

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 then is 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 into 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 16

Link-Layer Support on page 16

MPLS and Traffic Engineering on page 17

MPLS Applications on page 29

MPLS and Routing Tables on page 32

MPLS and Traffic Protection on page 34

MPLS Standards

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

ICMP Extensions for Multiprotocol Label Switching, Internet draft
draft-ietf-mpls-icmp-01.txt

The following documents provide a good overview of MPLS:

Multiprotocol Label Switching Architecture, Internet draft draft-ietf-mpls-arch-06.txt

A Framework for Multiprotocol Label Switching, Internet draft
draft-ietf-mpls-framework-05.txt

MPLS Label Stack Encoding, Internet draft draft-ietf-mpls-label-encaps-07.txt

The following documents provide information about traffic engineering:

RFC 2702, *Requirements for Traffic Engineering Over MPLS*

IS-IS Extensions for Traffic Engineering, Internet draft draft-ietf-isis-traffic-02.txt

Traffic Engineering Extensions to OSPF, Internet draft draft-katz-yeung-ospf-traffic-01.txt

Calculating IGP Routes over Traffic Engineering Tunnels, Internet draft
draft-hsmit-mpls-igp-spf-01.txt

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

The JUNOS software supports a proprietary MIB for MPLS objects; see the *JUNOS Internet Software Configuration Guide: Interfaces and Chassis* for more information.

Link-Layer Support

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

PPP—Protocol ID 0x0281, NCP protocol ID 0x8281.

Ethernet/Cisco HDLC—Ethernet type 0x8847.

ATM—SNAP-encoded Ethernet type 0x8847. Support is included for both point-to-point mode or NBMA mode. Support is not included for encoding MPLS labels as part of ATM VPI/VCI.

Frame Relay—SNAP-encoded, Ethernet type 0x8847. Support is not included for encoding MPLS labels as part of Frame Relay DLCI.

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 only BGP traffic (traffic whose destination is outside of an 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 IGP traffic; therefore, both intra-AS and inter-AS traffic is affected by the LSPs.

Label Description

Packets travelling along an LSP are identified by a *label*, a 20-bit unsigned integer in the range 0 through 1048575:

- 0 through 15—Reserved and have special semantics.

- 16 through 1023—Unused and unassigned by the software, a feature that is specific to the JUNOS software. You can use labels to manually configure static LSPs, and ensure that there are no conflicts with labels that are dynamically assigned by the software.

- 1024 through 99,999—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 100,000-799,999 range and per-interface labels are assigned in the 800,000-1,048,575 range.

Special Labels

Some of the reserved labels (in the 0 through 15 range) have well-defined meanings. For more complete details, see *MPLS Label Stack Encoding*, Internet draft draft-ietf-mpls-label-encaps-07.txt.

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 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 Implicit 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 IPv6 packet.

3, Implicit Null Label—This label is used in the control protocol (LDP, 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 a real label. 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. This means an egress router will not process a labelled packet; rather, it receives the payload (either IPv4, IPv6 or others) directly. This reduces one MPLS lookup at egress.

For label stacked packets, the egress router will receive an MPLS label packet with its top label already popped by the penultimate router. The egress router cannot receive label-stacked packets using label 0 or 2.

When functioning as an egress router, JUNOS software release 4.2 typically requests label 3. In JUNOS software release 4.0 and earlier, the egress router typically requests label 0 from the penultimate router. There are no interoperability problems among various JUNOS software versions because the software accepts any of these special labels and performs the requested operations.

Label Allocation

Note that earlier versions of JUNOS software allocate labels on a per-interface basis. Labels on different interfaces are assigned independently. This means that a particular label received on one interface is not related to the same label received on a different interface. For this reason, labels usually are preceded by an interface name in display output (in the format *interface.label*). For example, so-5/0/0.0.01024 indicates that the label value 01024 was received on interface so-5/0/0.0.

Label values can be allocated either on a per-router or per-interface basis. In the JUNOS software version 4.2, label values are allocated per router only. In this case, the display output will show only the label, for example, 01024.

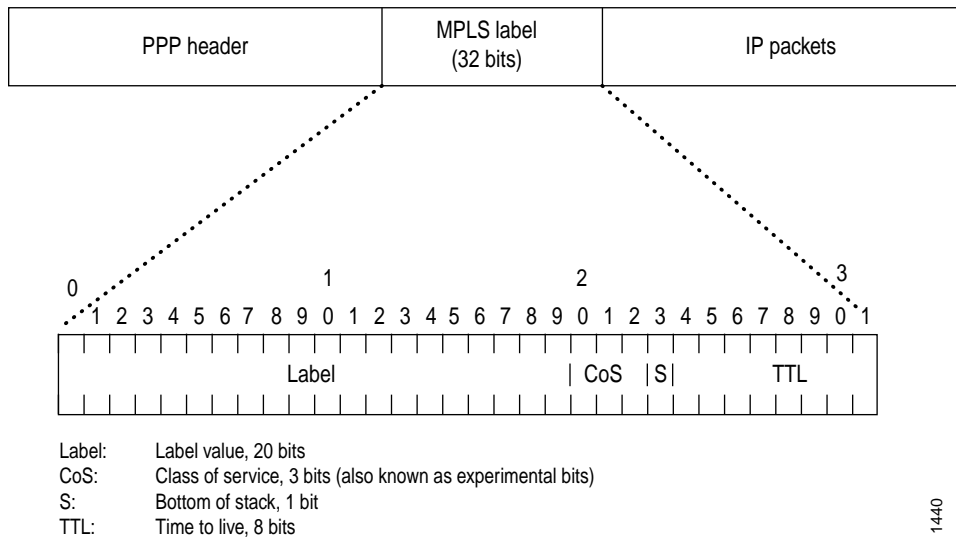
Labels for multicast packets are independent of those for unicast packets. Currently, the JUNOS software does not support multicast labels.

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 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 as to 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



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, S, and CoS fields are derived from the IP packet header. If the Push operation is performed on an existing MPLS packet, you will have a packet with 2 or more labels. This is called label stacking. The top label must have its S field set to 0, and might derive CoS and TTL from lower levels. Note that in JUNOS software release 4.2, 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 will yield another MPLS packet. The new top label might derive CoS and TTL from a previous top label. Note that in JUNOS software release 4.2, the popped TTL value from the previous top label is not written back to the new top label.

- Swap—Replaces 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 3) on top of existing packets. This is equivalent to doing Push multiple times.

- Swap and Push—Replace the existing top of the label stack with a new label, followed by pushing 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 then is 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. Then, they 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, no liveliness detection, and no statistics reporting.

LDP signaled LSPs—See LDP Overview on page 137.

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 strict and loose hops. 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 IS-IS or 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 LSPs that use constrained-path, the LSP computation is confined to one IGP area, 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 has no dependence 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

- Bandwidth requirements

- Hop limitations

- Administrative groups (that is, link color requirements)

- Priority (setup and hold)

- Explicit route (strict or loose)

Link attributes

- Reservable bandwidth of the links (static bandwidth minus the currently reserved bandwidth)

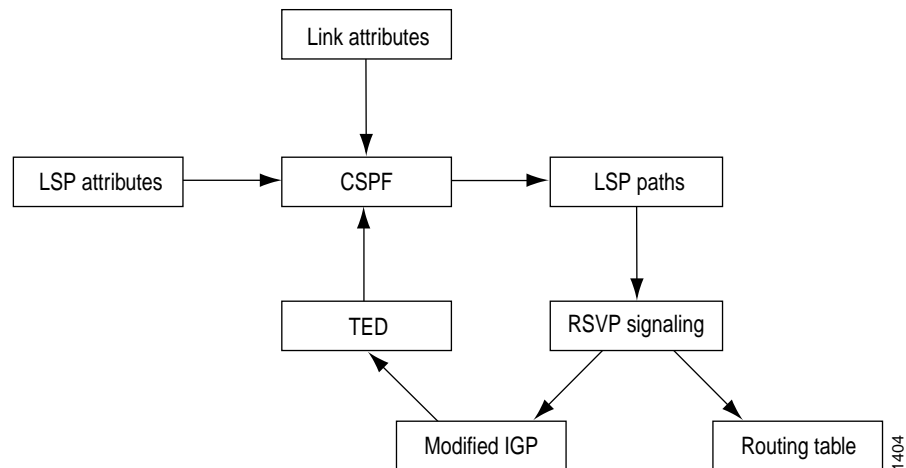
- Administrative groups (that is, link colors assigned to the link)

The data that CSPF considers comes from the:

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 2 for a diagram of the relationship among 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 2: CSPF Computation Process



How CSPF Selects a Path

To select a path, CSPF follows these steps:

1. Compute 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. Prune the topology database (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, prune all links that do not share any included colors.
4. If the LSP configuration includes the exclude statement, prune all links that contain excluded colors and do not contain a color.
5. Find the shortest path towards 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, choose the one whose last hop address is the same as the LSP's destination.
7. If several equal-cost paths remain, select the one with the fewest number of hops.
8. If several equal-cost paths remain, apply 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 applying the rules from the previous section, a tie-breaking rule is applied to choose the path for the LSP. There are three tie-breaking rules: random, least fill, and most fill. The rule used depends on the configuration. Random is the default rule. For 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

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.

Computing Paths Offline

The JUNOS software provides on-line, 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 totally 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

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 now configure backup paths that minimize the number of shared links and fiber paths with the primary paths as much as possible, to ensure that in the event of a fiber cut, the minimum amount of data is lost and that a path still exists to the destination.

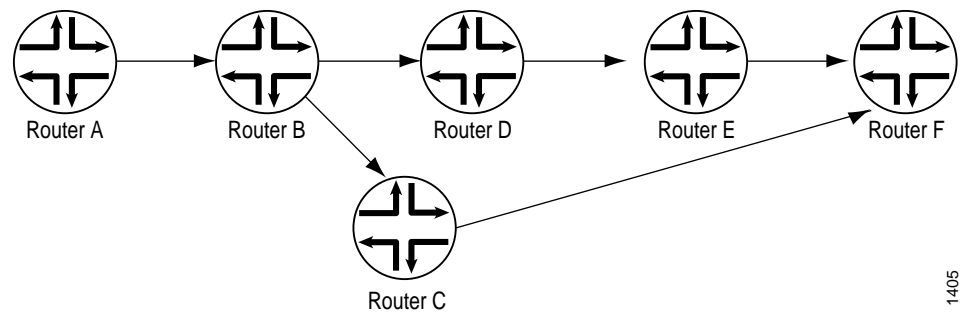
For a backup path to work optimally, it must not share links or physical fiber paths with the primary path. This ensures that a single point of failure will not affect the primary and backup paths at the same time.

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. 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 to 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 (Figure 3):

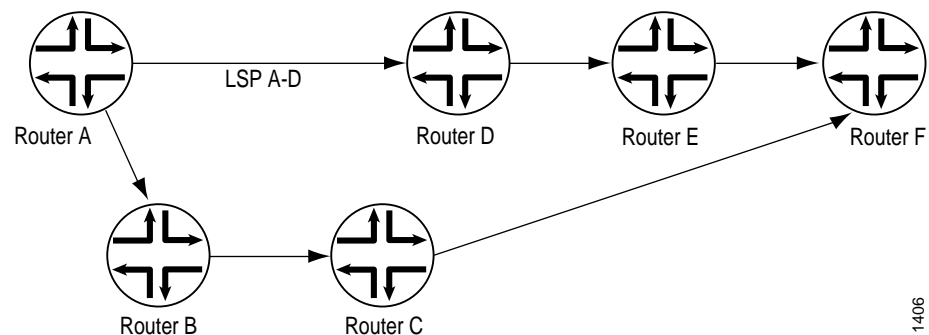
Figure 3: Typical SPF Tree, Sourced from Router A



1405

If an LSP connects Router A to Router D and if IGP shortcut is enabled on Router A, you might have the SPF tree shown in Figure 4.

Figure 4: Modified SPF Tree, Using LSP A-D as a Shortcut



1406

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

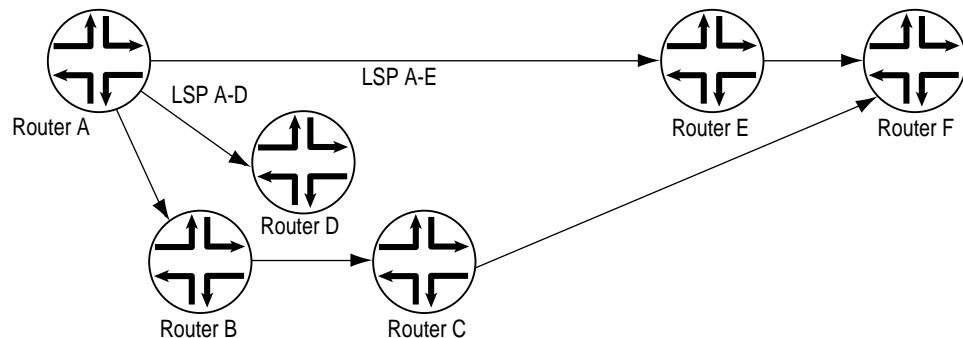
Use LSP A-D

The decision between the two choices is made 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 3). If the LSP metric is lower, LSP A-D is used (Figure 4). If both metrics are equal, Router A might share the load between the two 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 another LSP connects Router A to Router E, you might have the SPF tree shown in Figure 5.

Figure 5: Modified SPF Tree, Using Both LSP A-D and LSP A-E as Shortcuts



1441

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 the IGP shortcut for IS-IS and OSPF, see the *JUNOS Internet Software Configuration Guide: Routing and Routing Protocols*. You can enable an IGP shortcut on a per-router basis; there is no need to enable it 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.

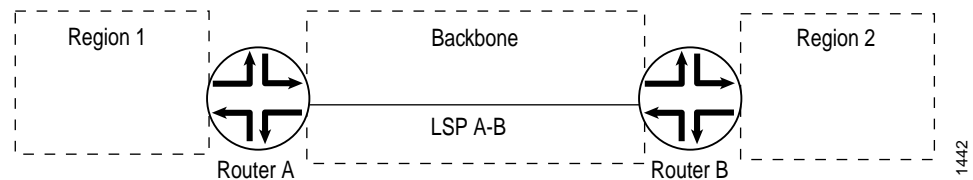
IGP Shortcut Applications

You can use shortcuts to engineer traffic traveling towards destination nodes that do not support MPLS LSPs. For example, in Figure 5, 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.

Use shortcuts to reduce the number of LSPs required to accomplish traffic engineering.

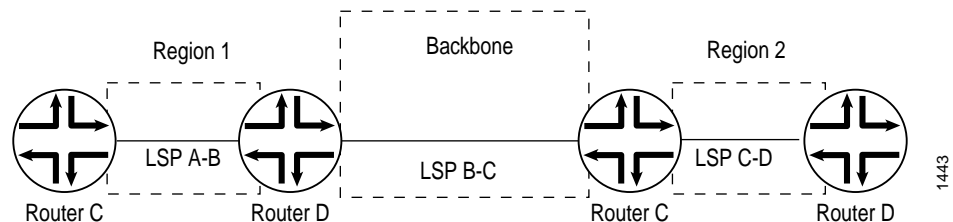
In Figure 6, all traffic from Region 1 to Region 2 traverses LSP A-B if IGP shortcuts are enabled on the ingress router (Router A), permitting aggregation of interregional traffic into one LSP. To perform traffic engineering on the interregional traffic, you only have to manipulate LSP A-B, which avoids creating N^2 LSPs from all the routers in Region 1 to all the routers in Region 2 and allows efficient resource controls on the backbone network.

Figure 6: IGP Shortcuts



Shortcuts allow deployment of LSPs into a network in an incremental, hierarchical fashion. In Figure 7, 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 7: 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 does two independent computations. The first one 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 with only LSPs as a logical interface in mind, producing routes that are reachable through LSPs only. The results are stored in the inet.3 table only. The routes produced in the second step are typically a subset of the first step.

If traffic engineering for IGP and BGP is enabled (see IGP and BGP Destinations on page 31), IGP moves all the routes in inet.3 into inet.0, merging all routes together, and at the same time emptying the inet.3 table. The number of routes in inet.0 will be exactly the same as before. Route next hops may traverse a physical interface, an LSP, or the combination of both if the metrics are equal.

Router Requirements

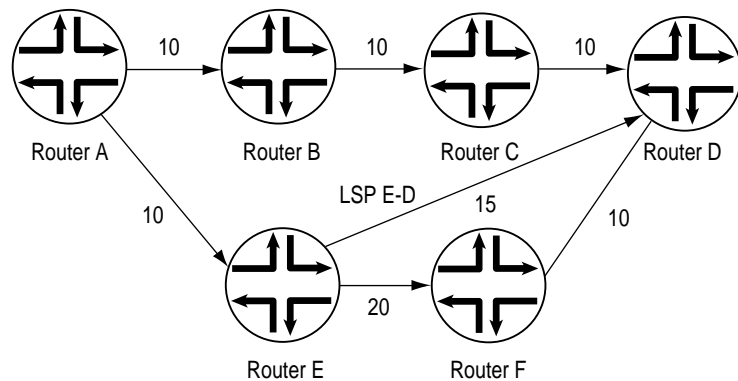
IGP shortcuts are enabled on a per-node basis. You do not need to coordinate with other nodes.

Advertise LSPs into IGP

IGP shortcuts allow an 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.

As an example, consider the network shown in Figure 8:

Figure 8: 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 metric 20; all other links have metric 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.

1458

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 Autonomous System topology and contains information about the router's local state (for example, the router's usable interfaces and reachable neighbors). In Figure 3, 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.

MPLS Applications

In the JUNOS 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 the route is more specific (that is, 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.

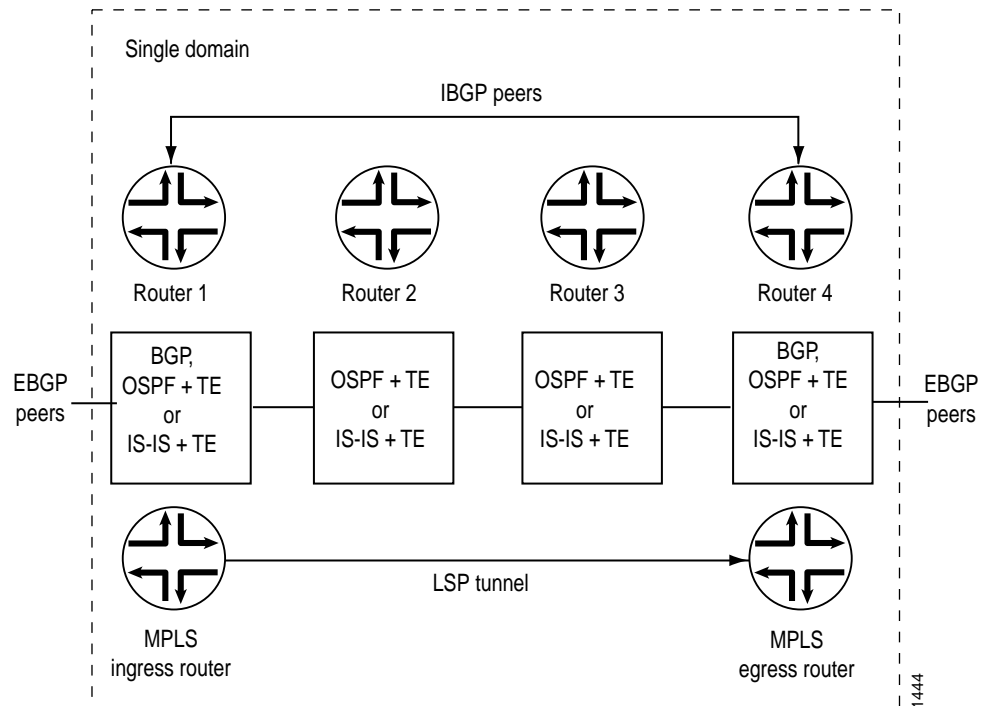
BGP Destinations

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

Both IBGP and 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 highest preference is chosen.

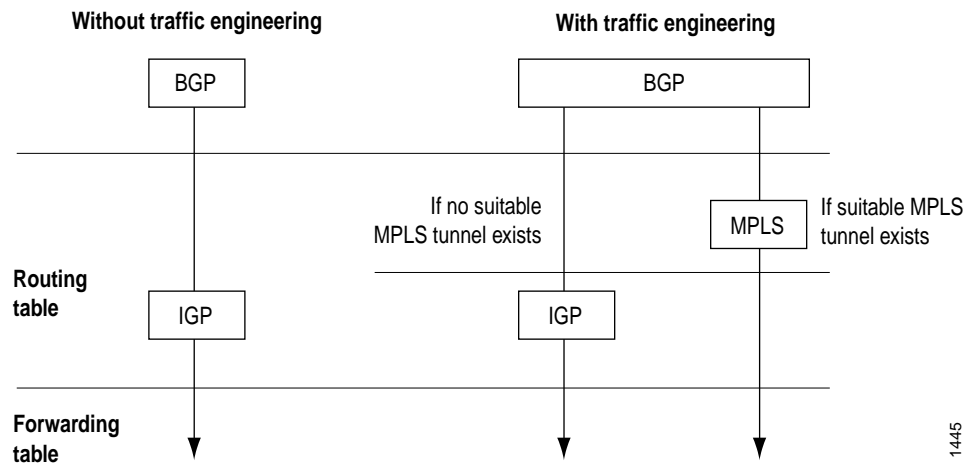
Figure 9 shows an MPLS topology that illustrates how MPLS and LSPs work. This topology consists of a single domain with four routers. The two routers at the edges of the domain, Router 1 and Router 4, are running EBGP to communicate with peers outside the domain and IBGP to communicate between themselves. For intradomain communication, all four routers are running an IGP. Finally, an LSP tunnel exists from Router 1 to Router 4.

Figure 9: 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 10.) However, when traffic engineering is enabled, if the BGP next hop matches the LSP tunnel end point (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 end point, those prefixes are sent following the IGP's shortest path. (See the right side of Figure 10.)

Figure 10: How BGP Determines How to Reach Next-Hop Addresses



1445

IGP and BGP Destinations

MPLS can be configured to control the paths that traffic takes to destination within an AS.

When traffic engineering is for IGP destinations only, the MPLS host routes are installed in the inet.3 routing table (see Figure 11), separate from the routes learned from other routing protocols. All inet.3 routes are not 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 impacts interdomain transit 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 9 (the LSP ingress router) to a destination outside of the domain, the packets use the LSP tunnel.

When traffic engineering for IGP and BGP destination is enabled, the MPLS host routes are installed in the inet.0 table (see Figure 12) 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.

Select Forwarding LSP Next Hop

If more than one LSP tunnel to a BGP next hop exists, then 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 *JUNOS Internet Software Configuration Guide: Routing and Routing Protocols*.

MPLS and Routing Tables

The IGP 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 IGP to use MPLS paths for forwarding traffic, MPLS path information is stored in the `inet.0` routing table. (Figure 11 and Figure 12 illustrate the routing tables in the two traffic-engineering configurations.)

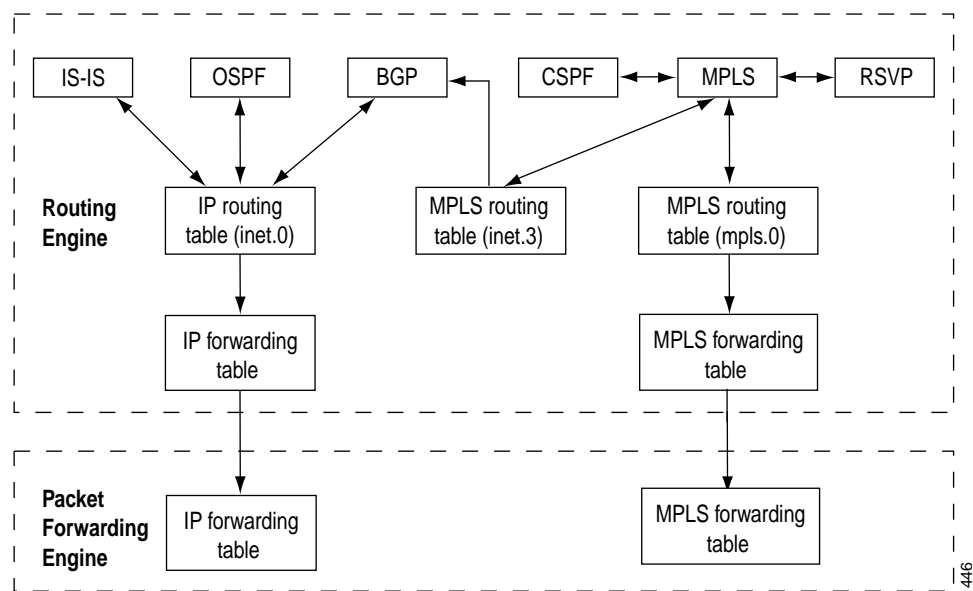
The `inet.3` routing table contains the host address of each LSP's egress router. This routing table is used on ingress routers to route packets to the destination egress router. BGP uses the `inet.3` routing table on the ingress router to help in resolving next-hop addresses.

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.

Figure 11: MPLS Routing and Forwarding Tables when `traffic-engineering bgp` is Configured

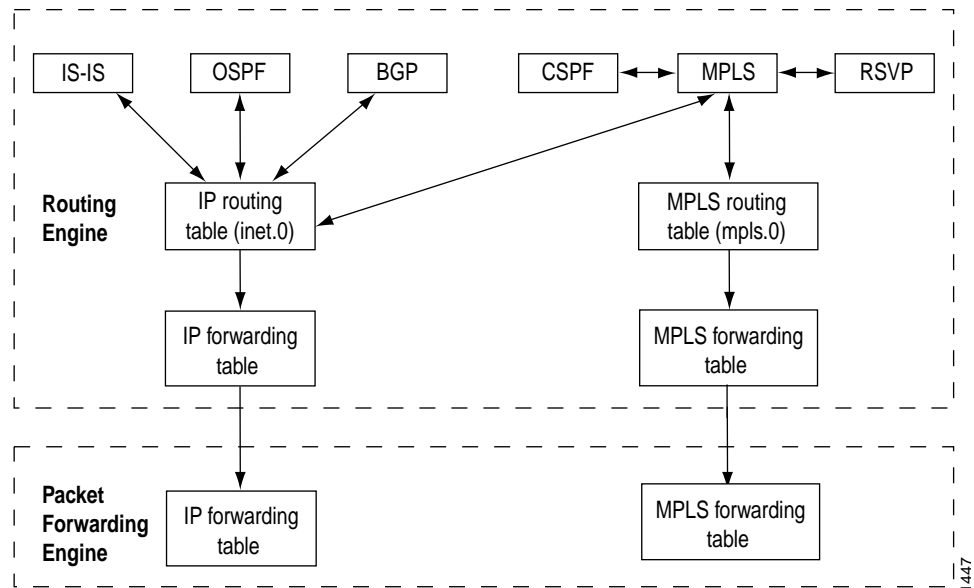


When BGP resolves a next-hop prefix, it examines both the inet.0 and inet.3 routing tables, seeking the next hop with the highest preference. If it finds a next-hop entry with equal preference in both routing tables, BGP prefers the entry found in the inet.3 routing table.

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 would revert to using a next hop from the inet.0 routing table.

Figure 12: MPLS Routing and Forwarding Tables when traffic-engineering bgp-igp is Configured



1447

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 two 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 more information about configuring standby LSPs, see “Configure the Standby State” on page 60.

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 more information about fast reroute, see “Configure Fast Reroute” on page 45.

When standby secondary path and fast reroute are both configured on the LSP, full traffic protection is enabled. When a failure occurs in an LSP, the router upstream of 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.

Per-Prefix Load Balancing

When there are multiple equal cost tunnels to a destination, load balancing can be controlled for each path. Load balancing is now proportional to the configured bandwidth per LSP. If an LSP has a larger bandwidth associated with it, that LSP will carry a larger number of prefixes. If you configure the bandwidth, the prefixes will automatically adjust themselves.

