

MPLS TE Technology Overview

In this chapter, you review the following topics:

- MPLS TE Introduction
- Basic Operation of MPLS TE
- DiffServ-Aware Traffic Engineering
- Fast Reroute

This chapter presents a review of *Multiprotocol Label Switching Traffic Engineering* (MPLS TE) technology. MPLS TE can play an important role in the implementation of network services with *quality of service* (QoS) guarantees. The initial sections describe the basic operation of the technology. This description includes the details of TE information distribution, path computation, and the signaling of TE LSPs. The subsequent sections present how *Differentiated Services (DiffServ)-Aware traffic engineering* (DS-TE) helps integrate the implementation of DiffServ and MPLS TE. This chapter closes with a review of the *fast reroute* (FRR) capabilities in MPLS TE. Chapter 4, “Cisco MPLS Traffic Engineering,” covers in depth the Cisco implementation of MPLS TE in Cisco IOS and Cisco IOS XR. In addition, Chapter 5, “Backbone Infrastructure,” discusses the different network designs that can combine QoS with MPLS TE.

MPLS TE Introduction

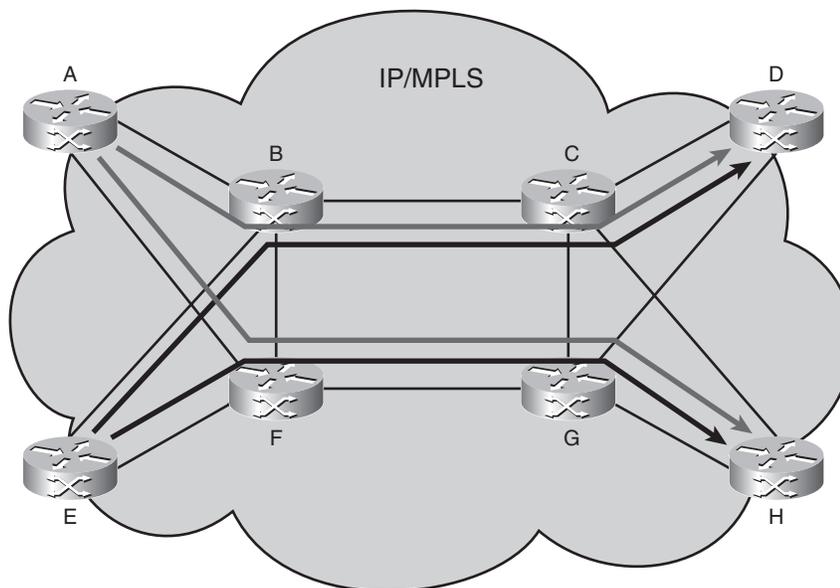
MPLS networks can use native TE mechanisms to minimize network congestion and improve network performance. TE modifies routing patterns to provide efficient mapping of traffic streams to network resources. This efficient mapping can reduce the occurrence of congestion and improves service quality in terms of the latency, jitter, and loss that packets experience. Historically, IP networks relied on the optimization of underlying network infrastructure or *Interior Gateway Protocol* (IGP) tuning for TE. Instead, MPLS extends existing IP protocols and makes use of MPLS forwarding capabilities to provide native TE. In addition, MPLS TE can reduce the impact of network failures and increase service availability. RFC 2702 discusses the requirements for TE in MPLS networks.

MPLS TE brings explicit routing capabilities to MPLS networks. An originating *label switching route* (LSR) (or headend) can set up a TE *label switched path* (LSP) to a terminating LSR (or tail end) through an explicitly defined path containing a list of

intermediate LSRs (or midpoints). IP uses destination-based routing and does not provide a general and scalable method for explicitly routing traffic. In contrast, MPLS networks can support destination-based and explicit routing simultaneously. MPLS TE uses extensions to RSVP and the MPLS forwarding paradigm to provide explicit routing. These enhancements provide a level of routing control that makes MPLS suitable for TE.

Figure 2-1 shows a sample MPLS network using TE. This network has multiple paths from nodes A and E toward nodes D and H. In this figure, traffic from A and E toward D follows explicitly routed LSPs through nodes B and C. Traffic from A and E toward H follows explicitly routed LSPs through nodes F and G. Without TE, the IGP would compute the shortest path using only a single metric or cost. You could tune that metric, but that would provide you limited capabilities to allocate network resources when compared with MPLS TE (specially, when you consider larger more complex network topologies). This chapter describes those routing and signaling enhancements that make MPLS TE possible.

Figure 2-1 Sample MPLS Network Using TE



MPLS TE also extends the MPLS routing capabilities with support for constraint-based routing. As mentioned earlier, IGP's typically compute routing information using a single metric. Instead of that simple approach, constraint-based routing can take into account more detailed information about network constraints, and policy resources. MPLS TE extends current link-state protocols (IS-IS and OSPF) to distribute such information.

Constraint-based routing and explicit routing allow an originating LSR to compute a path that meets some requirements (constraints) to a terminating LSR and then set up a TE LSP through that path. Constraint-based routing is optional within MPLS TE. An offline tool can perform path computation and leave TE LSP signaling to the LSRs.

MPLS TE supports preemption between TE LSPs of different priorities. Each TE LSP has a setup and a holding priority, which can range from zero (best priority) through seven (worst priority). When a node signals a new TE LSP, other nodes throughout the path compare the setup priority of the new TE LSPs with the holding priority of existing TE LSPs to make a preemption decision. A better setup priority can preempt worse-holding priorities a TE LSP can use hard or soft preemption. A node implementing hard preemption tears down the existing TE LSP to accommodate the new TE LSP. In contrast, a node implementing soft preemption signals back the pending preemption to the headend of the existing TE LSP. The headend can then reroute the TE LSP without impacting the traffic flow. RFC 3209 and draft-ietf-mpls-soft-preemption-07. define TE LSP preemption.

Basic Operation of MPLS TE

The operation of MPLS TE involves link information distribution, path computation, LSP signaling, and traffic selection. This section explains the most important concepts behind each of these four steps. LSRs implement the first two steps, link information distribution and path computation, when they need perform constraint-based routing. MPLS networks that do not use constraint-based routing (or use an offline tool for that purpose) perform only LSP signaling and traffic selection. MPLS TE does not define any new protocols even though it represents a significant change in how MPLS networks can route traffic. Instead, it uses extensions to existing IP protocols.

Link Information Distribution

MPLS TE uses extensions to existing IP link-state routing protocols to distribute topology information. An LSR requires detailed network information to perform constraint-based routing. It needs to know the current state of an extended list of link attributes to take a set of constraints into consideration during path computation for a TE LSP. Link-state protocols (IS-IS and OSPF) provide the flooding capabilities that are required to distribute these attributes. LSRs use this information to build a TE topology database. This database is separate from the regular topology database that LSRs build for hop-by-hop destination-based routing.

MPLS TE introduces available bandwidth, an administrative group (flags), and a TE metric as new link attributes. Each link has eight amounts of available bandwidth corresponding to the eight priority levels that TE LSPs can have. The administrative group (flags) acts as

a classification mechanism to define link inclusion and exclusion rules. The TE metric is a second link metric for path optimization (similar to the IGP link metric). In addition, LSRs distribute a TE ID that has a similar function to a router ID. The OSPF and IS-IS extensions mirror each other and have the same semantics. Table 2-1 shows the complete list of link attributes. RFC 3784 and RFC 3630 define the IS-IS and OSPF extensions for TE respectively.

Table 2-1 *Extended Link Attributes Distributed for TE*

Link Attribute	Description
Interface address	IP address of the interface corresponding to the link
Neighbor address	IP address of the neighbor's interface corresponding to the link
Maximum link bandwidth	True link capacity (in the neighbor direction)
Reservable link bandwidth	Maximum bandwidth that can be reserved (in the neighbor direction)
Unreserved bandwidth	Available bandwidth at each of the (eight) preemption priority levels (in the neighbor direction)
TE metric	Link metric for TE (may differ from the IGP metric)
Administrative group	Administratively value (flags) associated with the link for inclusion/exclusion policies

NOTE In addition to the attributes listed in Table 2-1, OSPF advertises the link type (point-to-point or multi-access) and a link ID. OSPF uses type 10 opaque (area-local scope) *link-state advertisements* (LSA) to distribute this information.

MPLS TE can still perform constraint-based routing in the presence of multiple IGP areas or multiple autonomous systems. OSPF and IS-IS use the concept of areas or levels to limit the scope of information flooding. An LSR in a network with multiple areas only builds a partial topology database. The existence of these partial databases has some implications for path computation, as the next section describes. LSRs in an inter-autonomous system TE environment also need to deal with partial network topologies. Fortunately, inter-area TE and inter-autonomous system TE use similar approaches for constraint-based routing in the presence of partial topology information.

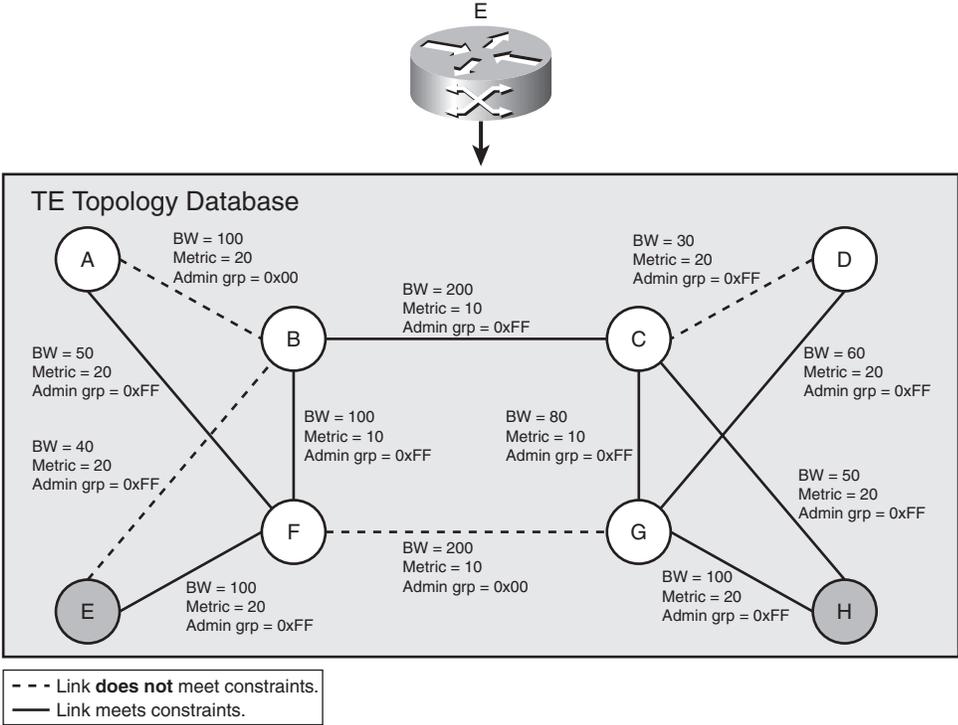
Path Computation

LSRs can perform path computation for a TE LSP using the TE topology database. A common approach for performing constraint-based routing on the LSRs is to use an extension of the *shortest path first* (SPF) algorithm. This extension to the original algorithm

generally receives the name of *constraint-based, shortest path first* (CSPF). The modified algorithm executes the SPF algorithm on the topology that results from the removal of the links that do not meet the TE LSP constraints. The algorithm may use the IGP link metric or the link TE metric to determine the shortest path. CSPF does not guarantee a completely optimal mapping of traffic streams to network resources, but it is considered an adequate approximation. MPLS TE specifications do not require that LSRs perform path computation or attempt to standardize a path computation algorithm.

Figure 2-2 illustrates a simplified version of CSPF on a sample network. In this case, node E wants to compute the shortest path to node H with the following constraints: only links with at least 50 bandwidth units available and an administrative group value of 0xFF. Node E examines the TE topology database and disregards links with insufficient bandwidth or administrative group values other than 0xFF. The dotted lines in the topology represent links that CSPF disregards. Subsequently, node E executes the shortest path algorithm on the reduced topology using the link metric values. In this case, the shortest path is {E, F, B, C, H}. Using this result, node E can initiate the TE LSP signaling.

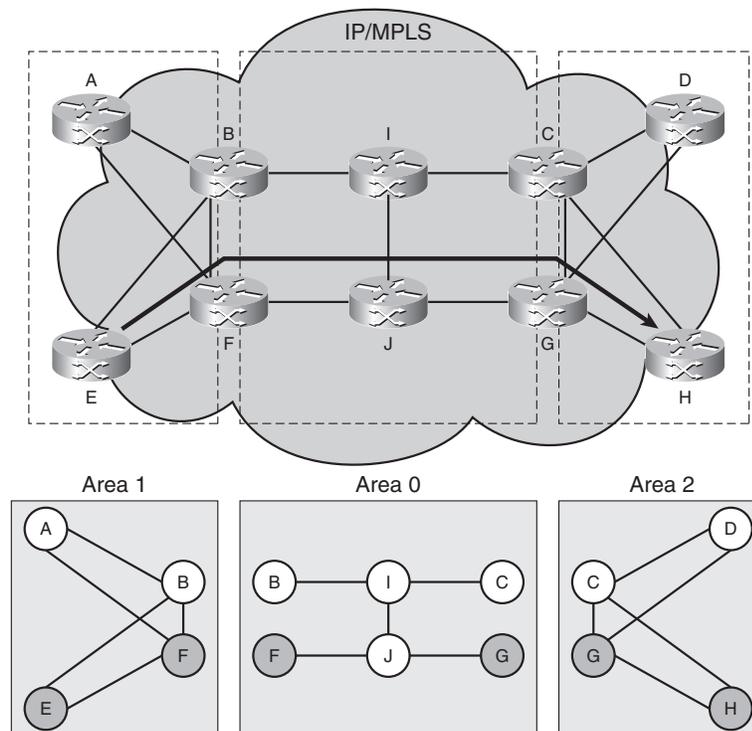
Figure 2-2 Path Computation Using the CSPF Algorithm



Path computation in multi-area or inter-autonomous system environments may involve several partial computations along the TE LSP. When the headend does not have a complete view of the network topology, it can specify the path as a list of predefined boundary LSR (*Area Border Router [ABR]* in the case of inter-area and *Autonomous System Boundary Router [ASBR]* in the case of inter-autonomous system). The headend can compute a path to the first boundary LSR (which must be in its topology database and initiate the signalling of the TE LSP signaling can be initiated). When the signaling reaches the boundary LSR, that LSR performs the path computation to the final destination if it is in its topology. If the destination is not in the topology, the signaling should indicate the next exit boundary LSR, and the path computation will take place to that boundary LSR. The process continues until the signaling reaches the destination. This process of completing path computation during TE LSP signaling is called loose routing.

Figure 2-3 shows path computation in a sample network with multiple IGP areas. All LSRs have a partial network topology. The network computes a path between nodes E and H crossing the three IGP areas in the network. Node E selected nodes F and G, which have as the boundary LSRs that the TE LSP will traverse. Node E computes the path to node F and initiates the TE LSP signaling. When node F receives the signaling message, it computes the next segment of the path toward node G. When the signaling arrives at node G, it completes the path computation toward node H in area 2. The next section explains how LSRs signal TE LSPs.

Figure 2-3 Multi-Area Path Computation



Signaling of TE LSPs

MPLS TE introduces extensions to RSVP to signal up LSPs. RSVP uses five new objects: LABEL_REQUEST, LABEL, EXPLICIT_ROUTE, RECORD_ROUTE, and SESSION_ATTRIBUTE. RSVP Path messages use a LABEL_REQUEST object to request a label binding at each hop. Resv messages use a LABEL object to perform label distribution. Network nodes perform downstream-on-demand label distribution using these two objects. The EXPLICIT_ROUTE object contains a hop list that defines the explicit routed path that the signaling will follow. The RECORD_ROUTE object collects hop and label information along the signaling path. The SESSION_ATTRIBUTE object lists the attribute requirements of the LSP (priority, protection, and so forth).

RFC 3209 defines these RSVP TE extensions. Table 2-2 summarizes the new RSVP objects and their function.

NOTE

The *Internet Engineering Task Force* (IETF) considered extensions to the *Label Distribution Protocol* (LDP) as a signaling protocol for TE LSPs in the early stages of MPLS TE. These protocol extensions were called *Constraint-based routed LDP* (CR-LDP). For some time, CR-LDP and RSVP TE specifications advanced simultaneously. In 2002, the MPLS working group at the IETF decided not to pursue new developments for CR-LDP and focused instead on RSVP TE as the prime protocol for MPLS TE.

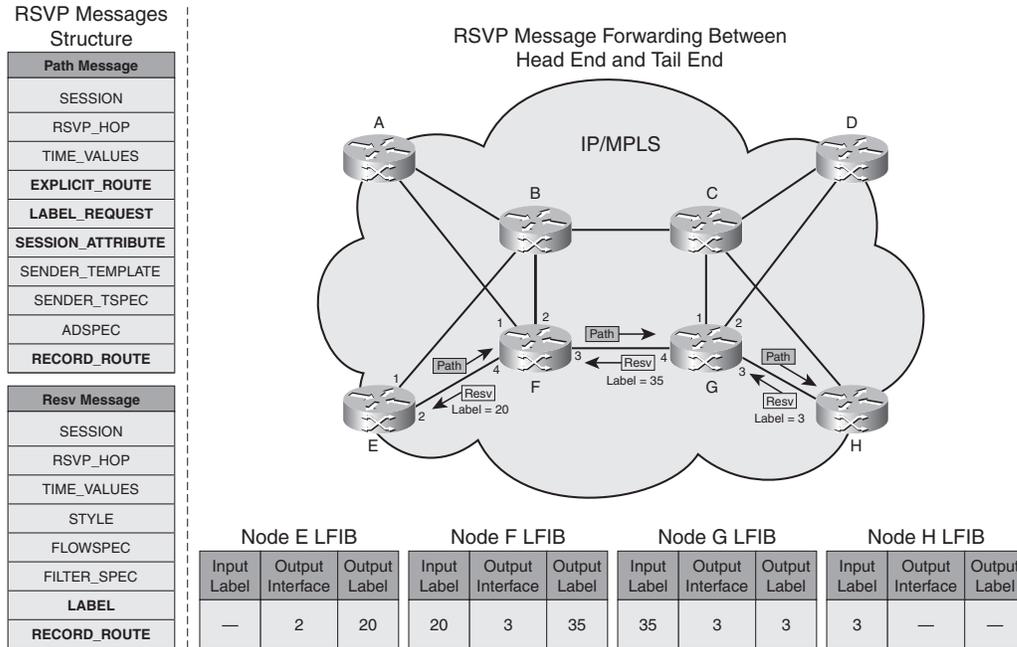
Table 2-2 *New RSVP Objects to Support MPLS TE*

RSVP Object	RSVP Message	Description
LABEL_REQUEST	Path	Label request to downstream neighbor
LABEL	Resv	MPLS label allocated by downstream neighbor
EXPLICIT_ROUTE	Path	Hop list defining the course of the TE LSP
RECORD_ROUTE	Path, Resv	Hop/label list recorded during TE LSP setup
SESSION_ATTRIBUTE	Path	Requested LSP attributes (priority, protection, affinities)

Figure 2-4 illustrates the setup of a TE LSP using RSVP. In this scenario, node E signals a TE LSP toward node H. RSVP Path messages flow downstream with a collection of objects, four of which are related to MPLS TE (EXPLICIT_ROUTE, LABEL_REQUEST, SESSION_ATTRIBUTE, and RECORD_ROUTE). Resv messages flow upstream and include two objects related to MPLS TE (LABEL and RECORD_ROUTE). Each node performs admission control and builds the *LSP forwarding information base* (LFIB) when processing the Resv messages. The structure of the LFIB and the MPLS forwarding

algorithm remain the same regardless of the protocols that populated the information (for example, RSVP in the case of MPLS TE).

Figure 2-4 TE LSP Setup Using RSVP



Traffic Selection

MPLS TE separates TE LSP creation from the process of selecting the traffic that will use the TE LSP. A headend can signal a TE LSP, but traffic will start flowing through the LSP after the LSR implements a traffic-selection mechanism. Traffic can enter the TE LSP only at the headend. Therefore, the selection of the traffic is a local head-end decision that can use different approaches without the need for standardization. The selection criteria can be static or dynamic. It can also depend on the packet type (for instance, IP or Ethernet) or packet contents (for example, class of service). An MPLS network can make use of several traffic-selection mechanisms depending on the services it offers.

DiffServ-Aware Traffic Engineering

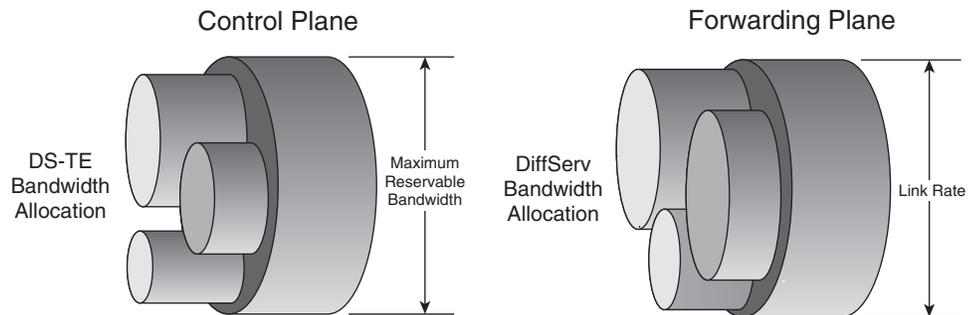
MPLS DS-TE enables per-class TE across an MPLS network. DS-TE provides more granular control to minimize network congestion and improve network performance. DS-TE retains the same overall operation framework of MPLS TE (link information

distribution, path computation, signaling, and traffic selection). However, it introduces extensions to support the concept of multiple classes and to make per-class constraint-based routing possible. These routing enhancements help control the proportion of traffic of different classes on network links. RFC 4124 introduces DS-TE protocol extensions.

Both DS-TE and DiffServ control the per-class bandwidth allocation on network links. DS-TE acts as a control-plane mechanism, while DiffServ acts in the forwarding plane. In general, the configuration in both planes will have a close relationship. However, they do not have to be identical. They can use a different number of classes and different relative bandwidth allocations to satisfy the requirements of particular network designs. Figure 2-5 shows an example of bandwidth allocation in DiffServ and DS-TE for a particular link. In this case, the link rate equals the maximum reservable bandwidth for TE. Each class receives a fraction of the total bandwidth amount in the control and forwarding planes. However, the bandwidth proportions between classes differ slightly in this case.

DS-TE does not imply the use of *Label-inferred-class LSP (L-LSPs)*. An MPLS node may signal a DS-TE LSP as an *EXP-inferred-class LSP (E-LSP)* or L-LSP. Furthermore, a DS-TE LSP may not signal any DiffServ information or not even count on the deployment of DiffServ on the network. You need to keep in mind that an instance of a class within DS-TE does not need to maintain a one-to-one relationship with a DiffServ class. Chapter 5 explains different models of interaction between TE and DiffServ.

Figure 2-5 *Bandwidth Allocation in DiffServ and DS-TE*



NOTE This section uses the term *aggregate MPLS TE* to refer to MPLS TE without the DS-TE extensions. Even though that name might not be completely accurate in some MPLS TE designs, it differentiates TE with a single bandwidth constraint from the approach that DS-TE uses.

Class-Types and TE-Classes

DS-TE uses the concept of *Class-Type* (CT) for the purposes of link bandwidth allocation, constraint-based routing, and admission control. A network can use up to eight CTs (CT0 through CT7). DS-TE retains support for TE LSP preemption, which can operate within a CT or across CTs. TE LSPs can have different preemption priorities regardless of their CT. CTs represent the concept of a class for DS-TE in a similar way that *per-hop behavior* (PHB) *scheduling class* (PSC) represents it for DiffServ. Note that flexible mappings between CTs and PSCs are possible. You can define a one-to-one mapping between CTs and PSCs. Alternatively, a CT can map to several PSCs, or several CTs can map to one PSC.

DS-TE provides flexible definition of preemption priorities while retaining the same mechanism for distribution of unreserved bandwidth on network links. DS-TE redefines the meaning of the unreserved bandwidth attribute discussed in the section “Link Information Distribution” without modifying its format. When DS-TE is in use, this attribute represents the unreserved bandwidth for eight TE classes. A TE-Class defines a combination of a CT and a corresponding preemption priority value. A network can use any 8 (TE-Class) combinations to use out of 64 possible combinations (8 CTs times 8 priorities). No relative ordering exists between the TE-Classes, and a network can use a subset of the 8 possible values. However, the TE-Class definitions must be consistent across the DS-TE network.

Tables 2-3, 2-4, and 2-5 include examples of three different TE-Class definitions:

- Table 2-3 illustrates a TE-Class definition that is backward compatible with aggregate MPLS TE. In this example, all TE-Classes support only CT0, with 8 different preemption priorities ranging from 0 through 7.
- Table 2-4 presents a second example where the TE-Class definition uses 4 CTs (CT0, CT1, CT2, and CT3), with 8 preemption priority levels (0 and 7) for each CT. This definition makes preemption possible within CTs but not across CTs.
- Table 2-5 contains a TE-Class definition with 2 CTs (CT0 and CT1) and 2 preemption priority levels (0 and 7). 2 third example defines some TE-Classes as *unused*. In this case, preemption is possible within and across CTs. With this design, preemption is possible within and across CTs, but you can signal CT1 TE LSPs (using priority zero) that no other TE LSP can preempt.

Table 2-3 TE-Class Definition Backward Compatible with Aggregate MPLS TE

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

Table 2-4 *TE-Class Definition with Four CTs and 8 Preemption Priorities*

TE Class	Class Type	Priority
0	0	7
1	0	6
2	1	5
3	1	4
4	2	3
5	2	2
6	3	1
7	3	0

Table 2-5 *TE-Class Definition with Two CTs and Two Preemption Priorities*

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

Table 2-6 *TE-Class Definition with Two CTs and Eight Preemption Priorities*

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

DS-TE introduces a new CLASSTYPE RSVP object. This object specifies the CT associated with the TE LSP and can take a value ranging from one to seven. DS-TE nodes must support this new object and include it in Path messages, with the exception of CT0 TE LSPs. The Path messages associated with those LSPs must not use the CLASSTYPE object to allow non-DS-TE nodes to interoperate with DS-TE nodes. Table 2-6 summarizes the CLASSTYPE object.

Table 2-7 *New RSVP Object for DS-TE*

RSVP Object	RSVP Message	FRR Function
CLASSTYPE	Path	CT associated with the TE LSP. Not used for CT0 for backward compatibility with non-DS-TE nodes.

Bandwidth Constraints

A set of *bandwidth constraints* (BC) defines the rules that a node uses to allocate bandwidth to different CTs. Each link in the DS-TE network has a set of BCs that applies to the CTs in use. This set may contain up to eight BCs. When a node using DS-TE admits a new TE LSP on a link, that node uses the BC rules to update the amount of unreserved bandwidth for each TE-Class. One or more BCs may apply to a CT depending on the model. DS-TE can support different BC models. The IETF has primarily defined two BC models: *maximum allocation model* (MAM) and *Russian dolls model* (RDM). These are discussed in the following subsections of this chapter.

DS-TE also defines a BC extension for IGP link advertisements. This extension complements the link attributes that Table 2-1 already described and applies equally to OSPF and IS-IS. Network nodes do not need this BC information to perform path computation. They rely on the unreserved bandwidth information for that purpose. However, they can optionally use it to verify DS-TE configuration consistency throughout the network or as a path computation heuristic (for instance, as a tie breaker for CSPF). A DS-TE deployment could use different BC models throughout the network. However, the simultaneous use of different models increases operational complexity and can adversely impact bandwidth optimization. Table 2-8 summarizes the BC link attribute that DS-TE uses.

Table 2-8 *Optional BC Link Attribute Distributed for DS-TE*

Link Attribute	Description
BCs	BC model ID and BCs (BC0 through BC n) that the link uses for DS-TE

Maximum Allocation Model

The MAM defines a one-to-one relationship between BCs and Class-Types. BC n defines the maximum amount of reservable bandwidth for CT n , as Table 2-9 shows. The use of preemption does not affect the amount of bandwidth that a CT receives. MAM offers

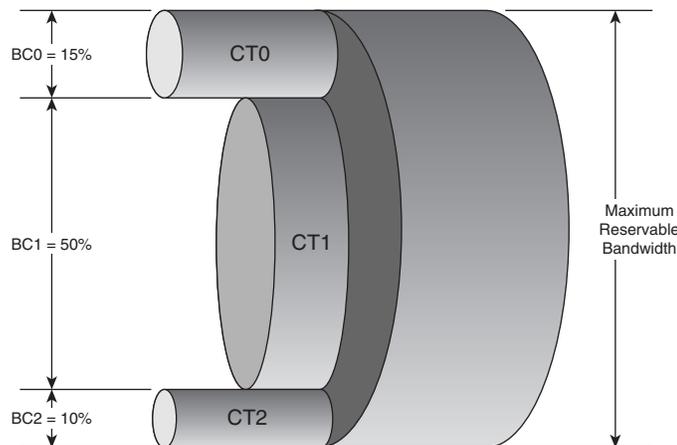
limited bandwidth sharing between CTs. A CT cannot make use of the bandwidth left unused by another CT. The packet schedulers managing congestion in the forwarding plane typically guarantee bandwidth sharing. To improve bandwidth sharing using MAM, you may make the sum of all BCs greater than the maximum reservable bandwidth. However, the total reserved bandwidth for all CTs cannot exceed the maximum reservable bandwidth at any time. RFC 4125 defines MAM.

Table 2-9 MAM Bandwidth Constraints for Eight CTs

Bandwidth Constraint	Maximum Bandwidth Allocation For
BC7	CT7
BC6	CT6
BC5	CT5
BC4	CT4
BC3	CT3
BC2	CT2
BC1	CT1
BC0	CT0

Figure 2-6 shows an example of a set of BCs using MAM. This DS-TE configuration uses three CTs with their corresponding BCs. In this case, BC0 limits CT0 bandwidth to 15 percent of the maximum reservable bandwidth. BC1 limits CT1 to 50 percent, and BC2 limits CT2 to 10 percent. The sum of BCs on this link is less than its maximum reservable bandwidth. Each CT will always receive its bandwidth share without the need for preemption. Preemption will not have an effect on the bandwidth that a CT can use. This predictability comes at the cost of no bandwidth sharing between CTs. The lack of bandwidth sharing can force some TE LSPs to follow longer paths than necessary.

Figure 2-6 MAM Constraint Model Example



Russian Dolls Model

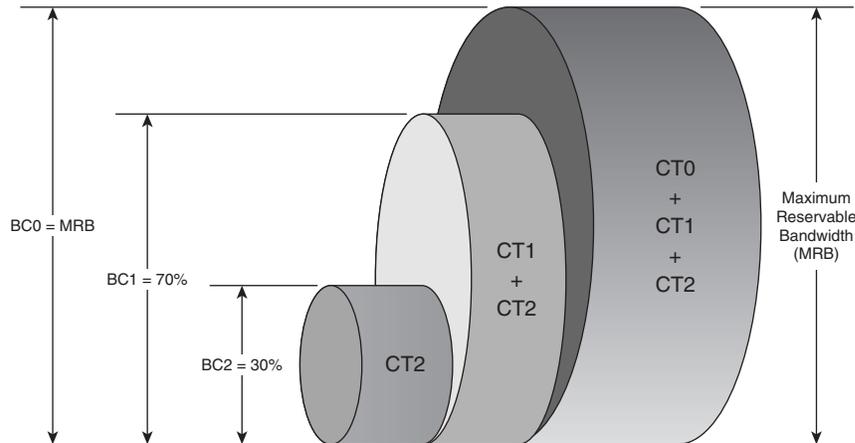
The RDM defines a cumulative set of constraints that group CTs. For an implementation with n CTs, BC_n always defines the maximum bandwidth allocation for CT_n . Subsequent lower BCs define the total bandwidth allocation for the CTs at equal or higher levels. BC_0 always defines the maximum bandwidth allocation across all CTs and is equal to the maximum reservable bandwidth of the link.

Table 2-10 shows the RDM BCs for a DS-TE implementation with eight CTs. The recursive definition of BCs improves bandwidth sharing between CTs. A particular CT can benefit from bandwidth left unused by higher CTs. A DS-TE network using RDM can rely on TE LSP preemption to guarantee that each CT gets a fair share of the bandwidth. RFC 4127 defines RDM.

Table 2-10 RDM Bandwidth Constrains for Eight CTs

Bandwidth Constraint	Maximum Bandwidth Allocation For
BC7	CT7
BC6	CT7+CT6
BC5	CT7+CT6+CT5
BC4	CT7+CT6+CT5+CT4
BC3	CT7+CT6+CT5+CT4+CT3
BC2	CT7+CT6+CT5+CT4+CT3+CT2
BC1	CT7+CT6+CT5+CT4+CT3+CT2+CT1
BC0 = Maximum reservable bandwidth	CT7+CT6+CT5+CT4+CT3+CT2+CT1+CT0

Figure 2-7 shows an example of a set of BCs using RDM. This DS-TE implementation uses three CTs with their corresponding BCs. In this case, BC2 limits CT2 to 30 percent of the maximum reservable bandwidth. BC1 limits CT2+CT1 to 70 percent. BC0 limits CT2+CT1+CT0 to 100 percent of the maximum reservable bandwidth, as is always the case with RDM. CT0 can use up to 100 percent of the bandwidth in the absence of CT2 and CT1 TE LSPs. Similarly, CT1 can use up to 70 percent of the bandwidth in the absence of TE LSPs of the other two CTs. CT2 will always be limited to 30 percent when no CT0 or CT1 TE LSPs exist. The maximum bandwidth that a CT receives on a particular link depends on the previously signaled TE LSPs, their CTs, and the preemption priorities of all TE LSPs. Table 2-11 compares MAM and RDM.

Figure 2-7 RDM Constraint Model Example**Table 2-11** Comparing MAM and RDM BC Models

MAM	RDM
1 BC per CT.	1 or more CTs per BC.
Sum of all BCs may exceed maximum reservable bandwidth.	BC0 always equals the maximum reservable bandwidth.
Preemption not required to provide bandwidth guarantees per CT.	Preemption required to provide bandwidth guarantees per CT.
Bandwidth efficiency and protection against QoS degradation are mutually exclusive.	Provides bandwidth efficiency and protection against QoS degradation simultaneously.

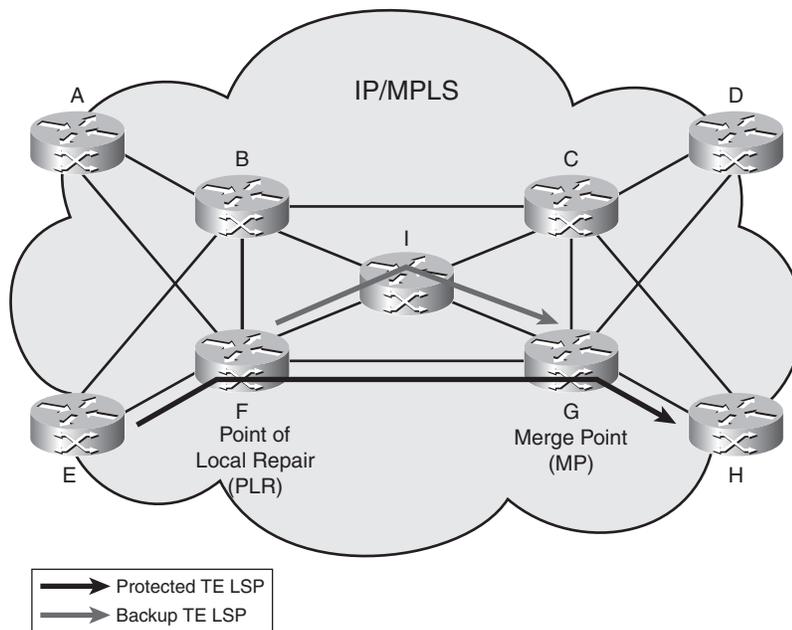
Fast Reroute

MPLS TE supports local repair of TE LSPs using FRR. Traffic protection in case of a network failure is critical for real-time traffic or any other traffic with strict packet-loss requirements. In particular, FRR uses a local protection approach that relies on a presignaled backup TE LSP to reroute traffic in case of a failure. The node immediately next to the failure is responsible for rerouting the traffic and is the headend of the backup TE LSP. Therefore, no delay occurs in the propagation of the failure condition, and no delay occurs in computing a path and signaling a new TE LSP to reroute the traffic. FRR can reroute traffic in tens of milliseconds. RFC 4090 describes the operation and the signaling extensions that MPLS TE FRR requires.

NOTE MPLS TE FRR specifications offer two protection techniques: facility backup and one-to-one backup. Facility backup uses label stacking to reroute multiple protected TE LSPs using a single backup TE LSP. One-to-one backup does not use label stacking, and every protected TE LSP requires a dedicated backup TE LSP. The remainder of this section focuses on the facility backup approach because of its greater scalability and wider use.

Figure 2-8 shows an example of an MPLS network using FRR. In this case, node E signals a TE LSP toward node H. The network protects this TE LSP against a failure of the link between nodes F and G. Given the local protection nature of FRR, node F is responsible for rerouting the traffic into the backup TE LSP in case the link between nodes F and G fails. This role makes node F the *point of local repair* (PLR). It has presignaled a backup TE LSP through node I toward node G to bypass the potential link failure. The PLR is always the headend of the backup TE LSP. Node G receives the name of *merge point* (MP) and is the node where the protected traffic will exit the backup TE LSP during the failure and retake the original path of the protected TE LSP.

Figure 2-8 MPLS Network Using FRR



MPLS TE FRR introduces a few RSVP extensions for the signaling of the protected TE LSP, as follows:

- A new FAST_REROUTE object defines the characteristics for the backup TE LSP. These characteristics include priorities (setup and holding), hop limit, bandwidth, and attributes. The FAST_REROUTE object also specifies whether nodes should use facility backup or one-to-one backup to protect the TE LSP.
- The extended RECORD_ROUTE object indicates protection availability at each hop and its type (link, node, or bandwidth protection).
- The extended SESSION_ATTRIBUTE object signals whether the TE LSP desires protection and its type (link, node, or bandwidth protection).

Table 2-12 summarizes these extensions.

Table 2-12 *RSVP Objects Used for MPLS TE FRR*

RSVP Object	RSVP Message	FRR Function
FAST_REROUTE	Path	Specifies the desired FRR technique (facility backup or one-to-one backup) and the desired characteristics (priority, bandwidth, attributes, and so on) of the backup TE LSP
RECORD_ROUTE	Path, Resv	Records a list of hops/labels for the protected TE LSP, including protection status and type at each hop
SESSION_ATTRIBUTE	Path	Indicates whether the TE LSP requires protection and the type of protection

NOTE

The one-to-one backup technique uses an additional RSVP object (DETOUR).

MPLS TE FRR can use global or local restoration of the protected TE LSP as a result of a network failure. The global restoration approach relies on the headend rerouting the protected TE LSP. When the failure of a protected facility occurs, the PLR sends a PathErr message toward the headend of the protected TE LSP. In addition to the RSVP notification, the headend may also learn about the failure condition from IGP updates if the failure happens in the same IGP area. When the headend receives the failure notification, it can reroute the protected TE LSP permanently around the failure. When a PLR uses local restoration instead, it reroutes the protected TE LSPs through the backup while the failure persists. When the facility is back in service, the PLR resignals the protected TE LSP through its original path. Global restoration is more desirable as it relies on the headend to re-optimize the protected TE LSP. That node typically has a more complete view of the network resources and TE LSP constraints.

Link Protection

Link protection uses a backup TE LSP destined to the PLR *next hop* (NHOP). When a node signals a TE LSP with link protection desired, nodes along the path attempt to associate the TE LSP with a backup TE LSP to the NHOP downstream. The backup TE LSP could exist already, or the node may attempt to compute a suitable path and signal it. Any node that finds a TE backup LSP becomes a potential PLR and signals back to the protected TE LSP headend the protection availability at that location using the RECORD_ROUTE object. When a link fails, the PLR reroutes all the identified TE LSPs using the backup TE LSP. The rerouting process involves pushing the protected TE LSP label (as done before the failure) and then stacking the backup TE LSP label on top.

Figure 2-9 illustrates the operation of link protection. Node E signals a TE LSP toward node H, indicating in the SESSION_ATTRIBUTE that the TE LSP desires protection for link failures. When node F processes the object, it finds a suitable backup to the NHOP (node G) through node I. When the link between nodes F and G fails, node F detects the failure locally and modifies the output encapsulation of the protected TE LSP. It continues to push label 35 as expected by the NHOP and, in addition, it pushes label 16 to reroute the traffic through the backup TE LSP. Node I switches the backup TE LSP packets without any knowledge of the protected TE LSP. In this case, node I performs a PHP operation and the packets finally arrive at the MP (node G) with label 35 to continue toward node H.

Link protection can also protect against the failure of *shared-risk link groups* (SLRG). In some cases, multiple links in a network have a high probability of failing at the same time. Generally, these SRLGs are the result of multiple links sharing the same underlying infrastructure (Layer 2, Layer 1, or actual physical facilities). The path computation for the backup TE LSP should take into account these SRLGs to avoid using links that could fail at the same time as the protected link. PLRs can learn about SRLGs dynamically from IGP extensions or through local configuration. SRLGs affect the path computation that the PLR may perform the backup TE LSP. However, they do not impact the operation of link protection.

Node Protection

Node protection uses a backup TE LSP destined to the PLR next-next hop (NNHOP). When a node signals a TE LSP with node protection desired, nodes along the path attempt to associate it with a backup TE LSP to the NNHOP downstream. The backup TE LSP could exist already, or the node may attempt to compute a suitable path and signal it. Nodes that find a TE backup LSP become a potential PLR and signal back to the protected TE LSP headend the protection availability at their location using the RECORD_ROUTE object. When the NHOP fails, the PLR reroutes all the identified TE LSPs using the backup TE LSP. The rerouting process involves pushing the protected TE LSP label expected by the NNHOP and then stacking the TE backup LSP label on top. The PLR learns the NNHOP label from the RECORD_ROUTE object in Resv messages. Node protection can also protect against SRLG failures. As described in the previous section, SRLGs affect backup path computation but have no impact on the operation FRR, and in this case, node protection.

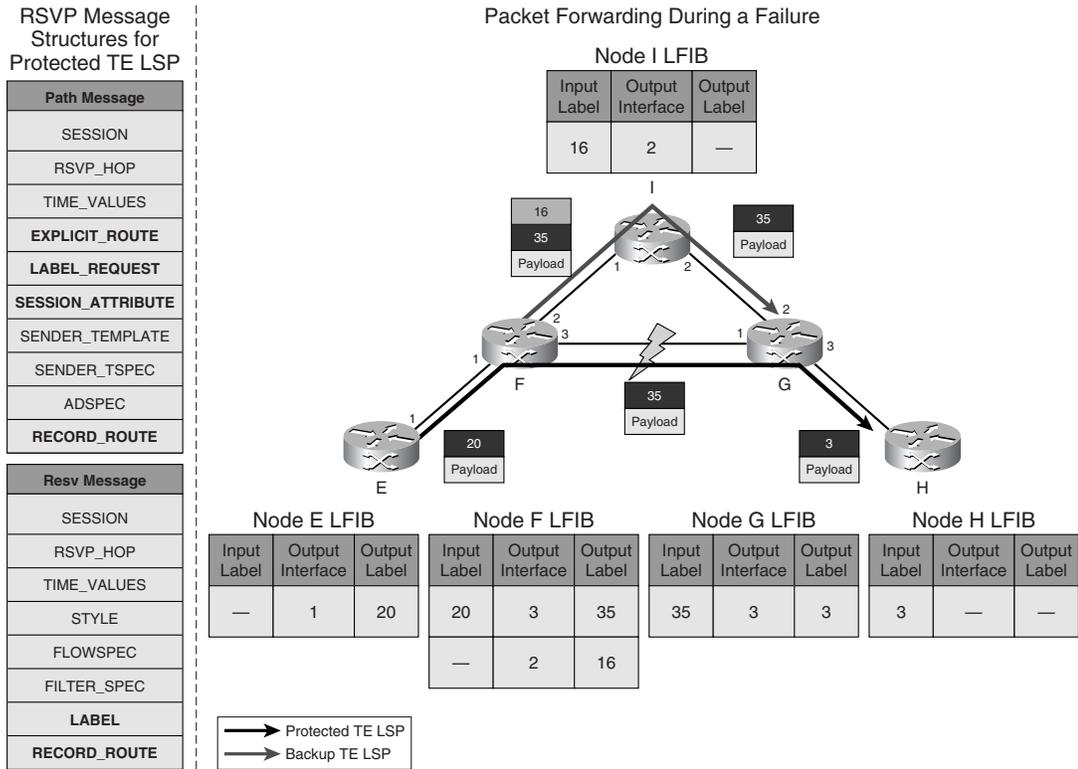
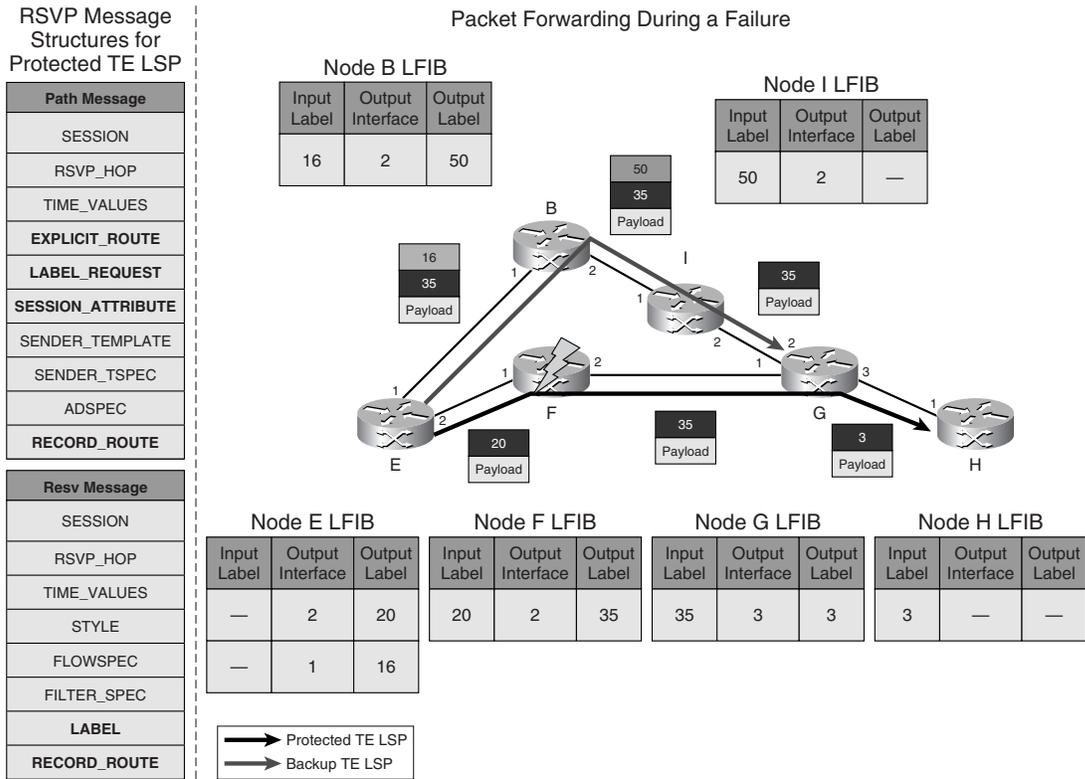
Figure 2-9 MPLS TE FRR Link Protection Operation

Figure 2-10 shows the operation of node protection. Node E signals a TE LSP toward node H, this time indicating in the SESSION_ATTRIBUTE that the TE LSP desires node protection. In this case, node E itself finds a suitable backup to the NNHOP (node G) through nodes B and I. When node F fails, node E detects the failure locally and modifies the output encapsulation of the protected TE LSP. Instead of pushing label 20 as performed before the failure, node E now pushes label 35 as expected by the node G and, in addition, it pushes label 16 to reroute the traffic through the backup TE LSP. Node B and I switch the backup TE LSP packets without any awareness of the protected TE LSP. Packets finally arrive at the MP (node G) with label 35 to continue toward node H.

Figure 2-10 MPLS TE FRR Node Protection Operation



Summary

MPLS provides native TE capabilities that can improve network efficiency and service guarantees. These MPLS TE capabilities bring explicit routing, constraint-based routing, and bandwidth reservation to MPLS networks. MPLS TE relies on extensions to existing IP protocols (IS-IS, OSPF, and RSVP). MPLS TE also supports its routing and bandwidth-reservation capabilities per class through the DS-TE extensions. DS-TE retains the same overall operation characteristics of MPLS TE but introduces minor protocol extensions to accommodate multiple classes. DS-TE enforces BCs on network links that complement the bandwidth allocation that DiffServ can provide in the forwarding plane provides. DS-TE can use two different BC models (MAM and RDM). Last, MPLS TE provides a fast protection mechanism for link and node failures using FRR. This mechanism relies on presignaled backup TE LSPs to provide fast protection (in milliseconds) in a scalable manner to other TE LSPs.

References

MPLS TE

Draft-ietf-mpls-soft-preemption-03 – MPLS Traffic Engineering Soft Preemption

Osborne, E., and A. Simha. *Traffic Engineering with MPLS*. Cisco Press; 2003.

RFC 3209, *RSVP-TE: Extensions to RSVP for LSP Tunnels*

RFC4090, *Fast Reroute Extensions to RSVP-TE for LSP Tunnels*

RFC 3784, *IS-IS extensions for Traffic Engineering*

RFC 3630, *Traffic Engineering (TE) Extensions to OSPF Version 2*

RFC 4124, *Protocol extensions for support of Differentiated-services-aware MPLS Traffic Engineering*

RFC 4125, *Maximum Allocation bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering*

RFC 4127, *Russian Dolls Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering*

RSVP

RFC 2205, *Resource Reservation Protocol (RSVP) -- Version 1 Functional Specification*

RFC 2209, *Resource Reservation Protocol (RSVP) -- Version 1 Message Processing Rules*

RFC 2747, *RSVP Cryptographic Authentication*

RFC 2961, *RSVP Refresh Overhead Reduction Extensions*

RFC 2996, *Format of the RSVP DCLASS Object*

RFC 3097, *RSVP Cryptographic Authentication—Updated Message Type Value*

RFC 3175, *Aggregation of RSVP for IPv4 and IPv6 Reservations*

RFC 3473, *Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions*