

Chapter 3

MPLS Overview

Multiprotocol Label Switching (MPLS) provides a mechanism for engineering network traffic patterns that is independent of routing tables. MPLS assigns short labels to network packets that describe how to forward them through the network. MPLS is independent of any routing protocol and can be used for unicast packets.

In the traditional Level 3 forwarding paradigm, as a packet travels from one router to the next, an independent forwarding decision is made at each hop. The IP network layer header is analyzed, and the next hop is chosen based on this analysis and on the information in the routing table. In an MPLS environment, the analysis of the packet header is performed just once, when a packet enters the MPLS cloud. The packet is then assigned to a stream, which is identified by a *label*, which is a short (20-bit), fixed-length value at the front of the packet. Labels are used as lookup indexes for the label forwarding table. For each label, this table stores forwarding information. You can associate additional information with a label—such as class-of-service (CoS) values—that can be used to prioritize packet forwarding.

This chapter discusses the following topics:

- MPLS Standards on page 24
- Link-Layer Support on page 25
- MPLS and Traffic Engineering on page 25
- IP and MPLS Packets on Aggregated Interfaces on page 42
- MPLS Applications on page 43
- MPLS and Routing Tables on page 46
- MPLS and Traffic Protection on page 48
- Automatic Bandwidth Allocation on page 53
- Point-to-Multipoint LSPs on page 54
- MPLS Load Balancing Based on the IP Header and MPLS Labels on page 55

MPLS Standards

The JUNOS software supports the following RFCs and Internet drafts related to MPLS:

- RFC 3031, *Multiprotocol Label Switching Architecture* (provides a good overview of MPLS)
- RFC 3032, *MPLS Label Stack Encoding*
- Internet draft draft-ietf-mpls-icmp-02.txt, *ICMP Extensions for Multiprotocol Label Switching* (expires February 2001)
- Internet draft draft-ietf-mpls-lsp-ping-06.txt, *Detecting MPLS Data Plane Failures* (only the LDP IPv4 prefix TLV, RSVP IPv4 Session Query TLV, and VPN IPv4 prefix TLV) (no expiration; revised July 2004)
- Internet draft draft-ietf-mpls-soft-preemption-02.txt, *MPLS Traffic Engineering Soft Preemption* (no expiration; revised March 2004)
- Internet draft draft-raggarwa-mpls-p2mp-te-02.txt (except non-adjacent signaling for sub-LSPs, make-before-break and fast reroute, and LSP hierarchy using P2P LSPs), *Establishing Point to Multipoint MPLS TE LSPs* (no expiration; revised January 2004)
- Internet draft draft-ietf-mpls-p2mp-requirement-01.txt, *Requirements for Point to Multipoint Extensions to RSVP-TE* (no expiration; revised January 2004)

The following documents provide information about traffic engineering:

- RFC 2702, *Requirements for Traffic Engineering Over MPLS*
- Internet draft draft-ietf-isis-traffic-04.txt, *IS-IS Extensions for Traffic Engineering* (expires February 2002)
- Internet draft draft-ietf-tewg-diff-te-mam-03.txt, *Maximum Allocation Bandwidth Constraints Model for Diff-Serv-aware MPLS Traffic Engineering* (expires September 2004)
- Internet draft draft-katz-yeung-ospf-traffic-06.txt, *Traffic Engineering Extensions to OSPF* (expires April 2002)
- Internet draft draft-ietf-mpls-rsvp-lsp-fastreroute-03.txt, *Fast Reroute Extensions to RSVP-TE for LSP Tunnels* (expires June 2003)
- Internet draft draft-ietf-mpls-rsvp-te-p2mp-01.txt, *Extensions to RSVP-TE for Point to Multipoint TE LSPs* (expires June 2005)

To access Internet RFCs and drafts, go to the IETF Web site at <http://www.ietf.org>.

The JUNOS software supports a proprietary Management Information Base (MIB) for MPLS objects; see the *JUNOS Network Management Configuration Guide* for more information.

Link-Layer Support

MPLS supports the following link-layer protocols, which are all supported in the JUNOS MPLS implementation:

- Point-to-Point Protocol (PPP)—Protocol ID 0x0281, Network Control Protocol (NCP) protocol ID 0x8281.
- Ethernet/Cisco High-level Data Link Control (HDLC)—Ethernet type 0x8847.
- Asynchronous Transfer Mode (ATM)—Subnetwork attachment point encoded (SNAP-encoded) Ethernet type 0x8847. Support is included for both point-to-point mode or nonbroadcast multiaccess (NBMA) mode. Support is not included for encoding MPLS labels as part of ATM virtual path identifier/virtual circuit identifier (VPI/VCI).
- Frame Relay—SNAP-encoded, Ethernet type 0x8847. Support is not included for encoding MPLS labels as part of Frame Relay data-link connection identifier (DLCI).
- Generic routing encapsulation (GRE) tunnel—Ethernet type 0x8847.

MPLS and Traffic Engineering

Traffic engineering allows you to control the path that data packets follow, bypassing the standard routing model, which uses routing tables. Traffic engineering moves flows from congested links to alternate links that would not be selected by the automatically computed destination-based shortest path. With traffic engineering, you can:

- Make more efficient use of expensive long-haul fibers.
- Control how traffic is rerouted in the face of single or multiple failures.
- Classify critical and regular traffic on a per-path basis.

The core of the traffic engineering design is based on building label-switched paths (LSPs) among routers. An LSP is connection-oriented, like a virtual circuit in Frame Relay or ATM. LSPs are not reliable: Packets entering an LSP do not have delivery guarantees, although preferential treatment is possible. LSPs also are similar to unidirectional tunnels in that packets entering a path are encapsulated in an envelope and switched across the entire path without being touched by intermediate nodes. LSPs provide fine-grained control over how packets are forwarded in a network. To provide reliability, an LSP can use a set of primary and secondary paths.

LSPs can be configured for Border Gateway Protocol (BGP) traffic only (traffic whose destination is outside of an autonomous system [AS]). In this case, traffic within the AS is not affected by the presence of LSPs. LSPs can also be configured for both BGP and interior gateway protocol (IGP) traffic; therefore, both intra-AS and inter-AS traffic is affected by the LSPs.

This section discusses the following topics:

- Label Description on page 26
- Label Allocation on page 28
- Routers in an LSP on page 30
- How a Packet Travels Along an LSP on page 30
- Types of LSPs on page 31
- Scope of LSPs on page 31
- Constrained-Path LSP Computation on page 32
- LSPs on an Overloaded Router on page 35
- Fate Sharing on page 36
- IGP Shortcuts on page 37
- Advertising LSPs into IGPs on page 41

Label Description

Packets travelling along an LSP are identified by a *label*—a 20-bit, unsigned integer in the range 0 through 1,048,575:

- 0 through 15—Reserved and have special semantics.
- 16 through 1023—Used by the JUNOS software for `vrf-table-label` statement operations.
- 10,000 through 99,999—Unused and unassigned by the software, a feature that is specific to the JUNOS software. You can use these labels to manually configure static LSPs and to ensure that there are no conflicts with labels that are dynamically assigned by the software.
- 1024 through 9999—Reserved for future applications.
- 100,000 through 1,048,575—Automatically negotiated, assigned, released, and reused by the software. Typically, per-box labels are assigned in the 100,000 through 799,999 range, and per-interface labels are assigned in the 800,000 through 1,048,575 range.

Special Labels

Some of the reserved labels (in the 0 through 15 range) have well-defined meanings. For more complete details, see RFC 3032, *MPLS Label Stack Encoding*.

- 0, IPv4 Explicit Null label—This value is legal only when it is the sole label entry (no label stacking). It indicates that the label must be popped upon receipt. Forwarding continues based on the IP version 4 (IPv4) packet.
- 1, Router Alert label—When a packet is received with a top label value of 1, it is delivered to the local software module for processing.
- 2, IPv6 Explicit Null label—This value is legal only when it is the sole label entry (no label stacking). It indicates that the label must be popped on receipt. Forwarding continues based on the IP version 6 (IPv6) packet.
- 3, Implicit Null label—This label is used in the control protocol (Label Distribution Protocol [LDP] or Resource Reservation Protocol [RSVP]) only to request label popping by the downstream router. It never actually appears in the encapsulation. Labels with a value of 3 should not be used in the data packet as real labels. No payload type (IPv4 or IPv6) is implied with this label.
- 4 through 15—Unassigned.

Special labels are commonly used between the egress and penultimate routers of an LSP. If the LSP is configured to carry IPv4 packets only, the egress router might signal the penultimate router to use 0 as a final-hop label. If the LSP is configured to carry IPv6 packets only, the egress router might signal the penultimate router to use 2 as a final-hop label.

The egress router might simply signal the penultimate router to use 3 as the final label, which is a request to perform penultimate-hop label popping. The egress router will not process a labelled packet; rather, it receives the payload (IPv4, IPv6, or others) directly, reducing one MPLS lookup at egress.

For label-stacked packets, the egress router receives an MPLS label packet with its top label already popped by the penultimate router. The egress router cannot receive label-stacked packets that use label 0 or 2. It typically requests label 3 from the penultimate router.

Label Allocation

In the JUNOS software, label values are allocated per router. The display output shows only the label (for example, 01024). Labels for multicast packets are independent of those for unicast packets. Currently, the JUNOS software does not support multicast labels.

Labels are assigned by downstream routers relative to the flow of packets. A router receiving labeled packets (the next-hop router) is responsible for assigning incoming labels. A received packet containing a label that is unrecognized (unassigned) is dropped. For unrecognized labels, the router does not attempt to unwrap the label to analyze the network layer header, nor does it generate an Internet Control Message Protocol (ICMP) destination unreachable message.

A packet can carry a number of labels, organized as a last-in, first-out stack. This is referred to as a *label stack*. At a particular router, the decision about how to forward a labeled packet is based exclusively on the label at the top of the stack.

Figure 1 shows the encoding of a single label. The encoding appears after data link layer headers, but before any network layer header.

Figure 1: Label Encoding

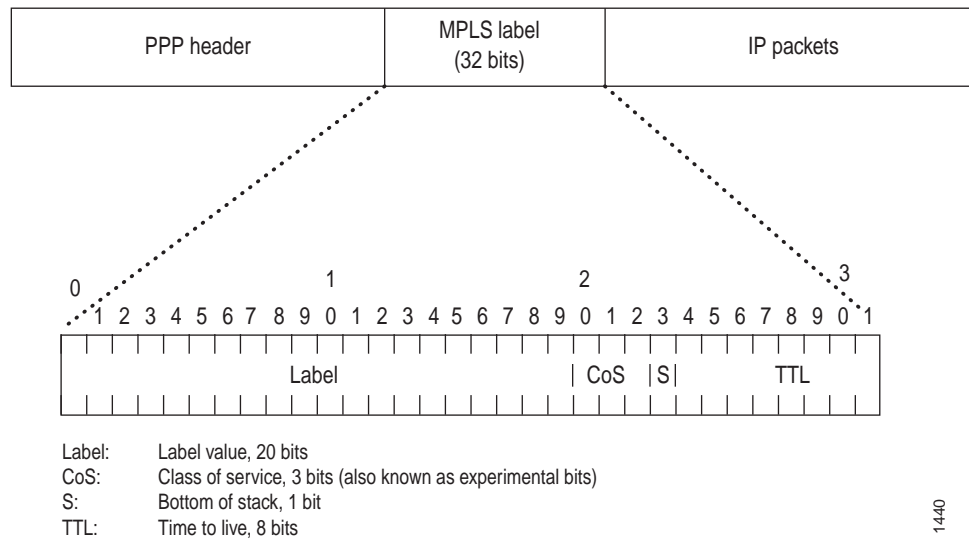
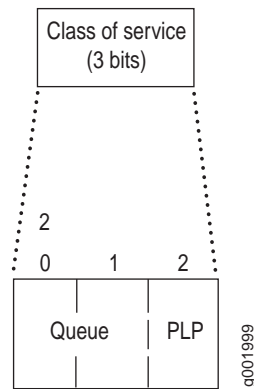


Figure 2 illustrates the purpose of the class-of-service bits (also known as the EXP or experimental bits). Bits 20 and 21 specify the queue number. Bit 22 is the packet loss priority (PLP) bit used to specify the random early detection (RED) drop profile. For more information about class of service and the class-of-service bits, see “Configuring Class of Service for MPLS” on page 95.

Figure 2: Class-of-Service Bits

Operations on Labels

The router supports the following label operations:

- **Push**—Add a new label to the top of the packet. For IPv4 packets, the new label is the first label. The TTL and S bits are derived from the IP packet header. The MPLS CoS is derived from the queue number. If the push operation is performed on an existing MPLS packet, you will have a packet with two or more labels. This is called label stacking. The top label must have its S bit set to 0, and might derive CoS and time to live (TTL) from lower levels. The new top label in a label stack always initializes its TTL to 255, regardless of the TTL value of lower labels.
- **Pop**—Remove the label from the beginning of the packet. Once the label is removed, the TTL is copied from the label into the IP packet header, and the underlying IP packet is forwarded as a native IP packet. In the case of multiple labels in a packet (label stacking), removal of the top label yields another MPLS packet. The new top label might derive CoS and TTL from a previous top label. The popped TTL value from the previous top label is not written back to the new top label.
- **Swap**—Replace the label at the top of the label stack with a new label. The S and CoS bits are copied from the previous label, and the TTL value is copied and decremented (unless the `no-decrement-ttl` or `no-propagate-ttl` statements are configured). A transit router supports a label stack of any depth.
- **Multiple Push**—Add multiple labels (up to three) on top of existing packets. This operation is equivalent to pushing multiple times.
- **Swap and Push**—Replace the existing top of the label stack with a new label, and then push another new label on top.

Routers in an LSP

Each router in an LSP performs one of the following functions:

- **Ingress router**—The router at the beginning of an LSP. This router encapsulates IP packets with an MPLS Layer 2 frame and forwards it to the next router in the path. Each LSP can have only one ingress router.
- **Egress router**—The router at the end of an LSP. This router removes the MPLS encapsulation, thus transforming it from an MPLS packet to an IP packet, and forwards the packet to its final destination using information in the IP forwarding table. Each LSP can have only one egress router. The ingress and egress routers in an LSP cannot be the same router.
- **Transit router**—Any intermediate router in the LSP between the ingress and egress routers. A transit router forwards received MPLS packets to the next router in the MPLS path. An LSP can contain zero or more transit routers, up to a maximum of 253 transit routers in a single LSP.

A single router can be part of multiple LSPs. It can be the ingress or egress router for one or more LSPs, and it also can be a transit router in one or more LSPs. The functions that each router supports depend on your network design.

How a Packet Travels Along an LSP

When an IP packet enters an LSP, the ingress router examines the packet and assigns it a label based on its destination, placing the label in the packet's header. The label transforms the packet from one that is forwarded based on its IP routing information to one that is forwarded based on information associated with the label.

The packet is then forwarded to the next router in the LSP. This router and all subsequent routers in the LSP do not examine any of the IP routing information in the labeled packet. Rather, they use the label to look up information in their label forwarding table. They then replace the old label with a new label and forward the packet to the next router in the path.

When the packet reaches the egress router, the label is removed, and the packet again becomes a native IP packet and is again forwarded based on its IP routing information.

Types of LSPs

There are three types of LSPs:

- Static LSPs—For static paths, you must manually assign labels on all routers involved (ingress, transit, and egress). No signaling protocol is needed. This procedure is similar to configuring static routes on individual routers. Like static routes, there is no error reporting, liveness detection, or statistics reporting.
- LDP-signaled LSPs—See “LDP Overview” on page 329.
- RSVP-signaled LSPs—For signaled paths, RSVP is used to set up the path and dynamically assign labels. (RSVP signaling messages are used to set up signaled paths.) You configure only the ingress router. The transit and egress routers accept signaling information from the ingress router, and they set up and maintain the LSP cooperatively. Any errors encountered while establishing an LSP are reported to the ingress router for diagnostics. For signaled LSPs to work, a version of RSVP that supports tunnel extensions must be enabled on all routers.

There are two types of RSVP-signaled LSPs:

- Explicit-path LSPs—All intermediate hops of the LSP are manually configured. The intermediate hops can be strict, loose, or any combination of the two. Explicit path LSPs provide you with complete control over how the path is set up. They are similar to static LSPs but require much less configuration.
- Constrained-path LSPs—The intermediate hops of the LSP are automatically computed by the software. The computation takes into account information provided by the topology information from the Intermediate System-to-Intermediate System (IS-IS) or Open Shortest Path First (OSPF) link-state routing protocol, the current network resource utilization determined by RSVP, and the resource requirements and constraints of the LSP. For signaled constrained-path LSPs to work, either the IS-IS or OSPF protocol and the IS-IS or OSPF traffic engineering extensions must be enabled on all routers.

Scope of LSPs

For constrained-path LSPs, the LSP computation is confined to one IGP domain, and cannot cross any AS boundary. This prevents an AS from extending its IGP into another AS.

Explicit-path LSPs, however, can cross as many AS boundaries as necessary. Because intermediate hops are manually specified, the LSP does not depend on the IGP topology or a local forwarding table.

Constrained-Path LSP Computation

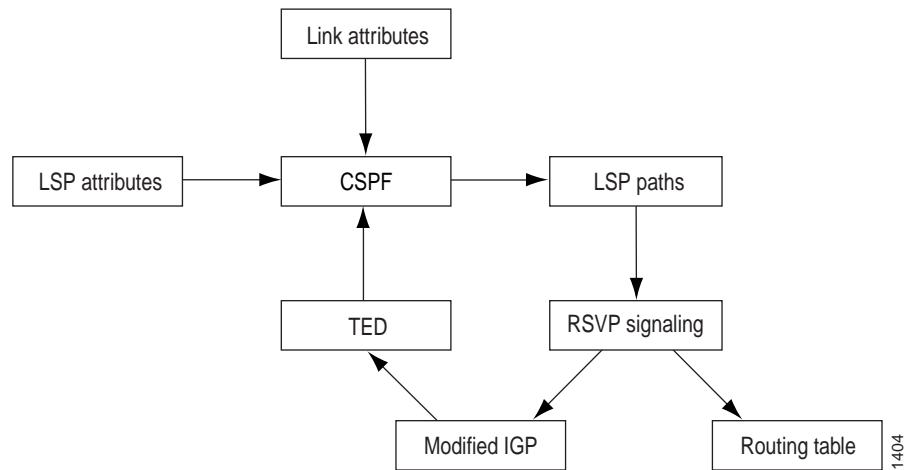
The Constrained Shortest Path First (CSPF) algorithm is an advanced form of the shortest-path-first (SPF) algorithm used in OSPF and IS-IS route computations. CSPF is used in computing paths for LSPs that are subject to multiple constraints. When computing paths for LSPs, CSPF considers not only the topology of the network, but also the attributes of the LSP and the links, and it attempts to minimize congestion by intelligently balancing the network load.

The constraints that CSPF considers include:

- LSP attributes
 - Administrative groups (that is, link color requirements)
 - Bandwidth requirements
 - Explicit route (strict or loose)
 - Hop limitations
 - Priority (setup and hold)
- Link attributes
 - Administrative groups (that is, link colors assigned to the link)
 - Reservable bandwidth of the links (static bandwidth minus the currently reserved bandwidth)

The data that CSPF considers comes from the following sources:

- Traffic engineering database (TED)—Provides CSPF with up-to-date topology information, the current reservable bandwidth of links, and the link colors. For the CSPF algorithm to perform its computations, a link-state IGP (such as OSPF or IS-IS) with special extensions is needed. For CSPF to be effective, the link-state IGP on all routers must support the special extensions. While building the topology database, the extended IGP must take into consideration the current LSPs and must flood the route information everywhere. Because changes in the reserved link bandwidth and link color cause database updates, an extended IGP tends to flood more frequently than a normal IGP. See Figure 3 for a diagram of the relationships between these components.
- Currently active LSPs—Includes all the LSPs that should originate from the router and their current operational status (up, down, or timeout).

Figure 3: CSPF Computation Process

This section discusses the following topics:

- How CSPF Selects a Path on page 33
- Path Selection Tie-Breaking on page 34
- Computing Paths Offline on page 35

How CSPF Selects a Path

To select a path, CSPF follows these steps:

1. Computes LSPs one at a time, beginning with the highest priority LSP (the one with the lowest setup priority value). Among LSPs of equal priority, CSPF starts with those that have the highest bandwidth requirement.
2. Prunes the TED of all the links that are not full duplex and do not have sufficient reservable bandwidth.
3. If the LSP configuration includes the `include` statement, prunes all links that do not share any included colors.
4. If the LSP configuration includes the `exclude` statement, prunes all links that contain excluded colors. If the link does not have a color, it is accepted.
5. Finds the shortest path toward the LSP's egress router, taking into account explicit-path constraints. For example, if the path must pass through Router A, two separate SPF's are computed, one from the ingress router to Router A, the other from Router A to the egress router.

6. If several paths have equal cost, chooses the one whose last-hop address is the same as the LSP's destination.
7. If several equal-cost paths remain, selects the one with the fewest number of hops.
8. If several equal-cost paths remain, applies the CSPF load-balancing rule configured on the LSP (least fill, most fill, or random).

Path Selection Tie-Breaking

If more than one path is available after the rules from the previous section have been applied, a tie-breaking rule is applied to choose the path for the LSP. There are three tie-breaking rules:

- Random—One of the remaining paths is picked at random. This rule tends to place an equal number of LSPs on each link, regardless of the available bandwidth ratio.
- Least fill—The path with the largest minimum available bandwidth ratio is preferred. This rule tries to equalize the reservation on each link.
- Most fill—The path with the smallest minimum available bandwidth ratio is preferred. This rule tries to fill a link before moving on to alternative links.

The rule used depends on the configuration. Random is the default rule.

For the other rules, the following definitions are needed:

- Reservable bandwidth = bandwidth of link x subscription factor of link
- Available bandwidth = reservable bandwidth – (sum of the bandwidths of the LSPs traversing the link)
- Available bandwidth ratio = available bandwidth/reservable bandwidth
- Minimum available bandwidth ratio (for a path) = the smallest available bandwidth ratio of the links in a path

Computing Paths Offline

The JUNOS software provides online, real-time CSPF computation only; each router performs CSPF calculations independent of the other routers in the network. These calculations are based on currently available topology information—information that is usually recent, but not completely accurate. LSP placements are locally optimized, based on current network status.

To optimize links globally across the network, you can use an offline tool to perform the CSPF calculations and determine the paths for the LSPs. You can create such a tool yourself, or you can modify an existing network design tool to perform these calculations. You should run the tool periodically (daily or weekly) and download the results into the router. An offline tool should take the following into account when performing the optimized calculations:

- All the LSP's requirements
- All link attributes
- Complete network topology

LSPs on an Overloaded Router

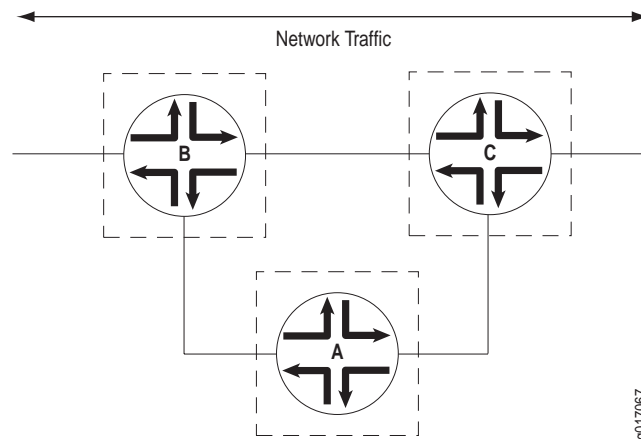
An overloaded router is a router running IS-IS with its overload bit set in its IS-IS configuration. In this case, an MPLS LSP specifically refers to an RSVP- or LDP-signaled LSP. In the case of RSVP, it applies to both CSPF and non-CSPF LSPs.

You cannot establish transit LSPs through an overloaded router. However, you can configure ingress and egress LSPs through an overloaded router.



NOTE: When you set the overload bit on an IS-IS router, all LSPs transiting through it are recomputed and rerouted away from it. If the recomputation fails, no additional attempt to reconfigure the LSP is made, and the affected LSPs are disconnected.

An example of when you might want to establish transit LSPs through an overloaded router is illustrated in Figure 4 on page 36, which shows an aggregation router (Router A) dual-homed on two core routers (Router B and Router C). You want to include the aggregation router in the LSP mesh, but transit LSPs should not pass through it, because it is a less capable router with relatively low-bandwidth uplinks to the core. Certain failure and rerouting scenarios could make it impossible for the aggregation router to establish some of its LSPs. Consequently, you run the router in a steady state with the overload bit set, but you are still able to establish ingress and egress LSPs through it.

Figure 4: Aggregation Router A Dual-Homed on Core Routers B and C

Fate Sharing

Fate sharing allows you to create a database of information that CSPF uses to compute one or more backup paths to use in case the primary path becomes unstable. The database describes the relationships between elements of the network, such as routers and links. You can specify one or more elements within a group.

Through fate sharing, you can configure backup paths that minimize the number of shared links and fiber paths with the primary paths as much as possible, to ensure that if a fiber is cut, the minimum amount of data is lost and a path still exists to the destination.

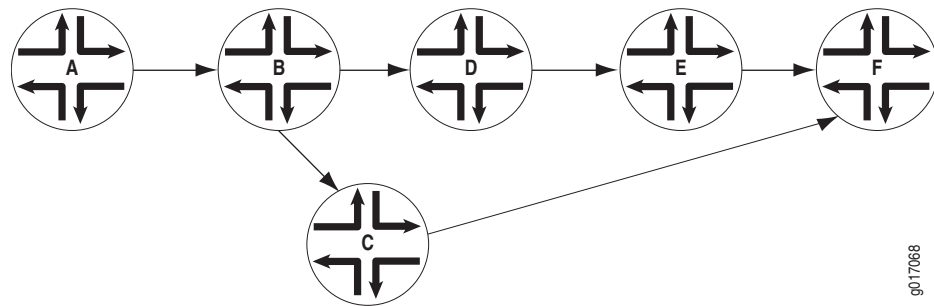
For a backup path to work optimally, it must not share links or physical fiber paths with the primary path, ensuring that a single point of failure will not affect the primary and backup paths simultaneously. For more information about fate sharing, see the *JUNOS Routing Protocols Configuration Guide*.

IGP Shortcuts

Link-state protocols, such as OSPF and IS-IS, use the SPF algorithm to compute the shortest-path tree to all nodes in the network. The results of such computations can be represented by the destination node, next-hop address, and output interface, where the output interface is a physical interface. LSPs can be used to augment the SPF algorithm, for the purposes of resolving BGP next hops. On the node performing the calculations, LSPs appear to be logical interfaces directly connected to remote nodes in the network. If you configure the IGP to treat LSPs the same as a physical interface and use the LSPs as a potential output interface, the SPF computation results are represented by the destination node and output LSP, effectively using the LSP as a shortcut through the network to the destination.

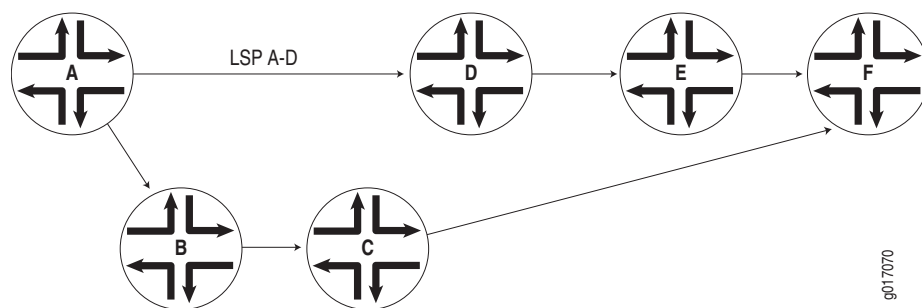
As an illustration, begin with a typical SPF tree (see Figure 5).

Figure 5: Typical SPF Tree, Sourced from Router A



If an LSP connects Router A to Router D and if IGP shortcuts are enabled on Router A, you might have the SPF tree shown in Figure 6.

Figure 6: Modified SPF Tree, Using LSP A-D as a Shortcut



Note that Router D is now reachable through LSP A-D. When computing the shortest path to reach Router D, Router A has two choices:

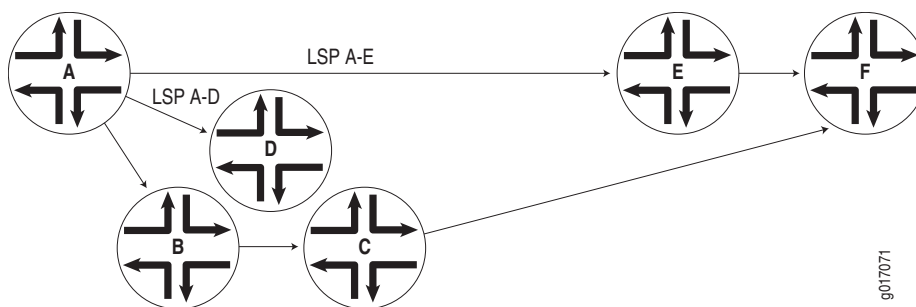
- Use IGP path A-B-D.
- Use LSP A-D.

Router A decides between the two choices by comparing the IGP metrics for path A-B-D with the LSP metrics for LSP A-D. If the IGP metric is lower, path A-B-D is chosen (Figure 5). If the LSP metric is lower, LSP A-D is used (Figure 6). If both metrics are equal, LSP A-D is chosen because LSP paths are preferred over IGP paths.

Note that Routers E and F are also reachable through LSP A-D, because they are downstream from Router D in the SPF tree.

Assuming that another LSP connects Router A to Router E, you might have the SPF tree shown in Figure 7.

Figure 7: Modified SPF Tree, Using LSP A-D and LSP A-E as Shortcuts



Enable IGP Shortcuts

IGP shortcuts are supported for both IS-IS and OSPF. A link-state protocol is required for IGP shortcuts. Shortcuts are disabled by default. For information about enabling IGP shortcuts for IS-IS and OSPF, see the *JUNOS Routing Protocols Configuration Guide*. You can enable IGP shortcuts on a per-router basis; you do not need to enable shortcuts globally. A router’s shortcut computation does not depend on another router performing similar computations, and shortcuts performed by other routers are irrelevant.

LSPs Qualified in Shortcut Computations

Not all LSPs are used in IGP shortcuts. Only those LSPs whose egress point (using the `to` statement) matches the router ID of the egress node are considered. Other LSPs, whose egress point matches the egress node interface address, are ignored in IGP shortcuts.

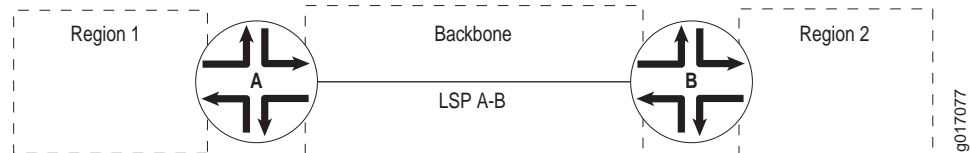
There are exceptions, however. If an LSP has an alias egress point (using the `install` statement) and it matches certain router IDs, it is included in the shortcut computation as well. If multiple equal metric LSPs destined to the same router ID exist, traffic can load-share among them.

IGP Shortcut Applications

You can use shortcuts to engineer traffic traveling toward destination nodes that do not support MPLS LSPs. For example, in Figure 7, traffic traveling toward Router F enters LSP A–E. You can control traffic between Router A and Router F by manipulating LSP A–E; you do not need to explicitly set up an LSP between Router A and Router F.

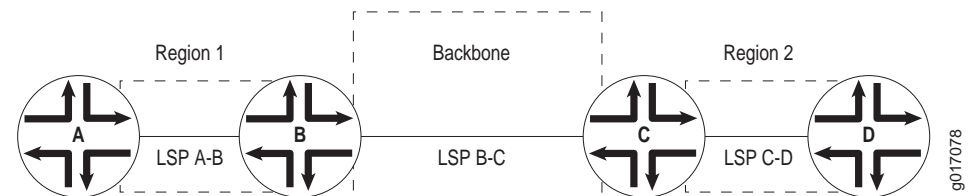
In Figure 8, all traffic from Region 1 to Region 2 traverses LSP A–B if IGP shortcuts are enabled on the ingress router (Router A), permitting aggregation of interregional traffic into one LSP. To perform traffic engineering on the interregional traffic, you have to manipulate LSP A-B only, which avoids creating n^2 LSPs from all routers in Region 1 to all routers in Region 2 and allows efficient resource controls on the backbone network.

Figure 8: IGP Shortcuts



Shortcuts allow you to deploy LSPs into a network in an incremental, hierarchical fashion. In Figure 9, each region can choose to implement traffic engineering LSPs independently, without requiring cooperation from other regions. Each region can choose to deploy intraregion LSPs to fit the region's bandwidth needs, at the pace appropriate for the region.

Figure 9: IGP Shortcuts in a Bigger Network



When intraregion LSPs are in place, interregional traffic automatically traverses the intraregion LSPs as needed, eliminating the need for a full mesh of LSPs between edge routers. For example, traffic from Router A to Router D traverses LSPs A–B, B–C, and C–D.

IGP Shortcuts and Routing Table

IGP typically performs two independent computations. The first is performed without considering any LSP. The result of the computation is stored in the `inet.0` table. This step is no different from traditional SPF computations and is always performed even if IGP shortcut is disabled.

The second computation is performed considering only LSPs as a logical interface. Each LSP's egress router is considered. The list of destinations whose shortest path traverses the egress router (established during the first computation) is placed in the `inet.3` routing table. These destinations are given the egress router of the LSP as a next hop, enabling BGP on the local router to use these LSPs to access BGP next hops beyond the egress router. Normally, BGP can use only LSPs that terminate at the BGP next hop. Note that BGP is the only protocol that uses the `inet.3` routing table. Other protocols will not route traffic through these LSPs.

If traffic engineering for IGP and BGP is enabled (see “IGP and BGP Destinations” on page 45), IGP moves all routes in `inet.3` into `inet.0`, merging all routes while emptying the `inet.3` table. The number of routes in `inet.0` will be exactly the same as before. Route next-hops can traverse a physical interface, an LSP, or the combination of the two if the metrics are equal.

Router Requirements

IGP shortcuts are enabled on a per-node basis. You do not need to coordinate with other nodes.

IGP Shortcuts and VPN Environments

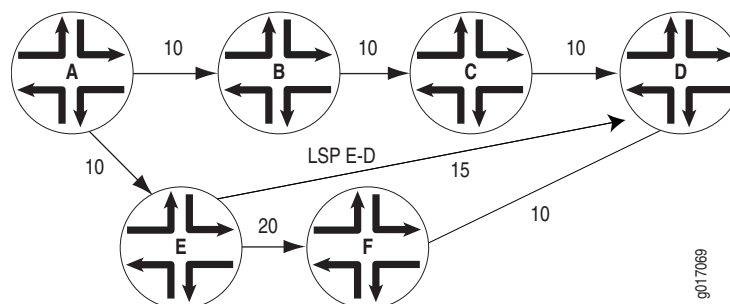
IGP shortcuts (configured under the `[protocols mpls traffic-engineering bgp-igp]` hierarchy level and under the `[protocols ospf traffic-engineering shortcuts]` hierarchy level) do not work in VPN environments. IGP shortcuts move routes in the `inet.3` routing table to the `inet.0` routing table. Virtual private network (VPN) IBGP (which belongs to family `inet-vpn`) relies on next hops that are in the `inet.3` table, so IGP shortcuts are incompatible with VPNs.

Advertising LSPs into IGPs

You can configure your IGP to treat an LSP as a link. IGP shortcuts allow only the ingress router of an LSP to use the LSP in its SPF computation. However, other routers on the network do not know of the existence of that LSP, so they cannot use it. This can lead to suboptimal traffic engineering. In addition, only BGP can use an IGP shortcut to an LSP. When you advertise an LSP as a link into the IGP, all traffic can traverse it, and all routers know about it.

As an example, consider the network shown in Figure 10.

Figure 10: SPF Computations with Advertised LSPs



Assume that Router A is computing a path to Router D. The link between Router E and Router F has a metric of 20; all other links have a metric of 10. Here, the path chosen by Router A is A-B-C-D, which has a metric of 30, instead of A-E-F-D, which has a metric of 40.

If Router E has an LSP to Router D with a metric of 15, you want traffic from Router A to Router D to use the path A-E-D, which has a metric of 25, instead of the path A-B-C-D. However, because Router A does not know about the LSP between Router E and Router D, it cannot route traffic through this path.

For all routers on the network to know about the LSP between Router E and Router D, you need to advertise it. This advertisement announces the LSP as a unidirectional, point-to-point link in the link-state database, and all routers can compute paths using the LSP. The link-state database maintains information about the AS topology and contains information about the router's local state (for example, the router's usable interfaces and reachable neighbors). In Figure 10, Router A will see the link from Router E to Router D and route traffic along this lower-metric path.

Because an LSP is announced as a unidirectional link, you might need to configure a reverse LSP (one that starts at the egress router and ends at the ingress router) so that the SPF bidirectionality check succeeds. As a step in the SPF computation, IS-IS considers a link from Router E to Router D. Before IS-IS uses any link, it verifies that there is a link from Router D to Router E (there is bidirectional connectivity between router E and D). Otherwise, the SPF computation will not use an announced LSP.

When an LSP is advertised to the IGP, the advertising router uses the LSP as the forwarding path for regular routes after installing them in the `inet.0` routing table. All packets traversing the router could be forwarded through the LSP. Conversely, IGP shortcuts are used only to forward packets that are following BGP routes.



NOTE: Do not configure IGP shortcuts and advertise LSPs to the IGP at the same time.

IP and MPLS Packets on Aggregated Interfaces

You can send IP and MPLS packets over aggregated interfaces. To the IP or MPLS session, there is a single LSP composed of the aggregated interfaces. Packets sent to an LSP that is part of an aggregated interface are redistributed over the aggregated member interfaces.

Sending IP and MPLS packets over aggregated interfaces has the following benefits:

- Bandwidth aggregation—You can increase the number of MPLS packet flows sent over each connection. In MPLS, a set of packets sharing the same label is considered a part of the same flow.
- Link redundancy—If a link or a line card failure affects an aggregate member link, the traffic flowing across that link is immediately forwarded across one of the remaining links.

JUNOS supports aggregated SONET and Ethernet interfaces.

Note that the JUNOS implementation of IP and MPLS over aggregated interfaces (aggregated Ethernet devices only) complies with IEEE 802.3ad.

For information about how to configure aggregated Ethernet or aggregated SONET interfaces, see the *JUNOS Network Interfaces Configuration Guide*.

MPLS Applications

In the JUNOS software implementation of MPLS, establishing an LSP installs on the ingress router a host route (a 32-bit mask) toward the egress router. The address of the host route is the destination address of the LSP. By default, the route has a preference value of 7, a value that is higher than all routes except direct interface and static routes. The 32-bit mask ensures that the route is more specific (that is, a longer match) than all other subnet routes. The host routes can be used to traffic-engineer BGP destinations only, or both IGP and BGP destinations.

This section discusses the following topics:

- BGP Destinations on page 43
- IGP and BGP Destinations on page 45
- Selecting a Forwarding LSP Next Hop on page 46

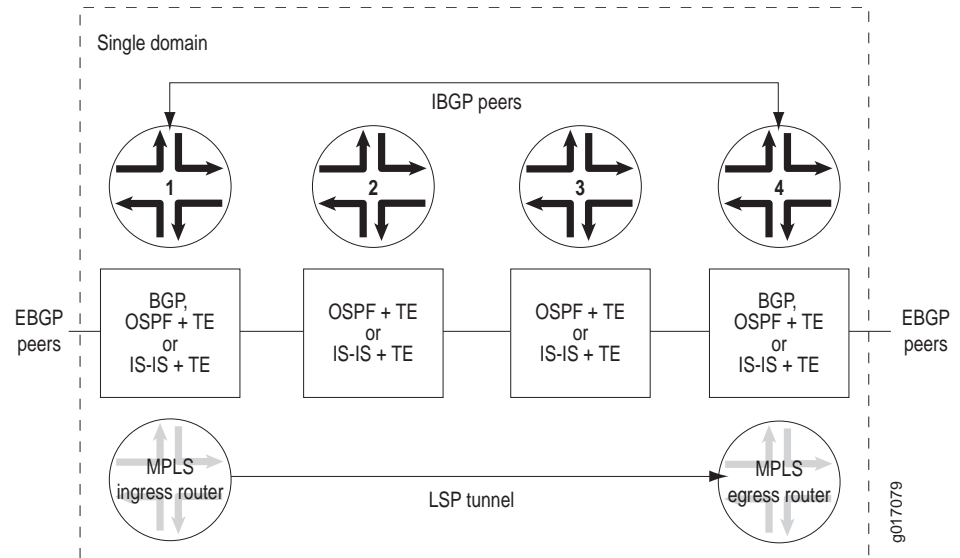
BGP Destinations

You can configure MPLS to control the paths that traffic takes to destinations outside an AS.

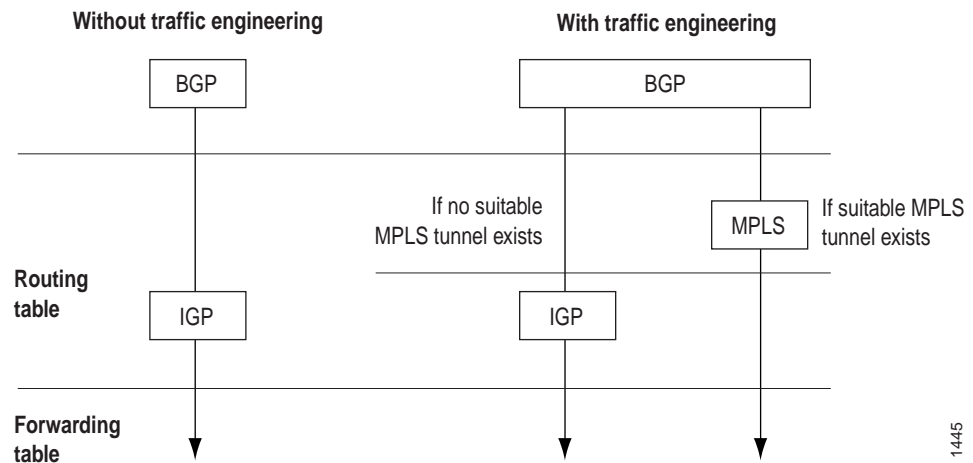
Both internal BGP (IBGP) and external BGP (EBGP) take advantage of the LSP host routes without requiring extra configuration. BGP compares the BGP next-hop address with the LSP host route. If a match is found, the packets for the BGP route are label-switched over the LSP. If multiple BGP routes share the same next-hop address, all the BGP routes are mapped to the same LSP route, regardless of which BGP peer the routes are learned from. If the BGP next-hop address does not match an LSP host route, BGP routes continue to be forwarded based on the IGP routes within the routing domain. In general, when both an LSP route and an IGP route exist for the same BGP next-hop address, the one with the lowest preference is chosen.

Figure 11 shows an MPLS topology that illustrates how MPLS and LSPs work. This topology consists of a single domain with four routers. The two routers at the edges of the domain, Router 1 and Router 4, are running EBGP to communicate with peers outside the domain and IBGP to communicate between themselves. For intradomain communication, all four routers are running an IGP. Finally, an LSP tunnel exists from Router 1 to Router 4.

Figure 11: MPLS Application Topology



When BGP on Router 1 receives prefixes from Router 4, it must determine how to reach a BGP next-hop address. Typically, when traffic engineering is not enabled, BGP uses IGP routes to determine how to reach next-hop addresses. (See the left side of Figure 12.) However, when traffic engineering is enabled, if the BGP next-hop matches the LSP tunnel endpoint (that is, the MPLS egress router), those prefixes enter the LSP tunnel. (To track these prefixes, look at the **Active Route** field in the `show mpls lsp` command output or at the output of the `show route label-switched-path path-name` command.) If the BGP next hop does not match an LSP tunnel endpoint, those prefixes are sent following the IGP's shortest path. (See Figure 12.)

Figure 12: How BGP Determines How to Reach Next-Hop Addresses

1445

IGP and BGP Destinations

You can configure MPLS to control the paths that traffic takes to destinations within an AS.

When traffic engineering is for BGP destinations only, the MPLS host routes are installed in the `inet.3` routing table (see Figure 13 on page 46), separate from the routes learned from other routing protocols. Not all `inet.3` routes are downloaded into the forwarding table. Packets directly addressed to the egress router do not follow the LSP, which prevents routes learned from LSPs from overriding routes learned from IGP or other sources.

Traffic within a domain, including BGP control traffic between BGP peers, is not affected by LSPs. MPLS affects interdomain traffic only; that is, it affects only those BGP prefixes that are learned from an external domain. MPLS does not disrupt intradomain traffic, so IS-IS or OSPF routes remain undisturbed. If you issue a `ping` or `traceroute` command to any destination within the domain, the `ping` or `traceroute` packets follow the IGP path. However, if you issue a `ping` or `traceroute` command from Router 1 in Figure 11 (the LSP ingress router) to a destination outside of the domain, the packets use the LSP tunnel.

When traffic engineering for IGP and BGP destinations is enabled, the MPLS host routes are installed in the `inet.0` table (see Figure 14) and downloaded into the forwarding table. Any traffic destined to the egress router could enter the LSP. In effect, it moves all the routes in `inet.3` into `inet.0`, causing the `inet.3` table to be emptied.

RSVP packets automatically avoid all MPLS LSPs, including those established by RSVP or LDP. This prevents placing one RSVP session into another LSP, or in other words, nesting one LSP into another.

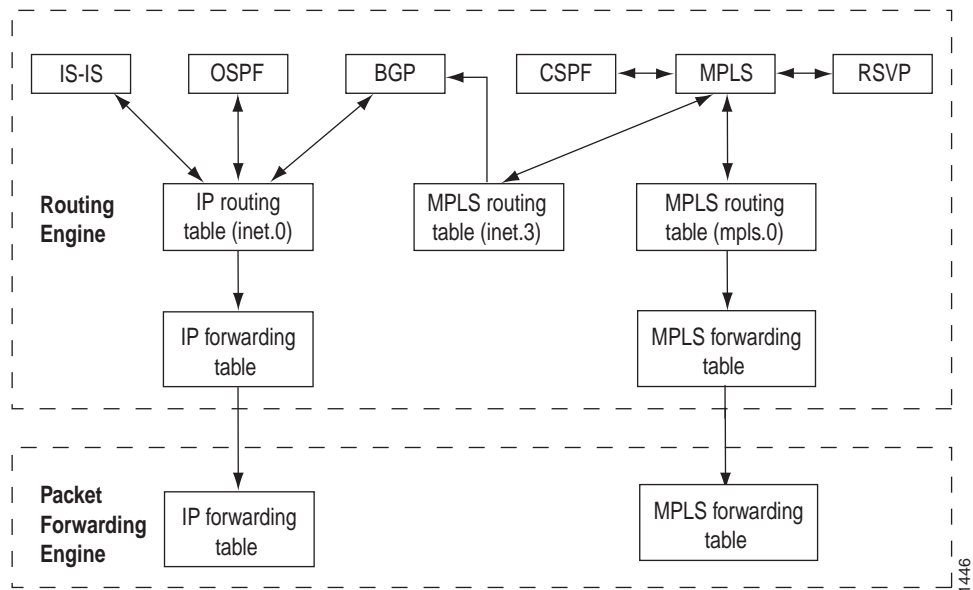
Selecting a Forwarding LSP Next Hop

If more than one LSP tunnel to a BGP next hop exists, the prefixes learned from the BGP next hop are randomly divided among the LSP tunnels. To control which LSP BGP uses to forward data for a given prefix, use the `install-nexthop` statement in the export policy applied to the forwarding table. For more information, see the *JUNOS Routing Protocols Configuration Guide*.

MPLS and Routing Tables

The IGPs and BGP store their routing information in the routing table `inet.0`, which is the main IP routing table. If `traffic-engineering bgp` is configured, thereby allowing only BGP to use MPLS paths for forwarding traffic, MPLS path information is stored in a separate routing table, `inet.3`. Only BGP accesses the `inet.3` routing table. BGP uses both `inet.0` and `inet.3` to resolve next-hop addresses. If `traffic-engineering bgp-igp` is configured, thereby allowing the IGPs to use MPLS paths for forwarding traffic, MPLS path information is stored in the `inet.0` routing table. (Figure 13 and Figure 14 illustrate the routing tables in the two traffic engineering configurations.)

Figure 13: Routing and Forwarding Tables, traffic-engineering bgp



The `inet.3` routing table contains the host address of each LSP’s egress router. This routing table is used on ingress routers to route packets to the destination egress router. BGP uses the `inet.3` routing table on the ingress router to help in resolving next-hop addresses.

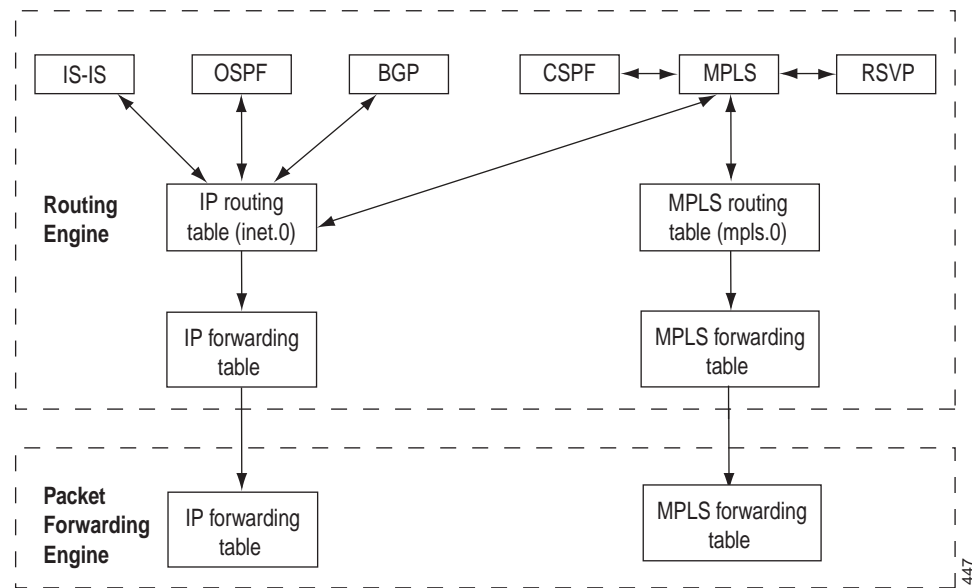
MPLS also maintains an MPLS path routing table (`mpls.0`), which contains a list of the next label-switched router in each LSP. This routing table is used on transit routers to route packets to the next router along an LSP.

Typically, the egress router in an LSP does not consult the `mpls.0` routing table. (This router does not need to consult `mpls.0` because the penultimate router in the LSP either changes the packet's label to a value of 0 or pops the label.) In either case, the egress router forwards it as an IPv4 packet, consulting the IP routing table, `inet.0`, to determine how to forward the packet.

When a transit or egress router receives an MPLS packet, information in the MPLS forwarding table is used to determine the next transit router in the LSP or to determine that this router is the egress router.

When BGP resolves a next-hop prefix, it examines both the `inet.0` and `inet.3` routing tables, seeking the next hop with the lowest preference. If it finds a next-hop entry with an equal preference in both routing tables, BGP prefers the entry in the `inet.3` routing table.

Figure 14: Routing and Forwarding Tables, traffic-engineering bgp-igp



Generally, BGP selects next-hop entries in the `inet.3` routing table because their preferences are always lower than OSPF and IS-IS next-hop preferences. When you configure LSPs, you can override the default preference for MPLS LSPs, which might alter the next-hop selection process.

When BGP selects a next-hop entry from the `inet.3` routing table, it installs that LSP into the forwarding table in the Packet Forwarding Engine, which causes packets destined for that next hop to enter and travel along the LSP. If the LSP is removed or fails, the path is removed from the `inet.3` routing table and from the forwarding table, and BGP reverts to using a next hop from the `inet.0` routing table.

MPLS and Traffic Protection

Typically, when an LSP fails, the router immediately upstream from the failure signals the outage to the ingress router. The ingress router calculates a new path to the egress router, establishes the new LSP, and then directs the traffic from the failed path to the new path. This rerouting process can be time-consuming and prone to failure. For example, the outage signals to the ingress router might get lost, or the new path might take too long to come up, resulting in significant packet drops. The JUNOS software provides several complementary mechanisms for protecting against LSP failures:

- Standby secondary paths—You can configure primary and secondary paths. You configure secondary paths with the `standby` statement. To activate traffic protection, you need to configure these standby paths only on the ingress router. If the primary path fails, the ingress router immediately reroutes traffic from the failed path to the standby path, thereby eliminating the need to calculate a new route and signal a new path. For information about configuring standby LSPs, see “Configuring the Standby State” on page 103.
- Fast reroute—You configure fast reroute on an LSP to minimize the effect of a failure in the LSP. Fast reroute enables a router upstream from the failure to route around the failure quickly to the router downstream of the failure. The upstream router then signals the outage to the ingress router, thereby maintaining connectivity before a new LSP is established. For a detailed overview of fast reroute, see “Fast Reroute Overview” on page 49. For information about configuring fast reroute, see “Configuring Fast Reroute” on page 77.
- Link protection—You can configure link protection to help ensure that traffic traversing a specific interface from one router to another can continue to reach its destination in the event that this interface fails. When link protection is configured for an interface and configured for an LSP that traverses this interface, a bypass LSP is created that will handle this traffic if the interface fails. The bypass LSP uses a different interface and path to reach the same destination. For information about configuring link protection, see “Configuring Node Protection or Link Protection” on page 286.

When standby secondary path, and fast reroute or link protection are configured on an LSP, full traffic protection is enabled. When a failure occurs in an LSP, the router upstream from the failure routes traffic around the failure and notifies the ingress router of the failure. This rerouting keeps the traffic flowing while waiting for the notification to be processed at the ingress router. After receiving the failure notification, the ingress router immediately reroutes the traffic from the patched primary path to the more optimal standby path.

Fast reroute and link protection provide a similar type of traffic protection. Both features provide a quick transfer service and employ a similar design. Fast reroute and link protection are both described in the same Internet draft `draft-ietf-mpls-rsvp-lsp-fastreroute-03.txt`, *Fast Reroute Extensions to RSVP-TE for LSP Tunnels*. However, you need to configure only one or the other. Although you can configure both, there is little, if any, benefit in doing so.

Fast Reroute Overview

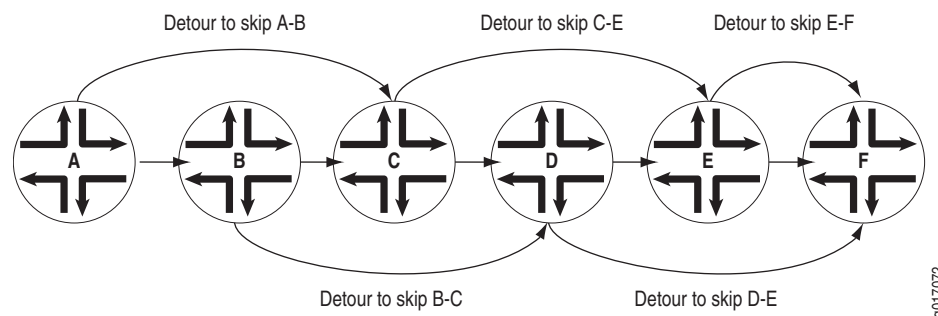
The following sections provide an overview of how fast reroute works:

- Fast Reroute Overview on page 49
- Detour Merging Process on page 52
- Detour Computations on page 53

Fast Reroute Overview

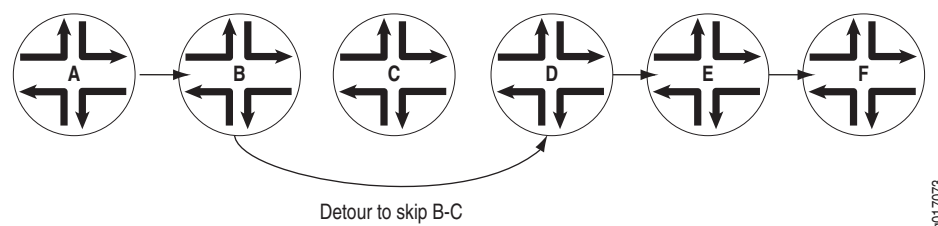
Fast rerouting is accomplished by precomputing and preestablishing a number of detours along the LSP. In case of a network failure on the current LSP path, the traffic is quickly routed to one of the detours. Figure 15 illustrates an LSP from Router A to Router F, showing the established detours. Each detour is established by an upstream node with the intent of avoiding the link toward the immediate downstream node and the immediate downstream node itself. Each detour might traverse through one or more label-switched routers that are not shown in the figure.

Figure 15: Detours Established for an LSP Using Fast Reroute



If a node detects that a downstream link has failed (using a link-layer-specific liveness detection mechanism) or that a downstream node has failed (for example, using the RSVP neighbor hello protocol), the node quickly switches the traffic to the detour and, at the same time, signals the ingress router about the link or node failure. Figure 16 on page 49 illustrates the detour taken when the link between Router B and Router C fails.

Figure 16: Detour After the Link from Router B to Router C Fails



If the network topology is not rich enough (there are not enough routers with sufficient links to other routers), some of the detours might not succeed. For example, the detour from Router A to Router C in Figure 15 cannot traverse link A-B and Router B. If such a path is not possible, the detour does not occur.

Note that after the node switches traffic to the detour, it might switch the traffic again to a newly calculated detour soon after. This is because the initial detour route might not be the best route. To make rerouting as fast as possible, the node switches traffic onto the initial detour without first verifying that the detour is valid. Once the switch is made, the node recomputes the detour. If the node determines that the initial detour is still valid, traffic continues to flow over this detour. If the node determines that the initial detour is no longer valid, it again switches the traffic to a newly computed detour.



NOTE: If you issue `show` commands after the node has switched traffic to the initial detour, the node might indicate that the traffic is still flowing over the original LSP. This situation is temporary and should correct itself quickly.

The time required for a fast-rerouting detour to take effect depends on two independent time intervals:

- Amount of time to detect that there is a link or node failure—This interval depends greatly on the link layer in use and the nature of the failure. For example, failure detection on an SONET/SDH link typically is much faster than on a Gigabit Ethernet link, and both are much faster than detection of a router failure.
- Amount of time required to splice the traffic onto the detour—This operation is performed by the Packet Forwarding Engine, which requires little time to splice traffic onto the detour. The time needed can vary depending on the number of LSPs being switched to detours.

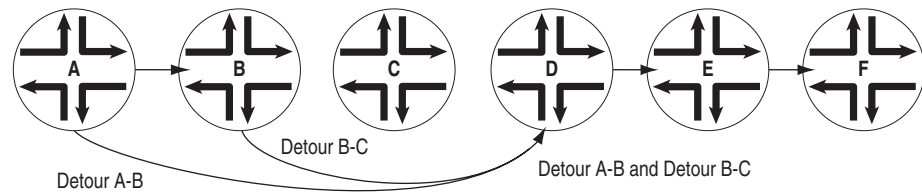
Fast reroute is a short-term patch to reduce packet loss. Because detour computation might not reserve adequate bandwidth, the detours might introduce congestion on the alternate links. The ingress router is the only router that is fully aware of LSP policy constraints and, therefore, is the only router able to come up with adequate long-term alternate paths.

Fast reroute protects traffic against any single point of failure between the ingress and egress routers. If there are multiple failures along an LSP, fast reroute itself might fail. Also, fast reroute does not protect against failure of the ingress or egress routers.

Detours are created by use of RSVP and, like all RSVP sessions, they require extra state and overhead in the network. For this reason, each node establishes at most one detour for each LSP that has fast reroute enabled. Creating more than one detour for each LSP increases the overhead, but serves no practical purpose.

To reduce network overhead further, each detour attempts to merge back into the LSP as soon as possible after the failed node or link. If you can consider an LSP that travels through n router nodes, it is possible to create $n - 1$ detours. For instance, in Figure 17, the detour tries to merge back into the LSP at Router D instead of at Router E or Router F. Merging back into the LSP makes the detour scalability problem more manageable. If topology limitations prevent the detour from quickly merging back into the LSP, detours merge with other detours automatically.

Figure 17: Detours Merging into Other Detours



Detour Merging Process

This section describes the process used by a router to determine which LSP to select when the router receives Path messages from different interfaces with identical Session and Sender Template objects. When this occurs, the router needs to merge the path states.

The router employs the following process to determine when and how to merge path states:

- When all the Path messages do not include a Fast Reroute or a Detour object, or when the router is the egress of the LSP, no merging is required. The messages are processed according to RSVP traffic engineering (TE).
- Otherwise, the router *must* record the path state in addition to the incoming interface. If the path messages do not share the same outgoing interface and next-hop router, the router considers them to be independent LSPs and does not merge them.
- For all the Path messages that share the same outgoing interface and next-hop router, the router uses the following process to select the final LSP:
 - If only one LSP originates from this node, select it as the final LSP.
 - If only one LSP contains a Fast Reroute object, select it as the final LSP.
 - If there are several LSPs and some of them have a Detour object, eliminate those containing a Detour object from the final LSP selection process.
 - If several final LSP candidates remain (that is, there are still both Detour and protected LSPs), select the LSPs with Fast Reroute objects.
 - If none of the LSPs have Fast Reroute objects, select the ones without Detour objects. If all the LSPs have Detour objects, select them all.
 - Of the remaining LSP candidates, eliminate from consideration those that traverse nodes that other LSPs avoid.
 - If several candidate LSPs still remain, select the one with the shortest ERO path length. If more than one LSP has the same path length, select one randomly.
- Once the final LSP has been identified, the router must transmit only the Path messages that correspond to this LSP. All other LSPs are considered merged at this node.

Detour Computations

Computing and setting up detours is done independently at each node. On a node, if an LSP has fast reroute enabled and if a downstream link or node can be identified, the router performs a Constrained Shortest Path First (CSPF) computation using the information in the local TED. For this reason, detours rely on your IGP supporting traffic engineering extensions. Without the TED, detours cannot be established.

CSPF initially attempts to find a path that skips the next downstream node. Attempting to find this path provides protection against downstream failures in either nodes or links. If a node-skipping path is not available, CSPF attempts to find a path on an alternate link to the next downstream node. Attempting to find an alternate link provides protection against downstream failures in links only. Detour computations might not succeed the first time. If a computation fails, the router recomputes detours approximately once every refresh interval until the computation succeeds. The RSVP metric for each detour is set to a value in the range from 10,000 through 19,999.

Automatic Bandwidth Allocation

Automatic bandwidth allocation allows an MPLS tunnel to automatically adjust its bandwidth allocation based on the volume of traffic flowing through the tunnel. You can configure an LSP with minimal bandwidth; this feature can dynamically adjust the LSP's bandwidth allocation based on current traffic patterns. The bandwidth adjustments do not interrupt traffic flow through the tunnel.

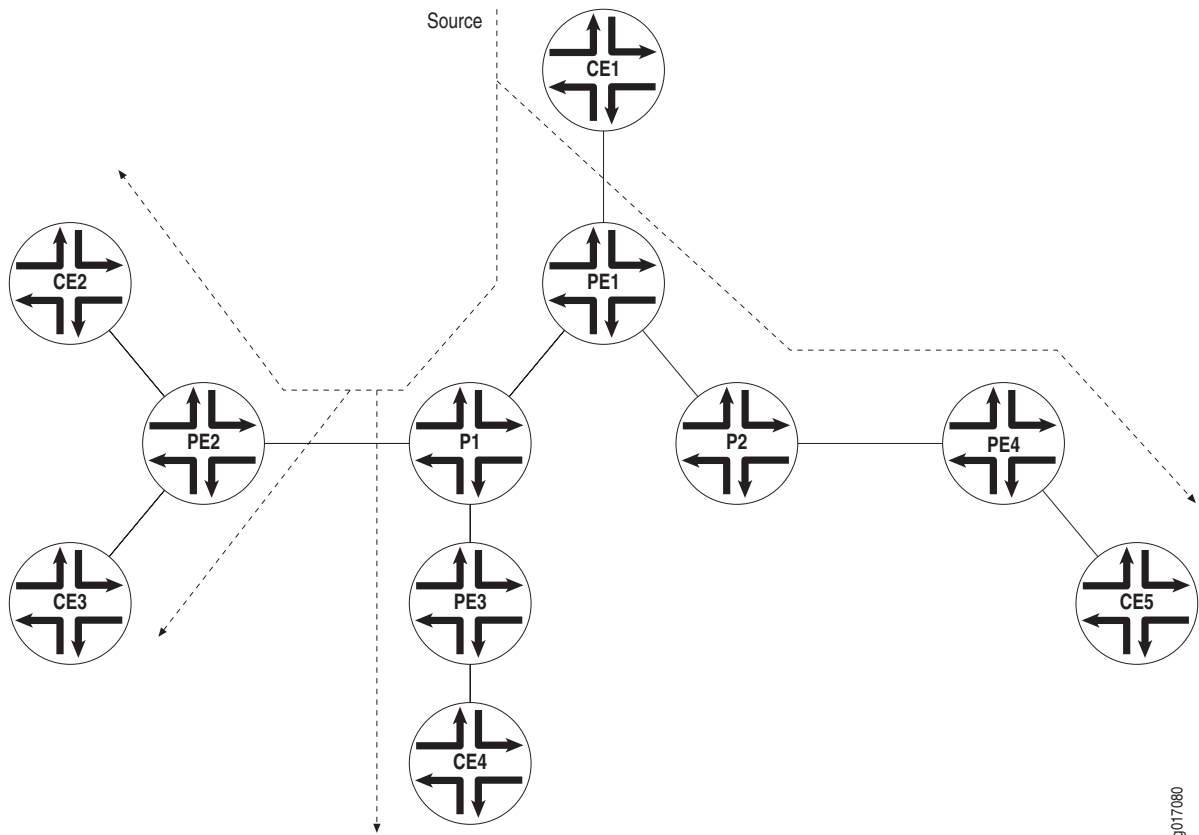
You set a sampling interval on an LSP configured with automatic bandwidth allocation. The average bandwidth is monitored during this interval. At the end of the interval, an attempt is made to signal a new path for the LSP with the bandwidth allocation set to the maximum average value for the preceding sampling interval. If the new path is successfully established and the original path is removed, the LSP is switched over to the new path. If a new path is not created, the LSP continues to use its current path until the end of the next sampling interval, when another attempt is made to establish a new path. Note that you can set minimum and maximum bandwidth values for the LSP.

Point-to-Multipoint LSPs

A point-to-multipoint MPLS LSP is an RSVP-signaled LSP with a single source and multiple destinations. By taking advantage of the MPLS packet replication capability of the network, point-to-multipoint LSPs avoid unnecessary packet replication at the ingress router. Packet replication takes place only when packets are forwarded to two or more different destinations requiring different network path.

This process is illustrated in Figure 18. Router PE1 is configured with a point-to-multipoint LSP to Routers PE2, PE3, and PE4. When Router PE1 sends a packet on the point-to-multipoint LSP to Routers P1 and P2, Router P1 replicates the packet and forwards it to Routers PE2 and PE3. Router P2 sends the packet to Router PE4.

Figure 18: Point-to-Multipoint LSPs



9017080

The following are some of the properties of point-to-multipoint LSPs:

- A point-to-multipoint LSP allows you to use MPLS for point-to-multipoint data distribution. This functionality is similar to that provided by IP multicast.
- You can add and remove sub-LSPs from a main point-to-multipoint LSP without disrupting traffic. The unaffected parts of the point-to-multipoint LSP continue to function normally.
- You can configure a node to be both a transit and an egress router for different sub-LSPs of the same point-to-multipoint LSP.
- You can enable link protection on a point-to-multipoint LSP. Link protection can provide a bypass LSP for each of the sub-LSPs that make up the point-to-multipoint LSP. If any of the primary paths fail, traffic can be quickly switched to the bypass.
- You can configure paths statically.
- You can enable graceful restart on point-to-multipoint LSPs.

For information on how to configure point-to-multipoint LSPs, see “Configuring Point-to-Multipoint LSPs” on page 108.

MPLS Load Balancing Based on the IP Header and MPLS Labels

It is possible to load-balance on a per-packet basis in MPLS. Load balancing can be performed on information in both the IP header and on up to three MPLS labels, providing a more uniform distribution of MPLS traffic to next hops.

This feature is only available on M-series and T-series routing platforms with enhanced FPCs. It requires no configuration.

The following information is extracted from the packet and used to load-balance the MPLS traffic:

- Interface index—24 bits
- MPLS label stack—bits 0 through 23 (the TTL bits are not examined)
- IPv4 header information:
 - Protocol—8 bits
 - Destination address—32 bits
 - Source address—32 bits
 - Source port—16 bits
 - Destination port—16 bits

- IPv6 header information:
 - Next header—8 bits
 - Least significant 4 bytes of destination address
 - Least significant 4 bytes of source address
 - Source port—16 bits
 - Destination port—16 bits

In summary, MPLS load balancing is performed using the following fields:

Interface index + MPLS label + IP header (IPv4 or IPv6)