP-VRF-1, where 10.0.0.1/32 is installed, may have an action to write tag-set TAG-2 on the same route.

In this case, SR Linux applies precedence rules to determine the tag-set that gets applied to the route. The following sections describe the precedence rules applied when multiple applications attempt to write an internal tag on the same route.

Writing a tag for MAC/IP routes generated for a MAC-VRF

The order in which the internal tag is applied is as follows:

(proxy-arp/nd tag (mac-vrf) + vrf-export-policy tag (mac-vrf level)) > (bgp-evpn.bgp-instance tag (mac-vrf) +vrf-export-policy tag (mac-vrf level))

That is, up to two tag-values can be written. However, if the route already has a tag-value, then the bgp-evpn.bgp-instance tag is not used. The vrf-export-policy tag action always appends to the route's tag (if already present on the route), or the default tag (under the bgp-instance) if configured.

Writing a tag for MAC/IP routes generated for a MAC-VRF with IRB

(arp/nd evpn tag (irb) + vrf-export-policy tag (mac-vrf level)) > (bgp-evpn.bgp-instance tag (mac-vrf) +vrf-export-policy tag (mac-vrf level))

In the same way as proxy-arp/nd tags, up to two tag-values can be written, and if the route already has a tag-value, then the bgp-evpn.bgp-instance tag is not used. The vrf-export-policy tag action always appends to the route's tag (if already present on the route), or the default tag (under the bgp-instance) if configured.

Writing a tag for ARP-ND host routes in an IP-VRF

Tags are appended until a maximum of two is reached, so the vrf-export policy action does not overwrite the tag if that is already present. The default tag of the instance is used when arpd does not add a tag.

A peer-export policy uses this precedence:

(arp/nd host-route tag (irb) + vrf-export policy tag) > (bgp-evpn.bgp-instance tag (ip-vrf) + vrf-export policy tag)

Writing a tag for IP routes in an IP-VRF (irrespective of the owner)

Tag-values are appended until a maximum of two is reached, so the vrf-export policy action does not overwrite the import-policy tag if that is already present. The default tag of the instance is used when import-policy does not add a tag.

A peer-export policy uses this precedence:

(peer-import-policy-tag + vrf-export policy tag) > (bgp-evpn.bgp-instance tag (ip-vrf) + vrf-export policy tag)

In cases where the routes are static routes, IS-IS routes, OSPF routes, and so on, when exported into EVPN IP prefix routes, there is no peer-import-policy.

In cases where there are BGP PE-CE routes that are imported into the IP-VRF route-table and readvertised as EVPN IP prefix routes, there are peer-import-policies, and the preference is as above.

Writing a tag for MAC/IP routes in MAC-VRF or routes in IP-VRF via multiple import policies

If multiple import-policies match the same route and try to write a tag, writing is allowed until a maximum of two tag-values is reached. Note that applying "multiple import-policies" refers to policy chaining at the peer-import level.

Writing a tag for routes via vrf-import and peer import policies

If both vrf-import and peer-import policies apply to a route, and each step inserts a tag-set, then the tag-set applied as a result of the vrf-import policy takes precedence.

Writing a tag for BGP routes in an ECMP set

If BGP creates an ECMP set for a prefix, when installing such a prefix in the FIB, BGP copies the tag-set of the best route into fib_mgr. This is what other protocols would see when importing the prefix from fib_mgr.

This also applies to BGP VPN routes that are received with different RDs and combined in an ECMP set.

4.10.2 Configuring route internal tags

Procedure

To configure route internal tags, you configure a tag-set and configure one or more applications to write the tag-set to routes.

Example: Configure a tag-set

```
--{ * candidate shared default }--[ ]--
# info routing-policy
  routing-policy {
    tag-set ts1 {
      tag-value [
        1
        2
      ]
    }
  }
}
```

Example: Match a tag-set in a routing policy

```
--{ +* candidate shared default }--[ ]--
# info routing-policy policy match_tag_1
  routing-policy {
    policy match_tag_1 {
      statement 1 {
        match {
          internal-tags {
            tag-set [
              ts1
            ]
          }
        }
      }
    }
  }
}
```

Example: Write a tag-set as an action in a routing policy

```
--{ * candidate shared default }--[ ]--
# info routing-policy policy d1
  routing-policy {
    policy t1 {
      default-action {
        internal-tags {
          set-tag-set [

```


Example: Write internal tags into ARP/ND entries that generate an EVPN MAC/IP advertisement route

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv4 arp evpn
  interface ethernet-1/1 {
    subinterface 1 {
      ipv4 {
        arp {
          evpn {
            advertise dynamic {
              internal-tags {
                set-tag-set [
                  ts1
                ]
              }
            }
          }
        }
      }
    }
  }
}
```

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv6 neighbor-discovery
  interface ethernet-1/1 {
    subinterface 1 {
      ipv6 {
        neighbor-discovery {
          evpn {
            advertise dynamic {
              internal-tags {
                set-tag-set [
                  ts1
                ]
              }
            }
          }
        }
      }
    }
  }
}
```

Example: Write internal tags into ARP/ND entries that generate an ARP/ND host route into the route table

```
--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv4 arp
  interface ethernet-1/1 {
    subinterface 1 {
      ipv4 {
        arp {
          host-route {
            populate evpn {
              internal-tags {
                set-tag-set [
                  ts1
                ]
              }
            }
          }
        }
      }
    }
  }
}
```



```

    }
  }
}

--{ * candidate shared default }--[ ]--
# info interface ethernet-1/1 subinterface 1 ipv6 neighbor-discovery
  interface ethernet-1/1 {
    subinterface 1 {
      ipv6 {
        neighbor-discovery {
          host-route {
            populate evpn {
              internal-tags {
                set-tag-set [
                  ts1
                ]
              }
            }
          }
        }
      }
    }
  }
}

```

Example: Write internal tag into routes to be exported as EVPN IP prefix routes

```

--{ * candidate shared default }--[ ]--
# info network-instance ip-vrf-1 protocols bgp-evpn bgp-instance 1
  network-instance ip-vrf-1 {
    protocols {
      bgp-evpn {
        bgp-instance 1 {
          internal-tags {
            set-tag-set [
              ts1
            ]
          }
        }
      }
    }
  }
}

```

5 IP-VPN services



Note: IP-VPN services are supported on 7250 IXR and 7730 SXR platforms.

This chapter describes the components of IP-VPN services on SR Linux. It contains the following topics:

- [Configuring IP-VPN services](#)

5.1 Configuring IP-VPN services

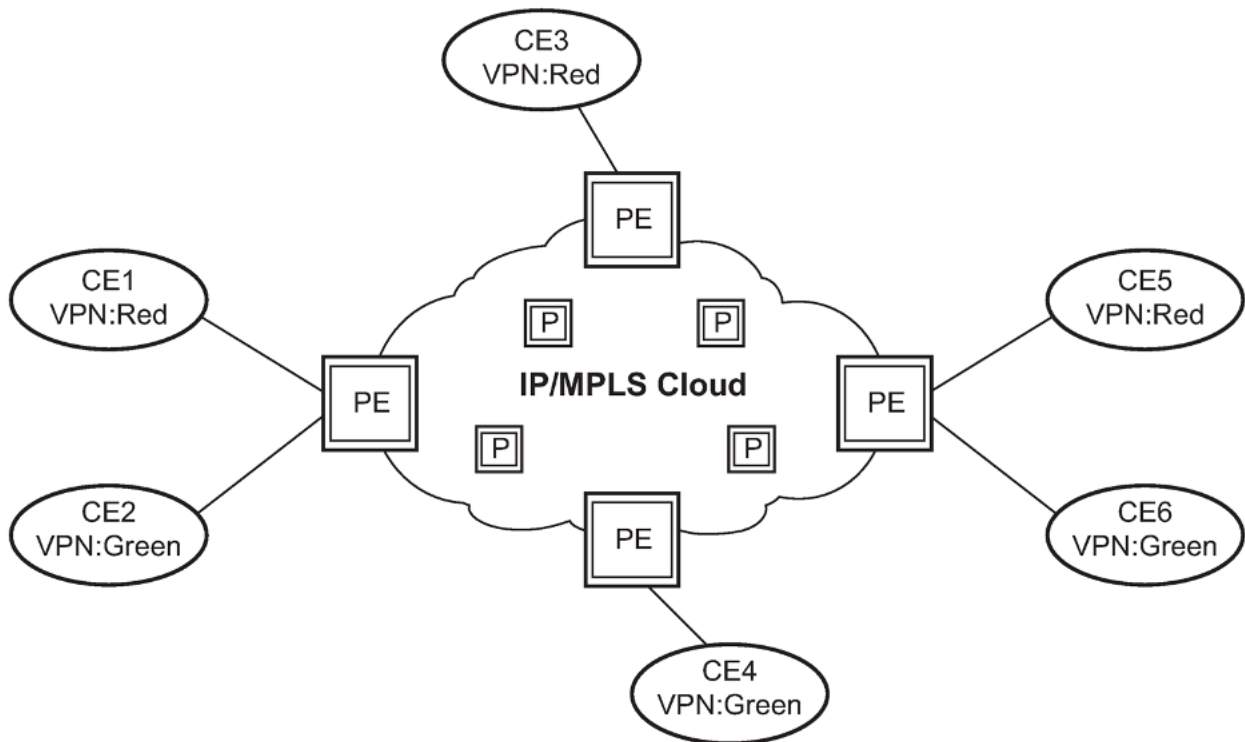
IP-VPN services use a combination of MP-BGP and MPLS to distribute IPv4/v6 routing information and provide Layer 3 VPN services.

Each IP-VPN consists of a set of customer end points connected to one or more PE routers. Each associated PE router maintains a separate IP forwarding table for each IP-VPN instance. Additionally, the PE routers exchange the routing information configured or learned from all customer sites via MP-BGP peering. Each route exchanged via the MP-BGP protocol includes a Route Distinguisher (RD), which identifies the IP-VPN association and handles any potential IP address overlap.

Multi-Protocol BGP (MP-BGP) is used to exchange the routes of a particular VPN among the PE routers that are attached to that VPN. This route exchange is done in a way that ensures that routes from different VPNs remain distinct and separate, even if two VPNs have an overlapping address space. When BGP distributes a VPN route it also distributes an MPLS label for that route to identify the advertising IP-VPN instance.

Before a customer data packet travels across the service provider's backbone, it is encapsulated with the MPLS label that corresponds, in the customer's IP-VPN, to the route that best matches the packet's destination address. The MPLS packet is further encapsulated with one or more MPLS labels corresponding to the resolving MPLS path to deliver the packet to the intended egress PE router. The following figure displays an IP-VPN network diagram example.

Figure 26: IP Virtual Private Network



OSSG024

Example: IP-VPN configuration

The following is an example of an IP-VPN configuration.

```
--{ * candidate shared default }--[ ]--
A:srl1# info interface ethernet-1/2
  interface ethernet-1/2 {
    admin-state enable
    subinterface 1 {
      type routed
      admin-state enable
      ipv4 {
        address 10.30.30.1/24 {
        }
      }
    }
  }
A:srl1# info network-instance Base
network-instance Base {
  type default
  protocols {
    bgp {
      admin-state enable
      autonomous-system 65550
      router-id 10.10.10.1
      afi-safi l3vpn-ipv4-unicast {
        admin-state enable
      }
    }
    group base-group {
```

```

        }
        neighbor 10.10.10.2 {
            peer-group base-group
        }
    }
}

--{ * candidate shared default }--[ ]--
A:sr11# info network-instance ip-vrf-red
network-instance ip-vrf-red {
    type ip-vrf
    interface ethernet-1/2 {
        interface-ref {
            interface ethernet-1/2
            subinterface 1
        }
    }
    protocols {
        bgp-ipvpn {
            bgp-instance 1 {
                admin-state enable
                ecmp 8
                mpls {
                    next-hop-resolution {
                        allowed-tunnel-types [
                            ldp
                            sr-isis
                        ]
                    }
                }
            }
        }
    }
}
bgp {
    admin-state enable
    autonomous-system 65551
    router-id 10.10.10.1
    afi-safi ipv4-unicast {
        admin-state enable
    }
    group ip-vrf-red-peers {
        admin-state enable
        afi-safi ipv4-unicast {
        }
    }
    neighbor 10.10.10.3 {
        peer-group ip-vrf-red-peers
    }
}
bgp-vpn {
    bgp-instance 1 {
        route-distinguisher {}
        route-target {}
    }
}
}
}

```

6 Multi-instance IP-VRF



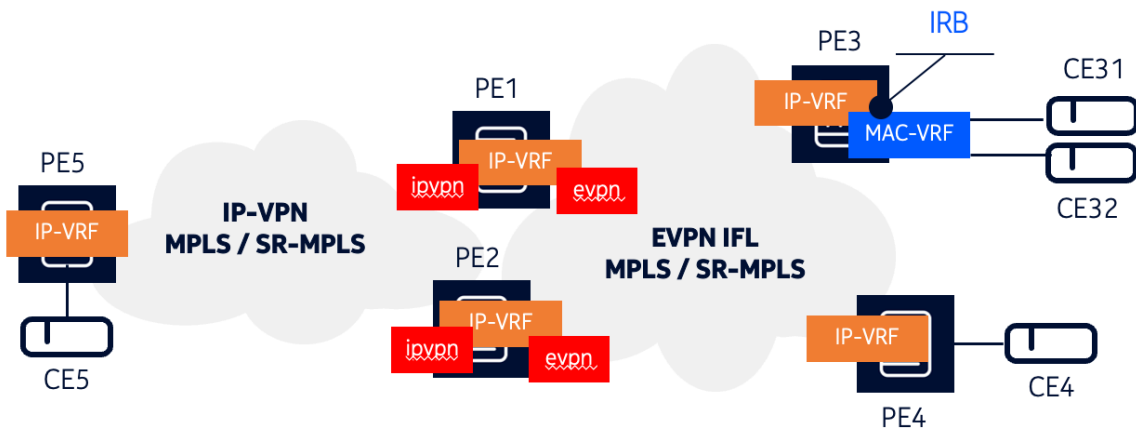
Note: Multi-instance IP-VRF is supported on 7250 IXR and 7730 SXR platforms.

Multi-instance IP-VRF refers to an IP-VRF network instance that supports multiple instances of BGP-VPN (a generalization of BGP families that are used for VPN services, EVPN, or IP-VPN). Multi-instance IP-VRF is configured using the command **network-instance protocols bgp-vpn bgp-instance**.

Multi-instance IP-VRFs are used on service gateway PE routers that stitch multiple domains together. These domains can use different BGP Layer 3 families or different tunnels in the data path.

The following figure illustrates PE1 and PE2 as service gateways that connect an EVPN-IFL domain to an IP-VPN domain. PE1 and PE2 use a single IP-VRF with two BGP-VPN instances.

Figure 27: Multi-instance IP-VRF structure



The following example shows the configuration of PE1.

```
--[ network-instance IP-VRF1]--
A:pe1# info
type ip-vrf
interface lo0.1 {
}
protocols {
  bgp-evpn {
    bgp-instance 2 {
      encapsulation-type mpls
      evi 10
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [ bgp te-policy-sr-policy-mpls-uncolored ]
        }
      }
    }
  }
}
```

```

    }
  }
  bgp-ipvpn {
    bgp-instance 1 {
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [ ldp sr-isis ]
        }
      }
    }
  }
  bgp-vpn {
    bgp-instance 1 {
      route-distinguisher { rd 1:1 }
      route-target {
        export-rt target:64500:10
        import-rt target:64500:10
      }
    }
    bgp-instance 2 {}
  }
}

```

The following shows the configuration of PE2.

```

--[ network-instance IP-VRF1]--
A:pe2# info
type ip-vrf
interface lo0.1 {}
protocols {
  bgp-evpn {
    bgp-instance 2 {
      encapsulation-type mpls
      evi 10
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [ bgp te-policy-sr-policy-mpls-uncolored ]
        }
      }
    }
  }
  bgp-ipvpn {
    bgp-instance 1 {
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [ ldp sr-isis ]
        }
      }
    }
  }
  bgp-vpn {
    bgp-instance 1 {
      route-distinguisher { rd 1:1 }
      route-target {
        export-rt target:64500:10
        import-rt target:64500:10
      }
    }
    bgp-instance 2 {}
  }
}

```

Up to two BGP instances are supported in the same network instance of type **ip-vrf**. One BGP instance is used for EVPN-IFL and the other BGP instance is used for IP-VPN.

If the same route is received on the same IP-VRF via different BGP owners, such as IP-VPN, EVPN-IFL, or BGP PE-CE, the selection and combined ECMP set is completed based on the best path selection for EVPN IFL routes. For more information, see [Best path selection for EVPN IFL routes](#).

7 Next-hop-self route reflector and inter-AS option B



Note: Next-hop-self route reflectors and inter-AS option B are supported on 7250 IXR and 7730 SXR platforms.

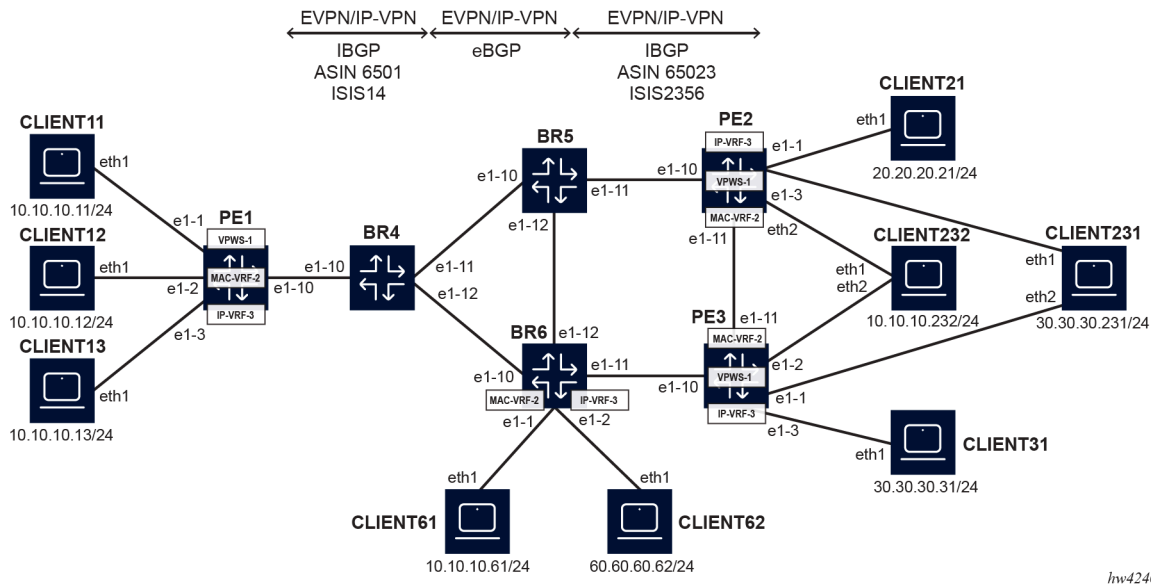
Next-hop-self route reflectors (NHS-RR) are used in BGP networks to reduce the need for fully meshed iBGP connections within a single autonomous system (AS). In a fully meshed AS, iBGP routers do not advertise routes to their neighbors, while NHS-RR allows a route reflector (RR) to advertise learned iBGP routes to iBGP neighbors with its own address as the next hop. This feature connects different domains within the same AS and improves the overall scale.

Inter-AS option B is a method for interconnecting VPN sites located in different ASes. Using this method, autonomous system border routers (ASBRs) are directly connected and routes are exchanged on a single interface.

EVPN IFL and VPN-IPv4/IPv6 routes both support NHS-RR capability and inter-AS option B.

The following figure shows the configuration of NHS-RR and inter-AS option B:

Figure 28: EVPN and IP-VPN services on inter-AS option B



The preceding figure shows two different ASes that are connected via inter-AS option B ASBRs br4, br5, and br6. This method allows for the extension of EVPN and IP-VPN services across different MPLS or segment routing MPLS domains without providing services on border routers.

Egress PEs advertise EVPN and IP-VPN routes to adjacent border routers.

Border routers carry out the following functions:

1. Import and redistribute routes to the remote border routers or local PEs where the addresses of the remote border routers are used as the next hops and their own service MPLS labels are used.

2. Program a label swap operation so that the ingress traffic service label is looked up and packets are forwarded with a new service label.

The following shows the configuration of the default network instance of pe1 in [Figure 28: EVPN and IP-VPN services on inter-AS option B](#), which is similar to the configuration of the other PEs in the figure:

```
--{ + candidate shared default }--[ network-instance default ]--
A:pe1# info
type default
interface ethernet-1/10.0 {
}
interface system0.0 {
}
protocols {
  bgp {
    admin-state enable
    autonomous-system 65001
    router-id 10.0.0.1
    bgp-label {
      bgp-ipvpn {
        next-hop-resolution {
          ipv4-next-hops {
            tunnel-resolution {
              allowed-tunnel-types [
                ldp
                sr-isis
              ]
            }
          }
        }
      }
    }
  }
  ebgp-default-policy {
    import-reject-all false
    export-reject-all false
  }
  afi-safi evpn {
    admin-state enable
    evpn {
      keep-all-routes true
      rapid-update true
    }
  }
  afi-safi l3vpn-ipv4-unicast {
    admin-state enable
    l3vpn-ipv4-unicast {
      keep-all-routes true
      rapid-update true
    }
  }
  group overlay-ibgp {
    peer-as 65001
    afi-safi evpn {
    }
    afi-safi l3vpn-ipv4-unicast {
    }
    timers {
      connect-retry 1
      minimum-advertisement-interval 1
    }
    trace-options {
      flag update {
        modifier detail
      }
    }
  }
}
```

```

    }
  }
}
neighbor 10.0.0.4 {
  peer-group overlay-ibgp
}
}
ldp {
  admin-state enable
  dynamic-label-block range-1-ldp
  discovery {
    interfaces {
      interface ethernet-1/10.0 {
        ipv4 {
          admin-state enable
        }
      }
    }
  }
}
}
isis {
  dynamic-label-block range-3-srgb
  instance i14 {
    admin-state enable
    instance-id 1
    level-capability L2
    iid-tlv true
    net [
      49.0001.0000.0000.0001.00
    ]
    trace-options {
      trace [
        adjacencies
        interfaces
        packets-all
      ]
    }
    segment-routing {
      mpls {
        dynamic-adjacency-sids {
          all-interfaces true
        }
      }
    }
    interface ethernet-1/10.0 {
      circuit-type point-to-point
      ipv4-unicast {
        admin-state enable
      }
    }
    interface system0.0 {
      passive true
      ipv4-unicast {
        admin-state enable
      }
    }
  }
}
}
}
segment-routing {
  mpls {
    global-block {
      label-range range-2-srgb
    }
  }
}

```

```

        local-prefix-sid 1 {
            interface system0.0
            ipv4-label-index 1
        }
    }
}

```

The following shows the configuration of the default network instance of RR pe2 in [Figure 28: EVPN and IP-VPN services on inter-AS option B](#):

```

--{ + candidate shared default }--[ network-instance default protocols bgp ]--
A:pe2# info
admin-state enable
autonomous-system 65023
router-id 10.0.0.2
bgp-label {
    bgp-ipvpn {
        next-hop-resolution {
            ipv4-next-hops {
                tunnel-resolution {
                    allowed-tunnel-types [
                        ldp
                        sr-isis
                    ]
                }
            }
        }
    }
}
ebgp-default-policy {
    import-reject-all false
    export-reject-all false
}
afi-safi evpn {
    admin-state enable
    evpn {
        keep-all-routes true
        rapid-update true
    }
}
afi-safi l3vpn-ipv4-unicast {
    admin-state enable
    l3vpn-ipv4-unicast {
        keep-all-routes true
        rapid-update true
    }
}
group overlay-ibgp {
    peer-as 65023
    afi-safi evpn {
    }
    afi-safi l3vpn-ipv4-unicast {
    }
    route-reflector {
        client true
        cluster-id 2.2.2.2
    }
    timers {
        connect-retry 1
        minimum-advertisement-interval 1
    }
    trace-options {
        flag update {

```

```

        modifier detail
    }
}
neighbor 10.0.0.3 {
    peer-group overlay-ibgp
}
neighbor 10.0.0.5 {
    peer-group overlay-ibgp
}
neighbor 10.0.0.6 {
    peer-group overlay-ibgp
}

```

The following commands enable NHS-RR and inter-AS option B functionality on SR Linux:

- **inter-as-vpn**
- **next-hop-self-route-reflector**

These commands affect all EVPN routes and trigger `bgp_mgr` to swap the service label for all EVPN MPLS routes. These commands also change the next hop to self in all routes with MPLS encapsulation.

For border router configuration, IGP is not enabled in interfaces to other border routers.

The **inter-as-vpn true** command allows received EVPN/IP-VPN routes to be retained in the BGP RIB and propagated to any eBGP or iBGP peer. To ensure label allocation, a dynamic label block must be configured for border routers. Label allocation re-advertises a received route into the adjacent AS with a local network instance MPLS label and ensures an MPLS label swap operation is completed.

Configure a dynamic label block using the following configuration:

```

--{ * candidate shared default }--[ ]--
A:srl1# info network-instance default
    network-instance default {
        type default
        protocols {
            bgp {
                bgp-label {
                    bgp-vpn {
                        dynamic-label-block range-6-bgp-lu
                    }
                }
            }
        }
    }

```

The label block is shared by the EVPN inter-AS model B and EVPN NHS-RR features.



Note:

The **inter-as-vpn true** command has the same function as the **keep-all-routes** command for keeping the routes in the RIB.

The following shows the configuration of the default network instance of border router br4 in [Figure 28: EVPN and IP-VPN services on inter-AS option B](#):

```

--{ + candidate shared default }--[ network-instance default ]--
A:br4# info
    type default
    interface ethernet-1/10.0 {
    }
    interface ethernet-1/11.0 {
    }
    interface ethernet-1/12.0 {
    }

```

```

interface system0.0 {
}
protocols {
  bgp {
    autonomous-system 65001
    router-id 10.0.0.4
    bgp-label {
      bgp-vpn {
        dynamic-label-block range-6-bgp-lu
      }
      bgp-ipvpn {
        next-hop-resolution {
          ipv4-next-hops {
            tunnel-resolution {
              allowed-tunnel-types [
                ldp
                sr-isis
              ]
            }
          }
        }
      }
    }
  }
  ebgp-default-policy {
    import-reject-all false
    export-reject-all false
  }
  afi-safi evpn {
    admin-state enable
    evpn {
      inter-as-vpn true
      rapid-update true
      default-received-encapsulation mpls
      next-hop-resolution {
        ipv4-next-hops {
          tunnel-resolution {
            allowed-tunnel-types [
              bgp
              ldp
              sr-isis
            ]
          }
        }
      }
    }
  }
  afi-safi l3vpn-ipv4-unicast {
    admin-state enable
    l3vpn-ipv4-unicast {
      inter-as-vpn true
      rapid-update true
    }
  }
  group overlay-ebgp {
    peer-as 65023
    afi-safi evpn {
    }
    timers {
      connect-retry 1
      minimum-advertisement-interval 1
    }
    trace-options {
      flag update {
        modifier detail
      }
    }
  }
}

```

```

    }
  }
}
group overlay-ibgp {
  peer-as 65001
  afi-safi evpn {
  }
  timers {
    connect-retry 1
    minimum-advertisement-interval 1
  }
  trace-options {
    flag update {
      modifier detail
    }
  }
}
neighbor 10.4.5.2 {
  peer-group overlay-ebgp
}
neighbor 10.4.6.2 {
  peer-group overlay-ebgp
}
neighbor 10.0.0.1 {
  peer-group overlay-ibgp
}
}
ldp {
  admin-state enable
  dynamic-label-block range-1-ldp
  discovery {
    interfaces {
      interface ethernet-1/10.0 {
        ipv4 {
          admin-state enable
        }
      }
      interface ethernet-1/11.0 {
        ipv4 {
          admin-state enable
        }
      }
      interface ethernet-1/12.0 {
        ipv4 {
          admin-state enable
        }
      }
    }
  }
}
}
isis {
  dynamic-label-block range-3-srgb
  instance i14 {
    admin-state enable
    instance-id 1
    level-capability L2
    iid-tlv true
    net [
      49.0001.0000.0000.0004.00
    ]
    segment-routing {
      mpls {
        dynamic-adjacency-sids {
          all-interfaces true
        }
      }
    }
  }
}
}

```



```

        end-label 122201
    }
}
services {
    evpn {
        dynamic-label-block range-4-evpn
    }
    network-instance {
        dynamic-label-block range-5-services
    }
}
}

```

The **next-hop-self-route-reflector** command requires the configuration of a border router as a RR. The behavior of this command is equivalent to the **inter-as-vpn** command, with one difference: the use of **next-hop-self-route-reflector** allows the border router to receive and readvertise routes to RR clients within the same AS.

BGP next-hop resolution for EVPN/IP-VPN routes

NHS-RR and inter-AS option B routes all use the **next-hop-resolution** grouping. This grouping is only used when the following options are enabled:

- **inter-as-vpn**
- **nhsrr-evpn**

The **next-hop-resolution** command allows the disabling of route resolution. By default, routes are not ignored. If a longest prefix match (LPM) FIB lookup provides a route that is not local or static, and there is no tunnel, the route is unresolved. This command also permits the:

- resolution of tunnels through a fallback to FIB lookup
- resolution of local and static routes to the next hop in the absence of resolving tunnels

BGP receives next-hop resolution information from the following contexts:

- For IP-VPN:

```

--{ * candidate shared default }--[ ]--
A:srl1# info network-instance default
network-instance default {
    protocols {
        bgp {
            bgp-label {
                bgp-ipvpn {
                    next-hop-resolution {
                    }
                }
            }
        }
    }
}

```

- For EVPN:

```

--{ * candidate shared default }--[ ]--
A:srl3# info network-instance default
network-instance default {
    protocols {
        bgp {
            afi-safi evpn {
                next-hop-resolution {
                }
            }
        }
    }
}

```



```

    }
  }
}

--{ candidate shared default }--[ ]--
A:br4# info from state network-instance default route-table next-hop-group 373030816119
network-instance default {
  route-table {
    next-hop-group 373030816119 {
      backup-next-hop-group 0
      fib-programming {
        last-successful-operation-type add
        last-successful-operation-timestamp 2024-04-10T11:57:16.328Z
        pending-operation-type none
        last-failed-operation-type none
      }
      next-hop 0 {
        next-hop 373030816115
        resolved true
      }
    }
  }
}

// In this example, the next hop is resolved to a local route 10.4.5.0 (ASBRs using single hop eBGP
// session)

--{ candidate shared default }--[ ]--
A:br4# info from state network-instance default route-table next-hop 373030816115
network-instance default {
  route-table {
    next-hop 373030816115 {
      type indirect
      ip-address 10.4.5.2
      resolving-route {
        ip-prefix 10.4.5.0/30
        route-type local
        route-owner net_inst_mgr
      }
      mpls {
        pushed-mpls-label-stack [
          122010
        ]
      }
    }
  }
}
}

```

Example: MPLS route table

The following example displays the the MPLS route table which summarizes the swap operation, incoming label, outgoing label, and associated next hop.

```

--{ candidate shared default }--[ ]--
A:br4# show network-instance default route-table mpls

```

Label	Operation	Type	Next Net-Inst	Next-hop IP (Type)	Next-hop Subinterface	Next-hop MPLS labels
100	POP	ldp	default			
100002	SWAP	sr-mpls	N/A	10.1.4.1 (mpls)	ethernet-1/10.0	100002
100005	POP	sr-mpls	default			

120002	SWAP	sr-mps	N/A	10.1.4.1 (mps)	ethernet-1/10.0	IMPLICIT_NULL
122001	SWAP	bgp	N/A	10.0.0.1 (indirect)		1002
122002	SWAP	bgp	N/A	10.0.0.1 (indirect)		1001
122003	SWAP	bgp	N/A	10.0.0.1 (indirect)		1003
122004	SWAP	bgp	N/A	10.4.5.2 (indirect)		122008
122005	SWAP	bgp	N/A	10.4.5.2 (indirect)		122011
122006	SWAP	bgp	N/A	10.4.5.2 (indirect)		122007
122007	SWAP	bgp	N/A	10.4.5.2 (indirect)		122010
122008	SWAP	bgp	N/A	10.4.5.2 (indirect)		122014
122009	SWAP	bgp	N/A	10.4.5.2 (indirect)		122009
122010	SWAP	bgp	N/A	10.4.5.2 (indirect)		122012
122011	SWAP	bgp	N/A	10.4.5.2 (indirect)		122013

// 122007 swapped with 122010 with next hop 10.4.5.2

8 Combined ECMP for BGP PE-CE and EVPN IFL/IP-VPN



Note: This feature is supported on 7220 IXR, 7250 IXR and 7730 SXR platforms.

SR Linux supports the combination of EVPN IFL, IP-VPN, and BGP PE-CE routes in a single ECMP set or next-hop group. The combination of routes is also supported in unequal ECMP.

Combined ECMP is supported along with IP Aliasing, unequal ECMP and resilient hashing for EVPN IFL routes.

Combined ECMP is represented in a route table as a single route with a combined next-hop group. When at least one of the next hops is coming from a BGP PE-CE route, the resulting route type is BGP.

[Table 6: All possible route combinations for combined BGP PE-CE weighted ECMP](#) shows all the possible route combinations and the resulting route types. The resulting route follows the best path selection result across the BGP owners of the route.

Table 6: All possible route combinations for combined BGP PE-CE weighted ECMP

Route combination	Resulting route type
<ul style="list-style-type: none"> • bgp • bgp-ipvpn 	BGP
<ul style="list-style-type: none"> • bgp • bgp-evpn 	BGP
<ul style="list-style-type: none"> • bgp-evpn • bgp-ipvpn 	BGP-EVPN
<ul style="list-style-type: none"> • bgp • bgp-evpn • bgp-ipvpn 	BGP

The range of the number of ECMP paths in a combined set is the minimum of the configured ECMP paths of the contributing routes. The individual ECMP path ranges are:

- `bgp-evpn.ecmp [=1..64]`
- `bgp-ipvpn.ecmp [=1..64]`
- `bgp.multipath.max-paths-level-1 [=1..64]`

To show the maximum ECMP number of paths, the following is an example configuration:

- `bgp-evpn.ecmp [=10]`
- `bgp-ipvpn.ecmp [=12]`

- `bgp.multipath.max-paths-level-1[=15]`

In this example, the following are true:

- If at least one `bgp-evpn` is present in the combined route, the maximum number of ECMP paths used is 10.
- Else, if at least one `bgp-ipvpn` route is present in the combined route, the maximum number of ECMP paths used is 12.
- Else, the maximum number of ECMP paths is 15.

The combination of multiple BGP PE-CE routes in a single ECMP set is configured using the **combined-ecmp** command.

```
--{ * candidate shared default }--[ ]--
A:srl1# info network-instance ip-vrf
  network-instance ip-vrf {
    protocols {
      bgp-vpn {
        combined-ecmp {
        }
      }
    }
  }
}
```

Candidate routes to be combined are compared against the BGP best path selection tie-breakers and routes can only be combined if all the best path selection criteria up to the MED comparison are equal. That is, two routes of different owners can be combined for ECMP purposes only if the following attributes match on both routes:

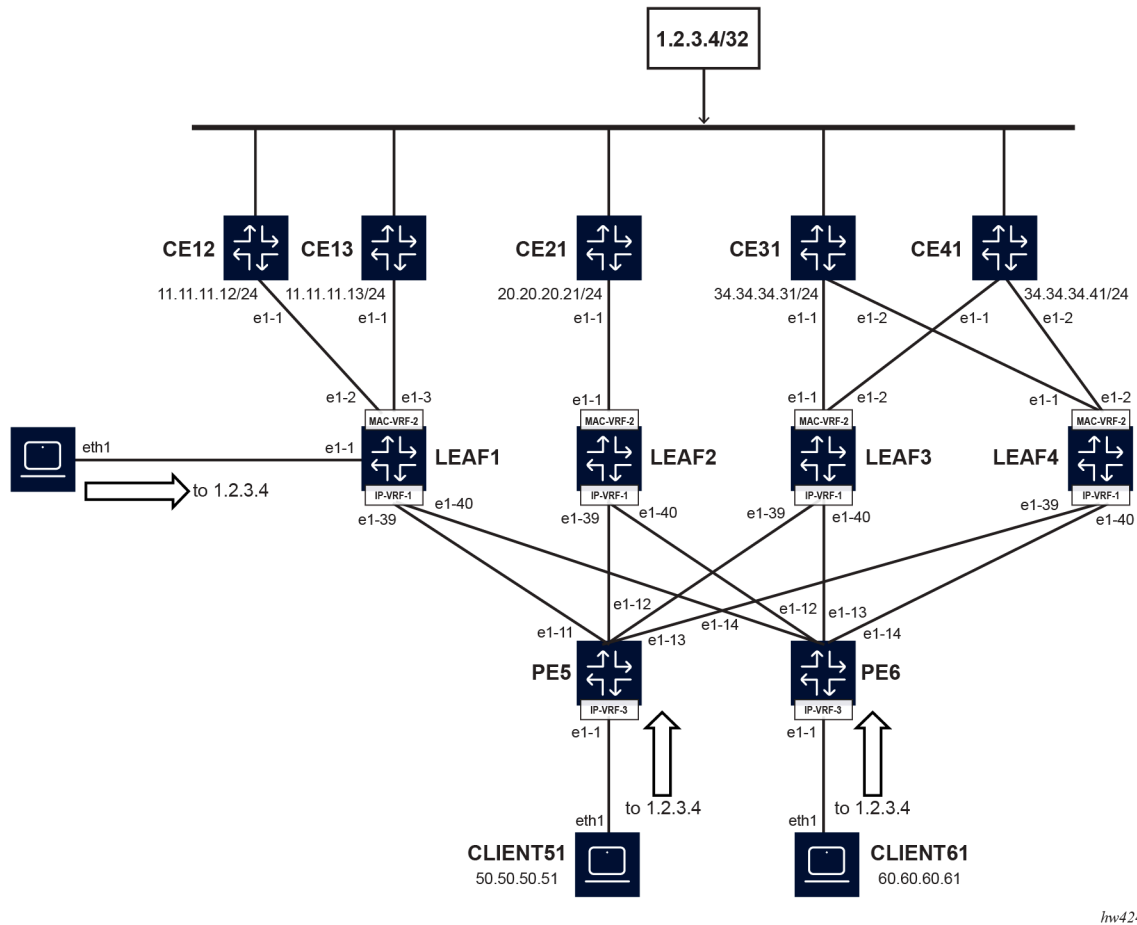
1. Valid routes
2. Origin validation state
3. Route table PREFERENCE
4. BGP local preference
5. BGP AS_PATH
6. BGP origin
7. BGP MED

When multiple routes of the same or different owner for the same prefix are received, the selection is affected by the configuration of **combined-ecmp**, as indicated in [Best path selection for EVPN IFL routes](#).

8.1 Combined ECMP example

The following figure shows an example of combined ECMP:

Figure 29: Combined ECMP



The figure shows that CE12, CE13, CE21, CE31 and CE41 are all connected to the host 1.2.3.4/32 and advertise this in BGP PE-CE to their connected leaf router. Leaf1 receives the 1.2.3.4/32 prefix via BGP and BGP-EVPN. By configuring **combined-ecmp**, leaf1 can install a single BGP route to 1.2.3.4/32 with a combined next-hop group that includes the PE-CE routes and remote next hops from EVPN IFL routes, with or without weights and with or without ESIs.

The following shows the configuration of leaf1:

```
--{ * candidate shared default }--[ network-instance IP-VRF-1 ]--
A:srl1# info
  type ip-vrf
  description "IP-VRF-2 EVPN-MPLS uecmp"
  ip-load-balancing {
    resilient-hash-prefix 1.2.0.0/16 {
      hash-buckets-per-path 6
      max-paths 10
    }
  }
  interface ethernet-1/1.1 {
  }
  interface irb0.1 {
  }
```

```

interface lo0.1 {
}
protocols {
  bgp-evpn {
    bgp-instance 1 {
      encapsulation-type mpls
      evi 1
      ecmp 64
      mpls {
        next-hop-resolution {
          allowed-tunnel-types [
            ldp
            sr-isis
          ]
        }
      }
    }
    routes {
      route-table {
        ip-prefix {
          evpn-link-bandwidth {
            advertise {
              weight dynamic
            }
            weighted-ecmp {
              admin-state enable
              max-ecmp-hash-buckets-per-next-hop-group 64
            }
          }
        }
      }
    }
  }
}
bgp {
  admin-state enable
  autonomous-system 64500
  router-id 1.1.1.1
  ebgp-default-policy {
    import-reject-all false
    export-reject-all false
  }
  afi-safi ipv4-unicast {
    admin-state enable
    multipath {
      maximum-paths 64
    }
  }
  trace-options {
    flag packets {
      modifier detail
    }
    flag update {
      modifier detail
    }
  }
}
group pe-ce {
  multihop {
    admin-state enable
    maximum-hops 10
  }
  afi-safi ipv4-unicast {
    ipv4-unicast {
      link-bandwidth {
        add-next-hop-count-to-received-bgp-routes 1
      }
    }
  }
}

```



```

    }
  }
  timers {
    connect-retry 1
    minimum-advertisement-interval 1
  }
  trace-options {
    flag update {
      modifier detail
    }
  }
  transport {
    local-address 1.1.1.1
  }
}
neighbor 11.11.11.12 {
  peer-as 64512
  peer-group pe-ce
}
neighbor 11.11.11.13 {
  peer-as 64513
  peer-group pe-ce
}
}
bgp-vpn {
  combined-ecmp {
  }
  bgp-instance 1 {
  }
}
}

```

The route table for prefix 1.2.3.4/32 is displayed as follows:

```

--{ + candidate shared default }--[ network-instance IP-VRF-1 ]--
A:leaf1# show route-table ipv4-unicast prefix 1.2.3.4/32
-----
IPv4 unicast route table of network instance IP-VRF-1
-----
---
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
--+
| Prefix      | ID | Route | Route | Active | Origin | Metric | Pref | Next-hop (Type) | Next-hop
| |
| Interface | | | | | | | | |
| |
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
==+
| 1.2.3.4/32 | 0  | bgp   | bgp_mgr | True  | IP-VRF-1 | 0      | 170 | 11.11.11.0/24 ( | irb0.1
| |
| |         | |     | |     | |     | |     | |     | |     | indirect/local) | irb0.1
| |
| |         | |     | |     | |     | |     | |     | |     | 11.11.11.0/24 ( |
| |
| |         | |     | |     | |     | |     | |     | |     | indirect/local) |
| |
| |         | |     | |     | |     | |     | |     | |     | 100.0.0.2/32 (  |
| |

```

									indirect/ldp)
									100.0.0.3/32 (
									indirect/ldp)
									100.0.0.4/32 (
									indirect/ldp)
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+									
--+									

The following shows the route table state for the route:

```
--{ + candidate shared default }--[ network-instance IP-VRF-1 ]--
A:leaf1# info from state route-table ipv4-unicast route 1.2.3.4/32 id 0 route-type bgp route-
owner bgp_mgr origin-network-instance IP-VRF-1
leakable false
  metric 0
  preference 170
  active true
  last-app-update "2024-07-02T09:06:02.838Z (an hour ago)"
  next-hop-group 347400906964
  next-hop-group-network-instance IP-VRF-1
  resilient-hash true
  fib-programming {
    suppressed false
    last-successful-operation-type modify
    last-successful-operation-timestamp "2024-07-02T09:06:02.839Z (an hour ago)"
    pending-operation-type none
    last-failed-operation-type none
  }
--{ + candidate shared default }--[ network-instance IP-VRF-1 ]--
A:leaf1# info from state route-table next-hop-group 347400906964
  backup-next-hop-group 0
  fib-programming {
    last-successful-operation-type modify
    last-successful-operation-timestamp "2024-07-02T09:06:02.839Z (an hour ago)"
    pending-operation-type none
    last-failed-operation-type none
  }
  next-hop 0 {
    next-hop 347400907006
    resolved true
  }
  next-hop 1 {
    next-hop 347400907007
    resolved true
  }
  next-hop 2 {
    next-hop 347400906982
    resolved true
  }
  next-hop 3 {
    next-hop 347400906999
    resolved true
  }
  next-hop 4 {
```

```

    next-hop 347400906996
    resolved true
  }
--{ + candidate shared default }--[ network-instance IP-VRF-1 ]--
A:leaf1# info from state route-table next-hop
{347400907006,347400907007,347400906982,347400906999,347400906996}
  next-hop 347400907006 {
    type indirect
    ip-address 11.11.11.12
    resolving-route {
      ip-prefix 11.11.11.0/24
      route-type local
      route-owner net_inst_mgr
    }
  }
  next-hop 347400907007 {
    type indirect
    ip-address 11.11.11.13
    resolving-route {
      ip-prefix 11.11.11.0/24
      route-type local
      route-owner net_inst_mgr
    }
  }
  next-hop 347400906982 {
    type indirect
    ip-address 100.0.0.2
    resolving-tunnel {
      ip-prefix 100.0.0.2/32
      tunnel-type ldp
      tunnel-owner ldp_mgr
    }
  }
  mpls {
    pushed-mpls-label-stack [
      1000
    ]
  }
}
  next-hop 347400906999 {
    type indirect
    ip-address 100.0.0.3
    resolving-tunnel {
      ip-prefix 100.0.0.3/32
      tunnel-type ldp
      tunnel-owner ldp_mgr
    }
  }
  mpls {
    pushed-mpls-label-stack [
      1002
    ]
  }
}
  next-hop 347400906996 {
    type indirect
    ip-address 100.0.0.4
    resolving-tunnel {
      ip-prefix 100.0.0.4/32
      tunnel-type ldp
      tunnel-owner ldp_mgr
    }
  }
  mpls {
    pushed-mpls-label-stack [
      1001
    ]
  }
}

```

```
}  
}
```

9 EVPN configuration examples

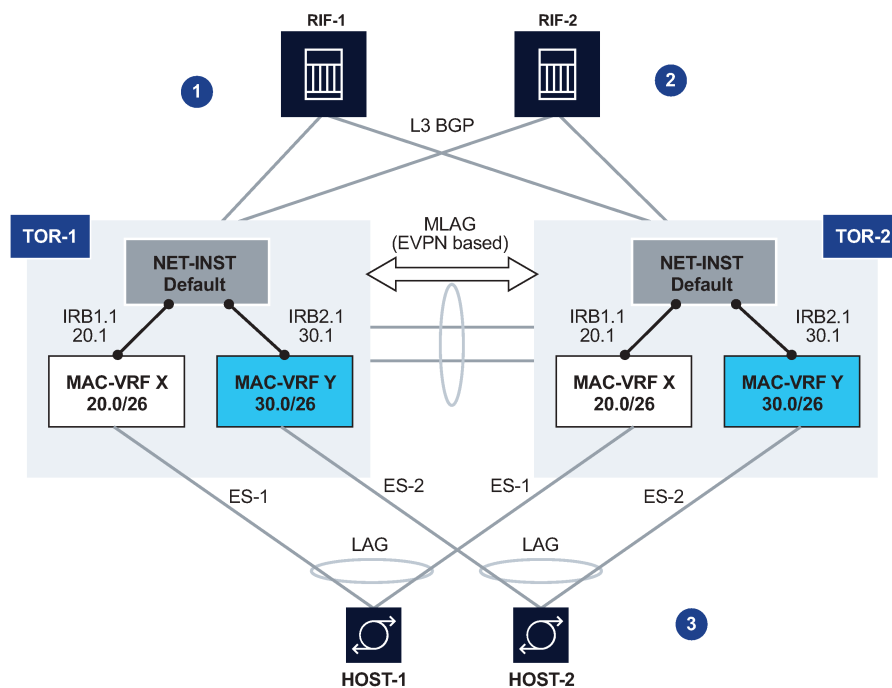
This chapter contains examples of configurations that use EVPN features.

- [All-active redundant connectivity example](#)
- [Hierarchical active-active connectivity example](#)
- [EVPN multihoming as standalone solution for MC-LAG](#)

9.1 All-active redundant connectivity example

The following figure shows an example of using EVPN multihoming as a standalone and self-contained multichassis solution.

Figure 30: Active-active connectivity example



This example uses the following features:

1. redundancy

TOR redundancy is based on an all-active redundancy model.

2. Layer 3 connectivity

- An anycast gateway solution is used on the IRB subinterfaces so that upstream traffic is always routed locally on the TOR receiving the traffic.

- South-to-north traffic is sent to the active link and routed by the local IRB subinterface. In case of failure on TOR-1, TOR-2 is ready to forward on the anycast gateway IRB, without the need to wait for any VRRP protocol to converge.
- North-to-south traffic is load-balanced by the fabric to the two TORs. ARP/ND entries are synchronized across both TORs. Host routes could be optionally created and advertised in BGP from the directly connected TOR to avoid tromboning in the downstream direction.

3. Layer 2 connectivity

- Servers do not need to run any xSTP protocols. The NDF TOR brings down the port and signals LOS to the server.
- This solution places no requirements on the servers.

This example imposes the use of a LAG on the server. The LAG can use LACP or not. The servers are unaware that the LAG is connected to two systems instead of only one.

9.1.1 Configuration for all-active connectivity example

Leaf 1 in [Figure 30: Active-active connectivity example](#) has the following configuration. Leaf 2 would have an equivalent configuration.

```
--{ [FACTORY] +* candidate shared default }--[ ]--
A:Leaf-1# info
// two IRB interfaces with anycast-gw configuration is added
interface irb1 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 20.0.0.1/24 {
        anycast-gw true
        primary
      }
    }
    arp {
      learn-unsolicited true
      evpn {
        advertise dynamic {
        }
      }
    }
  }
  anycast-gw {
  }
}
interface irb2 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 30.0.0.1/24 {
        anycast-gw true
        primary
      }
    }
    arp {
      learn-unsolicited true
      evpn {
        advertise dynamic {
        }
      }
    }
  }
  anycast-gw {
  }
}
```

```

    }
    }
    anycast-gw {
    }
}
// lags associated with the ethernet-segments.
// In this case static, but they can be lacp based too.
interface lag1 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    vlan {
      encap {
        single-tagged {
          vlan-id 20
        }
      }
    }
  }
  lag {
    lag-type static
  }
}
interface lag2 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    vlan {
      encap {
        single-tagged {
          vlan-id 30
        }
      }
    }
  }
  lag {
    lag-type static
  }
}
// ES configuration
system {
  network-instance {
    protocols {
      evpn {
        ethernet-segments {
          bgp-instance 1 {
            ethernet-segment ES-1 {
              admin-state enable
              esi 00:01:00:00:00:00:00:00:01
              interface lag1
            }
            ethernet-segment ES-2 {
              admin-state enable
              esi 00:02:00:00:00:00:00:00:02
              interface lag2
            }
          }
        }
      }
    }
  }
}
// MAC-VRFs

```

```

network-instance MAC-VRF-X {
  type mac-vrf
  admin-state enable
  interface irb1.1 {
  }
  interface lag1.1 {
  }
  vxlan-interface vxlan1.20 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.20
        evi 20
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
network-instance MAC-VRF-Y {
  type mac-vrf
  admin-state enable
  interface irb2.1 {
  }
  interface lag2.1 {
  }
  vxlan-interface vxlan1.30 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable
        vxlan-interface vxlan1.30
        evi 30
      }
    }
    bgp-vpn {
      bgp-instance 1 {
      }
    }
  }
}
// default network-instance configuration
network-instance default {
  type default
  admin-state enable
  description "Default network instance"
  router-id 1.1.1.1
  interface ethernet-1/1.1 {
  }
  interface ethernet-1/20.1 {
  }
  interface irb1.1 {
  }
  interface irb2.1 {
  }
  interface system0.0 {
  }
  protocols {
    bgp {

```



```

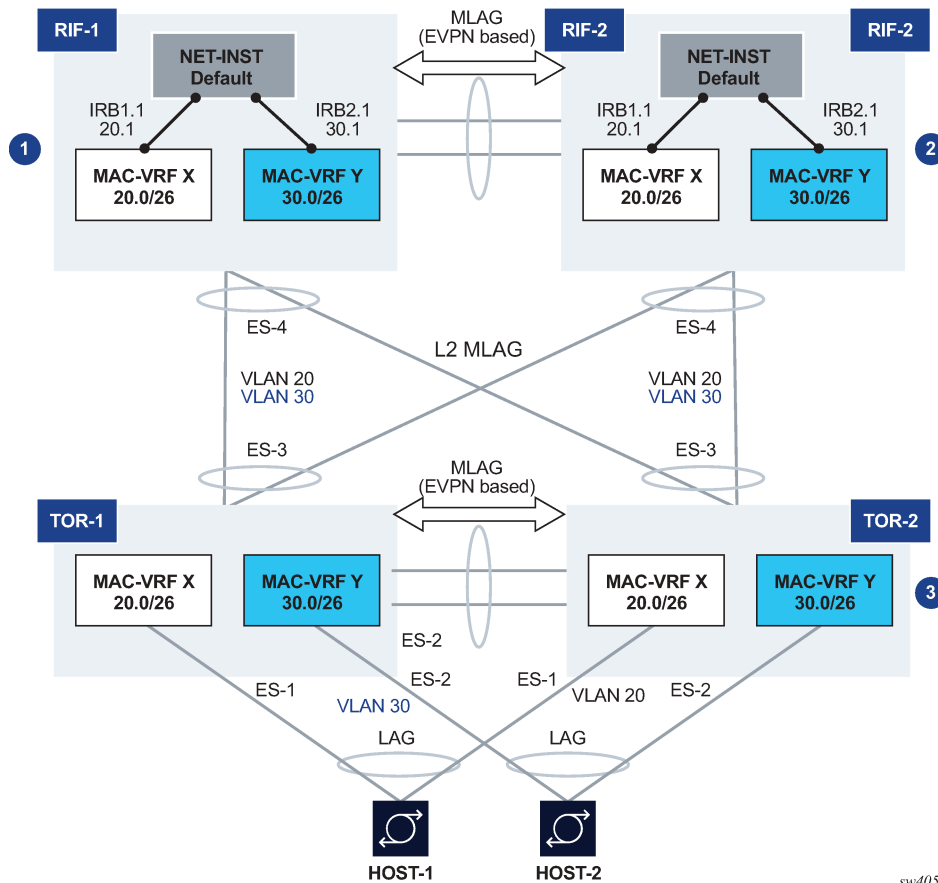
    admin-state enable
    afi-safi ipv4-unicast {
        admin-state enable
    }
    autonomous-system 1234
    router-id 1.1.1.1
    group eBGP-spines {
        admin-state enable
        export-policy export-all
        peer-as 4567
        afi-safi ipv4-unicast {
            admin-state enable
        }
    }
    group evpn-mh {
        admin-state enable
        export-policy export-all
        peer-as 1234
        afi-safi evpn {
            admin-state enable
        }
    }
    neighbor 2.2.2.2 {
        admin-state enable
        peer-group evpn-mh
    }
    neighbor 10.1.4.4 {
        admin-state enable
        peer-group eBGP-spines
    }
}
}
}
// vxlan interfaces for the MH configuration
tunnel-interface vxlan1 {
    vxlan-interface 20 {
        type bridged
        ingress {
            vni 20
        }
    }
    vxlan-interface 30 {
        type bridged
        ingress {
            vni 30
        }
    }
}
}
}

```

9.2 Hierarchical active-active connectivity example

The following figure shows an example that makes use of a Layer 2 and a Layer 3 multihoming solution.

Figure 31: Hierarchical active-active connectivity example



sw4059

This example uses the following features:

- **Leaf and Spine configuration**

- There are two multichassis pairs at the leaf and the spine levels.
- The leaf pair runs all-active multihoming on its own ESs.

The access hosts or spines are separate and independent ESs. The figure shows three ESs, one per host at the access, and one for the connectivity to the spine layer.



Note: This is only one ES in spite of the number of spine nodes, which could be 2 or 4.

- The spine layer runs all-active multihoming with a single ES to the leaf layer, associated with the LAG attached to the leaf layer.
- The two tiers are independent; there are no control plane protocols between them, except for LACP if used.

This means that the leaf and spine layer could theoretically be of a different vendor.

- **Layer 3 connectivity**

- An anycast gateway solution is used on the spine layer, on the IRB subinterfaces, so that upstream traffic is always routed locally on the leaf receiving the traffic.
 - South-to-north flows are sent to only one link at a time, so there is not duplication.
 - North-to-south traffic is load-balanced by the fabric to the two spines and for the spines load-balanced to the leafs. ARP/ND entries are synchronized across both spines. Host routes can optionally be created and advertised in BGP from the directly connected spine to avoid tromboning in the downstream direction.
- **Layer 2 connectivity**
 - Hosts are running LAG with or without LACP.
 - Leafs and spines are connected by standard Layer 2 LAGs that carry the VLANs for the two broadcast domains shown in [Figure 31: Hierarchical active-active connectivity example](#).

9.2.1 Configuration for hierarchical active-active connectivity example

Spine 1 in [Figure 31: Hierarchical active-active connectivity example](#) has the following configuration. The configuration for Spine 2 would be equivalent.

```
--{ [FACTORY] +* candidate shared default }--[ ]--
A:Spine-1# info
// two IRB interfaces with anycast-gw configuration is added
interface irb1 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 20.0.0.1/24 {
        anycast-gw true
        primary
      }
      arp {
        learn-unsolicited true
        evpn {
          advertise dynamic {
            // for ARP synchronization across MH leaf nodes
          }
        }
      }
    }
    anycast-gw {
    }
  }
}
interface irb2 {
  subinterface 1 {
    ipv4 {
      admin-state enable
      address 30.0.0.1/24 {
        anycast-gw true
        primary
      }
      arp {
        learn-unsolicited true
        evpn {
          advertise dynamic {
          }
        }
      }
    }
  }
}
```

```

    }
    anycast-gw {
    }
}
// lags associated with the ethernet-segment.
In this case static, but they can be lacp based too.
interface lag1 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    vlan {
      encap {
        single-tagged {
          vlan-id 20
        }
      }
    }
  }
  subinterface 2 {
    vlan {
      encap {
        single-tagged {
          vlan-id 30
        }
      }
    }
  }
  lag {
    lag-type static
  }
}
// ES configuration
system {
  network-instance {
    protocols {
      evpn {
        ethernet-segments {
          bgp-instance 1 {
            ethernet-segment ES-4 {
              admin-state enable
              esi 00:44:44:44:44:44:00:00:04
              interface lag1
            }
          }
        }
      }
    }
  }
}
// MAC-VRFs
network-instance MAC-VRF-X {
  type mac-vrf
  admin-state enable
  interface irb1.1 {
  }
  interface lag1.1 {
  }
  vxlan-interface vxlan1.20 {
  }
  protocols {
    bgp-evpn {
      bgp-instance 1 {
        admin-state enable

```

```

        vxlan-interface vxlan1.20
        evi 20
    }
}
    bgp-vpn {
        bgp-instance 1 {
        }
    }
}
network-instance MAC-VRF-Y {
    type mac-vrf
    admin-state enable
    interface irb2.1 {
    }
    interface lag1.2 {
    }
    vxlan-interface vxlan1.30 {
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.30
                evi 30
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
// default network-instance configuration
network-instance default {
    type default
    admin-state enable
    description "Default network instance"
    router-id 1.1.1.1
    interface ethernet-1/1.1 {
    }
    interface ethernet-1/20.1 {
    }
    interface irb1.1 {
    }
    interface irb2.1 {
    }
    interface system0.0 {
    }
    protocols {
        bgp {
            admin-state enable
            afi-safi ipv4-unicast {
                admin-state enable
            }
            autonomous-system 1234
            router-id 1.1.1.1
            group eBGP-spines {
                admin-state enable
                export-policy export-all
                peer-as 4567
            }
            afi-safi ipv4-unicast {
                admin-state enable
            }
        }
    }
}

```

```

    }
    group evpn-mh {
        admin-state enable
        export-policy export-all
        peer-as 1234
        afi-safi evpn {
            admin-state enable
        }
    }
    neighbor 2.2.2.2 {
        admin-state enable
        peer-group evpn-mh
    }
    neighbor 10.1.4.4 {
        admin-state enable
        peer-group eBGP-spines
    }
}
}
}
// vxlan interfaces for the MH configuration
tunnel-interface vxlan1 {
    vxlan-interface 20 {
        type bridged
        ingress {
            vni 20
        }
    }
    vxlan-interface 30 {
        type bridged
        ingress {
            vni 30
        }
    }
}
}
}

```

Leaf 1 in [Figure 31: Hierarchical active-active connectivity example](#) has the following configuration. The configuration for Leaf 2 would be equivalent.

```

--{ [FACTORY] +* candidate shared default }--[ ]--
A:Leaf-1# info
// lags associated with the ethernet-segments.
In this case static, but they can be lacp based too.
interface lag1 {
    admin-state enable
    vlan-tagging true
    subinterface 1 {
        vlan {
            encap {
                single-tagged {
                    vlan-id 20
                }
            }
        }
    }
}
lag {
    lag-type static
}
}
interface lag2 {
    admin-state enable
    vlan-tagging true
}

```

```

subinterface 1 {
  vlan {
    encap {
      single-tagged {
        vlan-id 30
      }
    }
  }
}
lag {
  lag-type static
}
}
interface lag3 {
  admin-state enable
  vlan-tagging true
  subinterface 1 {
    vlan {
      encap {
        single-tagged {
          vlan-id 20
        }
      }
    }
  }
  subinterface 2 {
    vlan {
      encap {
        single-tagged {
          vlan-id 30
        }
      }
    }
  }
  lag {
    lag-type static
  }
}
// ES configuration
system {
  network-instance {
    protocols {
      evpn {
        ethernet-segments {
          bgp-instance 1 {
            ethernet-segment ES-1 {
              admin-state enable
              esi 00:11:11:11:11:11:00:00:01
              interface lag1
            }
            ethernet-segment ES-2 {
              admin-state enable
              esi 00:22:22:22:22:22:00:00:02
              interface lag2
            }
            ethernet-segment ES-3 {
              admin-state enable
              esi 00:33:33:33:33:33:00:00:03
              interface lag3
            }
          }
        }
      }
    }
  }
}

```

```

    }
}
// MAC-VRFs
network-instance MAC-VRF-X {
    type mac-vrf
    admin-state enable
    interface lag1.1 {
    }
    interface lag3.1 {
    }
    vxlan-interface vxlan1.20 {
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.20
                evi 20
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
network-instance MAC-VRF-Y {
    type mac-vrf
    admin-state enable
    interface lag2.1 {
    }
    interface lag3.1 {
    }
    vxlan-interface vxlan1.30 {
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.30
                evi 30
            }
        }
        bgp-vpn {
            bgp-instance 1 {
            }
        }
    }
}
// default network-instance configuration
network-instance default {
    type default
    admin-state enable
    description "Default network instance"
    router-id 1.1.1.1
    interface ethernet-1/1.1 {
    }
    interface system0.0 {
    }
    protocols {
        bgp {
            admin-state enable
            afi-safi ipv4-unicast {
                admin-state enable
            }
        }
    }
}

```

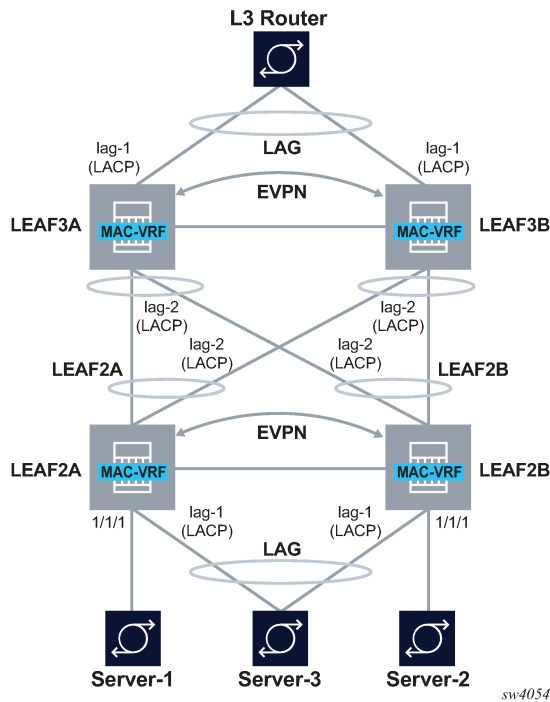


```
    }
    autonomous-system 1234
    router-id 1.1.1.1
    group evpn-mh {
        admin-state enable
        export-policy export-all
        peer-as 1234
        afi-safi evpn {
            admin-state enable
        }
    }
    neighbor 2.2.2.2 {
        admin-state enable
        peer-group evpn-mh
    }
}
}
// vxlan interfaces for the MH configuration
tunnel-interface vxlan1 {
    vxlan-interface 20 {
        type bridged
        ingress {
            vni 20
        }
    }
    vxlan-interface 30 {
        type bridged
        ingress {
            vni 30
        }
    }
}
}
```

9.3 EVPN multihoming as standalone solution for MC-LAG

EVPN multihoming is not only used in overlay DCs, but also as a standalone solution for multihoming in Layer 2 access networks with no VXLAN. The following figure shows this usage.

Figure 32: EVPN multihoming as standalone MC-LAG solution

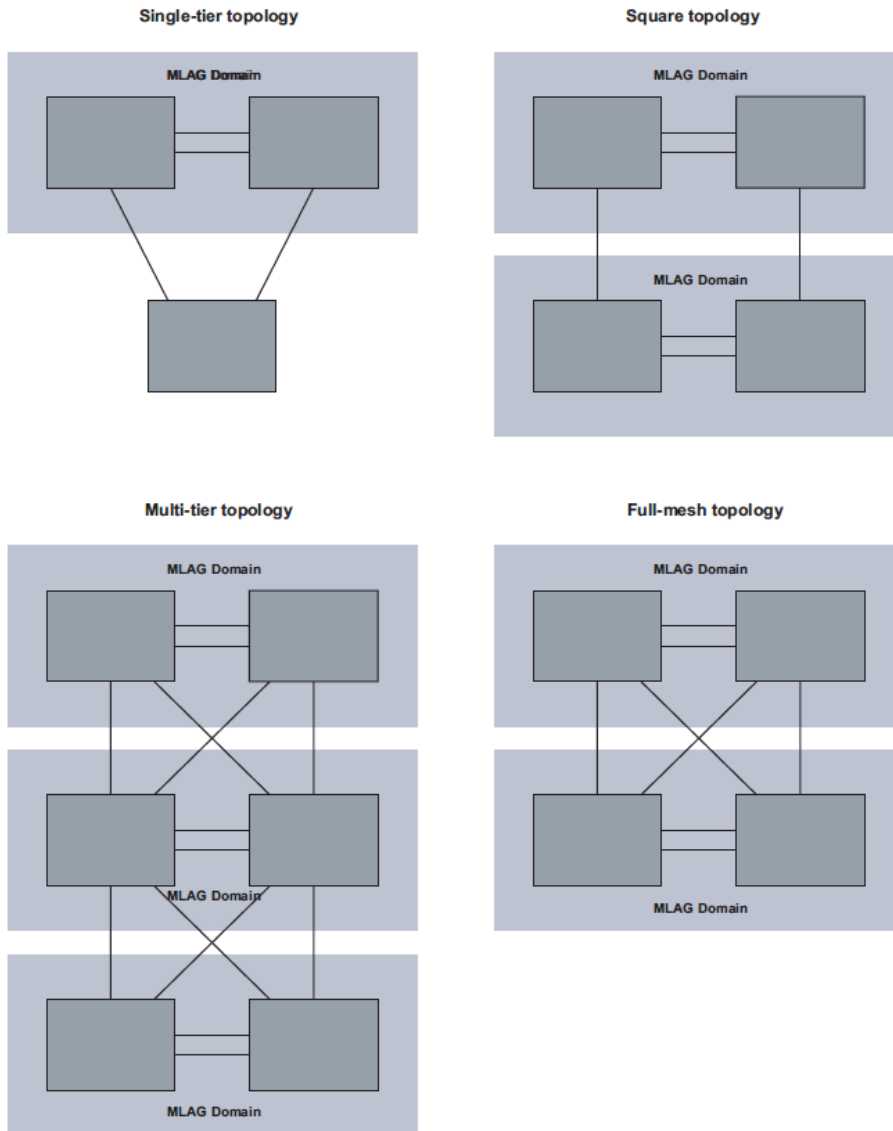


On the left side of the figure, EVPN multihoming is used as a standalone multihoming solution for leaf nodes connected via bridged subinterfaces.

Leaves of a layer do not use VXLAN to get connected to the higher layer. In this case, EVPN sessions are configured locally within each multihoming pair so that EVPN handles DF Election, split-horizon and synchronization of MAC and ARP entries. However, the leaves of different layers are not connected through any IP fabric, so no VXLAN or EVPN is needed end-to-end.

In this configuration, EVPN provides an alternative to MC-LAG solutions, being able to match all the topologies that other MC-LAG solutions support. These topologies include single-tier, multi-tier, square, or full-mesh/bow-tie topologies, as shown in the following figure. EVPN multihoming is supported in all of them as a replacement of MC-LAG.

Figure 33: MLAG topologies



9.3.1 Configuration for EVPN multihoming as standalone MC-LAG

LEAF2A in [Figure 32: EVPN multihoming as standalone MC-LAG solution](#) has the following configuration:

```
// lag1 connects LEAF2A to server-3
// and is associated to an all-active Ethernet Segment.
--{ candidate shared default }--[ ]--
A:LEAF2A# info interface lag*
  interface lag1 {
    admin-state enable
    vlan-tagging true
    subinterface 10 {
      type bridged
    }
  }
```

```

        vlan {
            encap {
                single-tagged {
                    vlan-id 10
                }
            }
        }
    }
    lag {
        lag-type static
        // lag-type could also be lACP, in which case the
        // system-id/key must match on lag1 of LEAF2B
        member-speed 10G
    }
}
// lag2 connects LEAF2A to LEAF3A and LEAF3B.
// This LAG is an access LAG (does not carry vxlan) associated to
// an all-active Ethernet Segment
interface lag2 {
    admin-state enable
    vlan-tagging true
    subinterface 10 {
        type bridged
        vlan {
            encap {
                single-tagged {
                    vlan-id 10
                }
            }
        }
    }
}
lag {
    lag-type static
    // lag-type could also be lACP, in which case the
    // system-id/key must match on lag2 of LEAF2B
    member-speed 10G
}
}
// A vxlan-interface is created for the inter-chassis traffic
--{ candidate shared default }--[ ]-
A:LEAF2A# info tunnel-interface vxlan1 vxlan-interface 10
    tunnel-interface vxlan1 {
        vxlan-interface 10 {
            type bridged
            ingress {
                vni 10
            }
        }
    }
}
// the Ethernet Segments associated to lag1 and lag2
--{ candidate shared default }--[ ]-
A:LEAF2A# info system network-instance protocols evpn
    system {
        network-instance {
            protocols {
                evpn {
                    ethernet-segments {
                        bgp-instance 1 {
                            ethernet-segment ES-leaf1-leaf2.CE1 {
                                admin-state enable
                                esi 00:12:12:12:12:12:00:00:01
                                interface lag1
                            }
                            ethernet-segment ES-leaf1-leaf2.Spines {

```

```

                                admin-state enable
                                esi 00:12:12:12:12:12:00:00:02
                                interface lag2
                            }
                        }
                    }
                }
            }
        }
    }
}
// the mac-vrf uses lag sub-interfaces for the connectivity to rest of
// the leaf nodes and servers, and a vxlan-subinterface for the connectivity
// to LEAF2B
--{ candidate shared default }--[ ]--
A:LEAF2A# info network-instance Blue-MAC-VRF-10
network-instance Blue-MAC-VRF-10 {
    type mac-vrf
    interface ethernet-1/1.10 {
        // this is connected to a single-homed access server-1
    }
    interface lag1.10 {
        // access lag - multi-homed to access server-3
    }
    interface lag2.10 {
        // multi-homed to spines
    }
    vxlan-interface vxlan1.10 {
        // vxlan-interface used for inter-chassis connectivity only
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.10
                evi 10
            }
        }
        bgp-vpn {
        }
    }
}
}

```

LEAF2A in [Figure 32: EVPN multihoming as standalone MC-LAG solution](#) has the following configuration:

```

// lag1 connects LEAF2B to server-3 and is associated to
// an all-active Ethernet Segment
--{ candidate shared default }--[ ]--
A:LEAF2B# info interface lag*
interface lag1 {
    admin-state enable
    vlan-tagging true
    subinterface 10 {
        type bridged
        vlan {
            encap {
                single-tagged {
                    vlan-id 10
                }
            }
        }
    }
}
lag {
    lag-type static
}

```

```
// lag-type could also be lacp, in which case the
system-id/key must match on lag1 of LEAF2A
    member-speed 10G
    }
}
// lag2 connects LEAF2B to LEAF3A and LEAF3B.
This LAG is an access LAG (does not carry vxlan) associated to
an all-active Ethernet Segment
interface lag2 {
    admin-state enable
    vlan-tagging true
    subinterface 10 {
        type bridged
        vlan {
            encap {
                single-tagged {
                    vlan-id 10
                }
            }
        }
    }
}
lag {
    lag-type static
// lag-type could also be lacp, in which case the
system-id/key must match on lag2 of LEAF2B
    member-speed 10G
}
}
// A vxlan-interface is created for the inter-chassis traffic
--{ candidate shared default }--[ ]--
A:LEAF2B# info tunnel-interface vxlan1 vxlan-interface 10
    tunnel-interface vxlan1 {
        vxlan-interface 10 {
            type bridged
            ingress {
                vni 10
            }
        }
    }
}
// the Ethernet Segments associated to lag1 and lag2
--{ candidate shared default }--[ ]--
A:LEAF2B# info system network-instance protocols evpn
system {
    network-instance {
        protocols {
            evpn {
                ethernet-segments {
                    bgp-instance 1 {
                        ethernet-segment ES-leaf1-leaf2.CE1 {
                            admin-state enable
                            esi 00:12:12:12:12:12:00:00:01
                            interface lag1
                        }
                        ethernet-segment ES-leaf1-leaf2.Spines {
                            admin-state enable
                            esi 00:12:12:12:12:12:00:00:02
                            interface lag2
                        }
                    }
                }
            }
        }
    }
}
}
```

```
// the mac-vrf uses lag sub-interfaces for the connectivity to rest of the
leaf nodes and servers, and a vxlan-subinterface for the connectivity
to LEAF2B
--{ candidate shared default }--[ ]--
A:LEAF2B# info network-instance Blue-MAC-VRF-10
network-instance Blue-MAC-VRF-10 {
    type mac-vrf
    interface ethernet-1/2.10 {
        !!! this is connected to a single-homed access server-2
    }
    interface lag1.10 {
        !!! access lag - multi-homed to access server-3
    }
    interface lag2.10 {
        !!! multi-homed to spines
    }
    vxlan-interface vxlan1.10 {
        !!! vxlan-interface used for inter-chassis connectivity only
    }
    protocols {
        bgp-evpn {
            bgp-instance 1 {
                admin-state enable
                vxlan-interface vxlan1.10
                evi 10
            }
        }
        bgp-vpn {
        }
    }
}
```

Customer document and product support



Customer documentation

[Customer documentation welcome page](#)



Technical support

[Product support portal](#)



Documentation feedback

[Customer documentation feedback](#)