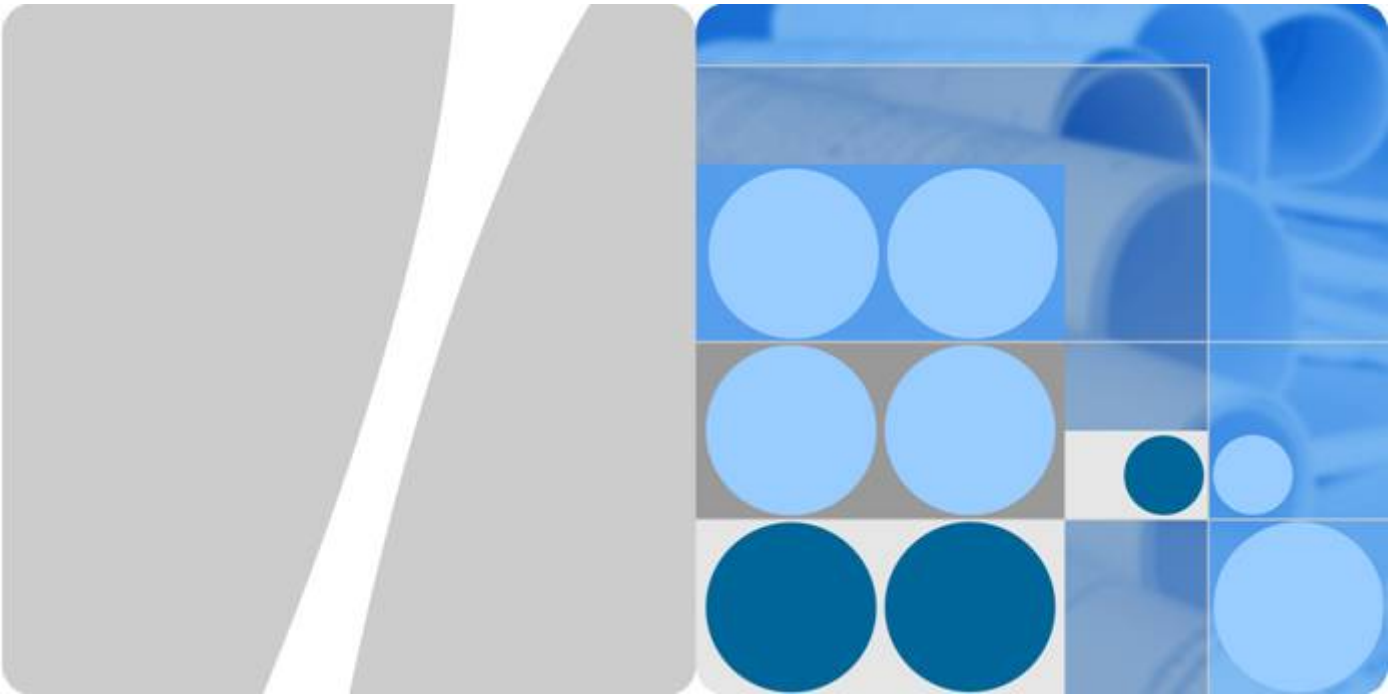


Enterprise Data Communication Products

Feature Description - MPLS

Issue 04

Date 2013-04-15

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://enterprise.huawei.com

About This Document

Intended Audience

This document describes the definition, purpose, and implementation of features on enterprise datacom products including the campus network switch, enterprise router, data center switch, and WLAN. For features supported by the device, see *Configuration Guide*.

This document is intended for:

- Network planning engineers
- Commissioning engineers
- Data configuration engineers
- System maintenance engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
DANGER	Indicates a hazard with a high level or medium level of risk which, if not avoided, could result in death or serious injury.
WARNING	Indicates a hazard with a low level of risk which, if not avoided, could result in minor or moderate injury.
CAUTION	Indicates a potentially hazardous situation that, if not avoided, could result in equipment damage, data loss, performance deterioration, or unanticipated results.
TIP	Provides a tip that may help you solve a problem or save time.
NOTE	Provides additional information to emphasize or supplement important points in the main text.

Command Conventions

The command conventions that may be found in this document are defined as follows.

Convention	Description
Boldface	The keywords of a command line are in **boldface**.
Italic	Command arguments are in *italics*.
[ ]	Items (keywords or arguments) in brackets [ ] are optional.
{ x | y | ... }	Optional items are grouped in braces and separated by vertical bars. One item is selected.
[x | y | ...]	Optional items are grouped in brackets and separated by vertical bars. One item is selected or no item is selected.
{ x | y | ... }*	Optional items are grouped in braces and separated by vertical bars. A minimum of one item or a maximum of all items can be selected.
[x | y | ...]*	Optional items are grouped in brackets and separated by vertical bars. You can select one or several items, or select no item.
&<1-n>	The parameter before the & sign can be repeated 1 to n times.
#	A line starting with the # sign is comments.

Change History

Changes between document issues are cumulative. Therefore, the latest document version contains all updates made to previous versions.

Changes in Issue 04 (2013-04-15)

This version has the following updates:

The following information is added:

- **2.2.12 LDP over GRE/mGRE**

Changes in Issue 03 (2013-01-31)

This version has the following updates:

The following information is added:

- **3.2.10 DS-TE**

The following information is modified:

- Descriptions and figures of MPLS TE are optimized.

Changes in Issue 02 (2012-12-31)

This version has the following updates:

The following information is modified:

- Descriptions and figures are optimized, improving availability.

Changes in Issue 01 (2012-09-30)

Initial commercial release.

Contents

1 MPLS Overview

About This Chapter

1.1 Introduction to MPLS

Definition

Multiprotocol Label Switching (MPLS) is technology used on IP backbone networks. MPLS uses connection-oriented label switching on connectionless IP networks. By combining Layer 3 routing technologies and Layer 2 switching technologies, MPLS leverages flexibility of IP routing and simplicity of Layer 2 switching.

MPLS is based on the Internet Protocol version 4 (IPv4). The core MPLS technology can be extended to multiple network protocols, such as the Internet Protocol version 6 (IPv6), Internet Packet Exchange (IPX), and Connectionless Network Protocol (CLNP). Multiprotocol in MPLS means that multiple network protocols are supported.

In fact, the MPLS technology is a tunneling technology but not a service or an application. It supports multiple protocols and services. Moreover, it ensures security of data transmission.

Purpose

The IP-based Internet in the middle 1990s stimulated data growth. However, IP technology is inefficient in forwarding packets because software must search for routes using the longest match algorithm. As a result, the forwarding capability of IP technology becomes a bottleneck of the network development.

Asynchronous transfer mode (ATM) technology has been created from the evolution of network technologies. It uses labels (particularly, cells) of fixed length and maintains a label table that is much smaller than a routing table. Compared to IP technology, ATM technology is much more efficient in forwarding packets. ATM technology, however, is a complex protocol with high deployment costs, which hinders its popularity and growth.

Traditional IP technology, however, is simple with less deployment costs. People are eager to use technology that combines advantages of IP and ATM technologies. The MPLS technology is used.

Initially, MPLS was created to increase forwarding rates. Different from the manner in which packets are routed and forwarded using IP technology, MPLS analyzes a packet header only on the edge of the network rather than at each hop. In this manner, the packet processing time is shortened.

Application-specific integrated circuit (ASIC) technology has now been developed and the routing rate is no longer a bottleneck to the network development. As a result, MPLS no longer has the high-speed forwarding advantages. MPLS supports multi-layer labels, and its forwarding plane is connection-oriented. MPLS is widely used in virtual private network (VPN), traffic engineering (TE), and quality of service (QoS).

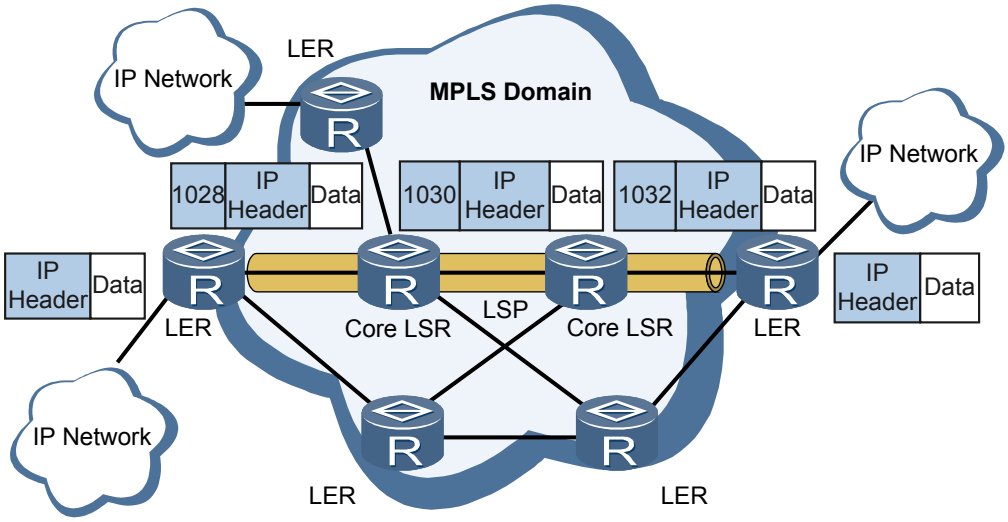
1.2 Principles

1.2.1 Basic MPLS Architecture

MPLS Network Structure

On a typical MPLS network shown in **Figure 1-1**, all routers function as label switching routers (LSRs) that exchange labels and forward packets. These LSRs construct an MPLS domain.LSRs that reside at the edge of the MPLS domain and connect to other networks are called label edge routers (LERs). LSRs within an MPLS domain are core LSRs.

Figure 1-1 MPLS network structure



On IP networks, packets are forwarded based on IP addresses; in MPLS domains, packets are forwarded based on labels.

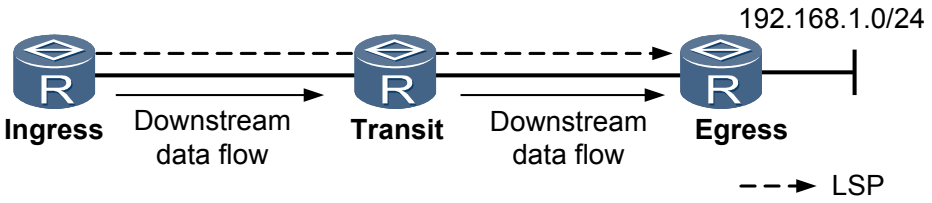
When receiving IP packets from the connected IP network, an LER tags labels on the packets and then forwards the labeled packets to a core LSR. When receiving labeled packets from the core LSR, the LER removes the labels and forwards the packets to the IP network. LSRs only forward packets based on labels.

LSPs are determined using different protocols and are established before packet forwarding. IP packets are transmitted through the specified label switched paths (LSPs) on an MPLS network.

As shown in **Figure 1-2**, an LSP is a unidirectional path whose direction is the same as the data flow. The nodes on an LSP include the ingress, transit, and egress nodes. The number of transit nodes on an LSP varies (none, one, or multiple), but only one ingress node and one egress node exist on the LSP.

To an LSR, all LSRs that send MPLS packets to the LSR are the upstream LSRs, and all next-hop LSRs that receive MPLS packets from the LSR are the downstream LSRs. As shown in **Figure 1-2**, for the data flow that are destined for 192.168.1.0/24, the ingress node is the upstream to the transit node, and the transit node is the downstream to the ingress node. Similarly, the transit node is the upstream to the egress node, and the egress node is the downstream to the transit node.

Figure 1-2 Upstream and downstream LSRs

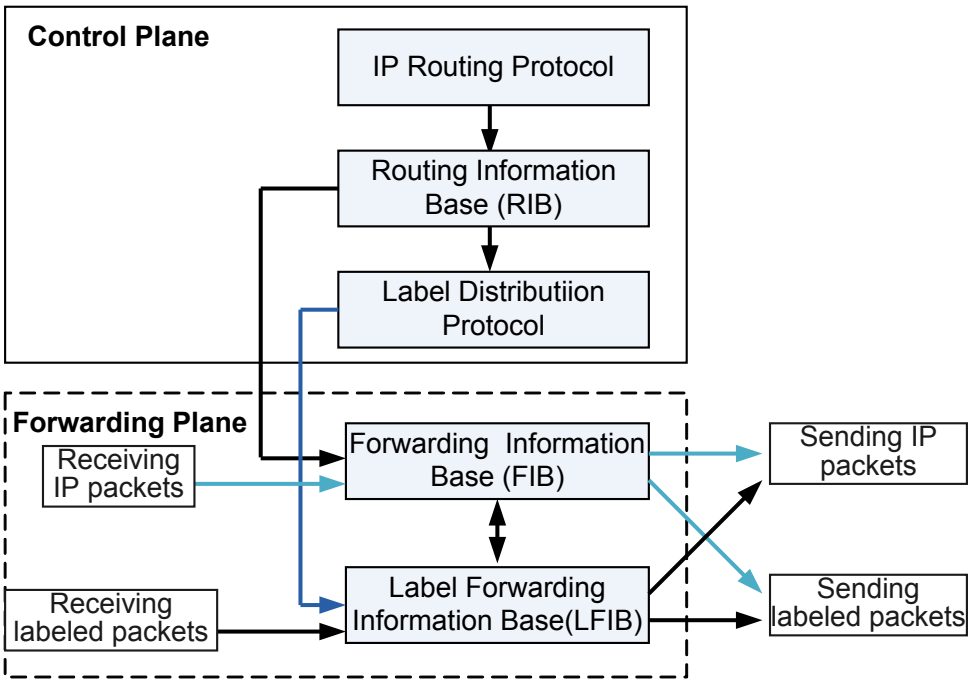


MPLS Architecture

The MPLS architecture consists of a control plane and a forwarding plane.

Figure 1-3 shows the MPLS architecture.

Figure 1-3 MPLS architecture



- The connectionless control plane generates and maintains routing information and labels.

  On the control plane, the IP Routing Protocol module transmits routing information and generates a routing information base (RIB); the Label Distribution Protocol module switches labels and establishes LSPs.

- The forwarding plane, also called data plane, is connection-oriented and forwards common IP packets and labeled MPLS packets.

  The forwarding plane consists of the modules IP forwarding information base (FIB) and label forwarding information base (LFIB). When receiving common IP packets, the forwarding plane forwards the packets based on the IP FIB or LFIB as required. When receiving labeled packets, the forwarding plane forwards the packets based on the LFIB.

If the destination locates on an IP network, the data plane removes the labels and forwards the packets based on the IP FIB.

1.2.2 MPLS Label

Forwarding Equivalence Class

Forwarding equivalence class (FEC) is a class-based forwarding technology that classifies the packets with the same forwarding mode based on the destination address or mask. Packets with the same FEC are forwarded in the same way on an MPLS network.

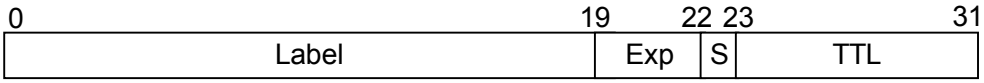
FEC can be defined based on the destination IP address and mask. For example, during IP forwarding, packets with the same destination belong to a FEC according to the longest match algorithm.

Label

A label is a short identifier that is 4 bytes long and has only local significance. It uniquely identifies a FEC to which a packet belongs. In some cases, such as load balancing, a FEC can be mapped to multiple incoming labels. Each label, however, represents only one FEC on a device.

Figure 1-4 shows the encapsulation structure of the label.

Figure 1-4 Structure of an MPLS label



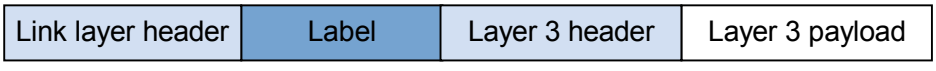
A label contains the following fields:

- Label: indicates the value field of a label. The length is 20 bits.
- Exp: indicates the bits used for extension. The length is 3 bits. Generally, this field is used for the class of service (CoS) that serves in a manner similar to Ethernet 802.1p.
- S: identifies the bottom of a label stack. The length is 1 bit. MPLS supports multiple labels, namely, the label nesting. When the S field is 1, the label is at the bottom of the label stack.
- TTL: indicates the time to live. The length is 8 bits. This field is the same as the TTL in IP packets.

Labels are encapsulated between the data link layer and the network layer. Labels can be supported by all data link layer protocols.

Figure 1-5 shows the position of the label in a packet.

Figure 1-5 Position of a label in a packet



Label Space

The label space is the value range of the label. The following describes the label space classification:

- 0 to 15: indicates special labels. For details about special labels, see **Table 1-1**.

Table 1-1 Special labels

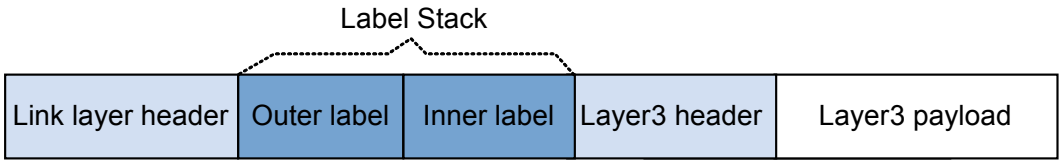
Label Value	Label	Description
0	IPv4 Explicit NULL Label	The label must be popped out, and the packets must be forwarded based on IPv4. If the egress node allocates a label whose value is 0 to the LSR at the penultimate hop, the LSR at the penultimate hop pushes label 0 to the top of the label stack and forwards the packet to the egress node. When the egress node recognizes that the value of the label carried in the packet is 0, the egress node pops it out. The label 0 is valid only at the bottom of the label stack.
1	Router Alert Label	A label that is only valid when it is not at the bottom of a label stack. The label is similar to the **Router Alert Option** field in IP packets. After receiving such a label, the node sends it to a local software module for further processing. Packet forwarding is determined by the next-layer label. If the packet needs to be forwarded continuously, the node pushes the Router Alert Label to the top of the label stack again.
2	IPv6 Explicit NULL Label	The label must be popped out, and the packets must be forwarded based on IPv6. If the egress node allocates a label with the value of 2 to the LSR at the penultimate hop, the LSR pushes label 2 to the top of the label stack and forwards the packet to the egress node. When the egress node recognizes that the value of the label carried in the packet is 2, the egress node immediately pops it out. The label 2 is valid only at the bottom of the label stack.
3	Implicit NULL Label	When the label with the value of 3 is swapped on an LSR at the penultimate hop, the LSR pops the label out and forwards the packet to the egress node. Upon receiving the packet, the egress node forwards the IP or VPN packet.
4 to 13	Reserved	None.
14	OAM Router Alert Label	A label for operation, administration and maintenance (OAM) packets over an MPLS network. MPLS OAM sends OAM packets to monitor LSPs and notify faults. OAM packets are transparent on transit nodes and the penultimate LSR.
15	Reserved	None.

- 16 to 1023: indicates the label space shared by static LSPs and static constraint-based routed LSPs (CR-LSPs).
- 1024 or above: indicates the label space for dynamic signaling protocols, such as Label Distribution Protocol (LDP), Resource Reservation Protocol-Traffic Engineering (RSVP-TE), and Multiprotocol Extensions for BGP (MP-BGP).

Label Stack

A label stack is a set of arranged labels. An MPLS packet can carry multiple labels at the same time. The label next to the Layer 2 header is called the top label or the outer label. The label next to the Layer 3 header is called the bottom label or inner label. Theoretically, MPLS labels can be nested without any limit.

Figure 1-6 Label stack



The label stack organizes labels according to the rule of Last-In, First-Out. The labels are processed from the top of the stack.

Label Operations

Information about basic label operations is a part of the label forwarding table. The operations are described as follows:

- Push: When an IP packet enters an MPLS domain, the ingress node adds a new label to the packet between the Layer 2 header and the IP header. Alternatively, an LSR adds a new label to the top of the label stack, namely, the label nesting.
- Swap: When a packet is transferred within the MPLS domain, a local node swaps the label at the top of the label stack in the MPLS packet for the label allocated by the next hop according to the label forwarding table.
- Pop: When a packet leaves the MPLS domain, the label is popped out of the MPLS packet. Alternatively, the top label of the label stack is popped out at the penultimate hop on an MPLS network to decrease the number of labels in the stack.

  In fact, the label is useless at the last hop of an MPLS domain. The penultimate hop popping (PHP) feature applies. On the penultimate node, the label is popped out of the packet to reduce the size of the packet that is forwarded to the last hop. Then, the last hop directly forwards the IP packet or forwards the packet by using the second label.

  PHP is configured on the egress node. The egress node supporting PHP allocates the label with the value of 3 to the penultimate hop.

The VPN Option C scenario supports the following action to process labels:

- Swappush: swaps an existing inner label for a new one and then pushes an outer label of other tunnel into a packet.

- Popgo: pops out an inner label from a packet and then pushes an outer label of other tunnel into the packet.

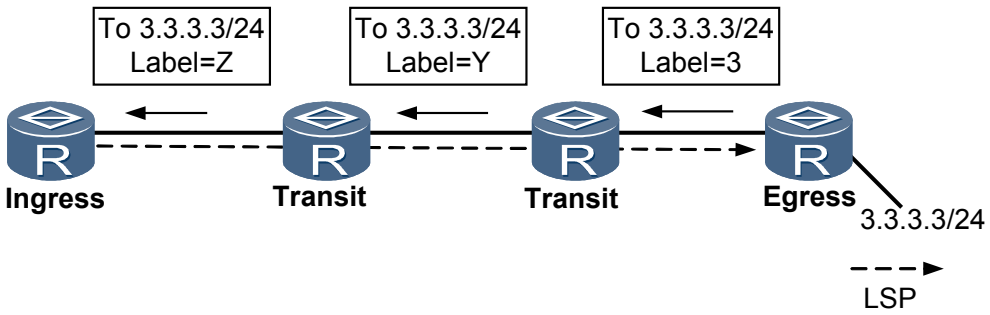
1.2.3 Establishing LSPs

Procedure for Establishing LSPs

Usually, MPLS allocates labels to packets and establishes an LSP through which MPLS forwards packets.

The downstream LSR allocate labels to packets sent to the upstream LSR. As shown in **Figure 1-7**, the downstream LSR identifies FEC based on the destination address, allocates a label to the specified FEC, and records the mapping between the label and FEC. The downstream LSR then encapsulates the mapping relationship into a message and sends it to the upstream LSR. A label forwarding table and an LSP are established.

Figure 1-7 Establishment of an LSP



LSPs are classified into the following types:

- Static LSP: set up by the administrator.
- Dynamic LSP: set up using the routing protocols and label distribution protocols.

Establishing Static LSPs

You can manually allocate labels to set up static LSPs. The value of the outgoing label of the upstream node is equal to the value of the incoming label of the downstream node.

The availability of a static LSP makes sense only for the local node that cannot detect the entire LSP.

A static LSP is set up without label distribution protocols or the exchanging of control packets. The static LSP costs little and is recommended for small-scale networks with the simple and stable topology. The static LSP cannot change with the network topology. Instead, it needs to be configured by an administrator.

Establishing Dynamic LSPs

Dynamic LSPs are established using label distribution protocols. As the control protocol or signaling protocol for MPLS, a label distribution protocol defines FECs, distributes labels, and establishes and maintains LSPs.

The following label distribution protocols apply to an MPLS network.

- LDP

  LDP is defined to distribute labels and used to dynamically establish LSPs. An LSR can use LDP to map routing information on the network layer to the LSP on the data link layer.

  For details about LDP, see **MPLS LDP**.

- RSVP-TE

  RSVP-TE is an extension to RSVP and used to establish or delete constraint-based LSPs.

  For details about RSVP-TE, see **MPLS TE**.

- MP-BGP

  MP-BGP is an extension to BGP and allocates labels to MPLS VPN routes and inter-AS VPN routes.

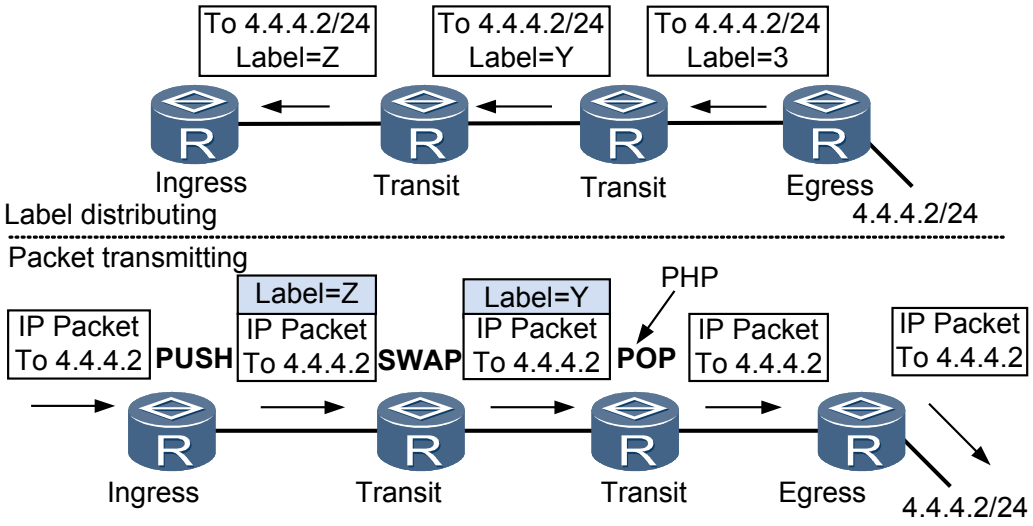
  For details about MP-BGP, see *Feature Description - IP Routing*.

1.2.4 MPLS Forwarding

MPLS Forwarding Principle

The LSP that supports the PHP is used in the following example to describe how MPLS packets are forwarded.

Figure 1-8 MPLS label distribution and packet forwarding



As shown in **Figure 1-8**, an LSP whose FEC is identified by the destination address 4.4.4.2/24 is set up on an MPLS network. MPLS packets are forwarded as follows:

1. The ingress node receives an IP packet destined for 4.4.4.2. Then, the ingress node adds Label Z to the packet and forwards it.
2. The transit node receives the labeled packet and swaps labels by popping Label Z out and pushing Label Y into the packet.
3. A transit node at the penultimate hop receives the packet with Label Y. The transit node pops Label Y out because the label value is 3. The transit node then forwards the packet to the egress node as an IP packet.

4. The egress node receives the IP packet and forwards it to 4.4.4.2/24.

Process of MPLS Packet Forwarding

- NHLFE

  The next hop label forwarding entry (NHLFE) can guide MPLS packet forwarding.

  An NHLFE contains the following information:

  - Tunnel ID
  - Outbound interface
  - Next hop
  - Outgoing label
  - Label operation

- FTN

  FTN is a short form of FEC-to-NHLFE. The FTN indicates the mapping between a FEC and a set of NHLFEs.

  Details about the FTN can be obtained by searching for the Tunnel ID values that are not 0x0 in a FIB. The FTN is available on the ingress only.

- ILM

  The incoming label map (ILM) indicates the mapping between an incoming label and a set of NHLFEs.

  The ILM contains the following information:

  - Tunnel ID
  - Incoming label
  - Inbound interface
  - Label operation

  The ILM on a transit node can bind the labels to NHLFEs. The function of an ILM table is similar to the FIB that is searched according to destination IP addresses. Therefore, you can obtain all label forwarding information by searching an ILM table.

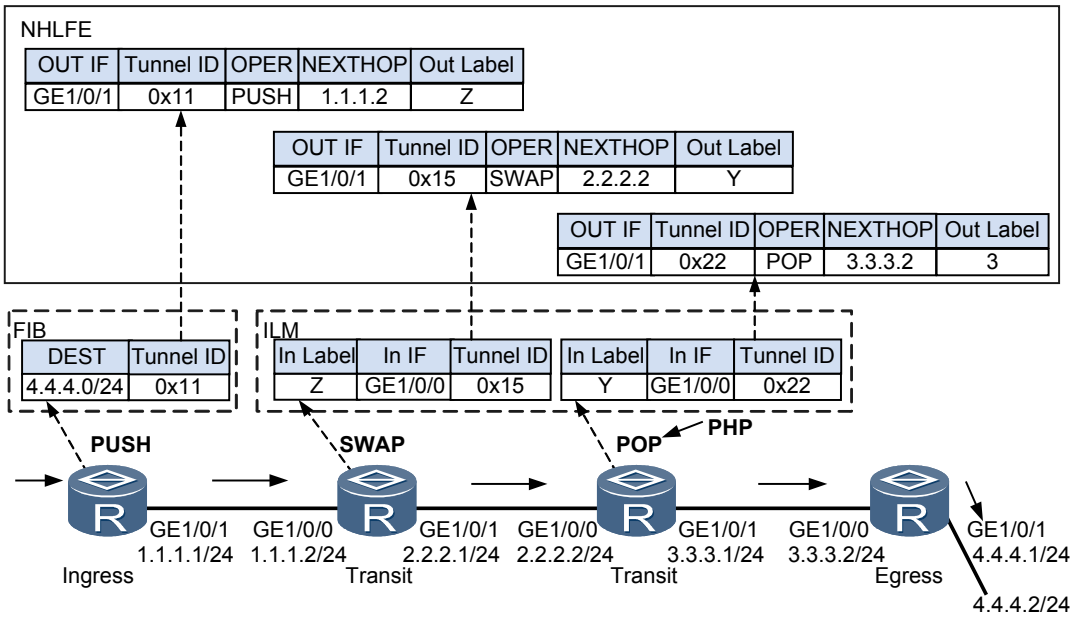
- Tunnel ID

  To provide the same interface of a tunnel used by upper layer applications such as the VPN and route management, the system automatically allocates an ID to each tunnel, referred to as the tunnel ID. The tunnel ID is 32 bits long and is valid only on the local end.

When an IP packet enters an MPLS domain, the ingress node searches the FIB to check whether the tunnel ID corresponding to the destination IP address is 0x0.

- If the tunnel ID is 0x0, the packet is forwarded along the IP link.
- If the tunnel ID is not 0x0, the packet is forwarded along an LSP.

Figure 1-9 Process of MPLS packet forwarding



MPLS packets are forwarded as follows on nodes along an LSP:

- The ingress node searches the FIB and NHLFE tables.
- The transit node searches the ILM and NHLFE tables.
- The egress node searches the ILM table or RIB.

During MPLS forwarding, FIB entries, ILM entries, and NHLFEs are associated with each other through the tunnel ID.

- Forwarding on the ingress node

  The ingress node processes the forwarding of MPLS packets as follows:

  1. Searches the FIB and finds the tunnel ID corresponding to the destination IP address.
  2. Finds the NHLFE corresponding to the tunnel ID in the FIB and associates the FIB entry with the NHLFE entry.
  3. Checks the NHLFE for information about the outbound interface, next hop, outgoing label, and label operation type. The label operation type is Push.
  4. Pushes the obtained label into IP packets, processes the EXP field according to QoS policy and TTL field, and sends the encapsulated MPLS packets to the next hop.

- Forwarding on the transit node

  The transit node forwards the received MPLS packets as follows:

  1. Checks the ILM table corresponding to an MPLS label and finds the Tunnel ID.
  2. Finds the NHLFE corresponding to the Tunnel ID in the ILM table.
  3. Checks the NHLFE for information about the outbound interface, next hop, outgoing label, and label operation type.
  4. Processes the MPLS packets according to the specific label value:

  - If the label value is equal to or greater than 16, a new label replaces the label in the MPLS packet. At the same time, the EXP field and TTL field are processed. The MPLS packet with the new label is forwarded to the next hop.
  - If the label value is 3, the label is popped out of the MPLS packet. At the same time, the EXP field and TTL field are processed. The packet is forwarded through IP routes, or in accordance with its next layer label.
- Forwarding on the egress node
  - When the egress node receives IP packets, it checks the FIB and performs IP forwarding.
  - When the egress node receives MPLS packets, it checks the ILM table for the label operation type. At the same time, the egress node processes the EXP field and TTL field.
    - When the S field in the label is equal to 1, the label is the stack's bottom label and the packet is directly forwarded through IP routes.
    - When the S field in the label is equal to 0, a next-layer label exists and the packet is forwarded according to the next layer label.

1.2.5 MPLS TTL Processing

This section describes how MPLS processes the TTL and responds to TTL timeout.

MPLS TTL Processing Modes

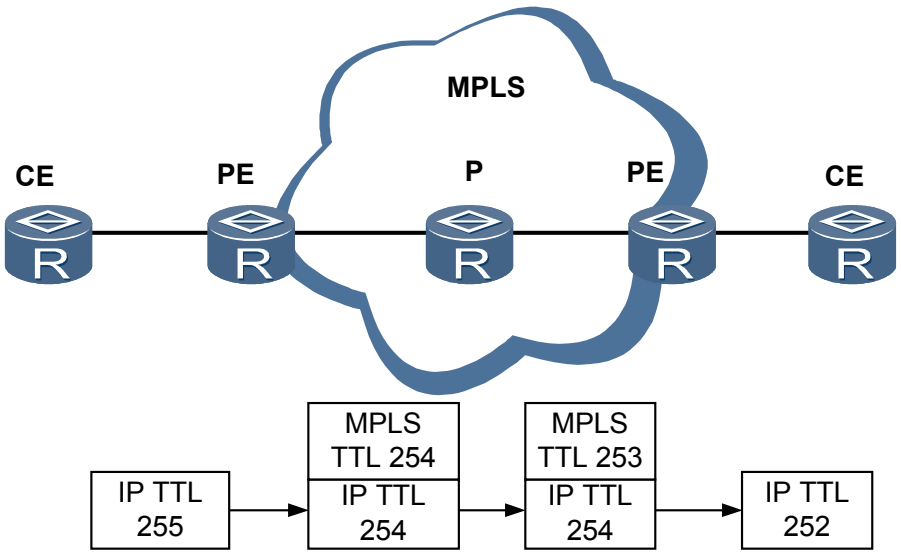
The TTL field in an MPLS label is 8 bits long. The TTL field is the same as that in an IP packet header. MPLS processes the TTL to prevent loops and implement traceroute.

RFC 3443 defines two modes to process the TTL in MPLS packets: Uniform mode and Pipe mode. By default, MPLS processes the TTL in Uniform mode.

- Uniform mode

  When IP packets enter an MPLS network, the ingress node decreases the IP TTL by one and copies it to the MPLS TTL field. The TTL field in MPLS packets is processed in standard mode. The egress node decreases the MPLS TTL by one and maps it to the IP TTL field. **Figure 1-10** shows how the TTL field is processed on the transmission path.

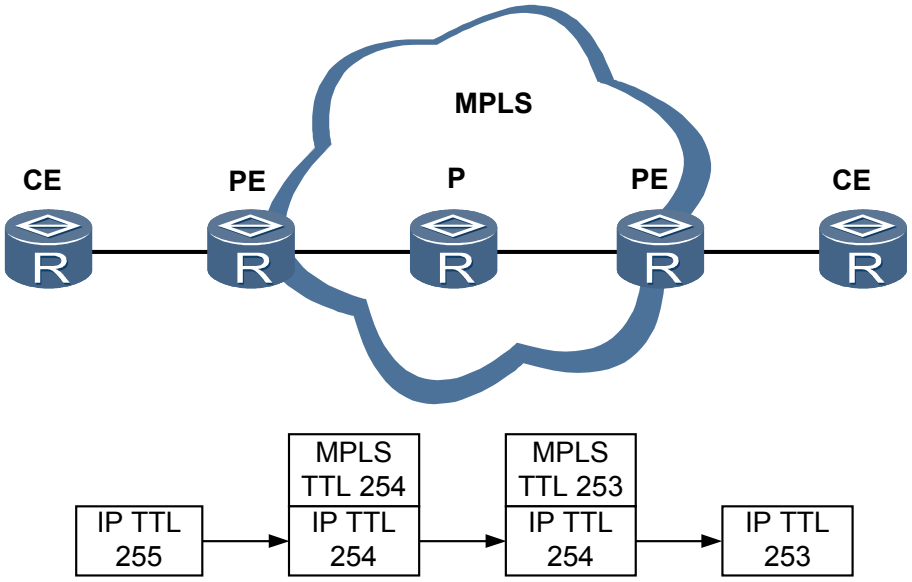
  Figure 1-10 TTL processing in Uniform mode



- Pipe mode

  As shown in **Figure 1-11**, the ingress node decreases the IP TTL by one and the MPLS TTL is constant. The TTL field in MPLS packets is processed in standard mode. The egress node decreases the IP TTL by one. In Pipe mode, the IP TTL only decreases by one on the ingress node and one on the egress node when packets travels across an MPLS network.

  Figure 1-11 TTL processing in Pipe mode



In MPLS VPN applications, the MPLS backbone network needs to be hidden to ensure network security. The Pipe mode is recommended for private network packets.

TTL Timeout Responding

On an MPLS network, an LSR receives labeled MPLS packets. The LSR generates an ICMP TTL-expired message when the TTL of an MPLS packet times out.

The LSR returns the TTL-expired message to the sender in the following ways:

- If the LSR has a reachable route to the sender, it directly sends the TTL-expired message to the sender through the IP route.
- If the LSR has no reachable route to the sender, it forwards the TTL-expired message along the LSP. The egress node forwards the TTL-expired message to the sender.

In most cases, the received MPLS packet contains only one label and the LSR responds to the sender with the TTL-expired message using the first method. If the MPLS packet contains multiple labels, the LSR uses the second method.

The MPLS VPN packets may contain only one label when they arrive at an autonomous system boundary router (ASBR) on the MPLS VPN, a superstratum PE (SPE) device in HoVPN networking, or a PE device in the VPN nesting networking. These devices have no IP routes to the sender, so they use the second method to reply to the TTL-expired messages.

1.2.6 MPLS QoS Implementation

MPLS QoS, an important part in the deployment of QoS services, implements QoS using the Differentiated Services (DiffServ) model in actual MPLS networking. MPLS QoS differentiates data flows based on the EXP field value, which ensures low delay and low packet loss ratio for voice and video data streams and increases network resource efficiency.

MPLS DiffServ

In the DiffServ model, network edge nodes map a service to a service class based on the QoS requirements of the service and use the DS field (ToS field) in IP packets to identify the service. Nodes on the backbone network apply preset policies to the service based on the DS field to ensure service quality. The service classification and label mechanism of DiffServ are similar to label distribution of MPLS. MPLS DiffServ combines DS distribution and MPLS label distribution.

MPLS DiffServ is implemented as the EXP field in an MPLS packet header carriers DiffServ per-hop behavior (PHB). An LSR must consider the MPLS EXP value when determining the forwarding policy. MPLS DiffServ provides the following plans for determining PHBs:

- E-LSP: an LSP whose PHB is determined by the EXP field. E-LSP applies to a network with less than eight PHBs. In this plan, a differentiated services code point (DSCP) is mapped to a specified EXP that identifies a PHB. Packets are forwarded based on labels, while the EXP field determines the scheduling type and drop priority at each hop. An LSP transmits a maximum of eight PHB flows that are differentiated based on the EXP field in the MPLS packet header. The EXP field can be determined by the ISP or mapped from the DSCP value in a packet. In this plan, PHB information does not need to be transmitted by signaling protocols, the label efficiency is high, and the label status is easy to maintain.
- L-LSP: an LSP whose PHB is determined by both the label and EXP field. L-LSP applies to a network with any number of PHBs. During packet forwarding, the label of a packet determines the forwarding path and scheduling type, while the EXP field determines the drop priority of the packet. Labels differentiate service flows, so multiple service flows can be transmitted over one LSP. This plan requires more labels and so occupies a large number of system resources.

NOTE

Currently, only the E-LSP plan is supported.

MPLS DiffServ Modes

An MPLS network provides tunnels for services. MPLS L3VPN DiffServ modes include: pipe, short pipe, and uniform.

- Pipe: The EXP field value that the ingress node adds to the MPLS label of packets is specified by the user. If the EXP field value of the packet is changed on the MPLS network, the change is valid only on the MPLS network. The egress node selects the PHB according to the EXP field value of the packet. When the packet leaves the MPLS network, the previous DSCP value becomes effective again.
- Short pipe: The EXP field value that the ingress node adds to the MPLS label of packets is specified by the user. If the EXP field value of the packet is changed on the MPLS network, the change is valid only on the MPLS network. The egress node selects the PHB according to the DSCP field value of the packet. When the packet leaves the MPLS network, the previous DSCP value becomes effective again.
- Uniform: The priorities of packets on the IP network and the MPLS network are uniformly defined, so the priorities of the packets on the two networks are globally valid. At the ingress

node, each packet is assigned a label and the lower 3 bits in the DSCP field are mapped to the EXP field. A change in the value of the EXP field on the MPLS network determines the PHB used when the packet leaves the MPLS network. The egress node maps the EXP field to the DSCP field.

On an L2VPN, the MPLS label is in the outer layer of an encapsulated packet. Therefore, the 802.1p field of VLAN packets needs to be mapped to the EXP field.

1.2.7 MPLS Ping/Tracert

Introduction to MPLS Ping/Tracert

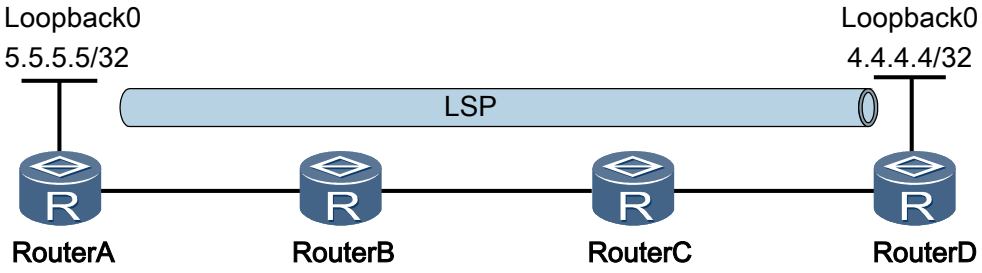
On an MPLS network, the control panel used for setting up an LSP cannot detect the failure in data forwarding of the LSP. This makes network maintenance difficult. The MPLS ping and tracert mechanisms detect LSP errors and locate faulty nodes.

MPLS ping is used to check network connectivity and host reachability. MPLS tracert is used to check the network connectivity and host reachability, and to locate network faults. Similar to IP ping and tracert, MPLS ping and tracert use MPLS echo request packets and MPLS echo reply packets to check LSP availability. MPLS echo request packets and echo reply packets are both encapsulated into User Datagram Protocol (UDP) packets. The UDP port number of the MPLS echo request packet is 3503, which can be identified only by MPLS-enabled devices.

An MPLS echo request packet carries FEC information to be detected, and is sent along the same LSP as other packets with the same FEC. In this manner, the connectivity of the LSP is checked. MPLS echo request packets are forwarded to the destination end using MPLS, while MPLS echo reply packets are forwarded to the source end using IP. Routers set the destination address in the IP header of the MPLS echo request packets to 127.0.0.1/8 (local loopback address) and the TTL value is 1. In this way, MPLS echo request packets are not forwarded using IP forwarding when the LSP fails so that the failure of the LPS can be detected.

MPLS Ping

Figure 1-12 MPLS network



As shown in **Figure 1-12**, RouterA establishes an LSP to RouterD. RouterA performs MPLS ping on the LSP by performing the following steps:

1. RouterA checks whether the LSP exists. (On a TE tunnel, the router checks whether the tunnel interface exists and the CR-LSP has been established.) If the LSP does not exist, an error message is displayed and the MPLS ping stops. If the LSP exists, RouterA performs the following operations.
2. RouterA creates an MPLS echo request packet and adds 4.4.4.4 to the destination FEC stack in the packet. In the IP header of the MPLS echo request packet, the destination

address is 127.0.0.1/8 and the TTL value is 1. RouterA searches for the corresponding LSP, adds the LSP label to the MPLS echo request packet, and sends the packet to RouterB.

3. Transit nodes RouterB and RouterC forward the MPLS echo request packet based on MPLS. If MPLS forwarding on a transit node fails, the transit node returns an MPLS echo reply packet carrying the error code to RouterA.
4. If no fault exists along the MPLS forwarding path, the MPLS echo request packet reaches the LSP egress node RouterD. RouterD returns a correct MPLS echo reply packet after verifying that the destination IP address 4.4.4.4 is the loopback interface address. MPLS ping is complete.

MPLS Tracert

As shown in **Figure 1-12**, RouterA performs MPLS tracert on RouterD (4.4.4.4/32) by performing the following steps:

1. RouterA checks whether an LSP exists to RouterD. (On a TE tunnel, the router checks whether the tunnel interface exists and the CR-LSP has been established.) If the LSP does not exist, an error message is displayed and the tracert stops. If the LSP exists, RouterA performs the following operations.
2. RouterA creates an MPLS echo request packet and adds 4.4.4.4 to the destination FEC stack in the packet. In the IP header of the MPLS echo request packet, the destination address is 127.0.0.1/8. Then RouterA adds the LSP label to the packet, sets the TTL value to 1, and sends the packet to RouterB. The MPLS echo request packet contains a downstream mapping TLV that carries downstream information about the LSP at the current node, such as next-hop address and outgoing label.
3. Upon receiving the MPLS echo request packet, RouterB decreases the TTL by one and finds that TTL times out. RouterB then checks whether the LSP exists and the next-hop address and whether the outgoing label of the downstream mapping TLV in the packet is correct. If so, RouterB returns a correct MPLS echo reply packet that carries the downstream mapping TLV of RouterB. If not, RouterB returns an incorrect MPLS echo reply packet.
4. After receiving the correct MPLS echo reply packet, RouterA resends the MPLS echo request packet that is encapsulated in the same way as step 2 and sets the TTL value to 2. The downstream mapping TLV of this MPLS echo request packet is replicated from the MPLS echo reply packet. RouterB performs common MPLS forwarding on this MPLS echo request packet. If TTL times out when RouterC receives the MPLS echo request packet, RouterC processes the MPLS echo request packet and returns an MPLS echo reply packet in the same way as step 3.
5. After receiving a correct MPLS echo reply packet, RouterA repeats step 4, sets the TTL value to 3, replicates the downstream mapping TLV in the MPLS echo reply packet, and sends the MPLS echo request packet. RouterB and RouterC perform common MPLS forwarding on this MPLS echo request packet. Upon receiving the MPLS echo request packet, RouterD repeats step 3 and verifies that the destination IP address 4.4.4.4 is the loopback interface address. RouterD returns an MPLS echo reply packet that does not carry the downstream mapping TLV. MPLS tracert is complete.

When routers return the MPLS echo reply packet that carries the downstream mapping TLV, RouterA obtains information about each node along the LSP.

1.3 Applications

1.3.1 MPLS-based VPN

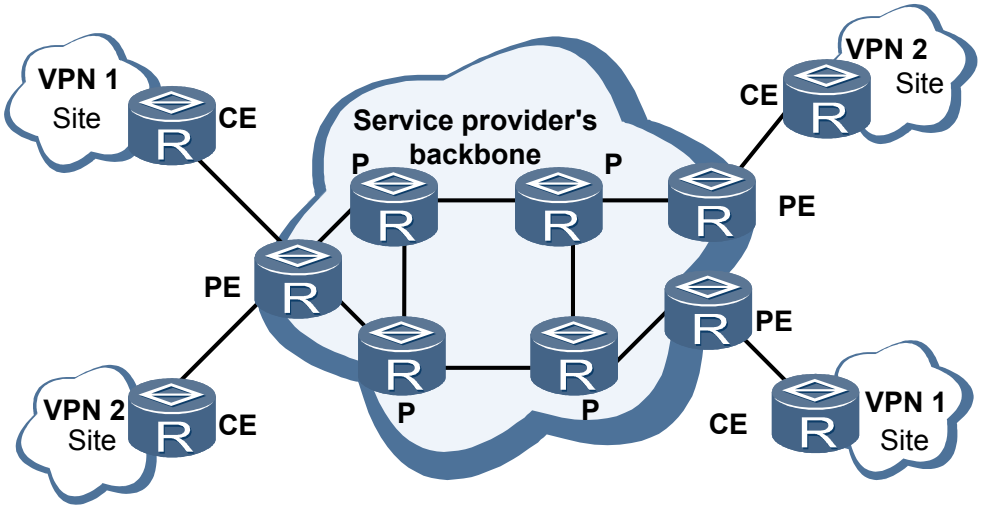
The traditional VPN transmits private network data over the public network using tunneling protocols, such as the Generic Routing Encapsulation (GRE), Layer 2 Tunneling Protocol (L2TP), and Point to Point Tunneling Protocol (PPTP).

An MPLS-based VPN has similar security as a frame relay (FR). Devices on the MPLS-based VPN do not require the configuration of the GRE or L2TP tunnel. Network delay is minimized because datagrams are not encapsulated or encrypted.

As shown in **Figure 1-13**, the MPLS-based VPN integrates private network branches through an LSP to form a unified network. The MPLS-based VPN controls the interconnection between VPNs. **Figure 1-13** shows the devices on the MPLS-based VPN.

- Customer edge (CE) is an edge device on a customer network. The CE can be a router, a switch, or a host.
- Provider edge (PE) is an edge device on a service provider network.
- Provider (P) is a backbone device on an SP network. A P is not directly connected to CEs. Ps only need to possess basic MPLS forwarding capabilities and do not maintain information about a VPN.

Figure 1-13 MPLS-based VPN



An MPLS-based VPN has the following characteristics:

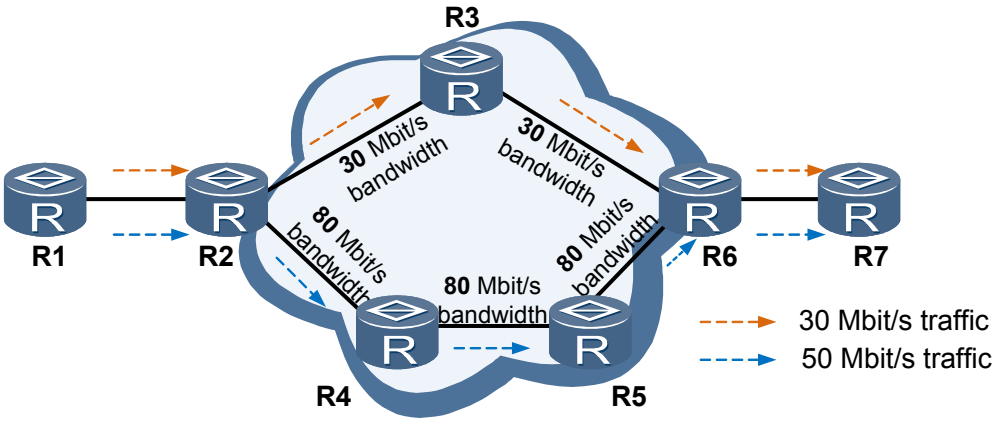
- PEs manage VPN users, set up LSPs between PEs, and allocate routes to sites on a VPN.
- The route allocation between PEs is implemented by LDP or MP-BGP.
- The MPLS-based VPN supports IP address multiplexing between sites as well as the interconnection of different VPNs.

1.3.2 MPLS-based TE

On traditional IP networks, routers select the shortest path as the route regardless of other factors such as bandwidth. Traffic on a path is not switched to other paths even if the path is congested. As more applications are deployed on the Internet, this shortest path first rule causes severe problems on networks.

Traffic engineering (TE) adjusts parameters including traffic management, routing, and resource restraint parameters in real time to dynamically monitor the network traffic and the load of the network components, which prevents network congestion caused by unbalanced traffic distribution.

Figure 1-14 MPLS-based TE



As shown in **Figure 1-14**, two paths are set up between R1 and R7: R1 -> R2 -> R3 -> R6 -> R7 and R1 -> R2 -> R4 -> R5 -> R6 -> R7. Bandwidth of the first path is 30 Mbit/s, and bandwidth of the second path is 80 Mbit/s. TE allocates traffic properly based on bandwidth, preventing link congestion. For example, 30 Mbit/s and 50 Mbit/s services are running between R1 and R7. TE distributes the 30 Mbit/s traffic to the 30 Mbit/s path and the 50 Mbit/s traffic to the 80 Mbit/s path.

The following characteristics of MPLS make TE implementation possible:

- Explicit paths can be specified for LSPs.
- Label forwarding is easier to manage and maintain than IP forwarding.
- MPLS TE occupies fewer resources than other TE implementations.

The MPLS TE technology integrates the MPLS technology with TE. MPLS TE can reserve resources by setting up LSPs along a specified path to prevent network congestion and balance network traffic. MPLS TE has the following advantages:

- MPLS TE can reserve resources to ensure the quality of services during the establishment of LSPs.
- The behaviors of an LSP can be easily controlled based on the attributes of the LSP such as priority and bandwidth.
- LSP establishment consumes a few resources and does not affect other network services.
- MPLS allows traffic aggregation and disaggregation, which is more flexible than IP forwarding.
- Backup path and fast reroute (FRR) protect the network communication upon a failure of a link or a node.

These advantages make MPLS TE the optimal TE solution. MPLS TE allows service providers (SPs) to fully leverage existing network resources to provide diverse services, optimize network resources, and efficiently manage the network.

1.3.3 MPLS-based 6PE

IPv6 Provider Edge (6PE) is a technology for transition from IPv4 to IPv6. The 6PE technology allows ISPs to provide access services for scattered IPv6 networks over the existing IPv4 backbone network. In this way, CEs on IPv6 islands can communicate with each other through existing IPv4 PEs.

On an MPLS 6PE network shown in **Figure 1-15**:

- 6PE routers exchange IPv6 routing information with CEs using IPv6 routing protocols.
- 6PE routers exchange IPv6 routing information with each other using MP-BGP and allocate MPLS labels to IPv6 prefixes.
- 6PE routers exchange IPv4 routing information with Ps using IPv4 routing protocols and establish LSPs between 6PE routers and Ps using MPLS.

Figure 1-15 Process of packet forwarding using MPLS 6PE

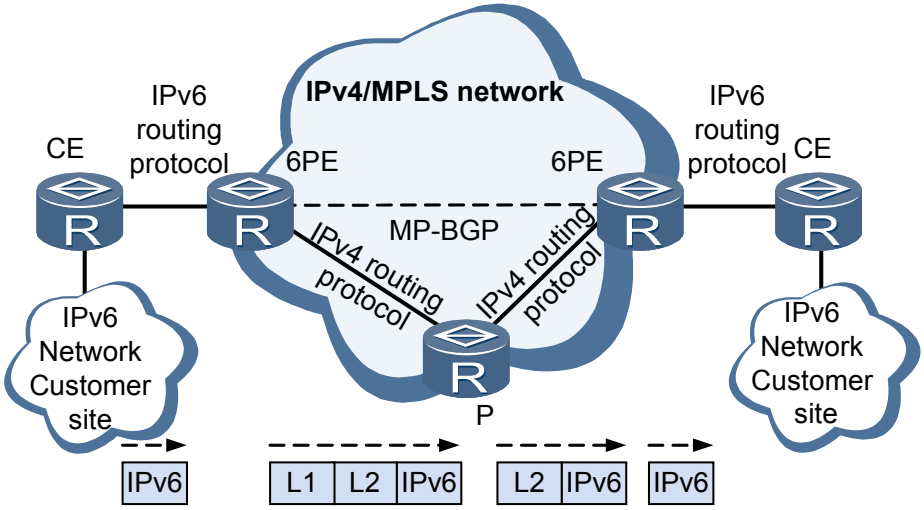


Figure 1-15 shows the process of IPv6 packet forwarding on an MPLS 6PE network. IPv6 packets must carry outer and inner labels when being forwarded on the IPv4 backbone network. The inner label maps the IPv6 prefix, while the outer label maps the LSP between 6PEs.

The MPLS 6PE technology allows ISPs to connect existing IPv4/MPLS networks to IPv6 networks by simply upgrading PEs. To ISPs, the MPLS 6PE technology is an efficient solution for transition to IPv6.

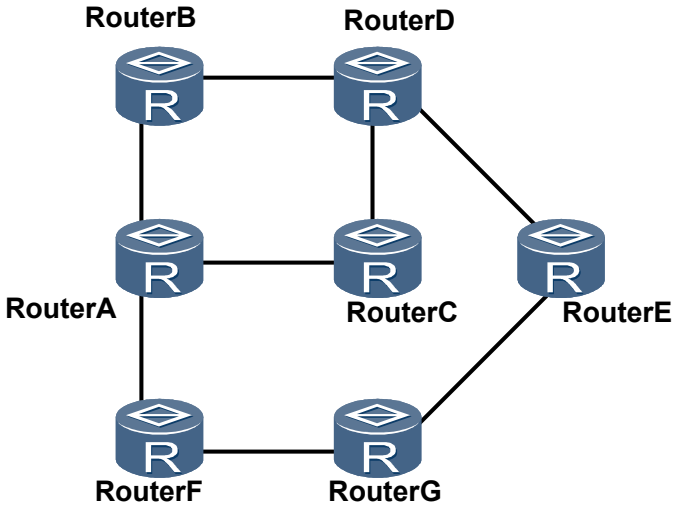
1.3.4 PBR to an LSP

Policy-based routing (PBR) selects a route according to a user-defined policy for security and load balancing. The router supports the PBR to an LSP. On an MPLS network, IP packets that meet the filtering policy can be forwarded through a specified LSP.

In **Figure 1-16**, RouterA, RouterB, RouterC, RouterD, and RouterE are on the existing network. RouterF and RouterG are added to provide new services. Traffic is forwarded as follows:

- Traffic for existing services is forwarded through the existing network.
- Traffic for new services is forwarded by RouterF and RouterG.

Figure 1-16 Application of the PBR to an LSP



To forward some traffic of new services through the existing network, configure the PBR to an LSP on RouterA. In this manner, traffic meeting the specified policy can be forwarded through the LSP on the existing network.

You can also use the PBR to the LSP with LDP FRR to divert some traffic to the backup LSP for load balancing when the backup LSP is idle relatively.

1.4 References

The following table lists the references.

Document No.	Description
RFC3031	Multiprotocol Label Switching Architecture
RFC3036	LDP Specification
RFC3032	MPLS Label Stack Encoding
RFC3443	Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks
RFC3034	Use of Label Switching on Frame Relay Networks Specification
RFC2702	Requirements for Traffic Engineering Over MPLS
RFC3209	RSVP-TE: Extensions to RSVP for LSP Tunnels
RFC4364	BGP/MPLS IP Virtual Private Networks (VPNs)
RFC2598	An Expedited Forwarding PHB

2 MPLS LDP

About This Chapter

2.1 Introduction to MPLS LDP

Definition

The Label Distribution Protocol (LDP) is a control protocol of Multiprotocol Label Switching (MPLS), which functions similarly to a signaling protocol on a traditional network. It classifies FECs, distributes labels, and establishes and maintains LSPs. LDP defines messages in the label distribution process as well as procedures for processing these messages.

Purpose

MPLS supports multiple labels and its forwarding plane is connection-oriented, and thus this excellent scalability enables the MPLS/IP-based network to provide various services. Through LDP, Label Switching Routers (LSRs) directly map routing information at the network layer to the switched paths at the data link layer, and thus establish LSPs at the network layer.

Currently, LDP is widely used to provide VPN services because it features simple networking and configurations, supports route-based establishment of LSPs, and supports high-capacity LSPs.

2.2 Principles

2.2.1 Basic Concepts

LDP Adjacency

When an LSR receives a Hello message from a peer, an LDP peer may exist. An LDP adjacency can be created to maintain the presence of the peer. There are two types of LDP adjacencies:

- Local adjacency: The adjacency is discovered by exchanging Link Hello messages.
- Remote adjacency: The adjacency is discovered by exchanging Target Hello messages.

LDP Peers

LDP peers refer to two LSRs that use LDP to set up an LDP session and then exchange label messages.

LDP peers learn labels from each other using the LDP session between them.

LDP Session

LSRs in an LDP session exchange messages such as label mapping messages and label release messages. LDP sessions are classified into the following types:

- Local LDP session: The LDP session is set up between local adjacencies. The two LSRs setting up the local LDP session are directly connected.
- Remote LDP session: The LDP session is set up between remote adjacencies. The two LSRs setting up the remote LDP session can be either directly or indirectly connected.

NOTE

LDP maintains the presence of peers using adjacencies. The type of peers depends on the type of adjacencies. A pair of peers can be maintained by multiple adjacencies. If a pair of peers is maintained by both local and remote adjacencies, the peers support coexistence of the local and remote adjacencies. An LDP session can only be established if such pairs of peers exist.

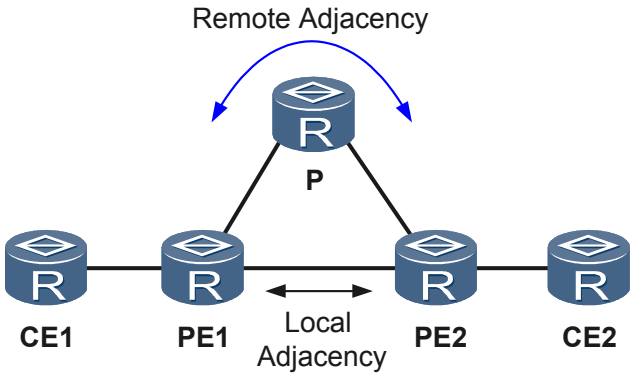
A local and a remote LDP session can be set up simultaneously.

The principle is that the local and remote LDP adjacencies can be connected to the same peer so that the peer is maintained by both the local and remote LDP adjacencies.

As shown in **Figure 2-1**, when the local LDP adjacency is deleted due to a failure on the link to which the adjacency is connected, the peer's type may change without affecting its presence or status. (The peer type is determined by the adjacency type. The types of adjacencies include local, remote, and coexistent local and remote.)

If the link becomes faulty or is recovering from a fault, the peer type may change while the type of the session associated with the peer changes accordingly. However, the session is not deleted and does not become Down. Instead, the session remains Up.

Figure 2-1 Networking diagram for a coexistent local and remote LDP session



A coexistent local and remote LDP session is typically applied to L2VPN. As shown in **Figure 2-1**, L2VPN services are transmitted between PE1 and PE2. When the directly-connected link between PE1 and PE2 recovers after being disconnected, the processing is as follows:

1. MPLS LDP is enabled on the directly-connected PE1 and PE2, and a local LDP session is set up between PE1 and PE2. PE1 and PE2 are configured as the remote peer of each other, and a remote LDP session is set up between PE1 and PE2. Local and remote adjacencies are then set up between PE1 and PE2. Since now, both local and remote LDP sessions exist between PE1 and PE2. L2VPN signaling messages are transmitted through the compatible local and remote LDP session.
2. When the physical link between PE1 and PE2 becomes Down, the local LDP adjacency also goes Down. The route between PE1 and PE2 is still reachable through the P, indicating that the remote LDP adjacency remains Up. The session changes to a remote session so that it can remain Up. The L2VPN does not detect the change in session status and therefore does not delete the session. This prevents the L2VPN from having to disconnect and recover services, and shortens service interruption time.
3. When the fault is rectified, the link between PE1 and PE2 as well as the local LDP adjacency can go Up again. The session changes to the compatible local and remote LDP session and

remains Up. Again, the L2VPN will not detect the change in session status and therefore does not delete the session. This shortens service interruption time.

Type of LDP Messages

LDP messages are classified into the following types:

- Discovery message: used to notify and maintain the existence of an LSR on a network.
- Session message: used to establish, maintain, and terminate sessions between LDP peers.
- Advertisement message: used to create, modify, and delete label mappings for FECs.
- Notification message: used to provide advisory and error information.

To ensure the reliability of message transmission, LDP uses the TCP transport for Session, Advertisement, and Notification messages. LDP uses the UDP transport only for transmitting the Discovery message.

Label space

A label space is a range of labels allocated between LDP peers, which can be categorized as follows:

- Per-platform label space: An entire LSR uses one label space. Currently, per-platform label space is mostly used.
- Per-interface label space: Each interface of an LSR is assigned a label space.

LDP identifier

An LDP identifier identifies the label space used by a specified LSR. An LDP identifier is 6 bytes in the format <LSR ID>:<Label space ID>.

- LSR ID: indicates the 4-byte LSR identifier.
- Label space ID: indicates the 2-byte label space identifier. The value 0 indicates the per-platform label space, while the value non-0 indicates the per-interface label space.

For example, the LDP ID is 192.168.1.1:0, indicating that the LSR ID is 192.168.1.1 and per-platform label space is used.

2.2.2 LDP Working Mechanism

LDP defines the label distribution process and messages transmitted during label distribution. An LSR can use LDP to map routing information on the network layer to on the data link layer, setting up an LSP. LDP working process goes through the following phases:

1. After discovering a neighbor, an LSR sets up an LDP session.
2. After the session is established, LDP notifies LDP adjacencies of the mappings between FECs and labels and sets up an LSP.

   RFC 5036 defines the label advertisement mode, label distribution control mode, and label retention mode to determine how the LSR advertises and manages labels.

LDP Session

LDP Discovery Mechanisms

LDP discovery mechanisms are used by LSRs to discover potential LDP peers. LDP discovery mechanisms are classified into the following types:

- Basic discovery mechanism: used to discover directly-connected LSR peers on a link.

  An LSR periodically sends LDP Hello messages to implement the mechanism and establish a local LDP session.

  The Hello messages are encapsulated in UDP packets with the multicast destination address and sent through LDP port 646. A Hello message carries an LDP ID and other information (such as the hello-hold time and the transport address). If an LSR receives an LDP Hello message on a specified interface, a potential LDP peer is connected to the same interface.
- Extended discovery mechanism: used to discover the LSR peers that are not directly connected on a link.

  An LSR periodically sends Target Hello messages to a specified destination address according to the mechanism to establish a remote LDP session.

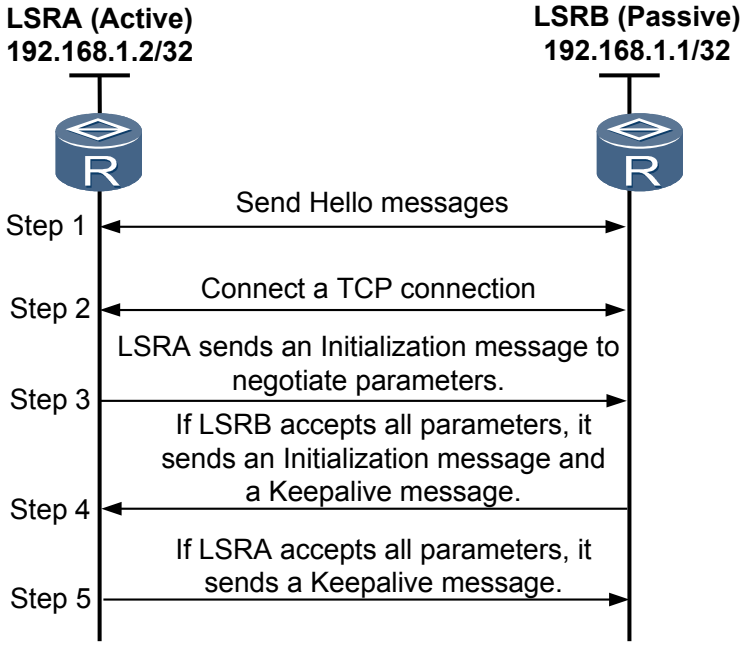
  The Target Hello messages are encapsulated in UDP packets and carry unicast destination addresses, sent using LDP port 646. A Target Hello message carries an LDP ID and other information (such as the hello-hold time and the transport address). If an LSR receives a Target Hello message, the LSR has a potential LDP peer.

Process of Establishing an LDP Session

Two LSRs exchange Hello messages to trigger the establishment of an LDP session.

Figure 2-2 shows the process of LDP session establishment.

Figure 2-2 Process for establishing an LDP session



1. Two LSRs send Hello messages to each other.
2. After receiving the Hello messages carrying the transport addresses, the two LSRs use the transport addresses to establish an LDP session. The LSR with the larger transport address serves as the active peer and initiates a TCP connection. As shown in **Figure 2-2**, LSRA serves as the active peer to initiate a TCP connection and LSRB serves as the passive peer to wait for the initiation of the TCP connection.

3. After the TCP connection is successfully established, LSRA sends an Initialization message to negotiate parameters used to establish an LDP session with LSRB. The main parameters include the LDP version, label advertisement mode, the Keepalive hold timer value, maximum PDU length, and label space.
4. If LSRB rejects some parameters, it sends a Notification message to terminate the establishment of the LDP session. If LSRB accepts all parameters, it sends an Initialization message carrying the LDP version, label advertisement mode, the Keepalive hold timer value, maximum PDU length, and label space, and sends a Keepalive message to LSRA.
5. If LSRA rejects certain parameters after receiving the Initialization message, it sends a Notification message to terminate LDP session establishment. If LSRA accepts all parameters, it sends a Keepalive message to LSRB.

After both LSRA and LSRB have accepted Keepalive messages from each other, the LDP session is successfully established.

Advertising and Managing Labels

Label Advertisement Modes

An LSR on an MPLS network assigns a label to a specified FEC and notifies its upstream LSRs of the label. This means that the label is specified by a downstream LSR, and is distributed from downstream to upstream.

As described in **Table 2-1**, two label advertisement modes are available.

Table 2-1 Label advertisement modes

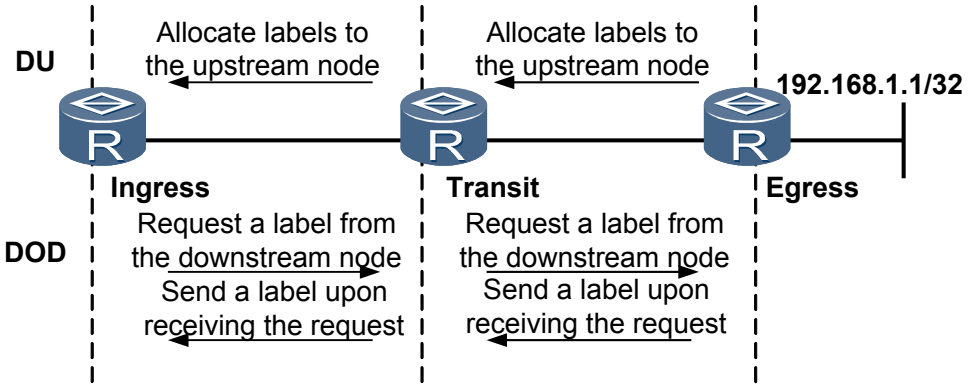
Label Advertisement Modes	Definition	Description
Downstream Unsolicited (DU) mode	An LSR distributes labels to a specified FEC without having to receive Label Request messages from its upstream LSR.	As shown in **Figure 2-3**, the downstream egress triggers the establishment of an LSP destined for the FEC 192.168.1.1/32 using a host route and sends a Label Mapping message to the upstream transit node to advertise the label of the host route to 192.168.1.1/32.
Downstream on Demand (DoD) mode	An LSR distributes labels to a specified FEC only after receiving Label Request messages from its upstream LSR.	As shown in **Figure 2-3**, the downstream egress triggers the establishment of an LSP destined for the FEC 192.168.1.1/32 in host mode. The upstream ingress sends a Label Request message to the downstream egress. After receiving the message, the downstream egress sends a Label Mapping message to the upstream LSR.

The label advertisement modes on upstream and downstream LSRs must be the same.

NOTE

When DU is used, LDP supports label distribution for all peers by default. Each node can send Label Mapping messages to all peers without distinguishing upstream and downstream nodes. If an LSR distributes labels only for upstream peers when it sends Label Mapping messages, the LSR checks the upstream/downstream relationship of the session in routing information. An upstream node cannot send Label Mapping messages to its downstream node along a route. If the route changes and the upstream/downstream relationship is switched, the new downstream node resends Label Mapping messages. In this process, the convergence is slow.

Figure 2-3 DU and DoD



Label Distribution Control Modes

The label distribution control mode refers to a method of label distribution on the LSR during LSP establishment.

As described in **Table 2-2**, two label distribution control modes are available.

Table 2-2 Label distribution control modes

Label Distribution Control Modes	Definition	Description
Independent mode	A local LSR can distribute a label bound to an FEC and then inform the upstream LSR, without waiting for the label distributed by the downstream LSR.	● As shown in **Figure 2-3**, if the label advertisement mode is DU and the label distribution control mode is Independent, a transit LSR can assign a label to the ingress node without waiting for the label assigned by the egress node. ● As shown in **Figure 2-3**, if the label advertisement mode is DoD and the label distribution control mode is Independent, the directly-connected ingress transit node that sends a Label Request message replies with a label without waiting for the label assigned by the egress node.

Label Distribution Control Modes	Definition	Description
Ordered mode	An LSR advertises the mapping between a label and an FEC to its upstream LSR only when this LSR is the outgoing node of the FEC or receives the Label Mapping message of the next hop for the FEC.	● As shown in **Figure 2-3**, the label distribution mode is DU and the label distribution control mode is ordered. Consequently, the LSR (the transit LSR in the diagram) must receive a Label Mapping message from the downstream LSR (the egress node in the diagram). Then, it can distribute a label to the ingress node in the diagram. ● As shown in **Figure 2-3**, if the label distribution mode is DoD and the label distribution control mode is Ordered, the directly-connected transit of the ingress node that sends the Label Request message must receive a Label Mapping message from the downstream (the egress node in the diagram). Then, it can distribute a label to the ingress node in the diagram.

Label Retention Modes

The label retention mode refers to the way an LSR processes the label mapping that it receives but does not immediately use.

The label mapping that an LSR receives may or may not originate at the next hop.

As described in **Table 2-3**, two label retention modes are available.

Table 2-3 Label retention modes

Label Retention Modes	Definition	Description
Liberal mode	When receiving a Label Mapping message from a neighbor LSR, an LSR retains the message regardless of whether the neighbor LSR is its next hop.	When the next hop of an LSR changes due to a change in network topology, note that: ● In Liberal mode, the LSR can use the previous label sent by a non-next hop to quickly reestablish an LSP. This requires more memory and label space than in conservative mode. ● In Conservative mode, the LSR only retains labels sent by the next hop. This saves memory and label space but slows down the reestablishment of an LSP. Conservative mode and DoD mode are used together to set up LSRs with limited label space.
Conservative mode	When receiving a Label Mapping message from a neighbor LSR, an LSR retains the message only when the neighbor LSR is its next hop.	

Currently, the combination of the following modes is supported:

- Combination of the DU label advertisement mode, ordered label control mode, and liberal label retention mode
- Combination of the DoD label advertisement mode, ordered label control mode, and conservative label retention mode

NOTE

On the device, LDP by default works in the DU label advertisement mode, ordered label control mode, and liberal label retention mode.

2.2.3 LDP Label Filtering Mechanism

By default, an LSR receives and sends Label Mapping messages for all FECs, resulting in the establishment of a large number of LDP LSPs. The establishment of a large number of LDP LSPs consumes a great deal of LSR resources. As a result, the LSR may be overburdened. An outbound or inbound LDP policy needs to be configured to reduce the number of Label Mapping messages to be sent or received, reducing the number of LSPs to be established and saving memory.

Outbound LDP Policy

LDP outbound policies are used to filter out Label Mapping messages sent to peers. If a FEC matches no outbound policy, neither a transit LSP nor an egress LSP can be established. If a pair

of or all peers have the same restriction on the FEC range when sending Label Mapping messages, the same outbound policy can be configured for the pair of or all peers.

An LDP outbound policy filters out Label Mapping messages only for the FEC, but not those for L2VPN. Meanwhile, the LDP outbound policy specifies the FEC range.

In addition, the outbound LDP policy supports split horizon. After split horizon is configured, an LSR distributes labels only to its upstream LDP peers.

Before sending Label Mapping messages only for the FEC to a peer, an LSR checks whether an outbound policy is configured.

- If no outbound policy is configured, the LSR sends the Label Mapping message.
- If an outbound policy is configured, the LSR checks whether the FEC in the Label Mapping message is within the range defined in the outbound policy. If the FEC is within the FEC range, the LSR sends a Label Mapping message for the FEC; if the FEC is not within the FEC range, the LSR does not send a Label Mapping message.

Inbound LDP Policy

LDP inbound policies are used to filter out Label Mapping messages received from peers. If a FEC matches no inbound policy, Label Mapping messages are not accepted. If a pair of or all peers have the same restriction on the FEC range when receiving Label Mapping messages, the same inbound policy can be configured for the pair of or all peers.

An LDP inbound policy filters out Label Mapping messages only for the FEC, but not those for L2VPN. Meanwhile, the LDP inbound policy specifies the FEC range for non-BGP routes.

An LSR checks whether an inbound policy mapped to a FEC is configured before receiving a Label Mapping message for the FEC.

- If no inbound policy is configured, the LSR receives the Label Mapping message.
- If an inbound policy is configured, the LSR checks whether the FEC in the Label Mapping message is within the range defined in the inbound policy. If the FEC is within the FEC range, the LSR receives the Label Mapping message for the FEC; if the FEC is not in the FEC range, the LSR does not receive the Label Mapping message.

If the FEC fails to pass an outbound policy on an LSR, the LSR receives no Label Mapping message for the FEC.

One of the following results may occur:

- If a DU LDP session is established between an LSR and its peer, a liberal LSP is established. This liberal LSP cannot function as a backup LSP after LDP FRR is enabled.
- If a DoD LDP session is established between an LSR and its peer, the LSR sends a Release message to tear down label-based bindings.

NOTE

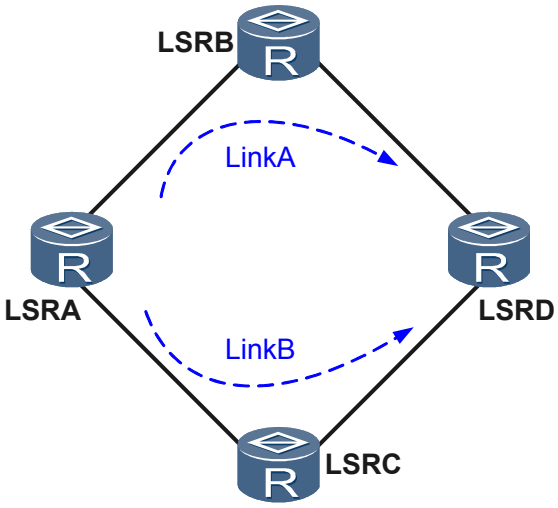
An LSP that is distributed with a label but is not successfully established called a liberal LSP.

2.2.4 Synchronization Between LDP and Static Routes

Synchronization between LDP and static routes applies to MPLS networks where primary and backup LSPs exist. LSPs are established between LSRs based on static routes. When the LDP session on the primary LSP fails (not due to a link failure) or the primary LSP is restored, MPLS traffic is interrupted for a short time.

As shown in **Figure 2-4**, LSRA and LSRD are connected using static routes. LDP establishes primary and backup LSPs between LSRA and LSRD based on static routes, and LinkA is the primary path.

Figure 2-4 LSP switchover based on synchronization between LDP and static routes



Synchronization between LDP and static routes implements LSP switchover in the following scenarios:

- The LDP session on the primary LSP fails (not due to a link failure).

  When an LDP session is established, MPLS traffic is forwarded through LinkA. If LDP is disabled or faulty on LSRB, the LDP session between LSRA and LSRB fails. However, the link between LSRA and LSRB is running properly and static routes are active. MPLS traffic is interrupted between LSRA and LSRD during LSP switchover to LinkB.

  After synchronization between LDP and static routes is enabled on LSRA, static routes automatically switch to LinkB when the LDP session is Down. This ensures uninterrupted MPLS traffic during an LSP switchover.
- The primary LSP recovers from a fault.

  If the link between LSRA and LSRB fails, the LSP switches to LinkB. When the link between LSRA and LSRB recovers, the LSP switches back to LinkA. At this time, the backup LSP cannot be used, but the new LSP has not been established. MPLS traffic between LSRA and LSRD is interrupted during this period.

  After synchronization between LDP and static routes is enabled on LSRA, static routes become active only when the LDP session is Up, which ensures uninterrupted traffic.

2.2.5 Synchronization Between LDP and IGP

Background

The LDP convergence speed depends on the convergence speed of IGP routes, which indicates IGP convergence is faster.

- On an MPLS network with the primary and backup links, the following problems occur:

  1. When the primary link fails, an IGP route of the backup link becomes reachable and a backup LSP over the backup link takes over traffic. After the primary link recovers,

the IGP route of the primary link becomes reachable before an LDP session is established over the primary link. As a result, traffic is dropped when being transmitted using the reachable IGP route along the unreachable LSP.

2. When the IGP route of the primary link is reachable and an LDP session between nodes on the primary link fails, traffic is directed using the IGP route of the primary link, whereas the LSP over the primary link is torn down. Because a preferred IGP route of the backup link is unavailable, an LSP over the backup link cannot be established, causing traffic loss.

- When the active/standby switchover occurs on a node, the LDP session establishment is later than the IGP GR completion. IGP advertises the maximum cost of the link, causing route flapping.

Related Concepts

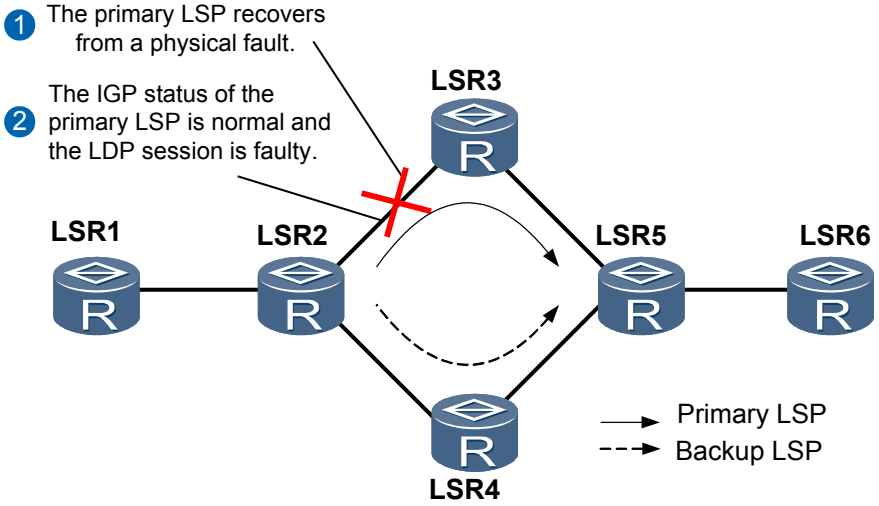
Synchronization between LDP and IGP is implemented by suppressing IGP from advertising normal routes to ensure convergence performed by synchronization between LDP and IGP.

Synchronization between LDP and IGP involves three timers:

- Hold-down timer: used to control the period for establishing the IGP neighbor relationship.
- Hold-max-cost timer: used to control the period for advertising the maximum cost of the link.
- Delay timer: used to control the period for waiting for the LSP establishment.

Implementation

Figure 2-5 Synchronization between LDP and IGP for revertive switchover



- During active/standby link switchover, synchronization between LDP and IGP takes effect.

  As shown in **Figure 2-5**, the processes of synchronization between LDP and IGP differ in the following scenarios:

  1. The primary link recovers from a physical fault.

     a. The faulty link recovers.

b. An LDP session is set up between LSR2 and LSR3. IGP suppresses the establishment of the neighbor relationship and starts the Hold-down timer as required.

c. Traffic keeps traveling through the LSP over the backup link.

d. After the LDP session is set up, Label Mapping messages are exchanged and then synchronization between IGP and LDP starts.

e. The IGP establishes a neighbor relationship and switches traffic back to the primary link, and the LSP is reestablished and its route converges on the primary link (in milliseconds).

2. IGP on the primary link is normal and the LDP session is faulty.

a. An LDP session between nodes along the primary link becomes defective.

b. LDP notifies the IGP primary link of the session fault. IGP starts the Hold-max-cost timer and advertises the maximum cost on the primary link.

c. The IGP route of the backup link becomes reachable.

d. An LSP is established over the backup link and the LDP module on LSR2 delivers forwarding entries.

The Hold-max-cost timer can be configured to always advertise the maximum cost of the primary link. This setting allows traffic to keep traveling through the backup link before the LDP session over the primary link is reestablished.

- During active/standby system switchover, the procedure for synchronization between LDP and IGP is as follows:

1. An IGP on the Restarter advertises a normal cost value and starts a Delay timer, waiting for an LDP session to be set up. Then IGP ends the GR process.
2. If the Delay timer expires before the LDP session is set up, IGP starts a Hold-max-cost timer, and advertises the maximum cost value of the link.
3. After the LDP session is established or the Hold-max-cost timer expires, IGP advertises the actual link cost and updates the IGP route.
4. The helper retains the IGP route and LSP. After the LDP session on the helper goes Down, the LDP module does not notify the IGP module of the session status change. This indicates that IGP keeps advertising the actual link cost, preventing traffic or LSP switchover.

2.2.6 BFD for LSP

A Bidirectional Forwarding Detection (BFD) session is established on an LSP. BFD is used to quickly detect faults on the LSP, providing end-to-end protection.

BFD is used to detect faults on the data plane of the MPLS LSP, and the format of BFD packets is fixed. When a BFD session is associated with a unidirectional LSP, the reverse link can be an IP link, an LSP, or a TE tunnel.

Implementation

BFD detects LSPs in asynchronous mode. The ingress and the egress nodes send BFD control packets to each other periodically.

- If any of the ingress and the egress nodes does not receive BFD control packets sent by the peer within a detection period, LSP status is considered to be Down and a message that the LSP is Down is sent to the LSP Management (LSPM) module.

- If the LSP status changes between Up and Down frequently, BFD sends two messages of LSP changes successively. Therefore, the detection can be performed flexibly.
- If the reverse link of the BFD control packets sent by the egress node to the ingress node fails, the BFD session is Down.

NOTE

BFD is a bidirectional detection mechanism, but BFD for LSP is unidirectional. BFD for LSP sends BFD control packets through LSPs on the ingress node and through IP links on the egress node. As a result, when the ingress node does not receive BFD control packets sent through the reverse path from the egress node, the system considers that the LSP fails no matter the fault occurs on LSP or on the reverse link.

BFD Session Setup

To check MPLS LSP connectivity, negotiation on a BFD session can be performed in the following modes:

- Static: The negotiation on a BFD session is performed using the local discriminator (LD) and remote discriminator (RD) that are manually configured.
- Dynamic: The negotiation on a BFD session is performed using the BFD discriminator TLV in an LSP ping packet.

BFD detects the following types of LSPs:

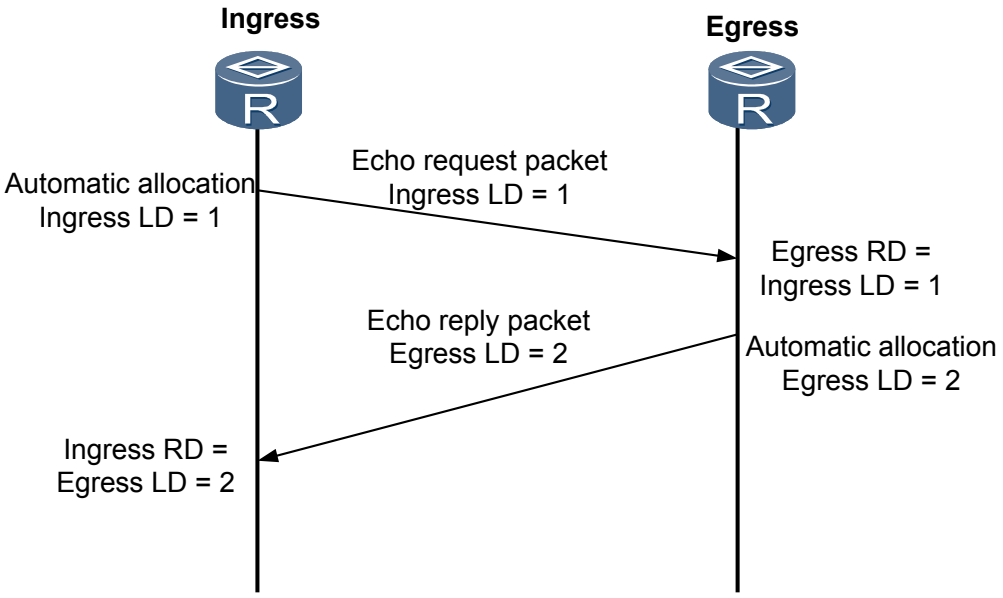
- Static BFD for static LSP
- Static BFD for LDP LSP
- Dynamic BFD for LDP LSP

Figure 2-6 shows the establishment of dynamic BFD sessions that detect LDP LSPs.

1. The ingress node sends an MPLS echo request packet that carries the type-length-value (TLV) with the type as 15 along an LSP. The packet contains an LD that the ingress node allocates to the BFD session.
2. The egress node receives the MPLS echo request packet sent from the ingress node and takes the contained LD as its own RD.
3. The egress node sends an MPLS echo reply packet to the ingress node. The packet contains an LD that the egress node allocates to the BFD session.
4. The ingress node receives the MPLS echo reply packet sent by the egress node and takes the contained LD as its own RD.

   The dynamic BFD session that detects the LDP LSP is created successfully.

Figure 2-6 Establishing a session of dynamic BFD for LDP LSP



2.2.7 LDP FRR

LDP Fast Reroute (FRR) provides the fast reroute function for MPLS networks by backing up local interfaces.

LDP FRR, in liberal label retention mode of LDP, obtains a liberal label, applies a forwarding entry for the label, and then forwards the forwarding entry to the forwarding plane as the backup forwarding entry for the primary LSP. When the interface is faulty (detected by the interface itself or according to BFD detection) or the primary LSP fails (according to BFD detection), LDP FRR fast switches traffic to the backup LSP to protect the primary LSP.

- Manually configured LDP FRR needs to be specified with the outbound interface and next hop of the backup LSP by running a command. When the source of the liberal label matches the outbound interface and next hop, a backup LSP can be established and its forwarding entries can be delivered.
- LDP auto FRR depends on the implementation of IP FRR. When the source of the preserved liberal label matches the outbound interface and next hop of the backup route, the requirement for the policy for establishing the backup LSP is met, and no backup LSP manually configured according to the backup route exists, a backup LSP can be established and its forwarding entries can be delivered. The default policy of LDP auto FRR is that LDP can use the 32-bit backup routes to establish backup LSPs. When both the manually configured LDP FRR and LDP auto FRR meet the establishment conditions, the manually configured LDP FRR is established preferentially.

Applicable Environment

Figure 2-7 A typical applicable environment of LDP FRR (triangle topology)

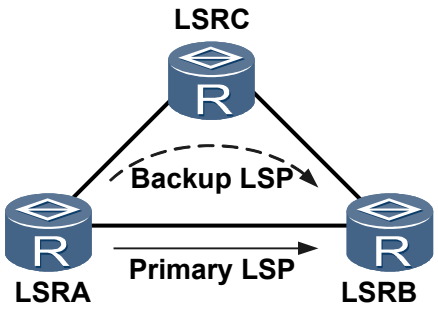
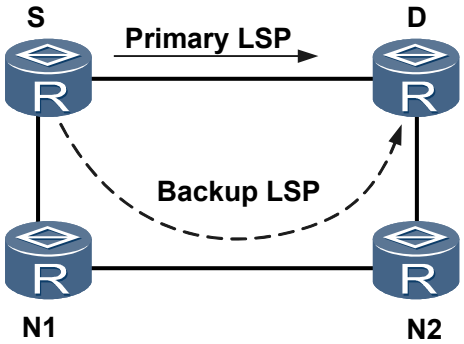


Figure 2-7 shows a typical applicable environment of LDP FRR. The optimal route from LSRA to LSRB is LSRA -> LSRB and the less optimal route is LSRA -> LSRC -> LSRB. A primary LSP along the path LSRA -> LSRB is established on LSRA, and a backup LSP along the path LSRA -> LSRC -> LSRB is established to protect the primary LSP. After receiving a label from LSRC, LSRA compares the label with the route from LSRA to LSRB and finds that LSRC is not the next hop of the route. LSRA preserves the label as a liberal label and applies for a forwarding entry as the backup forwarding entry of the primary LSP. LSRA forwards the forwarding entries of both the primary and backup LSPs to the forwarding plane. In this manner, the primary LSP is associated with the backup LSP.

When the interface detects faults by itself, BFD detects faults on the interface, or BFD detects that the primary LSP fails, LDP FRR is triggered. After LSP FRR is complete, traffic is switched to the backup LSP according to the backup forwarding entry. In this manner, LSP FRR takes effect. Then, the route is converged from LSRA-LSRB to LSRA-LSRC-LSRB. An LSP is established on the new LSP (the original backup LSP), and the original primary LSP is deleted, and then the traffic is forwarded along the new LSP LSRA -> LSRC -> LSRB.

Figure 2-8 A typical applicable environment of LDP FRR (rectangle topology)



As shown in Figure 2-7, all nodes in the triangle topology supports LDP FRR, but only parts of nodes in the rectangle topology supports LDP FRR. As shown in Figure 2-8, if the optimal route from N1 to D is N1 -> N2 -> D (load balancing is unavailable), S receives a liberal label from N1 and is configured with LDP FRR. When the link between S and D is faulty, traffic is switched to the route of S -> N1 -> N2 -> D without forming a loop.

However, if the optimal route from N1 to D is load balanced between N1 -> N2 -> D and N1 -> S -> D, S as the downstream neighbor of N1 does not necessarily receive the liberal label from N1. In addition, even though S receives the liberal label (LDP distributes labels for each peer) and is configured with LDP FRR, traffic may still go to S after traffic switches to N1, which leads to a loop. This occurs till the route from N1 to D is converged to N1 -> N2 -> D.

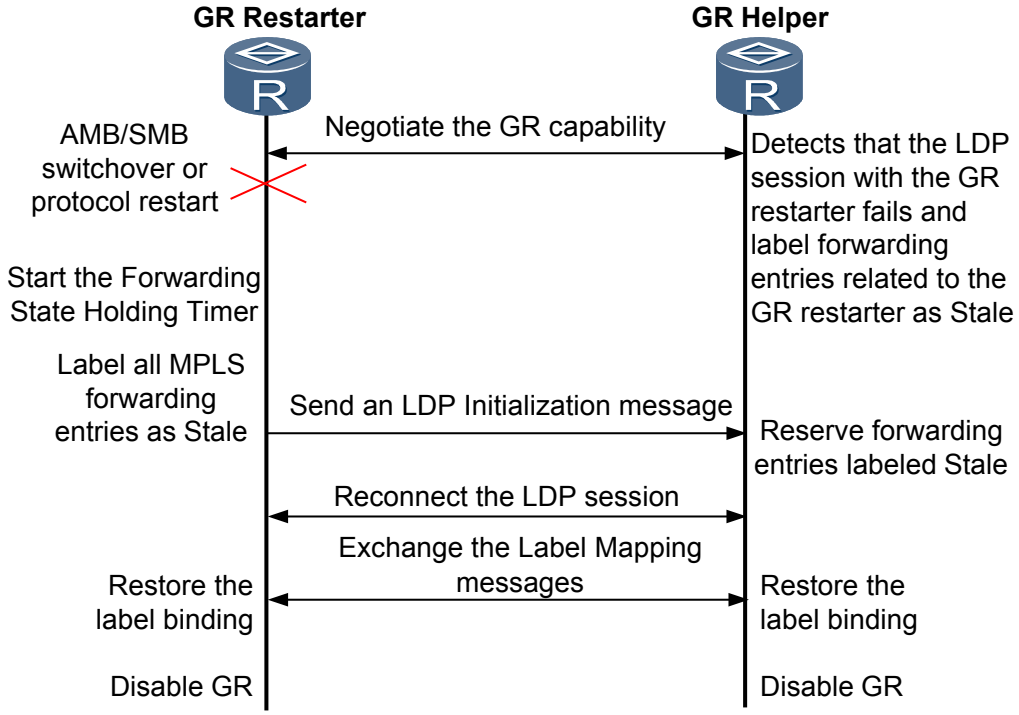
2.2.8 LDP GR

LDP graceful restart (GR) ensures uninterrupted traffic forwarding on the restarter with the help of a neighbor (Helper) when an active main board/standby main board (AMB/SMB) switchover or a protocol restart occurs on the restarter.

When the AMB/SMB switchover occurs on a device that is not capable of GR, the neighbor deletes the LSP because the LDP session becomes Down. As a result, traffic cannot be forwarded and services are interrupted for a short period. To prevent service interruption, LDP GR can be configured to keep labels consistent before and after the AMB/SMB switchover or the protocol restart. LDP GR ensures uninterrupted MPLS forwarding. **Figure 2-9** shows the detailed process.

1. Before the AMB/SMB switchover, LDP neighbors negotiate the GR capability during the LDP session establishment.
2. After the AMB/SMB switchover, the GR detects helper starts the LDP session failure and starts the GR Reconnect timer. The GR helper retains the forwarding entries related to the GR restarter and marks the entries with the stale tag.
3. After performing the AMB/SMB switchover, the GR restarter starts the Forwarding State Holding timer. Before the Forwarding State Holding timer times out, the GR restarter retains all MPLS forwarding entries before the restart and marks the entries with the stale tag. The GR restarter then sends an LDP Initialization message to the GR helper. When the Forwarding State Holding timer times out, the GR restarter performs step 6.
4. Before the GR Reconnect timer times out, the LDP session is reestablished. The GR helper deletes the Forwarding State Holding timer, starts the GR Recovery timer, and retains the forwarding entries with the stale tag.
5. Before the GR Recovery timer times out, the neighbors exchange Label Mapping messages with each other and restore the label binding before the AMB/SMB switchover. When the GR Recovery timer times out, the GR helper deletes all forwarding entries with the stale tag.
6. The GR process ends. The GR restarter deletes all forwarding entries with the stale tag.

Figure 2-9 LDP GR implementation



2.2.9 LDP NSR

The non-stop routing (NSR) technology is an innovation based on non-stop forwarding (NSF) technology. If a software or hardware fault occurs on the control plane, NSR ensures uninterrupted forwarding and connection of the control plane. In addition, the control plane of a neighbor will not detect any fault.

LDP NSR is implemented using the synchronization of the master and slave control boards. During the startup, the slave control board backs up data of the master control board in batches to ensure data consistency on both boards. LDP NSR simultaneously notifies the master and slave control boards of receipt of packets and backs up these packets in real time. In this manner, the slave control board synchronizes data with the master control board. NSR ensures that after switchover, the slave control board can quickly take over services from the original master control board, while the neighbor will not detect the fault on the local router.

LDP NSR synchronizes the following key data between the master and slave control boards:

- LSP forwarding entries
- Key resources such as labels and cross connections
- LDP protocol control blocks

2.2.10 LDP Security Mechanisms

MD5 Authentication

Message-digest algorithm 5 (MD5) is a standard digest algorithm defined in RFC 1321. A typical application of MD5 is to calculate a message digest to prevent message spoofing. The MD5

message digest is a unique result calculated by an irreversible character string conversion. If a message is modified during transmission, a different digest is generated. After the message arrives at the receiver, the receiver can determine whether the packet is modified by comparing the received digest with the pre-calculated digest.

LDP MD5 authentication prevents LDP packets from being modified by generating a unique digest for an information segment. This authentication is stricter than the common checksum verification of TCP connections.

Before an LDP message is sent over a TCP connection, LDP MD5 authentication is performed by padding the TCP header with a unique digest. This digest is a result calculated by MD5 based on the TCP header, LDP session message, and password set by the user.

When receiving this TCP packet, the receiver obtains the TCP header, digest, and LDP session message, and then uses MD5 to calculate a digest based on the received TCP header, received LDP session message, and locally stored password. The receiver compares the calculated digest with the received one to check whether the packet is modified.

A password can be set in either cipher text or plain text. The plain-text password is directly recorded in the configuration file. The cipher-text password is recorded in the configuration file after being encrypted using a special algorithm.

During the calculation of a digest, the manually entered character string is used regardless of whether the password is in plain text or cipher text. This indicates that a password calculated using an encryption algorithm does not participate in MD5 calculation, ensuring that LDP MD5 authentication implemented on Huawei devices is transparent to non-Huawei devices.

Keychain Authentication

Keychain, an enhanced encryption algorithm to MD5, calculates a message digest for the same LDP message to prevent the message from being modified.

During keychain authentication, a group of passwords are defined to form a password string. Each password is specified with encryption and decryption algorithms such as MD5 algorithm and SHA-1, and is configured with the validity period. When sending or receiving a packet, the system selects a valid password based on the user's configuration. Within the valid period of the password, the system uses the encryption algorithm matching the password to encrypt the packet before sending it out, or uses the decryption algorithm matching the password to decrypt the packet before accepting it. In addition, the system automatically uses a new password after the previous password expires, preventing the password from being decrypted.

The keychain authentication password, the encryption and decryption algorithms, and the password validity period that construct a keychain configuration node are configured using different commands. A keychain configuration node requires at least one password and encryption and decryption algorithms.

LDP GTSM

Generalized TTL Security Mechanism (GTSM) is a mechanism that protects the service by checking whether the TTL value in the IP header is within the pre-defined range. The prerequisites for using GTSM are as follows:

- The TTL of normal packets between routers is determined.
- The TTL value of packets can hardly be modified.

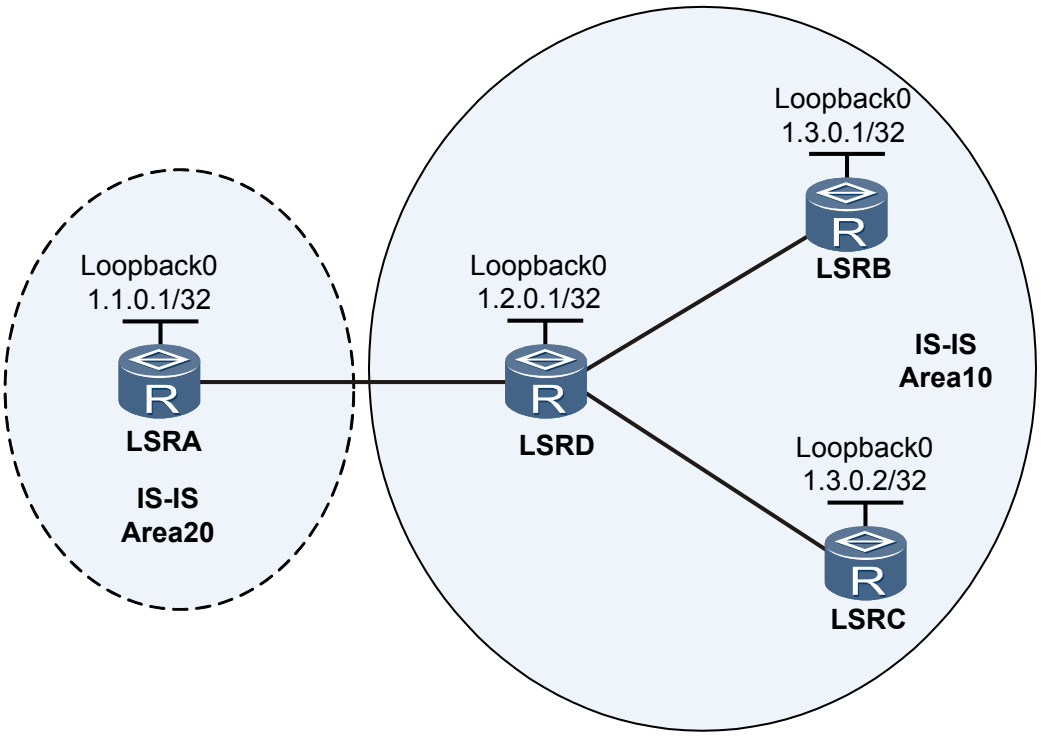
LDP GTSM refers to GTSM implementation over LDP.

To protect the router against attacks, GTSM checks the TTL in a packet to verify it. GTSM for LDP is applied to LDP packets between neighbor or adjacent (based on a fixed number of hops) routers. The TTL range is preset on each router for packets from other routers and GTSM is enabled. If the TTL of an LDP packet received by a router configured with LDP is out of the TTL range, the packet is considered invalid and is discarded. This protects the upper-layer protocols.

2.2.11 LDP Extension for Inter-Area LSP

This feature enables LDP to establish inter-area LDP LSPs to provide tunnels that traverse the public network.

Figure 2-10 Networking topology for LDP extension for inter-area LSP



As shown in **Figure 2-10**, there are two IGP areas: Area 10 and Area 20.

In the routing table of LSRD at the edge of Area 10, two host routes are reachable to LSRB and LSRC. You can use IS-IS to aggregate the two routes to one route to 1.3.0.0/24 and send this route to Area 20 to prevent a large number of routes from occupying too many resources on the LSRD. Consequently, there is only one aggregated route (1.3.0.0/24) but not 32-bit host routes in LSRA's routing table. By default, when establishing LSPs, LDP searches the routing table for the route that exactly matches the FEC in the received Label Mapping message. **Figure 2-10** shows routing entry information of LSRA and routing information carried in the FEC, as shown in **Table 2-4**.

Table 2-4 Routing entry information of LSRA and routing information carried in the FEC

Routing Entry Information of LSRA	FEC
1.3.0.0/24	1.3.0.1/32
	1.3.0.2/32

LDP establishes liberal LSPs, not inter-area LDP LSPs, for aggregated routes. In this situation, LDP cannot provide required backbone network tunnels for VPN services.

Therefore, in the situation shown in **Figure 2-10**, configure LDP to search for routes according to the longest match rule for establishing LSPs. There is already an aggregated route to 1.3.0.0/24 in the routing table of LSRA. When LSRA receives a Label Mapping message (such as the carried FEC is 1.3.0.1/32) from Area 10, LSRA searches for a route according to the longest match rule defined in RFC 5283. Then, LSRA finds information about the aggregated route to 1.3.0.0/24, and uses the outbound interface and next hop of this route as those of the route to 1.3.0.1/32. In this manner, LDP can establish inter-area LDP LSPs.

2.2.12 LDP over GRE/mGRE

GRE provides a mechanism to encapsulate packets of a protocol into packets of another protocol. This allows packets to be transmitted over heterogeneous networks. A channel for transmitting heterogeneous packets is called a tunnel.

A GRE tunnel can be established using the following tunnel interfaces:

- GRE tunnel interface

  A GRE tunnel interface is a point-to-point virtual interface used to encapsulate packets, and has the source address, destination address, tunnel interface IP address, and encapsulation type.
- mGRE tunnel interface

  An mGRE tunnel interface is a point-to-multipoint virtual interface used in DSVPN applications, and has the source address, destination address, and tunnel interface IP address.

  The destination IP address of a GRE tunnel interface is manually configured, whereas the destination IP address of an mGRE tunnel is resolved by the Next Hop Resolution Protocol (NHRP). An mGRE tunnel interface has multiple remote ends because there are multiple GRE tunnels on the interface.

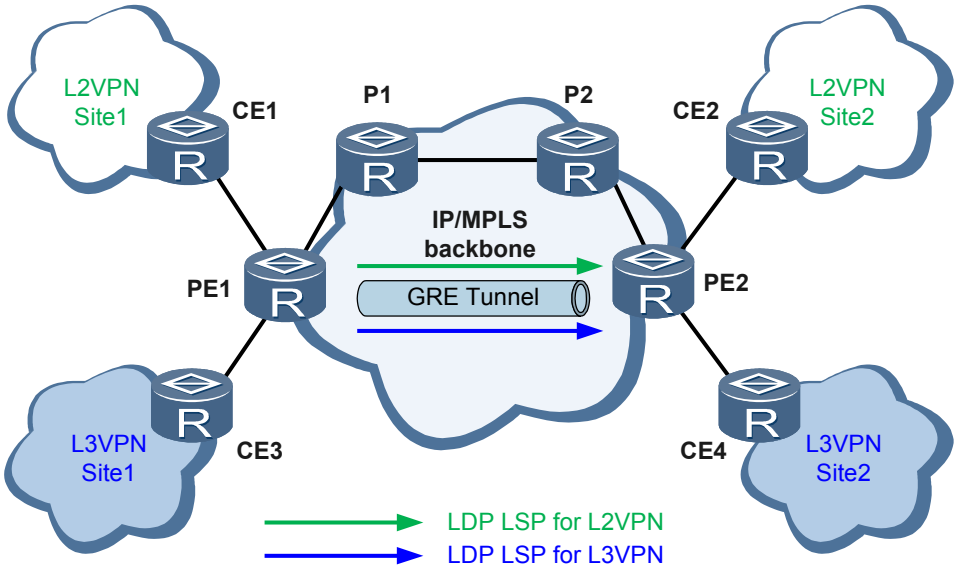
If backbone network devices are not enabled with MPLS or do not support MPLS, LDP LSPs cannot be established. As a result, L2VPN or L3VPN services cannot be deployed. LDP over GRE/mGRE addresses the preceding problem.

LDP over GRE

LDP over GRE technology establishes an LDP LSP on a GRE tunnel interface configured with MPLS LDP to transmit MPLS LDP packets.

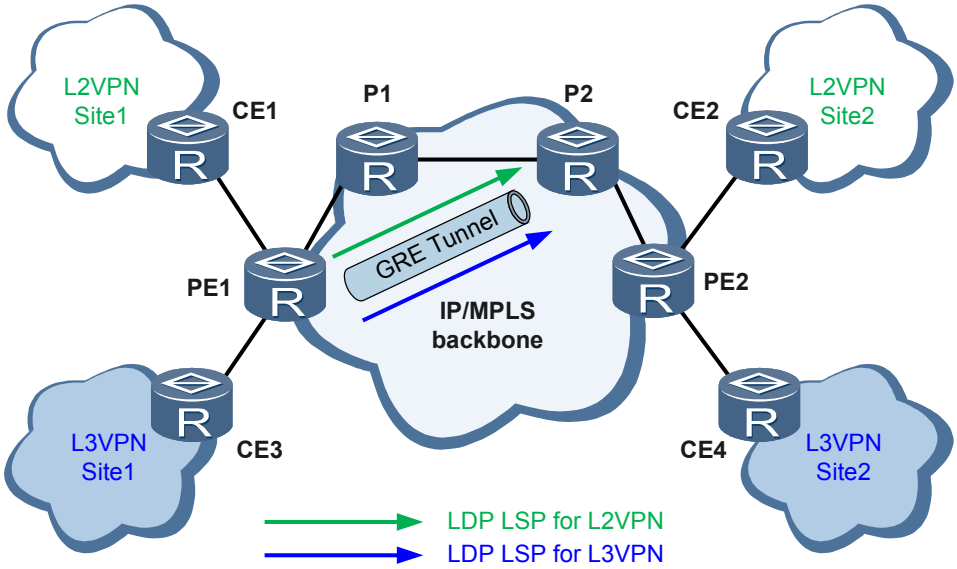
As shown in **Figure 2-11**, L2VPN or L3VPN services are deployed between PE1 and PE2 of an enterprise. Because backbone network devices may be not enabled with MPLS or do not support MPLS, an LDP LSP across a GRE tunnel needs to be set up between PE1 and PE2.

Figure 2-11 Deploying LDP over GRE on L2VPN/L3VPN networking (all P devices do not support MPLS)



As shown in **Figure 2-12**, backbone network device P2 supports MPLS, whereas P1 does not support MPLS. A GRE tunnel can be established between PE1 and P2 so that an LDP LSP is set up across the GRE tunnel.

Figure 2-12 Deploying LDP over GRE on L2VPN/L3VPN networking (some P devices do not support MPLS)



LDP over mGRE

As shown in **Figure 2-13**, the IP/MPLS backbone network is established in the enterprise headquarters. Enterprise branches in other areas need to connect to the IP/MPLS backbone

network through public network devices. If an L3VPN network needs to be established between the enterprise headquarters and branches, LDP over GRE can be used. There are the following problems:

- Public addresses of branch devices are not fixed.

  The branch Spoke uses a dynamically allocated public address. When LDP over GRE is used, GRE cannot be used to establish a GRE tunnel because public addresses of devices in multiple branches are not fixed. As a result, L3VPN services cannot be deployed.

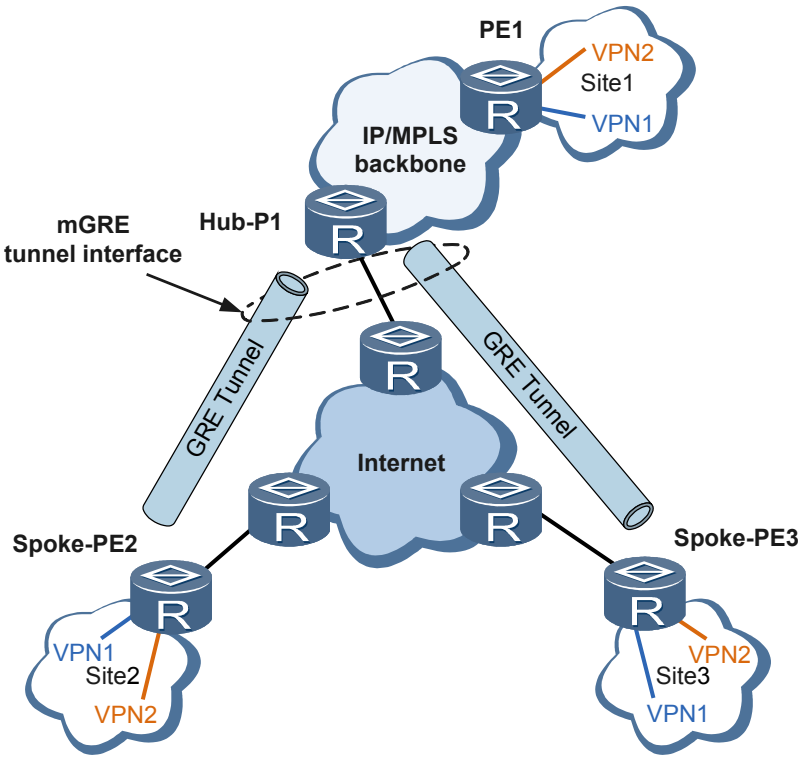
- There are many branches.

  When there are many branches, a large number of Spokes exist. If GRE is used to establish GRE tunnels (assume that devices use fixed public addresses), many GRE interfaces need to be created on Hubs. The configuration is complex and maintenance is difficult.

Dynamic Smart Virtual Private Network (DSVPN) technology solves the preceding problem. NHRP solves problems caused by non-fixed public addresses of branch devices and dynamically establishes a tunnel between the headquarters and a branch. mGRE allows multiple GRE tunnels to be set up on a tunnel interface, simplifying the Hub configuration.Similar to LDP over GRE, LDP over mGRE technology establishes an LDP LSP on an mGRE tunnel interface configured with MPLS LDP to transmit MPLS LDP packets.

As shown in **Figure 2-13**, the enterprise establishes a backbone network. The headquarters Hub-P uses a static address to connect to the public network and branch Spoke-PE uses dynamic addresses to connect to the public network. The enterprise requires that L3VPN services be deployed between all PEs. Because branch devices connect to the enterprise IP/MPLS backbone network through the public network and Spoke-PEs' public addresses are not fixed, LDP over mGRE is used to establish LDP LSPs between PEs across GRE tunnels.

Figure 2-13 Deploying LDP over mGRE in Hub-Spoke networking



When LDP over mGRE is used to construct an L3VPN network, traffic between branches is forwarded through Hubs.

There is a delay in transmitting traffic of voice services in this case because voice services require point-to-point transmission. DSVPN can be used to dynamically establish a tunnel so that branches can directly communicate. In addition to the preceding advantages, DSVPN has the following problems:

- During establishment of a tunnel between Spokes, traffic between branches is forwarded through the Hub. After the tunnel is set up, traffic is switched to the tunnel. In this process, packet mis-sequencing may occur on the receiver.
- A tunnel between branches needs to be dynamically maintained. If the Spoke performance is low, maintaining a large amount of tunnel information will cause the Spoke to deteriorate and affect services.

In addition, MPLS requires that labels in MPLS packets be swapped on LSPs between ingress and egress nodes. After LDP over mGRE is used, all Spokes are equivalent to PEs on an MPLS network. If Spokes directly communicate with each other (this scenario is not used), each two Spokes need to establish an LDP and distribute labels. Consequently, label resources of Spokes are insufficient. You can use the Hub to forward traffic between branches (the Hub, similar to the P on an MPLS network, maintains the LDP neighbor relationship and swaps packet labels). A Spoke only needs to establish one LSP with the Hub so that the Spoke can communicate with other Spokes. Because MPLS label swapping is used, only traffic on the Hub is increased. Hub performance is less affected.

Although LDP over mGRE does not use DSVPN to establish tunnels between Spokes, it provides the Spoke-Hub-Spoke path to better transmit traffic between Spokes. LDP over mGRE is often used when the MPLS network needs to be extended.

2.3 References

The following table lists the references.

Document No.	Description
RFC3036	LDP Specification
RFC3215	LDP State Machine
RFC5443	LDP IGP Synchronization
RFC3478	Graceful Restart Mechanism for Label Distribution Protocol
RFC1321	The MD5 Message-Digest Algorithm
RFC3037	LDP Applicability
RFC3899	Maximum Transmission Unit Signalling Extensions for the Label Distribution Protocol
RFC3270	Multi-Protocol Label Switching (MPLS) Support of Differentiated Services

3 MPLS TE

About This Chapter

3.1 Introduction to MPLS TE

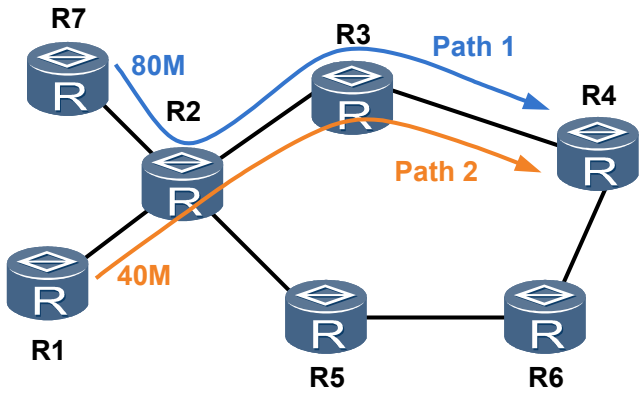
Definition

Multiprotocol Label Switching (MPLS) traffic engineering (TE) establishes label switched paths (LSPs) satisfying specific constraints and transparently transmits traffic over the LSPs based on labels.

Purpose

A node on a conventional IP network selects the shortest path as an optimal route, regardless of other factors, such as bandwidth. The shortest path may be congested with traffic, but other available paths are idle.

Figure 3-1 Conventional routing



Each link on the network shown in **Figure 3-1** has a bandwidth of 100 Mbit/s and the same metric value. R1 sends R4 traffic at 80 Mbit/s, and R7 sends R4 traffic at 40 Mbit/s. Interior Gateway Protocol (IGP) calculates the shortest path for traffic, such as Path1 and Path2 in **Figure 3-1**. Traffic on this network is forwarded along the path R2→R3→R4. As a result, the path R2→R3→R4 may be congested because of overload, while the path R2→R5→R6→R4 is idle.

Traffic engineering techniques can load balance traffic to the idle paths to prevent traffic congestion caused by uneven resource allocation.

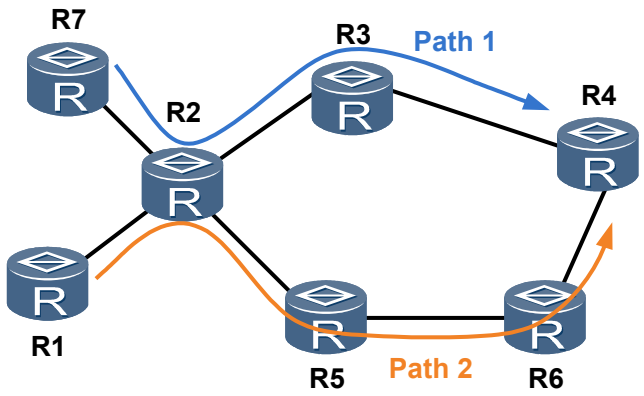
Conventional TE solutions are as follows:

- AP TE: AP TE controls network traffic by adjusting the metric of a path. This method eliminates congestion only on some links. Adjusting a metric is difficult on a complex network because a link change affects multiple routes.
- ATM TE: Because existing Interior Gateway Protocols (IGPs) are topology-driven and consider only network connectivity, they cannot present some dynamic factors such as bandwidth and traffic characteristics. The IP over asynchronous transfer mode (ATM) overlay model can solve this problem. ATM TE directs some traffic to virtual connections (VCs) using the overlay model so that traffic can be properly scheduled and allocated and QoS guarantee is ensured. However, ATM TE has high costs and low extensibility.

A scalable and simple solution is required to implement TE on a large-scale network. MPLS, an overlay model, allows a virtual topology to be established over a physical topology and maps traffic to the virtual topology. MPLS can be integrated with TE. MPLS TE is introduced.

MPLS TE can be used on the network shown in **Figure 3-1** to address congestion. MPLS TE sets up two LSPs: Path1 with bandwidth of 80 Mbit/s and Path2 with bandwidth of 40 Mbit/s. MPLS TE directs traffic to the two LSPs, preventing congestion.

Figure 3-2 MPLS TE



Benefits

MPLS TE effectively schedules, allocates, and uses existing network resources to provide sufficient bandwidth and support for quality of service (QoS). MPLS TE helps enterprises minimize expenditures without requiring hardware upgrades. TE is implemented based on MPLS techniques and is easy to deploy and maintain on live networks. MPLS TE supports a range of reliability techniques, which helps backbone networks achieve carrier- and device-class reliability.

3.2 Principles

3.2.1 Basic Concepts

This section describes the following basic concepts:

- **LSP Tunnel**
- **MPLS TE Tunnel**
- **Link Attributes**
- **Tunnel Attributes**

LSP Tunnel

After a label is added to a packet on the ingress node of an LSP, the packet is forwarded based on the label. Traffic forwarding is transparent to intermediate nodes, so an LSP can be considered as an LSP tunnel.

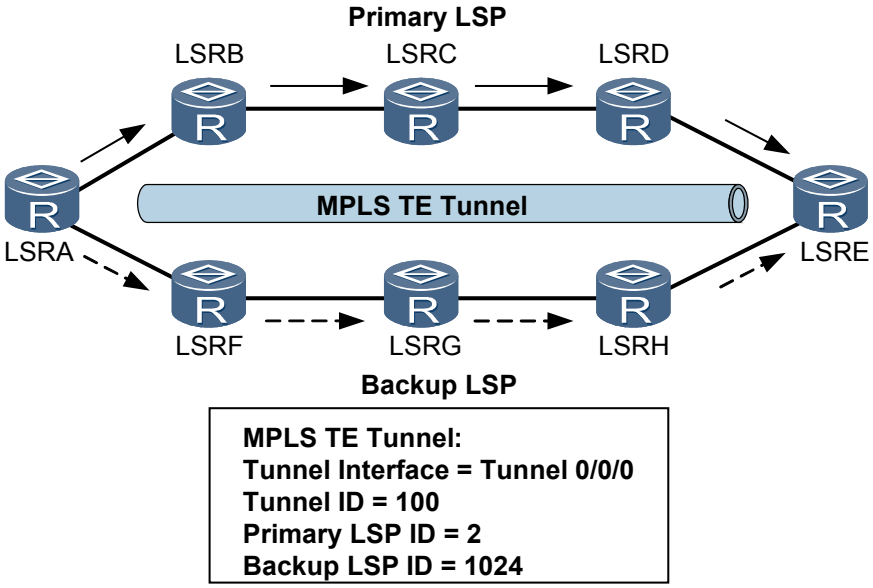
MPLS TE Tunnel

Multiple LSPs are bound together to form an MPLS TE tunnel. The MPLS TE tunnel involves the following entities:

- Tunnel interface: a P2P virtual interface that encapsulates packets. Similar to a loopback interface, a tunnel interface is a logical interface.
- Tunnel ID: a decimal number that identifies an MPLS TE tunnel and facilitates tunnel planning and management.
- LSP ID: a decimal number that identifies an LSP, which facilitates LSP planning and management.

A primary LSP with an LSP ID 2 is established along the path LSRA -> LSRB -> LSRC -> LSRD -> LSRE on the network shown in **Figure 3-3**. A backup LSP with an LSP ID 1024 is established along the path LSRA -> LSRF -> LSRG -> LSRH -> LSRE. The two LSPs are in a tunnel named Tunnel 0/0/0 with a tunnel ID 100.

Figure 3-3 MPLS TE Tunnel and LSPs



Link Attributes

MPLS TE link attributes describe bandwidth resources, route costs, and link reliability. The link attributes are as follows:

- Total link bandwidth: is physical link bandwidth.
- Maximum reservable bandwidth: is the maximum bandwidth that a link can reserve for an MPLS TE tunnel to be established. The maximum reservable bandwidth must be lower than or equal to the total link bandwidth.
- TE metric: is the TE cost of a link. To better control path calculation for an MPLS TE tunnel, MPLS TE provides the TE metric so that an IGP route can be selected independently. By default, a link uses the IGP metric as the TE metric.

- SRLG: is a set of links which are likely to fail concurrently when sharing a physical resource (for example, an optical fiber). Links in an SRLG share the same risk of faults. If one link fails, other links in the SRLG also fail.

  An SRLG enhances CR-LSP reliability on an MPLS TE network enabled with CR-LSP hot standby or TE FRR. For more information about the SRLG, see **SRLG**.
- Link administrative group: is also called link color. A link administrative group is a 32-bit vector, with each bit set to a specified value that is associated with a desired meaning. For example, a link administrative group attribute can be configured to describe link bandwidth, a performance parameter or a management policy. The policy can be a traffic type (multicast for example) or a flag indicating that an MPLS TE tunnel passes over the link. The link administrative group attribute is used together with **affinities** to control the paths for tunnels.

Tunnel Attributes

LSPs in an MPLS TE tunnel are constraint-based routed LSPs (CR-LSPs). These constraints are tunnel attributes.

Unlike Label Distribution Protocol (LDP) LSPs that are established using routing information, CR-LSPs are established based on bandwidth and path constraints in addition to routing information:

- Bandwidth constraint: is the tunnel bandwidth.
- Path constraint: includes the explicit path, priority and preemption, route pinning, affinity attribute, and hop limit.

The mechanism for establishing and managing these constraints is called Constraint-based Routing (CR). The device supports the following CRs:

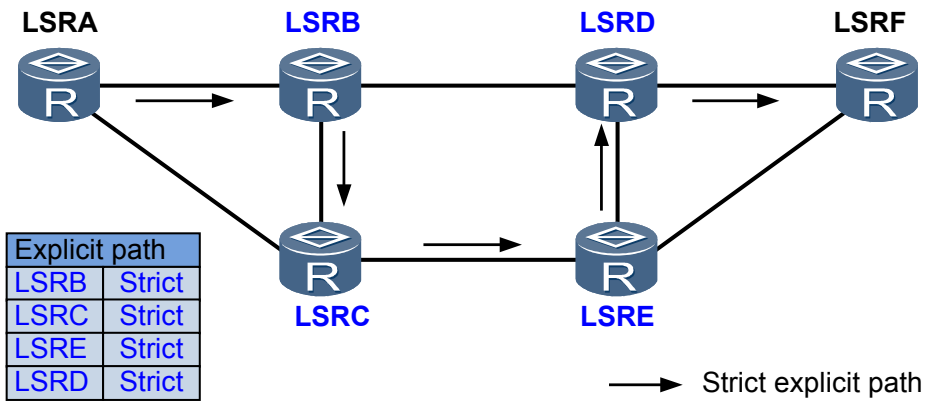
- Tunnel bandwidth

  the values are planned based on services that are to pass through a tunnel. The configured bandwidth is reserved on each node through which a tunnel passes.
- Explicit path

  An explicit path is used to establish a CR-LSP. Nodes to be included or excluded are specified on this path. Explicit paths are classified into the following types:
  - Strict explicit path

    A strict explicit path includes specified nodes through which a CR-LSP must pass. The next hop must be directly connected to the previous hop. By specifying a strict explicit path, the most accurate path is provided for a CR-LSP.

Figure 3-4 Strict explicit path

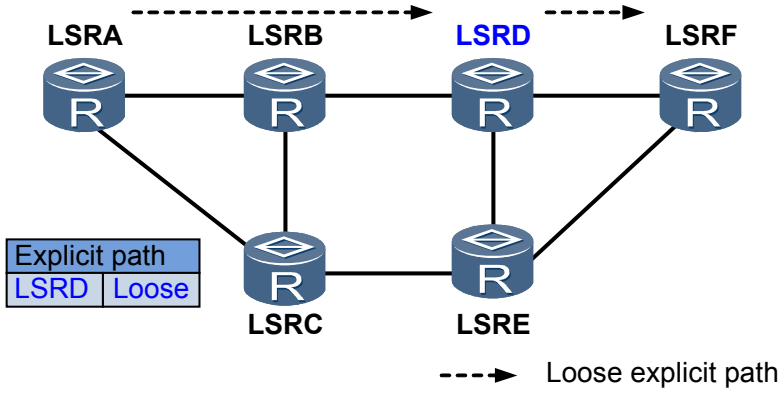


For example, a CR-LSP is set up between LSRA and LSRF on the network shown in **Figure 3-4**. LSRA is the ingress, and LSRF is the egress. "X Strict" specifies the LSR that the CR-LSP must travel through. For example, "LSRB Strict" indicates that the CR-LSP must travel through LSRB, and the previous hop of LSRB must be LSRA. "LSRC Strict" indicates that the CR-LSP must travel through LSRC, and the previous hop of LSRC must be LSRB. The procedure repeats. A path with each node specified is provided for the CR-LSP.

- Loose explicit path

  A loose explicit path contains specified nodes through which a CR-LSP must pass. Other routers that are not specified can also exist on the CR-LSP.

Figure 3-5 Loose explicit path



For example, a CR-LSP is set up over a loose explicit path between LSRA and LSRF on the network shown in **Figure 3-5**. LSRA is the ingress, and LSRF is the egress. "LSRD Loose" indicates that the CR-LSP must pass through LSRD and LSRD and LSRA may not be directly connected. This means that other LSRs may exist between LSRD and LSRA.

- Priorities and preemption

  They are used to allow TE tunnels to be established preferentially to transmit important services, preventing random resource competition during tunnel establishment.

  CR-LSPs use setup and holding priorities to determine whether to preempt resources. A new CR-LSP and an established CR-LSP compete for resources by comparing the priorities. The new path can succeed in preemption when its setup priority is higher than the holding priority of the established path. The priority value ranges from 0 to 7. A smaller value allows for a higher priority. The setup priority must be lower than or equal to the holding priority for a tunnel.

  If there is no path meeting the bandwidth requirement of a desired CR-LSP, a device can tear down an established CR-LSP and use the bandwidth assigned to that CR-LSP to establish a desired CR-LSP. This is called preemption. The following preemption modes are supported:

  - Hard preemption: A CR-LSP with a higher priority can directly preempt resources assigned to a CR-LSP with a lower priority. Some traffic is dropped on the CR-LSP with a lower priority during the hard preemption process.
  - Soft preemption: The **Make-Before-Break** mechanism applies. A CR-LSP with a higher priority has to wait until traffic over a lower-priority CR-LSP switches to another

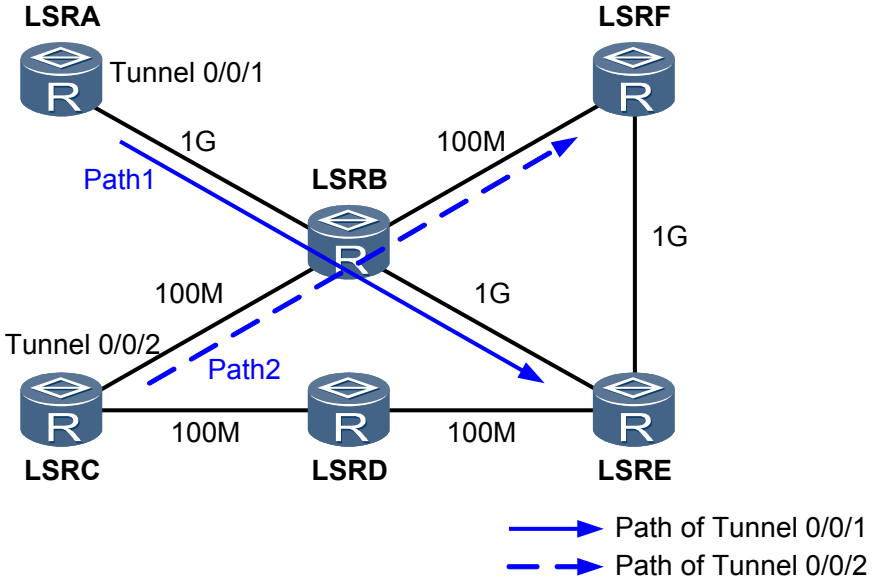
CR-LSP before the higher-priority CR-LSP preempts bandwidth assigned to the lower-priority CR-LSP.

The priority and preemption attributes are used in conjunction to determine resource preemption among tunnels. If multiple CR-LSPs are to be established, CR-LSPs with high priorities can be established by preempting resources. If resources (such as bandwidth) are insufficient, a CR-LSP with a higher setup priority can preempt resources of an established CR-LSP with a lower holding priority.

As shown in **Figure 3-6**, there are two TE tunnels on the network. The link bandwidth allocation is shown in **Figure 3-6** and the links have the same metric value.

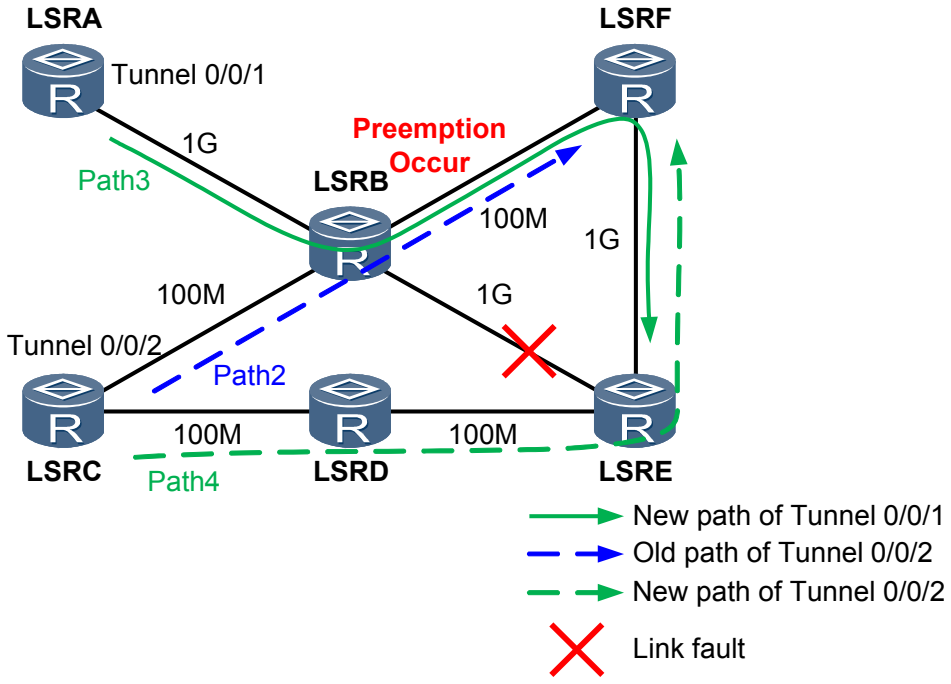
- Tunnel 0/0/1: established over the path LSRA -> LSRB -> LSRE. Its bandwidth is 100 Mbit/s, and its setup and holding priority values are 0.
- Tunnel 0/0/2: established over the path LSRC -> LSRB -> LSRF. Its bandwidth is 100 Mbit/s, and its setup and holding priority values are 7.

Figure 3-6 Before a link fault occurs



If the link between LSRB and LSRE fails, LSRA recalculates a path LSRA -> LSRB -> LSRF -> LSRE for tunnel 0/0/1. The link between LSRB and LSRF is shared by tunnel 0/0/1 and tunnel 0/0/2, but has insufficient bandwidth for these two tunnels. As a result, preemption is triggered, as shown in **Figure 3-7**.

Figure 3-7 After preemption is triggered



The process of establishing a new path for tunnel 0/0/1 is as follows:

1. After MPLS TE path calculation, Path messages are forwarded over the path LSRA -> LSRB -> LSRF -> LSRE and Resv messages are forwarded over the path LSRE -> LSRF -> LSRB -> LSRA.
2. After receiving Resv messages from LSRF, LSRB triggers preemption when it detects that bandwidth is insufficient for resource reservation. The preemption process differs in two preemption modes.
   - In hard preemption, LSRB directly tears down Path2 for tunnel 0/0/2 because the priority of tunnel 0/0/1 is higher than that of tunnel 0/0/2. LSRB sends a PathTear message to LSRF, requiring LSRF to remove Path2 information. In addition, LSRB sends a ResvTear message to LSRC, requiring LSRC to delete the node reservation state. In this case, some traffic on tunnel 0/0/2 is lost.
   - In soft preemption, LSRB sends a ResvTear message to LSRC and establishes a new path Path4 on the condition that LSRB and LSRC do not tear down Path2. After Path4 is established and traffic is switched to it, LSRB tears down Path2 of tunnel 0/0/2.

- Route pinning

  Any changes in the network topology or tunnel attributes may cause an established CR-LSP to be reestablished, leading to the following issues:

  - The reestablished CR-LSP may be over a path that is different from the original one, causing management difficulties.
  - Traffic must switch from the original CR-LSP to the new one, causing traffic loss.

  Route pinning can be used to resolve the preceding problems. Route pinning helps an established CR-LSP remain over a path regardless of route changes. This function improves service traffic continuity and reliability.

- Affinity attribute

  An affinity is a 32-bit vector, configured on the ingress of a tunnel. It must be used together with a link administrative group attribute.

  After a tunnel is configured with an affinity, a device compares the affinity with the administrative group value during link selection to determine whether a link with specified attributes is selected or not. The device implements two AND operations, one between a 32-bit mask and each affinity, and one between the 32-bit mask and the administrative group value. If the two AND operations yield the same results, the path is selected. If the results are different, the path is not selected. The following rules apply:

  - If some bits in a mask are 1s, at least one bit in the administrative group is 1 and the corresponding bit in the affinity must be 1. If some bits in the affinity are 0s, the corresponding bits in the administrative group cannot be 1.

    For example, an affinity is 0x0000FFFF and its mask is 0xFFFFFFFF. The higher-order 16 bits in the administrative group of available links are 0 and at least one of the lower-order 16 bits is 1. This means the administrative group attribute ranges from 0x00000001 to 0x0000FFFF.

  - If some bits in a mask are 0s, the corresponding bits in the administrative group are not compared with the affinity bits.

    For example, an affinity is 0xFFFFFFFF and its mask is 0xFFFF0000. At least one of the higher-order 16 bits in an administrative group attribute is 1 and the lower-order 16 bits can be 0s and 1s. This means that the administrative group attribute ranges from 0x00010000 to 0xFFFFFFFF.

  NOTE

  Understand specific comparison rules before deploying devices of different vendors because the comparison rules vary with the vendor.

  A network administrator can use the link administrative group and affinities to control the paths over which MPLS TE tunnels are established.

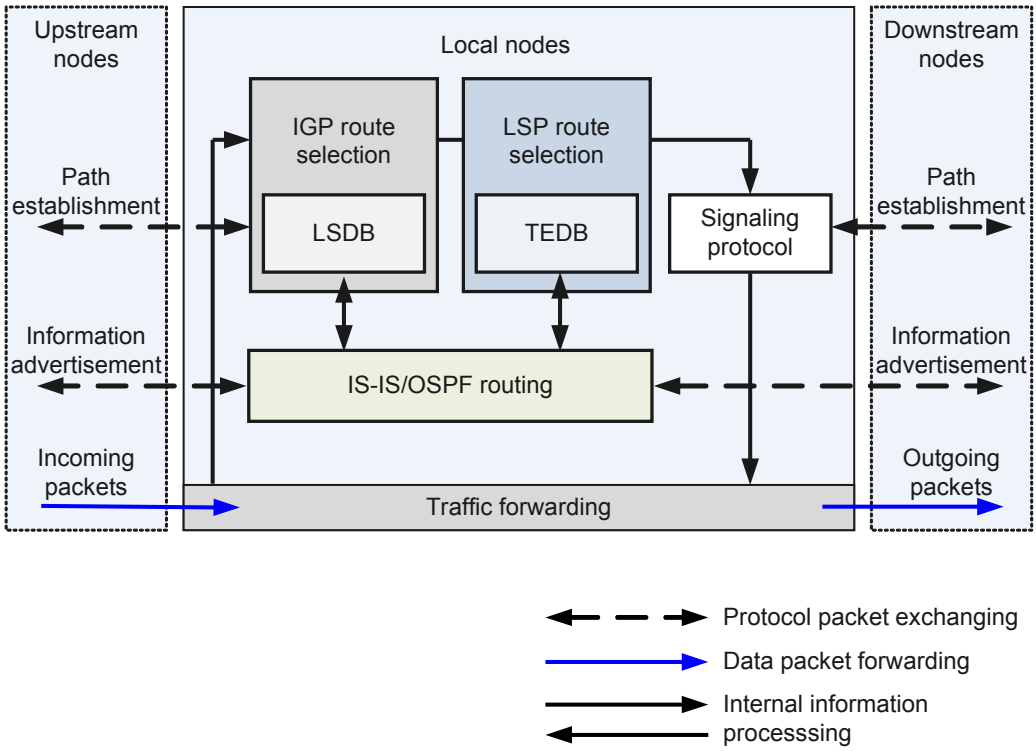
- Hop limit

  Hop limit is a condition for path selection during CR-LSP establishment. Similar to the administrative group and affinity attributes, a hop limit defines the number of hops that a CR-LSP allows.

3.2.2 Implementation

Figure 3-8 shows the MPLS TE implementation framework.

Figure 3-8 MPLS TE implementation framework



Information advertisement, path calculation, path establishment, and traffic forwarding are used for establishing an MPLS TE tunnel. The information advertisement function is used to collect TE related information using IGP. The path calculation function is used for calculating paths based on information collected. The path establishment function establishes paths by exchanging packets between upstream and downstream nodes using signaling protocols. The traffic forwarding function imports data packets to the MPLS TE tunnel and forwards the packets. **Table 3-1** details the functions.

Table 3-1 MPLS TE implementation process

No.	Function	Description
1	**Information Advertisement**	Extends an IGP to advertise TE information, in addition to routing information. TE information includes the maximum link bandwidth, maximum reservable bandwidth, reserved bandwidth, and link colors. Every node collects TE information about all nodes in a local area and generates a traffic engineering database (TEDB).
2	**Path Calculation**	Runs Constraint Shortest Path First (CSPF) and uses TEDB data to calculate a path that satisfies specific constraints. CSPF evolves from the Shortest Path First (SPF) protocol. CSPF excludes nodes and links that do not satisfy specific constraints and uses the same algorithm that SPF supports to calculate a path.

No.	Function	Description
3	**Path Establishment**	Establishes the following types of CR-LSPs: ● Static CR-LSP: set up by manually configuring labels and bandwidth, irrespective of signaling protocols or path calculation. Setting up a static CR-LSP consumes few resources because no MPLS control packets are exchanged between two ends of the CR-LSP. The static CR-LSP cannot be adjusted dynamically in a changing network topology; therefore, the static CR-LSP is not widely used. ● Dynamic CR-LSP: set up using RSVP-TE signaling. RSVP-TE carries parameters, such as the tunnel bandwidth, explicit path, and affinities. There is no need to manually configure each hop along a dynamic CR-LSP. Dynamic CR-LSPs apply to large-scale networks. You can use the **RSVP authentication mechanism** to improve security and reliability during path establishment.
4	**Traffic Forwarding**	Imports traffic to the MPLS TE tunnel and forwards traffic through the tunnel. The first three functions help establish an MPLS TE tunnel. This function forwards traffic after traffic is imported to the tunnel.

NOTE

- A static CR-LSP is manually established, and there is no need to use the information advertisement or the path calculation.
- A dynamic CR-LSP is dynamically established by signaling. Therefore, all the preceding functions are used to establish a dynamic CR-LSP.

When deploying MPLS TE on a network, you need to configure link attributes and tunnel attributes to automatically establish an MPLS TE tunnel. After a tunnel is established, traffic has to be imported to it and forwarded through it.

3.2.3 Information Advertisement

MPLS TE uses routing protocols to advertise resource allocation information about each node on a network. Each node on an MPLS TE network especially the ingress node of a tunnel determines the nodes through which a tunnel passes based on advertised information.

Contents to Be Advertised

The network resource information to be advertised includes the following items:

- Link status information: interface IP addresses, link types, and link metric values, which are collected by an Interior Gateway Protocol (IGP).
- Bandwidth information, such as maximum link bandwidth and maximum reservable bandwidth.
- TE metric: TE link metric, which is the same as the IGP metric by default.
- **Link administrative group**: link color.
- **Affinity Attributes**: color of the link required by TE.

- **SRLG**: shared risk link group. It is a constraint for path calculation of a backup path. SRLG helps prevent backup and primary paths from overlapping over links with the same risk level.

Advertisement Methods

MPLS TE advertises information using extended link-state-based routing protocols, including **OSPF TE** and **IS-IS TE**. Open Shortest Path First (OSPF) TE and Intermediate System to Intermediate System (IS-IS) TE automatically collect TE information and flood it to MPLS TE nodes.

OSPF TE

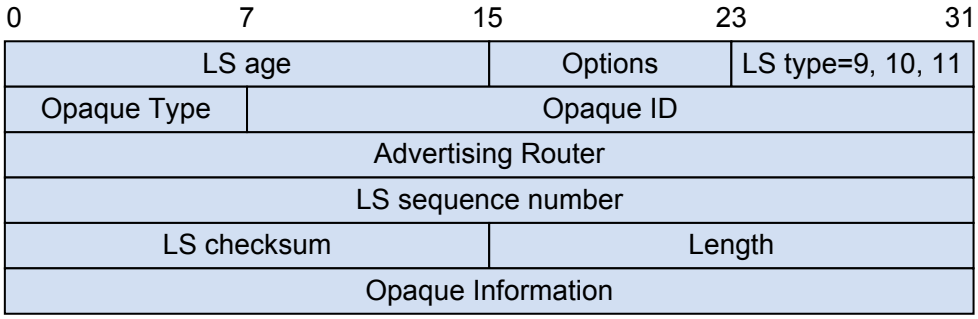
Open Shortest Path First (OSPF) is a link-state-based routing protocol and features strong scalability. OSPF defines label state advertisements (LSAs) of Type 1 to Type 5, and Type 7 to carry the routing information of the intra-area, inter-area, and autonomous system external (AS-external) for route calculation. These LSAs have fixed formats and cannot meet MPLS TE requirements. Therefore, Opaque LSA and TE LSA are introduced.

- **Opaque LSA**

  The Opaque LSA consists of three types of LSAs, that is, Type 9, Type 10, and Type 11. Type 9 Opaque LSA can be flooded only on an interface; Type 10 Opaque LSA can be flooded only within an area; Type 11 Opaque LSA that is similar to Type 5 LSA, can be flooded within the entire AS outside the stub area and the not-so-stubby area (NSSA).

  The Opaque LSA has the similar header format as that of other types of LSAs. The difference is that the 4-byte Link State ID field is divided into Opaque Type and Opaque ID, as shown in **Figure 3-9**.

Figure 3-9 Format of the Opaque LSA

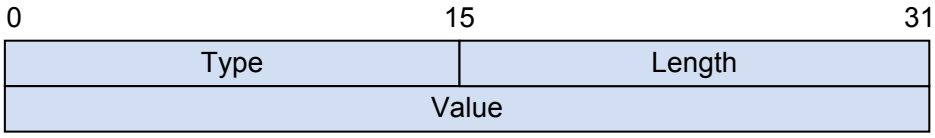


The first byte is Opaque Type that is used to differentiate between application types of this LSA; the other three bytes are Opaque ID that is used to differentiate LSAs of the same application type. The Opaque LSA of the same type may have 255 types of applications, and each application can have 16777216 LSAs in a flooding scope.

For example, the LSA applied in the OSPF graceful restart (GR) is Type 9 LSA with the application type as 3; the LSA applied in TE extensions is Type 10 LSA with the application type as 1.

The LSA carries information in the Opaque Information field. The information format is defined by the application with different requirements. Usually, the extensible type-length-value (TLV) format is used.

Figure 3-10 TLV format



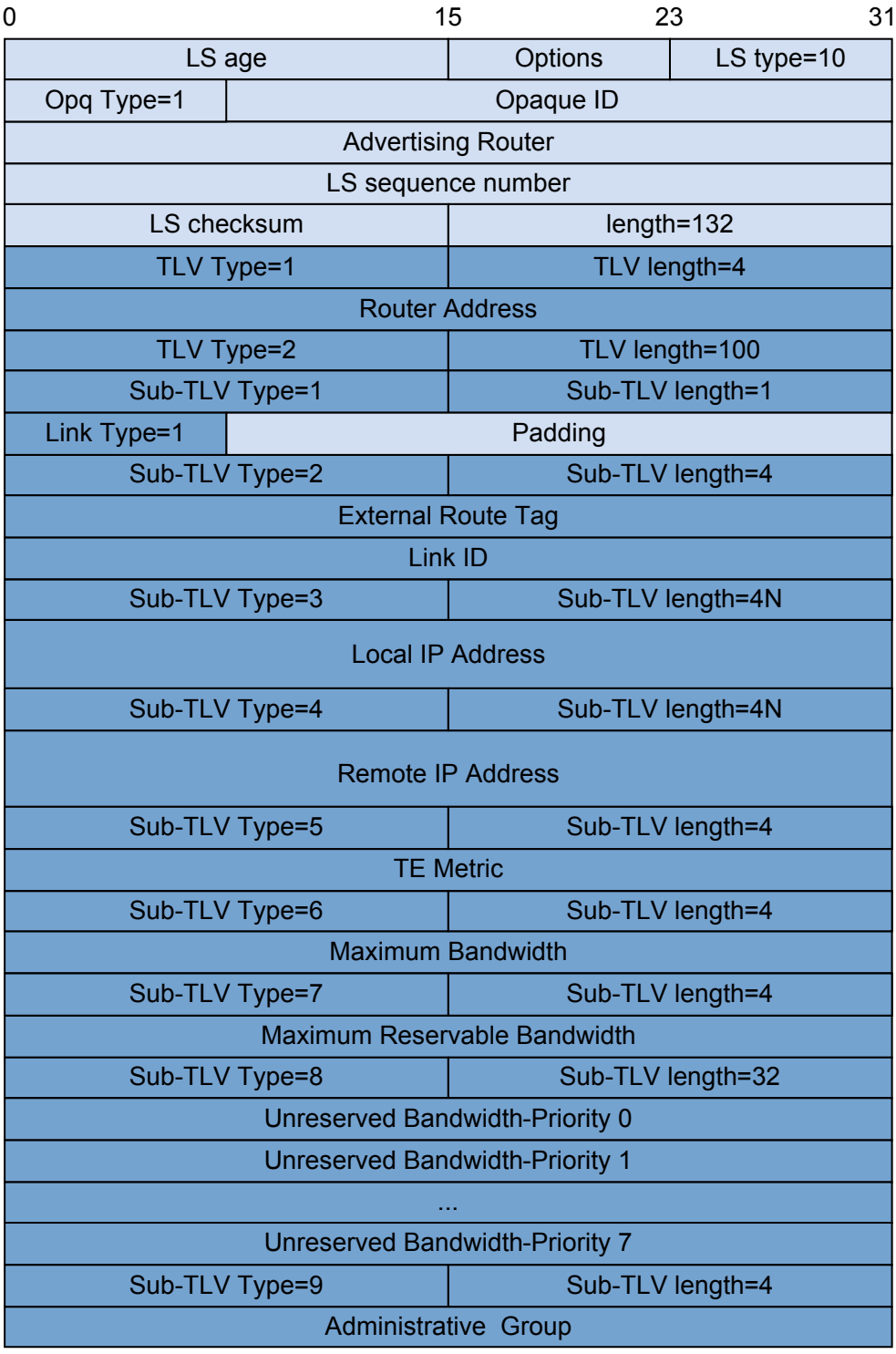
- Type: indicates the message type.
- Length: defines the length of the Value field, in bytes.
- Value: indicates the message carried in TLV. It can be a TLV format. The nested TLV is called sub-TLV.

- **TE LSA Extensions**

  The LSA applied in TE extensions is called TE LSA. The TE LSA is Type 10 LSA, with the application type as 1. Therefore, the TE LSA has the Link State ID in the format of 1.x.x.x; however, the flooding scope is restricted to an area.

  Figure 3-11 shows the typical format of the TE LSA.

Figure 3-11 TE LSA format



The TE LSA uses the TLV format to carry the needed information. At present, two types of TLVs are defined as follows:

- TLV Type 1

  Router address TLV: uniquely identifies an MPLS node. In CSPF, this is known as the router ID.

- TLV Type 2

  Link TLV: carries the attributes of a link enabled with MPLS TE. **Table 3-2** shows the sub-TLVs that can be carried in the Link TLV.

Table 3-2 Sub-TLVs that can be carried in the Link TLV

Sub-TLV	Description
Type1: Link Type (the length of the Value field is 1 byte)	Indicates the link type. ● 1: indicates point-to-point links. ● 2: indicates multi-access links. There is padding of three bytes after the value field of the Type1 sub-TLV.
Type2: Link ID (the length of the Value field is 4 bytes)	Indicates the link ID. It is in the format of an IP address. ● For a point-to-point link, this field indicates the OSPF router ID of the neighbor. ● For a multi-access link, this field indicates the interface IP address of the designated router (DR).
Type3: Local IP Address (the length of the Value field is 4N bytes)	Indicates the IP address of the local interface. It can be IP addresses of several local interfaces. Each IP address occupies 4 bytes.
Type4: Remote IP Address (the length of the Value field is 4N bytes)	Indicates the IP address of the remote interface. It can be IP addresses of several remote interfaces. Each IP address occupies 4 bytes. ● For a point-to-point link, this field is set as the remote IP address. ● For a multi-access link, this field can be set as 0.0.0.0 or skipped.
Type5: Traffic Engineering Metric (the length of the Value field is 4 bytes)	Indicates the TE metric configured on a TE link. The data format is ULONG.
Type6: Maximum Bandwidth (the length of the Value field is 4 bytes)	Indicates the maximum bandwidth of a link. The data format is 4 bytes in floating point.
Type7: Maximum Reservable Bandwidth (the length of the Value field is 4 bytes)	Indicates the maximum reservable bandwidth of a link. The data format is 4 bytes in floating point.
Type8: Unreserved Bandwidth (the length of the Value field is 32 bytes)	Indicates reservable bandwidth of eight priorities of a link. Each priority is in the format of 4 bytes in floating point.
Type9: Administrative Group (the length of the Value field is 4 bytes)	Indicates the administrative group attribute.

If a link is identified as an MPLS TE link, OSPF runs on the link, and OSPF neighbors are established; then the TE extensions to OSPF generates a corresponding TE LSA and advertises it to the area according to this TE link. If other devices in the area also support TE extensions, a network topology consisting of TE links is generated among these devices. Each device that advertises the TE LSA must have a unique Router Address.

The Opaque LSA of Type 10 is advertised within the OSPF area. Therefore, CSPF calculation is area-based. The inter-area LSPs need to be calculated in segments.

IS-IS TE

IS-IS is a routing protocol based on link status. It can be extended to advertise TE information.

In extended IS-IS, two TLVs are supported:

- Type 135: Wide Metric

  IS-IS has two metrics:

  - Narrow metric: 6 bits
  - Wide metric: 32 bits. This TLV is not used in route calculation. Instead, it is for TE information transmission only.

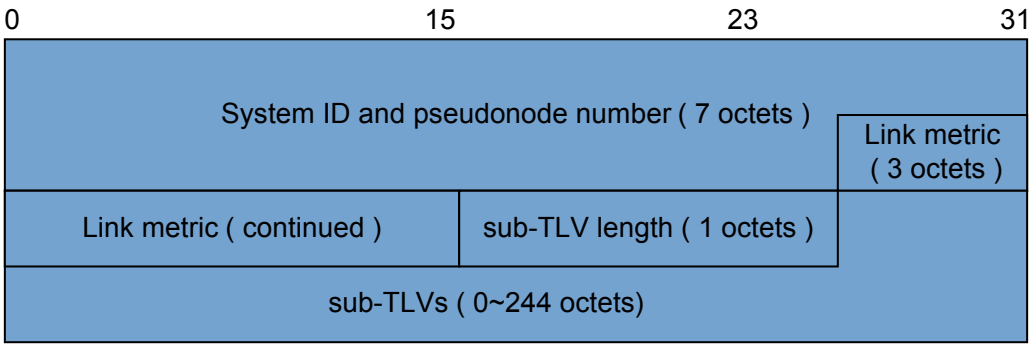
  Narrow metric provides only 64 metric values. So, it cannot meet the requirements of large-scale traffic engineering. Wide metric is introduced to transfer TE information.

  During the transition from the narrow metric to the wide metric, IS-IS TE must support the following metrics:

  - Compatible: capable of receiving and sending the packets with the metric types being narrow or wide
  - Wide Compatible: capable of receiving the packets with the metric types being narrow or wide and sending only the packets with the metric types being wide
- Type 22: IS reachability TLV

  The IS reachability TLV is TLV type 22. **Figure 3-12** shows the format of the IS reachability TLV.

Figure 3-12 Format of the IS reachability TLV



The IS reachability TLV consisting of:

- System ID and pseudo node ID
- Default link metric
- Length of sub-TLVs
- Sub-TLVs with changeable length

Table 3-3 describes sub-TLVs of different types.

Table 3-3 Sub-TLVs in IS-IS TE

Sub-TLV	Description
Type3: Administrative group (4 bytes)	Indicates the administrative group attribute. It has four bytes in length. Each set bit corresponds to one administrative group assigned to the interface.
Type6: IPv4 interface address (4N bytes)	Indicates the IP address of the local interface. It can be IP addresses of several local interfaces. Each IP address occupies four bytes.
Type8: IPv4 neighbor address (4N bytes)	Indicates the IP address of the remote interface. It can be IP addresses of several remote interfaces. Each IP address occupies four bytes. ● For a point-to-point link, this field is set as the remote IP address. ● For a multi-access link, this field is set as 0.0.0.0.
Type9: Maximum link bandwidth (4 bytes)	Indicates the maximum bandwidth of a link.
Type10: Reservable link bandwidth (4 bytes)	Indicates the maximum reservable bandwidth of a link.
Type11: Unreserved bandwidth (32 bytes)	Indicates reservable bandwidth of eight priorities of a link.
Type18: TE Default metric (3 bytes)	Indicates the TE metric configured on a TE link.

When to Advertise Information

OSPF TE or IS-IS TE floods link information so that each node can save area-wide link information in a traffic engineering database (TEDB). Information flooding is triggered by the establishment of an MPLS TE tunnel, or one of the following conditions:

- A specific IGP TE flooding interval elapses.
- A link is activated or deactivated.
- A CR-LSP in an MPLS TE tunnel fails to be established because of insufficient bandwidth.
- Link attributes, such as the administrative group attribute or affinity attribute change.
- The link bandwidth changes.

  When the available bandwidth of an MPLS interface changes, the system automatically updates information in the TEDB and floods it. When a lot of tunnels are to be established on a node, the node reserves bandwidth and frequently updates information in the TEDB and floods it. For example, the bandwidth of a link is 100 Mbit/s. If 100 TE tunnels, each with bandwidth of 1 Mbit/s, are established, the system floods link information 100 times.

  To help suppress the frequency at which TEDB information is updated and flooded, the flooding is triggered based on either of the following conditions:

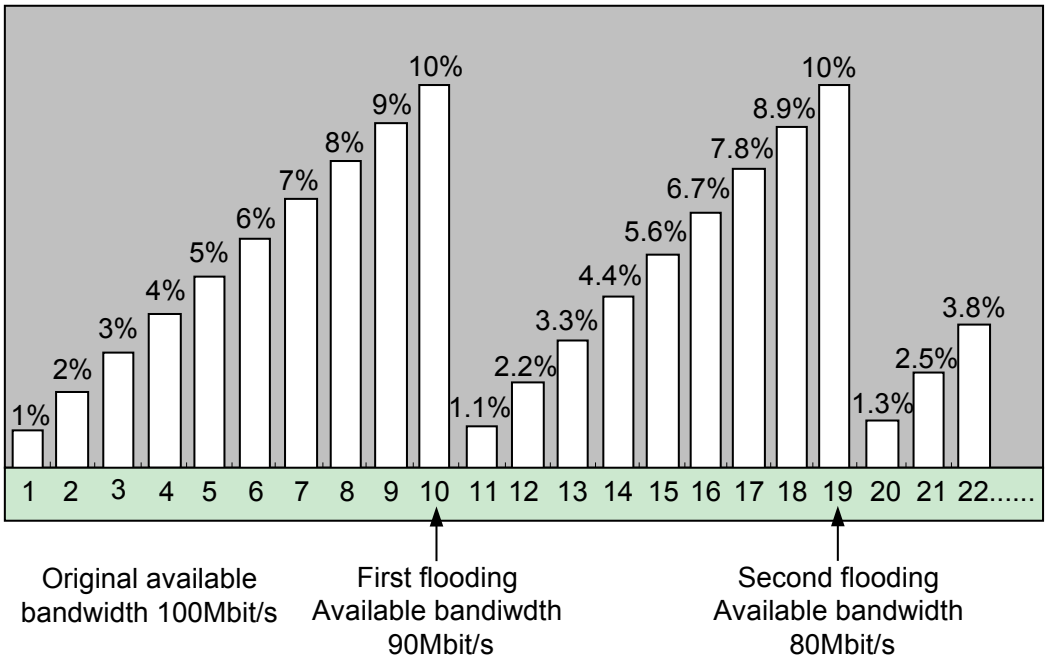
- The proportion of the bandwidth reserved for an MPLS TE tunnel to the available bandwidth in the TEDB is greater than or equal to a specific threshold.
- The proportion of the bandwidth released by an MPLS TE tunnel to the available bandwidth in the TEDB is greater than or equal to a specific threshold.

If either of the preceding conditions is met, an IGP floods link bandwidth information, and the device updates the TEDB.

For example, the available bandwidth of a link is 100 Mbit/s and 100 TE tunnels, each with bandwidth of 1 Mbit/s, are established over the link. The flooding threshold is 10%. The **Figure 3-13** shows the proportion of the bandwidth reserved for each MPLS TE tunnel to the available bandwidth in the TEDB.

Bandwidth flooding is not performed when tunnels 1 to 9 are created. After tunnel 10 is created, the bandwidth information (10 Mbit/s in total) on tunnels 1 to 10 is flooded. The available bandwidth is 90 Mbit/s. Similarly, no bandwidth information is flooded after tunnels 11 to 18 are created. After tunnel 19 is created, bandwidth information on tunnels 11 to 19 is flooded. The process repeats until tunnel 100 is established.

Figure 3-13 Proportion of the bandwidth reserved for each MPLS TE tunnel to the available bandwidth in the TEDB



Results Obtained After Information Advertisement

Every node creates a TEDB in an MPLS TE area after OSPF TE or IS-IS TE floods bandwidth information.

After MPLS TE is deployed on a network, related resource information needs to be advertised to each node. Each node collects information about link constraints and bandwidth usage in the local area to form a database covering network link attributes and topology attributes. This database is called TE Database (TEDB).

A node calculates the optimal path to another node in the MPLS TE area based on information in the TEDB. MPLS TE then establishes a CR-LSP over this optimal path.

The TEDB and IGP link-state data base (LSDB) are independent of each other. Both TEDB and LSDB contain information flooded by IGP, but they have different content and functions. In addition to information contained in the LSDB, TEDB contains TE information. An IGP uses information in an LSDB to calculate the shortest path, while MPLS TE uses information in a TEDB to calculate the optimal path.

3.2.4 Path Calculation

MPLS TE uses constrained shortest path first (CSPF) to calculate the optimal path to a specified node. CSPF, which is derived from SPF, is an algorithm that supports constraints.

CSPF Fundamentals

CSPF works based on the following parameters:

- Bandwidth of an LSP tunnel to be established, explicit path, setup priority, hold priority, and affinity attribute, which are configured on the ingress node of a tunnel
- Traffic engineering database (TEDB)

NOTE

A TEDB can be generated only after Interior Gateway Protocol (IGP) TE is configured. On an IGP TE-incapable network, CR-LSPs are established based on IGP routes, but not CSPF calculation results.

CSPF Calculation Process

CSPF checks the constraints for establishing an LSP to exclude the links that do not meet tunnel attribute requirements in the TEDB, and then calculates the shortest path to the destination of the tunnel by using SPF.

NOTE

When both OSPF TE and IS-IS TE are configured, CSPF attempts to use the OSPF TEDB to establish a path for a CR-LSP. If a path is successfully calculated using OSPF TEDB information, CSPF completes calculation and does not use the IS-IS TEDB to calculate a path. If path calculation fails, CSPF attempts to use IS-IS TEDB information to calculate a path.

CSPF can be configured to use the IS-IS TEDB to calculate a CR-LSP path. If path calculation fails, CSPF uses the OSPF TEDB to calculate a path.

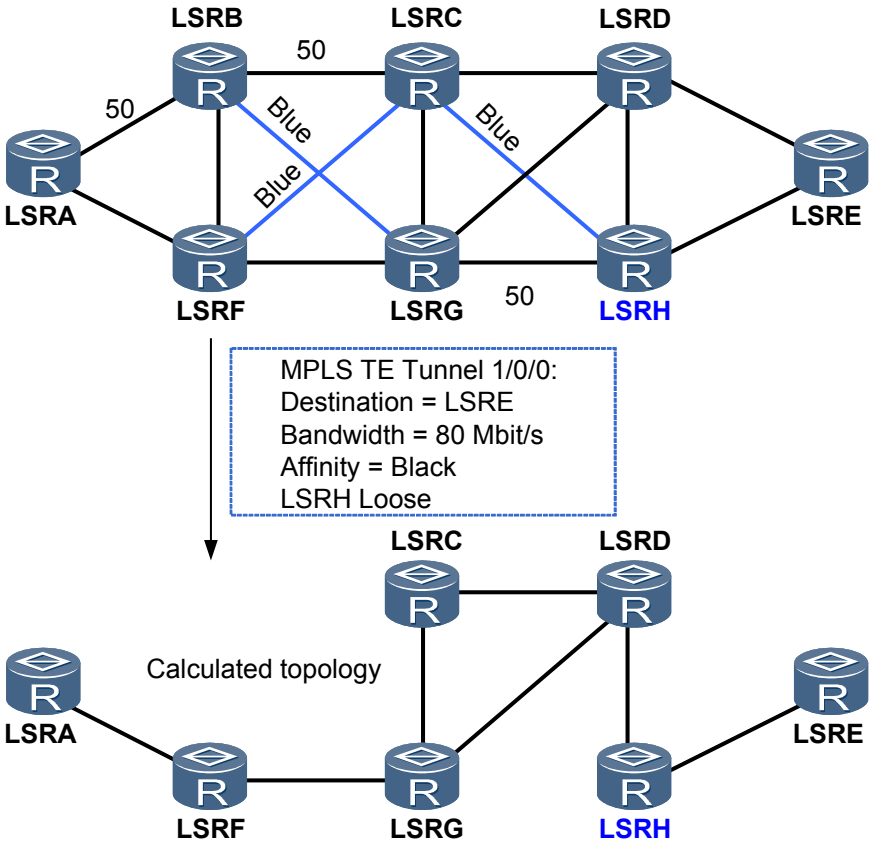
CSPF calculates the shortest path to a destination. If there are several shortest paths with the same metric, CSPF uses a tie-breaking policy to select one of them. The following tie-breaking policies for selecting a path are available:

- Most-fill: selects a link with the highest proportion of used bandwidth to the maximum reservable bandwidth, efficiently using bandwidth resources.
- Least-fill: selects a link with the lowest proportion of used bandwidth to the maximum reservable bandwidth, evenly using bandwidth resources among links.
- Random: selects links randomly, allowing LSPs to be established evenly over links, regardless of bandwidth distribution.

When several links have the same proportion of used bandwidth to the maximum reservable bandwidth, the link discovered first is selected, irrespective of whether most-fill or least-fill is configured.

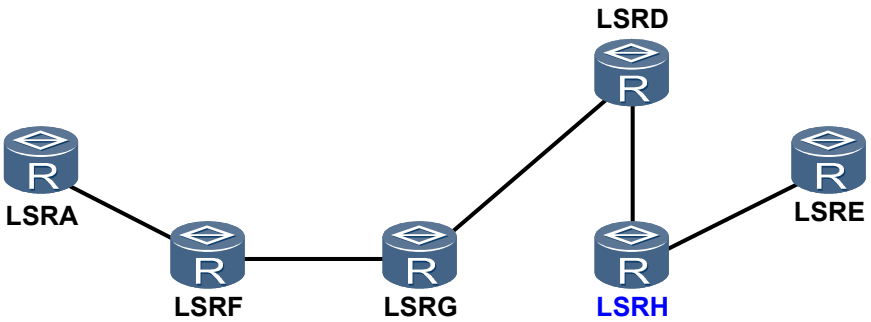
The network shown in **Figure 3-14** illustrates the CSPF path calculation process. CSPF removes links marked blue and links each with bandwidth of 50 Mbit/s based on tunnel constraints and uses other links each with bandwidth of 100 Mbit/s to calculate a path for an MPLS TE tunnel on the network shown in **Figure 3-14**. The constraints include the destination LSRE, bandwidth of 80 Mbit/s, and a transit node LSRH.

Figure 3-14 Process of link removal



CSPF calculates a path shown in **Figure 3-15** in the same way SPF would calculate it.

Figure 3-15 CSPF calculation result



Differences Between CSPF and SPF

CSPF is dedicated to calculating MPLS TE paths. It has similarities with SPF but they have the following differences:

- CSPF calculates the shortest path between the ingress and egress, and SPF calculates the shortest path between a node and each of other nodes on a network.
- CSPF uses tunnel constraints but not link costs between neighboring nodes as the metric.
- CSPF does not support load balancing and uses three tie-breaking policies to determine a path if multiple paths have the same metric.

3.2.5 Path Establishment

3.2.5.1 Path Establishment Modes

CR-LSP Establishment Modes

A CR-LSP can be established statically or dynamically.

Establishment of a static CR-LSP relies on manual configuration by a network administrator. This section describes how a dynamic CR-LSP is established using the RSVP-TE signaling protocol.

RSVP-TE Overview

The Resource Reservation Protocol (RSVP) is designed for the integrated service model and used on each node along a path for resource reservation. The bandwidth reservation capability makes it a suitable signaling protocol for establishing an MPLS TE tunnel.

RSVP-TE is an extension of RSVP to meet MPLS TE requirements. RSVP-TE extends RSVP in the following aspects:

- RSVP-TE appends Label Request objects to **Path messages** to request labels. **Resv messages** carry Label objects that are used to allocate labels.
- The extended RSVP messages can carry information about path constraint parameters, in addition to label binding information.
- RSVP-TE provides the resource reservation function by supporting MPLS TE bandwidth constraints through the extended objects.

RSVP Messages

RSVP has the following message types:

- Path message: This message is sent from a sender to receivers to collect path information of the passing nodes.
- Resv message: This message is sent upstream by the receiver hop-by-hop to respond to the Path message, require resource reservation.
- PathErr message: This message is sent upstream by an node to report errors in processing of the Path messages.
- ResvErr message: This message is sent downstream by an node if errors occur during the processing of the Resv messages.
- PathTear message: This message is sent to remove path state of the passing nodes.
- ResvTear message: This message is sent to remove resource reservation state on the node.

- ResvConf message: This message is sent downstream by the sender hop-by-hop to confirm the resource reservation requests. It is sent only when the Resv message contains the RESV_CONFIRM object.
- Srefresh message: This message refreshes the states of RSVP neighbors.

RSVP-TE Principles

Table 3-4 lists RSVP-TE principles.

Table 3-4 RSVP-TE principles

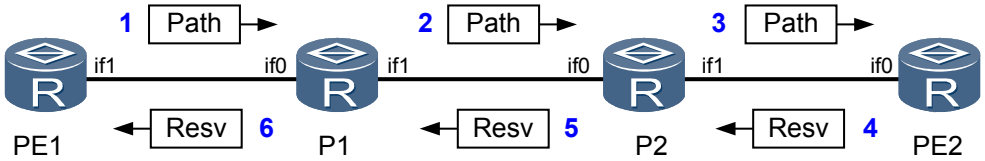
Function Module	Description
3.2.5.2 Establishment of Dynamic CR-LSPs	A CR-LSP is established over a path calculated by CSPF or an explicit path on the ingress.
3.2.5.3 Maintenance of Dynamic CR-LSPs	• **Path Status Maintenance** RSVP-TE sends messages to maintain path status on each node. • **Fault Advertisement** RSVP nodes send advertisements to notify upstream and downstream nodes of faults that occur during path establishment or maintenance. • **Path Teardown** A CR-LSP is torn down and releases labels and bandwidths on each node. The ingress initiates the request for a teardown.

3.2.5.2 Establishment of Dynamic CR-LSPs

To establish dynamic CR-LSPs, an ingress node sends **Path messages** to an egress node and the egress node sends **Resv messages** to the ingress node. Path messages are used to create RSVP sessions and maintain path states, so each node along the path that receives a Path message creates a path state block (PSB). Resv messages carry resource reservation information, so each node along the path that receives a Resv message creates a reserved state block (RSB) and the allocated label.

Figure 3-16 shows the process of establishing an RSVP-TE CR-LSP.

Figure 3-16 Process of establishing an RSVP-TE CR-LSP



1. PE1 uses CSPF to calculate a path between PE1 and PE2. The IP address of every hop on this path has been specified. PE1 generates a Path message and creates a PSB. PE1 then adds the explicit route object (ERO) field containing a list of IP addresses calculated by

CSPF, and sends the Path message to P1 along the path specified by the ERO. **Table 3-5** lists information carried in a Path message.

Table 3-5 Path message on PE1

Object	Value
SESSION	Source: PE1-if1; Destination: PE2-if0
RSVP_HOP	PE1-if1
EXPLICIT_ROUTE	P1-if0; P2-if0; PE2-if0
LABEL	LABEL_REQUEST

2. After P1 receives the Path message, P1 parses the message and creates PSB based on the Path message. P1 then generates a new Path message and sends it to P2 based on the ERO. **Table 3-6** lists information carried in a Path message.
   - In the previous step, PE1 updates the RSVP_HOP field in the Path message to the IP address of the outbound interface when the message is transmitted from PE1 to P1. Similarly, P1 updates the RSVP_HOP field in the Path message to the IP address of the outbound interface when the message is transmitted from P1 to P2.
   - P1 deletes the local LSR ID and IP addresses of the inbound and outbound interfaces from the ERO field in the Path message.

Table 3-6 Path message on P1

Object	Value
SESSION	Source: PE1-if1; Destination: PE2-if0
RSVP_HOP	P1-if1
EXPLICIT_ROUTE	P2-if0: PE2-if0
LABEL	LABEL_REQUEST

3. P2 deals with the received Path message in the same process as that on P1. P2 creates a PSB based on the Path message, updates the new Path message, and sends it to PE2. **Table 3-7** lists information carried in a Path message.

Table 3-7 Path message on P2

Object	Value
SESSION	Source: PE1-if1; Destination: PE2-if0
RSVP_HOP	P2-if1
EXPLICIT_ROUTE	PE2-if0
LABEL	LABEL_REQUEST

4. After PE2 receives a Path message, PE2 knows that itself is the egress of the CR-LSP to be set up based on the Tunnel Address field in the Session object. PE2 then allocates a label and bandwidth resources, and generates an RSB based on the Resv message. The Resv message is sent to P2 and carries the label which is allocated by PE2.

   PE2 extracts an IP address from the RSVP_HOP field of the received Path message and uses it as the destination IP address of the Resv message. The Resv message is forwarded along the reverse path. Therefore, the Resv message does not carry the ERO field. **Table 3-8** lists information carried in a Resv message.

   NOTE

   If the Resv message contains the RESV_CONFIRM object, nodes receiving the Resv message must send a ResvConf message to the generator of the Resv message to confirm the request for resource reservation.

Table 3-8 Resv message on PE2

Object	Value
SESSION	Source: PE2-if0; Destination: PE1-if1
RSVP_HOP	PE2-if0
LABEL	3
RECORD_ROUTE	PE2-if0

5. When P2 receives the Resv message, P2 create an RSB based on the Resv message, allocates a new label, updates the Resv message, and sends the message to P1. **Table 3-9** lists information carried in a Resv message.

Table 3-9 Resv message on P2

Object	Value
SESSION	Source: PE2-if0; Destination: PE1-if1
RSVP_HOP	P2-if0
LABEL	17
RECORD_ROUTE	P2-if0; PE2-if0

6. P1 deals with the received Resv message in the same process as that on P2. P1 updates the Resv message and sends the message to PE1. **Table 3-10** lists information carried in a Resv message.

   PE1 obtains the label allocated by P1 based on the received Resv message. Resource reservation succeeds and a CR-LSP is set up.

Table 3-10 Resv message on P1

Object	Value
SESSION	Source: PE2-if0; Destination: PE1-if1

Object	Value
RSVP_HOP	P1-if0
LABEL	18
RECORD_ROUTE	P1-if0; P2-if0; PE2-if0

3.2.5.3 Maintenance of Dynamic CR-LSPs

Path Status Maintenance

Soft State

Software state indicates that RSVP-TE periodically refreshes RSVP messages to maintain the resource reservation state.

The resource reservation state can be classified into two types: path state and reservation state. Path and Resv messages are created and refreshed periodically to maintain the two states respectively. These messages are called RSVP Refresh messages. RSVP Refresh messages contain PSB and RSB and are used for state synchronization on neighboring RSVP nodes. If a node does not receive any Refresh message about PSB or RSB within a specified time period, the node deletes the path or reservation state.

RSVP Refresh

RSVP messages are transmitted as IP datagrams; therefore the transmission is unreliable. After a CR-LSP is established, each node along the established CR-LSP periodically sends RSVP Refresh messages to its upstream and downstream nodes to synchronize states (including PSB and RSB) of neighboring RSVP nodes.

NOTE

A Refresh message is not a new type of message. Refresh messages are the messages that have already been advertised.

The refreshing interval is specified in the Time Value field.

If the PSB or RSB does not receive any Refresh message about a certain state block after the *keep-multiplier* refreshing intervals elapses, it deletes the state. *keep-multiplier* specifies the number of dropped successive RSVP Refresh message on a node. The default *keep-multiplier* is 3.

Sending of Path and Resv messages between neighboring RSVP nodes is independent to each other.

RSVP Srefresh

RSVP Refresh messages are used to synchronize path state block (PSB) and reservation state block (RSB) information between nodes. They can also be used to monitor the reachability between RSVP neighbors and maintain RSVP neighbor relationships. As the sizes of Path and Resv messages are larger, sending many messages to establish many CR-LSPs causes increased consumption of network resources. RSVP Srefresh can be used to address this problem.

RSVP Srefresh defines new objects based on the existing RSVP protocol:

- Message_ID extension and retransmission extension

According to the Message_ID extension mechanism defined in RFC 2961, RSVP messages carry extended objects, including Message_ID and Message_ID_ACK objects. The two objects are used to confirm RSVP messages and support reliable RSVP message delivery.

The Message_ID object can also be used to provide the RSVP retransmission mechanism. For example, a node initializes a retransmission interval as Rf seconds after it sends an RSVP message carrying the Message_ID object. If the node receives no ACK message within Rf seconds, the node retransmits an RSVP message after (1 + Delta) x Rf seconds. The Delta determines the increased rate of the transmission interval set by the sender. The node keeps retransmitting the message until it receives an ACK message or the retransmission times reach the threshold (called a retransmission increment value). By default, Rf is set to 500 milliseconds, Delta is set to 1, and the retransmission times is set to 3.

- Summary Refresh extension

  The Summary Refresh extension supports Srefresh messages to update the RSVP status, without the transmission of standard Path or Resv messages. The Srefresh extension builds on the Message_ID extension.

  Each Srefresh message carries a Message_ID object. Each object contains multiple messages IDs, each of which identifies a Path or Resv state to be refreshed. If a CR-LSP changes, its message ID value increases.

  Only the state that was previously advertised by Path and Resv messages containing Message_ID objects can be refreshed using the Srefresh extension.

  After a node receives an Srefresh message, the node compares the Message_ID with that saved in a local state block. If they match, the node does not change the state. If the Message_ID is greater than that saved in the local state block, the node sends a NACK message to the sender, refreshes the PSB or RSB based on the Path or Resv message, and updates the Message_ID.

Fault Advertisement

RSVP-TE uses the following messages to advertise LSP errors.

- PathErr message: sent upstream by an RSVP node if an error occurs while this node is processing a Path message. A PathErr message is forwarded by consecutive transit nodes and arrives at the ingress.
- ResvErr message: sent downstream by an RSVP node if an error occurs while this node is processing a Resv message. A ResvErr message is forwarded by consecutive transit nodes and arrives at the egress.

Path Teardown

After a user instructs an ingress to delete a CR-LSP or the ingress receives a ResvErr message, the ingress sends a PathTear message to a downstream node. The downstream node receives this message, tears down the CR-LSP, and replies to the ingress with a ResvTear message.

The functions of PathTear and ResvTear messages are as follows:

- A PathTear message instructs a node to remove saved path information. The PathTear message functions in the opposite way to a Path message.
- A ResvTear message instructs a node to remove resource reservation status. The ResvTear message functions in the opposite way to a Resv message.

3.2.5.4 RSVP-TE Message

RSVP-TE messages exchange information between nodes during the MPLS TE implementation process.

RSVP Message Format

Each type of RSVP messages contains a common header. The length and types of other fields are not fixed. **Figure 3-17** shows the format of RSVP messages.

Figure 3-17 RSVP message format

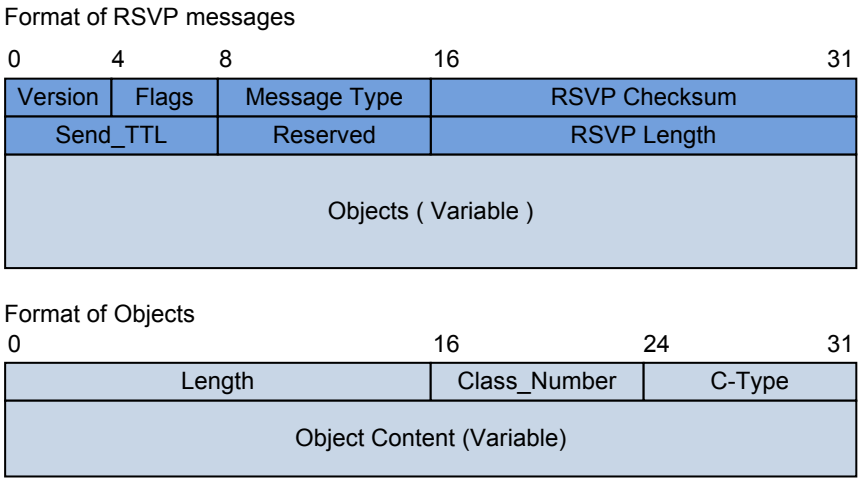


Table 3-11 describes each field in the format.

Table 3-11 Description of fields in RSVP messages

Field	Length	Description
Version	4 bits	Indicates the RSVP version number. Currently, the version is 1.
Flags	4 bits	Indicates the flag bit. Commonly, the value is 0. In RFC 2961, it is extended to identify whether Summary Refresh Extension (Srefresh) is supported. If Srefresh is supported, the value of the flag field is 0x01.
Message Type	8 bits	Indicates an RSVP message type. For example, 1 represents the Path message and 2 represents the Resv message.
RSVP Checksum	16 bits	Indicates the RSVP checksum. The value 0 indicates that no checksum was transmitted.
Send_TTL	8 bits	Indicates the TTL of the message. When a node receives an RSVP message, it compares the Send_TTL and the TTL in the IP header to calculate the hops that the message passes in a non-RSVP area.
Reserved	8 bits	Indicates that the field is reserved.

Field	Length	Description
RSVP Length	16 bits	Indicates the total length of the RSVP message, in bytes.
Objects	Variable	Indicates the object of the RSVP message. Each RSVP message contains kinds of objects. The carried objects vary with types of messages.
Length	16 bits	Indicates the total length of the object, in bytes. Its value must be a multiple of 4, and at least 4.
Class_Number	8 bits	Identifies an object class. Each object class has a name, such as SESSION, SENDER_TEMPLATE, and TIME_VALUE.
C-Type	8 bits	Indicates the object type, unique within the Class_Number. The Class-Number and C-Type is used together to define a unique type for each object.
Object Content	Variable	Indicates contents of objects. The length of this field is changeable.

NOTE

For details of each type of RSVP messages, refer to RFC 3209 and RFC 2205.

Path Message

In RSVP-TE, a Path message is used to create an RSVP session and maintain a path state. The Path message is sent from the ingress node to the egress node in the direction of data flows. On each node, the path state block (PSB) is created.

NOTE

The source IP address of a Path message is the LSR ID of the ingress node and the destination IP address is the LSR ID of the egress node.

Table 3-12 lists some objects carried in the Path message.

Table 3-12 Path message objects

Message Object	Class_Number	C-Type	Object Content
SESSION	1	1	Carries RSVP session information, including the destination address, tunnel ID, and extend tunnel ID.
RSVP_HOP	3	1	Identifies the IP address and the handle of the outgoing interface of the previous hop that sends the Path message.
TIME_VALUE	5	1	Carries the refreshing interval.

Message Object	Class_Number	C-Type	Object Content
SENDER_TEMPLATE	11	1	Specifies the sender IP address and LSP ID.
SENDER_TSPEC	12	2	Defines traffic characteristics of the data flow.
LABEL_REQUEST	19	1	Indicates LABEL_REQUEST object, which is carried only in Path messages.
ADSPEC	13	2	Collects actual QoS parameters about the path, such as estimation of bandwidth of the path, minimal path delay, and path MTU.
EXPLICIT_ROUTE	20	1	ERO, describes information about the path through which the LSP passes. The explicit paths can be strict or loose. Path messages are then forwarded along the specified ERO, without being restricted by IGP shortest path.
RECORD_ROUTE	21	1	RRO, lists the LSRs that the Path message passes when being transmitted. RRO can be used to collect path information and discover route loops. It can also be copied to the next Path message for implementing **Route Pinning**.
SESSION_ATTRIBUTE	207	• 1: LSP_TUNNEL_RA • 7: LSP Tunnel	Specifies the setup priority, hold priority, reservation style, affinity, and other information.

Resv message

After receiving a Path message, the egress node reply with Resv messages. The Resv message, carrying resource reservation information, is sent to the previous node hop-by-hop. Each passing node creates and maintains a reserved state block (RSB) and allocates a label. When the Resv message reaches the ingress node, an LSP is set up successfully.

Table 3-13 describes objects carried in the Resv message.

Table 3-13 Resv message object

Message Object	Class_Number	C-Type	Object Content
INTEGRITY	4	1	Carries cryptographic data to authenticate the originating node and to verify the contents of this RSVP message.
SESSION	1	1	Carries RSVP session information, including the destination address, tunnel ID, and extend tunnel ID.
RSVP_HOP	3	1	Identifies the IP address and the index of the outgoing interface that sends the Resv message.
TIME_VALUE	5	1	Carries the refreshing interval. By default, the value is 30 seconds.
STYLE	8	1	Indicates the resource reservation style. It is specified on the ingress node.
FLOW_SPEC	9	● 1: Reserved (obsolete) flowspec object ● 2: Inv-serv flowspec object	Specifies the QoS characteristics of the data flow.
FILTER_SPEC	10	1	Specifies the sender IP address and LSP ID of the node that sends the message.
RECORD_ROUTE	21	1	RRO, collects the IP address of the incoming interface, LSR-ID, and the IP address of the outgoing interface of the node along the path.
LABEL	16	1	Indicates the assigned label.
RESV_CONFIRM	15	1	Indicates a confirmation of the resource reservation is requested when this object is received. This object carries the IP address of the node that requests a confirmation of the resource reservation.

Reservation Styles

The treatment style of reserving resources for different senders within the same session is called a reservation style. The following reservation styles are supported:

- Fixed Filter (FF) style: creates a separate reservation for a tunnel from a particular sender. This sender does not share its resource reservation with other senders. A resource reservation on the same link is used by a specific CR-LSP.
- Shared Explicit (SE) style: creates a single reservation shared by a set of selected upstream senders. The same resource reservation on the same link is shared by different CR-LSPs.

3.2.6 Traffic Forwarding

Importing Traffic to an MPLS TE Tunnel

The traffic forwarding function imports traffic to a tunnel and forwards traffic over the tunnel. Although the information advertisement, path calculation, and path establishment are used to establish a CR-LSP in an MPLS TE tunnel, a CR-LSP (unlike an LDP LSP) cannot automatically import traffic. The traffic forwarding component must be used to import traffic to the CR-LSP before it forwards traffic.

- **Static Routes**: applies to scenarios with simple network topology or stable network environment.
- **Policy-based Routing**: applies to scenarios that require load balancing and security monitoring.
- **Tunnel Policies**: applies to scenarios in which TE tunnels need to be established to transmit VPN services.
- **Auto Routes**: applies to scenarios with complex network topology or unstable network environment.

Static Routes

Using static routes is the simplest method to import traffic to an MPLS TE tunnel. A TE static route works in the same way as a common static route and has a TE tunnel interface as an outbound interface.

Policy-based Routing

The policy-based routing (PBR) allows the system to select routes based on user-defined policies, improving security and load balancing traffic. If PBR is enabled on an MPLS network, IP packets are forwarded over specific CR-LSPs based on PBR rules.

MPLS TE PBR is implemented based on a set of matching rules and behaviors. The rules and behaviors are defined using an apply clause, in which the outbound interface is a specific tunnel interface. If packets do not match PBR rules, they are properly forwarded using IP; if they match PBR rules, they are forwarded over specific CR-LSPs.

Tunnel Policies

Generally, VPN traffic is forwarded through an LSP but not an MPLS TE tunnel. To import VPN traffic to the MPLS TE tunnel, you need to configure a tunnel policy. Two tunnel policies are available.

- Tunnel type prioritizing policy: Such a policy specifies the sequence in which different types of tunnels are selected by the VPN. You can specify the VPN to select the TE tunnel first.

- Tunnel binding policy: This policy binds a TE tunnel to a s specified VPN by binding a specified destination address to the TE tunnel to provide QoS guarantee.

Auto Routes

An Interior Gateway Protocol (IGP) uses an auto route related to a CR-LSP in a TE tunnel that functions as a logical link to calculate a path. The tunnel interface is used as an outbound interface in the auto route. The TE tunnel is considered a P2P link with a specified metric value. The following auto routes are supported:

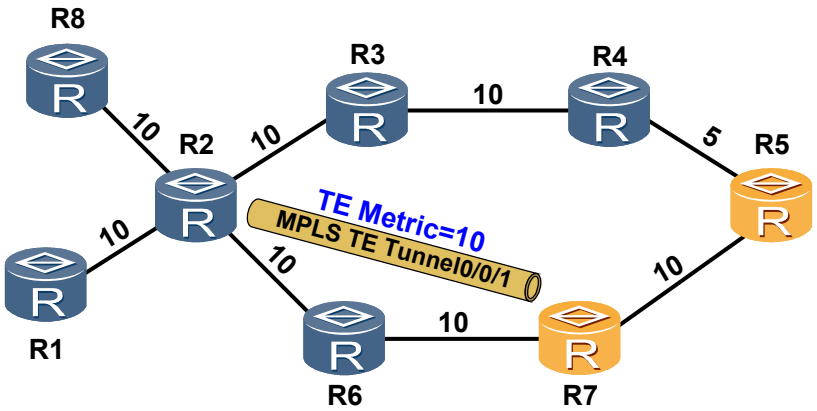
- IGP shortcut: A route related to a CR-LSP is not advertised to neighbor nodes, preventing other nodes from using the CR-LSP.
- Forwarding adjacency: A route related to a CR-LSP is advertised to neighbor nodes, allowing these nodes to use the CR-LSP.

  The forwarding adjacency advertises CR-LSP routes with neighbor IP addresses by sending link-state advertisements (LSAs) or IS-IS link state packets (LSPs). Type 10 Opaque LSAs carry the neighbor IP addresses in the Remote IP Address sub-type-length-value (sub-TLV), and LSPs carry the neighbor IP addresses in intermediate system (IS) reachability TLV's Remote IP Address sub-TLV.

  If the forwarding adjacency is used, nodes on both ends of a CR-LSP must be in the same area.

The following example demonstrates the IGP shortcut and forwarding adjacency.

Figure 3-18 Schematic diagram for IGP shortcut and forwarding adjacency



Node	Mode	Destination	Nexthop	Cost
R5	IGP Shortcut	R2	R4	25
		R1	R4	35
R7		R2	Tunnel0/0/1	10
		R1	Tunnel0/0/1	20
R5	Forwarding Adjacency	R2	R7	20
		R1	R7	30
R7		R2	Tunnel0/0/1	10
		R1	Tunnel0/0/1	20

A CR-LSP over the path R7 → R6 → R2 is established on the network shown in **Figure 3-18**, and the TE metric values are specified. When finding a route from R5 and R7 to R2 and R1 respectively, the result depends on whether the auto route is used.

- The auto route is not used. R5 uses R4 as the next hop in a route to R1 and a route to R2; R7 uses R6 as the next hop in a route to R1 and a route to R2.
- The auto route is used. Either IGP shortcut or forwarding adjacency can be configured:
  - The IGP shortcut is used to advertise the route of Tunnel0/0/1. R5 uses R4 as the next hop in the route to R1 and the route to R2; R7 uses Tunnel0/0/1 as the next hop in the route to R1 and the route to R2. R7, unlike R5, uses Tunnel0/0/1 in IGP path calculation.
  - The forwarding adjacency is used to advertise the route of Tunnel0/0/1. R5 uses R7 as the next hop in the route to R1 and the route to R2; R7 uses Tunnel0/0/1 as the next hop in the route to R1 and the route to R2. Both R5 and R7 use Tunnel0/0/1 in IGP path calculation.

3.2.7 Tunnel Re-optimization

An MPLS TE tunnel can be automatically reestablished over a new optimal path (if one exists) if topology information is updated.

Background

MPLS TE tunnels are used to optimize traffic distribution over a network. An MPLS TE tunnel is configured using static information, such as a bandwidth setting and a calculated path. Without the optimization function, an MPLS TE tunnel cannot be automatically updated after the service bandwidth or a tunnel management policy changes. This wastes network resources. MPLS TE tunnels need to be optimized after being established.

Implementation

A specific event that occurs on the ingress can trigger optimization for a CR-LSP bound to an MPLS TE tunnel. The optimization enables the CR-LSP to be reestablished over the optimal path with the smallest metric.

NOTE

- Re-optimization is disabled by default. If enabled, re-optimization is performed every 3600 seconds by default.
- The **fixed filter (FF) reservation style** and CR-LSP re-optimization cannot be configured together.
- Re-optimization cannot be performed for a CR-LSP that is established over an explicit path.

Re-optimization is classified into the following modes:

- Automatic re-optimization

  When the interval at which a CR-LSP is optimized elapses, Constraint Shortest Path First (CSPF) attempts to calculate a new path. If the calculated path has a metric smaller than that of the existing CR-LSP, a new CR-LSP is set up over the new path. After the CR-LSP is successfully set up, the ingress instructs the forwarding plane to switch traffic to the new CR-LSP and tear down the original CR-LSP. Re-optimization is then complete. If the CR-LSP fails to be set up, traffic is still forwarded along the existing CR-LSP.
- Manual re-optimization

  A re-optimization command is run in the user view to trigger re-optimization.

The **Make-Before-Break** mechanism is used to ensure uninterrupted service transmission during the re-optimization process. Traffic must switch to a new CR-LSP before the original CR-LSP is torn down.

3.2.8 MPLS TE Security

RSVP authentication verifies digest messages carried in RSVP messages to prevent attacks initiated by modified or forged messages. Authentication enhancements can also be used to prevent replay attacks and packet mis-sequence. RSVP authentication and its enhancements improve MPLS TE security.

Background

RSVP uses raw IP to transmit packets. Raw IP has no security mechanism and is prone to attacks. RSVP authentication can be used to verify packets based on keys to prevent attacks.

Original RSVP authentication, however, cannot prevent replay attacks or the problem of neighbor relationship termination resulted from RSVP message mis-sequence. The RSVP authentication enhancements are used to address this problem. The authentication lifetime, handshake, and message window are added as enhanced functions. The authentication enhancements improve security and RSVP neighborhood authentication in a harsh network environment, such as network congestion.

Related Concepts

- Raw IP: similar to UDP but unreliable. No control is provided for raw IP. Whether raw IP datagrams reach their destinations is uncertain.
- Spoofing attack: An unauthorized router establishes a neighbor relationship with a local router or attacks the local router by generating pseudo RSVP messages to establish an RSVP neighbor relationship. The pseudo RSVP messages can reserve lots of bandwidths.
- Replay attack: A remote router repeatedly sends a large number of packets with a sequence number less than the maximum sequence number on a local router. After the local router receives such RSVP packets, the local router terminates the RSVP neighbor relationship with the remote router and tears down the CR-LSP.

Implementation

- Key authentication

  RSVP authentication uses keys carried in packets exchanged between RSVP neighboring nodes to verify those packets, preventing spoofing attacks. The same key must be configured on two RSVP neighboring nodes before they perform RSVP authentication. A local node uses Keyed-Hashing for Message Authentication Message Digest 5 (HMAC-MD5) to calculate a digest for a key, adds this digest as an integrity object into an RSVP message, and sends that message to the remote node. After the remote node receives the message, the node uses the same key and algorithm to calculate a digest and checks whether the local digest is the same as the received one. If they match, the remote node accepts the message. If they do not match, the remote node discards the message.

- Authentication lifetime

  Authentication lifetime specifies how long the RSVP neighbor relationship can last. It provides the following functions:

- When no CR-LSP exists between RSVP neighbors, the RSVP adjacency remains until the RSVP authentication lifetime expires. The configuration of the RSVP authentication time does not affect the status of existing CR-LSPs.
- This function can avoid continuous RSVP authentication. For example, when RSVP authentication is enabled between RTA and RTB, but the key is damaged because the RSVP messages sent from RTA to RTB are incorrect, RTB receives and discards the messages. This can cause RTA to continuously send RTB the faulty RSVP messages and RTB to continuously discard these RSVP messages. The authentication relationship between the neighbors, however, cannot be torn down. In this case, the authentication lifetime needs to be configured. When a neighbor is able to receive a valid RSVP message within the lifetime, the RSVP authentication lifetime resets. Otherwise, the authentication relationship between RSVP neighbors is deleted after the authentication lifetime expires.

- Handshake mechanism

  The handshake mechanism maintains the RSVP authentication status. After RSVP neighboring nodes authenticate each other, they exchanged handshake packets. If they accept the packets, they record a successful handshake. If a local node receives a packet with the sequence number less than the local maximum sequence number, the local node processes the packet as follows:

  - Discards the packet if the packet shows that the handshake mechanism is not enabled on the remote node.
  - Discards the packet if the packet shows that the handshake mechanism is enabled on the remote node and the local node has a record about a successful handshake. If the local node does not have a record about a successful handshake, this packet is the first one arrives at the local node and the local node starts a handshake process.

- Message window

  A message window saves sequence numbers of received RSVP messages. The number of sequence numbers that can be saved ranges from 1 to 64. When the window size is 1, only the largest sequence number of RSVP messages from neighbors can be saved. When the window size is not 1, multiple sequence numbers of RSVP messages from neighbors can be saved. For example, a window size is set to 10, and the largest sequence number of a received RSVP message is 80. The sequence numbers between 71 and 80 can be saved if there is no packet mis-sequence. If a packet mis-sequence problem occurs, the local node arranges the messages and records the 10 largest sequence numbers.

  NOTE

  By default, the window size is 1. Packet processing in the handshake mechanism is based on the prerequisite that the the window size is 1. A non-1 window-size affects packet processing in the handshake mechanism.

RSVP Key Management Modes

RSVP keys can be managed in either of the following modes:

- MD5 key

  An MD5 key is entered in either ciphertext or plaintext on an RSVP interface or node. An MD5 key has the following characteristics:

  - A key cannot be shared. Each protocol is configured with a separate key.
  - An interface or a node is assigned only one key. The key can be reconfigured but cannot be changed.

- Keychain key

  Keychain is an enhanced encryption algorithm. A group of passwords are defined in the format of a password string during keychain authentication, and each password is assigned a specified encryption and decryption algorithm and configured with a validity period. When the system sends or receives a packet, the system selects a valid password. Within the validity period of the password, the system uses the encryption algorithm matching the password to encrypt the packet before sending it out, or uses the decryption algorithm matching the password to decrypt the packet before accepting it. In addition, the system automatically uses a new password after the previous password expires, minimizing password decryption risks.

  Keychain management has the following characteristics:

  - A keychain authentication password and the encryption and decryption algorithms must be configured. A password validity period can also be configured.
  - Keychain settings can be shared by separate protocols and features and can be managed uniformly.

  Keychain can be used on an RSVP interface and node and support HMAC-MD5.

Leveled RSVP Authentication

Leveled RSVP authentication is supported

- Neighbor-oriented authentication

  You can configure authentication information, such as authentication keys, based on different neighbor addresses. RSVP then authenticates each neighbor separately.

  The following configuration options available:

  - The IP address of an interface on an RSVP neighboring node as an RSVP neighbor address.
  - The LSR ID of an RSVP neighboring node is used as an RSVP neighbor address.
- Interface-oriented authentication

  Authentication is configured on interfaces, and RSVP authenticates messages based on inbound interfaces.

Neighbor-oriented authentication has a higher priority than interface-oriented authentication. A node discards messages if neighbor-oriented authentication fails and performs interface-oriented authentication only if neighbor-oriented authentication is not enabled.

3.2.9 MPLS TE Reliability

3.2.9.1 Reliability Overview

MPLS TE reliability techniques need to prevent or minimize packet loss that occurs in one of the following situations:

- If attributes, such as bandwidth, are modified when an MPLS TE tunnel is transmitting services, the tunnel is reestablished using new attributes, and services switch to the new path.
- If a node or link fails while an MPLS TE tunnel is transmitting services, a backup CR-LSP is established and takes over traffic.

- An MPLS TE tunnel has been established and is transmitting services. A fault occurs in the control plane but not the forwarding plane of a node along the path. Traffic needs to be forwarded uninterruptedly before the control plane recovers.

MPLS TE tunnels that transmit mission-critical services require high reliability. **Table 3-14** lists MPLS TE reliability functions.

Table 3-14 MPLS TE reliability functions

Technique Classification	Description	Functions
Reliability mechanism for updating MPLS TE attributes	Ensures reliable traffic transmission after attributes are updated and a new CR-LSP is established using the updated attribute and takes over traffic.	• **Make-Before-Break**
Fault detection	Rapidly detects MPLS TE network faults to speed up a protection switchover.	• **RSVP Hello** • **BFD for MPLS TE**
Traffic protection	Supports network-level reliability, including E2E path protection and local protection.	• **CR-LSP Backup** • **TE FRR** • **SRLG** • **TE Tunnel Protection Group**
	Supports device-level reliability, including uninterrupted traffic transmission on the forwarding plane while a fault occurs on the control plane of a node.	• **RSVP GR**

3.2.9.2 Make-Before-Break

The Make-Before-Break mechanism prevents traffic loss during a traffic switchover between two CR-LSPs. This mechanism improves MPLS TE tunnel reliability.

Background

If an MPLS TE tunnel is no longer the optimal path due to link attribute or tunnel attribute changes, a new CR-LSP will be established according to new attributes. After the new CR-LSP is established, traffic is switched to it. If traffic is switched away from the original MPLS TE tunnel before the new CR-LSP is successfully established, traffic loss occurs. MPLS TE provides the Make-Before-Break mechanism to prevent this problem.

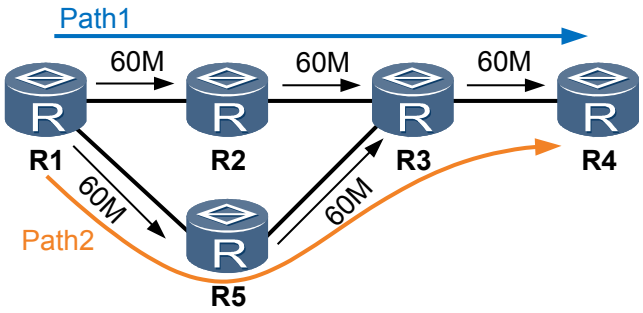
Principles

Make-Before-Break is a mechanism that allows a CR-LSP to be established using changed bandwidth and path attributes over a new path before the original CR-LSP is torn down. It helps minimize data loss and additional bandwidth consumption. The new CR-LSP is called a modified

CR-LSP. Make-Before-Break is implemented using the **shared explicit (SE)** resource reservation style.

The new CR-LSP competes with the original CR-LSP on some shared links for bandwidth. The new CR-LSP cannot be established if it fails the competition. The Make-Before-Break mechanism allows the system to reserve bandwidth used by the original CR-LSP for the new CR-LSP, without calculating the bandwidth to be reserved. Additional bandwidth is used if links on the new path do not overlap the links on the original path.

Figure 3-19 Schematic diagram for Make-Before-Break



In this example, the maximum reservable bandwidth on each link is 60 Mbit/s on the network shown in **Figure 3-19**. A CR-LSP along the path Path1 is established, with the bandwidth of 40 Mbit/s.

The path is expected to change to Path2 to forward data because R5 has a light load. The reservable bandwidth of the link between R3 and R4 is just 20 Mbit/s. The total available bandwidth for the new path is less than 40 Mbit/s. The Make-Before-Break mechanism can be used in this situation. The Make-Before-Break mechanism allows the newly established CR-LSP over the path Path2 to use the bandwidth of the original CR-LSP's link between R3 and R4. After the new CR-LSP is established over the path, traffic switches to the new CR-LSP, and the original CR-LSP is torn down.

In addition to the preceding method, another method of increasing the tunnel bandwidth can be used. If the reservable bandwidth of a shared link increases to a certain extent, a new CR-LSP can be established.

In the example shown in **Figure 3-19**, the maximum reservable bandwidth on each link is 60 Mbit/s. A CR-LSP along the path Path1 is established, with the bandwidth of 30 Mbit/s.

The path is expected to change to Path2 to forward data because R5 has a light load, and the bandwidth is expected to increase to 40 Mbit/s. The reservable bandwidth of the link between R3 and R4 is just 30 Mbit/s. The total available bandwidth for the new path is less than 40 Mbit/s. The Make-Before-Break mechanism can be used in this situation. The Make-Before-Break mechanism allows the newly established CR-LSP over the path Path2 to use the bandwidth of the original CR-LSP's link between R3 and R4. The bandwidth of the new CR-LSP is 40 Mbit/s, out of which 30 Mbit/s is released by the link between R3 and R4. After the new CR-LSP is established, traffic switches to the new CR-LSP and the original CR-LSP is torn down.

Delayed Switchover and Deletion

If an upstream node on an MPLS network is busy but its downstream node is idle or an upstream node is idle but its downstream node is busy, a CR-LSP may be torn down before the new CR-LSP is established, causing a temporary traffic interruption.

To prevent this temporary traffic interruption, the switching and deletion delays are used together with the Make-Before-Break mechanism. In this case, traffic switches to a new CR-LSP a specified delay time later after a new CR-LSP is established. The original CR-LSP is torn down a specified delay later after a new CR-LSP is established. The switching delay and deletion delay can be manually configured.

3.2.9.3 RSVP Hello

The RSVP Hello extension can rapidly monitor the reachability of RSVP nodes. If an RSVP node becomes unreachable, TE FRR protection is triggered. The RSVP Hello extension can also monitor whether an RSVP GR neighboring node is in the restart process.

Background

RSVP Refresh messages are used to synchronize path state block (PSB) and reservation state block (RSB) information between nodes. They can also be used to monitor the reachability between RSVP neighbors and maintain RSVP neighbor relationships.

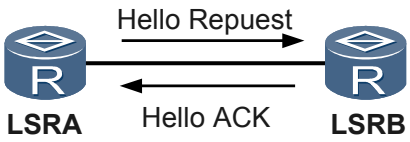
Using Path and Resv messages to monitor neighbor reachability delays a traffic switchover if a link fault occurs and therefore is slow. The RSVP Hello extension can address this problem.

Implementation

The principles of the RSVP Hello extension are as follows:

1. Hello handshake mechanism

   Figure 3-20 Hello handshake mechanism

   LSRA and LSRB are directly connected on the network shown in **Figure 3-20**.

   - If RSVP Hello is enabled on LSRA, LSRA sends a Hello Request message to LSRB.
   - After LSRB receives the Hello Request message and is also enabled with RSVP Hello, LSRB sends a Hello ACK message to LSRA.
   - After receiving the Hello ACK message, LSRA considers LSRB reachable.

2. Detecting neighbor loss

   After a successful Hello handshake is implemented, LSRA and LSRB exchange Hello messages. If LSRB does not respond to three consecutive Hello Request messages sent by LSRA, LSRA considers router B lost and re-initializes the RSVP Hello process.

3. Detecting neighbor restart

   If LSRA and LSRB are enabled with RSVP GR, and the Hello extension detects that LSRB is lost, LSRA waits for LSRB to send a Hello Request message carrying a GR extension. After receiving such a message, LSRA helps LSRB to restore the RSVP state and sends a Hello ACK message to LSRB. After receiving the Hello ACK message, LSRB performs the GR process and restores the RSVP soft state. LSRA and LSRB exchange Hello messages to maintain the restored RSVP soft state.

NOTE

On conditions when LSRA and LSRB are on the same CR-LSP:

- If GR is disabled and TE FRR is enabled, TE FRR switches traffic to a bypass CR-LSP after the Hello extension detects that the RSVP neighbor relationship is lost to ensure proper traffic transmission.
- If GR is enabled, the GR process is performed.

Deployment Scenarios

The RSVP Hello extension applies to networks enabled with both RSVP GR and TE FRR.

3.2.9.4 CR-LSP Backup

CR-LSP backup techniques protect E2E MPLS TE tunnels. If the ingress detects that the primary CR-LSP is unavailable, the ingress switches traffic to a backup CR-LSP. After the primary CR-LSP recovers, traffic switches back.

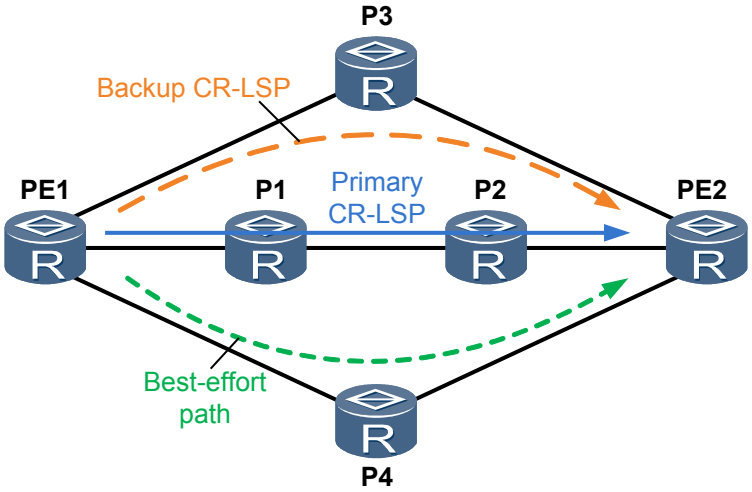
Related Concepts

CR-LSP backup functions include hot standby, ordinary backup, and the best-effort path function.

- Hot-standby: The backup CR-LSP is set up immediately after the primary CR-LSP is set up. If the primary CR-LSP fails, the backup CR-LSP takes over traffic from the primary CR-LSP.
- Ordinary backup: When the primary CR-LSP fails, the backup CR-LSP is set up and takes over traffic from the primary CR-LSP.
- Best-effort path: The failures of both primary and backup CR-LSPs can trigger the setup of a temporary CR-LSP, which is called the best-effort path. Traffic is then switched to the best-effort path.

  For example, the primary CR-LSP is established over the path PE1 → P1 → P2 → PE2, and the backup CR-LSP is established over the path PE1 → P3 → PE2 shown in **Figure 3-21**. If both CR-LSPs fail, PE1 establishes a best-effort path PE1 → P4 → PE2 to take over traffic.

Figure 3-21 Best-effort path



NOTE

A best-effort path has no bandwidth reserved for traffic, but has an affinity and a hop limit configured as needed.

Implementation

The procedure of CR-LSP backup is as follows:

1. Planning is implemented.

   Plan the paths, bandwidth values, and deployment modes. **Table 3-15** lists CR-LSP backup deployment items.

Table 3-15 CR-LSP backup deployment

Item	Hot Standby	Ordinary Backup	Best-Effort Path
Path	Determine whether the primary and hot-standby CR-LSPs partially overlap. A hot-standby CR-LSP can be established over an explicit path. The hot-standby CR-LSP supports the following constraints: ● Explicit path ● Affinity ● Hop limit ● **Overlapping Path for a Hot-standby CR-LSP**	The path of the backup CR-LSP partially overlaps the path of the primary CR-LSP regardless of whether the backup CR-LSP is set up along an explicit path. The ordinary backup CR-LSP supports the following constraints: ● Explicit path ● Affinity ● Hop limit	Automatically calculated by the ingress. The best-effort path supports the following constraints: ● Affinity ● Hop limit
Bandwidth	A hot-standby CR-LSP and a primary CR-LSP have the same bandwidth by default. **Dynamic Bandwidth Protection for Hot-standby CR-LSPs** is supported and ensures that a hot-standby CR-LSP does not use additional bandwidth when transmitting traffic.	An ordinary backup CR-LSP and a primary CR-LSP have the same bandwidth.	A best-effort path is only a protection path that does not have reserved bandwidth.
Deployment mode	Can be established without attribute templates.	Can be established without attribute templates.	Can be established without attribute templates.

Item	Hot Standby	Ordinary Backup	Best-Effort Path
	Can be established using attribute templates.	Can be established using attribute templates.	Automatically established and does not support attribute templates.
Configuration combination	● If established without an attribute template, a hot-standby CR-LSP can be used together with a best-effort path. ● If established using an attribute template, a hot-standby CR-LSP can be used together with both an ordinary backup CR-LSP and a best-effort path.	● If established without an attribute template, an ordinary CR-LSP can only be used alone. ● If established using an attribute template, an ordinary backup CR-LSP can be used together with a hot-standby backup CR-LSP and a best-effort path.	-

2. CR-LSPs are established in sequence.

   You can establish backup CR-LSPs using different modes on the same tunnel. To quickly establish a CR-LSP for service transmission, the system attempts to establish a backup CR-LSP using different modes in sequence until the backup CR-LSP is successfully established.

   The rules for establishing a CR-LSP are as follows:

   a. If new tunnel configuration is committed or a tunnel goes Down, the ingress first attempts to establish a primary CR-LSP. If the attempt fails, the ingress attempts to establish a hot-standby CR-LSP. If establishing the hot-standby CR-LSP fails, the ingress then attempts to establish an ordinary backup CR-LSP. If this attempt also fails, the ingress establishes a best-effort path.

   b. A maximum of three CR-LSP attribute templates can be configured for hot-standby CR-LSPs and three for ordinary backup CR-LSPs. These templates are prioritized. The ingress uses each in descending order by priority until a CR-LSP is successfully established.

   c. If a CR-LSP has been established using a lower-priority attribute template and the CR-LSP status changes, the ingress will attempt to establish a CR-LSP using a higher-priority attribute template. The Make-Before-Break mechanism ensures that traffic is uninterrupted when a new CR-LSP is being established.

   d. If a stable CR-LSP has been established using any of the attribute templates, you can lock the used backup CR-LSP attribute template. After the attribute template is locked, the ingress will not attempt to use a higher-priority attribute template to establish a CR-LSP. This locking function prevents unnecessary traffic switchovers and lowers system costs.

3. Backup CR-LSP attributes are modified.

   When the constraints for backup CR-LSPs are modified, the ingress triggers re-establishment of a backup CR-LSP. The system uses the Make-Before-Break mechanism to re-establish a backup CR-LSP. After that backup CR-LSP has been successfully

reestablished, traffic on the original backup CR-LSP (if it is transmitting traffic) switches to this new backup CR-LSP, and the original backup CR-LSP is torn down.

4. Fault detection is implemented.

   CR-LSP backup supports the following fault detection functions:

   - The RSVP-TE fault advertisement mechanism sends signaling packets to detect faults at a low speed.
   - Bidirectional forwarding detection (BFD) for CR-LSP rapidly detects faults. This is a recommended function.

5. A traffic switchover is implemented.

   If a primary CR-LSP fails, the ingress attempts to switch traffic from the primary CR-LSP to a hot-standby CR-LSP. If the hot-standby CR-LSP is unavailable, the ingress attempts to switch traffic to an ordinary backup CR-LSP. If the ordinary backup CR-LSP is unavailable, the ingress attempts to switch traffic to a best-effort path.

6. A traffic switchback is implemented.

   Traffic switches back to a path based on the available CR-LSPs. Traffic will switch first to the primary CR-LSP, which has the highest priority. If the primary CR-LSP is unavailable, traffic will switch to the hot-standby CR-LSP. The ordinary CR-LSP has the lowest priority.

Dynamic Bandwidth Protection for Hot-standby CR-LSPs

Hot-standby CR-LSPs support dynamic bandwidth protection. The dynamic bandwidth protection function allows a hot-standby CR-LSP to obtain bandwidth resources only after the hot-standby CR-LSP takes over traffic from a faulty primary CR-LSP. This function uses network resources efficiently and reduces network costs.

Dynamic bandwidth protection ensures that the hot-standby CR-LSP does not use bandwidth, while the primary CR-LSP is transmitting traffic. The dynamic bandwidth protection process is as follows:

1. If the primary CR-LSP fails, traffic immediately switches to the hot-standby CR-LSP with 0 bit/s bandwidth. The ingress uses the Make-Before-Break mechanism to establish a hot-standby CR-LSP.
2. After the new hot-standby CR-LSP has been successfully established, the ingress switches traffic to this CR-LSP and tears down the hot-standby CR-LSP with 0 bit/s bandwidth.
3. After the primary CR-LSP recovers, traffic switches back to the primary CR-LSP. The hot-standby CR-LSP then releases the bandwidth it uses and the ingress establishes another hot-standby CR-LSP with no bandwidth.

Overlapping Path for a Hot-standby CR-LSP

The path overlapping function can be configured for hot-standby CR-LSPs. This function allows the path of a hot-standby CR-LSP partially overlaps the path of the primary CR-LSP. After the hot-standby CR-LSP is established, it can protect traffic on the primary CR-LSP.

3.2.9.5 TE FRR

TE FRR protects links and nodes on CR-LSPs bound to an MPLS TE tunnel. If a link or node fails, TE FRR rapidly switches traffic to a backup path, minimizing traffic loss.

Background

A link or node failure triggers a primary/backup CR-LSP switchover. IGP routes of the backup path need to converge, and CSPF recalculates a path over which a CR-LSP is established. Traffic is dropped during this process.

TE FRR can be used to prevent traffic loss. After a link or node fails, TE FRR establishes a bypass CR-LSP, which excludes the faulty link or node. The bypass CR-LSP can rapidly take over traffic, minimizing traffic loss. The ingress can reestablish a primary CR-LSP.

Related Concepts

Figure 3-22 Local protection

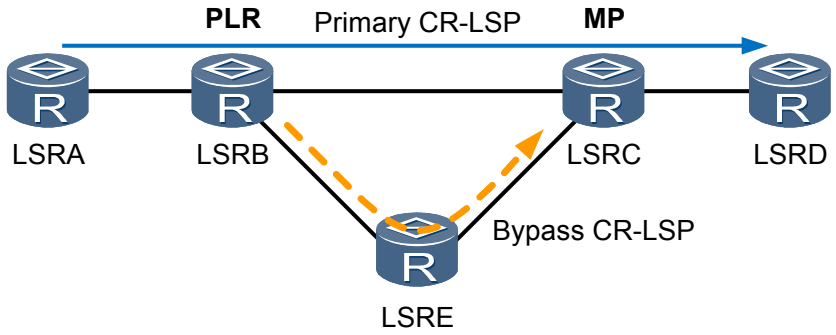


Table 3-16 describes TE FRR concepts.

Table 3-16 TE FRR concepts

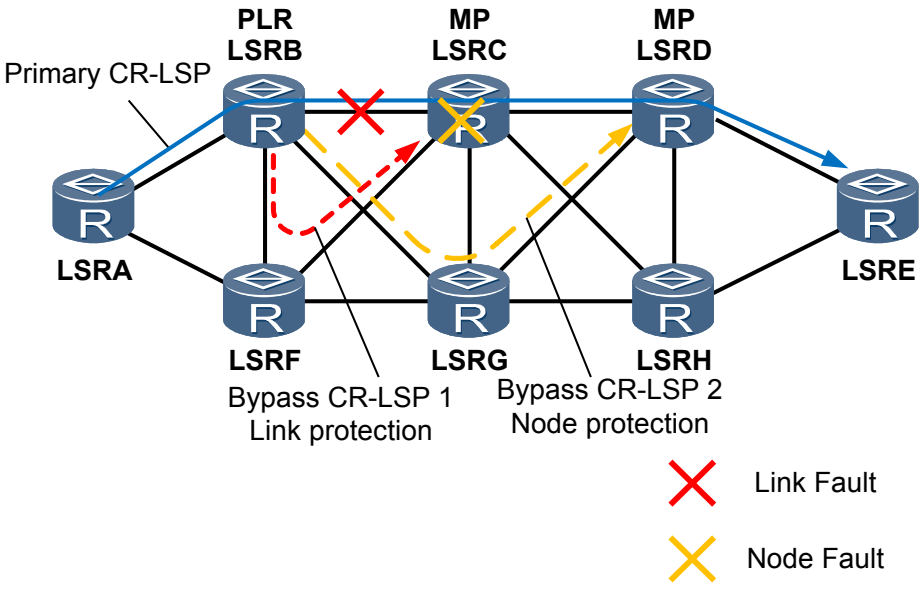
Concept	Description
Primary CR-LSP	A CR-LSP that is protected.
Bypass CR-LSP	A CR-LSP that protects the primary CR-LSP. The bypass CR-LSP is usually in the idle state and transmits few data. If the bypass CR-LSP needs to forward service data when it protects the primary CR-LSP, sufficient bandwidth must be allocated to the bypass CR-LSP.
Point of Local Repair (PLR)	The ingress of the bypass CR-LSP. It must be on the path of the primary CR-LSP. The PLR can be the ingress, not the egress of the primary CR-LSP.
Merge point (MP)	The egress of the bypass CR-LSP. It must be on the path of the primary CR-LSP. The MP cannot be the ingress of the primary CR-LSP.

Table 3-16 describes TE FRR protection functions.

Table 3-17 TE FRR protection functions

Classified By	Type	Description
Object to be protected	Link protection	As shown in **Figure 3-23**, the PLR (LSRB) and MP (LSRC) are directly connected, and the primary CR-LSP passes through the direct link. Bypass CR-LSP 1 protects the direct link.
	Node protection	As shown in **Figure 3-23**, a primary CR-LSP between the PLR (LSRB) and MP (LSRD) passes through LSRC. Bypass CR-LSP 2 protects LSRC on the primary CR-LSP.
Bandwidth	Bandwidth protection	The bandwidth of a bypass CR-LSP is higher than or equal to that of the primary CR-LSP. The bypass CR-LSP protects the primary CR-LSP and its bandwidth.
	Non-bandwidth protection	No bandwidth is assigned to a bypass CR-LSP. The bypass CR-LSP protects only the path of the primary CR-LSP.
Implementation	Manual protection	A manually configured bypass CR-LSP is established and bound to a CR-LSP that is to be protected. If a link or node on the protected CR-LSP fails, traffic automatically switches to the bypass CR-LSP.
	Auto FRR protection	An Auto FRR-enabled node automatically establishes a bypass CR-LSP. The node binds the bypass CR-LSP to a primary CR-LSP if the node receives an FRR protection request and the FRR topology requirements are met.

Figure 3-23 TE FRR link and node protection



NOTE

A bypass CR-LSP supports the combination of protection types. For example, manual protection, node protection, and bandwidth protection can be implemented together on a bypass CR-LSP.

Implementation

The PLR implements TE FRR as follows:

1. Establishes a primary CR-LSP.

   The establishment of a primary CR-LSP is the same as that of a common CR-LSP. The only difference is that the tunnel ingress adds SESSION_ATTRIBUTE related flags to the Path message when establishing a primary CR-LSP. For example, a local protection flag indicates a bypass CR-LSP that needs to be bound to the primary CR-LSP. A bandwidth protection flag indicates that bandwidth protection is required.

2. Binds a bypass CR-LSP to the primary CR-LSP.

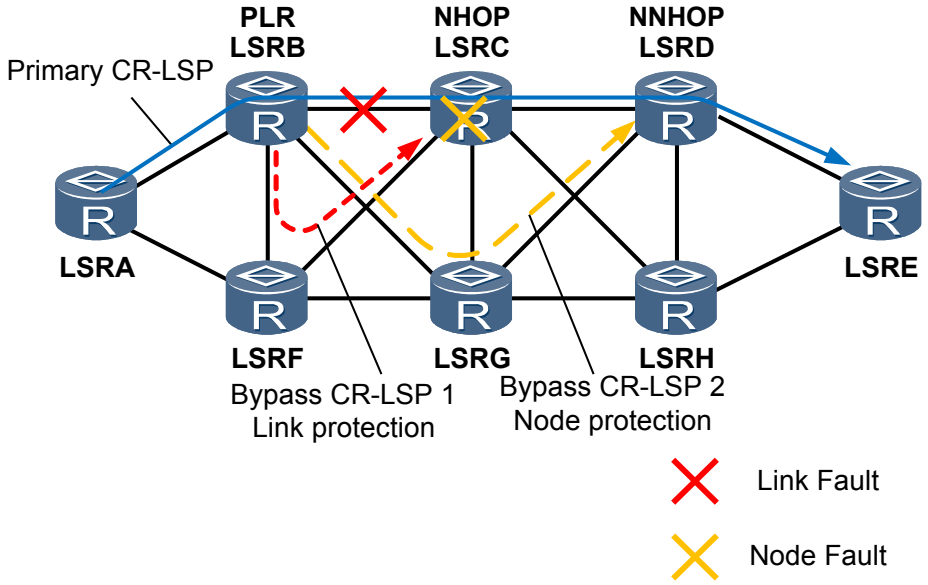
   Searching for a suitable bypass CR-LSP is also called bypass CR-LSP binding. This process is completed before a CR-LSP switchover is performed. A bypass CR-LSP can be bound to a primary CR-LSP only with the local protection flag.

   Before binding two CR-LSPs, a node must obtain the following information according to the RRO field in the Resv message. Some of them are listed as follow:

   - Outbound interface
   - Next Hop Label Forwarding Entry (NHLFE)
   - Label switching router (LSR) ID of the MP
   - Label allocated by the MP
   - Protection type

   The PLR node on a primary CR-LSP already obtains information about the next hop (NHOP) or next NHOP (NNHOP). If the egress LSR ID of the bypass CR-LSP is equal to the LSR ID of the NHOP, link protection is provided. If the egress LSR ID of the bypass CR-LSP is equal to the LSR ID of the NNHOP, node protection is provided. As shown in **Figure 3-24**, bypass CR-LSP 1 provides link protection and bypass CR-LSP 2 provides node protection.

Figure 3-24 Binding between bypass and primary CR-LSPs



If multiple bypass CR-LSPs are established, the PLR selects the one with the highest priority. The PLR prioritizes bypass CR-LSPs in the following order:

- Bandwidth protection
- Non-bandwidth protection
- Manual protection
- Auto FRR protection
- Node protection
- Link protection

Both bypass CR-LSPs 1 and 2 shown in **Figure 3-24** are manually configured and provide bandwidth protection. Bypass CR-LSP 1, which protects a link, has a lower priority than bypass CR-LSP 2, which protects a node. In such a scenario, bypass CR-LSP 2 is then bound to a primary CR-LSP. If bypass CR-LSP 1 only protects bandwidth and bypass CR-LSP 2 only protects a link, bypass CR-LSP 1 is then bound to the primary CR-LSP.

After the binding is complete, the primary CR-LSP NHLFE records the bypass CR-LSP NHLFE index and an inner label that the MP allocates for the primary CR-LSP. The label is used to forward traffic from the MP to the next hop along the primary CR-LSP.

3. Performs fault detection.

   - Link protection directly uses a data link layer protocol to detect and report faults. The speed of fault detection at the data link layer depends on the link type.
   - Node protection uses the link layer protocol to detect link faults. If no fault occurs on a link, the **RSVP Hello mechanism** is used to detect faults of the protected node or it is used with the **BFD for RSVP mechanism**.

   After a link or node fault is detected, FRR switching triggers immediately.

NOTE

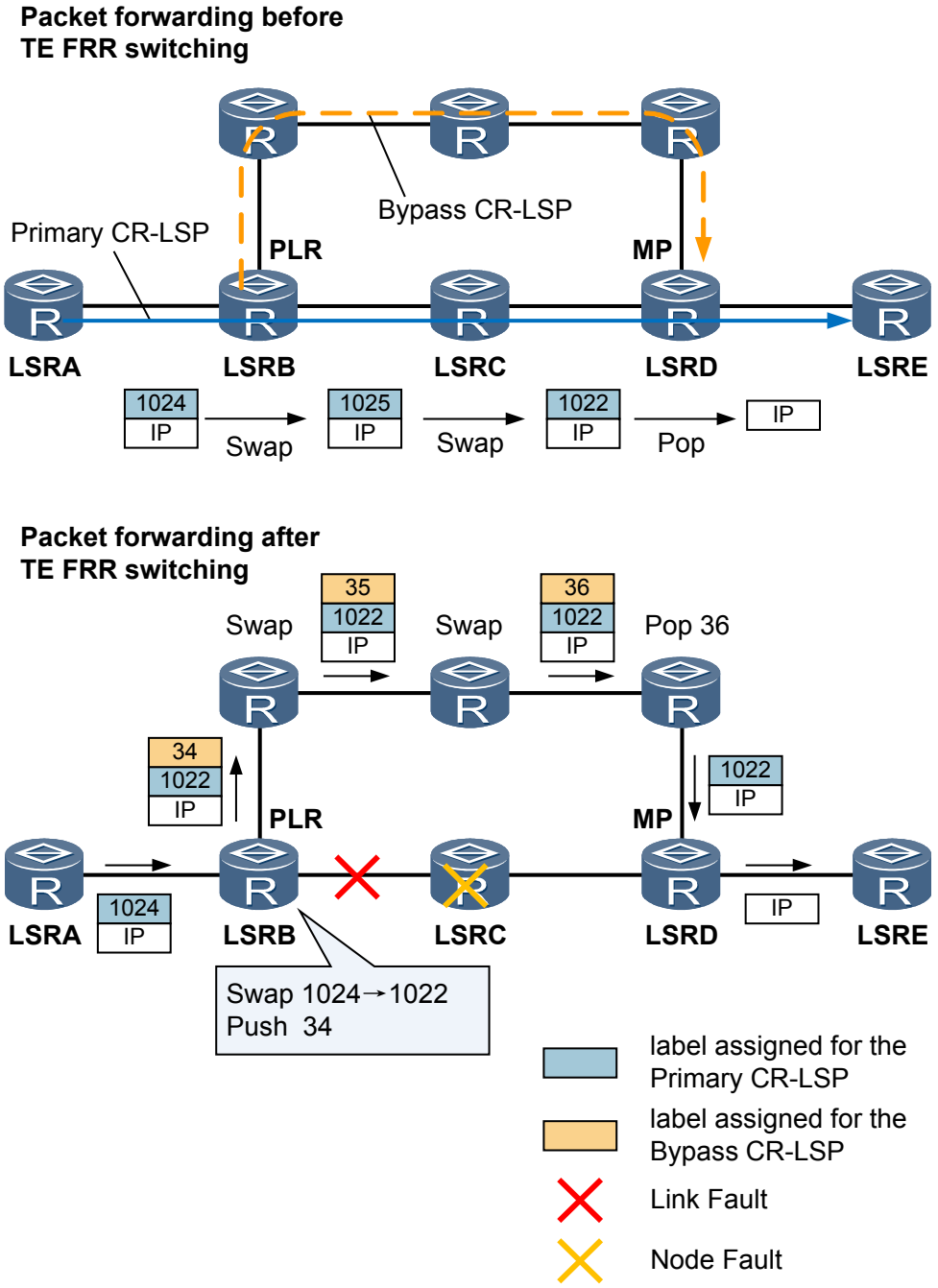
- If node protection is enabled, only the link between the protected node and PLR is protected. The PLR cannot detect faults in the link between the protected node and MP.
- Link fault detection, BFD detection, and RSVP Hello detection detect faults at a speed in descending order.

4. Performs a traffic switchover.

   If the primary CR-LSP fails, both data traffic and RSVP messages switch to the bypass CR-LSP, and the switchover event is reported upstream. The PLR pushes both an inner label that the MP assigns for the primary CR-LSP and an outer label assigned for the bypass CR-LSP into a packet. The outer label is removed at the penultimate hop of the bypass CR-LSP, and the packet, only with the inner label, arrives at the MP. The MP forwards the packet to the next hop along the primary CR-LSP.

   Figure 3-25 shows nodes on the primary and bypass CR-LSPs and their allocated labels and forwarding behaviors. The bypass CR-LSP provides node protection. If the link between LSRB and LSRC fails or LSRC fails, LSRB (PLR) swaps an inner label 1024 for an inner label 1022, pushes an outer label 34 into the packet, and forwards the packet over the bypass CR-LSP. After the packet arrives at LSRD, the LSRD forward the packet to the next hop LSRE. Packet forwarding after TE FRR switching in **Figure 3-25** shows the detailed forwarding process.

Figure 3-25 Schematic diagram of packet forwarding before and after TE FRR



5. Performs a traffic switchback.

   After TE FRR switching is complete, the PLR (ingress) attempts to reestablish the primary CR-LSP using the Make-Before-Break mechanism. Service traffic and RSVP messages switch from the bypass CR-LSP back to the primary CR-LSP after the primary CR-LSP is successfully reestablished. The reestablished CR-LSP is called a modified CR-LSP. The Make-Before-Break mechanism allows the original primary CR-LSP to be torn down only after the modified CR-LSP is set up successfully.

NOTE

FRR does not take effect if multiple nodes fail simultaneously. This means that after FRR switches data from the primary CR-LSP to the bypass CR-LSP, all nodes on the bypass CR-LSP must be working properly when transmitting data. If the bypass CR-LSP fails, the protected data cannot be forwarded, and the FRR function fails. Even if the bypass CR-LSP is reestablished, it cannot forward data. Data will be restored only after the primary CR-LSP is restored or reestablished.

Other Usage

- Board hot removal protection

  Board hot removal protection protects traffic on the primary CR-LSP's outbound interface on a PLR. If an interface board on which a protected outbound interface of a primary CR-LSP resides is removed from a PLR, the PLR rapidly switches traffic to a bypass CR-LSP. After the interface board is re-installed and the outbound interface of the primary CR-LSP becomes available, traffic switches back to the primary CR-LSP.

  Hot removal protection does not apply to an interface board, on which tunnel interfaces are configured. If an interface board configured with a tunnel interface is removed, CR-LSP information is lost and traffic is interrupted. The bypass CR-LSPs' tunnel interfaces and the bypass CR-LSP's outbound interface must be configured on boards different from the board configured with the bypass CR-LSP's outbound interface on the PLR.

  Configuring tunnel interfaces on the main control board of the PLR is recommended. If an interface board on which the primary CR-LSP's outbound interface is removed or fails, the primary CR-LSP's tunnel interface enters the Stale state, and resources allocated to the tunnel interface remain. After the interface board is re-installed, the tunnel interface recovers and a primary CR-LSP is reestablished.

- N:1 protection

  A single bypass CR-LSP can protect traffic over multiple primary CR-LSPs.

Coexistence of CR-LSP Backup and TE FRR

1. CR-LSP backup functions can be used together with TE FRR.
   - Ordinary backup and TE FRR: If TE FRR detects a link fault, traffic switches to a TE FRR bypass CR-LSP. If both the primary and TE FRR bypass CR-LSPs fail, an ordinary backup CR-LSP is established and takes over traffic.
   - Hot standby and TE FRR: If TE FRR detects a link fault, traffic switches to a TE FRR bypass CR-LSP and then to a hot-standby CR-LSP.
2. CR-LSP backup can be associated with TE FRR.

   The association improves tunnel security. The association provides the following functions based on backup modes:

   - Association between an ordinary backup CR-LSP and a TE FRR bypass CR-LSP provides the following functions:

     If a protected link or node fails, traffic switches to a bypass CR-LSP. The ingress attempts to reestablish the primary CR-LSP, while attempting to establish an ordinary backup CR-LSP.

     If the ordinary backup CR-LSP is established successfully before the primary CR-LSP is restored, traffic switches to the ordinary backup CR-LSP.

     After the primary CR-LSP recovers, traffic switches back to the primary CR-LSP.

     If the ordinary backup CR-LSP fails to be established and the primary CR-LSP does not recover, traffic passes still through the bypass CR-LSP.

- Association between a hot-standby CR-LSP and a TE FRR bypass CR-LSP provides the following functions:

  If a hot-standby CR-LSP is Up and a protected link or node fails, traffic switches to a TE FRR bypass CR-LSP and then immediately switches to the hot-standby CR-LSP. At the same time, the ingress attempts to restore the primary CR-LSP.

  If the hot-standby CR-LSP is Down, the traffic switching procedure is the same as that when the ordinary backup is used.

Association between ordinary backup CR-LSPs and TE FRR is recommended. An ordinary backup CR-LSP without additional bandwidth needed is established only after the primary CR-LSP enters the FRR-in-use state. Although the primary CR-LSP is Up, the system attempts to establish a hot-standby CR-LSP with additional bandwidth needed.

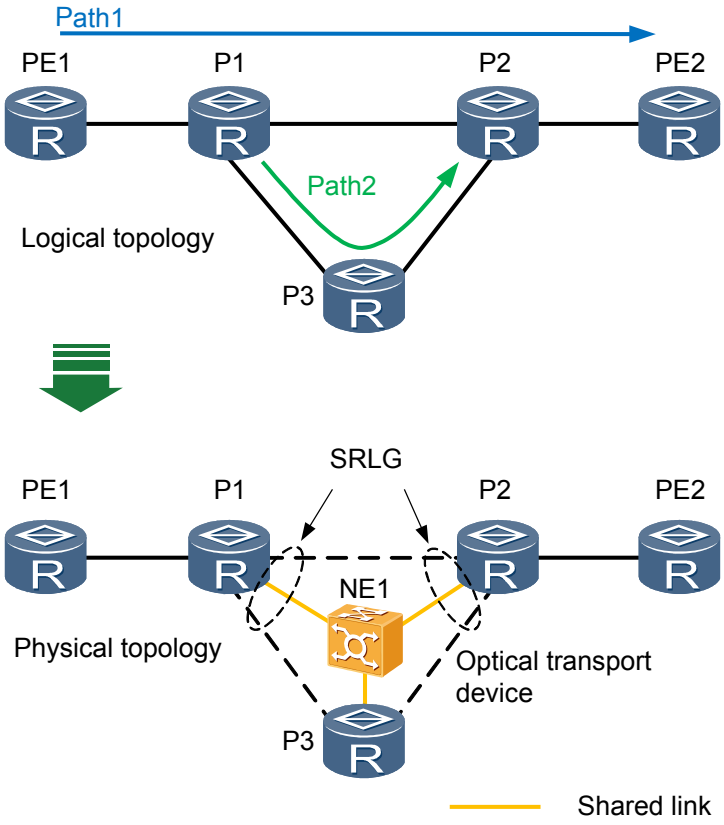
3.2.9.6 SRLG

The shared risk link group (SRLG) functions as a constraint that is used to calculate a backup path in the scenario where CR-LSP hot standby or TE FRR is used. This constraint helps prevent backup and primary paths from overlapping over links with the same risk level, improving MPLS TE tunnel reliability as a consequence.

Background

Network administrators use CR-LSP hot standby or TE FRR to improve MPLS TE tunnel reliability. However, in real-world situations protection failures can occur, requiring the SRLG technique to be configured as a preventative measure, as the following example demonstrates.

Figure 3-26 Networking diagram for an SRLG



The primary CR-LSP is established over the path Path1 on the network shown in **Figure 3-26**. The link between P1 and P2 is protected by a TE FRR bypass CR-LSP established over the path Path2.

In the lower part of **Figure 3-26**, core nodes P1, P2, and P3 are connected using a transport network device. They share some transport network links marked in yellow. If a fault occurs on a shared link, both the primary and bypass CR-LSPs are affected, causing an FRR protection failure. An SRLG can be configured to prevent the bypass CR-LSP from sharing a link with the primary CR-LSP, ensuring that FRR properly protects the primary CR-LSP.

An SRLG is a set of links at the same risk of faults. If a link in an SRLG fails, other links also fail. If a link in this group is used by a hot-standby CR-LSP or bypass CR-LSP, the hot-standby CR-LSP or bypass CR-LSP cannot provide protection.

Implementation

An SRLG link attribute is a number and links with the same SRLG number are in a single SRLG.

Interior Gateway Protocol (IGP) TE advertises SRLG information to all nodes in a single MPLS TE domain. The constraint shortest path first (CSPF) algorithm uses the SRLG attribute together with other constrains, such as bandwidth, to calculate a path.

The MPLS TE SRLG works in either of the following modes:

- Strict mode: The SRLG attribute is a necessary constraint used by CSPF to calculate a path for a hot-standby CR-LSP or an bypass CR-LSP.
- Preferred mode: The SRLG attribute is an optional constraint used by CSPF to calculate a path for a hot-standby CR-LSP or bypass CR-LSP. For example, if CSPF fails to calculate a path for a hot-standby CR-LSP based on the SRLG attribute, CSPF recalculates the path, regardless of the SRLG attribute.

Usage Scenario

The SRLG attribute is used in either the TE FRR or CR-LSP hot-standby scenario.

Benefits

The SRLG attribute limits the selection of a path for a hot-standby CR-LSP or bypass CR-LSP, which prevents the primary and bypass CR-LSPs from sharing links with the same risk level.

3.2.9.7 TE Tunnel Protection Group

A tunnel protection group protects E2E MPLS TE tunnels. If a working tunnel in a protection group fails, traffic switches to a protection tunnel, minimizing traffic interruptions.

Related Concepts

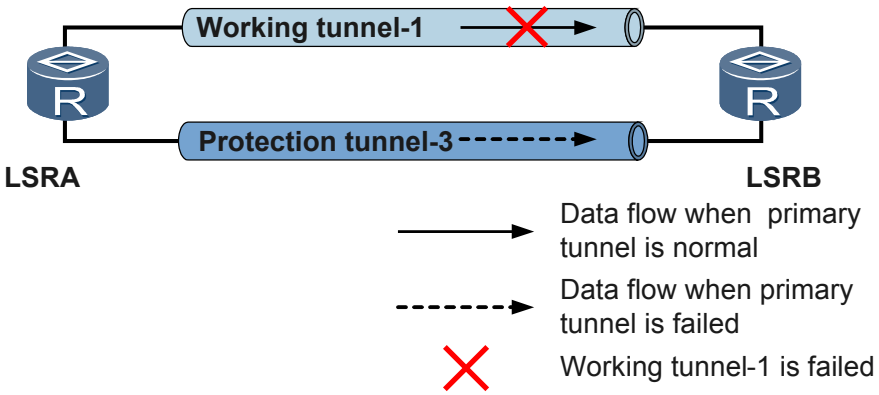
Concepts related to a tunnel protection group are as follows:

- Working tunnel: a tunnel to be protected.
- Protection tunnel: a tunnel that protects a working tunnel.

- Protection switchover: switches traffic from a faulty working tunnel to a protection tunnel in a tunnel protection group, which improves network reliability.

Figure 3-27 shows a tunnel protection group.

Figure 3-27 Tunnel protection group



Primary tunnels tunnel-1, and the bypass tunnel tunnel-3 are established on the ingress LSRA shown in Figure 3-27. Tunnel-3 is specified as a protection tunnel for primary CR-LSPs tunnel-1 on LSRA. If the configured fault detection mechanism on the ingress detects a fault in tunnel-1, traffic switches to tunnel-3. LSRA attempts to reestablish tunnel-1. If tunnel-1 is successfully established, traffic switches back to the primary CR-LSP.

Implementation

A TE tunnel protection group uses a configured protection tunnel to protect traffic on the working tunnel to improve tunnel reliability. To ensure the improved performance of the protection tunnel, the protection tunnel must exclude links and nodes through which the working tunnel passes during network planning.

Table 3-18 describes the implementation procedure of a tunnel protection group.

Table 3-18 Implementation procedure of a tunnel protection group

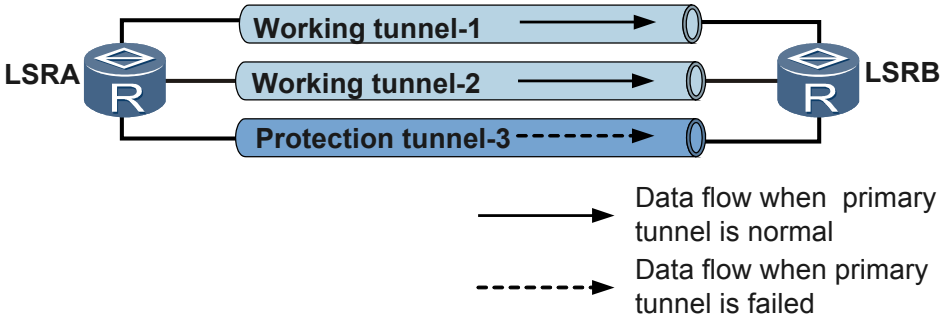
Process	Description
Establishment	The working and protection tunnels must have the same ingress and egress. The protection tunnel is established in the same procedure as a regular tunnel. The protection tunnel can use attributes that differ from those for the working tunnel. Ensure that the working and protection tunnels are established over different paths as much as possible. **NOTE** ● A protection tunnel cannot be protected or enabled with TE FRR. ● Attributes for a protection tunnel can be configured independently of those for the working tunnel, which facilitates the network planning.

Process	Description
Binding between the working and protection tunnels	The protection tunnel is bound to the tunnel ID of the working tunnel so that the two tunnels form a tunnel protection group.
Fault detection	In addition to MPLS TE's own detection mechanism, **MPLS OAM** and **BFD for CR-LSP** are used to detect faults in a tunnel protection group to speed up protection switching.
Protection switching	The tunnel protection group supports either of the following protection switching modes: ● Manual switching: Traffic is forcibly switched to the protection tunnel. ● Automatic switching: Traffic automatically switches to the protection tunnel if the working tunnel fails. A time interval can be set for automatic switching.
Switchback	After a traffic switchover is implemented, the ingress attempts to reestablish the working tunnel. If the working tunnel is reestablished, the ingress can switch traffic back to the working tunnel or still forward traffic over the protection tunnel.

Other Usage

A tunnel protection group works in either 1:1 or N:1 mode. The 1:1 mode enables a protection tunnel to protect only a single working tunnel. The N:1 mode enables a protection tunnel to protect more than one working tunnel.

Figure 3-28 N:1 protection mode



Differences Between CR-LSP Backup and a Tunnel Protection Group

CR-LSP backup and a tunnel protection group are both E2E protection mechanisms for MPLS TE. **Table 3-19** shows the comparison between these two mechanisms.

Table 3-19 Comparison between CR-LSP backup and a tunnel protection group

Item	CR-LSP Backup	Tunnel Protection Group
Object to be protected	Primary and backup CR-LSPs are established on the same tunnel interface. A backup CR-LSP protects traffic on a primary CR-LSP.	One tunnel protects traffic over another tunnel in a tunnel protection group.
TE FRR	A primary CR-LSP supports TE FRR. A backup CR-LSP does not support TE FRR.	A working tunnel supports TE FRR. A protection tunnel does not support TE FRR.
LSP attributes	Primary and backup CR-LSPs have the same attributes, except for the TE FRR attribute. In addition, the bandwidth for the backup CR-LSP can be set separately.	The attributes of one tunnel in a tunnel protection group are independent of the attributes of the other tunnel. For example, a protection tunnel with no bandwidth can protect traffic on a working tunnel that has a bandwidth.
Protection mode	The 1:1 protection mode is supported. Each primary CR-LSP is protected by a backup CR-LSP.	Apart from the 1:1 protection mode, the N:1 protection mode is supported. Multiple working tunnels share one protection tunnel. When any one working tunnel fails, data is switched to the protection tunnel.

3.2.9.8 BFD for MPLS TE

Bidirectional forwarding detection (BFD) can monitor MPLS TE tunnels, CR-LSPs bound to the MPLS TE tunnels, and RSVP neighbor relationships. If BFD detects a fault, the BFD module instructs the MPLS module to perform a traffic switchover, improving network reliability.

Background

TE FRR, CR-LSP backup, and tunnel protection groups can be used to improve the reliability of MPLS TE networks. A fault, however, occurs if no message arrives after the refresh period of RSVP Hello or RSVP messages, which leads to a slow detection speed. When a Layer 2 device (such as a switch or hub) exists on the faulty link, slow detection delays a traffic switchover and causes some traffic to be dropped. BFD can send packets to quickly detect faults in MPLS TE tunnels and trigger a rapid traffic switchover to minimize traffic loss.

Related Concepts

BFD sessions are classified into the following types:

- Static BFD session: Local and remote discriminators are configured manually.
- Dynamic BFD session: Local and remote discriminators are allocated automatically.

NOTE

For details about BFD, see the chapter "BFD" in the *Feature Description - Reliability*.

Implementation

The following BFD functions are supported for MPLS TE:

- **BFD for RSVP**

  BFD monitors RSVP. BFD can detect faults in links between RSVP neighboring nodes in milliseconds. BFD for RSVP applies to a TE FRR network, on which Layer 2 devices exist between the PLR and its RSVP neighboring nodes over the primary CR-LSP.

- **BFD for CR-LSP**

  BFD monitors CR-LSPs. After BFD detects a fault in a CR-LSP, the BFD module immediately instructs the forwarding plane to trigger a rapid traffic switchover. BFD for CR-LSP is used together with a hot-standby CR-LSP or a tunnel protection group.

- **BFD for TE tunnel**

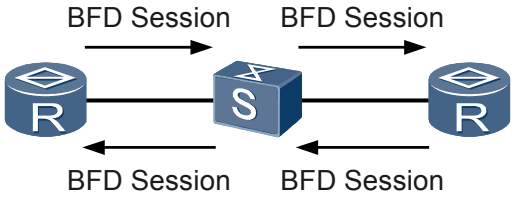
  BFD can monitor MPLS TE tunnels that are used as public network tunnels to transmit VPN traffic. BFD monitors a whole TE tunnel. If BFD detects a fault in a tunnel that transmits private network traffic, the BFD module instructs the VPN or virtual leased line (VLL) FRR module to perform a traffic switchover.

BFD for RSVP

When a Layer 2 device exists between RSVP neighboring nodes, the two nodes can detect a link fault only using the Hello mechanism in seconds. This process results in the loss of lots of data.

BFD monitors RSVP neighbor relationships. BFD for RSVP rapidly detects faults in a link between RSVP neighboring nodes within milliseconds. BFD for RSVP applies to TE FRR networks, on which Layer 2 devices exist on a primary CR-LSP between the PLR and its RSVP neighboring node, as shown in **Figure 3-29**.

Figure 3-29 BFD for RSVP



BFD for RSVP can share BFD sessions with BFD for Open Shortest Path First (OSPF), BFD for Intermediate System to Intermediate System (IS-IS), or BFD for Border Gateway Protocol (BGP). The local node selects the smallest values of parameters between the two ends of the shared BFD session as local BFD parameters. The parameters include the interval at which BFD packets are sent, interval at which BFD packets are received, and local detection multiplier.

BFD for CR-LSP

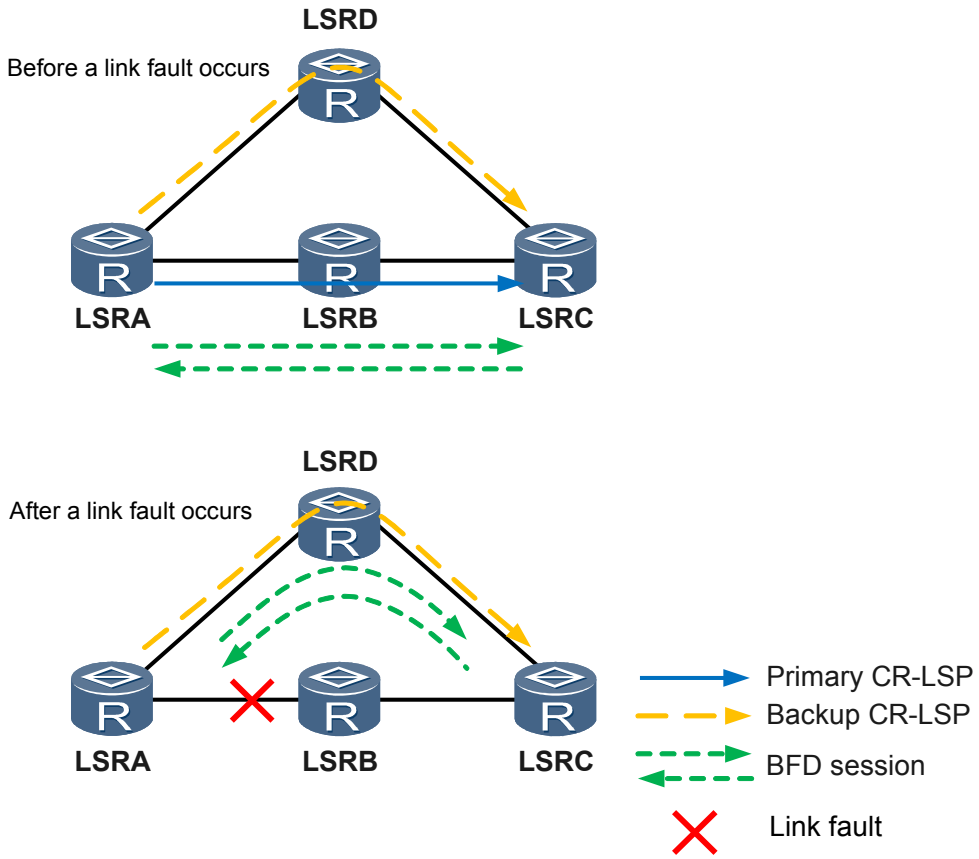
BFD monitors CR-LSPs. After BFD detects a fault in a CR-LSP, the BFD module immediately instructs the forwarding plane to trigger a rapid traffic switchover. BFD for CR-LSP is used together with a hot-standby CR-LSP or a tunnel protection group.

A BFD session is bound to a CR-LSP. This means that a BFD session is set up between the ingress and egress. A BFD packet is sent by the ingress to the egress along a CR-LSP. After the

egress receives the packet, the egress responds to the BFD packet. The ingress can rapidly detect the status of links through which the CR-LSP passes based on whether a reply packet is received.

If a link fault is detected, the BFD module notifies the forwarding plane of the fault. The forwarding plane searches for a backup CR-LSP and switches traffic to the backup CR-LSP. In addition, the forwarding plane reports the fault to the control plane. If dynamic BFD for CR-LSP is used, the control plane proactively creates a BFD session to monitor the backup CR-LSP. A static BFD session can also be used to monitor the backup CR-LSP.

Figure 3-30 Traffic forwarding of a BFD session before and after a traffic switchover



BFD for TE Tunnel

BFD for TE tunnel uses BFD sessions to detect the entire TE tunnel to trigger traffic switchover of the applications such as VPN FRR.

BFD for TE tunnel and BFD for CR-LSP send fault information to different objects. BFD for TE tunnel notifies applications (VPN for example) of faults and triggers a traffic switchover between different TE tunnel interfaces. BFD for CR-LSPs notifies a TE tunnel of faults and triggers a traffic switchover between different CR-LSPs in the same TE tunnel.

Differences

Table 3-20 lists differences between BFD for RSVP, BFD for CR-LSP, and BFD for TE tunnel.

Table 3-20 Differences between BFD for CR-LSP, BFD for RSVP, and BFD for TE tunnel

Detection Technique	Detection Object	Node	Usage Scenario	BFD Session Support
BFD for RSVP	RSVP neighbor relationships	Two ends of an RSVP session	Can be used with TE FRR	Dynamic
BFD for CR-LSP	CR-LSPs	Ingress and egress	Can be used with a hot-backup CR-LSP	● Dynamic ● Static
BFD for TE tunnel	MPLS TE tunnels	Ingress and egress	Can be used with VPN FRR	Static

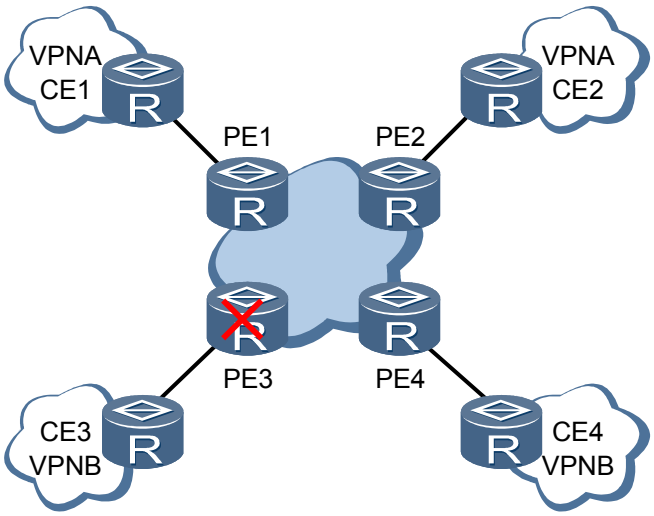
3.2.9.9 RSVP GR

RSVP graceful restart (GR) ensures uninterrupted transmission on the forwarding plane while an active main board (AMB)/standby main board (SMB) switchover is performed on the control plane.

Background

GR applies to provider edge (PE) routers on the provider network shown in **Figure 3-31**. User nodes access the provider network through only a single PE. MPLS TE tunnels are established between PEs on the network to implement TE or transmit VPN traffic. If a PE fails or a maintenance measure (such as a software upgrade) is taken, an AMB/SMB switchover is performed on the PE. To prevent traffic loss during a traffic switchover, RSVP GR can be implemented to ensure the uninterrupted transmission of critical services.

Figure 3-31 RSVP GR application scenario



Related Concepts

RSVP GR is a rapid status restoration mechanism for RSVP-TE that is implemented based on non-stop forwarding (NSF).

In the RSVP GR process, two roles are defined based on their functions. GR restarter performs a graceful restart; while GR helper assists the GR restarter in implementing a graceful restart.

RSVP GR supports the following messages:

- Hello message carrying a GR extension: This message is used to detect the GR status of a neighboring node.
- GR Path message: This message is sent by an upstream node and carries the contents of the latest refreshed Path message.
- Recovery Path message: This message is sent by a downstream node and carries the contents of the last Path message that is received by the downstream node.

Implementation

RSVP GR uses the Hello extension to monitor the GR status of neighboring nodes.

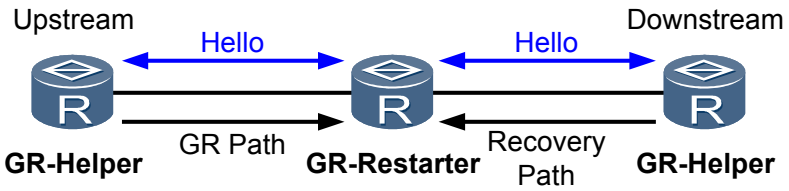
The principles of RSVP GR are as follows:

When the GR restarter performs GR on the network shown in **Figure 3-32**, it stops sending Hello messages to its neighboring nodes. If a GR-enabled neighboring node (GR-Helper) fails to receive consecutive three Hello messages, the neighbor node considers that the GR restarter is performing GR and retains all forwarding information. In addition, an interface board on the neighboring node continues transmitting services and waits for the GR restarter to restore the GR status.

After the GR restarter is restarted and receives a Hello message from its neighboring node, it replies with a Hello message. The processing modes of Hello messages are different on an upstream node and a downstream node.

- If an upstream GR helper receives a Hello message, it sends a GR Path message to the GR restarter.
- If a downstream GR helper receives a Hello message, it sends a Recovery Path message to the GR restarter.

Figure 3-32 Schematic diagram of RSVP GR



When the restarter receives GR Path and Recovery Path messages, the path state block (PSB) on the GR restarter restores CR-LSP status information on the local control plane is restored.

If a downstream GR helper cannot send a Recovery Path message, the local PSB can restore CR-LSP status information using only a GR Path message.

Deployment Scenarios

RSVP GR can be used on nodes that runs RSVP-TE to establish MPLS TE tunnels to improve device reliability.

Benefits

RSVP GR ensures uninterrupted data service transmission when the control plane performs an AMB/SMB switchover and supports device-level reliability for MPLS TE nodes.

3.2.10 DS-TE

3.2.10.1 Background

Traditional MPLS TE reserves resources for each node along the MPLS TE tunnel to ensure QoS, but cannot use one TE tunnel to provide differentiated services. When a tunnel transmits voice and data services, data services may be transmitted repeatedly. Therefore, data services must have higher drop priority than voice services. However, MPLS TE allocates the same drop priority to data and voice flows and cannot provide differentiated services.

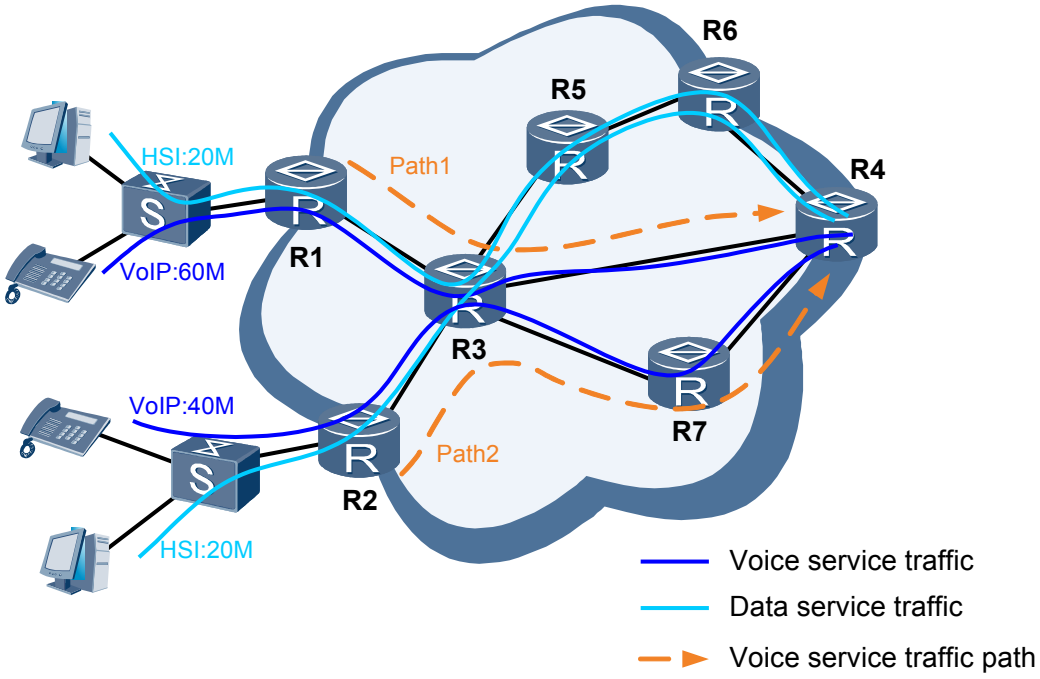
The Diff-Serv model controls and forwards traffic according to specific service classes, meeting different QoS requirements. The Diff-Serv model can reserve resources for a single node, but cannot guarantee the QoS over an entire path.

In certain scenarios, Diff-Serv and MPLS TE must be used together to meet service requirements. For example, a path may trasnmit both voice and data services. The total delay time of voice flows needs to be reduced to ensure QoS guarantee of voice services.

Assume that the Diff-Serv model is used to classify service types and a single MPLS TE tunnel is used to transmit a type of service. If the link or node becomes faulty, network topology changes, or LSP preemption occurs, voice traffic on a link may exceed the bandwidth and voice services are delayed.

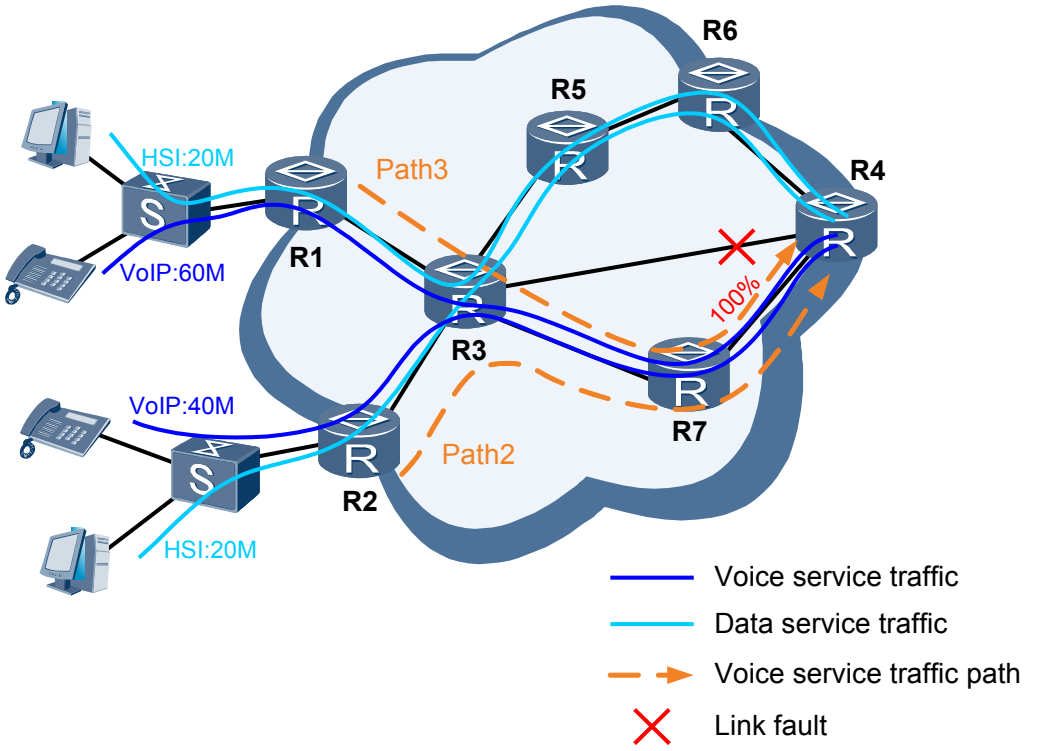
For example, in **Figure 3-33**, the bandwidth of each link is 100 Mbit/s and the link cost is the same. Voice flows pass through the links R1 -> R4 and R2 -> R4, and bandwidths of the links are 60 Mbit/s and 40 Mbit/s respectively. When voice flows on the link R1 -> R4 are transmitted through the TE tunnel of Path1, voice flows occupy 60% bandwidth on the link R3 -> R4. When voice flows on the link R2 -> R4 are transmitted through the TE tunnel of Path2, voice flows occupy 40% bandwidth on the link R7 -> R4.

Figure 3-33 Networking where MPLS TE and DiffServ are used



As shown in **Figure 3-34**, when the link between R3 and R4 becomes faulty, the CR-LSP between R1 and R4 is changed to Path3. The reason is that Path3 is the shortest path with sufficient bandwidth. In this case, voice flows on the link R7 -> R4 occupy 100% bandwidth, causing a long delay in transmitting voice flows.

Figure 3-34 Link failure



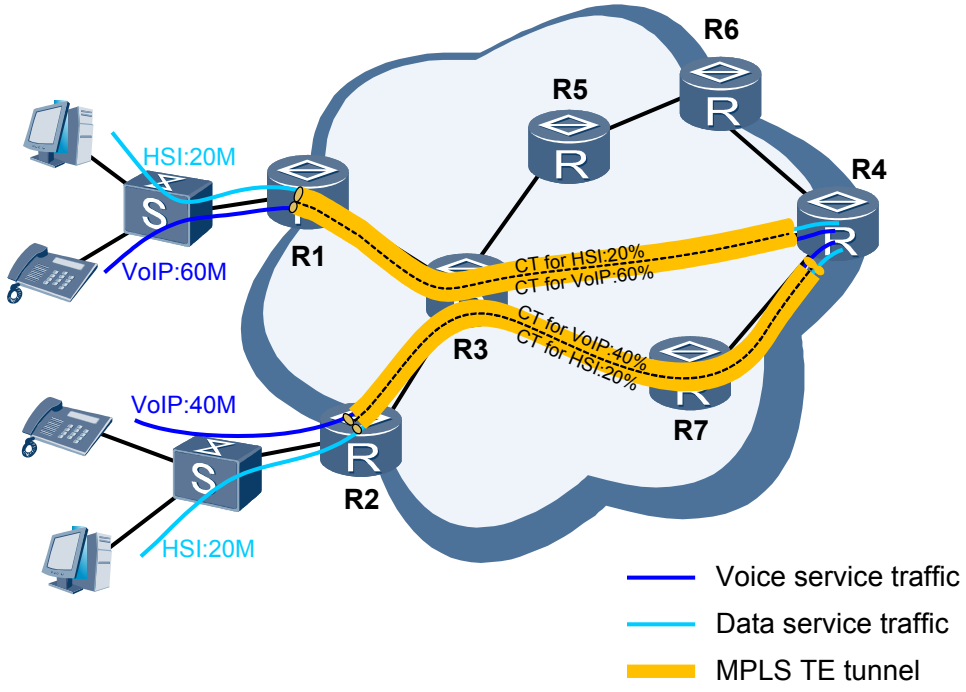
To solve the preceding problem, DiffServ-aware Traffic Engineering (DS-TE) is used. DS-TE can efficiently use network resources and reserve resources for different service flows.

MPLS DS-TE combines MPLS TE and Diff-Serv to provide QoS guarantee.

MPLS DS-TE uses the **Class Type (CT)** so that MPLS TE can allocate resources based on the type of traffic and provide differentiated services. To provide differentiated services, DS-TE divides the LSP bandwidth into one to eight parts, each part corresponding to one Class of Service (CoS). A set of bandwidth of an LSP or a group of LSPs with the same CoS are called a CT.

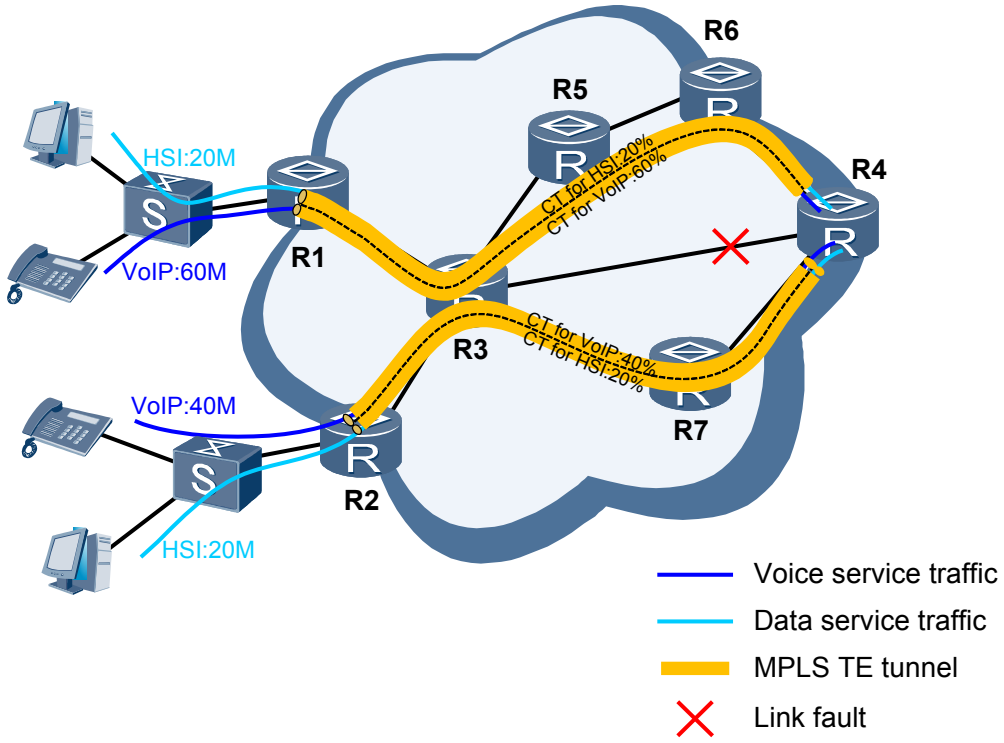
In **Figure 3-33**, multiple CT LSPs can be used. An LSP is divided into multiple CTs to transmit traffic of different CoS values. VoIP and HSI services on the links R1 -> R4 and R2 -> R4 are transmitted by different CTs of the same MPLS TE tunnel so that voice flows occupy a proper bandwidth percentage, as shown in **Figure 3-35**.

Figure 3-35 MPLS DS-TE



When the link R3 -> R4 becomes faulty, VoIP and HSI services on the link R3 -> R4 are switched to the link R1 -> R3 -> R5 -> R6 -> R4, as shown in **Figure 3-36**. After switching, voice flows on the link R1 -> R4 still occupy a proper bandwidth percentage.

Figure 3-36 Traffic switching after a link failure



3.2.10.2 Basic Concepts

DS Field

To carry out the Diff-Serv model, RFC 2474 redefines the ToS field in the IPv4 packet header as the Differentiated Services (DS) field. The high-order 2 bits in the DS field are reserved, and the low-order 6 bits specify the DS CodePoint (DSCP).

Per Hop Behavior

Per Hop Behavior (PHB) describes how the packets with the same DSCP value are forwarded to the next hop.

The IETF defines three standardized PHBs: expedited forwarding (EF), assured forwarding (AF), and best-effort (BE). BE is the default PHB.

CT

To carry out differentiated services, the DS-TE model divides the bandwidth of an LSP into one to eight parts. Each part of bandwidth is allocated with a different service class. The set of bandwidth of one LSP or a group of LSPs with the same service class is called a class type (CT). One CT can transmit the traffic of a single service type.

As defined in the IETF, the DS-TE supports a maximum of eight CTs. CTs can be represented as CTi. The value of "i" ranges from 0 to 7.

IGP Extension

To support DS-TE. RFC 4124 uses a Bandwidth Constraints Sub-TLV into the Interior Gateway Protocol (IGP) and redefines the Unreserved Bandwidth Sub-TLV. These Sub-TLVs are used to collect and advertise information about the reservable bandwidth for each CT along a link. For details, see RFC 4124.

Single-CT LSP and Multi-CT LSP

A single-CT LSP transmits traffic of only one CT.

A multi-CT LSP transmits traffic of multiple CTs.

For the multi-CT, the resource reservation, LSP establishment, or bandwidth preemption can be successfully performed only when the bandwidth of all the CTs is sufficient.

LSP Preemption and TE-Class Mapping

If no path meets the bandwidth requirement of a desired CR-LSP, a device can tear down an established CR-LSP and use the bandwidth assigned to that CR-LSP to establish a desired CR-LSP. This process is called preemption. DS-TE uses setup and holding priorities to determine whether to preempt resources.

DS-TE specifies a preemption priority for each CT. A TE-class is a combination of a CT and a priority, which is described as follows:

TE-Class[n] = <CTi, priority>

i and the priority value range from 0 to 7, and *n* identifies a TE class.

The priority indicates the priority of CR-LSP preemption, and is not the value of the EXP field in the MPLS packet header. The value of preemption priority ranges from 0 to 7. A smaller value indicates a higher priority. A CR-LSP can be set up only when both the combination of its CT and setup priority (<CT, setup-priority>) and the combination of its CT and holding priority (<CT, hold-priority>) exist in the TE-class mapping table. For example, the TE-class mapping table of a certain node contains only TE-class[0] = <CT0, 6> and TE-class[1] = <CT0, 7>.

Only the following types of CR-LSPs can be set up successfully:

- Class-Type = CT0, setup-priority = 6, hold-priority = 6
- Class-Type = CT0, setup-priority = 7, hold-priority = 6
- Class-Type = CT0, setup-priority = 7, hold-priority = 7

NOTE

The CR-LSPs of "Class-Type = CT0, setup-priority = 6, hold-priority = 7" cannot be configured. This is because the setup priority of the CR-LSP cannot be higher than its holding priority.

Each of eight CTs can be combined with any of eight priorities, so there are 64 TE-classes. On the device, eight TE-classes can be configured manually.

A TE-class mapping table consists of a set of TE-classes. You are advised to configure all the LSRs with the same TE-class mapping table over an MPLS network. The device has the default TE-class mapping table.

Table 3-21 Default TE-class mapping table

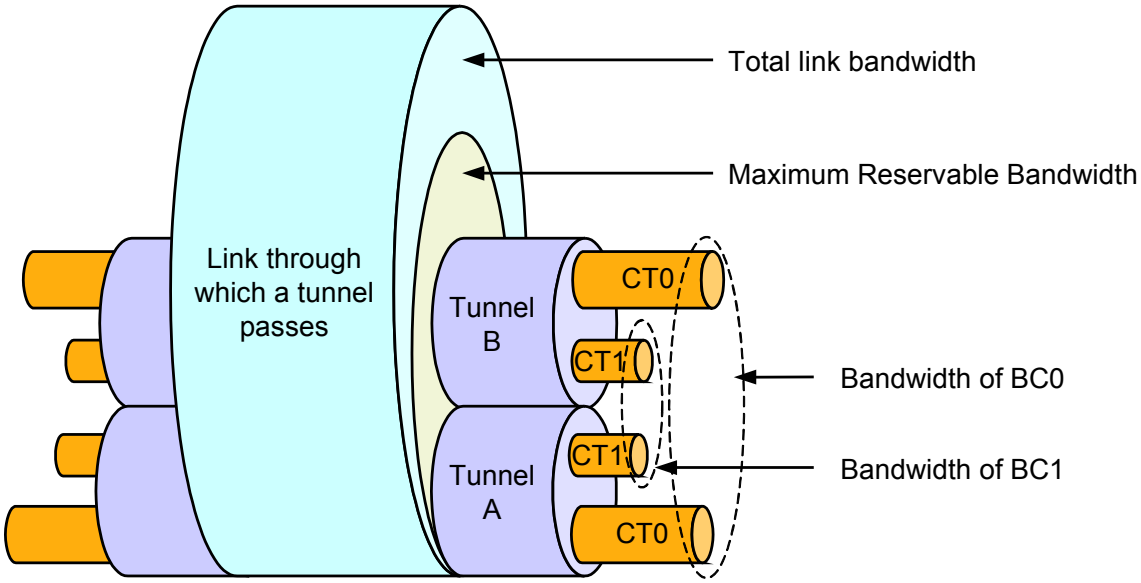
TE-Class	CT	Priority
TE-Class[0]	0	0
TE-Class[1]	1	0
TE-Class[2]	2	0
TE-Class[3]	3	0
TE-Class[4]	0	7
TE-Class[5]	1	7
TE-Class[6]	2	7
TE-Class[7]	3	7

Bandwidth

MPLS DS-TE involves the following types of bandwidth:

- Total link bandwidth: is physical link bandwidth.
- Maximum reservable bandwidth: is the maximum bandwidth that a link can reserve for an MPLS TE tunnel to be established. The maximum reservable bandwidth must be lower than or equal to the total link bandwidth.
- CT bandwidth: is the bandwidth of service traffic of each type on each DS-TE tunnel.
- BC bandwidth: is the bandwidth reserved for all CTs along a link.

Figure 3-37 Relationship between bandwidths



Bandwidth Constraints Model

Bandwidth constraint model defines the maximum number of bandwidth constraints and which CTs each bandwidth constraint applies to and how to use BC bandwidth.

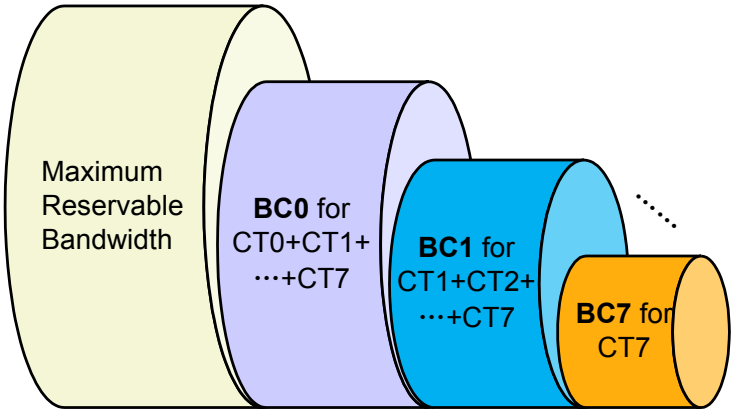
The IETF defines the following bandwidth constraints models:

- Russian Dolls Model (RDM): CTs can share bandwidth. The BC model ID of the RDM is 0.

  The bandwidth of BC0 is less than or equal to the maximum reservable bandwidth of a link. In **Figure 3-38**:

  - Total bandwidth of all LSPs from CT0, CT1, ... CT7 ≤ Bandwidth of BC0 ≤ Maximum reservable bandwidth
  - Total bandwidth of all LSPs from CT1, CT2, and CT7 ≤ Bandwidth of BC1
  - ...
  - Total bandwidth of all LSPs from CT7 ≤ Bandwidth of BC7

  Figure 3-38 RDM

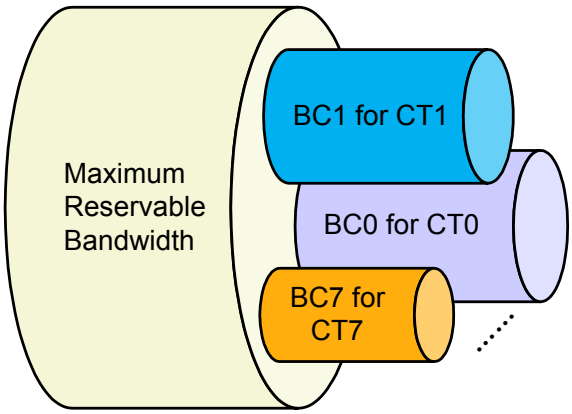
  

  For example, the bandwidth of a link is 100 Mbit/s, RDM is used, and three CTs are supported, that is, CT0, CT1, and CT2. CT0, CT1, and CT2 transmit BE, AF, and EF traffic respectively. The bandwidths of BC0, BC1, and BC2 are 100 Mbit/s, 50 Mbit/s, and 20 Mbit/s respectively. The total bandwidth of all LSPs transmitting EF traffic cannot be larger than 20 Mbit/s; the total bandwidth of all LSPs transmitting AF and EF traffic cannot be larger than 50 Mbit/s; the total bandwidth of all LSPs cannot be larger than 100 Mbit/s.

  The RDM allows bandwidth preemption between CTs. If $0 \leqslant m < n \leqslant 7$ and $0 \leqslant i < j \leqslant 7$, the CTi of priority *m* can preempt the bandwidth of CTi of priority *n* and the bandwidth of CTj of priority n. For example, CT0 with priority 3 can preempt the bandwidth of CT0 with priority 5 and the bandwidth of CT1 with priority 0. The total bandwidth of CTi of all LSPs cannot exceed the bandwidth of BCi.

- Maximum Allocation Model (MAM): One BC is mapped to one CT, and CTs cannot share bandwidth. The BC mode ID of the MAM is 1.

Figure 3-39 MAM



In the MAM, the total bandwidth of CTi along an LSP cannot be larger than that of BCi (0 ≤ i ≤ 7). The total bandwidth of BCs cannot be larger than the maximum reservable bandwidth.

For example, the bandwidth of a link is 100 Mbit/s, MAM is used, and three CTs are supported, that is, CT0, CT1, and CT2. BC0 is 20 Mbit/s and transmits CT0 traffic (for example, BE traffic); BC1 is 50 Mbit/s and transmits CT1 traffic (for example, AF traffic); BC2 is 30 Mbit/s and transmits CT2 traffic (for example, EF traffic). The total bandwidth of all LSPs transmitting BE traffic cannot be larger than 20 Mbit/s; the total bandwidth of all LSPs transmitting AF traffic cannot be larger than 50 Mbit/s; the total bandwidth of all LSPs transmitting EF traffic cannot be larger than 30 Mbit/s.

- Extended-MAM: Similar to the MAM, extended-MAM maps one BC to one CT and CTS cannot share the bandwidth. The BC mode ID of the extended-MAM is 254.

  The extended-MAM supports eight more implicit CTs (the combination of CT0 and eight priorities). This is different from the MAM.

  Extended-MAM redefines Unreserved Bandwidth Sub-TLV and Bandwidth Constraint Sub-TLV advertised by an IGP. Unreserved Bandwidth Sub-TLV carries unreserved bandwidth for eight TE-classes. Bandwidth Constraint Sub-TLV carries information about the BC model and unreserved bandwidth for eight TE-classes so that the device supports a maximum of 16 TE-classes.

  When device A that has Extended-MAM configured functions as the transit or egress node, device B configured with DS-TE in non-IETF mode functions as the ingress node and creates a dynamic CR-LSP, and the TE-class (<CT0, priority>, 0 ≤ priority ≤ 7) of the CR-LSP is not defined in the TE-class mapping table specified on device A, the CR-LSP creation request is valid.

Table 3-22 lists the comparisons between the three bandwidth constraints models.

Table 3-22 Comparisons between the three bandwidth constraints models

Item	RDM	MAM/Extended-MAM
BC-CT mapping	Maps one BC to one or more CTs.	Maps one BC to one CT, which is easy for bandwidth management.

Item	RDM	MAM/Extended-MAM
Bandwidth preemption	Is unable to divide CT bandwidth and requires preemption to provide sufficient bandwidth for CTs.	Divides CT bandwidth and provides sufficient bandwidth for CTs.
Bandwidth use efficiency	Efficiently uses bandwidth.	Wastes bandwidth.

3.2.10.3 Implementation

Basic Implementation

Edge nodes in the Diff-Serv model divide the traffic into several classes, and add class information into the DSCP field in packets. The internal node selects a proper PHB for a packet according to the DSCP field.

The EXP field in the MPLS packet header contains information relavant to the Diff-Serv model. The key to implement DS-TE is how to map the DSCP field (with a maximum of 64 values) to the EXP field (with a maximum of eight values). RFC 3270 defines the following solutions:

- Label-Only-Inferred-PSC LSP (L-LSP): The drop priority is specified in the EXP field and the PHB is determined by the label value. During packet forwarding, the label determines the packet forwarding path and allocates a PHB for the path.
- EXP-Inferred-PSC LSP (E-LSP): The PHB and the drop priority are specified in the EXP field of the MPLS label. During packet forwarding, the label value determines the packet forwarding path and the EXP value determines a PHB. The E-LSP is applicable to networks that supports a maximum of eight PHBs.

The device implements the E-LSP. The device maps DSCP or EXP priorities to local priorities. **Table 3-23** lists the default mapping. DS-TE map local priorities to CTs and separately allocates resources to each CT. Therefore, DS-TE LSPs are set up according to the CT. That is, the DS-TE calculates the path and reserves resources based on the CT and its bandwidth.

Table 3-23 Default mapping between DSCP priorities, local priorities, and EXP priorities

DSCP Priority	Local Priority	EXP Priority
0-7	0	0
8-15	1	1
16-23	2	2
24-31	3	3
32-39	4	4
40-47	5	5
48-55	6	6
56-63	7	7

RSVP Extensions

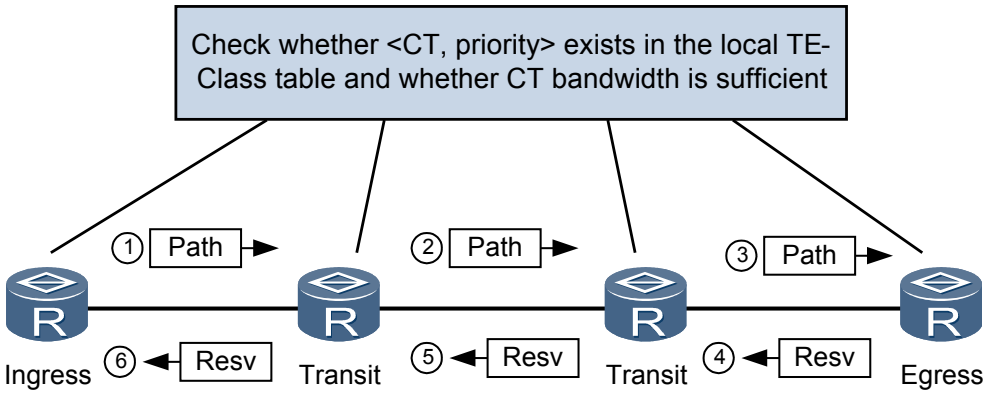
IETF extends RSVP to implement DS-TE in IETF mode. RFC 4124 defines a CLASSTYPE object for Path messages. The IETF draft (draft-minei-diffserv-te-multi-class-02) defines the extended-classtype object that carries CT information about the E-LSP. For details, see RFC 4124 and draft-minei-diffserv-te-multi-class-02.

DS-TE LSP Setup

The DS-TE LSP setup process is similar to **MPLS-TE LSP Setup**. The difference is as follows:

- The RSVP Path message contains CT information.
- When receiving the RSVP Path message carrying CT information, the LSR check whether <CT, priority> exists in the local TE-class mapping table and whether CT bandwidth is sufficient. If <CT, priority> exists in the local TE-class mapping table and CT bandwidth is sufficient, a new LSP can be set up.
- After the LSP is successfully set up, the LSR recalculates the reservable bandwidth for each CT. Information about the reservable bandwidth is then sent to an IGP, and the IGP advertises the information to other nodes over the network.

Figure 3-40 DS-TE LSP setup process



DS-TE Modes

The device provides the IETF mode and non-IETF mode:

- IETF mode: indicates the mode defined by the IETF. Eight CTs are combined with eight priorities and the combinations specify 64 TE-classes. A maximum of eight TE-classes can be configured on the device.
- Non-IETF mode: indicates the mode not defined by the IETF. Each of the two CTs is combined with each of eight priorities, so 16 TE-classes are available.

Table 3-24 describes their differences.

Table 3-24 Differences between the IETF mode and non-IETF mode

Item	Non-IETF Mode	IETF Mode
Bandwidth constraints model	Supports the MAM and RDM.	Supports the RDM, MAM, and extended-MAM.
CT	Supports CT0 and CT1.	Supports CT0 to CT7.
BC type	Supports BC0 and BC1.	Supports BC0 to BC7.
TE-class mapping table	A TE-class mapping table can be configured but cannot take effect.	Supports the configuration and application of the TE-class mapping table.
IGP message	● The Unreserved bandwidth Sub-TLV carries the unreserved bandwidth for eight TE-classes corresponding to CT0, in byte/s. ● The sub-TLV (Unreserved Bandwidth for Class-Type 1, type 0x8001) carries the unreserved bandwidth for eight TE-classes corresponding to CT1, in byte/s.	The CT information is carried in the sub-TLVs. The sub-TLVs are as follows: ● Unreserved Bandwidth Sub-TLV – For RDM and MAM, it carries the unreserved bandwidth for eight TE-classes, in byte/s. – For extended-MAM, it carries the unreserved bandwidth for eight TE-classes corresponding to CT0, in byte/s. ● Bandwidth Constraints Sub-TLV – For RDM and MAM, it carries information about the BC model and the BC bandwidth, in byte/s. – For extended-MAM, it carries information about the BC model and unreserved bandwidth for eight TE-classes, in byte/s.
RSVP messages	The ADSPEC object carries CT information.	Different objects carry CT information as follows: ● Single CT: The CLASSTYPE object carries CT information. ● Multi-CT: The EXTENDED_CLASS-TYPE object carries CT information.

DS-TE Mode Switching

On the device, the non-IETF mode and the IETF mode can be switched to each other. DS-TE mode switching is described in **Table 3-25**.

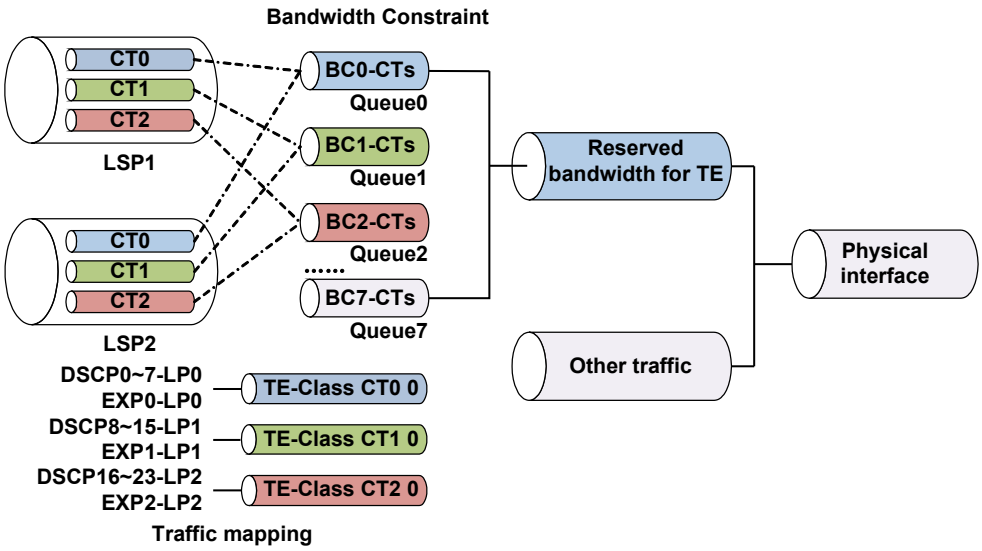
Table 3-25 DS-TE mode switching

Item	Non-IETF Mode → IETF Mode	IETF Mode → Non-IETF Mode
Change in the Bandwidth Constraints model	The bandwidth model is unchanged.	The bandwidth models are changed as follows: ● The extended-MAN is changed to MAM. ● The RDM is unchanged. ● The MAM is unchanged.
Change in the bandwidth	The bandwidth values of BC0 and BC1 are unchanged.	Other BC values are reset to zero except values of BC0 and BC1.
Change in the TE-Class mapping table	If the TE-class mapping table is configured, it is applied. Otherwise, the default one is applied. For information about the default TE-class mapping table, see **Table 3-21**.	The TE-class mapping table is not applied. ● If a TE-class mapping table is configured, it is not deleted. ● If no TE-class mapping table is configured, the default one is deleted.
LSP deletion	LSPs whose <CT, set-priority> or <CT, hold-priority> is not in the TE-class mapping table are deleted on the ingress node and the transit node.	The following LSPs are deleted on the ingress node and the transit node: ● Multi-CT LSPs ● LSPs of single CT from CT2 to CT7

DS-TE Scheduling

An ingress node marks local priorities of packets on the inbound interface based on priority mapping or complex traffic classification. Each local priority is mapped to a CT. Packets arrive at the outbound interface with local priorities. On the outbound interface, HQoS is used to allocate bandwidth for DS-TE traffic, as shown in **Figure 3-41**.

Figure 3-41 HQoS scheduling



- Reserved bandwidth for TE tunnels

  Certain bandwidth is reserved for all TE tunnels on a physical interface by a separate level of queuing. This prevents other types of traffic from occupying the reserved bandwidth.

- Bandwidth guarantee for CTs on TE tunnels

  A TE tunnel supports multiple CTs to transmit services of different types. End-to-end bandwidth is reserved for each CT when a CR-LSP is set up. The device allocate bandwidth to each CT from the bandwidth reserved for TE tunnels when they set up a CR-LSP. The allocated bandwidth conforms to the bandwidth constraints in the RDM or MAM model. The guaranteed bandwidth for a CT is specified by the committed information rate (CIR).

- Traffic scheduling between CTs (total traffic of the same CT on different LSPs)

  Each CT maps a service type. High-priority services (such as voice services) can be scheduled using priority queuing (PQ). Services requiring bandwidth guarantee (such as protocol and data services) can be scheduled using weighted fair queuing (WFQ). CTs are mapped to local priorities in a one-to-one mode. You can associate CTs with queue profiles and configure different scheduling modes in the queue profiles to provide differentiated services for CTs.

The AR provides 32 queue profiles globally configure scheduling modes for CTs. The following table lists the default mappings between CTs and fair queues (FQs), as shown in **Table 3-26**.

Table 3-26 CT scheduling

Local Priority	CT	FQ	Configurable Scheduling Mode
7 (CS7)	CT7	7	WFQ
6 (CS6)	CT6	6	WFQ
5 (EF)	CT5	5	WFQ
4 (AF4)	CT4	4	WFQ
3 (AF3)	CT3	3	WFQ

Local Priority	CT	FQ	Configurable Scheduling Mode
2 (AF2)	CT2	2	WFQ
1 (AF1)	CT1	1	WFQ
0 (BE)	CT0	0	WFQ

DS-TE Reliability

DS-TE provides the following reliability methods:

- TE FRR

  DS-TE is applied as follows:

  - When bandwidth protection is required, the CTs and bandwidth are configured manually on the bypass CR-LSP in manually-configured FRR. The QoS is guaranteed. The protection modes are the 1:1 and N:1. In the automatic FRR, the bypass CR-LSP inherits the CTs and bandwidth of the primary CR-LSP. The QoS is guaranteed. The protection mode is 1:1 only.
  - When bandwidth protection is not required, both manually-configured and automatic FRR support 1:1 protection and N:1 protection, irrespective of the CTs and their bandwidth on the bypass tunnel.

- CR-LSP backup

  The bypass CR-LSP inherits the CTs and their bandwidth from the primary CR-LSP. The best-effort path cannot guarantee QoS and it does not inherit the CTs and bandwidth from the primary CR-LSP.

Interworking Between Devices

During network deployment or device version upgrade, non-DS-TE devices may work with DS-TE devices, or devices in non-IETF mode work with devices in IETF mode.

The device supports the interworking between the following devices:

- Interworking between DS-TE devices and non-DS-TE devices
  - Supports the establishment of non-DS-TE tunnels from non-DS-TE devices to DS-TE devices.
  - Supports the establishment of non-DS-TE tunnels from DS-TE devices to non-DS-TE devices.
- Interworking between non-Huawei DS-TE devices that do not support the CLASSTYPE object

  The device can parse the following Path messages with CT information sent by non-Huawei devices:

  - L-LSP CT information that is carried by the EXTENDED_CLASSTYPE object
  - CTO information that is carried by the EXTENDED_CLASSTYPE object

3.3 Applications

3.3.1 MPLS TE Applications on an IP MAN

Service Overview

As technology advances, service bearing networks of carriers are becoming integrated and IP/MPLS technology is of great importance after network integration. Voice, video, leased line, and data services can be uniformly transmitted over the integrated IP/MPLS network. Services transmitted over MANs are classified into the following two types based on user types:

- Residential services: include high speed Internet (HSI), video on demand (VoD), and voice over IP (VoIP) services.
- Enterprise services: include VPN services for the headquarters and branches of large enterprises. L3VPN services include Business VPNs and L2VPN services include data, real-time video, and real-time voice services.

Table 3-27 lists services, their quality of service (QoS), reliability, and security requirements.

Table 3-27 IP MAN services

Service	QoS Requirements	Reliability Requirements	Security Requirements
HSI	● Bandwidth does not need to be reserved. ● Low QoS requirements	● A standby link is established for end-to-end (E2E) services. If an active link fails, the services can switch to the standby link. ● Voice services must be transmitted in real time and be rapidly switched to a standby link if an active link fails.	● Different types of services are separately transmitted. ● The IP bearer network infrastructure can defend against attacks and viruses to ensure proper network operation.
VoD	● Bandwidth needs to be reserved. ● Medium QoS requirements		
VoIP	● Bandwidth needs to be reserved. ● High QoS requirements		

Service	QoS Requirements	Reliability Requirements	Security Requirements
Business VPN	• Bandwidth needs to be reserved. • Medium QoS requirements		

Networking Description

The IP MAN consists of backbone and access subnetworks. The IP MAN sends services to users. **Figure 3-42** shows E2E residential service networking. **Figure 3-43** shows E2E enterprise service networking.

Figure 3-42 Residential service networking

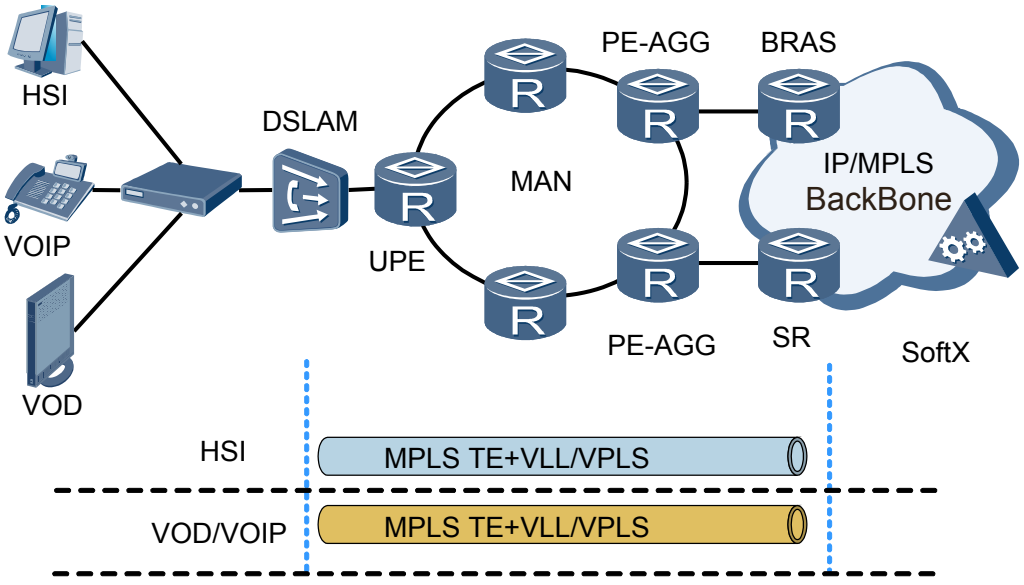
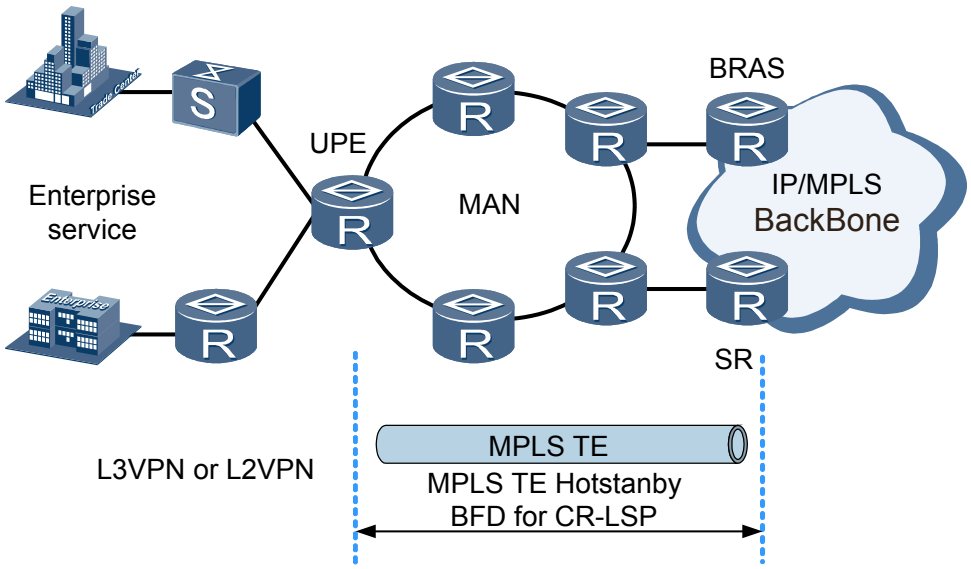


Figure 3-43 Enterprise service networking



Feature Deployment

Services shown in **Figure 3-42** and **Figure 3-43** are core services of carriers. These services have specific bandwidth, QoS, and reliability requirements. To meet these requirements, MPLS TE tunnels can be used over virtual private networks (VPNs). **Table 3-28** lists deployment solutions.

Table 3-28 MPLS TE over an IP MAN

Item	L3VPN	L2VPN
Service	Business VPN	● HSI ● VoD ● VoIP
Public network tunnels for VPN services	MPLS TE tunnels	MPLS TE tunnels

Item	L3VPN	L2VPN
Reliability	● Network reliability: – Link protection: provided using TE Hot-standby and bidirectional forwarding detection (BFD) for constraints-routed label switched path (CR-LSP). – Node protection: provided using VPN fast reroute (FRR) and BFD for TE tunnel. ● Device reliability: provided using RSVP graceful restart (GR) or non-stop routing (NSR).	● Network reliability: – Link protection: provided using TE Hot-standby and BFD for CR-LSP. – Node protection: provided using virtual leased line (VLL) FRR and BFD for TE tunnel. ● Device reliability: provided using RSVP GR or NSR.
QoS	E2E QoS functions need to be deployed on a link between a user-end provider edge device (UPE) and a broadband remote access server (BRAS) or a link between a UPE and a service router (SR).	
Security	Message digest 5 (MD5) or keychain is used to authenticate RSVP messages.	

Key deployment points are as follows:

- Explicit paths are configured to separately establish primary and backup CR-LSPs. The two paths do not overlap in important areas.

3.3.2 DS-TE Applications

Application Scenario: Access of Different Services to One VPN

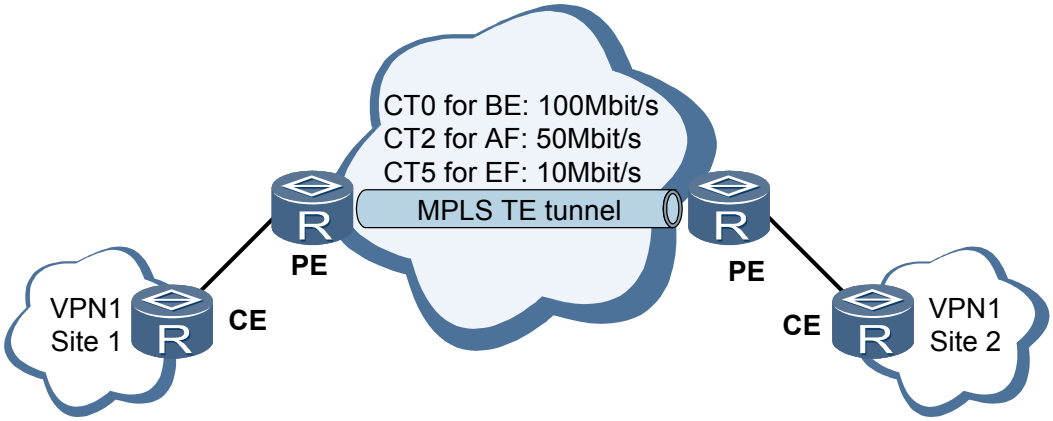
On VPNs with MPLS-TE tunnels, one VPN may transmit EF, AF, and BE services simultaneously. One MPLS TE tunnel may transmit different types of services with different QoS requirements.

To prevent services on one MPLS TE tunnel from affecting each other, set up specific VPNs and TE tunnels to transmit specific services. This is because resources may be wasted when multiple VPNs and tunnels are set up for transmitting difference types of services over a network simultaneously.

Alternatively, you can deploy DS-TE and use a multi-CT LSP to transmit services over one VPN. A multi-CT LSP can reserve up to eight CTs. Each CT can transmit one type of services of one VPN. Services among different CTs does not affect each other.

As shown in **Figure 3-44**, VPN1 transmits EF, AF, and BE services. One DS-TE tunnel needs to be set up and configured with CT0 (100 Mbit/s), CT2 (50 Mbit/s), and CT5 (10 Mbit/s). The tunnel is bound to VPN1 on the ingress. After traffic of VPN1 is classified, the traffic enters corresponding CT queues.

Figure 3-44 One MPLS TE tunnel transmitting different services on one VPN



Application Scenario: Access of Different Services to Different VPNs

On a VPN with MPLS-TE tunnels, multiple VPNs may share one TE tunnel. These VPNs require specific QoS. They compete with each other for resources, and QoS requirements of services over VPNs cannot be met.

The solutions to the preceding scenario are as follows:

- Multiple VPNs transmit different types of services.

  One TE tunnel can be used to transmit a maximum of eight types of services.

  For example, VPN1 and VPN2 can access the TE tunnel simultaneously. VPN1 transmits EF and BE services and VPN2 transmits AF services. One TE tunnel needs to be set up and each type of services on each VPN is configured with a specific CT. The number of CTs is equal to the sum of service types on VPN1 and the number of service types on VPN2. Three CTs are supported.
- Multiple VPNs transmit the same type of services.

  The number of TE tunnels to be set up is equal to the number of VPNs. The number of CTs on each tunnel is equal to the number of corresponding service types on VPNs.

  For example, VPN1 and VPN2 can access the TE tunnel simultaneousl. VPN1 bears EF and BE services and VPN2 also transmits EF and BE services. Two TE tunnels can be set up for VPN1 and VPN2. Each type of services on each tunnel is configured with a specific CT.
- Multiple VPNs transmit services (some services are the same).

  Each VPN needs a tunnel. The number of CTs on each tunnel is equal to the number of corresponding service types on VPNs.

Application Scenario: Access of Traffic to VPNs and Non-VPNs

QoS requirements vary with VPN traffic and non-VPN traffic. If one TE tunnel transmits all the traffic, the VPN traffic and non-VPN traffic may compete with each other for resources, and QoS requirements of services cannot be met.

The solutions to the preceding scenario are as follows:

- The VPN and non-VPN transmit different types of services.

One TE tunnel needs to be set up. Services on the VPN and non-VPN are configured with different CTs. The number of CTs is equal to the sum of the number of service types on the VPN and the number of service types on the non-VPN.

- The VPN and non-VPN transmit the same type of services.

  Two TE tunnels need to be set up for the VPN and non-VPN services. Specific types of services on each tunnel are configured with specific CTs.

- The VPN and non-VPN transmit services (some services are the same).

  Two TE tunnels need to be set up for the VPN and non-VPN services. Specific types of services on each tunnel are configured with specific CTs.

3.4 References

The following table lists the references of this document.

Document	Description	Remarks
RFC 2205	Resource Reservation Protocol	-
RFC 2209	Resource Reservation Protocol (RSVP) - Version 1 Message Processing Rules	-
RFC 2370	The OSPF Opaque LSA Option	-
RFC 2547	BGP/MPLS VPNs	-
RFC 2702	Requirements for Traffic Engineering Over MPLS	-
RFC 2747	RSVP Cryptographic Authentication	-
RFC 2961	RSVP Refresh Overhead Reduction Extensions	-
RFC 3031	Multiprotocol Label Switching Architecture	-
RFC 3032	MPLS Label Stack Encoding	-
RFC 3034	Use of Label Switching on Frame Relay Networks Specification	-
RFC 3209	RSVP-TE: Extensions to RSVP for LSP Tunnels	-
RFC 3210	Applicability Statement for Extensions to RSVP for LSP-Tunnels	-
RFC 3473	Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions	-
RFC 3630	Traffic Engineering (TE) Extensions to OSPF Version 2	-
RFC 3784	Intermediate System to Intermediate System (IS-IS) Extensions for Traffic Engineering (TE)	-
RFC 4124	Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering	-

Document	Description	Remarks
RFC 4127	Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering	-
RFC 4128	Bandwidth Constraints Models for Differentiated Services (Diffserv)-aware MPLS Traffic Engineering: Performance Evaluation	-
RFC 4139	Requirements for Generalized MPLS (GMPLS) Signaling Usage and Extensions for Automatically Switched Optical Network (ASON)	-
RFC 4090	Fast Reroute Extensions to RSVP-TE for LSP Tunnels	-
draft-ietf-mpls-nodeid-subobject-01	Definition of an RRO node-id subobject	-
draft-ietf-tewg-diff-te-proto-02	Protocol extensions for support of Diff-Serv-aware MPLS Traffic Engineering	-
draft-ietf-mpls-diff-te-reqts-00	Requirements for support of Diff-Serv-aware MPLS Traffic Engineering	-
draft-ietf-mpls-diff-ext-07	MPLS Support of Differentiated Services	-

4 MPLS OAM

About This Chapter

4.1 Introduction to MPLS OAM

Definition

Operation, Administration and Maintenance (OAM) is an important means to cut costs in network maintenance. The MPLS OAM mechanism manages operation and maintenance of Multiprotocol Label Switching (MPLS) networks.

MPLS supports different Layer 2 and Layer 3 protocols such as IP, Frame Relay (FR), and Asynchronous Transfer Mode (ATM). In an MPLS network, the OAM mechanism is provided totally independent of any upper or lower layer, which implements the following features on the MPLS user plane:

- Detects connectivity of label switched paths (LSPs).
- Assesses utilization and performance of an MPLS network.
- Performs protection switching when a defect or fault occurs on a link to provide services in compliance with the signed service level agreements (SLAs) signed.

Purpose

As an extensible key technology of the next generation network, MPLS provides multiple services guaranteed by quality of service (QoS). MPLS introduces a unique network layer that may cause faults. Therefore, MPLS networks need to support OAM.

The protocols (such as Synchronous Optical Network (SONET)/Synchronous Digital Hierarchy (SDH)) at the server layer below the MPLS network layer and the protocols (such as IP, FR, and ATM) at the client layer above the MPLS network layer have their respective OAM mechanisms. Failures of the MPLS network cannot be rectified thoroughly through the OAM mechanism of other layers. In addition, the network technology hierarchy also requires MPLS to have its independent OAM mechanism to decrease dependency between layers on each other.

The MPLS OAM mechanism can detect, identify, and locate a defect at the MPLS layer effectively. Then, the MPLS OAM mechanism the reports and handles the defect. In addition, when a failure occurs, the MPLS OAM mechanism can trigger protection switching.

4.2 Principles

4.2.1 MPLS OAM Detection

MPLS OAM packets can be classified into the following types:

- Connectivity detection packets
  - Fast Failure Detection (FFD) packets
  - Connectivity Verification (CV) packets
- Forward Defect Indication (FDI) packets
- Backward Defect Indication (BDI) packets

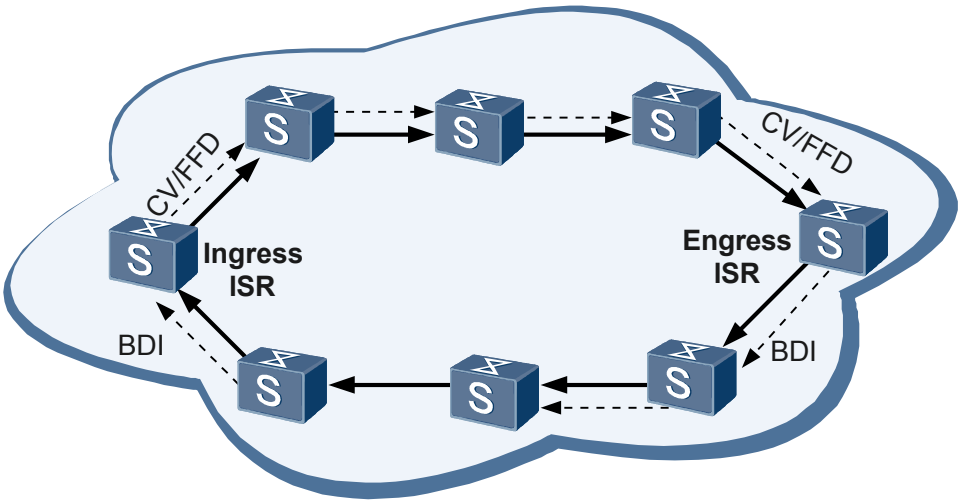
MPLS OAM monitors:

- **TE LSP Connectivity**

MPLS OAM periodically sends CV or FFD packets along a TE LSP.

MPLS OAM for a TE LSP

Figure 4-1 MPLS OAM for a TE LSP



As shown in **Figure 4-1**, MPLS OAM works as follows:

1. The ingress sends a CV packet or an FFD packet along an LSP to be detected. The packet passes through the LSP and arrives at the egress.
2. The egress compares the packet type, interval, and Trail Termination Source Identifier (TTSI) in the received packet with the local values to check the correctness of the packet. In addition, the egress counts the number of received correct and incorrect packets within a detection cycle. In this manner, MPLS OAM detects connectivity of the LSP.
3. The detection interval of CV packet is a fixed value, and the detection cycle of FFD packet is three times the detection interval.
4. When the egress detects an LSP defect, the egress analyzes the defect type and sends a BDI packet carrying the defect information to the ingress through a reverse tunnel. In this manner, the ingress is notified of the defect of a specific type in time. If the protection group is configured correctly, protection switching is triggered.

   The detected defect is of one of the following types:

   - Non-MPLS layer defects
     - dServer: indicates a server-layer defect. A dServer defect is the server-layer defect that occurs below an MPLS network. The defects of this type are reported by the server layer to MPLS OAM for handling.

       The lower layer network that bears MPLS services may have its own protection and defect detection mechanism. When a lower-layer defect occurs on an LSP, a downstream label switch router (LSR) that is closest to the defect can notify the egress of the defect. The lower-layer defect should not trigger the switchover but be only notified to the network management device. In addition, the lower-layer defect can be notified to the ingress through a proper method (of sending BDI packets).

  - dPeerME: indicates a peer maintenance entity defect. A dPeerME defect is the server-layer defect that occurs on a peer maintenance entity outside the MPLS subnet. The defects of this type are reported by other network layers connected to the MPLS subnet to MPLS OAM for handling.

- MPLS layer defects
  - dLOCV: indicates the defect of connectivity verification loss.

    A dLOCV defect occurs when no CV or FFD packet is received within three consecutive intervals for sending CV or FFD packets.
  - dTTSI_Mismatch: indicates the defect of TTSI mismatching.

    A dTTSI_Mismatch defect occurs when no CV or FFD packet with a correct TTSI is received within three consecutive intervals for sending CV or FFD packets.
  - dTTSI_Mismerge: indicates the defect of TTSI mismerging.

    A dTTSI_Mismerge defect occurs when CV or FFD packets with both correct and incorrect TTSIs are received within three consecutive intervals for sending CV or FFD packets.
  - dExcess: indicates the defect of the excessive rate of receiving connectivity detection packets.

    A dExcess defect occurs when five or more correct CV or FFD packets are received within three consecutive intervals for sending CV or FFD packets.
- Other defects

  dUnknown: indicates an unknown defect in an MPLS network.

  A defect can be defined as dUnknown. For example, if the egress detects a defect that both CV packets and FFD packets are sent along the same LSP, this special defect that is not defined in the protocol can be identified as a dUnknown defect.

4.2.2 Reverse Tunnel

When the basic OAM function is configured, the LSP to be detected needs to be bound to a reverse tunnel for transmitting BDI packets.

BDI packets are transmitted through the reverse tunnel. A reverse tunnel can be an LSP with the ingress and egress being opposite to those on the detected LSP. The reverse tunnel can also be a non-MPLS path that connects the ingress and the egress of the detected LSP.

The reverse tunnel transmitting BDI packets can be one of the following types:

- Private reverse LSP
- Shared reverse LSP
- Non-MPLS reverse path

NOTE

Currently, Huawei only supports a TE tunnel as the reverse tunnel.

4.2.3 MPLS OAM Auto Protocol

MPLS OAM defined in ITU-T Recommendation Y.1710 and Y.1711 has the following drawbacks:

- On an LSP, if the ingress is enabled with the OAM function later than the egress, or OAM is enabled on the egress and disabled on the ingress, a dLOCV defect occurs.

- The dLOCV defect also occurs when OAM is disabled. You must disable OAM on the ingress and egress before changing the type or updating the interval for sending detection packets.
- The OAM parameters need to be set on the ingress and egress respectively. This, however, may cause the detection packet type and the interval for sending detection packets on the ingress to be different from those on the egress.

On the Huawei devices, the OAM auto protocol can address the preceding problems.

If the OAM auto protocol is enabled on the egress, the functions of first packet triggering OAM and the dynamic enabling or disabling OAM are provided.

The MPLS OAM auto protocol is the patent of Huawei.

4.2.4 Protection Switching

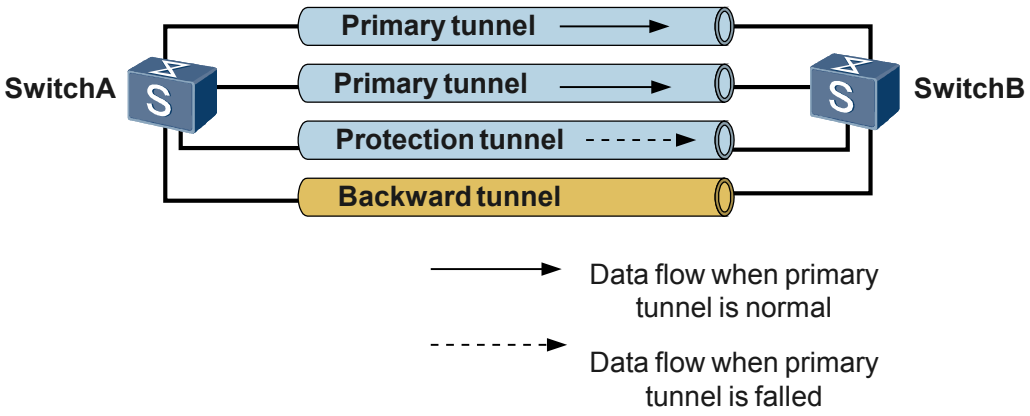
Protection switching refers to that a protection tunnel (namely, the bypass tunnel) is pre-set for the primary tunnel and assigned with bandwidth. The primary tunnel and the bypass tunnel form a protection group. When the primary tunnel is faulty, data traffic can be quickly switched to the bypass tunnel. This decreases the packet loss ratio or shortens the delay due to the LSP failure, and enhances reliability of networks. Protection switching refers to the end-to-end protection.

With MPLS OAM for fast fault detection, the protection switching can be performed in milliseconds.

On the Huawei devices, protection switching can be performed in 1:1 mode or N:1 mode.

- In 1:1 mode, a primary tunnel and a bypass tunnel are set up between the ingress and egress.
  - Normally, data is transmitted through the primary tunnel.
  - When the ingress detects a fault on the primary tunnel through MPLS OAM, protection switching is performed and the ingress switches data to the bypass tunnel for transmission.
- In N:1 mode, a tunnel functions as the bypass tunnel for multiple primary tunnels. When any primary tunnel fails, data is switched to the shared bypass tunnel. The N:1 mode is used to save bandwidth in a network with the mesh topology.

Figure 4-2 Schematic diagram of the MPLS OAM tunnel protection



4.3 References

The following table lists the references of this document.

Huawei implements MPLS OAM based on ITU-T recommendations; however, the Request for Comments (RFCs) are only for reference.

Document	Description	Remarks
ITU-T Recommendation Y.1710	Requirements for Operation & Maintenance functionality for MPLS networks	Huawei implement MPLS OAM in compliance with this recommendation.
ITU-T Recommendation Y.1711	Operation & Maintenance mechanism for MPLS networks	Huawei implement MPLS OAM in compliance with this recommendation.
ITU-T Recommendation Y.1720	Protection switching for MPLS networks	Huawei implement MPLS OAM in compliance with this recommendation.
RFC 3429	Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions	This RFC is only reference for Huawei to implement MPLS OAM.
RFC 4377	Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks	This RFC is only reference for Huawei to implement MPLS OAM.
RFC 4378	A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)	This RFC is only reference for Huawei to implement MPLS OAM.