MPLS and RSVP

In This Chapter

This chapter provides information to configure MPLS and RSVP.

MPLS

Multiprotocol Label Switching (MPLS) is a label switching technology that provides the ability to set up connection-oriented paths over a connectionless IP network. MPLS facilitates network traffic flow and provides a mechanism to engineer network traffic patterns independently from routing tables. MPLS sets up a specific path for a sequence of packets. The packets are identified by a label inserted into each packet. MPLS is not enabled by default and must be explicitly enabled.

MPLS is independent of any routing protocol but is considered multiprotocol because it works with the Internet Protocol (IP), Asynchronous Transport Mode (ATM), and frame relay network protocols.

MPLS Label Stack

MPLS requires a set of procedures to enhance network layer packets with label stacks which thereby turns them into labeled packets. Routers that support MPLS are known as Label Switching Routers (LSRs). In order to transmit a labeled packet on a particular data link, an LSR must support the encoding technique which, when given a label stack and a network layer packet, produces a labeled packet.

In MPLS, packets can carry not just one label, but a set of labels in a stack. An LSR can swap the label at the top of the stack, pop the stack, or swap the label and push one or more labels into the stack. The processing of a labeled packet is completely independent of the level of hierarchy. The processing is always based on the top label, without regard for the possibility that some number of other labels may have been above it in the past, or that some number of other labels may be below it at present.

As described in RFC 3032, *MPLS Label Stack Encoding*, the label stack is represented as a sequence of label stack entries. Each label stack entry is represented by 4 octets. Figure 1 displays the label placement in a packet.

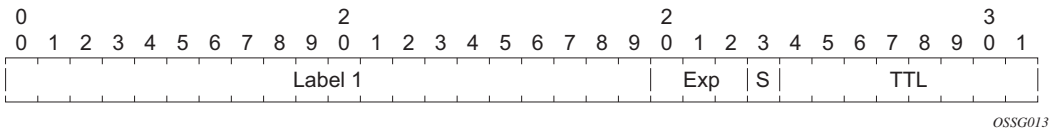


Figure 1: Label Placement

Table 2: Packet/Label Field Description

Field	Description
Label	This 20-bit field carries the actual value (unstructured) of the label.
Exp	This 3-bit field is reserved for experimental use. It is currently used for Class of Service (CoS).
S	This bit is set to 1 for the last entry (bottom) in the label stack, and 0 for all other label stack entries.
TTL	This 8-bit field is used to encode a TTL value.

A stack can carry several labels, organized in a last in/first out order. The top of the label stack appears first in the packet and the bottom of the stack appears last (Figure 2).

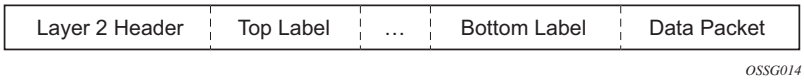


Figure 2: Label Packet Placement

The label value at the top of the stack is looked up when a labeled packet is received. A successful lookup reveals:

- The next hop where the packet is to be forwarded.
- The operation to be performed on the label stack before forwarding.

In addition, the lookup may reveal outgoing data link encapsulation and other information needed to properly forward the packet.

An empty label stack can be thought of as an unlabeled packet. An empty label stack has zero (0) depth. The label at the bottom of the stack is referred to as the Level 1 label. The label above it (if it exists) is the Level 2 label, and so on. The label at the top of the stack is referred to as the Level *m* label.

Labeled packet processing is independent of the level of hierarchy. Processing is always based on the top label in the stack which includes information about the operations to perform on the packet's label stack.

Label Values

Packets travelling along an LSP (see Label Switching Routers on page 25) are identified by its label, the 20-bit, unsigned integer. The range is 0 through 1,048,575. Label values 0-15 are reserved and are defined below as follows:

- A value of 0 represents the IPv4 Explicit NULL Label. This Label value is legal only at the bottom of the Label stack. It indicates that the Label stack must be popped, and the packet forwarding must be based on the IPv4 header.
- A value of 1 represents the router alert Label. This Label value is legal anywhere in the Label stack except at the bottom. When a received packet contains this Label value at the top of the Label stack, it is delivered to a local software module for processing. The actual packet forwarding is determined by the Label beneath it in the stack. However, if the packet is further forwarded, the router alert Label should be pushed back onto the Label stack before forwarding. The use of this Label is analogous to the use of the router alert option in IP packets. Since this Label cannot occur at the bottom of the stack, it is not associated with a particular network layer protocol.
- A value of 2 represents the IPv6 explicit NULL Label. This Label value is only legal at the bottom of the Label stack. It indicates that the Label stack must be popped, and the packet forwarding must be based on the IPv6 header.
- A value of 3 represents the Implicit NULL Label. This is a Label that a Label Switching Router (LSR) can assign and distribute, but which never actually appears in the encapsulation. When an LSR would otherwise replace the Label at the top of the stack with a new Label, but the new Label is Implicit NULL, the LSR pops the stack instead of doing the replacement. Although this value may never appear in the encapsulation, it needs to be specified in the Label Distribution Protocol (LDP), so a value is reserved.
- Values 4-15 are reserved for future use.

The router uses labels for MPLS, RSVP-TE, and LDP, as well as packet-based services such as VLL and VPLS.

Label values 16 through 1,048,575 are defined as follows:

- Label values 16 through 31 are reserved for future use.
- Label values 32 through 1,023 are available for static LSP label assignments.
- Label values 1,024 through 2,047 are reserved for future use.
- Label values 2,048 through 18,431 are available for static service label assignments
- Label values 18,432 through 262,14(131,071 in chassis modes lower than D)3 are assigned dynamically by RSVP, LDP, and BGP control planes for both MPLS LSP and service labels.
- Label values 262,14(131,072 in chassis modes lower than D)4 through 1,048,575 are reserved for future use.

Label Switching Routers

LSRs perform the label switching function. LSRs perform different functions based on it's position in an LSP. Routers in an LSP do one of the following:

- The router at the beginning of an LSP is the ingress label edge router (ILER). The ingress router can encapsulate packets with an MPLS header and forward it to the next router along the path. An LSP can only have one ingress router.
- A Label Switching Router (LSR) can be any intermediate router in the LSP between the ingress and egress routers. An LSR swaps the incoming label with the outgoing MPLS label and forwards the MPLS packets it receives to the next router in the MPLS path (LSP). An LSP can have 0-253 transit routers.
- The router at the end of an LSP is the egress label edge router (ELER). The egress router strips the MPLS encapsulation which changes it from an MPLS packet to a data packet, and then forwards the packet to its final destination using information in the forwarding table. Each LSP can have only one egress router. The ingress and egress routers in an LSP cannot be the same router.

A router in your network can act as an ingress, egress, or transit router for one or more LSPs, depending on your network design.

An LSP is confined to one IGP area for LSPs using constrained-path. They cannot cross an autonomous system (AS) boundary.

Static LSPs can cross AS boundaries. The intermediate hops are manually configured so the LSP has no dependence on the IGP topology or a local forwarding table.

LSP Types

The following are LSP types:

- Static LSPs — A static LSP specifies a static path. All routers that the LSP traverses must be configured manually with labels. No signaling such as RSVP or LDP is required.
- Signaled LSP — LSPs are set up using a signaling protocol such as RSVP-TE or LDP. The signaling protocol allows labels to be assigned from an ingress router to the egress router. Signaling is triggered by the ingress routers. Configuration is required only on the ingress router and is not required on intermediate routers. Signaling also facilitates path selection.

  There are two signaled LSP types:

  - Explicit-path LSPs — MPLS uses RSVP-TE to set up explicit path LSPs. The hops within the LSP are configured manually. The intermediate hops must be configured as either strict or loose meaning that the LSP must take either a direct path from the

previous hop router to this router (strict) or can traverse through other routers (loose). You can control how the path is set up. They are similar to static LSPs but require less configuration. See RSVP on page 71.

- Constrained-path LSPs — The intermediate hops of the LSP are dynamically assigned. A constrained path LSP relies on the Constrained Shortest Path First (CSPF) routing algorithm to find a path which satisfies the constraints for the LSP. In turn, CSPF relies on the topology database provided by the extended IGP such as OSPF or IS-IS.

  Once the path is found by CSPF, RSVP uses the path to request the LSP set up. CSPF calculates the shortest path based on the constraints provided such as bandwidth, class of service, and specified hops.

If fast reroute is configured, the ingress router signals the routers downstream. Each downstream router sets up a detour for the LSP. If a downstream router does not support fast reroute, the request is ignored and the router continues to support the LSP. This can cause some of the detours to fail, but otherwise the LSP is not impacted.

No bandwidth is reserved for the rerouted path. If the user enters a value in the bandwidth parameter in the **config>router>mpls>lsp>fast-reroute** context, it will have no effect on the LSP backup LSP establishment.

Hop-limit parameters specifies the maximum number of hops that an LSP can traverse, including the ingress and egress routers. An LSP is not set up if the hop limit is exceeded. The hop count is set to 255 by default for the primary and secondary paths. It is set to 16 by default for a bypass or detour LSP path.

MPLS Facility Bypass Method of MPLS Fast Re-Route (FRR)

The MPLS facility bypass method of MPLS Fast Re-Route (FRR) functionality is extended to the ingress node.

The behavior of an LSP at an ingress LER with both fast reroute and a standby LSP path configured is as follows:

- When a down stream detour becomes active at a point of local repair (PLR):

  The ingress LER switches to the standby LSP path. If the primary LSP path is repaired subsequently at the PLR, the LSP will switch back to the primary path. If the standby goes down, the LSP is switched back to the primary, even though it is still on the detour at the PLR. If the primary goes down at the ingress while the LSP is on the standby, the detour at the ingress is cleaned up and for one-to-one detours a "path tear" is sent for the detour path. In other words, the detour at the ingress does not protect the standby. If and when the primary LSP is again successfully re-signaled, the ingress detour state machine will be restarted.
- When the primary fails at the ingress:

  The LSP switches to the detour path. If a standby is available then LSP would switch to standby on expiration of **hold-timer**. If **hold-timer** is disabled then switchover to standby would happen immediately. On successful global revert of primary path, the LSP would switch back to the primary path.
- Admin groups are not taken into account when creating detours for LSPs.

Manual Bypass LSP

In prior releases, the router implemented dynamic bypass tunnels as per RFC 4090, *Fast Reroute Extensions to RSVP-TE for LSP Tunnels*. When an LSP is signaled and the local protection flag is set in the session_attribute object and/or the FRR object in the path message indicates that facility backup is desired, the PLR will establish a bypass tunnel to provide node and link protection. If a bypass LSP which merges in a downstream node with the protected LSP exist, and if this LSP satisfies the constraints in the FRR object, then this bypass tunnel is selected.

With the manual bypass feature, an LSP can be pre-configured from a PLR which will be used exclusively for bypass protection. When a path message for a new LSP requests bypass protection, the node will first check if a manual bypass tunnel satisfying the path constraints exists. If one is found, it will be selected. If no manual bypass tunnel is found, the router will dynamically signal a bypass LSP in the default behavior. Users can disable the dynamic bypass creation on a per node basis using the CLI.

A maximum of 1000 associations of primary LSP paths can be made with a single manual bypass by default. The **max-bypass-associations** *integer* command increases the number of associations.

If dynamic bypass creation is disabled on the node, it is recommended to configure additional manual bypass LSPs to handle the required number of associations.

Refer to Configuring Manual Bypass Tunnels on page 169 for configuration information.

PLR Bypass LSP Selection Rules

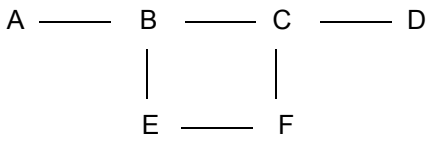


Figure 3: Bypass Tunnel Nodes

The PLR uses the following rules to select a bypass LSP among multiple manual and dynamic bypass LSPs at the time of establishment of the primary LSP path or when searching for a bypass for a protected LSP which does not have an association with a bypass tunnel:

1. The MPLS/RSVP task in the PLR node checks if an existing manual bypass satisfies the constraints. If the path message for the primary LSP path indicated node protection desired, which is the default LSP FRR setting at the head end node, MPLS/RSVP task searches for a node-protect' bypass LSP. If the path message for the primary LSP path indicated link protection desired, then it searches for a link-protect bypass LSP.
2. If multiple manual bypass LSPs satisfying the path constraints exist, it will prefer a manual-bypass terminating closer to the PLR over a manual bypass terminating further away. If multiple manual bypass LSPs satisfying the path constraints terminate on the same downstream node, it selects one with the lowest IGP path cost or if in a tie, picks the first one available.
3. If none satisfies the constraints and dynamic bypass tunnels have not been disabled on PLR node, then the MPLS/RSVP task in the PLR will check if any of the already established dynamic bypasses of the requested type satisfies the constraints.
4. If none do, then the MPLS/RSVP task will ask CSPF to check if a new dynamic bypass of the requested type, node-protect or link-protect, can be established.
5. If the path message for the primary LSP path indicated node protection desired, and no manual bypass was found after Step 1, and/or no dynamic bypass LSP was found after 3 attempts of performing Step 3, the MPLS/RSVP task will repeat Steps 1-3 looking for a suitable link-protect bypass LSP. If none are found, the primary LSP will have no protection and the PLR node must clear the "local protection available" flag in the IPv4 address sub-object of the RRO starting in the next Resv refresh message it sends upstream.

6. If the path message for the primary LSP path indicated link protection desired, and no manual bypass was found after step 1, and/or no dynamic bypass LSP was found after performing Step 3, the primary LSP will have no protection and the PLR node must clear the “local protection available” flag in the IPv4 address sub-object of the RRO starting in the next RESV refresh message it sends upstream. The PLR will not search for a node-protect’ bypass LSP in this case.
7. If the PLR node successfully makes an association, it must set the “local protection available” flag in the IPv4 address sub-object of the RRO starting in the next RESV refresh message it sends upstream.
8. For all primary LSP that requested FRR protection but are not currently associated with a bypass tunnel, the PLR node on reception of RESV refresh on the primary LSP path repeats Steps 1-7.

If the user disables dynamic-bypass tunnels on a node while dynamic bypass tunnels were activated and were passing traffic, traffic loss will occur on the protected LSP. Furthermore, if no manual bypass exist that satisfy the constraints of the protected LSP, the LSP will remain without protection.

If the user configures a bypass tunnel on node B and dynamic bypass tunnels have been disabled, LSPs which have been previously signaled and which were not associated with any manual bypass tunnel, for example, none existed, will be associated with the manual bypass tunnel if suitable. The node checks for the availability of a suitable bypass tunnel for each of the outstanding LSPs every time a RESV message is received for these LSPs.

If the user configures a bypass tunnel on node B and dynamic bypass tunnels have not been disabled, LSPs which have been previously signaled over dynamic bypass tunnels will not automatically be switched into the manual bypass tunnel even if the manual bypass is a more optimized path. The user will have to perform a make before break at the head end of these LSPs.

If the manual bypass goes into the down state in node B and dynamic bypass tunnels have been disabled, node B (PLR) will clear the “protection available” flag in the RRO IPv4 sub-object in the next RESV refresh message for each affected LSP. It will then try to associate each of these LSPs with one of the manual bypass tunnels that are still up. If it finds one, it will make the association and set again the “protection available” flag in the next RESV refresh message for each of these LSPs. If it could not find one, it will keep checking for one every time a RESV message is received for each of the remaining LSPs. When the manual bypass tunnel is back UP, the LSPs which did not find a match will be associated back to this tunnel and the protection available flag is set starting in the next RESV refresh message.

If the manual bypass goes into the down state in node B and dynamic bypass tunnels have not been disabled, node B will automatically signal a dynamic bypass to protect the LSPs if a suitable one does not exist. Similarly, if an LSP is signaled while the manual bypass is in the down state, the node will only signal a dynamic bypass tunnel if the user has not disabled dynamic tunnels. When the manual bypass tunnel is back into the UP state, the node will not switch the protected LSPs from the dynamic bypass tunnel into the manual bypass tunnel.

FRR Node-Protection (Facility)

The MPLS Fast Re-Route (FRR) functionality enables PLRs to be aware of the missing node protection and lets them regularly probe for a node-bypass. The following describes an LSP scenario:

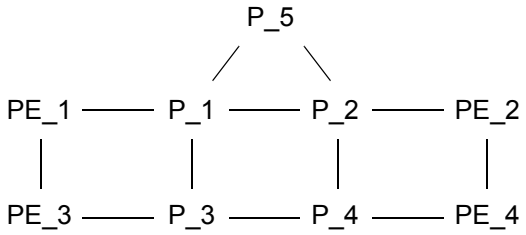


Figure 4: FRR Node-Protection Example

Where:

- LSP 1: between PE_1 to PE_2, with CSPF, FRR facility node-protect enabled.
- P_1 protects P_2 with bypass-nodes P_1 -P_3 - P_4 - PE_4 -PE_3.
- If P_4 fails, P_1 tries to establish the bypass-node three times.
- When the bypass-node creation fails, P_1 will protect link P_1-P_2.
- P_1 protects the link to P_2 through P_1 - P_5 - P_2.
- P_4 returns online.

Since LSP 1 had requested node protection, but due to lack of any available path, it could only obtain link protection. Therefore, every 60 seconds the PLR for LSP 1 will search for a new path that might be able to provide node protection. Once P_4 is back online and such a path is available, A new bypass tunnel will be signalled and LSP 1 will get associated with this new bypass tunnel.

Uniform FRR Failover Time

The failover time during FRR consists of a detection time and a switchover time. The detection time corresponds to the time it takes for the RSVP control plane protocol to detect that a network IP interface is down or that a neighbor/next-hop over a network IP interface is down. The control plane can be informed of an interface down event when event is due to a failure in a lower layer such in the physical layer. The control plane can also detect the failure of a neighbor/next-hop on its own by running a protocol such as Hello, Keep-Alive, or BFD.

The switchover time is measured from the time the control plane detected the failure of the interface or neighbor/next-hop to the time the IOM completed the reprogramming of all the impacted ILM or service records in the data path. This includes the time it takes for the control plane to send a down notification to all IOMs to request a switch to the backup NHLFE.

Uniform Fast-Reroute (FRR) failover enables the switchover of MPLS and service packets from the outgoing interface of the primary LSP path to that of the FRR backup LSP within the same amount of time regardless of the number of LSPs or service records. This is achieved by updating Ingress Label Map (ILM) records and service records to point to the backup Next-Hop Label to Forwarding Entry (NHLFE) in a single operation.

Automatic Bandwidth Allocation for RSVP LSPs

Enabling and Disabling Auto-Bandwidth Allocation on an LSP

This section discusses an auto-bandwidth hierarchy configurable in the **config>router>mpls>lsp** context.

Adding auto-bandwidth at the LSP level starts the measurement of LSP bandwidth described in Measurement of LSP Bandwidth on page 33 and allows auto-bandwidth adjustments to take place based on the triggers described in Periodic Automatic Bandwidth Adjustment on page 35.

When an LSP is first established, the bandwidth reserved along its primary path is controlled by the bandwidth parameter in the **config>router>mpls>lsp>primary** context, whether or not the LSP has auto-bandwidth enabled. When auto-bandwidth is enabled and a trigger occurs, the system will attempt to change the bandwidth of the LSP to a value between **min-bandwidth** and **max-bandwidth**, which are configurable values in the **lsp>auto-bandwidth** context. **min-bandwidth** is the minimum bandwidth that auto-bandwidth can signal for the LSP and **max-bandwidth** is the maximum bandwidth that can be signaled. The user can set the **min-bandwidth** to the same value as the primary path bandwidth but the system will not enforce this restriction. The system will allow:

- No **min-bandwidth** to be configured. In this case, the implicit minimum is 0 Mbps
- No **max-bandwidth** to be configured, as long as overflow-triggered auto-bandwidth is not configured. In this case, the implicit maximum is infinite (effectively 100 Gbps).
- The configured primary path bandwidth to be outside the range of min-bandwidth to max-bandwidth.
- **auto-bandwidth** parameters can be changed at any time on an operational LSP; in most cases the changes have no immediate impact but subsequent sections will describe some exceptions

All of the auto-bandwidth adjustments discussed are performed using MBB procedures.

Auto bandwidth can be added to an operational LSP at any time (without the need to shut down the LSP or path), but no bandwidth change occurs until a future trigger event. Auto bandwidth may also be removed from an operational LSP at any time and this causes an immediate MBB bandwidth change to be attempted using the configured primary path bandwidth.

Note that changing the configured bandwidth of an auto-bandwidth LSP has no immediate affect, it will only matters if the LSP/path goes down (due to failure or administrative action) and comes back up or if auto-bandwidth is removed from the LSP. The operator can force an auto-bandwidth LSP to be resized immediately to an arbitrary bandwidth using the appropriate tools commands.

Measurement of LSP Bandwidth

Automatic adjustment of RSVP LSP bandwidth based on measured traffic rate into the tunnel requires the LSP to be configured for egress statistics collection at the ingress LER. The following CLI shows an example:

```
config router mpls lsp name
      egress-statistics
      accounting-policy 99
      collect-stats
      no shutdown
     exit
```

All LSPs configured for accounting, including any configured for auto-bandwidth based on traffic measurements, must reference the same accounting policy. An example configuration of such an accounting-policy is shown below: in the CLI example below.

```
config log
     accounting-policy 99
     collection-interval 5
          record combined-mpls-lsp-egress
     exit
exit
```

Note that the record **combined-mpls-lsp-egress** command in the accounting policy has the effect of recording both egress packet and byte counts and bandwidth measurements based on the byte counts if auto-bandwidth is enabled on the LSP.

When egress statistics are enabled the CPM collects stats from of all IOMs involved in forwarding traffic belonging to the LSP (whether the traffic is currently leaving the ingress LER via the primary LSP path, a secondary LSP path, an FRR detour path or an FRR bypass path). The egress statistics have counts for the number of packets and bytes forwarded per LSP on a per-forwarding class, per-priority (in-profile vs. out-of-profile) basis. When auto-bandwidth is configured for an LSP the ingress LER calculates a traffic rate for the LSP as follows:

Average data rate of LSP[x] during interval[i] = F(x, i)—F(x, i-1)/sample interval

F(x, i) — The total number of bytes belonging to LSP[x], regardless of forwarding-class or priority, at time[i]

sample interval = time[i] — time [i-1], time[i+1] — time[i], etc.

The sample interval is the product of sample-multiplier and the collection-interval specified in the auto-bandwidth accounting policy. A default sample-multiplier for all LSPs may be configured using the **config>router>mpls>auto-bandwidth-defaults** command but this value can be overridden on a per-LSP basis at the **config>router>mpls>lsp>auto-bandwidth** context. The

default value of sample-multiplier (the value that would result from the no auto-bandwidth-defaults command) is 1, which means the default sample interval is 300 seconds.

Over a longer period of time called the adjust interval the router keeps track of the maximum average data rate recorded during any constituent sample interval. The adjust interval is the product of adjust-multiplier and the collection-interval specified in the auto-bandwidth accounting-policy. A default adjust-multiplier for all LSPs may be configured using the **config>router>mpls>auto-bandwidth-multiplier** command but this value can be overridden on a per-LSP basis at the **config>router>mpls>lsp>auto-bandwidth** context. The default value of adjust-multiplier (the value that would result from the no auto-bandwidth-mulitplier command) is 288, which means the default adjust interval is 86400 seconds or 24 hours. The system enforces the restriction that adjust-multiplier is equal to or greater than sample-multiplier. It is recommended that the **adjust-multiplier** be an integer multiple of the **sample-multiplier**.

The collection-interval in the auto-bandwidth accounting policy can be changed at any time, without disabling any of the LSPs that rely on that policy for statistics collection.

The sample-multiplier (at the **mpls>auto-bandwidth** level or the **lsp>auto-bandwidth** level) can be changed at any time. This will have no effect until the beginning of the next sample interval. In this case the adjust-interval does not change and information about the current adjust interval (such as the remaining adjust-multiplier, the maximum average data rate) is not lost when the sample-multiplier change takes effect.

The system allows adjust-multiplier (at the **mpls** level or the **lsp>auto-bandwidth** level) to be changed at any time as well but in this case the new value shall have no effect until the beginning of the next adjust interval.

Byte counts collected for LSP statistics include layer 2 encapsulation (Ethernet headers and trailers) and therefore average data rates measured by this feature include Layer 2 overhead as well.

Passive Monitoring of LSP Bandwidth

The system offers the option to measure the bandwidth of an RSVP LSP (see Measurement of LSP Bandwidth on page 33) without taking any action to adjust the bandwidth reservation, regardless of how different the measured bandwidth is from the current reservation. Passive monitoring is enabled using the **config>router>mpls>lsp>auto-bandwidth>monitor-bandwidth** command.

The **show>router>mpls>lsp detail** command can be used to view the maximum average data rate in the current adjust interval and the remaining time in the current adjust interval.

Periodic Automatic Bandwidth Adjustment

Automatic bandwidth allocation is supported on any RSVP LSP that has MBB enabled. MBB is enabled in the **config>router>mpls>lsp** context using the **adaptive** command. For automatic adjustments of LSP bandwidth to occur the monitor-bandwidth command must not be present at **config>router>mpls>lsp>auto-bandwidth** context, otherwise only passive measurements will occur.

If an eligible RSVP LSP is configured for auto-bandwidth, by entering auto-bandwidth at the config>router>mpls>lsp context, then the ingress LER decides every adjust interval whether to attempt auto-bandwidth adjustment. The following parameters are defined:

- current_bw — The currently reserved bandwidth of the LSP; this is the operational bandwidth that is already maintained in the MIB.
- measured_bw — The maximum average data rate in the current adjust interval.
- signaled_bw — The bandwidth that is provided to the CSPF algorithm and signaled in the SENDER_TSPEC and FLOWSPEC objects when an auto-bandwidth adjustment is attempted.
- min — The configured min-bandwidth of the LSP.
- max — The configured max-bandwidth of the LSP.
- up% — The minimum difference between measured_bw and current_bw, expressed as a percentage of current_bw, for increasing the bandwidth of the LSP.
- up — The minimum difference between measured_bw and current_bw, expressed as an absolute bandwidth relative to current_bw, for increasing the bandwidth of the LSP. This is an optional parameter; if not defined the value is 0.
- down% — The minimum difference between current_bw and measured_bw, expressed as a percentage of current_bw, for decreasing the bandwidth of the LSP.
- down — The minimum difference between current_bw and measured_bw, expressed as an absolute bandwidth relative to current_bw, for decreasing the bandwidth of the LSP. This is an optional parameter; if not defined the value is 0.

At the end of every adjust interval the system decides if an auto-bandwidth adjustment should be attempted. The heuristics are as follows:

- If the measured bandwidth exceeds the current bandwidth by more than the percentage threshold and also by more than the absolute threshold then the bandwidth is re-signaled to the measured bandwidth (subject to min and max constraints).
- If the measured bandwidth is less than the current bandwidth by more than the percentage threshold and also by more than the absolute threshold then the bandwidth is re-signaled to the measured bandwidth (subject to min and max constraints).

- If the current bandwidth is greater than the max bandwidth then the LSP bandwidth is re-signaled to max bandwidth, even if the thresholds have not been triggered.
- If the current bandwidth is greater than the min bandwidth then the LSP bandwidth is re-signaled to min bandwidth, even if the thresholds have not been triggered.

Changes to min-bandwidth, max-bandwidth and any of the threshold values (up, up%, down, down%) are permitted at any time on an operational LSP but the changes have no effect until the next auto-bandwidth trigger (for example, adjust interval expiry).

If the measured bandwidth exceeds the current bandwidth by more than the percentage threshold and also by more than the absolute threshold then the bandwidth is re-signaled to the measured bandwidth (subject to min and max constraints).

The adjust-interval and maximum average data rate are reset whether the adjustment succeeds or fails. If the bandwidth adjustment fails (for example, CSPF cannot find a path) then the existing LSP is maintained with its existing bandwidth reservation. The system does not retry the bandwidth adjustment (for example, per the configuration of the LSP retry-timer and retry-limit).

Overflow-Triggered Auto-Bandwidth Adjustment

For cases where the measured bandwidth of an LSP has increased significantly since the start of the current adjust interval it may be desirable for the system to preemptively adjust the bandwidth of the LSP and not wait until the end of the adjust interval.

The following parameters are defined:

- current_bw — The currently reserved bandwidth of the LSP.
- sampled_bw — The average data rate of the sample interval that just ended.
- measured_bw — The maximum average data rate in the current adjust interval.
- signaled_bw — The bandwidth that is provided to the CSPF algorithm and signaled in the SENDER_TSPEC and FLOWSPEC objects when an auto-bandwidth adjustment is attempted.
- max — The configured max-bandwidth of the LSP.
- %_threshold — The minimum difference between sampled_bw and current_bw, expressed as a percentage of the current_bw, for counting an overflow event.
- min_threshold — The minimum difference between sampled_bw and current_bw, expressed as an absolute bandwidth relative to current_bw, for counting an overflow event. This is an optional parameter; if not defined the value is 0.

When a sample interval ends it is counted as an overflow if:

- The sampled bandwidth exceeds the current bandwidth by more than the percentage threshold and by more than the absolute bandwidth threshold (if defined).
- When the number of overflow samples reaches a configured limit, an immediate attempt is made to adjust the bandwidth to the measured bandwidth (subject to the min and max constraints).

If the bandwidth adjustment is successful then the adjust-interval, maximum average data rate and overflow count are all reset. If the bandwidth adjustment fails then the overflow count is reset but the adjust-interval and maximum average data rate continue with current values. It is possible that the overflow count will once again reach the configured limit before the end of adjust-interval is reached and this will once again trigger an immediate auto-bandwidth adjustment attempt.

The overflow configuration command fails if the max-bandwidth of the LSP has not been defined.

The threshold limit can be changed on an operational auto-bandwidth LSP at any time and the change should take effect at the end of the current sample interval (for example, if the user decreases the overflow limit to a value lower than the current overflow count then auto-bandwidth adjustment will take place as soon as the sample interval ends). The threshold values can also be changed at any time (for example, %_threshold and min_threshold) but the new values will not take effect until the end of the current sample interval.

Manually-Triggered Auto-Bandwidth Adjustment

Manually-triggered auto-bandwidth adjustment feature is configured with the **tools>perform>router>mpls adjust-autobandwidth** [**lsp** *lsp-name* [**force** [**bandwidth** *mbps*]]] command to attempt immediate auto-bandwidth adjustment for either one specific LSP or all active LSPs. If the LSP is not specified then the system assumes the command applies to all LSPs. If an LSP name is provided then the command applies to that specific LSP only and the optional **force** parameter (with or without a bandwidth) can be used.

If **force** is not specified (or the command is not LSP-specific) then measured_bw is compared to current_bw and bandwidth adjustment may or may not occur

If **force** is specified and a bandwidth is not provided then the threshold checking is bypassed but the min and max bandwidth constraints are still enforced.

If **force** is specified with a bandwidth (in Mbps) then signaled_bw is set to this bandwidth. There is no requirement that the bandwidth entered as part of the command fall within the range of min-bandwidth to max-bandwidth.

The adjust-interval, maximum average data rate and overflow count are not reset by the manual auto-bandwidth command, whether or not the bandwidth adjustment succeeds or fails. The overflow count is reset only if the manual auto-bandwidth adjustment is successful.

MPLS Transport Profile (MPLS-TP)

MPLS can be used to provide a network layer to support packet transport services. In some operational environments, it is desirable that the operation and maintenance of such an MPLS based packet transport network follow operational models typical in traditional optical transport networks (e.g. SONET/SDH), while providing additional OAM, survivability and other maintenance functions targeted at that environment.

MPLS-TP defines a profile of MPLS targeted at transport applications. This profile defines the specific MPLS characteristics and extensions required to meet transport requirements, while retaining compliance to the standard IETF MPLS architecture and label switching paradigm. The basic requirements are architecture for MPLS-TP are described by the IETF in RFC 5654, RFC 5921 and RFC 5960, in order to meet two objectives:

1. To enable MPLS to be deployed in a transport network and operated in a similar manner to existing transport technologies.

2. To enable MPLS to support packet transport services with a similar degree of predictability to that found in existing transport networks.

In order to meet these objectives, MPLS-TP has a number of high level characteristics:

- It does not modify the MPLS forwarding architecture, which is based on existing pseudowire and LSP constructs. Point-to-point LSPs may be unidirectional or bi-directional. Bi-directional LSPs must be congruent (i.e. co-routed and folow the same pathe in each direction). The 7750 supports bidirectional co-routed MPLS-TP LSPs.
- There is no LSP merging.
- OAM, protection and forwarding of data packets can operate without IP forwarding support. When static provisioning is used, there is no dependency on dynamic routing or signaling.
- LSP and pseudowire monitoring is only achieved through the use of OAM and does not rely on control plane or routing functions to determine the health of a path. e.g. LDP hello failures, do not trigger protection.
- MPLS-TP can operate in the absence of an IP control plane and IP forwarding of OAM traffic. In release 11.0, MPLS-TP is only supported on static LSPs and PWs.

The 7750 SR supports MPLS-TP on LSPs and PWs with static labels. MPLS-TP is not supoprted on dynamically signalled LSPs and PWs. MPLS-TP is supported for EPIPE, APIPE and CPIPE VLLs, and EPIPE Spoke SDP termination on IES, VPRN and VPLS. Static PWs may use SDPs that use either static MPLS-TP LSPs or RSVP-TE LSPs.

The following MPLS-TP OAM and protection meachanisms, defined by the IETF, are supported:

- MPLS-TP Generic Associated Channel for LSPs and PWs (RFC 5586)

- MPLS-TP Identifiers (RFC 6370)
- Proactive CC, CV,and RDI using BFD for LSPs (RFC 6428)
- On-Demand CV for LSPs and PWs using LSP Ping and LSP Trace (RFC 6426)
- 1-for-1 Linear protection for LSPs (RFC 6378)
- Static PW Status Signaling (RFC 6478)

The 7750 SR can play the role of an LER and an LSR for static MPLS-TP LSPs, and a PE/T-PE and an S-PE for static MPLS-TP PWs. It can also act as a S-PE for MPLS-TP segments between an MPLS network that strictly follows the transport profile, and an MPLS network that supports both MPLS-TP and dynamic IP/MPLS.

MPLS-TP Model

Figure 5 shows a high level functional model for MPLS-TP in SROS. LSP A and LSP B are the working and protect LSPs of an LSP tunnel. These are modelled as working and protect paths of an MPLS-TP LSP in SROS. MPLS-TP OAM runs in-band on each path. 1:1 linear protection coordinates the working and protect paths, using a protection switching coordination protocol (PSC) that runs in-band on each path over a Generic Associated Channel (G-ACh) on each path. Each path can use either an IP numbered, IP unnumbered, or MPLS-TP unnumbered (i.e. non-IP) interface.

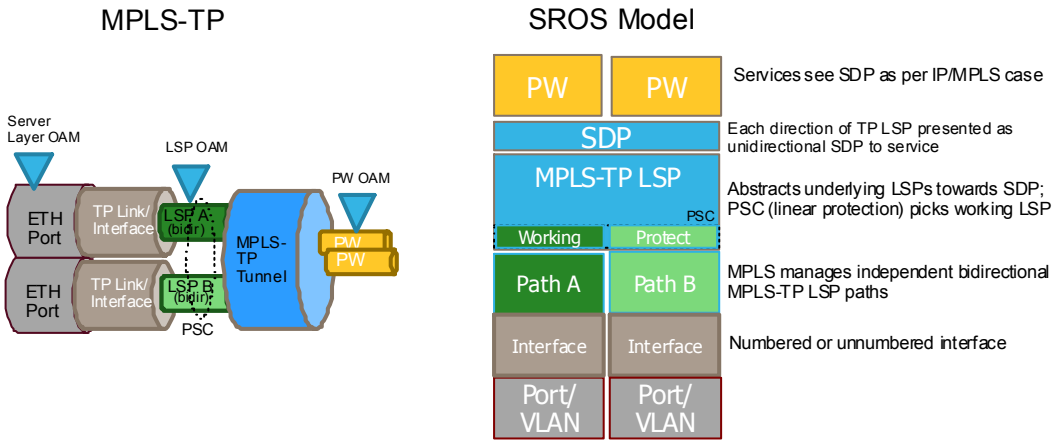


Figure 5: MPLS-TP Model

Note that in SR OS, all MPLS-TP LSPs are bidirectional co-routed, as detailed in RFC5654. That is, the forward and backward directions follow the same route (in terms of links and nodes) across the network. Both directions are setup, monitored and protected as a single entity. Therefore, both

ingress and egress directions of the same LSP segment are associated at the LER and LSR and use the same interface (although this is not enforced by the system).

In the above model, an SDP can use one MPLS-TP LSP. This abstracts the underlying paths towards the overlying services, which are transported on pseudowires. Pseudowires are modelled as spoke SDPs and can also use MPLS-TP OAM. PWs with static labels may use SDPs that in-turn use either signaled RSVP-TE LSPs, or one static MPLS-TP LSP.

MPLS-TP Provider Edge and Gateway

This section describes some example roles for the 7750 SR in an MPLS-TP network.

VLL Services

The 7750 SR may use MPLS TP LSPs, and PWs, to transport point to point virtual leased line services. The 7750 may play the role of a terminating PE or switching PE for VLLs. Epipe, Apipe and Cpipe VLLs are suported.

Figure 6 illustrates the use of the 7750 SR as a T-PE for services in an MPLS-TP domain, and as a S-PE for services between

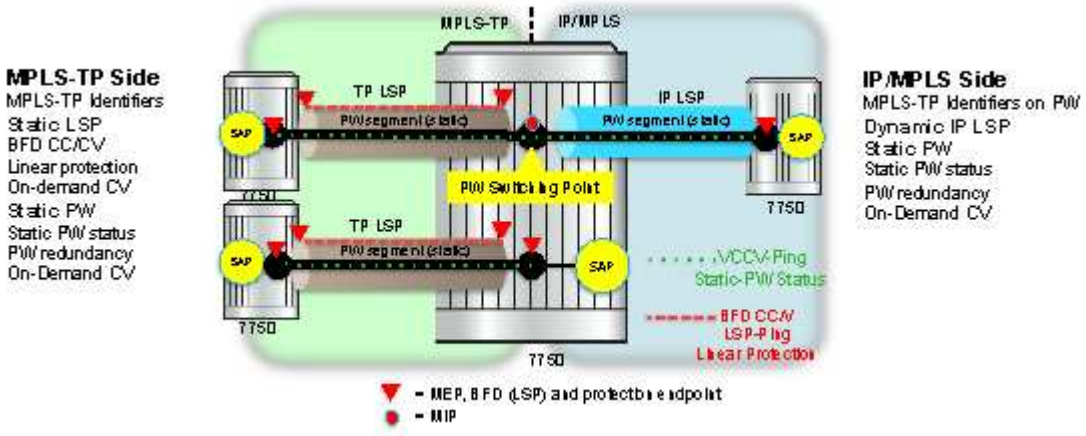


Figure 6: MPLS-TP Provider Edge and Gateway, VLL Services

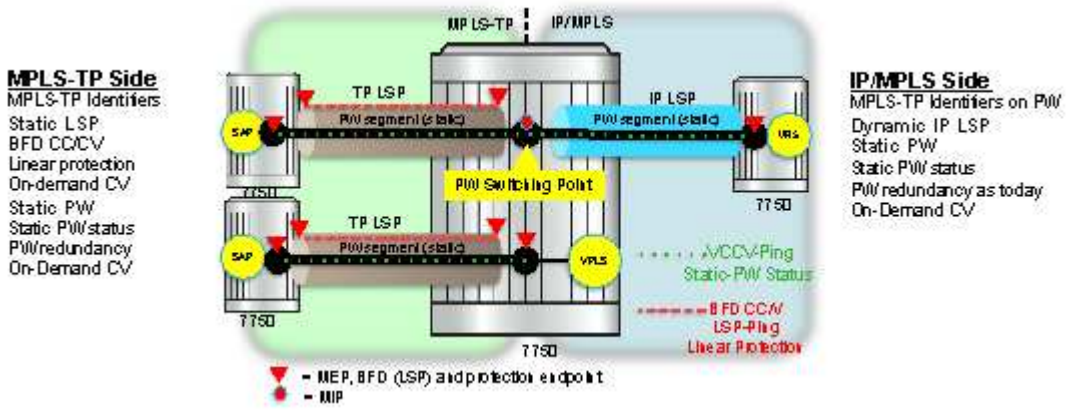


Figure 7: MPLS-TP Provider Edge and Gateway, spoke-SDP Termination on VPLS

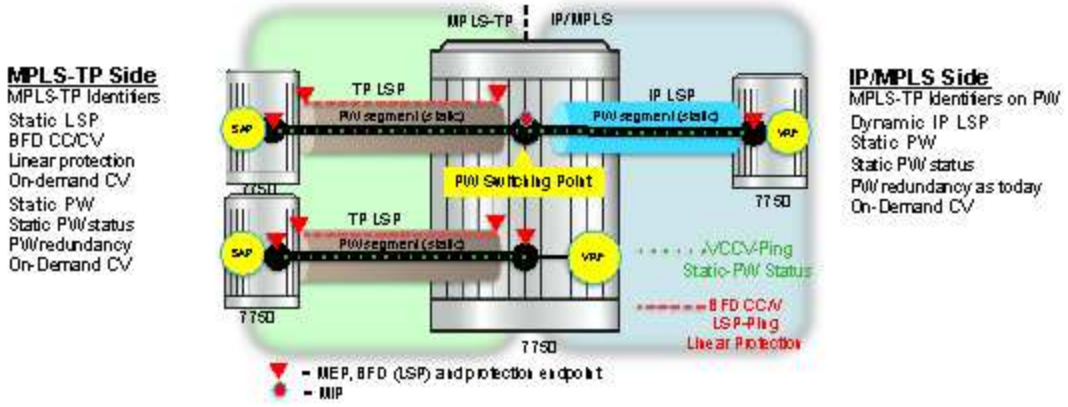


Figure 8: MPLS-TP Provider Edge and Gateway, spoke-SDP Termination on IES/VPRN

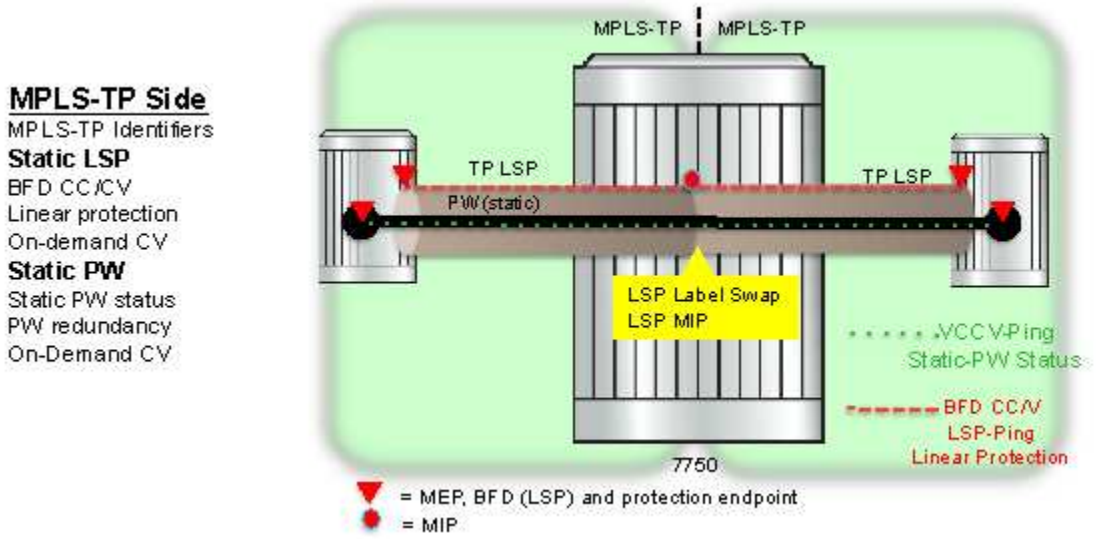


Figure 9: MPLS-TP LSR

Detailed Descriptions of MPLS-TP

MPLS-TP LSPs

SR OS supports the configuration of MPLS-TP tunnels, which comprise a working and, optionally, a protect LSP. In SROS, a tunnel is referred to as an LSP, while an MPLS-TP LSP is referred to as a path. It is then possible to bind an MPLS-TP tunnel to an SDP.

MPLS-TP LSPs (i.e. paths) with static labels are supported. MPLS-TP is not supported for signaled LSPs.

Both bidirectional associated (where the forward and reverse directions of a bidirectional LSP are associated at a given LER, but may take different routes through the intervening network) and bidirectional co-routed (where the forward and reverse directions of the LSP are associated at each LSR, and take the same route through the network) are possible in MPLS-TP. However, only bidirectional co-routed LSPs are supported.

It is possible to configure MPLS-TP identifiers associated with the LSP, and MPLS-TP OAM parameters on each LSP of a tunnel. MPLS-TP protection is configured for a tunnel at the level of the protect path level. Both protection and OAM configuration is managed via templates, in order to simplify provisioning for large numbers of tunnels.

The 7750 may play the role of either an LER or an LSR.

MPLS-TP on Pseudowires

MPLS-TP is supported on PWs with static labels. The provisioning model supports RFC6370-style PW path identifiers for MPLS-TP PWs.

MPLS-TP PWs reuse the static PW provisioning model of previous SR OS releases. Including the use of the pw-switching keywork to distinguish an S-PE. Therefore, the primary distinguishing feature for an "MPLS-TP" PW is the ability to configure MPLS-TP PW path identifiers, and to support MPLS-TP OAM and static PW status signaling.

The 7750 SR can perform the role of a T-PE or an S-PE for a PW with MPLS-TP.

A spoke-SDP with static PW labels and MPLS-TP identifiers and OAM capabilities can use an SDP that uses either an MPLS-TP tunnel, or that uses regular RSVP-TE LSPs. The control word is supported for all MPLS-TP PWs.

MPLS-TP Maintenance Identifiers

MPLS-TP is designed for use both with, and without, a control plane. MPLS-TP therefore specifies a set of identifiers that can be used for objects in either environment. This includes a path and maintenance identifier architecture comprising Node, Interface, PW and LSP identifiers, Maintenance Entity Groups (MEGs), Maintenance End Points (MEPs) and Maintenance Intermediate Points (MIPs). These identifiers are specified in RFC6370.

MPLS-TP OAM and protection switching operates within a framework that is designed to be similar to existing transport network maintenance architectures. MPLS-TP introduces concept of maintenance domains to be managed and monitored. In these, Maintenance Entity Group End Points (MEPs) are edges of a maintenance domain. OAM of a maintenance level must not leak beyond corresponding MEP and so MEPs typically reside at the end points of LSPs and PWs. Maintenance Intermediate Points (MIPS) define intermediate nodes to be monitored. Maintenance Entity Groups (MEGs) comprise all the MEPs and MIPs on an LSP or PW.

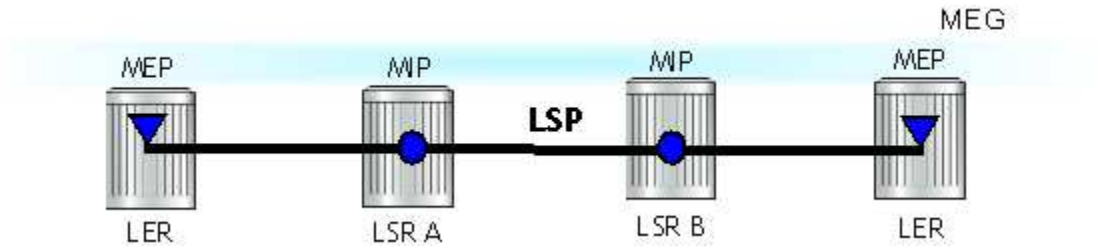


Figure 10: MPLS-TP Maintenance Architecture

Both IP-compatible and ICC (ITU-T carrier code) based identifiers for the above objects are specified in the IETF, but only the IP-compatible identifiers defined in RFC6370 are supported.

SROS suports the configuration of the following node and interface related identifiers:

- Global_ID: this is similar to the global ID that can be configured for Dynamic MS-PWs in Release 9.0 of SR OS. However, in MPLS-TP this should be set to the AS# of the node. If not explicitly configured, then it assumes the default value of 0. In SR OS, the source Global ID for an MPLS-TP Tunnel is taken to be the Global ID configured at the LER. The destination Global ID is optional in the tunnel configuration. If it is not configured, then it is taken as the same as the source Global ID.
- Node_ID: This is a 32-bit value assigned by the operator within the scope of the Global_ID. The 7750 SR supports the configuration of an IPv4 formatted address <a.b.c.d> or an unsigned 32-bit integer for the MPLS-TP Node ID at each node. The node ID must be unique within the scope of the global ID, but there is no requirement for it to be a valid routable IP address. Indeed, a node-id can represent a separate IP-compatible addressing space that may be separate from the IP addressing plan of the underlying network. If no node ID is configured, then the node ID is taken to be the system interface IPv4 address of the node. When configuring a tunnel at an LER, either an IPv4 or an unsigned integer Node ID can be configured as the source and destination identifiers, but both ends must be of the same type.

Statically configured LSPs are identified using GMPLS-compatible identifiers with the addition of a Tunnel_Num and LSP_Num. As in RSVP-TE, tunnels represent, for example, a set of working and protect LSPs. These are GMPLS-compatible because GMPLS chosen by the IETF as the control plane for MPLS-TP LSPs, although this is not supported in Release 11.0 of the 7750. PWs are identified using a PW Path ID which has the same structure as FEC129 AII Type 2.

SR OS derives the identifiers for MEPs and MIPs on LSPs and PWs based on the configured identifiers for the MPLS-TP Tunnel, LSP or PW Path ID, for use in MPLS-TP OAM and protection switching, as per RFC6370.

The information models for LSPs and PWs supported in Release 11.0 are illustrated in Figure 11 and Figure 12. The figures use the terminology defined in RFC6370.

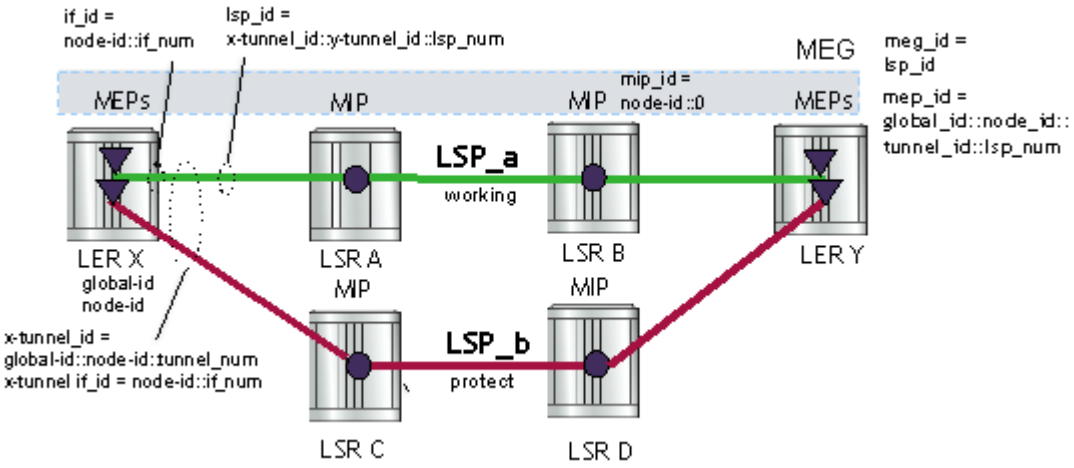


Figure 11: MPLS-TP LSP and Tunnel Information Model

The MPLS-TP Tunnel ID and LSP ID are not to be confused with the RSVP-TE tunnel id implemented on the 7x50 system. Table 3 shows how these map to the X and Y ends of the tunnel shown in the above figure for the case of co-routed bidirectional LSPs.

Table 3: Mapping from RSVP-TE to MPLS-TP Maintenance Identifiers

RSVP-TE Identifier	MPLS-TP Maintenance Identifier
Tunnel Endpoint Address	Node ID (Y)
Tunnel ID (X)	Tunnel Num (X)
Extended Tunnel ID	Node ID (X)
Tunnel Sender Address	Node ID (X)
LSP ID	LSP Num

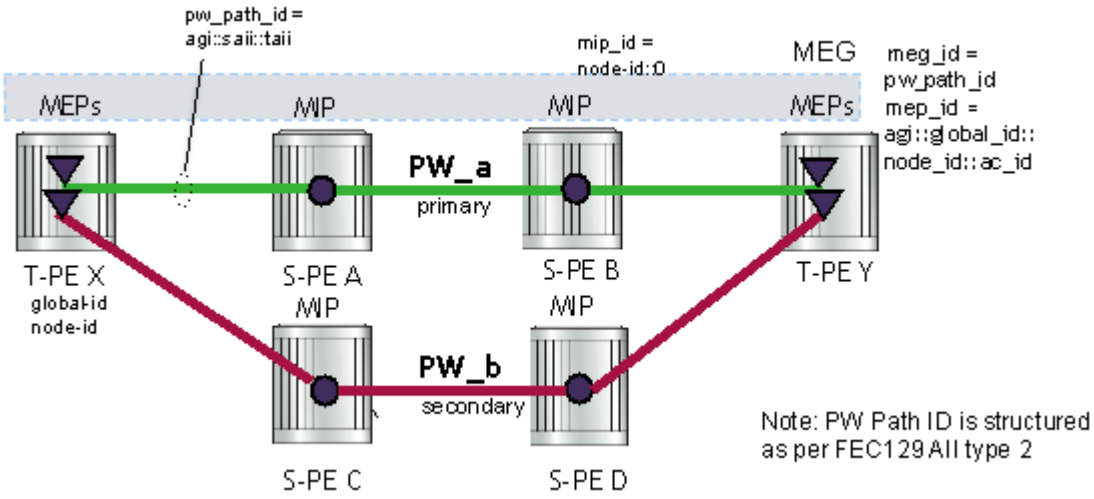


Figure 12: MPLS-TP PW Information Model

In the PW information model shown in Figure 12, the MS-PW is identified by the PW Path ID that is composed of the full AGI:SAII:TAII. The PW Path ID is also the MEP ID at the T-PEs, so a user does not have to explicitly configure a MEP ID; it is automatically derived by the system. For MPLS-TP PWs with static labels, although the PW is not signaled end-to-end, the directionality of the SAII and TAII is taken to be the same as for the equivalent label mapping message i.e. from downstream to upstream. This is to maintain consistency with signaled pseudowires using FEC 129.

On the 7750 SR, an S-PE for an MS-PW with static labels is configured as a pair of spoke-sdps bound together in an VLL service using the vc-switching command. Therefore, the PW Path ID configured at the spoke-sdp level at an S-PE must contain the Global-ID, Node-ID and AC-ID at the far end T-PEs, not the local S-PE. Note that the ordering of the SAII:TAII in the PW Path ID where static PWs are used should be consistent with the direction of signaling of the egress label to a spoke-SDP forming that segment, if that label were signaled using T-LDP (in downstream unsolicited mode). VCCV Ping will check the PW ID in the VCCV Ping echo request message against the configured PW Path ID for the egress PW segment.

Figure 13 shows an example of how the PW Path IDs can be configured for a simple two-segment MS-PW.

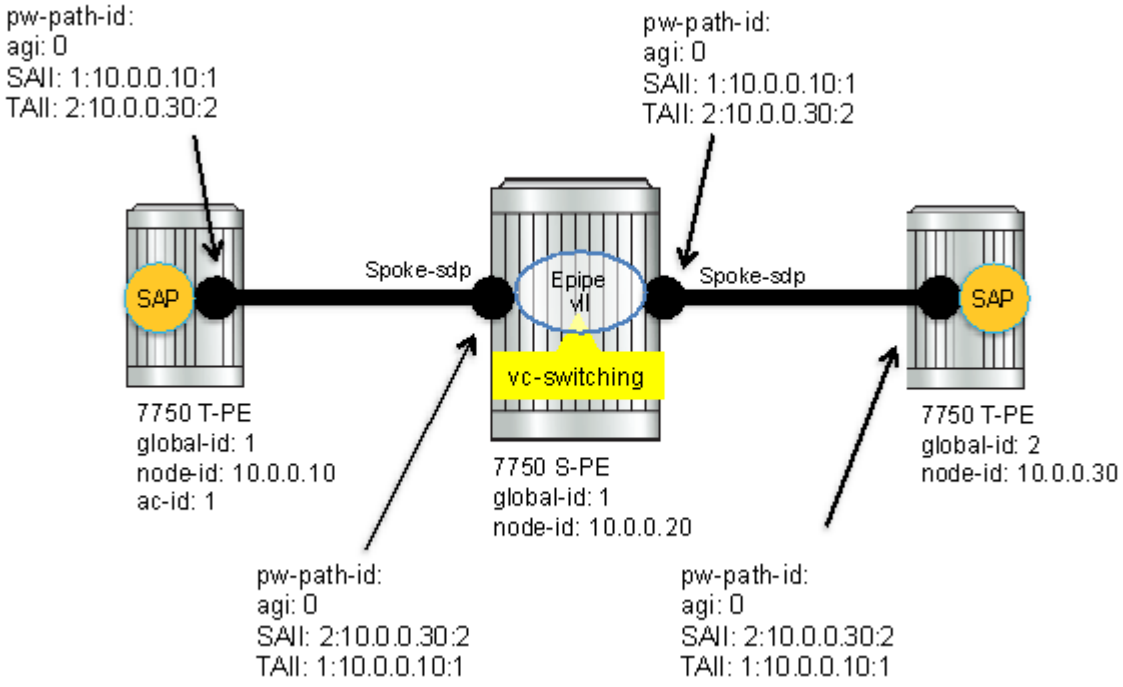


Figure 13: Example usage of PW Identifiers

Generic Associated Channel

MPLS-TP requires that all OAM traffic be carried in-band on both directions of an LSP or PW. This is to ensure that OAM traffic always shares fate with user data traffic. This is achieved by using an associated control channel on an LSP or PW, similar to that used today on PWs. This creates a channel, which is used for OAM, protection switching protocols (e.g. LSP linear protection protection switching coordination), and other maintenance traffic., and is known as the Generic Associated Channel (G-ACh).

RFC5586 specifies mechanisms for implementing the G-ACh, relying on the combination of a reserved MPLS label, the 'Generic-ACH Label (GAL)', as an alert mechanism (value=13) and Generic Associated Channel Header (G-ACH) for MPLS LSPs, and using the Generic Associated Channel Header, only, for MPLS PWs (although the GAL is allowed on PWs). The purpose of the GAL is to indicate that a G-ACH resides at the bottom of the label stack, and is only visible when the bottom non-reserved label is popped. The G-ACH channel type is used to indicate the packet type carried on the G-ACh. Packets on a G-ACh are targeted to a node containing a MEP by ensuring that the GAL is pushed immediately below the label that is popped at the MEP (e.g. LSP endpoint or PW endpoint), so that it can be inspected as soon as the label is popped. A G-ACh

packet is targeted to a node containing a MIP by setting the TTL of the LSP or PW label, as applicable, so that it expires at that node, in a similar manner to the SROS implementation of VCCV for MS-PWs.

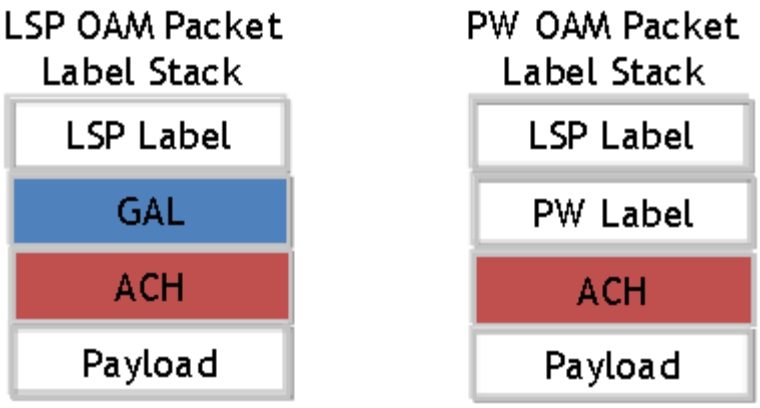


Figure 14: Label for LSP and PW G-ACh Packets

7750 SR supports the G-ACh on static pseudowires and static LSPs.

MPLS-TP Operations, Administration and Maintenance (OAM)

This section details the MPLS-TP OAM mechanisms that are supported.

On-Demand Connectivity Verification (CV) using LSP-Ping

MPLS–TP supports mechanisms for on demand CC/CV as well as route tracing for LSPs and PWs. These are required to enable an operator to test the initial configuration of a transport path, or to assist with fault isolation and diagnosis. On demand CC/CV and route tracing for MPLS-TP is based on LSP-Ping and is described in RFC6426. Three possible encapsulations are specified in that RFC:

- IP encapsulation, using the same label stack as RFC4379, or encapsulated in the IPv4 G-ACh channel with a GAL/ACH
- and non-IP encapsulation with GAL/ACH for LSPs and ACH for PWs.

In IP-encapsulation, LSP-Ping packets are sent over the MPLS LSP for which OAM is being performed and contain an IP/UDP packet within them. The On-demand CV echo response message is sent on the reverse path of the LSP, and the reply contains IP/UDP headers followed by the On-demand CV payload.

In non-IP environments, LSP ping can be encapsulated with no IP/UDP headers in a G-ACh and use a source address TLV to identify the source node, using forward and reverse LSP or PW associated channels on the same LSP or PW for the echo request and reply packets. In this case, no IP/UDP headers are included in the LSP-Ping packets.

The 7750 supportd the following encapsulations:

- IP encapsulation with ACH for PWs (as per VCCV type 1).
- IP encapsulation without ACH for LSPs using labeled encapsution
- Non-IP encapsulation with ACH for both PWs and LSPs.

LSP Ping and VCCV Ping for MPLS-TP use two new FEC sub-types in the target FEC stack in order to identify the static LSP or static PW being checked. These are the Static LSP FEC sub-type, which has the same format as the LSP identifier described above, and the Static PW FEC sub-type,. These are used in-place of the currently defined target FEC stack sub-TLVs.

In addition, MPLS-TP uses a source/destination TLV to carry the MPLS-TP global-id and node-id of the target node for the LSP ping packet, and the source node of the LSP ping packet.

LSP Ping and VCCV-Ping for MPLS-TP can only be launched by the LER or T-PE. The replying node therefore sets the TTL of the LSP label or PW label in the reply packet to 255 to ensure that it reaches the node that launched the LSP ping or VCCV Ping request.

Downstream Mapping Support

RFC4379 specifies four address types for the downstream mapping TLV for use with IP numbered and unnumbered interfaces:

Type #	Address Type	K Octets	Reference
1	IPv4 Numbered	16	RFC 4379
2	IPv4 Unnumbered	16	RFC 4379
3	IPv6 Numbered	40	RFC 4379
4	IPv6 Unnumbered	28	RFC 4379

RFC6426 adds address type 5 for use with Non IP interfaces, inlcuding MPLS-TP interfaces. In addiiton, this RFC specifies that type 5 must be used when non-IP ACH encapsulation is used for LSP Trace.

It is possible to send and respond to a DSMAP/DDMAP TLV in the LSP Trace packet for numbered IP interfaces as per RFC4379. In this case, the echo request message contains a downstream mapping TLV with address type 1 (IPv4 address) and the IPv4 address in the DDMAP/DSMAP TLV is taken to be the IP address of the IP interface that the LSP uses. The LSP

trace packet therefore contains a DSMAP TLV in addition to the MPLS-TP static LSP TLV in the target FEC stack.

DSMAP/DDMAP is not supported for pseudowires.

Proactive CC, CV and RDI

Proactive Continuity Check (CC) is used to detect a loss of continuity defect (LOC) between two MEPs in a MEG. Proactive Connectivity Verification (CV) is used to detect an unexpected connectivity defect between two MEPs (e.g. mis-merging or misconnection), as well as unexpected connectivity within the MEG with an unexpected MEP. This feature implements both functions using proactive generation of OAM packets by the source MEP that are processed by the peer sink MEP. CC and CV packets are always sent in-band such that they fate share with user traffic, either on an LSP, PW or section and are used to trigger protection switching mechanisms.

Proactive CC/CV based on bidirectional forwarding detection (BFD) for MPLS-TP is described in RFC6428. BFD packets are sent using operator configurable timers and encapsulated without UDP/IP headers on a standardized G-ACh channel on an LSP or PW. CC packets simply consist of a BFD control packet, while CV packets also include an identifier for the source MEP in order that the sink MEP can detect if it is receiving packets from an incorrect peer MEP, thus indicating a mis-connectivity defect. Other defect types (including period misconfiguration defect) should be supported. When a supported defect is detected, an appropriate alarm is generated (e.g. log, SNMP trap) at the receiving MEP and all traffic on the associated transport path (LSP or PW) is blocked. This is achieved using linear protection for CC defects, and by blocking the ingress data path for CV defects. The 7750 SR supports both a CC-only mode and a compbine CC / CV mode, as defined in RFC6428.

Note that when an LSP with CV is first configured, the LSP will be held in the CV defect state for 3.5 seconds after the first valid CV packet is received.

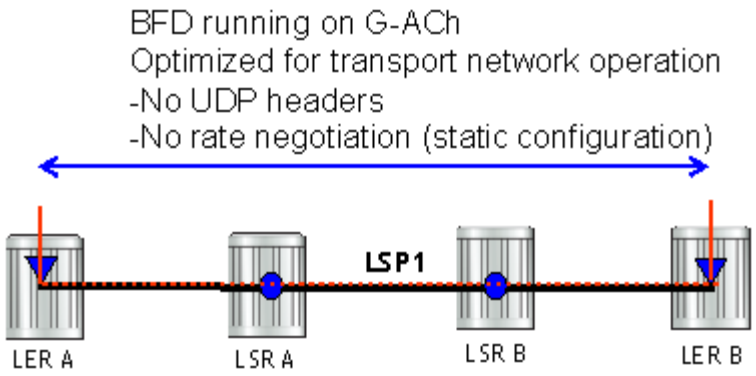


Figure 15: BFD used for proactive CC on MPLS-TP LSP

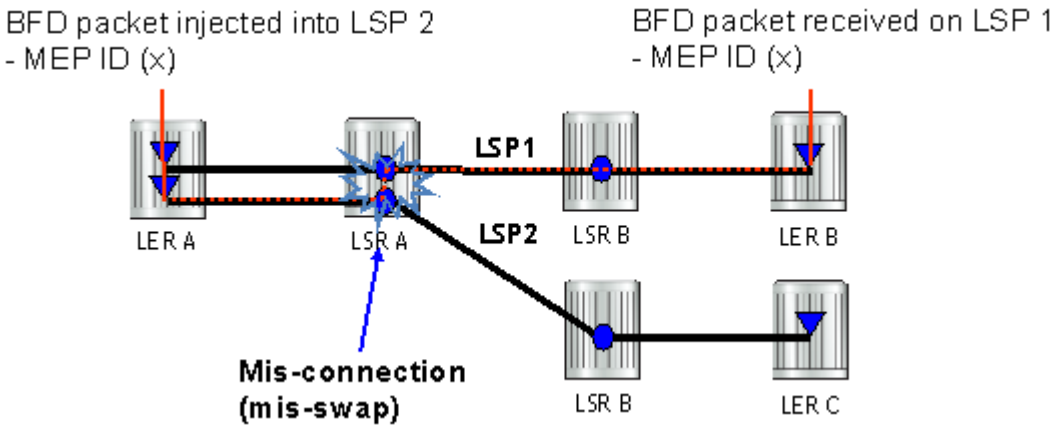


Figure 16: BFD used for proactive CV on MPLS-TP LSP

Linear protection switching of LSPs (see below) is triggered based on a CC or CV defect detected by BFD CC/CV.

Note that RFC6428 defines two BFD session modes: Coordinated mode, in which the session state on both directions of the LSP is coordinated and constructed from a single, bidirectional BFD session, and independent mode, in which two independent sessions are bound together at a MEP. Coordinated mode is supported.

BFD is supported on MPLS-TP LSPs. When BFD_CV detects a mis-connectivity on an LSP, the system will drop all incoming non-OAM traffic with the LSP label (at the LSP termination point) instead of forwarding it to the associated SAP or PW segment.

The following GACh channel types are supported for the combined CC/CV mode:

- 0x22 for BFD CC with no IP encapsulation
- 0x23 for BFD CV

The following G-ACh channel types are used for the CC-only mode:

- 0x07

BFD-based RDI

RDI provides a mechanism whereby the source MEP can be informed of a downstream failure on an LSP, and can thus either raise an alarm, or initiate a protection switching operation. In the case of BFD based CC/CV, RDI is communicated using the BFD diagnostic field in BFC CC/CV messages. The following diagnostic codes are suported:

"1 - Control Detection Time Expired"

"9 - mis-connectivity defect”

PW Control Channel Status Notifications (Static Pseudowire Status Signaling)

MPLS-TP introduces the ability to support a full range of OAM and protection / redundancy on PWs for which no dynamic T-LDP control plane exists. Static PW status signaling is used to advertise the status of a PW with statically configured labels by encapsulating the PW status TLV in a G-ACh on the PW. This mechanism enables OAM message mapping and PW redundancy for such PWs, as defined in RFC6478. This mechanism is known as control channel status signaling in SR OS.

PW control channel status notifications use a similar model to T-LDP status signaling. That is, in general, status is always sent to the nearest neighbor T-PE or S-PE and relayed to the next segment by the S-PE. To achieve this, the PW label TTL is set to 1 for the G-ACh packet containing the status message.

Control channel status notifications are disabled by default on a spoke-sdp. If they are enabled, then the default refresh interval is set to zero (although this value should be configurable in CLI). That is, when a status bit changes, three control channel status packets will be sent consecutively at one-second intervals, and then the transmitter will fall silent. If the refresh timer interval is non-zero, then status messages will continue to be sent at that interval. The system supports the configuration of a refresh timer of 0, or from 10-65535 seconds. The recommended value is 600 seconds.

The 7750 SR supports the optional acknowledgement of a PW control channel status message.

In order to constrain the CPU resources consumed processing control channel status messages, the system implements a credit-based mechanism. If a user enables control channel status on a PW[n], then a certain number of credits c_n are consumed from a CPM-wide pool of max_credit credits. The number of credits consumed is inversely proportional to the configured refresh timer (the first three messages at 1 second interval do not count against the credit). If the current_credit <= 0, then control channel status signaling cannot be configured on a PW (but the PW can still be configured and no shutdown).

If a PE with a non-zero refresh timer configured does not receive control channel status refresh messages for 3.5 time the specified timer value, then by default it will time out and assume a PW status of zero.

A trap is generated if the refresh timer times-out.

If PW redundancy is configured, the system will always consider the literal value of the PW status; a time-out of the refresh timer will not impact the choice of the active transit object for the VLL

service. The result of this is that if the refresh timer times-out, and a given PW is currently the active PW, then the system will not fail-over to an alternative PW if the status is zero and some lower-layer OAM mechanism e.g. BFD has not brought down the LSP due to a connectivity defect. It is recommended that the PW refresh timer be configured with a much longer interval than any proactive OAM on the LSP tunnel, so that the tunnel can be brought down before the refresh timer expires if there is a CC defect.

Note that a unidirectional continuity fault on a RSVP TE LSP may not result in the LSP being brought down before the received PW status refresh timer expires. It is therefore recommended that either bidirectional static MPLS-TP LSPs with BFD CC, or additional protection mechanisms e.g. FRR be used on RSVP-TE LSPs carrying MPLS-TP PWs. This is particularly important in active/standby PW dual homing configurations, where the active / standby forwarding state or operational state of every PW in the redundancy set must be accurately reflected at the redundant PE side fo the configuration.

Note that a PW with a refresh timer value of zero is always treated as having not expired.

The 7750 SR implements a hold-down timer for control-channel-status pw-status bits in order to suppress bouncing of the status of a PW. For a specific spoke-sdp, if the system receives 10 pw-status "change" events in 10 seconds, the system will "hold-down" the spoke-sdp on the local node with the last received non-zero pw-status bits for 20 seconds. It will update the local spoke with the most recently received pw-status. This hold down timer is not persistent across shutdown/no-shutdown events.

PW Control Channel Status Request Mechanism

The 7750 SR implements an optional PW control channel status request mechanism. This enhances the existing control channel status mechanism so that a peer that has “stale” PW status for the far-end of a PW can request that the peer PE send a static PW status update. Accurate and current information about the far end status of a PW is important for proper operation of PW redundancy. This mechanism ensures a consistent view of the control plane is maintained, as far as possible, between peer nodes. It is not intended to act as a continuity check between peer nodes.

Pseudowire Redundancy and Active / Standby Dual Homing

PW redundancy is supported for statuc MPLS-TP pseudowires. However, instead of using T-LDP status signaling to signal the forwarding state of a PW, control channel status signaling is used.

The following PW redundancy scenarios must be supported:

- MC-LAG and MC-APS with single and multi-segment PWs interconnecting the PEs.
- MS-PW (S-PE) Redundancy between VLL PEs with single-homed CEs.

- Dual-homing of a VLL service into redundant IES or VPRN PEs, with active/standby PWs.
- Dual-homing of a VLL service into a VPLS with active/standby PWs.

Note that active/standby dual-homing into routed VPLS is not supported in for MPLS-TP PWs. This is because it relies on PW label withdrawal of the standby PW in order to take down the VPLS instance, and hence the associated IP interface. Instead, it is possible to enable BGP multi-homing on a routed VPLS that has MPLS-TP PWs as spokes, and for the PW status of each spoke-sdp to be driven (using control channel status) from the active or standby forwarding state assigned to each PW by BGP.

It is possible to configure inter-chassis backup (ICB) PWs as static MPLS-TP PWs with MPLS-TP identifiers. Only MPLS-TP PWs are supported in the same endpoint. That is, PWs in an endpoint must either be all MPLS-TP, or none of them must be MPLS-TP. This implies that an ICB used in an endpoint for which other PWs are MPLS TP must also be configured as an MPLS-TP PW.

A failover to a standby pseudowire is initiated based on the existing supported methods (e.g. failure of the SDP).

MPLS-TP LSP Protection

Linear 1-for-1 protection of MPLS-TP LSPs is supported, as defined in RFC. This applies only to LSPs (not PWs).

This is supported edge-to-edge on an LSP, between two LERs, where normal traffic is transported either on the working LSP or on the protection LSP using a logical selector bridge at the source of the protected LSP.

At the sink LER of the protected LSP, the LSP that carries the normal traffic is selected, and that LSP becomes the working LSP. A protection switching coordination (PSC) protocol coordinates between the source and sink bridge, which LSP will be used, as working path and protection path. The PSC protocol is always carried on a G-ACh on the protection LSP.

The 7750 SR supports single-phased coordination between the LSP endpoints, in which the initiating LER performs the protection switchover to the alternate path and informs the far-end LER of the switch.

Bidirectional protection switching is achieved by the PSC protocol coordinating between the two end points to determine which of the two possible paths (i.e. the working or protect path), transmits user traffic at any given time.

It is possible to configure non-revertive or revertive behavior. For non-revertive, the LSP will not switch back to the working path when the PSC switchover requests end, while for revertive configurations, the LSP always returns back to the working path when the switchover requests end.

The following figures illustrate the behavior of linear protection in more detail.

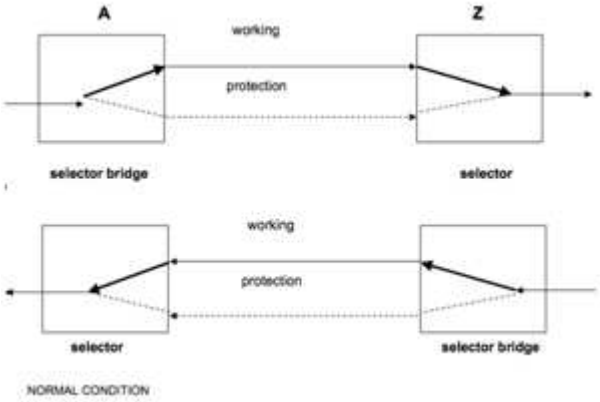


Figure 17: Normal Operation

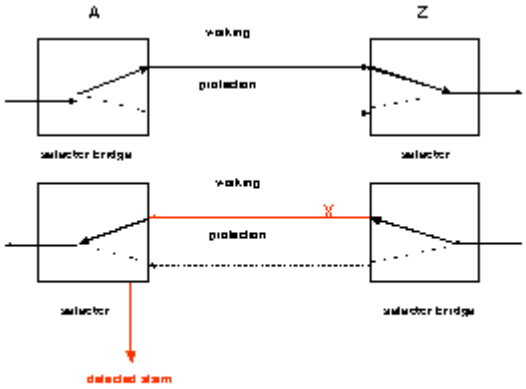


Figure 18: Failed Condition

In normal condition, user data packets are sent on the working path on both directions, from A to Z and Z to A.

A defect in the direction of transmission from node Z to node A impacts the working connection Z-to-A, and initiates the detection of a defect at the node A.

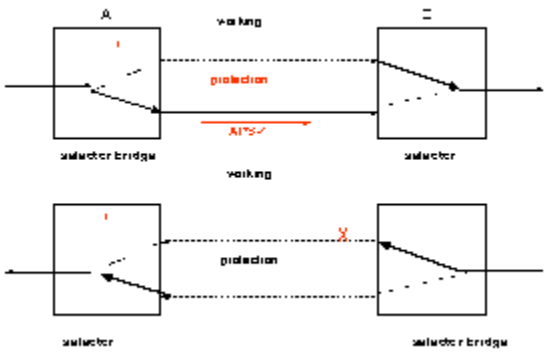


Figure 19: Failed Condition - Switching at A

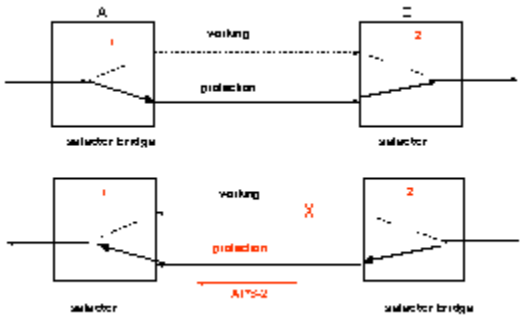


Figure 20: Failed Condition - Switching at Z

The unidirectional PSC protocol initiates protection switching: the selector bridge at node A is switched to protection connection A-to-Z and the selector at node A switches to protection connection Z to-A. The PSC packet, sent from node A to node Z, requests a protection switch to node Z.

After node Z validates the priority of the protection switch request, the selector at node Z is switched to protection connection A-to-Z and the selector bridge at the node Z is switched to protection connection Z-to-A. The PSC packet, sent from node Z to node A, is used as acknowledge, informing node A about the switching.

If BFD CC or CC/CV OAM packets are used to detect defects on the working and protection path, they are inserted on both working and protection paths. It should be noted that they are sent regardless of whether the selected as the currently active path.

The 7750 supports the following operator commands:

- Forced Switch
- Manual Switch
- Clear,
- Lockout of protection

Configuring MPLS-TP

This section describes the steps required to configured MPLS-TP.

Configuration Overview

The following steps must be performed in order to configure MPLS-TP LSPs or PWs.

At the 7x50 LER and LSR:

1. Create an MPLS-TP context, containing nodal MPLS-TP identifiers. This is configured under **config>router>mpls>mpls-tp**.
2. Ensure that a sufficient range of labels is reserved for static LSPs and PWs. This is configured under **config>router>mpls-labels>static-labels**.
3. Ensure that a range of tunnel identifiers is reserved for MPLS-TP LSPs under **config>router>mpls-mpls-tp>tp-tunnel-id-range**.
4. A user may optionally configure MPLS-TP interfaces, which are interfaces that no not use IP addressing or ARP for next hop resolution. These can only be used by MPLS-TP LSPs.

At the 7x50 LER, configure:

1. OAM Templates. These contain generic parameters for MPLS-TP proactive OAM. An OAM template is configured under **config>router>mpls>mpls-tp>oam-template**.
2. BFD templates. These contain generic parameters for BFD used for MPLS-TP LSPs. A BFD template is configured under **config>router>bfd>bfd-template**.
3. Protection templates. These contain generic parameters for MPLS-TP 1-for-1 linear protection. A protection template is configured under **config>router>mpls>mpls-tp>protection-template**.
4. MPLS-TP LSPs are configured under **config>router>mpls>lsp mpls-tp**
5. Pseudowires using MPLS-TP are configured as spoke-sdps with static PW labels.

At an LSR, a use must configure an LSP transit-path under **config>router>mpls>mpls-tp>transit-path**.

The following sections describe these configuration steps in more detail.

Node-Wide MPLS-TP Parameter Configuration

Generic MPLS-TP parameters are configured under **config>router>mpls>mpls-tp**. If a user configures **no mpls**, normally the entire mpls configuration is deleted. However, in the case of mpls-tp a check that there is no other mpls-tp configuration e.g. services or tunnels using mpls-tp on the node, will be performed.

The mpls-tp context is configured as follows:

```
config
   router
      mpls
         [no] mpls-tp
            . . .
            [no] shutdown
```

MPLS-TP LSPs may be configured if the mpls-tp context is administratively down (shutdown), but they will remain down until the mpls-tp context is configured as administratively up. No programming of the data path for an MPLS-TP Path occurs until the following are all true:

- Mpls-tp context is **no shutdown**
- Mpls-tp LSP context is **no shutdown**
- MPLS-TP Path context is **no shutdown**

A **shutdown** of mpls-tp will therefore bring down all MPLS-TP LSPs on the system.

The mpls-tp context cannot be deleted if MPLS-TP LSPs or SDPs exist on the system.

Node-Wide MPLS-TP Identifier Configuration

MPLS-TP identifiers are configured for a node under the following CLI tree:

```
config
   router
     mpls
       mpls-tp
          global-id <global-id>
          node-id {<ipv4address> | | <1.. .4,294,967,295>}
          [no] shutdown
          exit
```

The default value for the global-id is 0. This is used if the global-id is not explicitly configured. If a user expects that inter domain LSPs will be configured, then it is recommended that the global ID should be set to the local ASN of the node, as configured under **config>system**. If two-byte ASNs are used, then the most significant two bytes of the global-id are padded with zeros.

The default value of the node-id is the system interface IPv4 address. The MPLS-TP context cannot be administratively enabled unless at least a system interface IPv4 address is configured because MPLS requires that this value is configured.

These values are used unless overridden at the LSP or PW end-points, and apply only to static MPLS-TP LSPs and PWs.

In order to change the values, **config>router>mpls>mpls-tp** must be in the shutdown state. This will bring down all of the MPLS-TP LSPs on the node. New values are propagated to the system when a **no shutdown** is performed.

Static LSP and pseudowire (VC) Label and Tunnel Ranges

SR OS reserves a range of labels for use by static LSPs, and a range of labels for use by static pseudowires (SVCs) i.e. LSPs and pseudowires with no dynamic signaling of the label mapping. These are configured as follows:

```
config
   router
      mpls-labels
         [no] static-label max-lsp-labels <number>
              static-svc-label <number>
```

<number>: indicates the maximum number of labels for the label type.

The minimum label value for the static LSP label starts at 32 and expands all the way to the maximum number specified. The static VC label range is contiguous with this. The dynamic label range exists above the static VC label range (the label ranges for the respective label type are contiguous). This prevents fragmentation of the label range.

The MPLS-TP tunnel ID range is configured as follows:

```
config
   router
      mpls
         mpls-tp
            [no] tp-tunnel-id-range <start-id> <end-id>
```

The tunnel ID range referred to here is a contiguous range of RSVP-TE Tunnel IDs is reserved for use by MPLS TP, and these IDs map to the MPLS-TP Tunnel Numbers. There are some cases where the dynamic LSPs may have caused fragmentation to the number space such that contiguous range {max-min} is not available. In these cases, the command will fail.

There is no default value for the tunnel id range, and it must be configured to enable MPLS-TP.

If a configuration of the tunnel ID range fails, then the system will give a reason. This could be that the initially requested range, or the change to the allocated range, is not available i.e. tunnel IDs in that range have already been allocated by RSVP-TE. Allocated Tunnel IDs are visible using a show command.

Note that changing the LSP or static VC label ranges does not require a reboot.

Note also that the static label ranges for LSPs, above, apply only to static LSPs configured using the CLI tree for MPLS-TP specified in this section. Different scalability constraints apply to static LSPs configured using the following CLI introduced in earlier SR OS releases:

config>router>mpls>static-lsp

config>router>mpls>interface>label-map

The scalability applying to labels configured using this CLI is enforced as follows:

- A maximum of 1000 static LSP names may be configured with a PUSH operation.
- A maximum of 1000 LSPs with a POP or SWAP operation may be configured.

These two limits are independent of one another, giving a combined limit of 1000 PUSH and 1000 POP/SAP operations configured on a node.

The static LSP and VC label spaces are contiguous. Therefore, the dimensioning of these label spaces requires careful planning by an operator as increasing the static LSP label space impacts the start of the static VC label space, which may already-deployed

Interface Configuration for MPLS-TP

It is possible for MPLS-TP paths to use both numbered IP numbered interfaces that use ARP/static ARP, or IP unnumbered interfaces. MPLS-TP requires no changes to these interafces. It is also possible to use a new type of interface that does not require any IP addressing or next-hop resolution.

Draft-ietf-mpls-tp-next-hop-addressing provides guidelines for the usage of various Layer 2 next-hop resolution mechanisms with MPLS-TP. If protocols such as ARP are supported, then they should be used. However, in the case where no dynamic next hop resolution protocol is used, it should be possible to configure a unicast, multicast or broadcast next-hop MAC address. The rationale is to minimize the amount of configuration required for upstream nodes when downstream interfaces are changes. A default multicast MAC address for use by MPLS-TP point-to-point LSPs has been assigned by IANA (Value: 01-00-5e-90-00-00). This value is configurable on the 7x50 to support interoperability with 3rd party implementations that do not default to this value, and this no default value is implemented on the 7x50.

In order to support these requirements, a new interface type, known as an unnumbered MPLS-TP interface is introduced. This is an unnumbered interface that allows a broadcast or multicast destination MAC address to be configured. An unnumbered MPLS-TP interface is configured using the **unnumbered-mpls-tp** keyword, as follows:

```
config
   router
      interface <if-name> [unnumbered-mpls-tp]
         port <port-id>[:encap-val]
         mac <local-mac-address>
         static-arp <remote-mac-addr>
         //ieee-address needs to support mcast and bcast
                  exit
```

The **remote-mac-address** may be any unicast, broadcast of multicast address. However, a broadcast or multicast remote-mac-address is only allowed in the **static-arp** command on Ethernet unnumbered interfaces when the **unnumbered-mpls-tp** keyword has been configured. This also allows the interface to accept packets on a broadcast or any multicast MAC address. Note that if a packet is received with a unicast destination MAC address, then it will be checked against the configured <local-mac-address> for the interface, and dropped if it does not match. When an interface is of type **unnumbered-mpls-tp**, only MPLS-TP LSPs are allowed on that interface; other protocols are blocked from using the interface.

An unnumbered MPLS-TP interface is assumed to be point-to-point, and therefore users must ensure that the associated link is not broadcast or multicast in nature if a multicast or broadcast remote MAC address is configured.

The following is a summary of the constraints of an unnumbered MPLS-TP interface:

- It is unnumbered and may borrow/use the system interface address
- It prevents explicit configuration of a borrowed address
- It prevents IP address configuration
- It prevents all protocols except mpls
- It prevents Deletion if an MPLS-TP LSP is bound to the Interface
- It is allowed only in network chassis mode D

MPLS-TP is only supported over Ethernet ports in Release 11.0. The system will block the association of an MPLS-TP LSP to an interface whose port is non-Ethernet.

LER Configuration for MPLS-TP

LSP and Path Configuration

MPLS-TP tunnels are configured using the **mpls-tp** LSP type at an LER under the LSP configuration, using the following CLI tree:

```
config
   router
      mpls
         lsp <xyz> [bypass-only|p2mp-lsp|mpls-tp <src-tunnel-num>]
            to node-id {<a.b.c.d> | <1.. .4,294,967,295>}
            dest-global-id <global-id>
             dest-tunnel-number <tunnel-num>
             [no] working-tp-path
                lsp-num <lsp-num>
                in-label <in-label>
                out-label <out-label> out-link <if-name>
                          [next-hop <ipv4-address>]
                [no] mep
                   [no] oam-template <name>
                   [no] bfd-enable [cc | cc_cv]  // defaults to cc
                   [no] shutdown
                   exit
                [no] shutdown
                exit
             [no] protect-tp-path
                lsp-num <lsp-num>
                in-label <in-label>
                out-label <out-label> out-link <if-name>
                          [next-hop <ipv4-address> ]
                [no] mep
                   [no] protection-template <name>
                   [no] oam-template <name>
                   [no] bfd-enable [cc | cc_cv]  //defaults to cc
                   [no] shutdown
                   exit
                [no] shutdown
                exit
```

<*if-name*> could be numbered or unnumbered interface using an Ethernet port.

<*src-tunnel-num*> is a mandatory create time parameter for mpls-tp tunnels, and has to be assigned by the user based on the configured range of tunnel ids. The *src-global-id* used for the LSP ID is derived from the node-wide *global-id* value configured under config>router>mpls>mpls-tp. A tunnel can not be **un shutdown** unless the *global-id* is configured.

The from address of an LSP to be used in the tunnel identifier is taken to be the local node's node-id/global-id, as configured under config>router>mpls>mpls-tp. If that is not explicitly configured, either, then the default value of the system interface IPv4 address is used

The **to node-id** address may be entered in 4-octet IPv4 address format or unsigned 32-bit format. This is the far-end node-id for the LSP, and does do need to be routable IP addresses.

The **from** and **to** addresses are used as the from and to node-id in the MPLS-TP Tunnel Identifier used for the MEP ID.

Each LSP consists of a working-tp-path and, optionally, a protect-tp-path. The protect-tp-path provides protection for the working-tp-path is 1:1 linear protection is configured (see below). Proactive OAM, such as BFD, is configured under the MEP context of each path. Protection for the LSP is configured under the protect-tp-path mep context.

The ‘to’ global-id is an optional parameter. If it is not entered, then the dest global ID takes the default value of 0. Global ID values of 0 are allowed and indicate that the node’s configured Global ID should be used. If the local global id value is 0, then the remote ‘to’ global ID must also be 0. The ‘to’ global ID value cannot be changed if an LSP is in use by an SDP.

The ‘to’ tunnel number is an optional parameter. If it is not entered, then it is taken to be the same value as the source tunnel number.

LSPs are assumed to be bidirectional and co-routed in Relase 11.0. Therefore, system will assume that the incoming interface is the same as the out-link.

The next-hop <ip-address> can only be configured if the out-link if-name refers to a numbered IP interface. In this case, the system will determine the interface to use to reach the configured next-hop, but will check that the user-entered value for the out-link corresponds to the link returned by the system. If they do not correspond, then the path will not come up. Note that if a user changes the physical port referred to in the interface configuration, then BFD, if configured on the LSP, will go down. Users should therefore ensure that an LSP is moved to a different interface with a different port configuration in order to change the port that it uses. This is enforced by blocking the next-hop configuration for an unnumbered interface.

There is no check made that a valid ARP entry exists before allowing a path to be un shut. Therefore, a path will only be held down if BFD is down. If static ARP is not configured for the interface, then it is assumed that dynamic ARP is used. The result is that if BFD is not configured, a path can come up before ARP resolution has completed for an interface. If BFD is not used, then it is recommended that the connectivity of the path is explicitly checked using on-demand CC/CV prior to sending user traffic on it.

The following is a list of additional considerations for the configuration of MPLS-TP LSPs and paths:

- The working-tp-path must be configured before the protect-tp-path.
- Likewise, the protect-tp-path has to be deleted first before the working-tp-path.
- The *lsp-num* parameter is optional. It’s default value is ‘1’ for the working-tp-path and ‘2’ for protect-tp-path.
- The **mep** context must be deleted before a path can be deleted.
- An MPLS interface needs to be created under **config>router>mpls>interface** before using/specifying the out-label/out-link in in the Forward path for an MPLS-TP LSP. Creation of the LSP will fail if the corresponding mpls interface doesn't exist even though the specified router interface may be valid.

- The system will program the MPLS-TP LSP information upon a '**no shutdown**' of the TP-Path only on the very first **no shutdown**. The Working TP-Path is programmed as the Primary and the Protection TP-Path is programmed as the 'backup'.
- The system will not de-program the IOM on an 'admin shutdown' of the MPLS-TP path. Traffic will gracefully move to the other TP-Path if valid, as determined by the proactive MPLS-TP OAM. This should not result in traffic loss. However it is recommended that the user does moves traffic to the other TP-Path through a tools command before doing 'admin shut' of an Active TP-Path.
- Deletion of the out-label/out-link sub-command under the MPLS-TP Path is not allowed once configured. These can only be modified.
- MPLS will allow the deletion of an 'admin shutdown' TP-Path. This will cause MPLS to de-program the corresponding TP-Path forwarding information from IOM. This can cause traffic loss for certain users that are bound to the MPLS-TP LSP.
- MPLS will not de-program the IOM on a specific interface admin shut/clear unless the interface is a System Interface. However, if mpls informs the TP-OAM module that the mpls interface has gone down, then it triggers a switch to the standby tp-path if the associated interface went down and if it is valid.
- If a MEP is defined and shutdown, then the corresponding path is also operationally down. Note, however, that the MEP admin state is applicable only when a MEP is created from an MPLS-TP path.
- It is not mandatory to configure BFD or protection on an MPLS-TP path in order to bring the LSP up.
- If **bfd-enable cc** is configured, then CC-only mode using ACh channel 0x07 is used. If **bfd-enable cc_v** is configured, then BFD CC packets use channel 0x22 and CV packets use channel 0x23.

The protection template is associated with a LSP as a part of the MEP on the protect path. If only a working path is configured, then the protection template is not configured.

BFD cannot be enabled under the MEP context unless a named BFD template is configured.

Proactive CC/CV (using BFD) Configuration

Generally applicable proactive OAM parameters are configured using templates.

Proactive CC and CV uses BFD parameters such as Tx/Rx timer intervals, multiplier and other session/fault management parameters which are specific to BFD. These are configured using a BFD Template. The BFD Template may be used for non-MPLS-TP applications of BFD, and therefore contains the full set of possible configuration parameters for BFD. Only a sub-set of these may be used for any given application.

Generic MPLS-TP OAM and fault management parameters are configured in the OAM Template.

Named templates are referenced from the MPLS-TP Path MEP configuration, so different parameter values are possible for the working and protect paths of a tunnel.

The BFD Template is configured as follows:

```
config
   router
      bfd
         [no] bfd-template <name>
            [no] transmit-interval <transmit-interval>
            [no] receive-interval <receive-interval>
            [no] echo-receive <echo-interval>
            [no] multiplier <multiplier>
            [no] type <cpm-np>
            exit
```

The parameters are as follows:

- **transmit-interval** *transmit-interval* and the **rx** *receive-interval*: These are the transmit and receive timers for BFD packets. If the template is used for MPLS-TP, then these are the timers used by CC packets. Values are in milliseconds: 10ms to 100,000ms, with 1ms granularity. Default 10ms for CPM3 or better, 1 sec for other hardware. Note that for MPLS-TP CV packets, a transmit interval of 1 sec is always used.
- **multiplier** *multiplier*: Integer 3 – 20. Default: 3. This parameter is ignored for MPLS-TP combined cc-v BFD sessions, and the default of 3 used, as per RFC6428..
- **echo-receive** *echo-interval*: Sets the minimum echo receive interval, in milliseconds, for a session. Values: 100ms – 100,000ms. Default: 100. This parameter is not used by a BFD session for MPLS-TP.
- **type cpm-np**: This selects the CPM network processor as the local termination point for the BFD session. This is enabled by default.

Note that if the above BFD timer values are changed in a given template, any BFD sessions on MEPs to which that template is bound will try to renegotiate their timers to the new values. Note that the BFD implementations in some MPLS-TP peer nodes may not be able handle this renegotiation, as allowed by Section 3.7.1 of RFC6428 and may take the BFD session down. This could result in undesired behavior, for example an unexpected protection switching event. It is therefore recommended that in these circumstances, user of the 7750 SR exercise care in modifying the BFD timer values after a BFD session is UP.

Commands within the Bfd-template use a begin-commit model. To edit any value within the BFD template, a <begin> needs to be executed once the template context has been entered. However, a value will still be stored temporarily until the commit is issued. Once the commit is issued, values will actually be used by other modules like the mpls-tp module and bfd module.

A BFD template is referenced from the OAM template. The OAM Template is configured as follows:

```
config
```

```
router
   mpls
      mpls-tp
         [no] oam-template <name>
            [no] bfd-template <name>
            [no] hold-time-down <interval>
            [no] hold-time-up <interval>
         exit
```

- **hold-time-down** *interval*: 0-5000 deciseconds, 10ms steps, default 0. This is equivalent to the standardized hold-off timer.
- **hold-time-up** *interval*: 0-500 centiseconds in 100ms steps, default 2 seconds This is an additional timer that can be used to reduce BFD bouncing.
- **bfd-template** *name*: This is the named BFD template to use for any BFD sessions enabled under a MEP for which the OAM template is configured.

An OAM template is then applied to a MEP as described above.

Protection templates and Linear Protection Configuration

Protection templates defines the generally applicable protection parameters for an MPLS-TP tunnel. Only linear protection is supported, and so the application of a named template to an MPLS-TP tunnel implies that linear protection is used.

A template is configured as follows:

```
config
   router
      mpls
         mpls-tp
            protection-template <name>
              [no] revertive
               [no] wait-to-restore <interval>
               rapid-psc-timer <interval>
               slow-psc-timer <interval>
               exit
```

The allowed values are as follows:

- **wait-to-restore** *interval*: 0-720 seconds, 1 sec steps, default 300 seconds. This is applicable to revertive mode only.
- **rapid-psc-timer** *interval*: [10, 100, 1000ms]. Default 100ms
- **slow-psc-timer** *interval*: 5s-60s. Default: 5s
- **revertive**: Selects revertive behavior. Default: no revertive.

LSP Linear Protection operations are enacted using the following **tools>perform** commands.

```
tools>perform>router>mpls
            tp-tunnel
               clear {<lsp-name> | id <tunnel-id>}
               force {<lsp-name> | id <tunnel-id>}
               lockout {<lsp-name> | id <tunnel-id>}
               manual {<lsp-name> | id <tunnel-id>}
            exit
         exit
```

To minimize outage times, users should use the “mpls-tp protection command” (e.g. force/manual) to switch all the relevant MPLS-TP paths before executing the following commands:

- clear router mpls interface <>
- config router mpls interface <> shut

Intermediate LSR Configuration for MPLS-TP LSPs

The forward and reverse directions of the MPLS-TP LSP Path at a transit LSR are configured using the following CLI tree:

```
config
   router
      mpls
         mpls-tp
            transit-path <path-name>
               [no] path-id {lsp-num <lsp-num>|working-path|protect-path

                  [src-global-id <global-id>]
                  src-node-id {<ipv4address> | <1.. .4,294,967,295>}
                  src-tunnel-num <tunnel-num>
                  [dest-global-id <global-id>]
                  dest-node-id {<ipv4address> | <1.. .4,294,967,295>}
                  [dest-tunnel-num <tunnel-num>]}

               forward-path
                  in-label <in-label> out-label <out-label>
                       out-link <if-name> [next-hop <ipv4-next-hop>]
               reverse-path
                  in-label <in-label> out-label <out-label>
                      [out-link <if-name> [next-hop <ipv4-next-hop>]
               [no] shutdown
```

Note that the *src-tunnel-num* and *dest-tunnel-num* are consistent with the source and destination of a label mapping message for a signaled LSP.

If *dest-tunnel-num* is not entered in CLI, the *dest-tunnel-num* value is taken to be the same as the src-tunnel-num value.

If any of the *global-id* values are not entered, the value is taken to be 0.

If the *src-global-id* value is entered, but the *dest-global-id* value is not entered, *dest-global-id* value is the same as the *src-global-id* value.

Note that the *lsp-num* must match the value configured in the LER for a given path. If no explicit lsp-num is configured, then working-path or protect-path must be specified (equating to 1 or 2 in the system).

The forward path must be configured before the reverse path. The configuration of the reverse path is optional.

The LSP-ID (path-id) parameters apply with respect to the downstream direction of the forward LSP path, and are used to populate the MIP ID for the path at this LSR.

The reverse path configuration must be deleted before the forward path.

The forward-path (and reverse-path if applicable) parameters can be configured with or without the path-id, but they must be configured if MPLS-TP OAM is to be able to identify the LSR MIP.

The transit-path can be no shutdown (as long as the forward-path/reverse-path parameters have been configured properly) with or without identifiers.

The path-id and path-name must be unique on the node. There is a one to one mapping between a given path-name and path-id.

Traffic can not pass through the transit-path if the transit-path is in the **shutdown** state.

RSVP

The Resource Reservation Protocol (RSVP) is a network control protocol used by a host to request specific qualities of service from the network for particular application data streams or flows. RSVP is also used by routers to deliver quality of service (QoS) requests to all nodes along the path(s) of the flows and to establish and maintain state to provide the requested service. RSVP requests generally result in resources reserved in each node along the data path. MPLS leverages this RSVP mechanism to set up traffic engineered LSPs. RSVP is not enabled by default and must be explicitly enabled.

RSVP requests resources for simplex flows. It requests resources only in one direction (unidirectional). Therefore, RSVP treats a sender as logically distinct from a receiver, although the same application process may act as both a sender and a receiver at the same time. Duplex flows require two LSPs, to carry traffic in each direction.

RSVP is not a routing protocol. RSVP operates with unicast and multicast routing protocols. Routing protocols determine where packets are forwarded. RSVP consults local routing tables to relay RSVP messages.

RSVP uses two message types to set up LSPs, PATH and RESV. Figure 21 depicts the process to establish an LSP.

- The sender (the ingress LER (ILER)), sends PATH messages toward the receiver, (the egress LER (ELER)) to indicate the FEC for which label bindings are desired. PATH messages are used to signal and request label bindings required to establish the LSP from ingress to egress. Each router along the path observes the traffic type.

  PATH messages facilitate the routers along the path to make the necessary bandwidth reservations and distribute the label binding to the router upstream.
- The ELER sends label binding information in the RESV messages in response to PATH messages received.
- The LSP is considered operational when the ILER receives the label binding information.

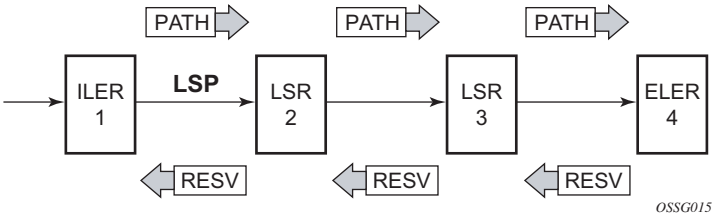


Figure 21: Establishing LSPs

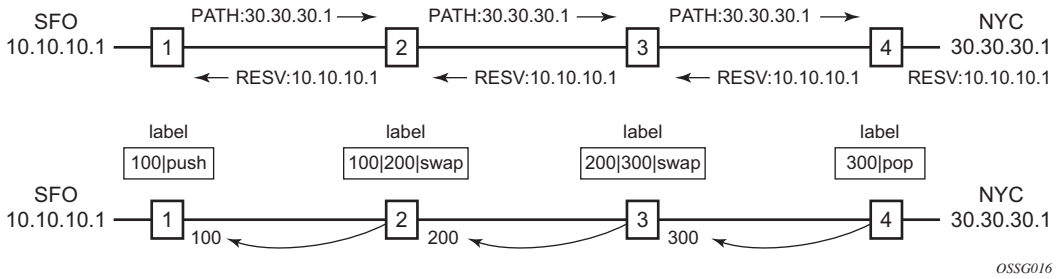


Figure 22: LSP Using RSVP Path Set Up

Figure 22 displays an example of an LSP path set up using RSVP. The ingress label edge router (ILER 1) transmits an RSVP path message (path: 30.30.30.1) downstream to the egress label edge router (ELER 4). The path message contains a label request object that requests intermediate LSRs and the ELER to provide a label binding for this path.

In addition to the label request object, an RSVP PATH message can also contain a number of optional objects:

- Explicit route object (ERO) — When the ERO is present, the RSVP path message is forced to follow the path specified by the ERO (independent of the IGP shortest path).
- Record route object (RRO) — Allows the ILER to receive a listing of the LSRs that the LSP tunnel actually traverses.
- A session attribute object controls the path set up priority, holding priority, and local-rerouting features.

Upon receiving a path message containing a label request object, the ELER transmits a RESV message that contains a label object. The label object contains the label binding that the downstream LSR communicates to its upstream neighbor. The RESV message is sent upstream towards the ILER, in a direction opposite to that followed by the path message. Each LSR that processes the RESV message carrying a label object uses the received label for outgoing traffic associated with the specific LSP. When the RESV message arrives at the ingress LSR, the LSP is established.

Using RSVP for MPLS

Hosts and routers that support both MPLS and RSVP can associate labels with RSVP flows. When MPLS and RSVP are combined, the definition of a flow can be made more flexible. Once an LSP is established, the traffic through the path is defined by the label applied at the ingress node of the LSP. The mapping of label to traffic can be accomplished using a variety of criteria. The set of packets that are assigned the same label value by a specific node are considered to belong to the same FEC which defines the RSVP flow.

For use with MPLS, RSVP already has the resource reservation component built-in which makes it ideal to reserve resources for LSPs.

RSVP Traffic Engineering Extensions for MPLS

RSVP has been extended for MPLS to support automatic signaling of LSPs. To enhance the scalability, latency, and reliability of RSVP signaling, several extensions have been defined. Refresh messages are still transmitted but the volume of traffic, the amount of CPU utilization, and response latency are reduced while reliability is supported. None of these extensions result in backward compatibility problems with traditional RSVP implementations.

Hello Protocol

The Hello protocol detects the loss of a neighbor node or the reset of a neighbor's RSVP state information. In standard RSVP, neighbor monitoring occurs as part of RSVP's soft-state model. The reservation state is maintained as cached information that is first installed and then periodically refreshed by the ingress and egress LSRs. If the state is not refreshed within a specified time interval, the LSR discards the state because it assumes that either the neighbor node has been lost or its RSVP state information has been reset.

The Hello protocol extension is composed of a hello message, a hello request object and a hello ACK object. Hello processing between two neighbors supports independent selection of failure detection intervals. Each neighbor can automatically issue hello request objects. Each hello request object is answered by a hello ACK object.

MD5 Authentication of RSVP Interface

When enabled on an RSVP interface, authentication of RSVP messages operates in both directions of the interface.

A node maintains a security association with its neighbors for each authentication key. The following items are stored in the context of this security association:

- The HMAC-MD5 authentication algorithm.
- Key used with the authentication algorithm.
- Lifetime of the key. A key is user-generated key using a third party software/hardware and enters the value as static string into CLI configuration of the RSVP interface. The key will continue to be valid until it is removed from that RSVP interface.
- Source Address of the sending system.
- Latest sending sequence number used with this key identifier.

The RSVP sender transmits an authenticating digest of the RSVP message, computed using the shared authentication key and a keyed-hash algorithm. The message digest is included in an Integrity object which also contains a Flags field, a Key Identifier field, and a Sequence Number field. The RSVP sender complies to the procedures for RSVP message generation in RFC 2747, *RSVP Cryptographic Authentication*.

An RSVP receiver uses the key together with the authentication algorithm to process received RSVP messages.

When a PLR node switches the path of the LSP to a bypass LSP, it does not send the Integrity object in the RSVP messages over the bypass tunnel. If an integrity object is received from the MP node, then the message is discarded since there is no security association with the next-next-hop MP node.

The MD5 implementation does not support the authentication challenge procedures in RFC 2747.

Reservation Styles

LSPs can be signaled with explicit reservation styles. A reservation style is a set of control options that specify a number of supported parameters. The style information is part of the LSP configuration. SR OS supports two reservation styles:

- Fixed Filter (FF) — The Fixed Filter (FF) reservation style specifies an explicit list of senders and a distinct reservation for each of them. Each sender has a dedicated reservation that is not shared with other senders. Each sender is identified by an IP address and a local identification number, the LSP ID. Because each sender has its own reservation, a unique label and a separate LSP can be constructed for each sender-receiver pair. For traditional RSVP applications, the FF reservation style is ideal for a video distribution application in which each channel (or source) requires a separate pipe for each of the individual video streams.
- Shared Explicit (SE) — The Shared Explicit (SE) reservation style creates a single reservation over a link that is shared by an explicit list of senders. Because each sender is explicitly listed in the RESV message, different labels can be assigned to different sender-receiver pairs, thereby creating separate LSPs.

  Note that if FRR option is enabled for the LSP and selects the facility FRR method at the head-end node, only the SE reservation style is allowed. Furthermore, if a PLR node receives a path message with fast-reroute requested with facility method and the FF reservation style, it will reject the reservation. The one-to-one detour method supports both FF and SE styles.

RSVP Message Pacing

When a flood of signaling messages arrive because of topology changes in the network, signaling messages can be dropped which results in longer set up times for LSPs. RSVP message pacing controls the transmission rate for RSVP messages, allowing the messages to be sent in timed intervals. Pacing reduces the number of dropped messages that can occur from bursts of signaling messages in large networks.

RSVP Overhead Refresh Reduction

The RSVP refresh reduction feature consists of the following capabilities implemented in accordance to RFC 2961, *RSVP Refresh Overhead Reduction Extensions*:

- RSVP message bundling — This capability is intended to reduce overall message handling load. The system supports receipt and processing of bundled message only, but no transmission of bundled messages.
- Reliable message delivery: — This capability consists of sending a message-id and returning a message-ack for each RSVP message. It can be used to detect message loss and support reliable RSVP message delivery on a per hop basis. It also helps reduce the refresh rate since the delivery becomes more reliable.
- Summary refresh — This capability consists of refreshing multiples states with a single message-id list and sending negative ACKs (NACKs) for a message_id which could not be matched. The summary refresh capability reduce the amount of messaging exchanged and the corresponding message processing between peers. It does not however reduce the amount of soft state to be stored in the node.

These capabilities can be enabled on a per-RSVP-interface basis are referred to collectively as “refresh overhead reduction extensions”. When the refresh-reduction is enabled on a system RSVP interface, the node indicates this to its peer by setting a refresh-reduction- capable bit in the flags field of the common RSVP header. If both peers of an RSVP interface set this bit, all the above three capabilities can be used. Furthermore, the node monitors the settings of this bit in received RSVP messages from the peer on the interface. As soon as this bit is cleared, the node stops sending summary refresh messages. If a peer did not set the “refresh-reduction-capable” bit, a node does not attempt to send summary refresh messages.

The RSVP Overhead Refresh Reduction is supported with both RSVP P2P LSP path and the S2L path of an RSVP P2MP LSP instance over the same RSVP interface.

RSVP Graceful Restart Helper

This **gr-helper** command enables the RSVP Graceful Restart Helper feature.

The RSVP-TE Graceful Restart helper mode allows the SR OS based system (the helper node) to provide another router that has requested it (the restarting node) a grace period, during which the system will continue to use RSVP sessions to neighbors requesting the grace period. This is typically used when another router is rebooting its control plane but its forwarding plane is expected to continue to forward traffic based on the previously available Path and Resv states.

The user can enable Graceful Restart helper on each RSVP interface separately. When the GR helper feature is enabled on an RSVP interface, the node starts inserting a new Restart_Cap Object in the Hello packets to its neighbor. The restarting node does the same and indicates to the helper node the desired Restart Time and Recovery Time.

The GR Restart helper consists of a couple of phases. Once it loses Hello communication with its neighbor, the helper node enters the Restart phase. During this phase, it preserves the state of all RSVP sessions to its neighbor and waits for a new Hello message.

Once the Hello message is received indicating the restarting node preserved state, the helper node enters the recovery phase in which it starts refreshing all the sessions that were preserved. The restarting node will activate all the stale sessions that are refreshed by the helper node. Any Path state that did not get a Resv message from the restarting node once the Recovery Phase time is over is considered to have expired and is deleted by the helper node causing the proper Path Tear generation downstream.

The duration of the restart phase (recovery phase) is equal to the minimum of the neighbor's advertised Restart Time (Recovery Time) in its last Hello message and the locally configured value of the max-restart (max-recovery) parameter.

When GR helper is enabled on an RSVP interface, its procedures apply to the state of both P2P and P2MP RSVP LSP to a neighbor over this interface.

Enhancements to RSVP control plane congestion control

The RSVP control plane makes use of a global flow control mechanism to adjust the rate of Path messages for unmapped LSP paths sent to the network under congestion conditions. When a Path message for establishing a new LSP path or retrying an LSP path that failed is sent out, the control plane keeps track of the rate of successful establishment of these paths and adjusts the number of Path messages it sends per second to reflect the success ratio.

In addition, an option to enable an exponential back-off retry-timer is available. When an LSP path establishment attempt fails, the path is put into retry procedures and a new attempt will be performed at the expiry of the user-configurable retry-timer. By default, the retry time is constant. The exponential back-off timer procedures will double the value of the user configurable retry-timer value at every failure of the attempt to adjust to the potential network congestion that caused the failure. An LSP establishment fails if no Resv message was received and the Path message retry-timer expired, or a PathErr message was received before the timer expired.

Three enhancements to this flow-control mechanism to improve congestion handling in the rest of the network are supported.

The first enhancement is the change to the LSP path retry procedure. If the establishment attempt failed due to a Path message timeout and no Resv was received, the next attempt will be performed at the expiry of a new LSP path initial retry-timer instead of the existing retry-timer. While the LSP path initial retry-timer is still running, a refresh of the Path message using the same path and the same LSP-id is performed according to the configuration of the refresh-timer. Once the LSP path initial retry-timer expires, the ingress LER then puts this path on the regular retry-timer to schedule the next path signaling using a new computed path by CSPF and a new LSP-id.

The benefits of this enhancement is that the user can now control how many refreshes of the pending PATH state can be performed before starting a new retry-cycle with a new LSP-id. This is all done without affecting the ability to react faster to failures of the LSP path, which will continue to be governed by the existing retry-timer. By configuring the LSP path initial retry-timer to values that are larger than the retry-timer, the ingress LER will decrease the probability of overwhelming a congested LSR with new state while the previous states installed by the same LSP are lingering and will only be removed after the refresh timeout period expires.

The second enhancement consists of applying a jitter +/- 25% to the value of the retry-timer similar to how it is currently done for the refresh timer. This will further decrease the probability that ingress LER nodes synchronize their sending of Path messages during the retry-procedure in response to a congestion event in the network.

The third enhances the RSVP flow control mechanism by taking into account new parameters: outstanding CSPF requests, Resv timeouts and Path timeouts.

RSVP LSP Statistics

This feature provides the following counters:

- Per forwarding class forwarded in-profile packet count
- Per forwarding class forwarded in-profile byte count
- Per forwarding class forwarded out of profile packet count
- Per forwarding class forwarded out of profile byte count

The counters are available for an RSVP LSP at the egress datapath of an ingress LER and at the ingress datapath of an egress LER. No LSR statistics are provided.

This feature is supported on IOM-2 and IOM-3 and requires chassis mode C or higher.

Configuring Implicit Null

The implicit null label option allows a 7x50 egress LER to receive MPLS packets from the previous hop without the outer LSP label. The operation of the previous hop is referred to as penultimate hop popping (PHP).

This option is signaled by the egress LER to the previous hop during the LSP signaling with RSVP control protocol. In addition, the egress LER can be configured to receive MPLS packet with the implicit null label on a static LSP.

The user can configure your router to signal the implicit null label value over all RSVP interfaces and for all RSVP LSPs for which this node is the egress LER using the **implicit-null-label** command in the **config>router>rsvp** context.

The user must shutdown RSVP before being able to change the implicit null configuration option.

The user can also override the RSVP level configuration for a specific RSVP interface:

config>router>rsvp>interface>implicit-null-label {enable | disable}

All LSPs for which this node is the egress LER and for which the path message is received from the previous hop node over this RSVP interface will signal the implicit null label. This means that if the egress LER is also the merge-point (MP) node, then the incoming interface for the path refresh message over the bypass dictates if the packet will use the implicit null label or not. The same for a 1-to-1 detour LSP.

By default, an RSVP interface inherits the RSVP level configuration. The user must shutdown the RSVP interface before being able to change the implicit null configuration option. Note that the RSVP interface must be shutdown regardless if the new value for the interface is the same or different than the one it is currently using.

The egress LER does not signal the implicit null label value on P2MP RSVP LSPs. However, the PHP node can honor a Resv message with the label value set to the implicit null value when the egress LER is a third party implementation.

The implicit null label option is also supported on a static label LSP. The following commands can be used to cause the node to push or to swap to an implicit null label on the MPLS packet:

config>router>mpls>static-lsp>push implicit-null-label nexthop *ip-address*

config>router>mpls>interface>label-map>swap implicit-null-label nexthop *ip-address*

Using Unnumbered Point-to-Point Interface in RSVP

This feature introduces the use of unnumbered IP interface as a Traffic Engineering (TE) link for the signaling of RSVP P2P LSP and P2MP LSP.

An unnumbered IP interface is identified uniquely on a router in the network by the tuple {router-id, ifIndex}. Each side of the link assigns a system-wide unique interface index to the unnumbered interface. ISIS, OSPF, RSVP, and OAM modules will use this tuple to advertise the link information, signal LSP paths over this unnumbered interface, or send and respond to an MPLS echo request message over an unnumbered interface.

The interface borrowed IP address is used exclusively as the source address for IP packets that are originated from the interface and needs to be configured to an address different from system interface for the FRR bypass LSP to come up at the ingress LER.

The borrowed IP address for an unnumbered interface is configured using the following CLI command with a default value set to the system interface address:

configure> router>interface>unnumbered [*ip-int-name* | *ip-address*].

The support of unnumbered TE link in IS-IS consists of adding a new sub-TLV of the extended IS reachability TLV, which encodes the Link Local and Link Remote Identifiers as defined in RFC 5307.

The support of unnumbered TE link in OSPF consists of adding a new sub-TLV, which encodes the same Link Local and Link Remote Identifiers in the Link TLV of the TE area opaque LSA and sends the local Identifier in the Link Local Identifier TLV in the TE link local opaque LSA as per RFC 4203.

The support of unnumbered TE link in RSVP implements the signaling of unnumbered interfaces in ERO/RRO as per RFC 3477 and the support of IF_ID RSVP_HOP object with a new Ctype as per Section 8.1.1 of RFC 3473. The IPv4 Next/Previous Hop Address field is set to the borrowed IP interface address.

The unnumbered IP is advertised by IS-IS TE and OSPF TE, and CSPF can include them in the computation of a path for a P2P LSP or for the S2L of a P2MP LSP. This feature does not, however, support defining an unnumbered interface a hop in the path definition of an LSP.

A router creates an RSVP neighbor over an unnumbered interface using the tuple {router-id, ifIndex}. The router-id of the router that advertised a given unnumbered interface index is obtained from the TE database. As as result, if traffic engineering is disabled in IS-IS or OSPF, a non-CSPF LSP with the next-hop for its path is over an unnumbered interface will not come up at the ingress LER since the router-id of the neighbor that has the next-hop of the path message cannot be looked up. In this case, the LSP path will remain in operationally down state with a reason ‘noRouteToDestination’. If a PATH message was received at the LSR in which traffic engineering was disabled and the next-hop for the LSP path is over an unnumbered interface, a

PathErr message will be sent back to the ingress LER with the "Routing Problem" error code of 24 and an error value of 5 “No route available toward destination”.

All MPLS features available for numbered IP interfaces are supported, with the exception of the following:

- Configuring a router-id with a value other than system.
- Signaling of an LSP path with an ERO based a loose/strict hop using an unnumbered TE link in the path hop definition.
- Signaling of one-to-one detour LSP over unnumbered interface.
- Soft pre-emption of LSP path using unnumbered interface.
- Inter-area LSP.
- Unnumbered RSVP interface registration with BFD.
- RSVP Hello and all Hello related capabilities such as Graceful-restart helper.
- RSVP refresh reduction on an unnumbered interface
- The user SRLG database feature. The user-srlg-db option under MPLS allows the user to manually enter the SRLG membership of any link in the network in a local database at the ingress LER. The user cannot enter an unnumbered interface into this database and as such, all unnumbered interfaces will be considered as having no SRLG membership if the user enabled the user-srlg-db option.

This feature also extends the support of lsp-ping, p2mp-lsp-ping, lsp-trace, and p2mp-lsptrace to P2P and P2MP LSPs that have unnumbered TE links in their path.

Operation of RSVP FRR Facility Backup over Unnumbered Interface

When the Point-of-Local Repair (PLR) node activates the bypass LSP by sending a PATH message to refresh the path state of protected LSP at the Merge-Point (MP) node, it must use an "IPv4 tunnel sender address" in the sender template object that is different than the one used by the ingress LER in the PATH message. These are the procedures specified in RFC 4090 and that are followed in the 7x50 implementation.

The 7x50 uses the address of the outgoing interface of the bypass LSP as the "IPv4 tunnel sender address" in the sender template object. This address will be different from the system interface address used in the sender template of the protected LSP by the ingress LER and thus there are no conflicts when the ingress LER acts as a PLR.

When the PLR is the ingress LER node and the outgoing interface of the bypass LSP is unnumbered, it is required that the user assigns to the interface a borrowed IP address that is different from the system interface. If not, the bypass LSP will not come up.

In addition, the PLR node will include the IPv4 RSVP_HOP object (C-Type=1) or the IF_ID RSVP_HOP object (C-Type=3) in the PATH message if the outgoing interface of the bypass LSP is numbered or unnumbered respectively.

When the MP node receives the PATH message over the bypass LSP, it will create the merge-point context for the protected LSP and associate it with the existing state if any of the following is satisfied:

- Change in C-Type of the RSVP_HOP object, or
- C-Type is IF_ID RSVP_HOP and did not change but IF_ID TLV is different, or
- Change in IPv4 Next/Previous Hop Address in RSVP_HOP object regardless of the C-Type value.

These procedures at PLR and MP nodes are followed in both link-protect and node-protect FRR. Note that if the MP node is running a pre-R11 implementation, it will reject the new IF_ID C-Type and will drop the PATH over bypass. This will result in the protected LSP state expiring at the MP node, which will tear down the path. This will be the case in general when node-protect FRR is enabled and the MP node does not support unnumbered RSVP interface.

Traffic Engineering

Without traffic engineering, routers route traffic according to the SPF algorithm, disregarding congestion or packet types.

With traffic engineering, network traffic is routed efficiently to maximize throughput and minimize delay. Traffic engineering facilitates traffic flows to be mapped to the destination through a different (less congested) path other than the one selected by the SPF algorithm.

MPLS directs a flow of IP packets along a label switched path (LSP). LSPs are simplex, meaning that the traffic flows in one direction (unidirectional) from an ingress router to an egress router. Two LSPs are required for duplex traffic. Each LSP carries traffic in a specific direction, forwarding packets from one router to the next across the MPLS domain.

When an ingress router receives a packet, it adds an MPLS header to the packet and forwards it to the next hop in the LSP. The labeled packet is forwarded along the LSP path until it reaches the destination point. The MPLS header is removed and the packet is forwarded based on Layer 3 information such as the IP destination address. The physical path of the LSP is not constrained to the shortest path that the IGP would choose to reach the destination IP address.

TE Metric (IS-IS and OSPF)

When the use of the TE metric is selected for an LSP, the shortest path computation after the TE constraints are applied will select an LSP path based on the TE metric instead of the IGP metric. The user configures the TE metric under the MPLS interface. Both the TE and IGP metrics are advertised by OSPF and IS-IS for each link in the network. The TE metric is part of the traffic engineering extensions of both IGP protocols.

A typical application of the TE metric is to allow CSPF to represent a dual TE topology for the purpose of computing LSP paths.

An LSP dedicated for real-time and delay sensitive user and control traffic has its path computed by CSPF using the TE metric. The user configures the TE metric to represent the delay figure, or a combined delay/jitter figure, of the link. In this case, the shortest path satisfying the constraints of the LSP path will effectively represent the shortest delay path.

An LSP dedicated for non delay sensitive user and control traffic has its path computed by CSPF using the IGP metric. The IGP metric could represent the link bandwidth or some other figure as required.

When the use of the TE metric is enabled for an LSP, CSPF will first prune all links in the network topology that do not meet the constraints specified for the LSP path. These constraints include bandwidth, admin-groups, and hop limit. CSPF will then run an SPF on the remaining links. The shortest path among the all SPF paths will be selected based on the TE metric instead of the IGP metric which is used by default. Note that the TE metric is only used in CSPF computations for MPLS paths and not in the regular SPF computation for IP reachability.

Admin Group Support on Facility Bypass Backup LSP

This feature provides for the inclusion of the LSP primary path admin-group constraints in the computation of a Fast ReRoute (FRR) facility bypass backup LSP to protect the primary LSP path by all nodes in the LSP path.

This feature is supported with the following LSP types and in both intra-area and inter-area TE where applicable:

- Primary path of a RSVP P2P LSP.
- S2L path of an RSVP P2MP LSP instance
- LSP template for an S2L path of an RSVP P2MP LSP instance .

Procedures at Head-End Node

The user enables the signaling of the primary LSP path admin-group constraints in the FRR object at the ingress LER with the following CLI command:

configure>router>mpls>lsp>fast-reroute>propagate-admin-group

When this command is enabled at the inress LER, the admin-group constraints configured in the context of the P2P LSP primary path, or the ones configured in the context of the LSP and inherited by the primary path, are copied into the FAST_REROUTE object. The admin-group constraints are copied into the ‘include-any’ or ‘exclude-any’ fields.

The ingress LER thus propagates these constraints to the downstream nodes during the signaling of the LSP to allow them to include the admin-group constraints in the selection of the FRR backup LSP for protecting the LSP primary path.

The ingress LER will insert the FAST_REROUTE object by default in a primary LSP path message. If the user disables the object using the following command, the admin-group constraints will not be propagated: **configure>router>mpls>no frr-object**.

Note that the same admin-group constraints can be copied into the Session Attribute object. They are intended for the use of an LSR, typically an ABR, to expand the ERO of an inter-area LSP path. They are also used by any LSR node in the path of a CSPF or non-CSPF LSP to check the admin-group constraints against the ERO regardless if the hop is strict or loose. These are governed strictly by the command:

configure>router>mpls>lsp>propagate-admin-group

In other words, the user may decide to copy the primary path admin-group constraints into the FAST_REROUTE object only, or into the Session Attribute object only, or into both.

Note however, that the PLR rules for processing the admin-group constraints can make use of either of the two object admin-group constraints.

Procedures at PLR Node

The user enables the use of the admin-group constraints in the association of a manual or dynamic bypass LSP with the primary LSP path at a Point-of-Local Repair (PLR) node using the following global command:

configure>router>mpls>admin-group-frr

When this command is enabled, each PLR node reads the admin-group constraints in the FAST_REROUTE object in the Path message of the LSP primary path. If the FAST_REROUTE object is not included in the Path message, then the PLR will read the admin-group constraints from the Session Attribute object in the Path message.

If the PLR is also the ingress LER for the LSP primary path, then it just uses the admin-group constraint from the LSP and/or path level configurations.

Whether the PLR node is also the ingress LER or just an LSR for the protected LSP primary path, the outcome of the ingress LER configuration dictates the behavior of the PLR node and is summarized in Table 4.

Table 4: Bypass LSP Admin-Group Constraint Behavior

	Ingress LER Configuration	Session Attribute	FRR Object	Bypass LSP at PLR (LER/LSF) follows admin-group constraints
1	frr-object lsp>no propagate-admin group lsp>frr>propagate-admin-group	Admin color constraints not sent	Admin color constraints sent	yes
2	frr-object lsp>propagate-admin-group lsp>frr>propagate-admin group	Admin color constraints sent	Admin color constraints sent	yes

Table 4: Bypass LSP Admin-Group Constraint Behavior

3	frr-object lsp>propagate-admin group lsp>frr>no propagate-admin-group	Admin color constraints sent	Admin color constraints not sent	no
4	frr-object lsp>propagate-admin group lsp>frr>no propagate-admin-group	Admin color constraints sent	Not present	yes
5	No frr-object lsp>no propagate-admin group lsp>frr>propagate-admin-group	Admin color constraints not sent	Not present	no
6	No frr-object lsp>propagate-admin group lsp>frr>no propagate-admin-group	Admin color constraints sent	Not present	yes

The PLR node then uses the admin-group constraints along with other constraints, such as hop-limit and SRLG, to select a manual or dynamic bypass among those that are already in use.

If none of the manual or dynamic bypass LSP satisfies the admin-group constraints, and/or the other constraints, the PLR node will request CSPF for a path that merges the closest to the protected link or node and that includes or excludes the specified admin-group IDs.

If the user changes the configuration of the above command, it will not have any effect on existing bypass associations. The change will only apply to new attempts to find a valid bypass.

Diff-Serv Traffic Engineering

Diff-Serv traffic engineering provides the ability to manage bandwidth on a per Traffic Engineering (TE) class basis as per RFC 4124. In the base traffic engineering, LER computes LSP paths based on available BW of links on the path. Diff-Serv TE adds ability to perform this on a per TE class basis.

A TE class is a combination of Class Type and LSP priority. A Class Type is mapped to one or more system Forwarding Classes using a configuration profile. The operator sets different limits for admission control of LSPs in each TE class over each TE link. Eight TE classes are supported. Admission control of LSP paths bandwidth reservation is performed using the Maximum Allocation Bandwidth Constraint Model as per RFC 4125.

Mapping of Traffic to a Diff-Serv LSP

An LER will allow the operator to map traffic to a Diff-Serv LSP through one of the following methods:

1. Explicit RSVP SDP configuration of a VLL, VPLS, or VPRN service.
2. Class-based forwarding in an RSVP SDP. The operator can enable the checking by RSVP that a Forwarding Class (FC) mapping to an LSP under the SDP configuration is compatible with the Diff-Serv Class Type (CT) configuration for this LSP.
3. Auto-bind RSVP-TE option in a VPRN service.
4. Static routes with indirect next-hop being an RSVP LSP name.

Admission Control of Classes

There are a couple of admission control decisions made when an LSP with a specified bandwidth is to be signaled. The first is in the head-end node. CSPF will only consider network links that have sufficient bandwidth. Link bandwidth information is provided by IGP TE advertisement by all nodes in that network.

Another decision made is local CAC and is performed when the RESV message for the LSP path is received in the reverse direction by a SR OS node in that path. The bandwidth value selected by the egress LER will be checked against link bandwidth, otherwise the reservation is rejected. If accepted, the new value for the remaining link bandwidth will be advertised by IGP at the next advertisement event.

Both of these admission decisions are enhanced to be performed at the TE class level when Diff-Serv TE is enabled. In other words, CSPF in the head-end node will need to check the LSP bandwidth against the 'unreserved bandwidth' advertised for all links in the path of the LSP for that TE class which consists of a combination of a CT and a priority. Same for the admission control at SR OS node receiving the Resv message.

Maximum Allocation Model

The admission control rules for this model are described in RFC 4125. Each CT shares a percentage of the Maximum Reservable Link Bandwidth through the user-configured BC for this CT. The Maximum Reservable Link Bandwidth is the link bandwidth multiplied by the RSVP interface subscription factor.

The sum of all BC values across all CTs will not exceed the Maximum Reservable Link Bandwidth. In other words, the following rule is enforced:

SUM (BCc) =< Max-Reservable-Bandwidth, 0 <= c <= 7

An LSP of class-type CTc, setup priority p, holding priority h (h=<p), and bandwidth B is admitted into a link if the following condition is satisfied:

B <= Unreserved Bandwidth for TE-Class[i]

where TE-Class [i] maps to < CTc, p > in the definition of the TE classes on the node. The bandwidth reservation is effected at the holding priority, i.e., in TE-class [j] = <CTc, h>. Thus, the reserved bandwidth for CTc and the unreserved bandwidth for the TE classes using CTc are updated as follows:

Reserved(CTc) = Reserved(CTc) + B

Unreserved TE-Class [j] = BCc - SUM (Reserved(CTc,q)) for 0<= q <= h

Unreserved TE-Class [i] = BCc - SUM (Reserved(CTc,q)) for 0<= q <= p

The same is done to update the unreserved bandwidth for any other TE class making use of the same CTc. These new values are advertised to the rest of the network at the next IGP-TE flooding.

When Diff-Serv is disabled on the node, this model degenerates into a single default CT internally with eight pre-emption priorities and a non-configurable BC equal to the Maximum Reservable Link Bandwidth. This would behave exactly like CT0 with eight pre-emption priorities and BC= Maximum Reservable Link Bandwidth if Diff-Serv was enabled.

Russian Doll Model

The RDM model is defined using the following equations:

SUM (Reserved (CTc)) <= BCb,

where the SUM is across all values of **c** in the range **b <= c <= (MaxCT - 1)**, and **BCb** is the bandwidth constraint of **CTb**.

BC0= Max-Reservable-Bandwidth, so that:

SUM (Reserved(CTc)) <= Max-Reservable-Bandwidth,

where the **SUM** is across all values of **c** in the range **0 <= c <= (MaxCT - 1)**

An LSP of class-type **CTc**, setup priority **p**, holding priority **h (h=<p)**, and bandwidth **B** is admitted into a link if the following condition is satisfied:

B <= Unreserved Bandwidth for TE-Class[i],

where **TE-Class [i]** maps to < **CTc , p** > in the definition of the TE classes on the node. The bandwidth reservation is effected at the holding priority, i.e., in **TE-class [j] = <CTc, h>**. Thus, the reserved bandwidth for CTc and the unreserved bandwidth for the TE classes using CTc are updated as follows:

Reserved(CTc) = Reserved(CTc) + B

Unreserved TE-Class [j] = Unreserved (CTc, h) = Min [
BCc - SUM (Reserved (CTb, q) for 0<=q <= h, c <= b <= 7,
BC(c-1) – SUM (Reserved (CTb, q) for 0<=q <= h, (c-1) <= b <= 7,
…….
BC0 - SUM (Reserved (CTb, q) for 0<=q <= h, 0 <= b <= 7]

Unreserved TE-Class [i] = Unreserved (CTc, p) = Min [
BCc - SUM (Reserved (CTb, q) for 0<=q <= p, c <= b <= 7,
BC(c-1) – SUM (Reserved (CTb, q) for 0<=q <= p, (c-1) <= b <= 7,
…….
BC0 - SUM (Reserved (CTb, q) for 0<=q <= p, 0 <= b <= 7]

The same is done to update the unreserved bandwidth for any other TE class making use of the same CTc. These new values are advertised to the rest of the network at the next IGP-TE flooding.

Example CT Bandwidth Sharing with RDM

Below is a simple example with two CT values (CT0, CT1) and one priority 0 as shown in Figure 23.

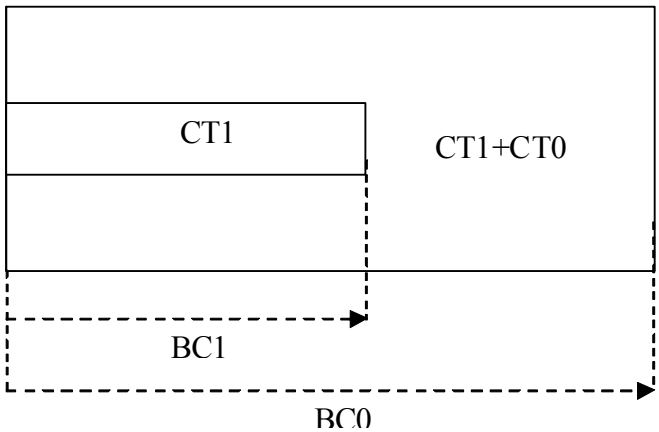


Figure 23: RDM with Two Class Types

Suppose CT1 bandwidth, or the CT1 percentage of Maximum Reservable Bandwidth to be more accurate is 100 Mbps and CT2 bandwidth is 100 Mbps and link bandwidth is 200 Mbps. BC constraints can be calculated as follows.

BC1 = CT1 Bandwidth = 100 Mbps.

BC0 = {CT1 Bandwidth} + {CT0 Bandwidth} = 200 Mbps.

Suppose an LSP comes with CT1, setup and holding priorities of 0 and a bandwidth of 50 Mbps.

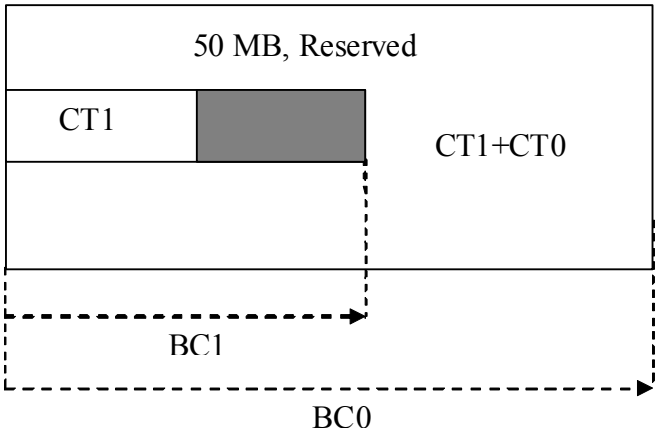


Figure 24: First LSP Reservation

According to the RDM admission control policy:

Reserved (CT1, 0) = 50 <= 100 Mbps

Reserved (CT0, 0) + Reserved (CT1, 0) = 50 <= 200 Mbps

This results in the following unreserved bandwidth calculation.

Unreserved (CT1, 0) = BC1 – Reserved (CT1, 0) = 100 – 50 = 50 Mbps

Unreserved (CT0, 0) = BC0 – Reserved (CT0, 0) – Reserved (CT1, 0) = 200 – 0 – 50= 150 Mbps.

Note that bandwidth reserved by a doll is not available to itself as well any of the outer dolls.

Suppose now another LSP comes with CT0, setup and holding priorities of 0 and a bandwidth 120 Mbps.

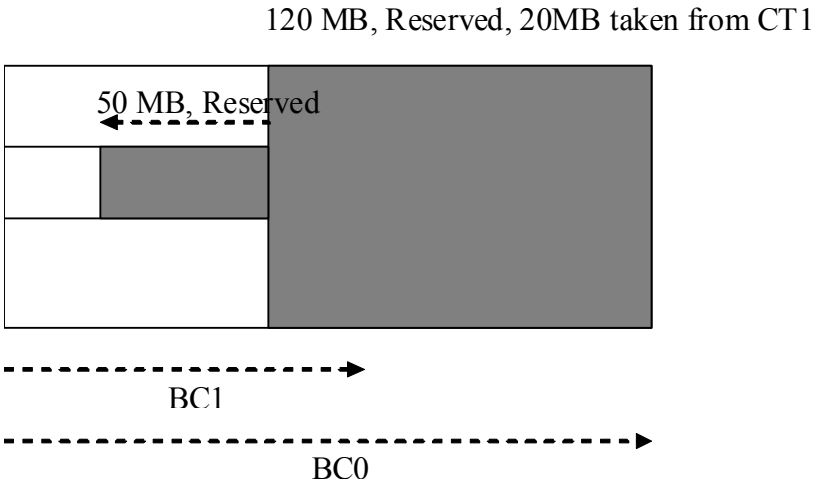


Figure 25: Second LSP Reservation

Reserved (CT0, 0) = 120 <= 150 Mbps

Reserved (CT0, 0) + Reserved (CT1, 0) = 120 + 50 = 170 <= 200 Mbps

Unreserved (CT0, 0) = 150 -120 = 30 Mbps

If we simply checked BC1, the formula would yield the wrong results:

Unreserved (CT1, 0) = BC1 – Reserved (CT1, 0) = 100 -50 = 50 Mbps

Because of the encroaching of CT0 into CT1, we would need to deduct the overlapping reservation. This would then yield:

Unreserved (CT1, 0) = BC0 – Reserved (CT0, 0) – Reserved (CT1, 0) = 200 – 120 - 50 = 30 Mbps,

which is the correct figure.

Extending the formula with both equations:

Unreserved (CT1, 0) = Min [BC1 – Reserved (CT1, 0), BC0 – Reserved (CT0, 0) – Reserved (CT1, 0)] = Min [100 – 50, 200 – 120 – 50] = 30 Mbps

Note that an outer doll can encroach into inner doll reducing the bandwidth available for inner dolls.

RSVP Control Plane Extensions

RSVP will use the Class Type object to carry LSP class-type information during path setup. Eight values will be supported for class-types 0 through 7 as per RFC 4124. Class type 0 is the default class which is supported today on the router.

One or more forwarding classes will map to a Diff-Serv class type trough a system level configuration.

IGP Extensions

IGP extensions are defined in RFC 4124. Diff-Serv TE advertises link available bandwidth, referred to as unreserved bandwidth, by OSPF TE or IS-IS TE on a per TE class basis. A TE class is a combination of a class type and an LSP priority. In order to reduce the amount of per TE class flooding required in the network, the number of TE classes is set to eight. This means that eight class types can be supported with a single priority or four class types with two priorities, etc. In that case, the operator configures the desired class type on the LSP such that RSVP-TE can signal it in the class-type object in the path message.

IGP will continue to advertise the existing Maximum Reservable Link Bandwidth TE parameter to mean the maximum bandwidth that can be booked on a given interface by all classes. The value advertised is adjusted with the link subscription factor.

Diff-Serv TE Configuration and Operation

RSVP Protocol Level

The following are the configuration steps at the RSVP protocol level:

1. The operator enables Diff-Serv TE by executing the **diffserv-te** command in the **config>router>rsvp** context. When this command is enabled, IS-IS and OSPF will start advertising available bandwidth for each TE class configured under the **diffserv-te** node. The operator can disable Diff-Serv TE globally by using the no form of the command.

2. The enabling or disabling of Diff-Serv on the system requires that the RSVP and MPLS protocol be shutdown. The operator must execute the **no shutdown** command in each context once all parameters under both protocols are defined. When saved in the configuration file, the **no shutdown** command is automatically inserted under both protocols to make sure they come up after a node reboot.

3. IGP will advertise the available bandwidth in each TE class in the unreserved bandwidth TE parameter for that class for each RSVP interface in the system.

4. In addition, IGP will continue to advertise the existing Maximum Reservable Link Bandwidth TE parameter so the maximum bandwidth that can be booked on a given interface by all classes. The value advertised is adjusted with the link subscription factor configured in the **config>router>rsvp>interface>subscription** *percentage* context.

5. The operator can overbook (underbook) the maximum reservable bandwidth of a given CT by overbooking (underbooking) the interface maximum reservable bandwidth by configuring the appropriate value for the **subscription** *percentage* parameter.

6. The **diffserv-te** command will only have effect if the operator has already enabled traffic engineering at the IS-IS and/or OSPF routing protocol levels:
   config>router>isis>traffic-engineering
   and/or:
   config>router>ospf>traffic-engineering

7. The following Diff-Serv TE parameters are configured globally under the **diffserv-te** node. They apply to all RSVP interfaces on the system. Once configured, these parameters can only be changed after shutting down the MPLS and RSVP protocols:

   a. Definition of TE classes, TE Class = {Class Type (CT), LSP priority}. Eight TE classes can be supported. There is no default TE class once Diff-Serv is enabled. The operator must explicitly define each TE class. However, when Diff-Serv is disabled there will be an internal use of the default CT (CT0) and eight pre-emption priorities as shown in Table 5.

Table 5: Internal TE Class Definition when Diff-Serv TE is Disabled

Class Type (CT internal)	LSP Priority
0	7
0	6
0	5
0	4
0	3
0	2
0	1
0	0

b. A mapping of the system forwarding class to CT. The default settings are shown in Table 6.

Table 6: Default Mapping of Forwarding Class to TE Class

FC ID	FC Name	FC Designation	Class Type (CT)
7	Network Control	NC	7
6	High-1	H1	6
5	Expedited	EF	5
4	High-2	H2	4
3	Low-1	L1	3
2	Assured	AF	2
1	Low-2	L2	1
0	Best Effort	BE	0

c. Configuration of the percentage of RSVP interface bandwidth each CT shares, for example, the Bandwidth Constraint (BC), using the **class-type-bw** command. The absolute value of the CT share of the interface bandwidth is derived as the percentage of the bandwidth advertised by IGP in the maximum reservable link bandwidth TE parameter, for example, the link bandwidth multiplied by the RSVP interface **subscription** *percentage* parameter. Note that this configuration also exists at the RSVP

interface level and the interface specific configured value overrides the global configured value. The BC value can be changed at any time. The operator can specify the BC for a CT which is not used in any of the TE class definition but that does not get used by any LSP originating or transiting this node.

d. Configuration of the Admission Control Policy to be used: only the Maximum Allocation Model (MAM) is supported. The MAM value represents the bandwidth constraint models for the admission control of an LSP reservation to a link.

RSVP Interface Level

The following are the configuration steps at the RSVP interface level:

1. The operator configures the percentage of RSVP interface bandwidth each CT shares, for example, the BC, using the **class-type-bw** command. The value entered at the interface level overrides the global value configured under the **diffserv-te** node.
2. The operator can overbook (underbook) the maximum reservable bandwidth of a given CT by overbooking (underbooking) the interface maximum reservable bandwidth via configuring the appropriate value for the **subscription** *percentage* parameter in the **config>router>rsvp>interface** context.
3. .Both the BC value and the subscription parameter can be changed at any time.

LSP and LSP Path Levels

The following are the configuration steps at the LSP and LSP path levels:

1. The operator configures the CT in which the LSP belongs by configuring the **class-type** *ct-number* command at the LSP level and/or the path level. The path level value overrides the LSP level value. By default, an LSP belongs to CT0.
2. Only one CT per LSP path is allowed per RFC 4124, *Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering*. A multi-class LSP path is achieved through mapping multiple system Forwarding Classes to a CT.
3. The signaled CT of a dynamic bypass must always be CT0 regardless of the CT of the primary LSP path. The setup and hold priorities must be set to default values, for example, 7 and 0 respectively. This assumes that the operator configured a couple of TE classes, one which combines CT0 and a priority of 7 and the other which combines CTO and a priority of 0. If not, the bypass LSP will not be signaled and will go into the down state.
4. The operator cannot configure the CT, setup priority, and holding priority of a manual bypass. They are always signaled with CT0 and the default setup and holding priorities.

5. The signaled CT, setup priority and holding priority of a detour LSP matches those of the primary LSP path it is associated with.
6. The operator can also configure the setup and holding priorities for each LSP path.
7. An LSP which does not have the CT explicitly configured will behave like a CT0 LSP when Diff-Serv is enabled.

If the operator configured a combination of a CT and a setup priority and/or a combination of a CT and a holding priority for an LSP path that are not supported by the user-defined TE classes, the LSP path will be kept in a down state and error code will be shown within the show command output for the LSP path.

Diff-Serv TE LSP Class Type Change under Failure

An option to configure a main Class Type (CT) and a backup CT for the primary path of a Diff-Serv TE LSP is provided. The main CT is used under normal operating conditions, for example, when the LSP is established the first time and when it gets re-optimized due to timer based or manual re-signal. The backup CT is used when the LSP retries under failure.

The use of backup Class Type (CT) by an LSP is enabled by executing the **config>router>mpls>lsp>primary>backup-class-type** *ct-number* command at the LSP primary path level.

When this option is enabled, the LSP will use the CT configured using the following commands (whichever is inherited at the primary path level) as the main CT:

- **config>router>mpls>lsp>class-type** *ct-number*
- **config>router>mpls>lsp>primary>class-type** *ct-number*

The main CT is used at initial establishment and during a manual or a timer based re-signal Make-Before-Break (MBB) of the LSP primary path. The backup CT is used temporarily to signal the LSP primary path when it fails and goes into retry.

Note that any valid values may be entered for the backup CT and main CT, but they cannot be the same. No check is performed to make sure that the backup CT is a lower CT in Diff-Serv Russian-Doll Model (RDM) admission control context.

The secondary paths of the same LSP are always signaled using the main CT as in existing implementation.

LSP Primary Path Retry Procedures

This feature behaves according to the following procedures.

- When a LSP primary path retries due a failure, for example, it fails after being in the up state, or undergoes any type of MBB, MPLS will retry a new path for the LSP using the main CT. If the first attempt failed, the head-end node performs subsequent retries using the backup CT. This procedure must be followed regardless if the currently used CT by this path is the main or backup CT. This applies to both CSPF and non-CSPF LSPs.
- The triggers for using the backup CT after the first retry attempt are:
  - A local interface failure or a control plane failure (hello timeout, etc.).
  - Receipt of a PathErr message with a notification of a FRR protection becoming active downstream and/or receipt of a Resv message with a ‘Local-Protection-In-Use’ flag set. This invokes the FRR Global Revertive MBB.

- Receipt of a PathErr message with error code=25 (Notify) and sub-code=7 (Local link maintenance required) or a sub-code=8 (Local node maintenance required). This invokes the TE Graceful Shutdown MBB. Note that in this case, only a single attempt is performed by MBB as in current implementation; only the main CT will be retried.
- Receipt of a Resv refresh message with the 'Preemption pending' flag set or a PathErr message with error code=34 (Reroute) and a value=1 (Reroute request soft preemption). This invokes the soft pre-emption MBB.
- Receipt of a ResvTear message.
- A configuration change MBB.

- When an unmapped LSP primary path goes into retry, it uses the main CT until the number of retries reaches the value of the new main-ct-retry-limit parameter. If the path did not come up, it must start using the backup CT at that point in time. By default, this parameter is set to infinite value. The new main-ct-retry-limit parameter has no effect on an LSP primary path, which retries due to a failure event. This parameter is configured using the **main-ct-retry-limit** command in the **config>router>mpls>lsp** context. If the user entered a value of the **main-ct-retry-limit** parameter that is greater than the LSP retry-limit, the number of retries will still stop when the LSP primary path reaches the value of the LSP retry-limit. In other words, the meaning of the LSP retry-limit parameter is not changed and always represents the upper bound on the number of retries. The unmapped LSP primary path behavior applies to both CSPF and non-CSPF LSPs.
- An unmapped LSP primary path is a path that never received a Resv in response to the first path message sent. This can occur when performing a "shut/no-shut" on the LSP or LSP primary path or when the node reboots. An unmapped LSP primary path goes into retry if the retry timer expired or the head-end node received a PathErr message before the retry timer expired.
- When the **clear>router>mpls>lsp** command is executed, the retry behavior for this LSP is the same as in the case of an unmapped LSP.
- If the value of the parameter main-ct-retry-limit is changed, the new value will only be used at the next time the LSP path is put into a "no-shut" state.
- The following is the behavior when the user changes the main or backup CT:
  - If the user changes the LSP level CT, all paths of the LSP are torn down and re-signaled in a break-before-make fashion. Specifically, the LSP primary path will be torn down and re-signaled even if it is currently using the backup CT.
  - If the user changes the main CT of the LSP primary path, the path will be torn down and re-signaled even if it is currently using the backup CT.
  - If the user changes the backup CT of an LSP primary path when the backup CT is in use, the path is torn down and is re-signaled.
  - If the user changes the backup CT of an LSP primary path when the backup CT is not in use, no action is taken. If however, the path was in global Revertive, gshut, or soft pre-emption MBB, the MBB is restarted. This actually means the first attempt will be with the main CT and subsequent ones, if any, with the new value of the backup CT.

- Consider the following priority of the various MBB types form highest to lowest: Delayed Retry, Preemption, Global Revertive, Configuration Change, and TE Graceful Shutdown. If an MBB request occurs while a higher priority MBB is in progress, the latter MBB will be restarted. This actually means the first attempt will be with the main CT and subsequent ones, if any, with the new value of the backup CT.

- If the least-fill option is enabled at the LSP level, then CSPF must use least-fill equal cost path selection when the main or backup CT is used on the primary path.
- When the re-signal timer expires, CSPF will try to find a path with the main CT. The head-end node must re-signal the LSP even if the new path found by CSPF is identical to the existing one since the idea is to restore the main CT for the primary path. If a path with main CT is not found, the LSP remains on its current primary path using the backup CT. This means that the LSP primary path with the backup CT may no longer be the most optimal one. Furthermore, if the least-fill option was enabled at the LSP level, CSPF will not check if there is a more optimal path, with the backup CT, according to the least-fill criterion and will thus raise no trap to indicate the LSP path is eligible for least-fill re-optimization.
- When the user performs a manual re-signal of the primary path, CSPF will try to find a path with the main CT. The head-end node must re-signal the LSP as in current implementation.
- If a CPM switchover occurs while an the LSP primary path was in retry using the main or backup CT, for example, was still in operationally down state, the path retry will be restarted with the main CT until it comes up. This is because the LSP path retry count is not synchronized between the active and standby CPMs until the path becomes up.
- When the user configured secondary standby and non-standby paths on the same LSP, the switchover behavior between primary and secondary is the same as in existing implementation.

This feature is not supported on a P2MP LSP.

Bandwidth Sharing Across Class Types

In order to allow different levels of booking of network links under normal operating conditions and under failure conditions, it is necessary to allow sharing of bandwidth across class types.

This feature introduces the Russian-Doll Model (RDM) Diff-Serv TE admission control policy described in RFC 4127, *Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering*. This mode is enabled using the following command: **config>router>rsvp>diffserv-te rdm**.

The Russian Doll Model (RDM) LSP admission control policy allows bandwidth sharing across Class Types (CTs). It provides a hierarchical model by which the reserved bandwidth of a CT is the sum of the reserved bandwidths of the numerically equal and higher CTs.

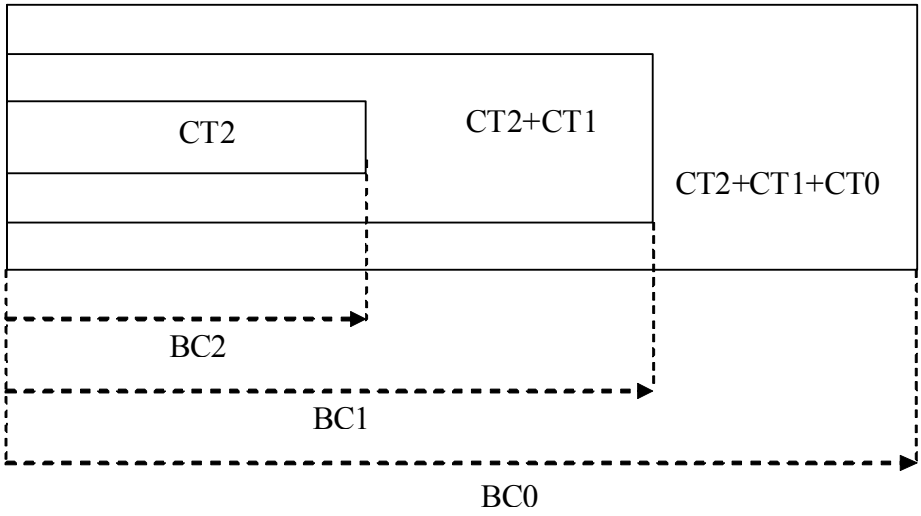


Figure 26: RDM Admission Control Policy Example

CT2 has a bandwidth constraint BC2 which represents a percentage of the maximum reservable link bandwidth. Both CT2 and CT1 can share BC1 which is the sum of the percentage of the maximum reservable bandwidth values configured for CT2 and CT1 respectively. Finally, CT2, CT1, and CT0 together can share BC0 which is the sum of the percentage of the maximum reservable bandwidth values configured for CT2, CT1, and CT0 respectively. The maximum value for BC0 is of course the maximum reservable link bandwidth.

What this means in practice is that CT0 LSPs can use up to BC0 in the absence of LSPs in CT1 and CT2. When this occurs and a CT2 LSP with a reservation less than or equal to BC2 requests admission, it is only admitted by preempting one or more CT0 LSPs of lower holding priority than this LSP setup priority. Otherwise, the reservation request for the CT2 LSP will be rejected.

It is required that multiple paths of the same LSP share common link bandwidth since they are signaled using the Shared Explicit (SE) style. Specifically, two instances of a primary path, one with the main CT and the other with the backup CT, must temporarily share bandwidth while MBB is in progress. Also, a primary path and one or many secondary paths of the same LSP must share bandwidth whether they are configured with the same or different CTs.

Downgrading the CT of Bandwidth Sharing LSP Paths

Consider a link configured with two class types CT0 and CT1 and making use of the RDM admission control model as shown in Figure 27.

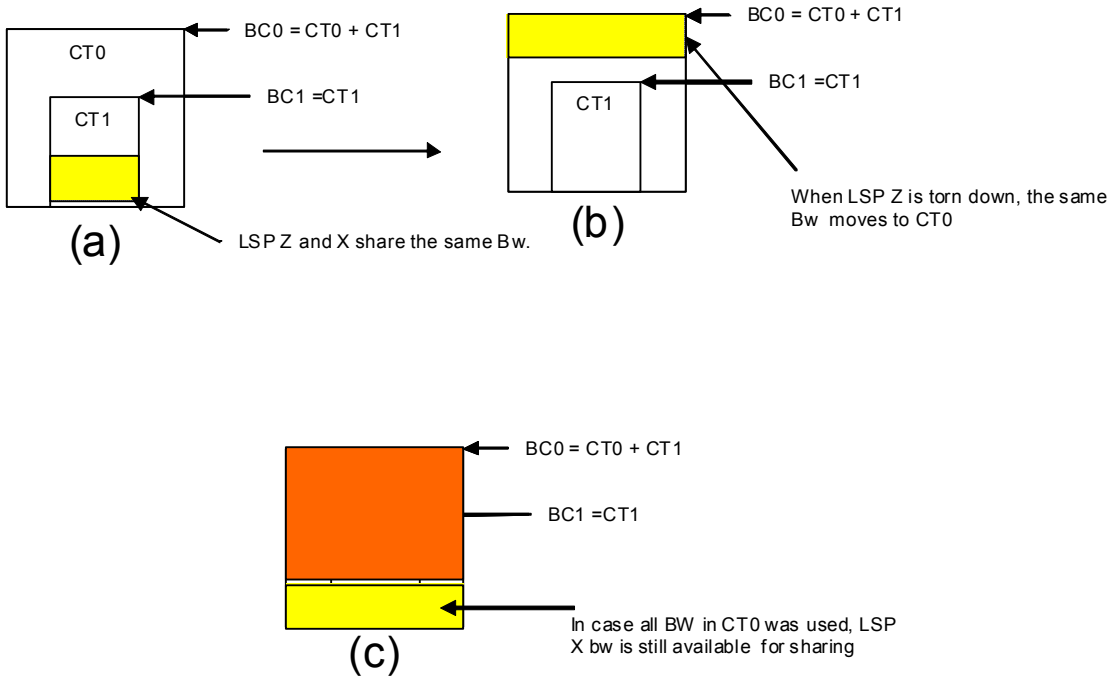


Figure 27: Sharing bandwidth when an LSP primary path is downgraded to backup CT

Consider an LSP path Z occupying bandwidth B at CT1. BC0 being the sum of all CTs below it, the bandwidth occupied in CT1 is guaranteed to be available in CT0. Thus when new path X of the same LSP for CT0 is setup, it will use the same bandwidth B as used by path Z as shown in Figure 27 (a). When path Z is torn down the same bandwidth now occupies CT0 as shown in Figure 27 (b). Even if there were no new BW available in CT0 as can be seen in Figure 27 (c), path X can always share the bandwidth with path Z.

CSPF at the head-end node and CAC at the transit LSR node will share bandwidth of an existing path when its CT is downgraded in the new path of the same LSP.

Upgrading the CT of Bandwidth Sharing LSP Paths

When upgrading the CT the following issue can be apparent. Assume an LSP path X exists with CT0. An attempt is made to upgrade this path to a new path Z with CT1 using an MBB.

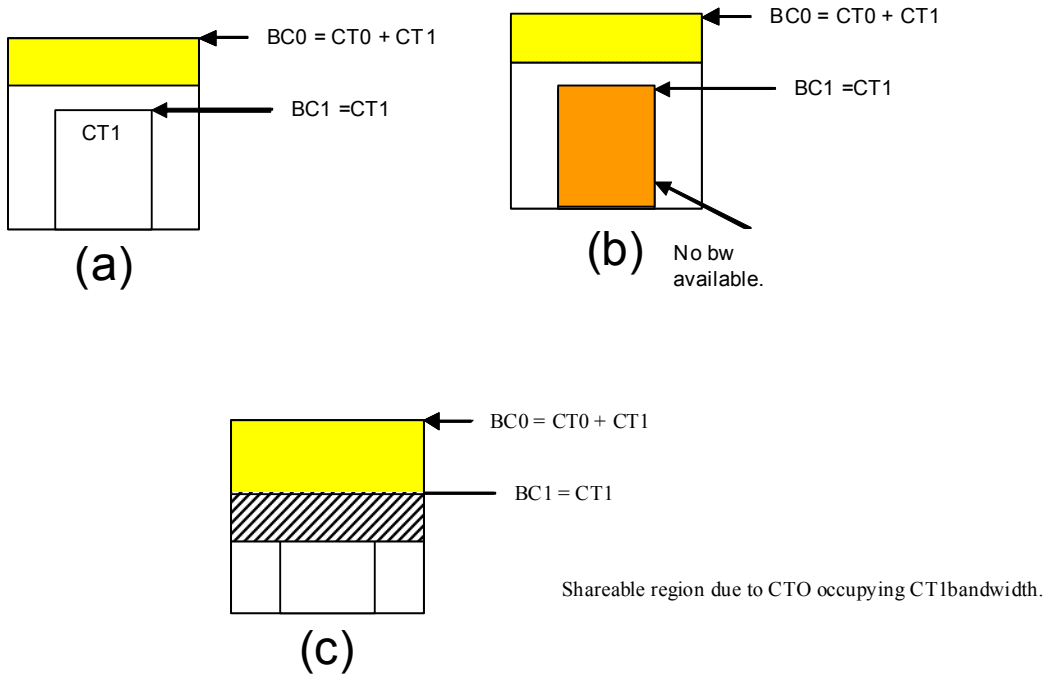


Figure 28: Sharing Bandwidth When an LSP Primary Path is Upgraded to Main CT

In Figure 28 (a), if the path X occupies the bandwidth as shown it can not share the bandwidth with the new path Z being setup. If a condition exists, as shown in Figure 28, (b) the path Z can never be setup on this particular link.

Consider Figure 28 (c). The CT0 has a region that overlaps with CT1 as CT0 has incursion into CT1. This overlap can be shared. However, in order to find whether such an incursion has occurred and how large the region is, it is required to know the reserved bandwidths in each class. Currently, IGP-TE advertises only the unreserved bandwidths. Hence, it is not possible to compute these overlap regions at the head end during CSPF. Moreover, the head end needs to then try and mimic each of the traversed links exactly which increases the complexity.

CSPF at the head-end node will only attempt to signal the LSP path with an upgraded CT if the advertised bandwidth for that CT can accommodate the bandwidth. In other words, it will assume that in the worst case this path will not share bandwidth with another path of the same LSP using a lower CT.

Advanced MPLS/RSVP Features

Extending RSVP LSP to use Loopback Interfaces Other Than router-id

It is possible to configure the address of a loopback interface, other than the router-id, as the destination of an RSVP LSP, or a P2MP S2L sub-LSP. In the case of a CSPF LSP, CSPF searches for the best path that matches the constraints across all areas and levels of the IGP where this address is reachable. If the address is the router-id of the destination node, then CSPF selects the best path across all areas and levels of the IGP for that router-id; regardless of which area and level the router-id is reachable as an interface.

In addition, the user can now configure the address of a loopback interface, other than the router-id, as a hop in the LSP path hop definition. If the hop is strict and corresponds to the router-id of the node, the CSPF path can use any TE enabled link to the downstream node, based on best cost. If the hop is strict and does not correspond to the router-id of the node, then CSPF will fail.

LSP Path Change

The **tools perform router mpls update-path** {**lsp** *lsp-name* **path** *current-path-name* **new-path** *new-path-name*} command instructs MPLS to replace the path of the primary or secondary LSP.

The primary or secondary LSP path is indirectly identified via the current-path-name value. In existing implementation, the same path name cannot be used more than once in a given LSP name.

This command is also supported on an SNMP interface.

This command applies to both CSPF LSP and to a non-CSPF LSP. However, it will only be honored when the specified current-path-name has the adaptive option enabled. The adaptive option can be enabled the LSP level or at the path level.

The new path must be first configured in CLI or provided via SNMP. The **configure router mpls path** *path-name* CLI command is used to enter the path.

The command fails if any of the following conditions are satisfied:

- The specified current-path-name of this LSP does not have the adaptive option enabled.
- The specified new-path-name value does not correspond to a previously defined path.
- The specified new-path-name value exists but is being used by any path of the same LSP, including this one.

When the command is executed, MPLS performs the following procedures:

- MPLS performs a single MBB attempt to move the LSP path to the new path.
- If the MBB is successful, MPLS updates the new path.
  - MPLS writes the corresponding NHLFE in the data path if this path is the current backup path for the primary.
  - If the current path is the active LSP path, it will update the path, write the new NHLFE in the data path, which will cause traffic to switch to the new path.
- If the MBB is not successful, the path retains it current value.
- The update-path MBB has the same priority as the manual re-signal MBB.

Manual LSP Path Switch

This feature provides a new command to move the path of an LSP from a standby secondary to another standby secondary.

The base version of the command allows the path of the LSP to move from a standby (or an active secondary) to another standby of the same priority. If a new standby path with a higher priority or a primary path comes up after the **tools perform** command is executed, the path re-evaluation command runs and the path is moved to the path specified by the outcome of the re-evaluation.

The CLI command for the base version is:

tools perform router mpls switch-path lsp *lsp-name* **path** *path-name*

The sticky version of the command can be used to move from a standby path to any other standby path regardless of priority. The LSP remains in the specified path until this path goes down or the user performs the no form of the **tools perform** command.

The CLI commands for the sticky version are:

tools perform router mpls force-switch-path lsp *lsp-name* **path** *path-name*
tools perform router mpls no force-switch-path lsp *lsp-name*

Make-Before-Break (MBB) Procedures for LSP/Path Parameter Configuration Change

When an LSP is switched from an existing working path to a new path, it is desirable to perform this in a hitless fashion. The Make-Before-Break (MBB) procedure consist of first signaling the new path when it is up, and having the ingress LER move the traffic to the new path. Only then the ingress LER tears down the original path.

MBB procedure is invoked during the following operations:

1. Timer based and manual re-signal of an LSP path.
2. Fast-ReRoute (FRR) global revertive procedures.
3. Soft Pre-emption of an LSP path.
4. Traffic-Engineering (TE) graceful shutdown procedures.
5. Update of secondary path due to an update to primary path SRLG.
6. LSP primary or secondary path name change.
7. LSP or path configuration parameter change.

In a prior implementation, item (7) covers the following parameters:

1. Changing the primary or secondary path **bandwidth** parameter on the fly.
2. Enabling the **frr** option for an LSP.

This feature extends the coverage of the MBB procedure to most of the other LSP level and Path level parameters as follows:

1. Changes to include/exclude of admin groups at LSP and path levels.
2. Enabling/disabling LSP level cspf option.
3. Enabling/disabling LSP level use-te-metric parameter when cspf option is enabled.
4. Enabling/disabling LSP level propagate-admin-group option.
5. Enabling/disabling LSP level hop-limit option in the fast-reroute context.
6. Enabling the LSP level least-fill option.
7. Enabling/disabling LSP level adspec option.
8. Changing between node-protect and “no node-protect” (link-protect) values in the LSP level fast-reroute option.
9. Changing LSP primary or secondary path priority values (setup-priority and hold-priority).

10. Changing LSP primary or secondary path class-type value and primary path backup-class-type value.
11. Changing LSP level and path level hop-limit parameter value.
12. Enabling/disabling primary or secondary path record and record-label options.

This feature is not supported on a manual bypass LSP.

P2MP Tree Level Make-before-break operation is supported if changes are made to the following parameters on LSP-Template:

1. Changing Bandwidth on P2MP LSP-Template.
2. Enabling Fast -Re-Route on P2MP LSP-Template.

Automatic Creation of RSVP-TE LSP Mesh

This feature enables the automatic creation of an RSVP point-to-point LSP to a destination node whose router-id matches a prefix in the specified peer prefix policy. This LSP type is referred to as auto-LSP of type mesh.

The user can associate multiple templates with the same or different peer prefix policies. Each application of an LSP template with a given prefix in the prefix list will result in the instantiation of a single CSPF computed LSP primary path using the LSP template parameters as long as the prefix corresponds to a router-id for a node in the TE database. Each instantiated LSP will have a unique LSP-id and a unique tunnel-ID.

Up to five (5) peer prefix policies can be associated with a given LSP template at all times. Each time the user executes the above command with the same or different prefix policy associations, or the user changes a prefix policy associated with an LSP template, the system re-evaluates the prefix policy. The outcome of the re-evaluation will tell MPLS if an existing LSP needs to be torn down or if a new LSP needs to be signaled to a destination address that is already in the TE database.

If a /32 prefix is added to (removed from) or if a prefix range is expanded (shrunk) in a prefix list associated with a LSP template, the same prefix policy re-evaluation described above is performed.

The trigger to signal the LSP is when the router with a router-id the matching a prefix in the prefix list appears in the Traffic Engineering database. The signaled LSP is installed in the Tunnel Table Manager (TTM) and is available to applications such as LDP-over-RSVP, resolution of BGP label routes, resolution of BGP, IGP, and static routes. It can also be used for provisioning an SDP; it is however, not available to be used as a provisioned SDP for explicit binding or auto-binding by services.

If the **one-hop** option is specified instead of a prefix policy, this command enables the automatic signaling of one-hop point-to-point LSPs using the specified template to all directly connected neighbors. This LSP type is referred to as auto-LSP of type one-hop. Although the provisioning model and CLI syntax differ from that of a mesh LSP only by the absence of a prefix list, the actual behavior is quite different. When the above command is executed, the TE database will keep track of each TE link that comes up to a directly connected IGP neighbor whose router-id is discovered. It then instructs MPLS to signal an LSP with a destination address matching the router-id of the neighbor and with a strict hop consisting of the address of the interface used by the TE link. Thus, the **auto-lsp** command with the **one-hop** option will result in one or more LSPs signaled to the neighboring router.

An auto-created mesh or one-hop LSP can have egress statistics collected at the ingress LER by adding the **egress-statistics** node configuration into the LSP template. The user can also have ingress statistics collected at the egress LER using the same **ingress-statistics** node in CLI used

with a provisioned LSP. The user must specify the full LSP name as signaled by the ingress LER in the RSVP session name field of the Session Attribute object in the received Path message.

RSVP-TE LSP Shortcut for IGP Resolution

RSVP-TE LSP shortcut for IGP route resolution allows forwarding of packets to IGP learned routes using an RSVP-TE LSP. This is also referred to as IGP shortcut. This feature is enabled by entering the following command at the IS-IS routing protocol level or at the OSPF routing protocol instance level:

- **config>router>isis>rsvp-shortcut**
- **config>router>ospf>rsvp-shortcut**

These commands instruct IS-IS or OSPF to include RSVP LSPs originating on this node and terminating on the router-id of a remote node as direct links with a metric equal to the operational metric provided by MPLS. If the user enabled the relative-metric option for this LSP, IGP will apply the shortest IGP cost between the endpoints of the LSP plus the value of the offset, instead of the LSP operational metric, when computing the cost of a prefix which is resolved to the LSP.

When a prefix is resolved to a tunnel next-hop, the packet is sent labeled with the label stack corresponding to the NHLFE of the RSVP LSP. Any network event causing an RSVP LSP to go down will trigger a full SPF computation which may result in installing a new route over another RSVP LSP shortcut as tunnel next-hop or over a regular IP next-hop.

When rsvp-shortcut is enabled at the IGP instance level, all RSVP LSPs originating on this node are eligible by default as long as the destination address of the LSP, as configured in **configure>router>mpls>lsp>to**, corresponds to a router-id of a remote node. RSVP LSPs with a destination corresponding to an interface address or any other loopback interface address of a remote node are automatically not considered by IS-IS or OSPF. The user can, however, exclude a specific RSVP LSP from being used as a shortcut for resolving IGP routes by entering the command:

- **config>router>mpls>lsp>no igp-shortcut**

The SPF in OSPF or IS-IS will only use RSVP LSPs as forwarding adjacencies, IGP shortcuts, or as endpoints for LDP-over-RSVP. These applications of RSVP LSPs are mutually exclusive at the IGP instance level. If the user enabled two or more options in the same IGP instance, then forwarding adjacency takes precedence over the shortcut application that takes precedence over the LDP-over-RSVP application.

Table 7 summarizes the outcome in terms of RSVP LSP role of mixing these configuration options.

Table 7: RSVP LSP Role As Outcome of LSP level and IGP level configuration options

	IGP Instance level configurations					
LSP level configuration	advertise-tunnel-link enabled / rsvp-short-cut enabled / ldp-over-rsvp enabled	advertise-tunnel-link enabled / rsvp-short-cut enabled / ldp-over-rsvp disabled	advertise-tunnel-link enabled / rsvp-short-cut disabled / ldp-over-rsvp disabled	advertise-tunnel-link disabled / rsvp-short-cut disabled / ldp-over-rsvp disabled	advertise-tunnel-link disabled / rsvp-short-cut enabled / ldp-over-rsvp enabled	advertise-tunnel-link disabled / rsvp-short-cut disabled / ldp-over-rsvp enabled
igp-shortcut enabled / ldp-over-rsvp enabled	Forwarding Adjacency	Forwarding Adjacency	Forwarding Adjacency	None	IGP Shortcut	LDP-over-RSVP
igp-shortcut enabled / ldp-over-rsvp disabled	Forwarding Adjacency	Forwarding Adjacency	Forwarding Adjacency	None	IGP Shortcut	None
igp-shortcut disabled / ldp-over-rsvp enabled	None	None	None	None	None	LDP-over-RSVP
igp-shortcut disabled / ldp-over-rsvp disabled	None	None	None	None	None	None

The resolution and forwarding of IPv6 prefixes to IPv4 IGP shortcuts is not supported.

The **no** form of this command disables the resolution of IGP routes using RSVP shortcuts.

Using LSP Relative Metric with IGP Shortcut

By default, the absolute metric of the LSP is used to update the SPF tree when the user enables IGP shortcut by configuring the rsvp-shortcut option in IGP. The absolute metric is the operational metric of the LSP populated by MPLS in the Tunnel Table Manager (TTM). This corresponds to the cumulative IGP-metric of the LSP path returned by CSPF or the static admin metric value of the LSP if the user configured one using the **config>router>mpls>lsp>metric** command. Note that MPLS populates the TTM with the maximum metric value of 16777215 in the case of a CSPF LSP using the TE-metric and a non-CSPF LSP with a loose or strict hop in the path. A non-CSPF LSP with an empty hop in the path definition returns the IGP cost for the destination of the LSP.

The user enables the use of the relative metric for an IGP shortcut with the following new CLI command:

config>router>mpls>lsp>igp-shortcut relative-metric [*offset*]

IGP will apply the shortest IGP cost between the endpoints of the LSP plus the value of the offset, instead of the LSP operational metric, when computing the cost of a prefix which is resolved to the LSP.

The offset value is optional and it defaults to zero. An offset value of zero is used when the **relative-metric** option is enabled without specifying the offset parameter value.

The minimum net cost for a prefix is capped to the value of one (1) after applying the offset:

Prefix cost = max(1, IGP cost + relative metric offset)

Note that the TTM continues the show the LSP operational metric as provided by MPLS. In other words, applications such as LDP-over-RSVP (when IGP shortcut is disabled) and BGP and static route shortcuts will continue to use the LSP operational metric.

The **relative-metric** option is mutually exclusive with the **lfa-protect** or the **lfa-only** options. In other words, an LSP with the **relative-metric** option enabled cannot be included in the LFA SPF and vice-versa when the **rsvp-shortcut** option is enabled in the IGP.

Finally, it should be noted that the **relative-metric** option is ignored when forwarding adjacency is enabled in IS-IS or OSPF by configuring the **advertise-tunnel-link** option. In this case, IGP advertises the LSP as a point-to-point unnumbered link along with the LSP operational metric capped to the maximum link metric allowed in that IGP.

The resolution and forwarding of IPv6 prefixes to IPv4 forwarding adjacency LSP is not supported.

ECMP Considerations

When ECMP is enabled on the system and multiple equal-cost paths exist for a prefix, the following selection criteria are used to pick up the set of next-hops to program in the data path:

- for a destination = tunnel-endpoint (including external prefixes with tunnel-endpoint as the next-hop):
  - select tunnel with lowest tunnel-index (ip next-hop is never used in this case)
- for a destination != tunnel-endpoint:
  - exclude LSPs with metric higher than underlying IGP cost between the endpoint of the LSP
  - prefer tunnel next-hop over ip next-hop

- within tunnel next-hops:
  i. select lowest endpoint to destination cost
  ii. if same endpoint to destination cost, select lowest endpoint node router-id
  iii. if same router-id, select lowest tunnel-index
- within ip next-hops:
  i. select lowest downstream router-id
  ii. if same downstream router-id, select lowest interface-index

- Note though no ECMP is performed across both the IP and tunnel next-hops the tunnel end-point lies in one of the shortest IGP paths for that prefix. In that case, the tunnel next-hop is always selected as long as the prefix cost using the tunnel is equal or lower than the IGP cost.

The ingress IOM will spray the packets for a prefix over the set of tunnel next-hops and IP next-hops based on the hashing routine currently supported for IPv4 packets.

Handling of Control Packets

All control plane packets that require an RTM lookup and whose destination is reachable over the RSVP shortcut will be forwarded over the shortcut. This is because RTM keeps a single route entry for each prefix unless there is ECMP over different outgoing interfaces.

Interface bound control packets are not impacted by the RSVP shortcut since RSVP LSPs with a destination address different than the router-id are not included by IGP in its SPF calculation.

Forwarding Adjacency

The forwarding adjacency feature can be enabled independently from the IGP shortcut feature in CLI. To enable forwarding adjacency, the user enters the following command in IS-IS or OSPF:

- **configure>router>isis>advertise-tunnel-link**
- **configure>router>ospf>advertise-tunnel-link**

If both **rsvp-shortcut** and **advertise-tunnel-link** options are enabled for a given IGP instance, then the **advertise-tunnel-link** will win. With this feature, ISIS or OSPF advertises an RSVP LSP as a link so that other routers in the network can include it in their SPF computations. The RSVP LSP is advertised as an unnumbered point-to-point link and the link LSP/LSA has no Traffic Engineering opaque sub-TLVs as per RFC 3906 *Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels*.

The forwarding adjacency feature can be enabled independently from the IGP shortcut feature in CLI. If both rsvp-shortcut and advertise-tunnel-link options are enabled for a given IGP instance, then the advertise-tunnel-link will win.

When the forwarding adjacency feature is enabled, each node advertises a p2p unnumbered link for each best metric tunnel to the router-id of any endpoint node. The node does not include the tunnels as IGP shortcuts in SPF computation directly. Instead, when the LSA/LSP advertising the corresponding P2P unnumbered link is installed in the local routing database, then the node performs an SPF using it like any other link LSA/LSP. The link bi-directional check requires that a link, regular link or tunnel link, exists in the reverse direction for the tunnel to be used in SPF.

Note that the **igp-shortcut** option under the LSP name governs the use of the LSP with both the **rsvp-shortcut** and the **advertise-tunnel-link** options in IGP. The interactions of these options are summarized in Table 8:

Table 8: Impact of LSP level configuration on IGP shortcut and forwarding adjacency features

LSP level configuration	Actions with IGP Shortcut Feature	Actions with Forwarding Adjacency Feature
igp-shortcut	Tunnel is used in main SPF, but is not used in LFA SPF	Tunnel is advertised as p2p link if it has best LSP metric, is used in main SPF if advertised, but is not used in LFA SPF
igp-shortcut lfa-protect	Tunnel is used in main SPF, and is used in LFA SPF	Tunnel is advertised as p2p link if it has best LSP metric, is used in main SPF if advertised, and is used in LFA SPF regardless if it is advertised or not
igp-shortcut lfa-only	Tunnel is not used in main SPF, but is used in LFA SPF	Tunnel is not advertised as p2p link, if not used in main SPF, but is used in LFA SPF

LDP Forwarding over IGP Shortcut

The user can enable LDP FECs over IGP shortcuts by configuring T-LDP sessions to the destination of the RSVP LSP. In this case, LDP FEC is tunneled over the RSVP LSP, effectively implementing LDP-over-RSVP without having to enable the **ldp-over-rsvp** option in OSPF or IS-IS. The **ldp-over-rsvp** and **igp-shortcut** options are mutually exclusive under OSFP or IS-IS.

Handling of Multicast Packets

This feature supports multicast Reverse-Path Check (RPF) in the presence of IGP shortcuts. When the multicast source for a packet is reachable via an IGP shortcut, the RPF check fails since PIM requires a bi-directional path to the source but IGP shortcuts are unidirectional.

The implementation of the IGP shortcut feature provides IGP with the capability to populate the multicast RTM with the prefix IP next-hop when both the **rsvp-shortcut** option and the **multicast-import** option are enabled in IGP.

This change is made possible with the enhancement introduced by which SPF keeps track of both the direct first hop and the tunneled first hop of a node that is added to the Dijkstra tree.

Note that IGP will not pass LFA next-hop information to the mcast RTM in this case. Only ECMP next-hops are passed. As a consequence, features such as PIM Multicast-Only FRR (MoFRR) will only work with ECMP next-hops when IGP shortcuts are enabled.

Finally, note that the concurrent enabling of the **advertise-tunnel-link** option and the **multicast-import** option will result a multicast RTM that is a copy of the unicast RTM and is thus populated with mix of IP and tunnel NHs. RPF will succeed for a prefix resolved to a IP NH, but will fail for a prefix resolved to a tunnel NH. Table 9 summarizes the interaction of the **rsvp-shortcut** and **advertise-tunnel-link** options with unicast and multicast RTMs.

Table 9: Impact of IGP Shortcut and Forwarding Adjacency on Unicast and Multicast RTM

		Unicast RTM (Primary SPF)	Multicast RTM (Primary SPF)	Unicast RTM (LFA SPF)	Multicast RTM (LFA SPF)
OSPF	rsvp-shortcut	√	√ **(1)**	√	X **(3)**
	advertise-tunnel-link	√	√ **(2)**	√	√ **(4)**
IS-IS	rsvp-shortcut	√	√ **(1)**	√	X **(3)**
	advertise-tunnel-link	√	√ **(2)**	√	√ **(4)**

Notes:

1. Multicast RTM is different from unicast RTM as it is populated with IP NHs only, including ECMP IP NHs. RPF check can be performed for all prefixes.

2. Multicast RTM is a copy of the unicast RTM and is thus populated with mix of IP and tunnel NHs. RPF will succeed for a prefix resolved to a IP NH but will fail for a prefix resolved to a tunnel NH.

3. LFA NH is not computed for the IP primary next-hop of a prefix passed to multicast RTM even if the same IP primary next-hop ends up being installed in the unicast RTM. The LFA next-hop will, however, be computed and installed in the unicast RTM for a primary IP next-hop of a prefix.
4. Multicast RTM is a copy of the unicast RTM and is thus populated with mix of IP and tunnel LFA NHs. RPF will succeed for a prefix resolved to a primary or LFA IP NH but will fail for a prefix resolved to a primary or LFA tunnel NH.

Disabling TTL Propagation in an LSP Shortcut

This feature provides the option for disabling TTL propagation from a transit or a locally generated IP packet header into the LSP label stack when an RSVP LSP is used as a shortcut for BGP next-hop resolution, a static-route next-hop resolution, or for an IGP route resolution.

A transit packet is a packet received from an IP interface and forwarded over the LSP shortcut at ingress LER.

A locally-generated IP packet is any control plane packet generated from the CPM and forwarded over the LSP shortcut at ingress LER.

TTL handling can be configured for all RSVP LSP shortcuts originating on an ingress LER using the following global commands:

config>router>mpls>[no] shortcut-transit-ttl-propagate
config>router>mpls>[no] shortcut-local-ttl-propagate

These commands apply to all RSVP LSPs which are used to resolve static routes, BGP routes, and IGP routes.

When the **no** form of the above command is enabled for local packets, TTL propagation is disabled on all locally generated IP packets, including ICMP Ping, trace route, and OAM packets that are destined to a route that is resolved to the LSP shortcut. In this case, a TTL of 255 is programmed onto the pushed label stack. This is referred to as pipe mode.

Similarly, when the **no** form is enabled for transit packets, TTL propagation is disabled on all IP packets received on any IES interface and destined to a route that is resolved to the LSP shortcut. In this case, a TTL of 255 is programmed onto the pushed label stack.

RSVP-TE LSP Signaling using LSP Template

LSP template can be used for signaling RSVP-TE LSP to far-end PE node that is detected based on auto-discovery method by a client application. RSVP-TE P2MP LSP signaling based on LSP template is supported for Multicast VPN application on SROS platform. LSP template avoids an explicit LSP or LSP S2L configuration for a node that is dynamically added as a receiver.

LSP template has option to configure traffic engineering parameters that apply to LSP that is setup using the template. Traffic engineering options that are currently supported are:

- adaptive
- admin-group
- bandwidth
- CSPF calculation
- fast-reroute
- hop-limit
- record-label
- retry-timer

Shared Risk Link Groups

Shared Risk Link Groups (SRLGs) is a feature that allows the user to establish a backup secondary LSP path or a FRR LSP path which is disjoint from the path of the primary LSP. Links that are members of the same SRLG represent resources sharing the same risk, for example, fiber links sharing the same conduit or multiple wavelengths sharing the same fiber.

When the SRLG option is enabled on a secondary path, CSPF includes the SRLG constraint in the computation of the secondary LSP path. This requires that the primary LSP already be established and up since the head-end LER needs the most current ERO computed by CSPF for the primary path. CSPF would return the list of SRLG groups along with the ERO during primary path CSPF computation. At a subsequent establishment of a secondary path with the SRLG constraint, the MPLS/RSVP task will query again CSPF providing the list of SLRG group numbers to be avoided. CSPF prunes all links with interfaces which belong to the same SRLGs as the interfaces included in the ERO of the primary path. If CSPF finds a path, the secondary is setup. If not, MPLS/RSVP will keep retrying the requests to CSPF.

When the SRLG option is enabled on FRR, CSPF includes the SRLG constraint in the computation of a FRR detour or bypass for protecting the primary LSP path. CSPF prunes all links with interfaces which belong to the same SRLG as the interface which is being protected, for example, the outgoing interface at the PLR the primary path is using. If one or more paths are found, the MPLS/RSVP task will select one based on best cost and will signal the bypass/detour. If not and the user included the strict option, the bypass/detour is not setup and the MPLS/RSVP task will keep retrying the request to CSPF. Otherwise, if a path exists which meets the other TE constraints, other than the SRLG one, the bypass/detour is setup.

A bypass or a detour LSP path is not guaranteed to be SRLG disjoint from the primary path. This is because only the SRLG constraint of the outgoing interface at the PLR that the primary path is using is avoided.

Enabling Disjoint Backup Paths

A typical application of the SRLG feature is to provide for an automatic placement of secondary backup LSPs or FRR bypass/detour LSPs that minimizes the probability of fate sharing with the path of the primary LSP (Figure 29).

The following details the steps necessary to create shared risk link groups:

- For primary/standby SRLG disjoint configuration:
  - Create an SRLG-group, similar to admin groups.
  - Link the SRLG-group to MPLS interfaces.

- Configure primary and secondary LSP paths and enable SRLG on the secondary LSP path. Note that the SRLG secondary LSP path(s) will *always* perform a strict CSPF query. The **srlg-frr** command is irrelevant in this case (see srlg-frr on page 312).

- For FRR detours/bypass SRLG disjoint configuration:
  - Create an SRLG group, similar to admin groups.
  - Link the SRLG group to MPLS interfaces.
  - Enable the **srlg-frr** (strict/non-strict) option, which is a system-wide parameter, and it force every LSP path CSPF calculation, to take the configured SRLG membership(s) (and propagated through the IGP opaque-te-database) into account.
  - Configure primary FRR (one-to-one/facility) LSP path(s). Consider that each PLR will create a detour/bypass that will only avoid the SRLG membership(s) configured on the primary LSP path egress interface. In a one-to-one case, detour-detour merging is out of the control of the PLR, thus the latter will not ensure that its detour will be prohibited to merge with a colliding one. For facility bypass, with the presence of several bypass type to bind to, the following priority rules will be followed:
    1. Manual bypass disjoint
    2. Manual bypass non-disjoint (eligible only if srlg-frr is non-strict)
    3. Dynamic disjoint
    4. Dynamic non-disjoint (eligible only if srlg-frr is non-strict)

  Non-CSPF manual bypass is not considered.

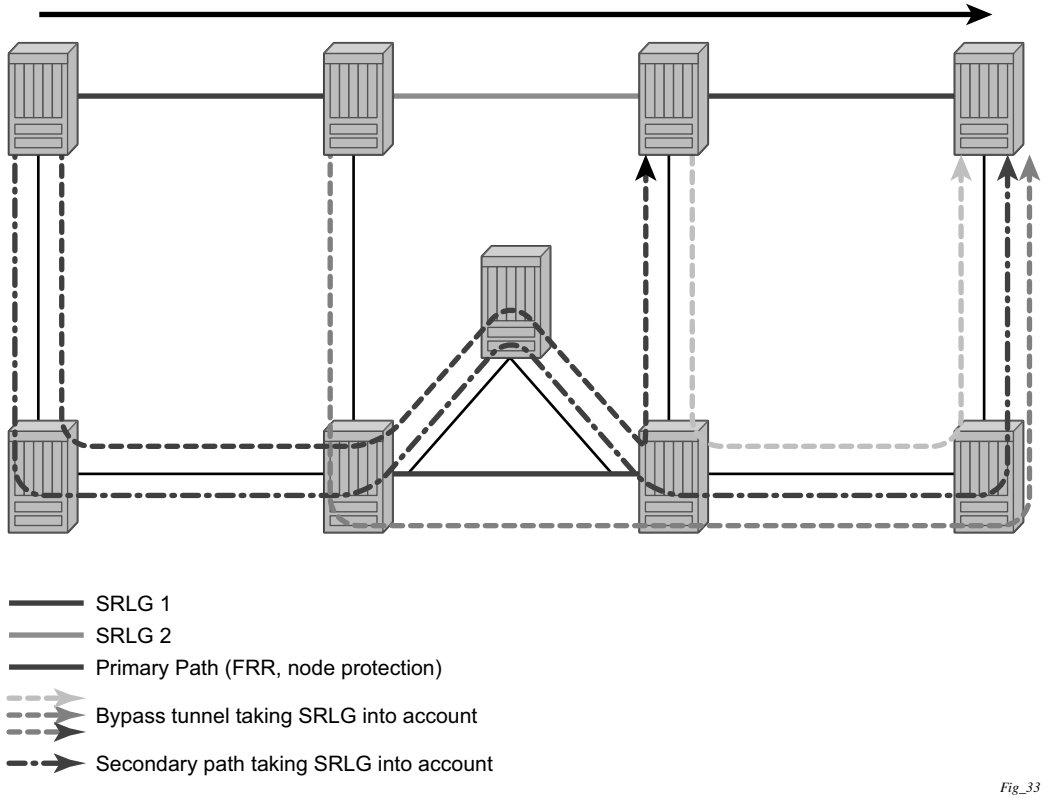


Figure 29: Shared Risk Link Groups

This feature is supported on OSPF and IS-IS interfaces on which RSVP is enabled.

Static Configurations of SRLG Memberships

This feature provides operations with the ability to manually enter the link members of SRLG groups for the entire network at any SR OS node which will need to signal LSP paths (for example, a head-end node).

The operator may explicitly enables the use by CSPF of the SRLG database. In that case, CSPF will not query the TE database for IGP advertised interface SRLG information.

Note, however, that the SRLG secondary path computation and FRR bypass/detour path computation remains unchanged.

There are deployments where the SR OS will interoperate with routers that do not implement the SRLG membership advertisement via IGP SRLG TLV or sub-TLV.

In these situations, the user is provided with the ability to enter manually the link members of SRLG groups for the entire network at any SR OS node which will need to signal LSP paths, for example, a head-end node.

The user enters the SRLG membership information for any link in the network by using the **interface** *ip-int-name* **srlg-group** *group-name* command in the **config>router>mpls> srlg-database>router-id** context. An interface can be associated with up to 5 SRLG groups for each execution of this command. The user can associate an interface with up to 64 SRLG groups by executing the command multiple times. The user must also use this command to enter the local interface SRLG membership into the user SRLG database. The user deletes a specific interface entry in this database by executing the **no** form of this command.

The *group-name* must have been previously defined in the SRLG **srlg-group** *group-name* **value** *group-value* command in the **config>router>mpls**. The maximum number of distinct SRLG groups the user can configure on the system is 1024.

The parameter value for *router-id* must correspond to the router ID configured under the base router instance, the base OSPF instance or the base IS-IS instance of a given node. Note however that a single user SLRG database is maintained per node regardless if the listed interfaces participate in static routing, OSPF, IS-IS, or both routing protocols. The user can temporarily disable the use by CSPF of all interface membership information of a specific router ID by executing the **shutdown** command in the **config>router>mpls> srlg-database> router-id** context. In this case, CSPF will assume these interfaces have no SRLG membership association. The operator can delete all interface entries of a specific router ID entry in this database by executing the **no router-id** *router-address* command in the **config>router>mpls> srlg-database** context.

CSPF will not use entered SRLG membership if an interface is not listed as part of a router ID in the TE database. If an interface was not entered into the user SRLG database, it will be assumed that it does not have any SRLG membership. CSPF will not query the TE database for IGP advertised interface SRLG information.

The operator enables the use by CSPF of the user SRLG database by entering the user-srlg-db enable command in the **config>router>mpls** context. When the MPLS module makes a request to CSPF for the computation of an SRLG secondary path, CSPF will query the local SRLG and computes a path after pruning links which are members of the SRLG IDs of the associated primary path. Similarly, when MPLS makes a request to CSPF for a FRR bypass or detour path to associate with the primary path, CSPF queries the user SRLG database and computes a path after pruning links which are members of the SRLG IDs of the PLR outgoing interface.

The operator can disable the use of the user SRLG database by entering the user-srlg-db disable in command in the **config>router>mpls** context. CSPF will then resumes queries into the TE database for SRLG membership information. However, the user SRLG database is maintained

The operator can delete the entire SRLG database by entering the **no srlg-database** command in the **config>router>mpls** context. In this case, CSPF will assume all interfaces have no SRLG membership association if the user has not disabled the use of this database.

TE Graceful Shutdown

Graceful shutdown provides a method to bulk re-route transit LSPs away from the node during software upgrade of a node. A solution is described in RFC 5817, *Graceful Shutdown in MPLS and Generalized MPLS Traffic Engineering Networks*. This is achieved in this RFC by using a PathErr message with a specific error code Local Maintenance on TE link required flag. When a LER gets this message, it performs a make-before-break on the LSP path to move the LSP away from the links/nodes which IP addresses were indicated in the PathErr message.

Graceful shutdown can flag the affected link/node resources in the TE database so other routers will signal LSPs using the affected resources only as a last resort. This is achieved by flooding an IGP TE LSA/LSP containing link TLV for the links under graceful shutdown with the traffic engineering metric set to 0xffffffff and 0 as unreserved bandwidth.

Soft Pre-emption of Diff-Serv RSVP LSP

A Diff-Serv LSP can pre-empt another LSP of the same or of a different CT if its setup priority is strictly higher (numerically lower) than the holding priority of that other LSP.

Least-Fill Bandwidth Rule in CSPF ECMP Selection

When multiples equal-cost paths satisfy the constraints of a given RSVP LSP path, CSPF in the router head-end node will select a path so that LSP bandwidth is balanced across the network links. In releases prior to R7.0, CSPF used a random number generator to select the path and returned it to MPLS. In the course of time, this method actually balances the number of LSP paths over the links in the network; it does not necessarily balance the bandwidth across those links.

The least-fill path selection algorithm identifies the single link in each of the equal cost paths which has the least available bandwidth in proportion to its maximum reserved bandwidth. It then selects the path which has the largest value of this figure. The net affect of this algorithm is that LSP paths will be spread over the network links over time such that percentage link utilization is balanced. When the least-fill option is enabled on an LSP, during a manual reset CSPF will apply this method to all path calculations of the LSP, also at the time of the initial configuration.

Inter Area TE LSP (ERO Expansion Method)

Inter area contiguous LSP scheme provides end-to-end TE path. Each transit node in an area can set up a TE path LSP based on TE information available within its local area.

A PE node initiating an inter area contiguous TE LSP does partial CSPF calculation to include its local area border router as a loose node.

Area border router on receiving a PATH message with loose hop ERO does a partial CSPF calculation to the next domain border router as loose hop or CSPF to reach the final destination.

The **cspf-on-loose-hop** option must be enabled on LSR to allow CSPF calculation for the next segment on receiving a PATH message with loose hop in ERO.

Area Border Node FRR Protection for Inter Area LSP

This feature enhances the prior implementation of an inter-area RSVP P2P LSP by making the ABR selection automatic at the ingress LER. The user will not need to include the ABR as a loose-hop in the LSP path definition.

CSPF adds a new capability to compute all segments of a multi-segment intra-area or inter-area LSP path in one operation. In previous releases, MPLS makes a request to CSPF for each segment separately.

Figure 7 1 illustrates the role of each node in the signaling of an inter-area LSP with automatic ABR node selection.

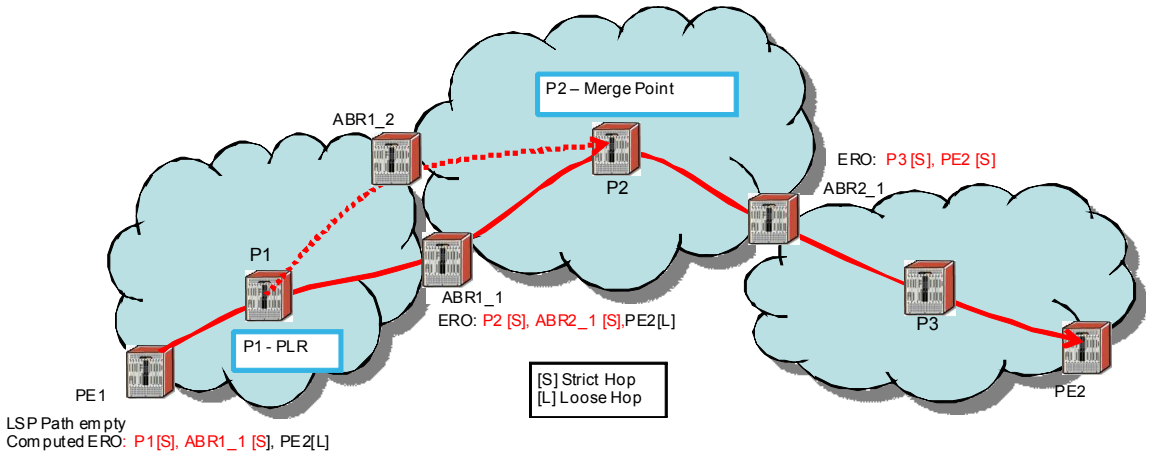


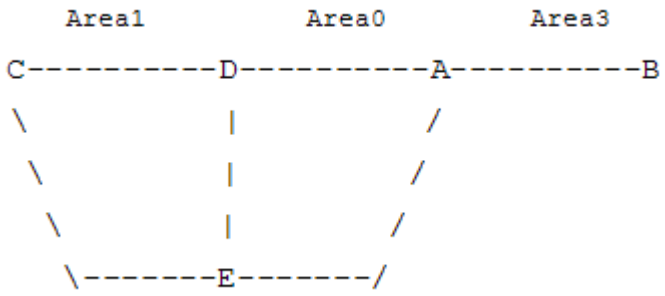
Figure 30: Automatic ABR Node Selection for Inter-Area LSP

CSPF for an inter-area LSP operates as follows:

1. CSPF in the Ingress LER node determines that an LSP is inter-area by doing a route lookup with the destination address of a P2P LSP (i.e., the address in the to field of the LSP configuration). If there is no intra-area route to the destination address, the LSP is considered as inter-area.
2. When the path of the LSP is empty, CPSF will compute a single-segment intra-area path to an ABR node that advertised a prefix matching with the destination address of the LSP.
3. When the path of the LSP contains one or more hops, CSPF will compute a multi-segment intra-area path including the hops that are in the area of the Ingress LER node.
4. When all hops are in the area of the ingress LER node, the calculated path ends on an ABR node that advertised a prefix matching with the destination address of the LSP.
5. When there are one or more hops that are not in the area of the ingress LER node, the calculated path ends on an ABR node that advertised a prefix matching with the first hop-address that is not in the area of the ingress LER node.
6. Note the following special case of a multi-segment inter-area LSP. If CSPF hits a hop that can be reached via an intra-area path but that resides on an ABR, CSPF only calculates a path up to that ABR. This is because there is a better chance to reach the destination of the LSP by first signaling the LSP up to that ABR and continuing the path calculation from there on by having the ABR expand the remaining hops in the ERO.

   This behavior can be illustrated in the following example. The TE link between ABR nodes D and E is in area 0. When node C computes the path for LSP from C to B which path specified nodes C and D as loose hops, it would fail the path computation if CSPF

attempted a path all the way to the last hop in the local area, node E. Instead, CSPF stops the path at node A which will further expand the ERO by including link D-E as part of the path in area 0.



7. If there is more than 1 ABR that advertized a prefix, CSPF will calculate a path for all ABRs. Only the shortest path will be withheld. If more than one path has the shortest path, CSPF will pick a path randomly or based on the least-fill criterion if enabled. If more than one ABR satisfies the least-fill criterion, CSPF will also pick one path randomly.
8. The path for an intra-area LSP path will not be able to exit and re-enter the local area of the ingress LER. This behavior was possible in prior implementation when the user specified a loose hop outside of the local area or when the only available path was via TE links outside of the local area.

Rerouting of Inter-Area LSP

In prior implementation, an inter-area LSP path would have been re-routed if a failure or a topology change occurred in the local or a remote area while the ABR loose-hop in the path definition was still up. If the exit ABR node went down, went into IS-IS overload, or was put into node TE graceful shutdown, the LSP path will remain down at the ingress LER.

One new behavior introduced by the automatic selection of ABR is the ability of the ingress LER to reroute an inter-area LSP primary path via a different ABR in the following situations:

- When the local exit ABR node fails, There are two cases to consider:
  - The primary path is not protected at the ABR and is thus torn down by the previous hop in the path. In this case the ingress LER will retry the LSP primary path via the ABR which currently has the best path for the destination prefix of the LSP.
  - The primary path is protected at the ABR with a manual or dynamic bypass LSP. In this case the ingress LER will receive a Path Error message with a notification of a protection becoming active downstream and a RESV with a *Local-Protection-In-Use*

flag set. At the receipt of first of these two messages, the ingress LER will then perform a Global Revertive Make-Before-Break (MBB) to re-optimize the LSP primary path via the ABR which currently has the best path for the destination prefix of the LSP.

- When the local exit ABR node goes into IS-IS overload or is put into node TE Graceful Shutdown. In this case, the ingress LER will perform a MBB to re-optimize the LSP primary path via the ABR which currently has the best path for the destination prefix of the LSP. The MBB is performed at the receipt of the PathErr message for the node TE shutdown or at the next timer or manual re-optimization of the LSP path in the case of the receipt of the IS-IS overload bit.

Behavior of MPLS Options in Inter-Area LSP

The automatic ABR selection for an inter-area LSP does not change prior implementation inter-area LSP behavior of many of the LSP and path level options. There is, however, a number of enhancements introduced by the automatic ABR selection feature as explained in the following.

- Features such as path bandwidth reservation and admin-groups continue to operate within the scope of all areas since they rely on propagating the parameter information in the Path message across the area boundary.
- The TE graceful shutdown and soft pre-emption features will continue to support MBB of the LSP path to avoid the link or node that originated the PathErr message as long as the link or node is in the local area of the ingress LER. If the PathErr originated in a remote area, the ingress LER will not be able to avoid the link or node when it performs the MBB since it computes the path to the local ABR exit router only. There is, however, an exception to this for the TE graceful shutdown case only. An enhancement has been added to cause the upstream ABR nodes in the current path of the LSP to record the link or node to avoid and will use it in subsequent ERO expansions. This means that if the ingress LER computes a new MBB path which goes via the same exit ABR router as the current path and all ABR upstream nodes of the node or link which originated the PathErr message are also selected in the new MBB path when the ERO is expanded, the new path will indeed avoid this link or node. The latter is a new behavior introduced with the automatic ABR selection feature.
- The support of MBB to avoid the ABR node when the node is put into TE Graceful Shutdown is a new behavior introduced with the automatic ABR selection feature.
- The **use-te-metric** option in CSPF cannot be propagated across the area boundary and thus will operate within the scope of the local area of the ingress LER node. This is a new behavior introduced with the automatic ABR selection feature.
- The **srlg** option on bypass LSP will continue to operate locally at each PLR within each area. The PLR node protecting the ABR will check the SRLG constraint for the path of the bypass within the local area.

- The **srlg** option on secondary path is allowed to operate within the scope of the local area of the ingress LER node with the automatic ABR selection feature.
- The **least-fill** option support with an inter-area LSP is introduced with the automatic ABR selection feature. When this option is enabled, CSPF applies the least-fill criterion to select the path segment to the exit ABR node in the local area.
- 1The PLR node must indicate to CSPF that a request to one-to-one detour LSP path must remain within the local area. If the destination for the detour, which is the same as that of the LSP, is outside of the area, CSPF must return no path.
- The **propagate-admin-group** option under the LSP will still need to be enabled on the inter-area LSP if the user wants to have admin-groups propagated across the areas.
- With the automatic ABR selection feature, timer based re-signal of the inter-area LSP path will be supported and will re-signal the path if the cost of the path segment to the local exit ABR changed. The cost shown for the inter-area LSP at ingress LER will be the cost of the path segments to the ABR node.

Inter-Area LSP support of OSPF Virtual Links

The OSPF virtual link extends area 0 for a router that is not connected to area 0. As a result, it makes all prefixes in area 0 reachable via an intra-area path but in reality, they are not since the path crosses the transit area through which the virtual link is set up to reach the area 0 remote nodes.

The TE database in a router learns all of the remote TE links in area 0 from the ABR connected to the transit area, but an intra-area LSP path using these TE links cannot be signaled within area 0 since none of these links is directly connected to this node.

This inter-area LSP feature can identify when the destination of an LSP is reachable via a virtual link. In that case, CSPF will automatically compute and signal an inter-area LSP via the ABR nodes that is connected to the transit area.

However, when the ingress LER for the LSP is the ABR connected to the transit area and the destination of the LSP is the address corresponding to another ABR's router-id in that same transit area, CSPF will compute and signal an intra-area LSP using the transit area TE links, even when the destination router-id is only part of area 0.

Area Border Node FRR Protection for Inter Area LSP

For protection of the area border router, the upstream node of the area border router acts as a point-of-local-repair (PLR), and the next-hop node to the protected domain border router is the merge-point (MP). Both manual and dynamic bypass are available to protect area border node.

Manual bypass protection works only when a proper completely strict path is provisioned that avoids the area border node.

Dynamic bypass protection provides for the automatic computation, signaling, and association with the primary path of an inter-area P2P LSP to provide ABR node protection. Figure 31 illustrates the role of each node in the ABR node protection using a dynamic bypass LSP.

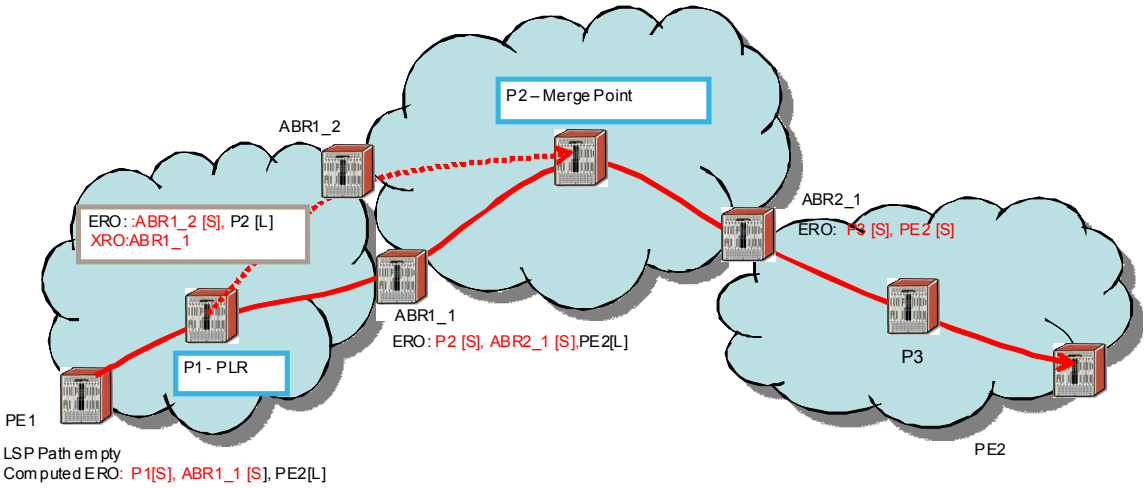


Figure 31: ABR Node Protection Using Dynamic Bypass LSP

In order for a PLR node within the local area of the ingress LER to provide ABR node protection, it must dynamically signal a bypass LSP and associate it with the primary path of the inter-area LSP using the following new procedures:

- The PLR node must inspect the node-id RRO of the LSP primary path to determine the address of the node immediately downstream of the ABR in the other area.
- The PLR signals an inter-area bypass LSP with a destination address set to the address downstream of the ABR node and with the XRO set to exclude the node-id of the protected ABR node.
- The request to CSPF is for a path to the merge-point (i.e., the next-next-hop in the RRO received in the RESV for the primary path) along with the constraint to exclude the protected ABR node and the include/exclude admin-groups of the primary path. If CSPF returns a path that can only go to an intermediate hop, then the PLR node signals the dynamic bypass and will automatically include the XRO with the address of the protected ABR node and propagate the admin-group constraints of the primary path into the Session Attribute object of the bypass LSP. Otherwise, the PLR signals the dynamic bypass directly to the merge-point node with no XRO object in the Path message.

- If a node-protect dynamic bypass cannot be found or signaled, the PLR node attempts a link-protect dynamic bypass LSP. As in existing implementation of dynamic bypass within the same area, the PLR attempts in the background to signal a node-protect bypass at the receipt of every third Resv refresh message for the primary path.
- Refresh reduction over dynamic bypass will only work if the node-id RRO also contains the interface address. Otherwise the neighbor will not be created once the bypass is activated by the PLR node. The Path state will then time out after three refreshes following the activation of the bypass backup LSP.

Note that a one-to-one detour backup LSP cannot be used at the PLR for the protection of the ABR node. As a result, a 7x50 PLR node will not signal a one-to-one detour LSP for ABR protection. In addition, an ABR node will reject a Path message, received from a third party implementation, with a detour object and with the ERO having the next-hop loose. This is performed regardless if the **cspf-on-loose** option is enabled or not on the 7x50 node. In other words, the 7x50 as a transit ABR for the detour path will reject the signaling of an inter-area detour backup LSP.

Automatic Creation of a RSVP Mesh LSP

Feature Configuration

The user first creates an LSP template of type mesh P2P:

config>router>mpls>lsp-template *template-name* **mesh-p2p**

Inside the template the user configures the common LSP and path level parameters or options shared by all LSPs using this template.

Then the user references the peer prefix list which is defined inside a policy statement defined in the global policy manager.

config>router>mpls>auto-lsp lsp-template *template-name* **policy** *peer-prefix-policy*

The user can associate multiple templates with same or different peer prefix policies. Each application of an LSP template with a given prefix in the prefix list will result in the instantiation of a single CSPF computed LSP primary path using the LSP template parameters as long as the prefix corresponds to a router-id for a node in the TE database. This feature does not support the automatic signaling of a secondary path for an LSP. If the user requires the signaling of multiple LSPs to the same destination node, he/she must apply a separate LSP template to the same or different prefix list which contains the same destination node. Each instantiated LSP will have a unique LSP-id and a unique tunnel-ID. This feature also does not support the signaling of a non-CSPF LSP. The selection of the '**no cspf**' option in the LSP template is thus blocked.

Up to 5 peer prefix policies can be associated with a given LSP template at all times. Each time the user executes the above command, with the same or different prefix policy associations, or the user changes a prefix policy associated with an LSP template, the system re-evaluates the prefix policy. The outcome of the re-evaluation will tell MPLS if an existing LSP needs to be torn down or a new LSP needs to be signaled to a destination address which is already in the TE database.

If a /32 prefix is added to (removed from) or if a prefix range is expanded (shrunk) in a prefix list associated with a LSP template, the same prefix policy re-evaluation described above is performed.

The user must perform a **no shutdown** of the template before it takes effect. Once a template is in use, the user must shutdown the template before effecting any changes to the parameters except for those LSP parameters for which the change can be handled with the Make-Before-Break (MBB) procedures. These parameters are **bandwidth** and enabling **fast-reroute** without the **hop-limit** or **node-protect** options. For all other parameters, the user shuts down the template and once a it is added, removed or modified, the existing instances of the LSP using this template are torn down and re-signaled.

Finally the auto-created mesh LSP can be signaled over both numbered and unnumbered RSVP interfaces.

Feature Behavior

Whether the prefix list contains one or more specific /32 addresses or a range of addresses, an external trigger is required to indicate to MPLS to instantiate an LSP to a node which address matches an entry in the prefix list. The objective of the feature is to provide an automatic creation of a mesh of RSVP LSP to achieve automatic tunneling of LDP-over-RSVP. The external trigger is when the router with the router-id matching an address in the prefix list appears in the Traffic Engineering database. In the latter case, the TE database provides the trigger to MPLS which means this feature operates with CSPF LSP only.

Each instantiation of an LSP template results in RSVP signaling and installing state of a primary path for the LSP to the destination router. The auto- LSP is installed in the Tunnel Table Manager (TTM) and is available to applications such as LDP-over-RSVP, resolution of BGP label routes, resolution of BGP, IGP, and static routes. The auto-LSP can also be used for auto-binding by services such as VPRN, BGP-AD VPLS, and FEC129 VLL service. The auto-LSP is however not available to be used in a provisioned SDP for explicit binding by services.

If the user changes the **bandwidth** parameter in the LSP template, an MBB is performed for all LSPs using the template. If however the **auto-bandwidth** option was enabled in the template, the bandwidth **parameter** change will be saved but will only take effect at the next time the LSP bounces or is re-signaled.

Except for the MBB limitations to the configuration parameter change in the LSP template, MBB procedures for manual and timer based re-signaling of the LSP, for TE Graceful Shutdown and for soft pre-emption are supported.

Note that the use of the '**tools perform router mpls update-path**' command with a mesh LSP is not supported.

The **one-to-one** option under **fast-reroute** is also not supported.

If while the LSP is UP, with the bypass backup path activated or not, the TE database loses the router-id, it will perform an update to MPLS module which will state router-id is no longer in TE database. This will cause MPLS to tear down all mesh LSPs to this router-id. Note however that if the destination router is not a neighbor of the ingress LER and the user shuts down the IGP instance in the destination router, the router-id corresponding to the IGP instance will only be deleted from the TE database in the ingress LER after the LSA/LSP ages out. If the user brought back up the IGP instance before the LSA/LSP aged out, the ingress LER will delete and re-install the same router-id at the receipt of the updated LSA/LSP. In other words, the RSVP LSPs destined to this router-id will get deleted and re-established. All other failure conditions will cause the LSP to activate the bypass backup LSP or to go down without being deleted.

There is no overall chassis mode restrictions enforced with the mesh LSP feature. If the chassis-mode, network chassis-mode or IOM type requirements for a feature are not met, the configuration of the corresponding command will not be allowed into the LSP template on the system.

Multi-Area and Multi-Instance Support

A router which does not have TE links within a given IGP area/level will not have its router-id discovered in the TE database by other routers in this area/level. In other words, an auto-LSP of type P2P mesh cannot be signaled to a router which does not participate in the area/level of the ingress LER.

A mesh LSP can however be signaled using TE links all belonging to the same IGP area even if the router-id of the ingress and egress routers are interfaces reachable in a different area. In this case, the LSP is considered to be an intra-area LSP.

If multiple instances of ISIS or OSPF are configured on a router, each with its own router-id value, the TE database in other routers will be able to discover TE links advertised by each instance. In such a case, an instance of an LSP can be signaled to each router-id with a CSPF path computed using TE links within each instance.

Finally, if multiple instances of ISIS or OSPF are configured on a destination router each with the same router-id value, a single instance of LSP will be signaled from other routers. If the user shuts down one IGP instance, this will be no op as long as the other IGP instances remain up. The LSP will remain up and will forward traffic using the same TE links. The same behavior exists with a provisioned LSP.

Mesh LSP Name Encoding and Statistics

When the ingress LER signals the path of a mesh auto-LSP, it includes the name of the LSP and that of the path in the Session Name field of the Session Attribute object in the Path message. The encoding is as follows:

Session Name: <lsp-name::path-name>, where lsp-name component is encoded as follows:

"***TemplateName-DestIpv4Address-TunnelId***"

Where ***DestIpv4Address*** is the address of the destination of the auto-created LSP.

At ingress LER, the user can enable egress statistics for the auto-created mesh LSP by adding the following configuration to the LSP template:

```
config
     router
          [no] mpls
               lsp-template template-name mesh-p2p]
               no lsp-template template-name
```

```
[no] egress-statistics
     accounting-policy policy-id
     no accounting-policy
     [no] collect-stats
```

If there are no stat indices available when an LSP is instantiated, the assignment is failed and the egress-statistics field in the show command for the LSP path will be in operational DOWN state but in admin UP state.

An auto-created mesh LSP can also have ingress statistics enabled on the egress LER as long as the user specifies the full LSP name following the above syntax.

configure>router>mpls>ingress-statistics>lsp *lsp-name* sender *ip-address*

Automatic Creation of a RSVP One-Hop LSPs

Feature Configuration

The user first creates an LSP template of type mesh P2P:

config>router>mpls>lsp-template *template-name* **mesh-p2p**

Inside the template the user configures the common LSP and path level parameters or options shared by all LSPs using this template.

Then the user references the peer prefix list which is defined inside a policy statement defined in the global policy manager.

config>router>mpls>auto-lsp lsp-template *template-name* **policy** *peer-prefix-policy*

The user can associate multiple templates with same or different peer prefix policies. Each application of an LSP template with a given prefix in the prefix list will result in the instantiation of a single CSPF computed LSP primary path using the LSP template parameters as long as the prefix corresponds to a router-id for a node in the TE database. This feature does not support the automatic signaling of a secondary path for an LSP. If the user requires the signaling of multiple LSPs to the same destination node, he/she must apply a separate LSP template to the same or different prefix list which contains the same destination node. Each instantiated LSP will have a unique LSP-id and a unique tunnel-ID. This feature also does not support the signaling of a non-CSPF LSP. The selection of the '**no cspf**' option in the LSP template is thus blocked.

Up to 5 peer prefix policies can be associated with a given LSP template at all times. Each time the user executes the above command, with the same or different prefix policy associations, or the user changes a prefix policy associated with an LSP template, the system re-evaluates the prefix policy. The outcome of the re-evaluation will tell MPLS if an existing LSP needs to be torn down or a new LSP needs to be signaled to a destination address which is already in the TE database.

If a /32 prefix is added to (removed from) or if a prefix range is expanded (shrunk) in a prefix list associated with a LSP template, the same prefix policy re-evaluation described above is performed.

The user must perform a **no shutdown** of the template before it takes effect. Once a template is in use, the user must shutdown the template before effecting any changes to the parameters except for those LSP parameters for which the change can be handled with the Make-Before-Break (MBB) procedures. These parameters are **bandwidth** and enabling **fast-reroute** without the **hop-limit** or **node-protect** options. For all other parameters, the user shuts down the template and once a it is added, removed or modified, the existing instances of the LSP using this template are torn down and re-signaled.

Finally the auto-created mesh LSP can be signaled over both numbered and unnumbered RSVP interfaces.

Feature Behavior

Whether the prefix list contains one or more specific /32 addresses or a range of addresses, an external trigger is required to indicate to MPLS to instantiate an LSP to a node which address matches an entry in the prefix list. The objective of the feature is to provide an automatic creation of a mesh of RSVP LSP to achieve automatic tunneling of LDP-over-RSVP. The external trigger is when the router with the router-id matching an address in the prefix list appears in the Traffic Engineering database. In the latter case, the TE database provides the trigger to MPLS which means this feature operates with CSPF LSP only.

Each instantiation of an LSP template results in RSVP signaling and installing state of a primary path for the LSP to the destination router. The auto- LSP is installed in the Tunnel Table Manager (TTM) and is available to applications such as LDP-over-RSVP, resolution of BGP label routes, resolution of BGP, IGP, and static routes. The auto-LSP can also be used for auto-binding by services such as VPRN, BGP-AD VPLS, and FEC129 VLL service. The auto-LSP is however not available to be used in a provisioned SDP for explicit binding by services.

If the user changes the **bandwidth** parameter in the LSP template, an MBB is performed for all LSPs using the template. If however the **auto-bandwidth** option was enabled in the template, the bandwidth **parameter** change will be saved but will only take effect at the next time the LSP bounces or is re-signaled.

Except for the MBB limitations to the configuration parameter change in the LSP template, MBB procedures for manual and timer based re-signaling of the LSP, for TE Graceful Shutdown and for soft pre-emption are supported.

Note that the use of the '**tools perform router mpls update-path**' command with a mesh LSP is not supported.

The **one-to-one** option under **fast-reroute** is also not supported.

If while the LSP is UP, with the bypass backup path activated or not, the TE database loses the router-id, it will perform an update to MPLS module which will state router-id is no longer in TE database. This will cause MPLS to tear down all mesh LSPs to this router-id. Note however that if the destination router is not a neighbor of the ingress LER and the user shuts down the IGP instance in the destination router, the router-id corresponding to the IGP instance will only be deleted from the TE database in the ingress LER after the LSA/LSP ages out. If the user brought back up the IGP instance before the LSA/LSP aged out, the ingress LER will delete and re-install the same router-id at the receipt of the updated LSA/LSP. In other words, the RSVP LSPs destined to this router-id will get deleted and re-established. All other failure conditions will cause the LSP to activate the bypass backup LSP or to go down without being deleted.

There is no overall chassis mode restrictions enforced with the mesh LSP feature. If the chassis-mode, network chassis-mode or IOM type requirements for a feature are not met, the configuration of the corresponding command will not be allowed into the LSP template on the system.

Multi-Area and Multi-Instance Support

A router which does not have TE links within a given IGP area/level will not have its router-id discovered in the TE database by other routers in this area/level. In other words, an auto-LSP of type P2P mesh cannot be signaled to a router which does not participate in the area/level of the ingress LER.

A mesh LSP can however be signaled using TE links all belonging to the same IGP area even if the router-id of the ingress and egress routers are interfaces reachable in a different area. In this case, the LSP is considered to be an intra-area LSP.

If multiple instances of ISIS or OSPF are configured on a router, each with its own router-id value, the TE database in other routers will be able to discover TE links advertised by each instance. In such a case, an instance of an LSP can be signaled to each router-id with a CSPF path computed using TE links within each instance.

Finally, if multiple instances of ISIS or OSPF are configured on a destination router each with the same router-id value, a single instance of LSP will be signaled from other routers. If the user shuts down one IGP instance, this will be no op as long as the other IGP instances remain up. The LSP will remain up and will forward traffic using the same TE links. The same behavior exists with a provisioned LSP.

Mesh LSP Name Encoding and Statistics

When the ingress LER signals the path of a mesh auto-LSP, it includes the name of the LSP and that of the path in the Session Name field of the Session Attribute object in the Path message. The encoding is as follows:

Session Name: <lsp-name::path-name>, where lsp-name component is encoded as follows:

"TemplateName-DestIpv4Address-TunnelId"

Where ***DestIpv4Address*** is the address of the destination of the auto-created LSP.

At ingress LER, the user can enable egress statistics for the auto-created mesh LSP by adding the following configuration to the LSP template:

```
config
     router
```

```
         [no] mpls
lsp-template template-name mesh-p2p]
no lsp-template template-name
                   [no] egress-statistics
                        accounting-policy policy-id
                        no accounting-policy
                        [no] collect-stats
```

If there are no stat indices available when an LSP is instantiated, the assignment is failed and the egress-statistics field in the show command for the LSP path will be in operational DOWN state but in admin UP state.

An auto-created mesh LSP can also have ingress statistics enabled on the egress LER as long as the user specifies the full LSP name following the above syntax.

configure>router>mpls>ingress-statistics>lsp *lsp-name* **sender** *ip-address*

Automatic Creation of a RSVP One-Hop LSP

Feature Configuration

The user first creates an LSP template of type one-hop:

config>router>mpls>lsp-template t*emplate-name* **one-hop-p2p**

Then the user enables the automatic signaling of one-hop LSP to all direct neighbors using the following command:

config>router>mpls>auto-lsp lsp-template *template-name* **one-hop**

The LSP and path parameters and options supported in a LSP template of type **one-hop-p2p** are that same as in the LSP template of type **mesh-p2p** except for the parameter **from** which is not allowed in a template of type **one-hop-p2p**. The show command for the auto-LSP will display the actual outgoing interface address in the ‘from’ field. The full list of template parameters is shown in the CLI Section. Also, the rules for adding or modifying the template parameters are as described in 7.1.1.

Finally the auto-created one-hop LSP can be signaled over both numbered and unnumbered RSVP interfaces.

Feature Behavior

Although the provisioning model and CLI syntax differ from that of a mesh LSP only by the absence of a prefix list, the actual behavior is quite different. When the above command is

executed, the TE database will keep track of each TE link which comes up to a directly connected IGP neighbor which router-id is discovered. It then instructs MPLS to signals an LSP with a destination address matching the router-id of the neighbor and with a strict hop consisting of the address of the interface used by the TE link. Thus the **auto-lsp** command with the **one-hop** option will result in one or more LSPs signaled to the IGP neighbor.

Only the router-id of the first IGP instance of the neighbor which advertises a TE link will cause the LSP to be signaled. If subsequently another IGP instance with a different router-id advertises the same TE link, no action is taken and the existing LSP is kept up. If the router-id originally used disappears from the TE database, the LSP is kept up and is associated now with the other router-id.

The state of a one-hop LSP once signaled follows the following behavior:

- If the interface used by the TE link goes down or BFD times out and the RSVP interface registered with BFD, the LSP path moves to the bypass backup LSP if the primary path is associated with one.
- If while the one-hop LSP is UP, with the bypass backup path activated or not, the association of the TE-link with a router-id is removed in the TE databases, the one-hop LSP is torn down. This would be the case if the interface used by the TE link is deleted or if the interface is shutdown in the context of RSVP.
- If while the LSP is UP, with the bypass backup path activated or not, the TE database loses the router-id, it will perform two separate updates to MPLS module. The first one updates the loss of the TE link association which will cause action (B) above for the one-hop LSP. The other update will state router-id is no longer in TE database which will cause MPLS to tear down all mesh LSPs to this router-id as explained in Section 7.1.2. Note however that a shutdown at the neighbor of the IGP instance which advertised the router-id will cause the router-id to be removed from the ingress LER node immediately after the last IGP adjacency is lost and is not subject to age-out as for a non-directly connected destination router.

All other feature behavior, limitations, and statistics support are the same as for an auto-LSP of type **mesh-p2p**.

Point-to-Multipoint (P2MP) RSVP LSP

Point-to-multipoint (P2MP) RSVP LSP allows the source of multicast traffic to forward packets to one or many multicast receivers over a network without requiring a multicast protocol, such as PIM, to be configured in the network core routers. A P2MP LSP tree is established in the control plane which path consists of a head-end node, one or many branch nodes, and the leaf nodes. Packets injected by the head-end node are replicated in the data plane at the branching nodes before they are delivered to the leaf nodes.

Application in Video Broadcast

Figure 32 illustrates the use of the SR product family in triple play application (TPSDA). The Broadband Service Router (BSR) is a 7750 SR and the Broadband Service Aggregator (BSA) is the 7450 ESS.

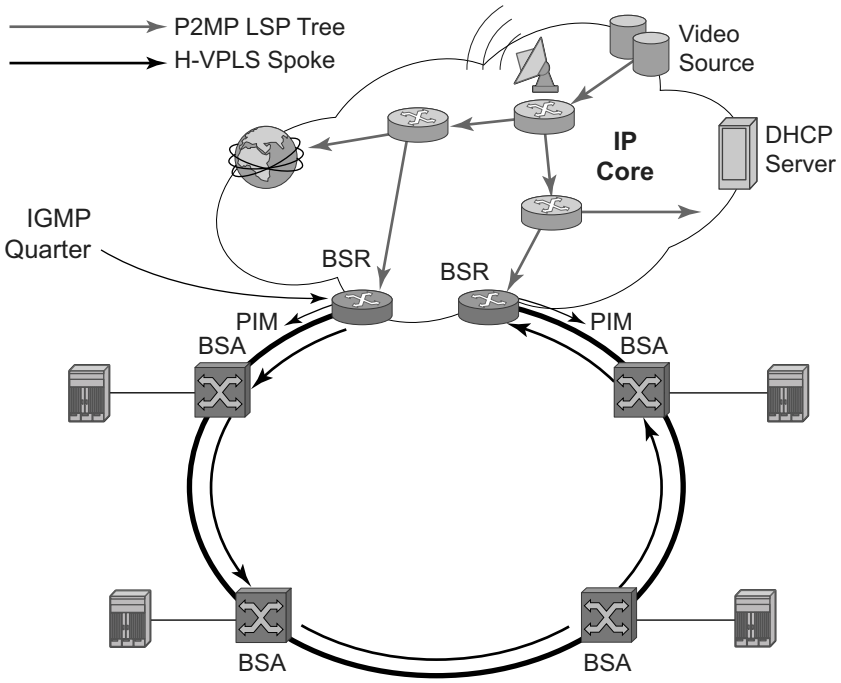


Figure 32: Application of P2MP LSP in Video Broadcast

A PIM-free core network can be achieved by deploying P2MP LSPs using other core routers. The router can act as the ingress LER receiving the multicast packets from the multicast source and forwarding them over the P2MP LSP.

A router can act as a leaf for the P2MP LSP tree initiated from the head-end router co-located with the video source. The router can also act as a branch node serving other leaf nodes and supports the replication of multicast packets over P2MP LSPs.

P2MP LSP Data Plane

A P2MP LSP is a unidirectional label switched path (LSP) which inserts packets at the root (ingress LER) and forwards the exact same replication of the packet to one or more leaf nodes (egress LER). The packet can be replicated at the root of P2MP LSP tree and/or at a transit LSR which acts as a branch node for the P2MP LSP tree.

Note that the data link layer code-point, for example Ethertype when Ethernet is the network port, continues to use the unicast codepoint defined in RFC 3032, *MPLS Label Stack Encoding*, and which is used on P2P LSP. This change is specified in draft-ietf-mpls-multicast-encaps, *MPLS Multicast Encapsulations*.

When a router sends a packet over a P2MP LSP which egresses on an Ethernet-based network interface, the Ethernet frame uses a MAC unicast destination address when sending the packet over the primary P2MP LSP instance or over a P2P bypass LSP). Note that a MAC multicast destination address is also allowed in the draft-ietf-mpls-multicast-encaps. Thus, at the ingress network interface on an Ethernet port, the router can accept both types of Ethernet destination addresses.

Procedures at Ingress LER Node

The following procedures occur at the root of the P2MP LSP (head-end or ingress LER node):

1. First, the P2MP LSP state is established via the control plane. Each leaf of the P2MP LSP will have a next-hop label forwarding entry (NHLFE) configured in the forwarding plane for each outgoing interface.

1. The user maps a specific multicast destination group address to the P2MP LSP in the base router instance by configuring a static multicast group under a tunnel interface representing the P2MP LSP.
2. An FTN entry is programmed at the ingress of the head-end node that maps the FEC of a received user IP multicast packet to a list of outgoing interfaces (OIF) and corresponding NHLFEs.
3. The head-end node replicates the received IP multicast packet to each NHLFE. Replication is performed at ingress toward the fabric and/or at egress forwarding engine depending on the location of the OIF.
4. At ingress, the head-end node performs a PUSH operation on each of the replicated packets.

Procedures at LSR Node

The following procedures occur at an LSR node that is not a branch node:

- The LSR performs a label swapping operation on a leaf of the P2MP LSP. This is a conventional operation of an LSR in a P2P LSP. An ILM entry is programmed at the ingress of the LSR to map an incoming label to a NHLFE.

The following is an exception handling procedure for control packets received on an ILM in an LSR.

- Packets that arrive with the TTL in the outer label expiring are sent to the CPM for further processing and are not forwarded to the egress NHLFE.

Procedures at Branch LSR Node

The following procedures occur at an LSR node that is a branch node:

- The LSR performs a replication and a label swapping for each leaf of the P2MP LSP. An ILM entry is programmed at the ingress of the LSR to map an incoming label to a list of OIF and corresponding NHLFEs.
- There is a limit of 127 OIF/NHLFEs per ILM entry.

The following is an exception handling procedure for control packets received on an ILM in a branch LSR:

- Packets that arrive with the TTL in the outer label expiring are sent to the CPM for further processing and not copied to the LSP branches.

Procedures at Egress LER Node

The following procedures occur at the leaf node of the P2MP LSP (egress LER):

- The egress LER performs a pop operation. An ILM entry is programmed at the ingress of the egress LER to map an incoming label to a list of next-hop/OIF.

The following is an exception handling procedure for control packets received on an ILM in an egress LER.

- The packet is sent to the CPM for further processing if there is any of the IP header exception handling conditions set after the label is popped: 127/8 destination address, router alert option set, or any other options set.

Procedures at BUD LSR Node

The following are procedures at an LSR node which is both a branch node and an egress leaf node (bud node):

- The bud LSR performs a pop operation on one or many replications of the received packet and a swap operation of the remaining replications. An ILM entry is programmed at ingress of the LSR to map the incoming label to list of NHLFE/OIF and next-hop/OIF.

  Note however, the exact same packets are replicated to an LSP leaf and to a local interface.

The following are the exception handling procedures for control packets received on an ILM in a bud LSR:

- Packets which arrive with the TTL in the outer label expiring are sent to the CPM and are not copied to the LSP branches.
- Packets whose TTL does not expire are copied to all branches of the LSP. The local copy of the packet is sent to the CPM for further processing if there is any of the IP header exception handling conditions set after the label is popped: 127/8 destination address, router alert option set, or any other options set.

Ingress Path Management for P2MP LSP Packets

The SR OS provides the ingress multicast path management (IMPM) capability that allows users to manage the way IP multicast streams are forwarded over the router's fabric and to maximize the use of the fabric multicast path capacity.

IMPM consists of two components, a bandwidth policy and a multicast information policy. The bandwidth policy configures the parameters of the multicast paths to the fabric. This includes the rate limit and the multicast queue parameters of each path. The multicast information policy configures the bandwidth and preference parameters of individual multicast flows corresponding to a channel, for example, a <*,G> or a <S,G>, or a bundle of channels.

By defaultboth, the IOM-2 and IOM-3 ingress data path provide two multicast paths through the fabric referred to as high-priority path and low-priority path respectively. When a multicast packet is received on an ingress network or access interface or on a VPLS SAP, the packet's classification will determine its forwarding class and priority or profile as per the ingress QoS policy. This then determines which of the SAP or interface multicast queues it must be stored in. By default SAP and interface expedited forwarding class queues forward over the high-priority multicast path and the non expedited forwarding class queues forward over the low-priority multicast path.

When IMPM on the ingress MDA is enabled, one or more multicast paths are enabled depending on the IOM type. In addition, multicast flows managed by IMPM will be stored in a separate shared multicast queue for each multicast path. These queues are configured in the bandwidth policy.

IMPM maps a packet to one of the paths dynamically based on monitoring the bandwidth usage of each packet flow matching a <*,G> or <S,G> record. The multicast bandwidth manager assigns multicast flows to a primary path, and ancillary path for IOM-2, based on the flow preference until the rate limits of each path is reached. At that point in time, a multicast flow is mapped to the secondary flow. If a path congests, the bandwidth manager will remove and black-hole lower preference flows to guarantee bandwidth to higher preference flows. The preference of a multicast flow is configured in the multicast info policy.

A packet received on a P2MP LSP ILM is managed by IMPM when IMPM is enabled on the ingress MDA and the packet matches a specific multicast record. When IMPM is enabled but the packet does not match a multicast record, or when IMPM is disabled, a packet received on a P2MP LSP ILM is mapped to a multicast path differently depending if the ingress IOM is an IOM-2 or IOM-3.

Ingress P2MP Path Management on IOM-3

On an ingress IOM-3, there are 16 multicast paths available to forward multicast packets. Each path has a set of multicast queues and associated with it. Paths 0 and 15 are enabled by default and represent the high-priority and low-priority paths respectively. Each VPLS SAP, access interface,

and network interface will have a set of per forwarding class multicast and/or broadcast queues which are defined in the ingress QoS policy associated with them. The expedited queues will be attached to Path 0 while the non-expedited queues will be attached to Path 15.

When IMPM is enabled and/or when a P2MP LSP ILM exists on the ingress IOM-3, the remaining 14 multicast paths are also enabled for a total of 16 paths. The first 15 paths are renamed as primary paths while the 16th path is renamed as a secondary path.

A separate pair of shared multicast queues is created on each of the 15 primary paths, one for IMPM managed packets and one for P2MP LPS packets not managed by IMPM. The secondary path does not forward IMPM managed packets or P2MP LSP packets. These queues have default rate (PIR=CIR) and CBS/MBS/Hi-Priority-Only thresholds but can be changed away from default under the bandwidth policy.

A VPLS snooped packet, a PIM routed packet, or a P2MP LSP packet is managed by IMPM if it matches a <*,G> or a <S,G> multicast record in the ingress forwarding table and IMPM is enabled on the ingress MDA where the packet is received. The user enables IMPM on the ingress MDA data path using the **config>card>mda>ingress>mcast-path-management** command.

A packet received on an IP interface and to be forwarded to a P2MP LSP NHLFE or a packet received on a P2MP LSP ILM is not managed by IMPM when IMPM is disabled on the ingress MDA where the packet is received or when IMPM is enabled but the packet does not match any multicast record. A P2MP LSP packet duplicated at a branch LSR node is an example of a packet not managed by IMPM even when IMPM is enabled on the ingress MDA where the P2MP LSP ILM exists. A packet forwarded over a P2MP LSP at an ingress LER and which matches a <*,G> or a <S<G> is an example of a packet which is not managed by IMPM if IMPM is disabled on the ingress MDA where the packet is received.

When a P2MP LSP packet is not managed by IMPM, it is stored in the unmanaged P2MP shared queue of one of the 15 primary multicast paths.

By default, non-managed P2MP LSP traffic is distributed across the IMPM primary paths using hash mechanisms. This can be optimized by enabling IMPM on any forwarding complex, which allows the system to redistributed this traffic on all forwarding complexes across the IMPM paths to achieve a more even capacity distribution. Be aware that enabling IMPM will cause routed and VPLS (IGMP and PIM) snooped IP multicast groups to be managed by IMPM.

The above ingress data path procedures apply to packets of a P2MP LSP at ingress LER, LSR, branch LSR, bud LSR, and egress LER. Note that in the presence of both IMPM managed traffic and unmanaged P2MP LSP traffic on the same ingress forwarding plane, the user must account for the presence of the unmanaged traffic on the same path when setting the rate limit for an IMPM path in the bandwidth policy.

Ingress P2MP Path Management on IOM-2

The following procedures apply at the ingress data path for packets received from or to be forwarded to a P2MP LSP at ingress LER, LSR, branch LSR, bud LSR, and egress LER.

On ingress IOM-2, there are 3 multicast paths which are available for forwarding multicast packets. Each path has a set of multicast queues and a multicast VoQ associated with it. Paths 0 and 2 are enabled by default and represent the high-priority and low-priority paths respectively. Each VPLS SAP, access interface, and network interface will have a set of per forwarding class multicast and/or broadcast queues which are defined in the ingress QoS policy associated with them. The expedited queues will be attached to Path 0 while the non-expedited queues will be attached to Path 2.

When IMPM is disabled, packets of P2MP LSP arriving on a network interface will be queued in that interface queue corresponding to the forwarding class of the packet.

When the user enables IMPM on the ingress MDA, a third multicast path, referred to as ancillary path, is added on the ingress IOM-2. This path reuses unused capacity from the unicast paths. The high-priority and low-priority paths are renamed as primary and secondary paths respectively.

A VPLS snooped packet or a PIM routed packet is managed by IMPM if it matches a <*,G> or a <S,G> multicast record in the ingress IOM-2 forwarding table and IMPM is enabled on the ingress MDA where the packet is received. The user enables IMPM on the ingress MDA data path using the **config>card>mda>ingress>mcast-path-management** command.

A P2MP LSP packet which matches a multicast record is also managed by IMPM on ingress IOM-2 and is thus distributed to one of the primary, ancillary, or secondary path according to the congestion level of the paths and the preference of the packet's multicast flow as configured in the multicast info policy 2.

RSVP Control Plane in a P2MP LSP

P2MP RSVP LSP is specified in RFC 4875, *Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)*.

A P2MP LSP is modeled as a set of root-to-leaf (S2L) sub-LSPs. The root, for example the head-end node, triggers signaling using one or multiple path messages. A path message can contain the signaling information for one or more S2L sub-LSPs. The leaf sub-LSP paths are merged at branching points.

A P2MP LSP is identified by the combination of <P2MP ID, tunnel ID, extended tunnel ID> part of the P2MP session object, and <tunnel sender address, LSP ID> fields in the P2MP sender_template object.

A specific sub-LSP is identified by the <S2L sub-LSP destination address> part of the S2L_SUB_LSP object and an ERO and secondary ERO (SERO) objects.

The following are characteristics of this feature:

1. Supports the de-aggregated method for signaling the P2MP RSVP LSP. Each root to leaf is modeled as a P2P LSP in the RSVP control plane. Only data plane merges the paths of the packets.

2. Each S2L sub-LSP is signaled in a separate path message. Each leaf node responds with its own resv message. A branch LSR node will forward the path message of each S2L sub-LSP to the downstream LSR without replicating it. It will also forward the resv message of each S2L sub-LSP to the upstream LSR without merging it with the resv messages of other S2L sub-LSPs of the same P2MP LSP. The same is done for subsequent refreshes of the path and resv states.
3. The node will drop aggregated RSVP messages on the receive side if originated by another vendor's implementation.
4. The user configures a P2MP LSP by specifying the optional create-time parameter **p2mp-lsp** following the LSP name. Next, the user creates a primary P2MP instance using the keyword **primary-p2mp-instance**. Then a path name of each S2L sub-LSP must added to the P2MP instance using the keyword **s2l-path**. The paths can be empty paths or can specify a list of explicit hops. The path name must exist and must have been defined in the **config>router>mpls>path** context.
5. The same path name can be re-used by more than one S2L of the primary P2MP instance. However the to keyword must have a unique argument per S2L as it corresponds to the address of the egress LER node.
6. The user can configure a secondary instance of the P2MP LSP to backup the primary one. In this case, the user enters the name of the secondary P2MP LSP instance under the same LSP name. One or more secondary instances can be created. The trigger for the head-end

node to switch the path of the LSP from the primary P2MP instance to the secondary P2MP instance is to be determined. This could be based on the number of leaf LSPs which went down at any given time.

7. The following parameters can be used with a P2MP LSP: adaptive, cspf, exclude, fast-reroute, from, hop-limit, include, metric, retry-limit, retry-timer, resignal-timer.
8. The following parameters cannot be used with a P2MP LSP: adspec, primary, secondary, to.
9. The node ingress LER will not inset an adspec object in the path message of an S2L sub-LSP. If received in the resv message, it will be dropped. The operational MTU of an S2L path is derived from the MTU of the outgoing interface of that S2L path.
10. The **to** parameter is not available at the LSP level but at the path level of each S2L sub-LSP of the primary or secondary instance of this P2MP LSP.
11. The hold-timer configured in the **config>router>mpls>hold-timer** context applies when signaling or re-signaling an individual S2L sub-LSP path. It does not apply when the entire tree is signaled or re-signaled.
12. The head-end node can add and/or remove a S2L sub-LSP of a specific leaf node without impacting forwarding over the already established S2L sub-LSPs of this P2MP LSP and without re-signaling them.
13. The head-end node performs a make-before break (MBB) on an individual S2L path of a primary P2MP instance whenever it applies the FRR global revertive procedures to this path. If CSPF finds a new path, RSVP signals this S2L path with the same LSP-ID as the existing path.
14. All other configuration changes, such as adaptive/no-adaptive, use-te-metric, no-frr, cspf/no-cspf, result in the tear-down and re-try of all affected S2L paths as is the case for P2P LSP paths.
15. MPLS requests CSPF to re-compute the whole set of S2L paths of a given active P2MP instance each time the P2MP re-signal timer expires. The P2MP re-signal timer is configured separately from the P2P LSP. MPLS performs a global MBB and moves each S2L sub-LSP in the instance into its new path using a new P2MP LSP ID if the global MBB is successful. This is regardless of the cost of the new S2L path.
16. MPLS will request CSPF to re-compute the whole set of S2L paths of a given active P2MP instance each time the user performs a manual re-signal of the P2MP instance. MPLS then always performs a global MBB and moves each S2L sub-LSP in the instance into its new path using a new P2MP LSP ID if the global MBB is successful. This is regardless of the cost of the new S2L path. The user executes a manual re-signal of the P2MP LSP instance using the command: **tools>perform>router>mpls>resignal p2mp-lsp** *lsp-name* **p2mp-instance** *instance-name*.
17. When performing global MBB, MPLS runs a separate MBB on each S2L in the P2MP LSP instance. If an S2L MBB does not succeed the first time, MPLS will re-try the S2L using the re-try timer and re-try count values inherited from P2MP LSP configuration.

However, there will be a global MBB timer set to 600 seconds and which is not configurable. If the global MBB succeeds, for example, all S2L MBBs have succeeded, before the global timer expires, MPLS moves the all S2L sub-LSPs into their new path. Otherwise when this timer expires, MPLS checks if all S2L paths have at least tried once. If so, it then aborts the global MBB. If not, it will continue until all S2Ls have re-tried once and then aborts the global MBB. Once global MBB is aborted, MPLS will move all S2L sub-LSPs into the new paths only if the set of S2Ls with a new path found is a superset of the S2Ls which have a current path which is up.

18. While make-before break is being performed on individual S2L sub-LSP paths, the P2MP LSP will continue forwarding packets on S2L sub-LSP paths which are not being re-optimized and on the older S2L sub-LSP paths for which make-before-break operation was not successful. MBB will thus result in duplication of packets until the old path is torn down.
19. The MPLS data path of an LSR node, branch LSR node, and bud LSR node will be able to re-merge S2L sub-LSP paths of the same P2MP LSP in case their ILM is on different incoming interfaces and their NHLFE is on the same or different outgoing interfaces. This could occur anytime there are equal cost paths through this node for the S2L sub-LSPs of this P2MP LSP.
20. Link-protect FRR bypass using P2P LSPs is supported. In link protect, the PLR protecting an interface to a branch LSR will only make use of a single P2P bypass LSP to protect all S2L sub-LSPs traversing the protected interface.
21. Refresh reduction on RSVP interface and on P2P bypass LSP protecting one or more S2L sub-LSPs.
22. A manual bypass LSP cannot be used for protecting S2L paths of a P2MP LSP.
23. The following MPLS features do operate with P2MP LSP:
    - BFD on RSVP interface.
    - MD5 on RSVP interface.
    - IGP metric and TE metric for computing the path of the P2MP LSP with CSPF.
    - SRLG constraint for computing the path of the P2MP LSP with CSPF. SRLG is supported on FRR backup path only.
    - TE graceful shutdown.
    - Admin group constraint.
24. The following MPLS features are not operable with P2MP LSP:
    - Class based forwarding over P2MP RSVP LSP.
    - LDP-over-RSVP where the RSVP LSP is a P2MP LSP.
    - Diff-Serv TE.
    - Soft pre-emption of RSVP P2MP LSP.

Forwarding Multicast Packets over RSVP P2MP LSP in the Base Router

Multicast packets are forwarded over the P2MP LSP at the ingress LER based on a static join configuration of the multicast group against the tunnel interface associated with the originating P2MP LSP. At the egress LER, packets of a multicast group are received from the P2MP LSP via a static assignment of the specific <S,G> to the tunnel interface associated with a terminating LSP.

Procedures at Ingress LER Node

The forwarding of multicast packets over a P2MP LSP follows the following procedures:

1. The user creates a tunnel interface associated with the P2MP LSP: **configure>router>tunnel-interface rsvp-p2mp** *lsp-name*. The configure>router>pim>tunnel-interface command has been discontinued.

2. The user adds static multicast group joins to the PIM interface, either as a specific <S,G> or as a <*,G>: **configure>router>igmp>tunnel-interface>static>group>source** *ip-address* and **configure>router>igmp>tunnel-interface>static>group>starg**.

The tunnel interface identifier consists of a string of characters representing the LSP name for the RSVP P2MP LSP. Note that MPLS will actually pass to PIM a more structured tunnel interface identifier. The structure will follow the one BGP uses to distribute the PMSI tunnel information in BGP multicast VPN as specified in draft-ietf-l3vpn-2547bis-mcast-bgp, *Multicast in MPLS/BGP IP VPNs*.The format is: <extended tunnel ID, reserved, tunnel ID, P2MP ID> as encoded in the RSVP-TE P2MP LSP session_attribute object in RFC 4875.

The user can create one or more tunnel interfaces in PIM and associate each to a different RSVP P2MP LSP. The user can then assign static multicast group joins to each tunnel interface. Note however that a given <*,G> or <S,G> can only be associated with a single tunnel interface.

A multicast packet which is received on an interface and which succeeds the RPF check for the source address will be replicated and forwarded to all OIFs which correspond to the branches of the P2MP LSP. The packet is sent on each OIF with the label stack indicated in the NHLFE of this OIF. The packets will also be replicated and forwarded natively on all OIFs which have received IGMP or PIM joins for this <S,G>.

The multicast packet can be received over a PIM or IGMP interface which can be an IES interface, a spoke SDP-terminated IES interface, or a network interface.

In order to duplicate a packet for a multicast group over the OIF of both P2MP LSP branches and the regular PIM or IGMP interfaces, the tap mask for the P2MP LSP and that of the PIM based interfaces will need to be combined into a superset MCID.

Procedures at Egress LER Node

Procedures with a Primary Tunnel Interface

The user configures a tunnel interface and associates it with a terminating P2MP LSP leaf using the command: **config>router>tunnel-interface rsvp-p2mp lsp-name sender** *sender-address*. The **configure>router>pim>tunnel-interface** command has been discontinued.

The tunnel interface identifier consists of a couple of string of characters representing the LSP name for the RSVP P2MP LSP followed by the system address of the ingress LER. The LSP name must correspond to a P2MP LSP name configured by the user at the ingress LER and must not contain the special character ":" Note that MPLS will actually pass to PIM a more structured tunnel interface identifier. The structure will follow the one BGP uses to distribute the PMSI tunnel information in BGP multicast VPN as specified in draft-ietf-l3vpn-2547bis-mcast-bgp.The format is: <extended tunnel ID, reserved, tunnel ID, P2MP ID> as encoded in the RSVP-TE P2MP LSP session_attribute object in RFC 4875.

The egress LER accepts multicast packets the following methods:

1. The regular RPF check on unlabeled IP multicast packets, which is based on routing table lookup.

2. The static assignment which specifies the receiving of a multicast group <*,G> or a specific <S,G> from a primary tunnel-interface associated with an RSVP P2MP LSP.

One or more primary tunnel interfaces in the base router instance can be configured. In other words, the user will be able to receive different multicast groups, <*,G> or specific <S,G>, from different P2MP LSPs. This assumes that the user configured static joins for the same multicast groups at the ingress LER to forward over a tunnel interface associated with the same P2MP LSP.

A multicast info policy CLI option allows the user to define a bundle and specify channels in the bundle that must be received from the primary tunnel interface. The user can apply the defined multicast info policy to the base router instance.

At any given time, packets of the same multicast group can be accepted from either the primary tunnel interface associated with a P2MP LSP or from a PIM interface. These are mutually exclusive options. As soon as a multicast group is configured against a primary tunnel interface in the multicast info policy, it is blocked from other PIM interfaces.

However, if the user configured a multicast group to be received from a given primary tunnel interface, there is nothing preventing packets of the same multicast group from being received and accepted from another primary tunnel interface. However, an ingress LER will not allow the same multicast group to be forwarded over two different P2MP LSPs. The only possible case is that of two ingress LERs forwarding the same multicast group over two P2MP LSPs towards the same egress LER.

A multicast packet received on a tunnel interface associated with a P2MP LSP can be forwarded over a PIM or IGMP interface which can be an IES interface, a spoke SDP terminated IES interface, or a network interface.

Note that packets received from a primary tunnel-interface associated with a terminating P2MP LSP cannot be forwarded over a tunnel interface associated with an originating P2MP LSP.

MPLS Service Usage

Alcatel-Lucent routers enable service providers to deliver virtual private networks (VPNs) and Internet access using Generic Routing Encapsulation (GRE) and/or MPLS tunnels, with Ethernet and/or SONET/SDH interfaces.

Service Distribution Paths

A service distribution path (SDP) acts as a logical way of directing traffic from one router to another through a uni-directional (one-way) service tunnel. The SDP terminates at the far-end router which directs packets to the correct service egress service access point (SAP) on that device. All services mapped to an SDP use the same transport encapsulation type defined for the SDP (either GRE or MPLS).

For information about service transport tunnels, refer to the Service Distribution Paths (SDPs) section in the OS Services Guide. They can support up to eight forwarding classes and can be used by multiple services. Multiple LSPs with the same destination can be used to load-balance traffic.

MPLS/RSVP Configuration Process Overview

Figure 33 displays the process to configure MPLS and RSVP parameters.

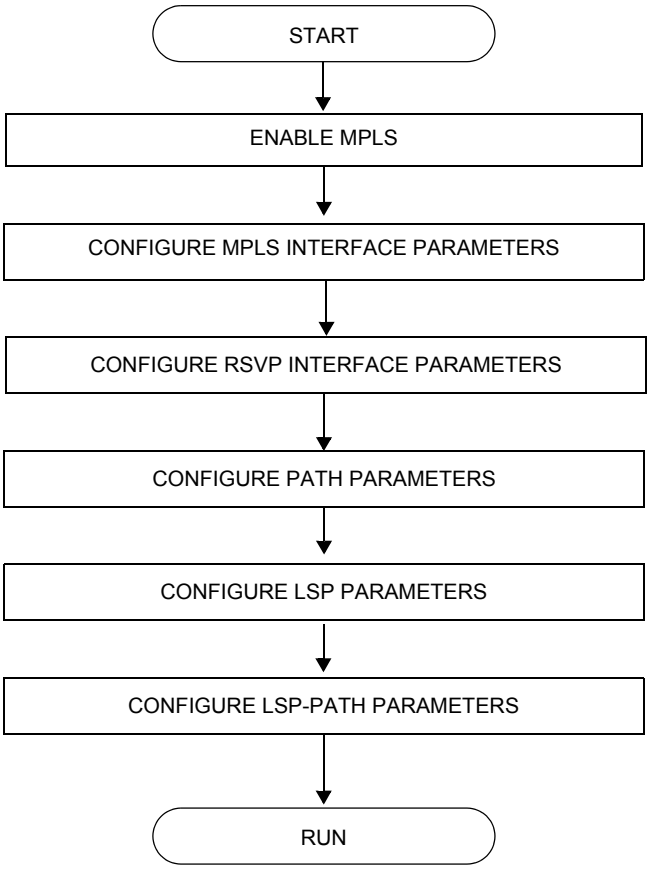


Figure 33: MPLS and RSVP Configuration and Implementation Flow

Configuration Notes

This section describes MPLS and RSVP caveats.

- Interfaces must already be configured in the `config>router>interface` context before they can be specified in MPLS and RSVP.
- A router interface must be specified in the `config>router>mpls` context in order to apply it or modify parameters in the `config>router>rsvp` context.
- A system interface must be configured and specified in the `config>router>mpls` context.
- Paths must be created before they can be applied to an LSP.