Technology Overview

MPLS

MPLS was viewed until recently as a service provider routing technology. Next generation enterprise networks relying on intelligent network infrastructure for solutions such as IP telephony, storage, wireless, and the applications and services that surround them demand network resiliency and features no less than and at times exceeding what is available in service provider networks. Using MPLS to build VPNs in enterprise networks addresses new requirements such as network segmentation, extending segmentation across campuses, address transparency, and shared services in the most scalable way while leveraging the benefits and flexibility of IP. MPLS application components include Layer 3 VPNs, Layer 2 VPNs, QoS, and Traffic Engineering. The following sections focus on Layer 3 and Layer 2 VPNs as these are the key applications for Enterprise networks.

MPLS Layer 3 VPNs

The following components perform a specific role in successfully building an MPLS VPN network:

- Interior Gateway Protocol (IGP)—This routing protocol is used in an MPLS core to learn about internal routes.

  The IGP table is the global routing table that includes routes to the provider edge (PE) routers or any provider (P) router. Note that these routes are not redistributed into the VPN (external site) routes.

  Although any routing protocol including static routes can be used in the MPLS core, using a dynamic routing protocol such as EIGRP or OSPF that gives sub-second convergence is more desirable. If the customer is required to support MPLS Traffic Engineering applications, then a link-state protocol such as OSPF or IS-IS is required.

- Cisco Express Forwarding table—Derived from FIB and LFIB tables and used to forward VPN traffic.

- Label Distribution Protocol (LDP)—Tag Distribution Protocol (TDP) is the precursor to LDP and was invented by Cisco Systems. TDP is a proprietary protocol. TDP and LDP use the same label format but the message format is different.

  LDP supports the following features that are not part of TDP:

  - Extension mechanism for vendor-private and experimental features
  - Backoff procedures for session establishment failures
  - Abort mechanism for LSP setup
  - Optional use of TCP MD5 signature option for (more) secure operation

- Optional path-vector-based loop detection mechanism to prevent setup of looping LSPs
- More combinations of modes of operation

It is recommended to use LDP when possible.

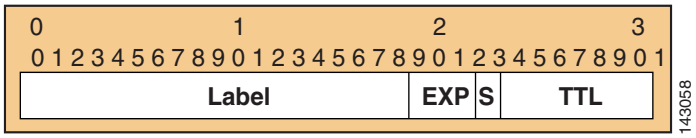
LDP is enabled on the core routers (P) and the PE core-facing interfaces to generate and distribute labels appropriately for the prefixes that were learned from the core IGP.

After the IGP in the core converges, labels are generated and bound to these prefixes and kept in the table that is referenced when the packets are forwarded. This table is called the Label Forwarding Information Base (LFIB).

Note that packets are switched based on pre-calculated labels and not routed through an MPLS core. Ingress Edge LSR (PE) appends a label before forwarding a packet to its neighbor LSR(P). The neighbor LSR swaps incoming label with outgoing label and forwards the packet to its neighbor. If this neighbor is Egress Edge LSR(PE), the core LSR(P) pops the label to avoid double look up (MPLS and IP) at the Egress Edge LSR and forwards the packet as an IP packet. This action of removing the label one hop prior to reaching the egress LSR is called Penultimate Hop Popping (PHP).

- MPLS labels and label stacking—The MPLS label is 32 bits and is used as a shim header in the forwarding plane (See Figure 2-1). Each application is responsible for generating labels. Labels generated in the control plane are used in the forwarding plane and encapsulated between Layer 2 and Layer 3 packet headers.

Figure 2-1 MPLS Labels



- Label = 20 bits
- CoS/EXP = Class of Service, 3 bits
- S = Bottom of stack, 1 bit
- TTL = Time to live, 8 bits

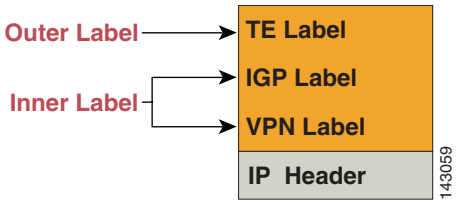
Label stacking occurs if more than one MPLS application is present in the network. For example:

- IGP labels (outer label if TE is not in use)—Used for forwarding packets in the core
- BGP labels (inner label)—Used for assigning end users/communities
- RSVP labels (outer label)—Used for TE tunnels

If TE, VPN, and MPLS are enabled, the headend and the tailend LSR are in charge of applying and removing TE label. If not, IGP is the outer most label and since PHP is on by default, an LSR one hop prior to the Egress Edge LSR removes this outer label. For VPN traffic, Ingress Edge LSR appends VPN label and Egress Edge LSR removes VPN label.

MPLS label stacking in Figure 2-2 demonstrates label stacking when multiple MPLS services are enabled in a network.

Figure 2-2 MPLS Label Stacking Example



- CE-PE routing protocols—Distribute VPN site routing information to a PE that is adjacently connected to a VPN site. Any of the EGP and IGP routing protocols including static, RipV2, EIGRP, OSPF, IS-IS, or eBGP are supported.
- Route distinguisher (RD)—A PE acquires knowledge about routes for multiple VPNs through a single BGP process. This process builds a BGP table containing prefixes belonging to VPNs that can possibly have overlapping address spaces. To enforce uniqueness of the prefixes held in the BGP table, a 64-bits per-VRF RD is prepended to the IP prefixes. The RD helps keep routes separate for VPNs so that customers in different VPNs do not see each others routes. It is a good practice to use the same RD for a VPN on all PEs.
- Route target (RT)—For each route, MP-BGP carries an extended community attribute (RT) that determines who is allowed to import or export that router. When a PE builds a VRF, it imports only the BGP routes that have the specific RT it is configured to import. Similarly, a VRF tells MP-BGP which RT value to use to advertise its routes, known as exporting routes.

  For intranet routing, a VPN should export and import the same RT for an optimal memory use by the BGP table. For extranet or overlapping VPNs, one VPN imports the RT of another VPN and vice versa. Route maps may be used to further tune which routes to import/export. Access to a service can also be announced using a dedicated RT.
- Virtual routing forwarding instance (VRF)—PEs use a separate routing table called VRF per-VPN. A VRF is built by using information from the BGP table built by MP-iBGP updates.
- Multi Protocol BGP (MP-iBGP) as described in RFC 2547 to create Layer 3 VPNs— Conventional BGP is designed to carry routing information for the IPv4 address family. MP-iBGP includes multi-protocol extensions such as RD, RT, and VPN label information as part of the Network Layer Reachability Information (NLRI).

  MP-iBGP carries VPNv4 routes from an ingress PE and relays it to an egress PE. At the ingress PE, MP-iBGP allocates labels for VPNv4 prefixes and installs them in the LFIB table. At the egress PE, MP-iBGP installs VPN prefixes and labels in the VRF FIB table. The associated label is appended to a packet when a packet within that VPN needs to be forwarded from the ingress PE to an egress PE. Note that this is an inner label. At the ingress PE, the outer label is derived from IGP plus LDP and is used as an outer header to switch the traffic from the ingress PE to the egress PE.
- Multi-VRF— Also known as VRF-Lite, this is a lightweight segmentation solution and a label-free solution that works without LDP and MP-iBGP. To keep traffic separate for each VPN segment, a VRF is created and mapped to a pair of ingress/egress Layer 3 interfaces. Routing information for each VPN is kept in its associated VRF instance.

Multipath Load Balancing

Current networks more commonly have redundant links and devices to the same destination. It is essential that traffic use multiple available paths for better traffic load balancing. Several load balancing mechanisms are available, including Cisco IOS software-based mechanisms such as Cisco Express Forwarding, unequal cost load balancing, OER, GLBP, eBGP, iBGP, and PBR. Per flow or per packet

Cisco Express Forwarding is employed to use multiple links on a router. Per destination (per flow) is the recommended method because per packet can introduce packet reordering and non-predictive latency within a session.

IGP can easily load balance based on the path metrics. Because BGP does not have a metric, load balancing over BGP paths becomes more challenging. Typically, BGP chooses only one best path using the complicated path selection algorithm and installs this path in the routing table. In addition, unlike what happens in IGPs, next hops of BGP routes may not be directly connected. eBGP and iBGP multipath mechanisms allow the installation of multiple BGP next hops in the routing tables (VRF routing tables).

There are two types of BGP multipath mechanisms available: eBGP and iBGP. eBGP multipath is used if a network has multiple exit points to a VPN site (destination in a VPN site). If eBGP is used to connect VPN sites to an MPLS cloud, eBGP multipath can be used on a CPE facing the cloud and iBGP multipath load balancing on PEs facing VPN sites. If IGP is used to connect VPN sites to an MPLS cloud, iBGP multipath load balancing on PEs suffices. iBGP multipath is used for dual-homed PEs. If at the egress point traffic can be sent via two PEs to the VPN site, the ingress PE needs to load balance using both exit points. iBGP multipath load balancing works with both equal cost and unequal cost paths to the destination (egress) PEs.

Note that the BGP multipath mechanism does not interfere with the BGP best path selection process; it installs multiple paths, but designates one of the paths as the best path. Multiple paths are installed in both RIB and FIB tables. Unequal cost paths are used proportionally to the link bandwidth. For BGP multipaths to work, all the path selection attributes such as weight, local preference, AS path, origin code, multi-exit discriminator (MED), and IGP distance need to be identical for both paths.

Route reflectors reflect only the best path to their clients. If the route reflectors are used to get multiple paths reflected to all the PEs, it is essential to use a unique RD value for the same VPN on each ingress PE for a VRF. Note that this does not affect RT values. For a fully-meshed VRF, the same RTs can be used for both importing and exporting VPN routes.

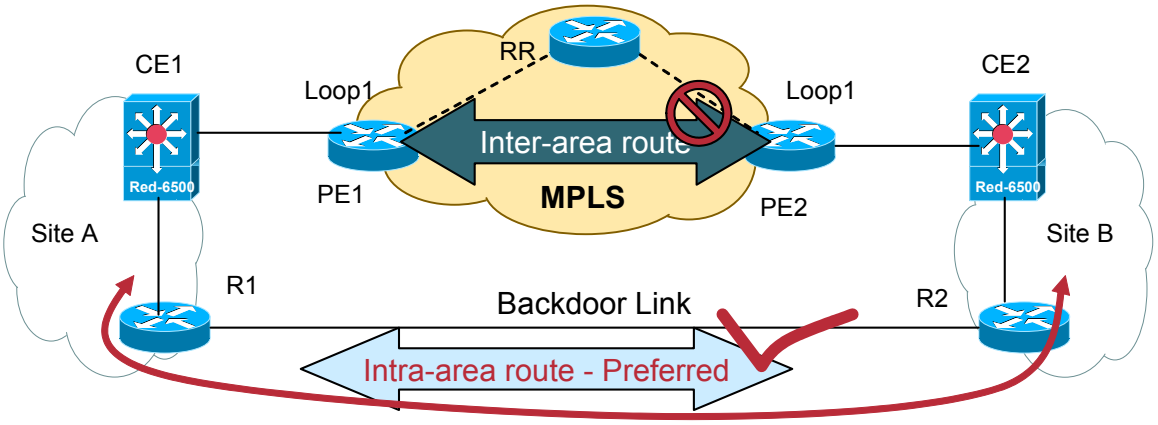
OSPF as the PE-CE Routing Protocol

When OSPF is used to connect PE and CE routers, all routing information learned from a VPN site is placed in the VPN routing and forwarding (VRF) instance associated with the incoming interface. The PE routers that attach to the VPN use BGP to distribute VPN routes to each other. When OSPF routes are propagated over the MPLS VPN backbone, additional information about the prefix in the form of BGP extended communities (route type, domain ID extended communities) is appended to the BGP update. This community information is used by the receiving PE router to decide the type of link-state advertisement (LSA) to be generated when the BGP route is redistributed to the OSPF PE-CE process. In this way, internal OSPF routes that belong to the same VPN and are advertised over the VPN backbone are seen as interarea routes on the remote sites.

OSPF and Backdoor Links

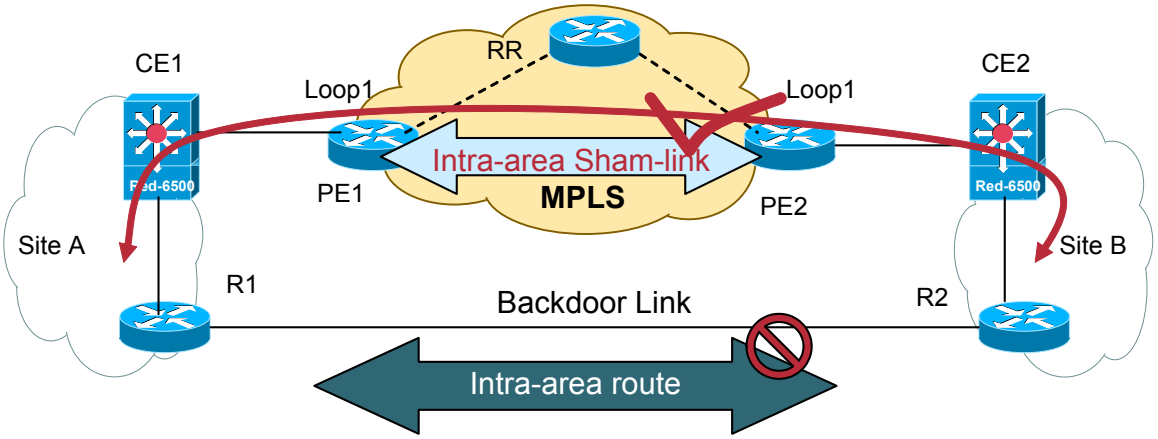
Although OSPF PE-CE connections assume that the only path between two sites is across the MPLS VPN backbone, backdoor paths between VPN sites (shown in Figure 2-3) may exist. If these sites belong to the same OSPF area, the path over a backdoor link is always selected because OSPF prefers intra-area paths to inter-area paths. (PE routers advertise OSPF routes learned over the VPN backbone as inter-area paths.) For this reason, OSPF backdoor links between VPN sites must be taken into account so that routing is performed based on policy.

Figure 2-3 OSPF—Backdoor Link Without Sham Link Support



If the backdoor links between sites are used only for backup purposes and do not participate in the VPN service, then the default route selection is not acceptable. To reestablish the desired path selection over the MPLS VPN backbone, create an additional OSPF intra-area (logical) link between ingress and egress VRFs on the relevant PE routers. This link is called a sham-link. A cost is configured with each sham-link and is used to decide whether traffic is sent over the backdoor path or the sham-link path. When a sham-link is configured between PE routers, the PEs can populate the VRF routing table with the OSPF routes learned over the sham-link.

Figure 2-4 OSPF—Backdoor Link With Sham Link Support



Because the sham-link is seen as an intra-area link between PE routers, an OSPF adjacency is created and database exchange (for the particular OSPF process) occurs across the link. The PE router can then flood LSAs between sites from across the MPLS VPN backbone. As a result, the desired intra-area connectivity is created.

Before you create a sham-link between PE routers in an MPLS VPN, you must:

1. Configure a separate /32 address on the remote PE so that OSPF packets can be sent over the VPN backbone to the remote end of the sham-link. You can use the /32 address for other sham-links.The /32 address must meet the following criteria:
   - Belong to a VRF
   - Not be advertised by OSPF

  - Be advertised by BGP

2. Associate the sham-link with an existing OSPF area.

EIGRP as PE-CE Routing Protocol

When EIGRP is used as the PE-CE protocol, EIGRP metrics are preserved across the MPLS VPN backbone though use of MP-BGP extended community attributes. The EIGRP route type and vector metric information is encoded in a series of well-known attributes. These attributes are transported across the MPLS VPN backbone and used to recreate the EIGRP route when received by the target PE router. There are no EIGRP adjacencies, EIGRP updates, or EIGRP queries sent across the MPLS VPN backbone. Only EIGRP metric information is carried across the MPLS VPN backbone via the MP-BGP extended communities.

Routes are recreated by the PE router and sent to the CE router as an EIGRP route. The same route type and cost basis as the original route are used to recreate the EIGRP route. The metric of the recreated route is increased by the link cost of the interface. On the PE router, if a route is received via BGP and the route has no extended community information for EIGRP, the route is advertised to the customer edge router as an external EIGRP route using the default metric. If no default metric is configured, the route is not advertised to the customer edge router.

EIGRP and Backdoor Links

The SoO extended community is a BGP extended community attribute that is used to identify routes that have originated from a site so that the re-advertisement of that prefix back to the source site can be prevented. The SoO extended community uniquely identifies the site from which a PE router has learned a route. SoO support provides the capability to filter MPLS VPN traffic on a per-EIGRP site basis.

If all of the routers in the customer's sites between the provider edge routers and the backdoor routers support the SoO feature, and the SoO values are defined on both the provider edge routers and the backdoor links, the provider edge routers and the backdoor routers all play a role in supporting convergence across the two (or more) sites. Routers that are not provider edge routers or backdoor routers must only propagate the SoO value on routes as they forward them to their neighbors, but they play no other role in convergence beyond the normal dual-attachment stations. The next two sections describe the operation of the PE routers and backdoor routers in this environment.

PE Router Operations

When this SoO is enabled, the EIGRP routing process on the PE router checks each received route for the SoO extended community and filters based on the following conditions:

- A received route from BGP or a CE router contains a SoO value that matches the SoO value on the receiving interface—If a route is received with an associated SoO value that matches the SoO value that is configured on the receiving interface, the route is filtered out because it was learned from another PE router or from a back door link. This behavior is designed to prevent routing loops.
- A received route from a CE router is configured with a SoO value that does not match—If a route is received with an associated SoO value that does not match the SoO value that is configured on the receiving interface, the route is accepted into the EIGRP topology table so that it can be redistributed into BGP. If the route is already installed to the EIGRP topology table but is associated with a different SoO value, the SoO value from the topology table is used when the route is redistributed into BGP.

- A received route from a CE router does not contain a SoO value—If a route is received without a SoO value, the route is accepted into the EIGRP topology table, and the SoO value from the interface that is used to reach the next hop CE router is appended to the route before it is redistributed into BGP.

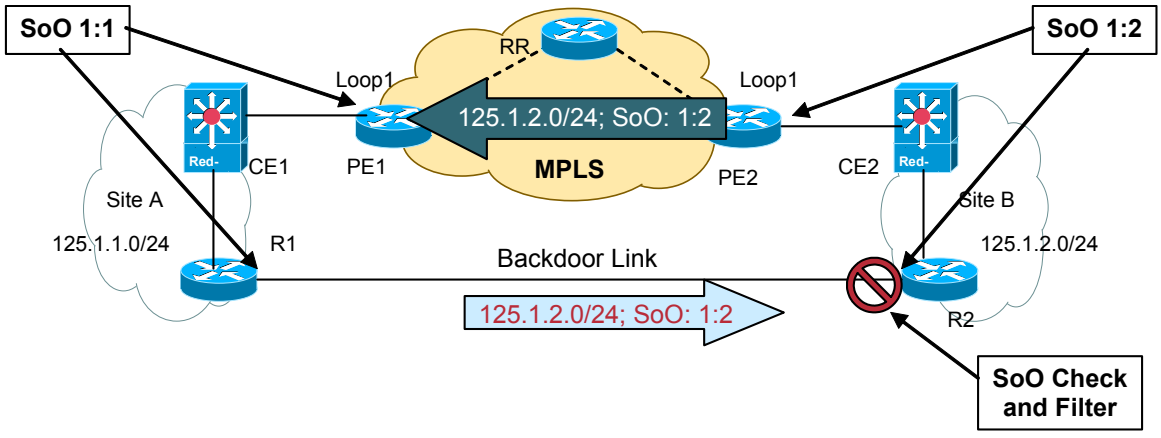
When BGP and EIGRP peers that support the SoO extended community receive these routes, they also receive the associated SoO values and pass them to other BGP and EIGRP peers that support the SoO extended community.This filtering is designed to prevent transient routes from being relearned from the originating site, which prevents transient routing loops from occurring.

The "pre-bestpath" point of insertion (POI) was introduced in the BGP Cost Community feature to support mixed EIGRP VPN network topologies that contain VPN and backdoor links. This POI is applied automatically to EIGRP routes that are redistributed into BGP. The "pre-best path" POI carries the EIGRP route type and metric. This POI influences the best path calculation process by influencing BGP to consider this POI before any other comparison step. When BGP has a prefix in the BGP table that is locally sourced and it receives the same prefix from a BGP peer, BGP compares the cost community values of the two paths. The path that has the best (or lowest) cost community value is selected as the best path.

Backdoor Link Router Operation

When a backdoor router receives EIGRP updates (or replies) from a neighbor across the backdoor link, it checks each received route to verify that it does not contain an SoO value that matches the one defined on the interface. If it finds a route with an SoO value that matches, the route is rejected and not put into the topology table. Typically, the reason that a route would be received with a matching SoO value would be that it was learned by the other site via the VPN connection and advertised back to the original site over the backdoor link. By filtering these routes based on the SoO value at the backdoor link, short term invalid routing is avoided.

Figure 2-5 EIGRP—Backdoor Link Support



In Figure 2-5, routes originating in site B are tagged with the SoO value 1:3 when the PE2 redistributes them into BGP. When the routes are redistributed from BGP into EIGRP on PE1, the SoO value is pulled out of the BGP table and retained on the routes as they are sent to site A. Routes are forwarded within site A and eventually advertised out backdoor router R1 to R2. The routes with the SoO value 1:1 are filtered out when updates are received by R2, stopping them from being relearned in Site A via the backdoor, thus preventing routing loops.

MPLS Network Convergence

Convergence can be defined as the time taken for routing nodes within a particular domain to learn about the complete topology and to recompute an alternative path (if one exists) to a particular destination after a network change has occurred. This process involves the routers adapting to these changes through synchronization of their view of the network with other routers within the same domain.

In an MPLS network, the convergence times of the following three network components can have an effect on application performance:

- Backbone convergence time—Convergence behavior in the backbone varies based on the core IGP and LDP operational mode. Core convergence time is dictated by its IGP convergence time. LDP convergence time is almost insignificant.
- VPN site convergence— Convergence behavior in a VPN site varies based on the IGP in use.
- VPN site route distribution convergence time—The redistribution delay comes from redistributing VPN site routes to MP-iBGP and redistributing MP-iBGP routes back to VPN sites. The delay is dictated by MP-iBGP.

Convergence behavior in the backbone varies based on the core IGP and LDP operational mode. Core convergence time is dictated by its IGP convergence time. LDP convergence time is almost insignificant. Convergence behavior in a VPN site varies based on the IGP in use.

The redistribution delay comes from redistributing VPN site routes to MP-iBGP and redistributing MP-iBGP routes back to VPN sites. The delay is dictated by MP-iBGP.

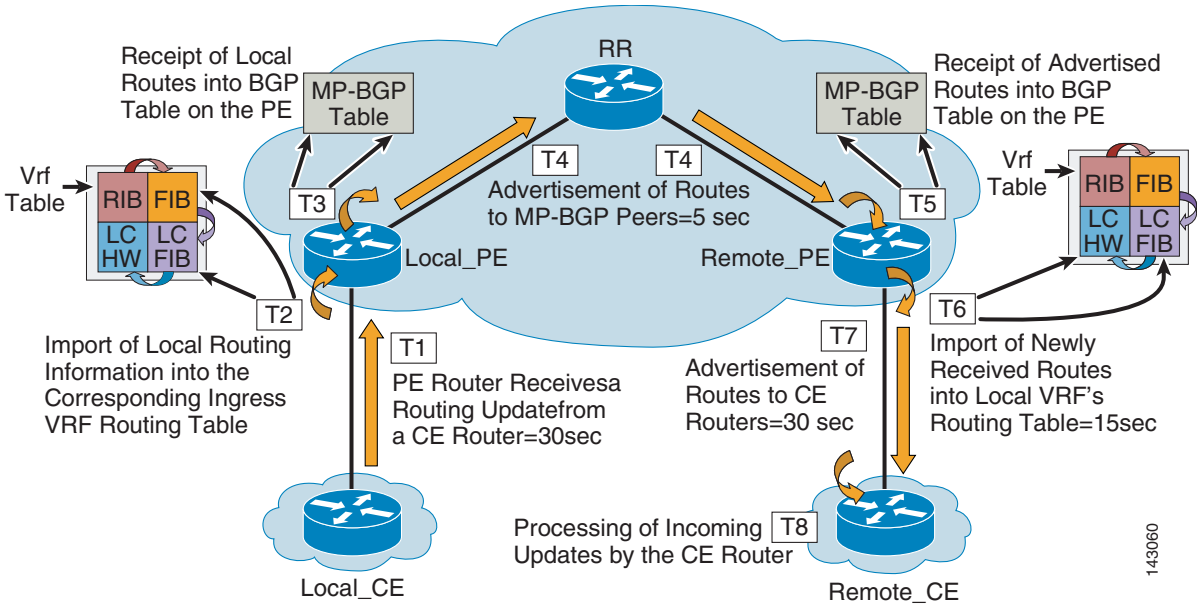
Also note that the convergence times vary for initial site updates (up convergence) and convergence occurring because of the failures in the network after the initial setup (down convergence). Only the key parameters that can be used to tune the network are highlighted in this guide.

For the most part, intelligent network design can help create faster converging networks. Some of the variables that can help tune the times are listed for each routing protocol, although the timers should be tuned after a careful examination of the current (default) network convergence times based on network design, network load, and application requirements.

Site-to-Site VPN Convergence Description with Default Timers

Figure 2-6 shows an example of site-to-site VPN convergence.

Figure 2-6 Site-to-Site VPN Convergence



The timers can be summarized in two categories:

- The first set of timers includes T1, T4, T6, and T7, which contribute higher convergence times unless tuned down.
- The second set of timers includes T2, T3, T5, and T8, and add smaller times.

Table 2-1 summarizes the maximum site-to-site convergence times with default timers for different routing protocols.

Table 2-1 Maximum Site-to-Site Convergence Times

PE-CE Protocol	Max Convergence Time (Default Settings) Where x= T2+T3+T5+T8	Max Convergence Time (Timers Tweaked Scan=5, Adv=0) Where x= T2+T3+T5+T8
BGP	~85+x seconds	~5+x seconds
OSPF	~25+x seconds	~5+x seconds
EIGRP	~25+x seconds	~5+x seconds
RIP	~85+x seconds	~5+x seconds

MPLS Network Convergence Tuning Parameters

EIGRP

EIGRP inherently provides sub-second convergence if the network is designed properly. An EIGRP network can be designed using feasible successors, summarization to bound queries, and (if applicable) using stub routers to fine tune overall network convergence time. For more information, see the following URLs:

- http://www.cisco.com/en/US/partner/tech/tk365/technologies_white_paper09186a0080094cb7.shtml

- http://www.cisco.com/application/pdf/en/us/guest/tech/tk207/c1550/cdccont_0900aecd801e4ab6.pdf
- http://www.cisco.com/en/US/partner/tech/tk365/technologies_white_paper0900aecd8023df6f.shtml

OSPF

Fast IGP convergence enhancements permit detection of link/node failure, propagation of the route change in the network, and recalculation and installation of the new routes in the routing and forwarding table as soon as possible. Some of the OSPF enhancements are OSPF event propagation, OSPF sub-second hellos tuning, OSPF LSA generation exponential backoff, and OSPF exponential backoff. LSA generation and SPF run wait times can be changed to allow faster convergence times. For more information, see the following URLs:

- http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1829/products_feature_guide09186a0080161064.html
- http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1838/products_feature_guide09186a0080134ad8.html

BGP

- BGP scanner time—By default, BGP scans the BGP table and routing table every 60 seconds for all the address-families that are configured under the BGP process. The next-hop validation is performed via this process. So if there is any route whose next-hop is not reachable anymore, this scan process marks the route as invalid and withdraws it. The scanner interval can be modified using the following command under vpnv4 address-family:

  ```
  bgp scan-time <5-60 seconds>
  ```

- BGP scan-time import—A number of vpnv4 routes might be learned from across the backbone, which are then subjected to best path calculation. Once best path is calculated, it gets imported into the respective VRF routing table. This import cycle runs every 15 seconds by default. Hence it can take up to a max of 15 seconds for vpnv4 routes learned by a PE from RR or another PE to make it into the local VRF routing table.

  This import process is a separate invocation and does not occur at the same time as the scan process. BGP import scan-time can be modified under vpnv4 address-family using the following command:

  ```
  bgp scan-time import <5-60 seconds>
  ```

- BGP VRF maximum-paths import—By limiting numbers of routes in a VRF, convergence time can be improved. For more information, see the following URL: http://cco.cisco.com/en/US/products/sw/iosswrel/ps5187/products_command_reference_chapter09186a008017d029.html#wp1058523.

LDP

LDP convergence mainly depends on IGP convergence. It is insignificant compared to IGP convergence.

Bidirectional Forwarding Detection (BFD)

Bi-directional Forwarding Detection (BFD) provides rapid failure detection times between forwarding engines, while maintaining low overhead. It also provides a single, standardized method of link/device/protocol failure detection at any protocol layer and over any media. The Internet draft for

BFD defines two modes for session initiation, Active and Passive. An Active node sends BFD control packets in an effort to establish a BFD session. A Passive node does not send any BFD packets until it receives BFD packets from an Active node.

Once the BFD session and appropriate timers have been negotiated, the BFD peers send BFD control packets to each other at the negotiated interval. As long as each BFD peer receives a BFD control packet within the detect-timer period, the BFD session remains up and any routing protocol associated with BFD maintains its adjacencies. If a BFD peer does not receive a control packet within the detect interval [(Required Minimum RX Interval) * (Detect Multiplier)], it informs any clients of that BFD session (i.e., any routing protocols) about the failure. A BFD session is associated with one client only even if it is configured between the same set of peers.

BFD is currently supported for BGP, OSPF, ISIS and EIGRP protocols. Refer to the following URL for supported platforms/releases:

http://www.cisco.com/en/US/tech/tk365/technologies_white_paper0900aecd80244005.shtml

Scalability of an MPLS Network

RFC 2547 architecture calls for supporting over one million VPNs in an MPLS network, although the number of VPNs supported per PE is limited by platform resources, the type and number of services supported on the platform, CE-PE routing protocol used, traffic patterns, and traffic load. The number of CPE, PE, and P devices needed in the network depends on the size of the organization and how the sites are dispersed geographically. CPE, PE, and P devices should be sized carefully based on the network size, number of VPN sites, and traffic load. For example, memory utilized by different components:

- Hardware IDB requires 4692 Bytes (One Per Physical Interface)
- Software IDB requires 2576 Bytes (One Per Interface and Per Sub-Interface)
- MPLS Forwarding Memory (LFIB) consumes one “taginfo” (64 Bytes) per route, plus one Forwarding Entry (104 Bytes) for each path
- Minimum OSPF protocol memory needed is 168KB per process
- Need about 60-70KB per VRF and about 800-900 bytes per route
- Each BGP prefix entry with multiple iBGP paths needs 350 bytes of additional memory

MPLS Layer 2 VPNs—AToM

Any Transport over MPLS (AToM) is an industry solution for transporting Layer 2 packets over an IP/MPLS backbone. AToM is provided as part of the Unified VPN portfolio of leading-edge VPN technologies available over the widest breadth of Cisco routers and is based on the Martini draft described in the following URL:
http://www.ietf.org/Internet-drafts/draft-martini-l2circuit-trans-mpls-07.txt.

AToM is an attractive solution for the customers with ATM, Frame Relay, PPP, HDLC, or Ethernet networks that need point-to-point Layer 2 connectivity. With point-to-point virtual circuits, the Layer 2 connections retain their character as VPNs. The VPN site controls traffic routing within the network and the routing information resides on the VPN site edge router. As a result, the complexity of redistributing VPN site routing to and from the MPLS network is reduced. The MPLS PE supplies point-to-point connections or an emulated pseudowire (PW). A pseudowire is a connection between two PE devices

that connect two PW services of the same or disparate transport types. Note that Layer 2 and Layer 3 VPNs can be supported on the same PE device, but the CE-PE on a PE interface can only be a Layer 3 or Layer 2 VPN interface.

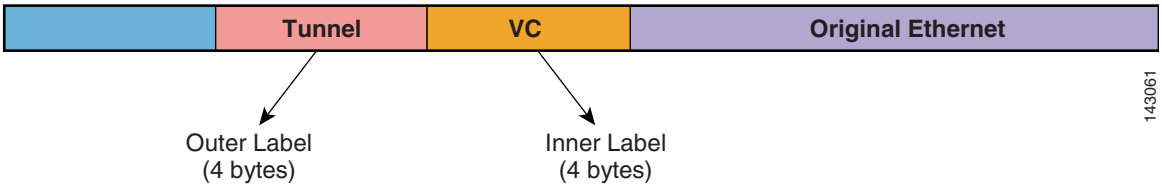
Ethernet over MPLS

Ethernet over MPLS (EoMPLS) is a popular method for creating Ethernet Virtual LAN services because it allows multiple services such as transparent LAN services (TLS) or bridging between sites, IP VPN services, and transport of desktop protocols such as SNA without interfering with the routing of the site. Ethernet traffic (unicast, broadcast, and multicast) from a source 802.1Q VLAN to a destination 802.1Q VLAN is transported over an MPLS core by mapping the VLANs to MPLS LSPs.

EoMPLS virtual circuits are created using LDP. EoMPLS uses targeted LDP sessions to dynamically set up and tear down LSPs over an MPLS core for dynamic service provisioning. No MAC address learning is required because this is a point-to-point connection that appears to be on the same wire.

Figure 2-7 shows an example of EoMPLS.

Figure 2-7 EoMPLS Example



EoMPLS comprises the following:

- Two levels of labels (8 bytes) are used:
  - Tunnel label—Outer label to forward the packet across the network
  - Virtual circuit (VC)—Inner label to bind Layer 2 interface where packets must be forwarded. The label is provided from the disposition PE. The imposition PE prepends this label so that the disposition router knows to which output interface and VC to route a packet. The egress PE uses the VC label to identify the VC and output interface to which the packet belongs.

    A VC is a 32-bit identifier used uniquely to identify the VC per tunnel and is configured between two different interfaces on PEs. The VC is used to define a point-to-point circuit over which Layer 2 PDUs are transported.

EoMPLS can operate in two modes:

- Port mode
- VLAN mode

VC type-0x0004 is used for VLAN over MPLS application and VC type-0x0005 is used for Ethernet port tunneling application (port transparency).

Port mode allows a frame coming into an interface to be packed into an MPLS packet and transported over the MPLS backbone to an egress interface. The entire Ethernet frame is transported without the preamble or FCS as a single packet.

VLAN mode transports Ethernet traffic from a source 802.1q to destination 802.1q VLAN over an MPLS core. The AToM control word is supported. However, if the peer PE does not support a control word, the control word is disabled. This negotiation is done by LDP label binding. Ethernet packets with hardware level cyclic redundancy check (CRC) errors, framing errors, and runt packets are discarded on input.

Port mode and Ethernet VLAN mode are mutually exclusive. If you enable a main interface for port-to-port transport, you cannot also enter commands on a subinterface.

EoMPLS operation is as follows:

1. The ingress PE router receives an Ethernet frame and encapsulates the packet by removing the preamble, the start of frame delimiter (SFD), and the frame check sequence (FCS). The rest of the packet header is not changed.
2. The ingress PE router adds a point-to-point virtual connection (VC) label and a label switched path (LSP) tunnel label for normal MPLS routing through the MPLS backbone.
3. The network core routers use the LSP tunnel label to move the packet through the MPLS backbone and do not distinguish Ethernet traffic from any other types of packets in the MPLS backbone.
4. At the other end of the MPLS backbone, the egress PE router receives the packet and de-encapsulates the packet by removing the LSP tunnel label if one is present. The PE router also removes the VC label from the packet.
5. The PE router updates the header, if necessary, and sends the packet out the appropriate interface to the destination switch.

QoS in AToM

The same QoS classification and marking mechanisms that are inherent in an MPLS network are used in AToM. Experimental bits in the MPLS header are used to create priority levels. For example, based on the type of service of the attachment VC, the MPLS EXP field can be set to a higher priority that allows better delivery of Layer 2 frames across the MPLS network. Layer 2 QoS, such as the 802.1P field in the IP header, can be easily mapped to MPLS EXP to translate QoS from Layer 2 to MPLS, thereby providing bandwidth, delay, and jitter guarantees. In the case of Frame Relay and ATM, the EXP values can be set by reference to the discard eligible (DE) bit marking in the frame header and to the cell loss priority (CLP) bit marking in the ATM cell header.

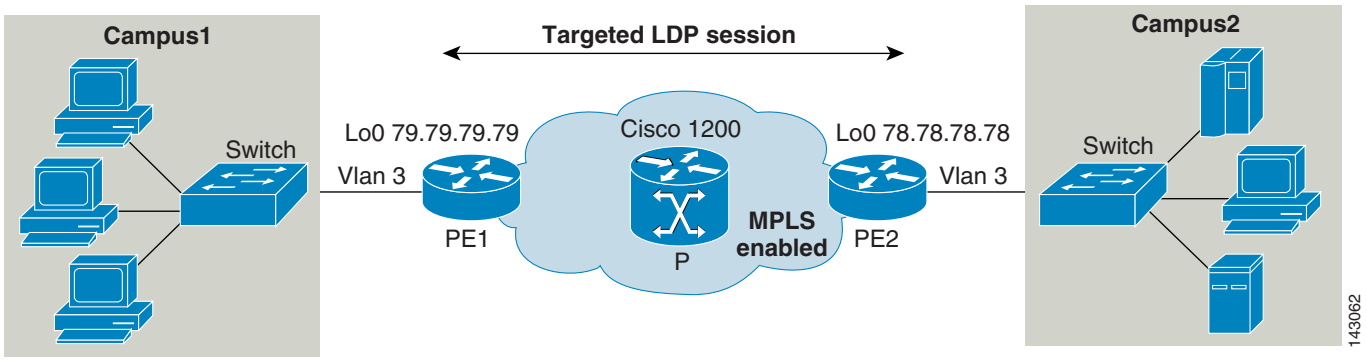
Scalability

MPLS AToM scales well because it does not require VC state information to be maintained by core MPLS devices. This is accomplished by label stacking to direct multiple connections bound for the same destination onto a single VC. The number of virtual circuits/VPNs serviced does not affect the MPLS core network. AToM, as per the IETF draft *Transport of Layer 2 Frames over MPLS,* calls for unlimited virtual circuits to be created: “This technique allows an unbounded number of Layer 2 ‘VCs’ to be carried together in a single tunnel.” Thus, it scales quite well in the network backbone. Although there are no hardware IDB limitations, the number of Layer 2 VCs supported per device (PE) is limited by the device (PE) resources, traffic load, and additional services enabled on the device (PE). From the provisioning perspective, if a fully-meshed connectivity between the sites is required, depending on the total number of sites, this solution can be labor-intensive to provision because it requires manually setting up n*(n-1) site meshes.

EoMPLS Sample Configuration

Figure 2-8 shows an example of an EoMPLS configuration topology.

Figure 2-8 EoMPLS Sample Configuration Typology



The following is the configuration on PE1:

```
mpls label protocol tdp
mpls ldp discovery directed-hello accept from 1

mpls ldp router-id Loopback0
!
interface FastEthernet2/11.3
 encapsulation dot1Q 3
 no ip directed-broadcast
 mpls l2transport route 78.78.78.78 300
 no cdp enable
!
access-list 1 permit 78.78.78.78
```

The following is the configuration on PE2:

```
mpls label protocol tdp
mpls ldp discovery directed-hello accept from 1

mpls ldp router-id Loopback0
!
interface FastEthernet2/11.3
 encapsulation dot1Q 3
 no ip directed-broadcast
 mpls l2transport route 79.79.79.79 300
 no cdp enable
!
access-list 1 permit 79.79.79.79
```