Multiprotocol Label Switching Overview

This chapter describes the Multiprotocol Label Switching (MPLS) distribution protocol. MPLS combines the performance and capabilities of Layer 2 (data link layer) switching with the proven scalability of Layer 3 (network layer) routing. It enables service providers to meet challenges brought about by explosive growth and provides the opportunity for differentiated services without necessitating the sacrifice of existing infrastructure.

The MPLS architecture is remarkable for its flexibility:

- Data can be transferred over any combination of Layer 2 technologies

- Support is offered for all Layer 3 protocols

- Scaling is possible well beyond anything offered in today's networks.

Specifically, MPLS can efficiently enable the delivery of IP services over an ATM switched network. It supports the creation of different routes between a source and a destination on a purely router-based Internet backbone. Service providers who use MPLS can save money and increase revenue and productivity.

Procedures for configuring MPLS are provided in the "Configuring Multiprotocol Label Switching" chapter later in this publication.

This chapter describes MPLS. It contains the following sections:

- MPLS/Tag Switching Terminology

- Benefits

- Label Functions

- Distribution of Label Bindings

- MPLS and Routing

- MPLS Traffic Engineering

- MPLS VPN Services

- MPLS Class of Service

- Label Switch Controller

MPLS/Tag Switching Terminology

Beginning with Cisco IOS Release 12.1, the Tag Switching distribution protocol has been replaced with the Multiprotocol Label Switching (MPLS) distribution protocol. MPLS supports the following:

- Tag Switching features

- Tag Switching command-line interface (CLI) commands

Table 13 lists the new MPLS standard terms which replace the old Tag Switching terms found in earlier releases of this document.

Table 13   MPLS Standard Terminology

Former Term	New Term
Tag Switching	Multiprotocol Label Switching (MPLS)
Tag (item or packet)	Label
Tag Switched	Label Switched
Tag Distribution Protocol (TDP)[1]	Label Distribution Protocol (LDP)
Tag Forwarding Information Base (TFIB)	Label Forwarding Information Base (LFIB)
Tag Switching Router (TSR)	Label Switching Router (LSR)
Tag Switch Controller (TSC)	Label Switch Controller (LSC)
ATM Tag Switch Router (ATM-TSR)	ATM Label Switch Router (ATM-LSR) For example, BPX 8650.
Tag Virtual Circuit, Tag VC, (TVC)	Label Virtual Circuit, Label VC (LVC)
Tag Switch Protocol (TSP)	Label Switch Protocol (LSP)
Extended Tag ATM Port (XTagATM)	Extended MPLS ATM Port (XmplsATM)

1. Cisco TDP and LDP (MPLS Label Distribution Protocol)) are nearly identical in function, but use incompatible message formats and some different procedures. Cisco will be changing from TDP to a fully compliant LDP.

Benefits

MPLS offers the following benefits:

- IP over ATM scalability—Enables service providers to keep up with Internet growth

- IP services over ATM—Brings Layer 2 benefits to Layer 3, such as traffic engineering capability

- Standards—Supports multivendor solutions

- Architectural flexibility—Offers choice of ATM or router technology, or a mix of both

Label Functions

In conventional Layer 3 forwarding, as a packet traverses the network, each router extracts all the information relevant to forwarding the packet from the Layer 3 header. This information is then used as an index for a routing table lookup to determine the packet's next hop.

In the most common case, the only relevant field in the header is the destination address field, but in some cases other header fields may also be relevant. As a result, the header analysis must be done independently at each router through which the packet passes, and a complicated lookup must also be done at each router.

In MPLS, the analysis of the Layer 3 header is done just once. The Layer 3 header is then mapped into a fixed length, unstructured value called a *label*.

Many different headers can map to the same label, as long as those headers always result in the same choice of next hop. In effect, a label represents a *forwarding equivalence class*—that is, a set of packets, which, however different they may be, are indistinguishable to the forwarding function.

The initial choice of label need not be based exclusively on the contents of the Layer 3 header; it can also be based on policy. This allows forwarding decisions at subsequent hops to be based on policy as well.

Once a label is chosen, a short label header is put at the front of the Layer 3 packet, so that the label value can be carried across the network with the packet. At each subsequent hop, the forwarding decision can be made simply by looking up the label. There is no need to re-analyze the header. Since the label is a fixed length and unstructured value, looking it up is fast and simple.

Distribution of Label Bindings

Each label switching router (LSR) makes an independent, local decision as to which label value is used to represent which forwarding equivalence class. This association is known as a *label binding*. Each LSR informs its neighbors of the label bindings it has made. This is done by means of the Label Distribution Protocol (LDP).

When a labeled packet is being sent from LSR A to neighboring LSR B, the label value carried by the packet is the label value that B assigned to represent the packet's forwarding equivalence class. Thus, the label value changes as the packet travels through the network.

MPLS and Routing

A label represents a forwarding equivalence class, but it does not represent a particular path through the network. In general, the path through the network continues to be chosen by the existing Layer 3 routing algorithms such as OSPF, Enhanced IGRP, and BGP. That is, at each hop when a label is looked up, the next hop chosen is determined by the dynamic routing algorithm.

MPLS Traffic Engineering

MPLS is an integration of Layer 2 and Layer 3 technologies. By making traditional Layer 2 features available to Layer 3, MPLS enables traffic engineering. Thus, you can offer in a one-tier network what now can be achieved only by overlaying a Layer 3 network on a Layer 2 network.

MPLS traffic engineering automatically establishes and maintains the tunnel across the backbone, using RSVP. The path used by a given tunnel at any point in time is determined based on the tunnel resource requirements and network resources, such as bandwidth.

Available resources are flooded via extensions to a link-state based Interior Gateway Protocol (IGP).

Tunnel paths are calculated at the tunnel head based on a fit between required and available resources (constraint-based routing). The IGP automatically routes the traffic into these tunnels. Typically, a packet crossing the MPLS traffic engineering backbone travels on a single tunnel that connects the ingress point to the egress point.

MPLS traffic engineering is built on the following Cisco IOS mechanisms:

- LSP tunnels, which are signalled through RSVP, with traffic engineering extensions. LSP tunnels are represented as Cisco IOS tunnel interfaces, have a configured destination, and are unidirectional.

- A link-state IGPs (such as IS-IS and OSPF) with extensions for the global flooding of resource information, and extensions for the automatic routing of traffic onto LSP tunnels as appropriate.

- An MPLS traffic engineering path calculation module that determines paths to use for LSP tunnels.

- An MPLS traffic engineering link management module that does link admission and bookkeeping of the resource information to be flooded.

- Label switching forwarding, which provides routers with a Layer 2-like ability to direct traffic across multiple hops as directed by the resource-based routing algorithm.

One approach to engineer a backbone is to define a mesh of tunnels from every ingress device to every egress device. The IGP, operating at an ingress device, determines which traffic should go to which egress device, and steers that traffic into the tunnel from ingress to egress. The MPLS traffic engineering path calculation and signalling modules determine the path taken by the LSP tunnel, subject to resource availability and the dynamic state of the network. For each tunnel, counts of packets and bytes sent are kept.

Sometimes, a flow is so large that it cannot fit over a single link, so it cannot be carried by a single tunnel. In this case multiple tunnels between a given ingress and egress can be configured, and the flow is load shared among them.

The following sections describe how conventional hop-by-hop link-state routing protocols interact with new traffic engineering capabilities. In particular, these sections describe how Dijkstra's shortest path first (SPF) algorithm has been adapted so that a link-state IGP can automatically forward traffic over tunnels that are set up by traffic engineering.

Mapping Traffic into Tunnels

Link-state protocols such as Integrated IS-IS use Dijkstra's SPF algorithm to compute a shortest path tree to all nodes in the network. Routing tables are derived from this shortest path tree. The routing tables contain ordered sets of destination and first-hop information. If a router does normal hop-by-hop routing, the first hop is a physical interface attached to the router.

New traffic engineering algorithms calculate explicit routes to one or more nodes in the network. These explicit routes are viewed as logical interfaces by the originating router. In the context of this document, these explicit routes are represented by LSPs and referred to as traffic engineering tunnels (TE tunnels).

The following sections describe how link-state IGPs can make use of these shortcuts, and how they can install routes in the routing table that point to these TE tunnels. These tunnels use explicit routes, and the path taken by a TE tunnel is controlled by the router that is the headend of the tunnel. In the absence of errors, TE tunnels are guaranteed not to loop, but routers must agree on how to use the TE tunnels. Otherwise traffic might loop through two or more tunnels.

Enhancement to the SPF Computation

During each step of the SPF computation, a router discovers the path to one node in the network. If that node is directly connected to the calculating router, the first-hop information is derived from the adjacency database. If a node is not directly connected to the calculating router, the node inherits the first-hop information from the parent(s) of that node. Each node has one or more parents and each node is the parent of zero or more downstream nodes.

For traffic engineering purposes, each router maintains a list of all TE tunnels that originate at this router. For each of those TE tunnels, the router at the tailend is known.

During the SPF computation, when a router finds the path to a new node, the new node is moved from the TENTative list to the PATHS list. The router must determine the first-hop information. There are three possible ways to do this:

1. Examine the list of tailend routers directly reachable by way of a TE tunnel. If there is a TE tunnel to this node, use the TE tunnel as the first-hop.

2. If there is no TE tunnel, and the node is directly connected, use the first-hop information from the adjacency database.

3. If the node is not directly connected, and is not directly reachable by way of a TE tunnel, the first-hop information is copied from the parent node(s) to the new node.

As a result of this computation, traffic to nodes that are the tailend of TE tunnels flows over those TE tunnels. Traffic to nodes that are downstream of the tailend nodes also flows over those TE tunnels. If there is more than one TE tunnel to different intermediate nodes on the path to destination node X, traffic flows over the TE tunnel whose tailend node is closest to node X.

Special Cases and Exceptions

The SPF algorithm finds equal-cost parallel paths to destinations. The enhancement previously described does not change this. Traffic can be forwarded over one or more native IP paths, over one or more TE tunnels, or over a combination of native IP paths and TE tunnels.

Figure 16 illustrates a special situation occurring in the topology.

Figure 16   Sample MPLS Topology Mapping Traffic into Tunnels
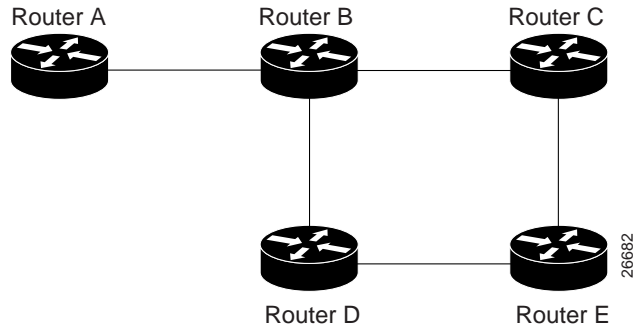
Assume that all links have the same cost and that a TE tunnel is set up from Router A to Router D. When the SPF calculation puts Router C on the TENTative list, it realizes that Router C is not directly connected. It uses the first-hop information from the parent, which is Router B. When the SPF calculation on Router A puts Router D on the TENTative list, it realizes that Router D is the tailend of a TE tunnel. Thus Router A installs a route to Router D by way of the TE tunnel, and not by way of Router B.

When Router A puts Router E on the TENTative list, it realizes that Router E is not directly connected, and that Router E is not the tailend of a TE tunnel. Therefore Router A copies the first-hop information from the parents (Router C and Router D) to the first-hop information of Router E.

Traffic to Router E now load-balances over the native IP path by way of Router A to Router B to Router C, and the TE tunnel Router A to Router D.

If parallel native IP paths and paths over TE tunnels are available, these implementations allow you to force traffic to flow over TE tunnels only or only over native IP paths.

Additional Enhancements to SPF Computation Using Configured Tunnel Metrics

When an IGP route is installed into a router information base (RIB) by means of TE tunnels as next hops, the distance or metric of the route must be calculated. Normally, you could make the metric the same as the IGP metric over native IP paths as if the TE tunnels did not exist. For example, Router A can reach Router C with the shortest distance of 20. X is a route advertised in IGP by Router C. Route X is installed in Router A's RIB with the metric of 20. When a TE tunnel from Router A to Router C comes up, by default the route is installed with a metric of 20, but the next-hop information for X is changed.

Although the same metric scheme can work well in other situations, for some applications it is useful to change the TE tunnel metric. For instance, when there are equal cost paths through TE tunnel and native IP links. You can adjust TE tunnel metrics to force the traffic to prefer the TE tunnel, to prefer the native IP paths, or to load share among them.

Again, suppose that multiple TE tunnels go to the same or different destinations. TE tunnel metrics can force the traffic to prefer some TE tunnels over others, regardless of IGP distances to those destinations.

Setting metrics on TE tunnels does not affect the basic SPF algorithm. It affects only two questions in two areas: (1) whether the TE tunnel is installed as one of the next hops to the destination routers, and (2) what the metric value is of the routes being installed into the RIB. You can modify the metrics for determining the first-hop information in the following instances:

- If the metric of the TE tunnel to the tailend routers is higher than the metric for the other TE tunnels or native hop-by-hop IGP paths, this tunnel is not installed as the next hop.

- If the metric of the TE tunnel is equal to the metric of either other TE tunnels or native hop-by-hop IGP paths, this tunnel is added to the existing next hops.

- If the metric of the TE tunnel is lower than the metric of other TE tunnels or native hop-by-hop IGP paths, this tunnel replaces them as the only next hop.

In each of the above cases, routes associated with those tailend routers and their downstream routers are assigned metrics related to those tunnels.

This mechanism is loop free because the traffic through the TE tunnel is basically source routed. The end result of TE tunnel metric adjustment is the control of traffic loadsharing. If there is only one way to reach the destination through a single TE tunnel, then no matter what metric is assigned, the traffic has only one way to go.

You can represent the TE tunnel metric in two different ways: (1) as an absolute (or fixed) metric or (2) as a relative (or floating) metric.

If you use an absolute metric, the routes assigned with the metric are fixed. This metric is used not only for the routes sourced on the TE tunnel tailend router, but also for each route downstream of this tailend router that uses this TE tunnel as one of its next hops.

For example, if you have TE tunnels to two core routers in a remote point of presence (POP), and one of them has an absolute metric of 1, all traffic going to that POP traverses this low-metric TE tunnel.

If you use a relative metric, the actual assigned metric value of routes is based on the IGP metric. This relative metric can be positive or negative, and is bounded by minimum and maximum allowed metric values. See Figure 17.

Figure 17   Sample MPLS Topology Using Configured Tunnel Metrics
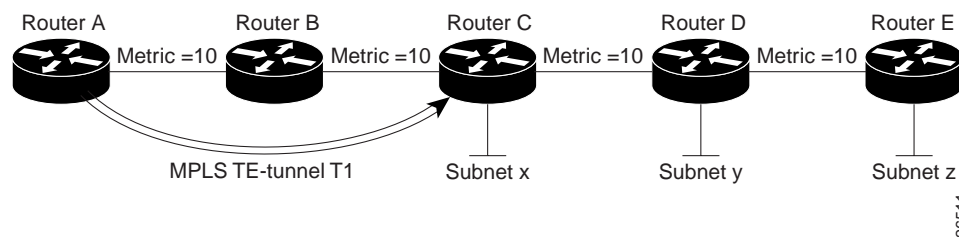
If there is no TE tunnel, Router A installs routes x, y, and z and assigns metrics 20, 30, and 40 respectively. Suppose that Router A has a TE tunnel T1 to Router C. If the relative metric -5 is used on tunnel T1, the routers x, y, and z have the installed metric of 15, 25, and 35. If an absolute metric of 5 is used on tunnel T1, routes x, y and z have the same metric 5 installed in the RIB for Router A. The assigning of no metric on the TE tunnel is a special case, a relative metric scheme where the metric is 0.

Migrating an IS-IS Network to a New Technology

This section discusses two different ways to migrate an existing IS-IS network from the standard ISO 10589 protocol, toward a new flavor of IS-IS with extensions.

New Extensions for the IS-IS Routing Protocol

Recently new extensions have been designed and implemented for the IS-IS routing protocol. The extensions serve multiple purposes.

One goal is to remove the 6-bit limit on link metrics. A second goal is to allow for inter-area IP routes. A third goal is to enable IS-IS to carry different kinds of information for the purpose of traffic engineering. In the future, more extensions might be needed.

To serve all these purposes, two new type, length, and value objects (TLVs) are defined. One TLV (TLV #22) describes links (or rather adjacencies). It serves the same purpose as the "IS neighbor option" in ISO 10589 (TLV #2). The second new TLV (TLV #135) describes reachable IP prefixes, similar to the IP Neighbor options from RFC 1195 (TLVs #128 and #130).

Both new TLVs have a fixed length part, followed by optional sub-TLVs. The metric space in these new TLVs has been enhanced from 6 bits to 24 or 32 bits. The sub-TLVs allow you to add new properties to links and prefixes. Traffic engineering is the first technology to make use of this ability to describe new properties of a link.

For the purpose of briefness, these two new TLVs, #22 and #135, are referred to as "new-style TLVs." TLVs #2, #128 and #130 are referred to as "old-style TLVs."

The Problem in Theory

Link-state routing protocols compute loop-free routes. This can be guaranteed because all routers calculate their routing tables based on the same information from the link-state database (LSPDB). The problem arises when some routers look at old-style TLVs and some routers look at new-style TLVs. In that case, the information on which they base their SPF calculation can be different. This different view of the world can cause routing loops among routers. Network administrators have to take great care to make sure that routers see the same view of the world.

The Problem in Practice

The easiest way to migrate from old-style TLVs towards new-style TLVs would be to introduce a "flag day." A flag day means you reconfigure all routers during a short period of time, during which service is interrupted. Assuming the implementation of a flag day is not an acceptable solution, a network administrator needs to find a viable solution for modern existing networks

Therefore, it becomes necessary to find a way to smoothly migrate a network from using IS-IS with old-style TLVs to IS-IS with new-style TLVs.

Another problem that arises and requires a solution is the need for new traffic engineering software to be tested in existing networks. Network administrators want the ability to test this software on a limited number of routers. They cannot upgrade all of their routers before they start testing—not in their production networks and not in their test networks.

The new extensions allow for a network administrator to use old-style TLVs in one area, and new-style in another area. However, this is not a solution for administrators that need or want to run their network in one single area.

Network administrators need a way to run an IS-IS network where some routers are advertising and using the new-style TLVs, and, at the same time, other routers are only capable of advertising and using old-style TLVs.

First Solution

One solution when you are migrating from old-style TLVs toward new-style TLVs is to advertise the same information twice—once in old-style TLVs and once in new-style TLVs. This ensures that all routers have the opportunity to understand what is advertised.

However, with this approach the two obvious drawbacks are as follows:

1. The size of the LSPs—During transition the LSPs grow roughly twice in size. This might be a problem in networks where the LSPDB is large. An LSPDB can be large because there are many routers and thus LSPs. Or the LSPs are large because of many neighbors or IP prefixes per router. A router that advertises a lot of information causes the LSPs to be fragmented.

   A large network in transition is pushing the limits regarding LSP flooding and SPF scaling. During transition you can expect some extra network instability. During this time, you especially do not want to test how far you can push an implementation. There is also the possibility that the traffic engineering extensions might cause LSPs to be reflooded more often. For a large network, this solution could produce unpredictable results.

2. The problem of ambiguity—If you choose this solution, you may get ambiguous answers to questions such as these:

   What should a router do if it encounters different information in the old-style TLVs and new-style TLVs?

This problem can be largely solved in an easy way by using:

- All information in old-style and new-style TLVs in an LSP.

- The adjacency with the lowest link metric if an adjacency is advertised more than once.

The main benefit is that network administrators can use new-style TLVs before all routers in the network are capable of understanding them.

Transition Steps During the First Solution

Here are some steps you can follow when transitioning from using IS-IS with old-style TLVs to new-style TLVs:

- Advertise and use only old-style TLVs if all routers run old software.

- Upgrade some routers to newer software.

- Configure some routers with new software to advertise both old-style and new-style TLVs. They accept both styles of TLVs. Configure other routers (with old software) to remain advertising and using only old-style TLVs.

- Test traffic engineering in parts of their network; however, wider metrics cannot be used yet.

- If the whole network needs to migrate, upgrade and configure all remaining routers to advertise and accept both styles of TLVs.

- Configure all routers to only advertise and accept new-style TLVs.

- Configure metrics larger than 63.

Second Solution

Routers advertise only one style of TLVs at the same time, but are able to understand both types of TLVs during migration.

One benefit is that LSPs stay roughly the same size during migration. Another benefit is that there is no ambiguity between the same information advertised twice inside one LSP.

The drawback is that all routers must understand the new-style TLVs before any router can start advertising new-style TLVs. So this transition scheme is useful when transitioning the whole network (or a whole area) to use wider metrics. It does not help the second problem, where network administrators want to use the new-style TLVs for traffic engineering, while some routers are still only capable of understanding old-style TLVs.

Transition Steps During the Second Solution

- Advertise and use only old-style TLVs if all routers run old software.
- Upgrade all routers to newer software.
- Configure all routers one-by-one to advertise old-style TLVs, but to accept both styles of TLVs.
- Configure all routers one-by-one to advertise new-style TLVs, but to accept both styles of TLVs.
- Configure all routers one-by-one to only advertise and to accept new-style TLVs.
- Configure metrics larger than 63.

Configuration Commands

Cisco IOS software has a new "router isis" command-line interface (CLI) subcommand called **metric-style**. Once you are in the router isis subcommand mode, you have the option to choose the following:

- **metric-style narrow**—Enables the router to advertise and accept only old-style TLVs
- **metric-style wide**—Enables the router to advertise and accept only new-style TLVs
- **metric-style transition**—Enables the router to advertise and accept both styles
- **metric-style narrow transition**—Enables the router to advertise old-style TLVs and accept both styles
- **metric-style wide transition**—Enables the router to advertise new-style TLVs and accept both styles

There are two transition schemes that you can employ using the metric-style commands. They are:

1. Narrow to transition to wide
2. Narrow to narrow transition to wide transition to wide

Cisco IOS software implements both transition schemes. Network administrators can choose the scheme that suits them best. For test networks, the first solution is ideal. For real transition, both schemes can be used. The first scheme requires less steps and thus less configuration. Only the largest of largest networks that do not want to risk doubling their LSPDB during transition need to use the second solution.

MPLS Virtual Private Networks

Using MPLS Virtual Private Networks (VPNs) in a Cisco IOS network provide the capability to deploy and administer scalable Layer 3 VPN backbone services including applications, data hosting network commerce, and telephony services to business customers. A VPN is a secure IP-based network that shares resources on one or more physical networks. A VPN contains geographically dispersed sites that can communicate securely over a shared backbone.

A one-to-one relationship does not necessarily exist between customer sites and VPNs; a given site can be a member of multiple VPNs. However, a site can associate with only one routing/forwarding instance (VRF). Each VPN is associated with one or more VPN VRFs. A VRF includes routing and forwarding tables and rules that define the VPN membership of customer devices attached to CE routers. A VRF consists of the following:

- IP routing table
- Cisco Express Forwarding (CEF) table
- Set of interfaces that use the CEF forwarding table
- Set of rules and routing protocol parameters to control the information in the routing tables

VPN routing information is stored in the IP routing table and the CEF table for each VRF. A separate set of routing and CEF tables is maintained for each VRF. These tables prevent information from being forwarded outside a VPN and also prevent packets that are outside a VPN from being forwarded to a router within the VPN.

MPLS VPN Services

MPLS VPNs allow service providers to deploy scalable VPNs and build the foundation to deliver value-added services, including:

- Connectionless service—A significant technical advantage of MPLS VPNs is that they are connectionless. The Internet owes its success to its basic technology, TCP/IP. TCP/IP is built on packet-based, connectionless network paradigm. This means that no prior action is necessary to establish communication between hosts, making it easy for two parties to communicate. To establish privacy in a connectionless IP environment, current VPN solutions impose a connection-oriented, point-to-point overlay on the network. Even if it runs over a connectionless network, a VPN cannot take advantage of the ease of connectivity and multiple services available in connectionless networks. When you create a connectionless VPN, you do not need tunnels and encryption for network privacy, thus eliminating significant complexity.

- Centralized service—Building VPNs in Layer 3 allows delivery of targeted services to a group of users represented by a VPN. A VPN must give service providers more than a mechanism for privately connecting users to intranet services. It must also provide a way to flexibly deliver value-added services to targeted customers. Scalability is critical, because customers want to use services privately in their intranets and extranets. Because MPLS VPNs are seen as private intranets, you may use new IP services such as:

  - Multicast
  - Quality of service (QoS)
  - Telephony support within a VPN
  - Centralized services including content and web hosting to a VPN

You can customize several combinations of specialized services for individual customers. For example, a service that combines IP multicast with a low-latency service class enables videoconferencing within an intranet.

- Scalability—If you create a VPN using connection-oriented, point-to-point overlays, Frame Relay, or ATM virtual connections (VCs), the VPN's key deficiency is scalability. Specifically, connection-oriented VPNs without fully meshed connections between customer sites, are not optimal. MPLS-based VPNs instead use the peer model and Layer 3 connectionless architecture to leverage a highly scalable VPN solution. The peer model requires a customer site to only peer with one provider edge (PE) router as opposed to all other CPE or customer edge (CE) routers that are members of the VPN. The connectionless architecture allows the creation of VPNs in Layer 3, eliminating the need for tunnels or VCs.

  Other scalability issues of MPLS VPNs are due to the partitioning of VPN routes between PE routers and the further partitioning of VPN and IGP routes between PE routers and provider (P) routers in a core network.

  - PE routers must maintain VPN routes for those VPNs who are members.

  - P routers do not maintain any VPN routes.

  This increases the scalability of the provider's core and ensures that no one device is a scalability bottleneck.

- Security—MPLS VPNs offer the same level of security as connection-oriented VPNs. Packets from one VPN do not inadvertently go to another VPN. Security is provided

  - At the edge of a provider network, ensuring packets received from a customer are placed on the correct VPN.

  - At the backbone, VPN traffic is kept separate. Malicious spoofing (an attempt to gain access to a PE router) is nearly impossible because the packets received from customers are IP packets. These IP packets must be received on a particular interface or subinterface to be uniquely identified with a VPN label.

- Easy to create—To take full advantage of VPNs, it must be easy for customers to create new VPNs and user communities. Because MPLS VPNs are connectionless, no specific point-to-point connection maps or topologies are required. You can add sites to intranets and extranets and form closed user groups. When you manage VPNs in this manner, it enables membership of any given site in multiple VPNs, maximizing flexibility in building intranets and extranets.

- Flexible addressing—To make a VPN service more accessible, customers of a service provider can design their own addressing plan, independent of addressing plans for other service provider customers. Many customers use private address spaces and do not want to invest the time and expense of converting to public IP addresses to enable intranet connectivity. MPLS VPNs allow customers to continue to use their present address spaces without network address translation (NAT) by providing a public and private view of the address. A NAT is required only if two VPNs with overlapping address spaces want to communicate. This enables customers to use their own unregistered private addresses, and communicate freely across a public IP network.

- Integrated Class of Service (CoS) support—CoS is an important requirement for many IP VPN customers. It provides the ability to address two fundamental VPN requirements:

  - Predictable performance and policy implementation

 - Support for multiple levels of service in a MPLS VPN

  Network traffic is classified and labeled at the edge of the network before traffic is aggregated according to policies defined by subscribers and implemented by the provider and transported across the provider core. Traffic at the edge and core of the network can then be differentiated into different classes by drop probability or delay.

- Straightforward migration—For service providers to quickly deploy VPN services, use a straightforward migration path. MPLS VPNs are unique because you can build them over multiple network architectures, including IP, ATM, Frame Relay, and hybrid networks.

  Migration for the end customer is simplified because there is no requirement to support MPLS on the CE router and no modifications are required to a customer's intranet.

  Figure 18 shows an example of a VPN with a service provider (P) backbone network, service provider edge routers (PE), and customer edge routers (CE).

Figure 18   VPNs with a Service Provider Backbone
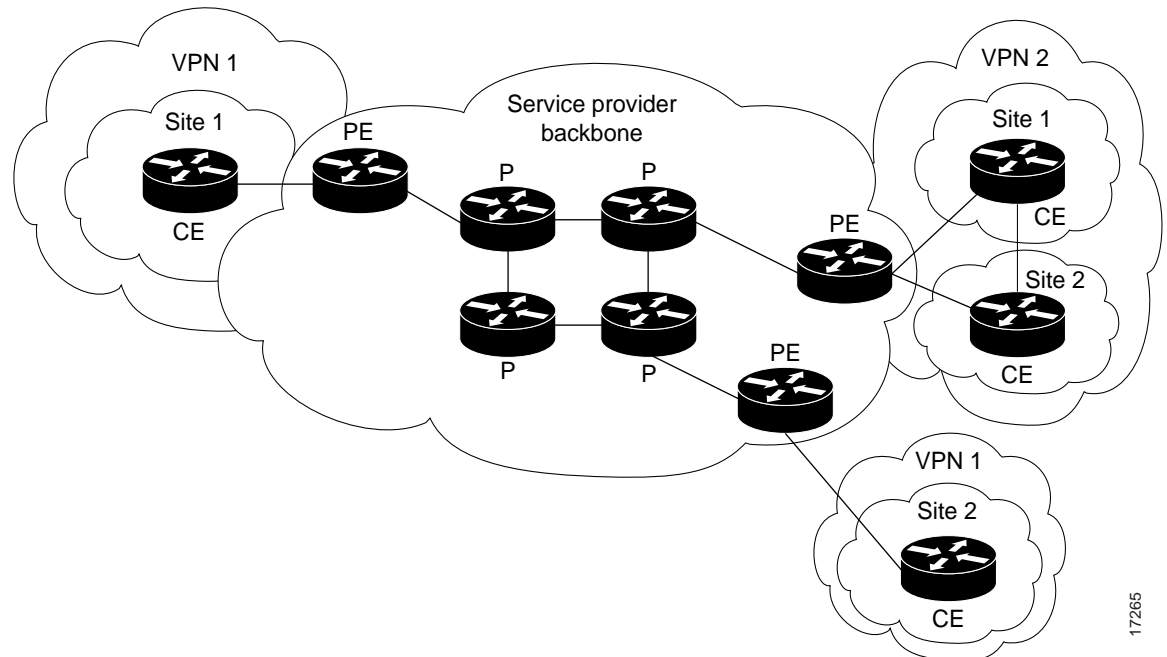
A VPN contains customer devices attached to the CE routers. These customer devices use VPNs to exchange information between devices. Only the PE routers are aware of the VPNs.

Figure 19 shows five customer sites communicating within three VPNs. The VPNs can communicate with the following sites:

- VPN1—Sites 2 and 4
- VPN2—Sites 1, 3, and 4
- VPN3—Sites 1,3, and 5

Figure 19   Customer Sites within VPNs
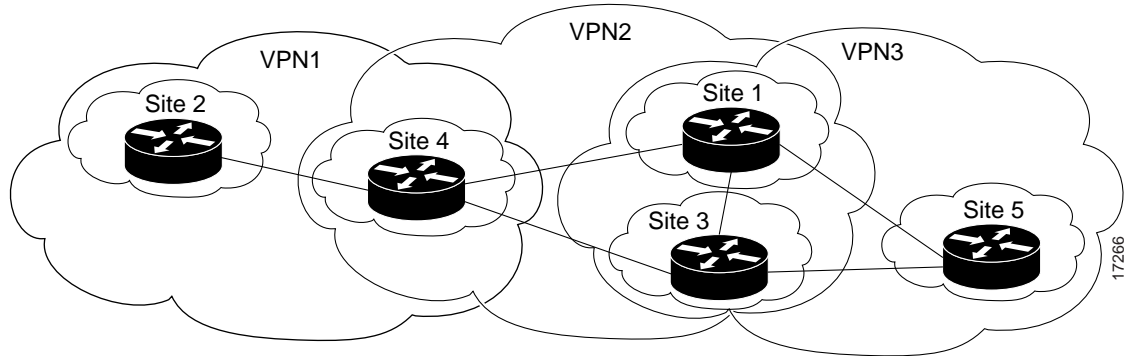
Increased BGP Functionality

- Configuring BGP hub and spoke connections—Configuring PE routers in a hub and spoke configuration allows a CE router to readvertise all prefixes containing duplicate autonomous system numbers (ASNs) to neighboring PE routers. Using duplicate ASNs in a hub and spoke configuration provides faster convergence of routing information within geographically dispersed locations.

- Configuring faster convergence for BGP VRF routes—Configuring scanning intervals of BGP routers decreases import processing time of VPNv4 routing information, thereby providing faster convergence of routing information. Routing tables are updated with routing information about VPNv4 routes learned from PE routers or route reflectors.

- Limiting VPN VRFs—Limiting the number of routes in a VRF prevents a PE router from importing too many routes, thus diminishing the router's performance. This enhancement can also be used to enforce the maximum number of members that can join a VPN from a particular site. A threshold is set in the VRF routing table to limit the number of VRF routes imported.

- Reuse ASNs in an MPLS VPN environment—Configuring a PE router to reuse an existing ASN allows customers to configure BGP routes with the same ASNs in multiple geographically dispersed sites, providing better scalability between sites.

- Distributing BGP OSPF routing information—Setting a separate router ID for each interface or subinterface on a PE router attached to multiple CE routers within a VPN provides increased flexibility through OSPF when routers exchange routing information between sites.

Table 14 lists the MPLS VPN features and the associated BGP commands.

Table 14   MPLS VPN Features and the Associated BGP Commands

Name of Feature	Command	Description
Configuring Faster Convergence for BGP VPN Routing/Forwarding Instance (VRF) Routes	**bgp scan-time import**	Configures scanning intervals of BGP routers to decrease import processing time of routing information.
Limiting VRF Routes	**maximum routes**	Limits the number of routes in a VRF to prevent a PE router from importing too many routes.

Table 14   MPLS VPN Features and the Associated BGP Commands (continued)

Name of Feature	Command	Description
Configuring Border Gateway Protocol (BGP) Hub and Spoke Connections	**neighbor allowas-in**	Configures provider edge (PE) routers to allow customer edge (CE) routers to readvertise all prefixes that contain duplicate autonomous system numbers (ASNs) to neighboring PE routers.
Reusing ASNs in an MPLS VPN Environment	**neighbor as-override**	Configures a PE router to reuse the same ASN on all sites within an MPLS VPN by overriding private ASNs.
Distributing BGP Open Shortest Path First (OSPF) Routing Information	**set ospf router-id**	Sets a separate router ID for each interface or subinterface on the PE router for each directly attached CE router.

VPN Operation

Each VPN is associated with one or more VPN routing/forwarding instances (VRFs). A VRF defines the VPN membership of a customer site attached to a PE router. A VRF consists of an IP routing table, a derived Cisco Express Forwarding (CEF) table, a set of interfaces that use the forwarding table, and a set of rules and routing protocol parameters that control the information that is included into the routing table.

A one-to-one relationship does not necessarily exist between customer sites and VPNs. A given site can be a member of multiple VPNs, as shown in Figure 19. However, a site can only associate with one (and only one) VRF. A customer site's VRF contains all the routes available to the site from the VPNs of which it is a member.

Packet forwarding information is stored in the IP routing table and the CEF table for each VRF. A separate set of routing and CEF tables is maintained for each VRF. These tables prevent information from being forwarded outside a VPN, and also prevent packets that are outside a VPN from being forwarded to a router within the VPN.

Distribution of VPN Routing Information

The distribution of VPN routing information is controlled through the use of VPN route target communities, implemented by border gateway protocol (BGP) extended communities. Distribution of VPN routing information works as follows:

- When a VPN route learned from a CE router is injected into BGP, a list of VPN route target extended community attributes is associated with it. Typically the list of route target community values is set from an export list of route targets associated with the VRF from which the route was learned.

- An import list of route target extended communities is associated with each VRF. The import list defines route target extended community attributes that a route must have in order for the route to be imported into the VRF. For example, if the import list for a particular VRF includes route target communities A, B, and C, then any VPN route that carries any of those route target extended communities—A, B, *or* C—is imported into the VRF.

BGP Distribution of VPN Routing Information

A service provider edge (PE) router can learn an IP prefix from a customer edge (CE) router by static configuration, through a BGP session with the CE router, or through the routing information protocol (RIP) exchange with the CE router. The IP prefix is a member of the IPv4 address family. After it learns the IP prefix, the PE converts it into a VPN-IPv4 prefix by combining it with an 8-byte route distinguisher (RD). The generated prefix is a member of the VPN-IPv4 address family. It uniquely identifies the customer address, even if the customer site is using globally nonunique (unregistered private) IP addresses.

The route distinguisher used to generate the VPN-IPv4 prefix is specified by a configuration command associated with the VRF on the PE router.

BGP distributes reachability information for VPN-IPv4 prefixes for each VPN. BGP communication takes place at two levels: within IP domains, known as autonomous systems (interior BGP or IBGP) and between autonomous systems (external BGP or EBGP). PE-PE or PE-RR (route reflector) sessions are IBGP sessions, and PE-CE sessions are EBGP sessions.

BGP propagates reachability information for VPN-IPv4 prefixes among PE routers by means of the BGP multiprotocol extensions, which define support for address families other than IPv4. It does this in a way that ensures the routes for a given VPN are learned only by other members of that VPN, enabling members of the VPN to communicate with each other.

MPLS Forwarding

Based on routing information stored in the VRF IP routing table and VRF CEF table, packets are forwarded to their destination using MPLS.

A PE router binds a label to each customer prefix learned from a CE router and includes the label in the network reachability information for the prefix that it advertises to other PE routers. When a PE router forwards a packet received from a CE router across the provider network it labels the packet with the label learned from the destination PE router. When the destination PE router receives the labeled packet it pops the label and uses it to direct the packet to the correct CE router. Label forwarding across the provider backbone, is based on either dynamic label switching or traffic engineered paths. A customer data packet carries two levels of labels when traversing the backbone:

- Top label directs the packet to the correct PE router
- Second label indicates how that PE router should forward the packet to the CE router

MPLS Class of Service

The Class of Service (CoS) feature for Multiprotocol Label Switching (MPLS) enables network administrators to provide differentiated types of service across an MPLS network. Differentiated service satisfies a range of requirements by supplying for each packet transmitted the particular kind of service specified for that packet by its CoS. Service can be specified in different ways, for example, using the IP precedence bit settings in IP packets.

In supplying differentiated service, MPLS CoS offers packet classification, congestion avoidance, and congestion management. Table 15 lists these functions and their descriptions.

Table 15   CoS Services and Features

Service	CoS Function	Description
Packet classification	Committed access rate (CAR). Packets are classified at the edge of the network before labels are assigned.	CAR uses the type of service (TOS) bits in the IP header to classify packets according to input and output transmission rates. CAR is often configured on interfaces at the edge of a network in order to control traffic into or out of the network. You can use CAR classification commands to classify or reclassify a packet.
Congestion avoidance	Weighted random early detection (WRED). Packet classes are differentiated based on drop probability.	WRED monitors network traffic, trying to anticipate and prevent congestion at common network and internetwork bottlenecks. WRED can selectively discard lower priority traffic when an interface begins to get congested. It can also provide differentiated performance characteristics for different classes of service.
Congestion management	Weighted fair queueing (WFQ). Packet classes are differentiated based on bandwidth and bounded delay.	WFQ is an automated scheduling system that provides fair bandwidth allocation to all network traffic. WFQ uses weights (priorities) to determine how much bandwidth each class of traffic is allocated.

For more information on configuration of the CoS functions (CAR, WRED, and WFQ), see the *Cisco IOS Quality of Service Solutions Configuration Guide*.

For complete command syntax information for CAR, WRED, and WFQ, see the *Cisco IOS Quality of Service Solutions Command Reference.*

MPLS CoS lets you duplicate Cisco IOS IP CoS (Layer 3) features as closely as possible in MPLS devices, including label edge routers (LERs), label switch routers (LSRs), and asynchronous transfer mode LSRs (ATM LSRs). MPLS CoS functions map nearly one-for-one to IP CoS functions on all interface types.

Label Switch Controller

The Label Switch Controller (LSC) with Cisco's BPX 8620 wide area switch and BPX 8650 IP+ATM switch delivers scalable integration of IP services over an ATM network.

The LSC enables the BPX 8620 and 8650 to:

- Participate in a MPLS network
- Directly peer with IP edge routers
- Support the full suite of IP features available in Cisco IOS

The Label Switch Controller (LSC) is a label switch router (LSR) that controls the operation of a separate ATM switch. Together, the router and ATM switch function as a single ATM MPLS router (ATM-LSR). A Cisco 7200 or 7500 series router acts as the LSC, and a Cisco BPX 8600 Service Node or a partner's switch acts as the VSI-controlled ATM switch. The LSC controls the ATM switch using the Cisco Virtual Switch Interface (VSI), which runs over an ATM link connecting the two.

MPLS's highly scalable IP+ATM integration is created by the LSC using a direct peer relationship between the BPX 8620 or 8650 and IP edge routers. This removes the limit placed on the number of IP edge routers, seen in traditional IP-over-ATM networks, allowing service providers to keep pace with the growing demand for IP services. The LSC also supports the easy, quick, and direct implementation of advanced IP services over ATM networks with BPX 8620s and 8650s.

The combination of a LSC and the ATM switch it controls is shown in Figure 20.

Figure 20   Label Switch Controller and Controlled ATM Switch
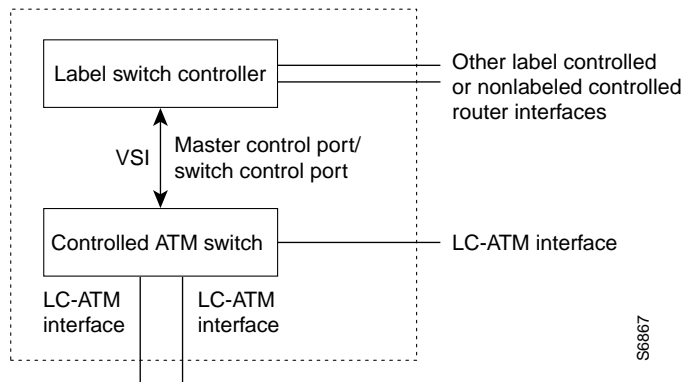
In Figure 20, the dotted line represents the external interface of the LSC and controlled switch as seen in the IP routing topology. The controlled ATM switch shows one or more TC-ATM interfaces at this external interface and the LSC itself may have additional interfaces that may or may not be label controlled.