# MPLS based Transparent LAN services

Traditional metropolitan services are based upon TDM technologies, like SONET, and are optimized for voice services. But with data traffic becoming prominent, new Metro Service Providers (MSPs), also known as IP CLECs, now offer data services based upon Ethernet and IP technologies. The drive to use Ethernet as an communications technology comes from the economic benefits and flexibility that Ethernet offers.[1] MSPs offering Ethernet access can typically offer customers much more bandwidth for much less money.

But a serious challenge faces Ethernet-based MSPs: offering business customers the advanced services already available on ATM or frame relay networks. Today, MSPs rely on VLAN technology and IP networks to offer Virtual Leased Line (VLL) and Transparent LAN Services (TLS). Unfortunately, this is clearly a short term solution. VLANs were never designed for this usage. The IEEE 802.1Q specification allows a maximum of 4096 unique VLANs; as soon as more than 4096 customers need to be supported, MSPs will need new technologies. Furthermore, IP tunnels do not offer the kind of QoS guarantees available with ATM VCs, nor the level of protection that SONET offers. To truly compete with TDM technologies, new mechanisms are required.

MPLS-based TLS and VLL offer an attractive answer. Using MPLS-TLS and VLL allows MSPs to offer security, traffic engineering and QoS services to customers across the Metro network and into the core network. TLS allows MSPs to create a VPN tunnel for every customer through the network. Each VPN tunnel can be provisioned to a customer specified bandwidth and delay. VLL services, in turn, allow MSPs to compete with traditional LECs by offering a bandwidth provisioning point-to-point circuit within the metro. Instead of connecting metro buildings with traditional T1 circuits from a LEC, customers can obtain an Ethernet VLL from an MSP to perform the same service, for far less money.

This paper describes the operation of MPLS-TLS, as implemented by Riverstone Networks. Using MPLS-based TLS and VLL services herein described will allow MSPs to compete effectively against incumbent LECs.

[1] According to "Optical Access in the Public Network," a report from Communications Industry Researchers Inc., gigabit Ethernet, 10-Gbit/s Ethernet, and techniques for running IP directly over lightwaves will comprise at least 40 percent of ports shipped for metro access in the U.S. by 2004.

## New VPN Technologies in Metro Area Networks

Several new technologies address some or all of these problems:

- Stackable VLAN (SVLAN)
- Ethernet in IP or GRE
- MPLS

*SVLANs* solve the 4096 VLAN limitation by allowing a stack of two 802.1Q headers to be carried in an Ethernet frame, effectively extending the number of VLANs to more than 16 million (4096 * 4096). At the same time multiple VLANs can now be multiplexed within a single core VLAN (the top .1Q tag). The Generic Attribute Registration Protocol (GARP) and the GARP VLAN Registration Protocol (GVRP) can be used to automatically provision VLANs across the backbone. Spanning Tree extensions allow fast convergence (in the order of 1 sec). However, SVLANs only provide a partial solution. If customers are given the option to define their own VLAN ID spaces, the core of the network remains limited to 4096 VLANs. Also, without extended tunneling options such as Ethernet-in-Ethernet tunneling or without the use of a CPE router, the number of MAC addresses handled by the core will not be manageable.

*Ethernet in IP or GRE* offers the strength and scalability of a routed backbone, while allowing each customer site to define multiple private VLAN's that can be tunneled within a very large number of IP tunnels. Provisioning of IP tunnels is not automatic and the number of IP tunnel address pairs to manage is a major issue. For protection, new protocols are being devised. IP routing protocols take several seconds at best to converge when a failure occurs. The Link Management Protocol, being defined within the IETF, will monitor the link state of any underlying technology and provide fast failure detection. For scalability purposes, the number of tunnels in the core could be minimized by defining hierarchical IP VPNs. But this lead to bandwidth inefficiency as the original Ethernet frame needs to be encapsulated into two IP headers, where the inner IP tunnel is used for intra-POP connectivity and the outer IP tunnel for inter-POP connectivity.

*MPLS* offers the strength and scalability of IP tunnels while providing means to dynamically provision MPLS tunnels known as Label Switch Paths (LSPs). These LSPs can be used for traffic engineering, to create differentiated services, and to offer unique protection schemes. The Martini Internet draft specifies how to transport Ethernet, ATM and Frame Relay protocol data units (PDUs), and TDM signals over MPLS. This Internet draft focuses on point-to-point connectivity. In this paper, we will discuss extensions to this model in order to provide multipoint-to-multipoint support, that is, broadcasting and multicasting support.

Note that all these schemes can be combined. For instance, SVLANs can be used within the POP while IP or MPLS tunneling is used in the core. A possible scenario would be to start deploying MPLS in the core while the edge continues to use more mature technologies such as VLANs or stackable VLANs.

## MPLS Packet Flow Overview

A customer's Ethernet frame is either switched or routed by a CPE device to a Provider Edge (PE) router known as an MPLS Label Edge Router (LER). The PE router determines which VLAN the frame belongs to, either by looking at the 802.1q header or by determining the VLAN associated with the incoming port. Filters can be applied to the frame so that undesired frames get dropped. For instance, if a CPE router is used, the PE device can check that the source MAC address corresponds to the CPE MAC address. Once the frame is deemed valid, the packet is mapped to a user-defined Forwarding Equivalence Class (FEC) which defines how specific packets get forwarded. The FEC lookup yields the outgoing port and two labels. The first label at the top of the stack is the tunnel label and is used to carry the frame across the provider backbone. The second label at the bottom of the stack is the VC label and is used by the egress switch to determine how to process the frame. After adding the two MPLS headers, one for each label, the frame is encapsulated into the proper format corresponding to the outgoing interface.

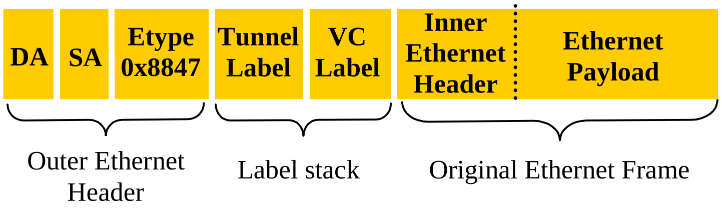

**Figure 1: MPLS Label Stack**

The backbone Label Switch Routers (LSRs) only look at the top label to switch the labeled frame across the MPLS domain. It is possible that additional labels get pushed along the way. The top tunnel label is typically removed by the penultimate hop i.e. the hop prior the egress LER. The egress LER infers from the VC label how to process the frame and then forwards it to the appropriate outgoing port.

In the previous section, it was assumed that tunnel and VC LSPs had been already established. VC LSPs are usually set up statically or dynamically via the Label Distribution Protocol LDP. LDP allows best effort LSPs to be established. When traffic engineering LSPs are required, the CR-LDP (Constraint-Based LDP) and RSVP-TE (Resource ReSerVation Protocol-Traffic Engineering) signaling protocols are used instead. Since resources in metro area networks are usually plentiful, traffic engineering is not necessary, making LDP a good choice for setting up VC LSPs. In the core backbone, since resources are not as easily available, traffic engineering is often required. For this reason, tunnel LSPs are usually established via RSVP-TE. One VC LSP, or multiple differentiated VC LSPs as described in the "Quality of Service" section in this paper, is established between each customer site belonging to the same VLAN. A single tunnel LSP carries all the traffic from multiple customers between two locations. By nesting LSPs, i.e. by building a forwarding hierarchy, and by limiting the number of core LSPs to the number of locations to interconnect, MPLS offers a very scalable solution.

In figure 2, two different customers are provided with TLS services. Customer A has three different sites, one in San Francisco, one in Chicago and one in New York. Customer B has facilities in San Francisco and New York. The MSP backbone consists of a full mesh of three LSPs (three pairs as discussed below). An end-to-end LSP, established between each location for each customer, is tunneled through a core LSP. For customer A, there are two VC LSPs established at each POP. From the San Francisco POP, one VC LSP carries traffic to Chicago and another LSP carries traffic to New York. Similarly, there are two VC LSPs in Chicago and New York set up exclusively for customer A. This full mesh of LSPs forms a unique broadcast domain, VLAN A, for customer A. For customer B, only one VC LSP is needed in San Francisco and New York. Customers A and B share the same tunnel LSP between San Francisco and New York LSRs.

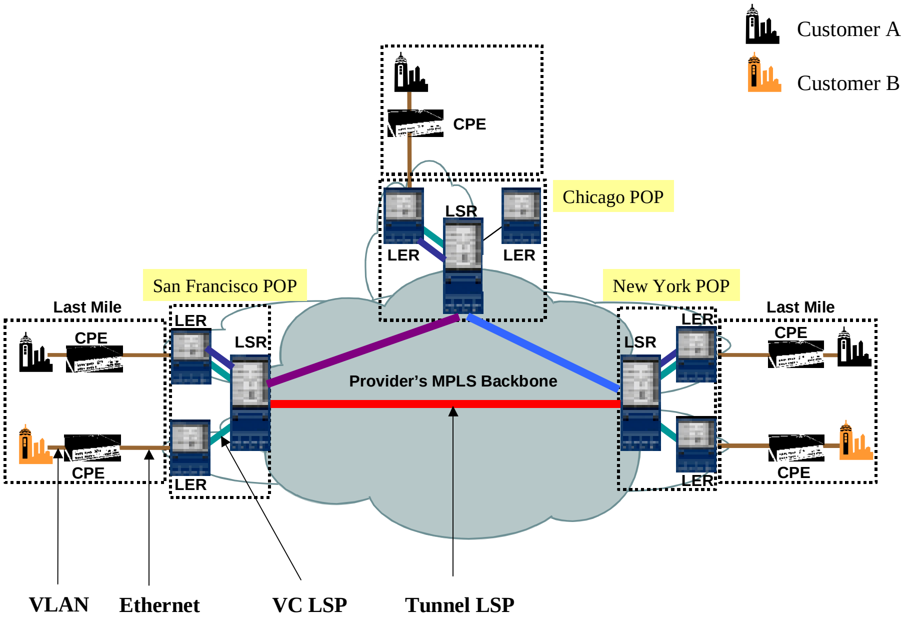

**Figure 2: TLS across MANs**

It should be noted that since LSPs are unidirectional, a pair of LSPs is actually needed to create a bi-directional pipe. Extensions to the LDP and RSVP-TE signaling protocols are being proposed in order to automatically set up either the reverse path LSP when the first simplex LSP is established or set up bi-directional LSPs.

The ability to treat pairs of LSPs as virtual interfaces that can be added to a VLAN allows transparent bridging to operate. When a broadcast frame or a frame with an unknown destination needs to be sent, the frame is flooded on all the LSPs that are part of the VLAN. The LER performs the packet replication across the LSPs as the frame enters the MPLS domain. Once MAC addresses have been learned, frames are only sent on the proper LSP. When a new MAC address is learned on an inbound LSP, it needs to be associated with the outbound LSP that is part of the same pair.

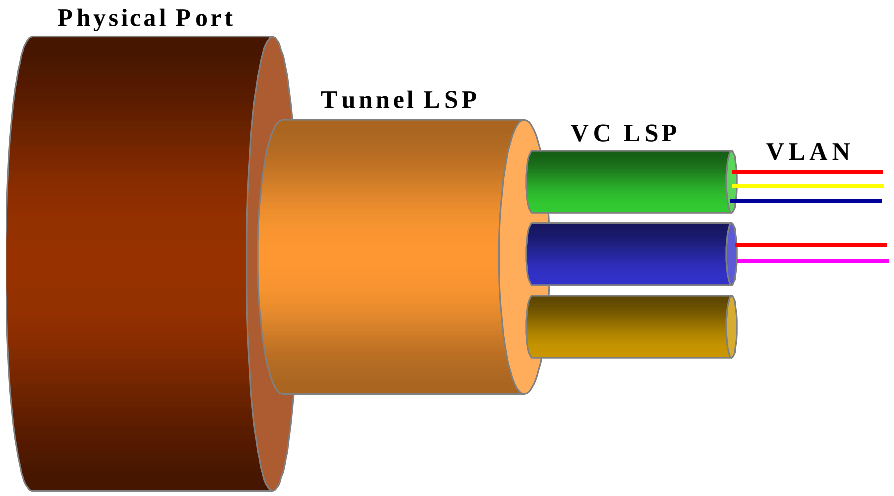

**Figure 3: MPLS Tunneling Hierarchy**

Because of the MPLS tunneling hierarchy, the VC label is not visible until the frame reaches the egress LER. The egress LER infers from the VC label the type of traffic being carried, such as ATM, Frame Relay, or Ethernet, and how to handle the corresponding frame. For ATM AAL5 traffic, the frame needs to be carried across the fabric to the proper output port and VPI/VCI. For Ethernet traffic, the VC label can be used to determine the VLAN the frame belongs to and the outgoing port or to perform an extended L2 lookup. The VC LSP creates a per customer tunnel that isolates traffic from other customers and offers the same level of security as a Frame Relay or ATM virtual circuit.

## Frame format

As a frame crosses an MPLS domain, several headers get added and several fields get changed. Figure 4 shows how an Ethernet frame originated from a customer site is transformed into a labeled frame and sent across other Ethernet links. The first hop, the CPE device, is an Ethernet switch in our example that will not change any field. As the frame enters the Service Provider network, the LER adds a two- label MPLS header to the original frame. The LER then adds another Ethernet header since the outgoing interface is also an Ethernet link. This outer Ethernet header contains the source MAC address of the LER and the destination MAC address of the next MPLS hop, and the MPLS Ethernet type (0x8847 for unicast traffic and 0x8848 for multicast). The original Ethernet is obviously untouched and carries the MAC addresses of the original sender and the actual recipient. The tunnel label is swapped by each transit LSR as the labeled frame crosses the MPLS cloud. At the same time, outer source and destination MAC addresses get also changed for the current hop and next hop MAC addresses, exactly like a traditional router. When the frame reaches the penultimate hop, the tunnel label is popped off and the labeled frame is sent to the egress LER. The LER uses the VC label to infer the output port, pops off the last label, removes the outer Ethernet header, and transmits the original Ethernet frame towards the recipient.

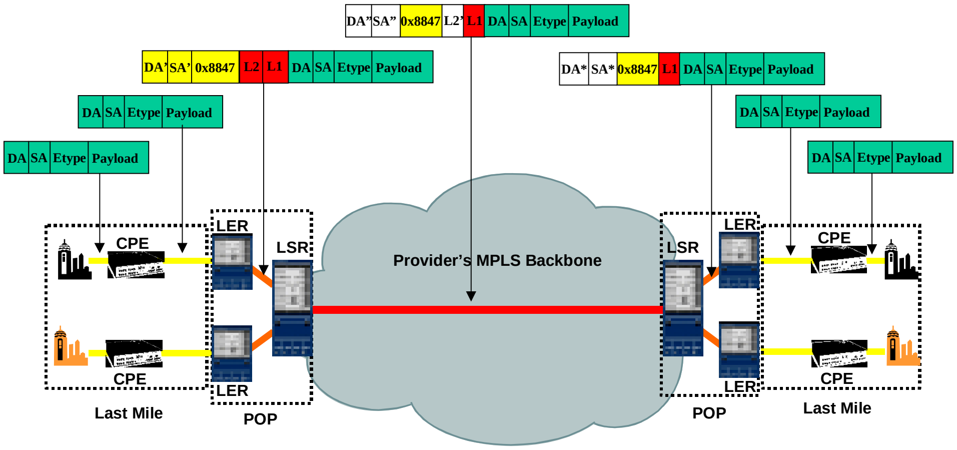

**Figure 4: MPLS Encapsulation**

## Fragmentation

The addition of a label stack and outer header may cause the maximum frame size to be exceeded. For instance, if an original Ethernet frame size is 1518 bytes, the maximum Ethernet frame size, new headers cannot be added. If jumbo frames are supported across all hops in the LSP path, the frame can be sent as is. For instance, if the links across the LSP path are SONET links, this is not an issue since the maximum frame size is 64KB. If jumbo frames are not supported, the frame could either be dropped or it could be pre-fragmented. If the original payload contains an IP packet, IP fragmentation can be applied at the very edge of the network before the frame gets tunneled over MPLS. Since the ingress LER is an IP router, it can fragment the original IP packet into multiple smaller IP packets as if the original sender had sent them, provided that the “Don’t Fragment” bit is not set. Note that the original sender is supposed to perform IP MTU path discovery or use the default IP MTU of 576 bytes if it can not, but there are several applications that violate this requirement. The advantage of pre-fragmenting packets is that the burden of re-assembling these IP packets falls into the recipient responsibility. IP re-assembly is an expensive processing and is rarely performed at wire speed, it should be minimized and handled by the actual recipient.

## Quality of Service and Resiliency

### Quality of Service

Quality of Service (QoS) and Class of Service (Cos) can be offered with TLS services in two different ways. Once the priority of an L2 frame is determined, based on the 802.1p priority or based on LER classification, a frame can either be marked with the appropriate class of service or it can be mapped to a specific QoS LSP. The MPLS header consists of a 20 bit label, 3 bit CoS field (also known as the EXP or Experimental bits), 1 bit bottom of stack, and an 8 bit TTL field as shown in figure 5.

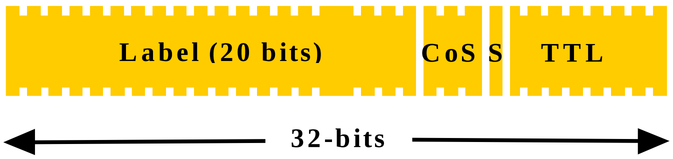

**Figure 5: MPLS Header**

The CoS bits are used at every hop along the LSP path to determine queueing characteristics. When a single LSP carries multiple classes of services, the CoS bits are used to determine a specific queueing, scheduling, and drop policy. This model corresponds to the E-LSP model in the MPLS support of differentiated services (Diff-Serv) draft.

The main intent of using multiple LSPs of different classes of service is to meet specific traffic engineering, quality of service, and protection requirements. In this case, the CoS bits can be used to only define a specific drop precedence, as specified in the L-LSP model.

The Riverstone switch/router family supports the E-LSP model and a variant of the L-LSP model. When multiple QoS LSPs are configured, the RS switches will not only infer the drop precedence but also the scheduling treatment of each packet.

When only soft QoS or CoS is needed, an MSP can rely upon the LSRs to prioritize best effort traffic based on the CoS bits allowing higher priority traffic to be subjected to lower delays. If necessary, fairness can be added by enabling weighted fair queuing or weighted round robin scheduling algorithms. When stricter or guaranteed services are required, an MSP can provision different paths across his network such that bandwidth and delay requirements are met. Additionally, an MSP can specify the relative priority of the different LSPs such that low priority LSPs can be preempted in the case of failure of a higher priority LSP if network resources are no longer available to restore the high priority LSP.

Figure 6 shows an example of three core differentiated LSPs. The gold LSP has been traffic engineered such that bandwidth requirements are guaranteed. This LSP can be used to carry highly critical traffic such as packetized voice. A silver LSP is available for medium priority traffic and a bronze LSP is used for best effort traffic with no service guarantee. The bronze LSP is typically oversubscribed.

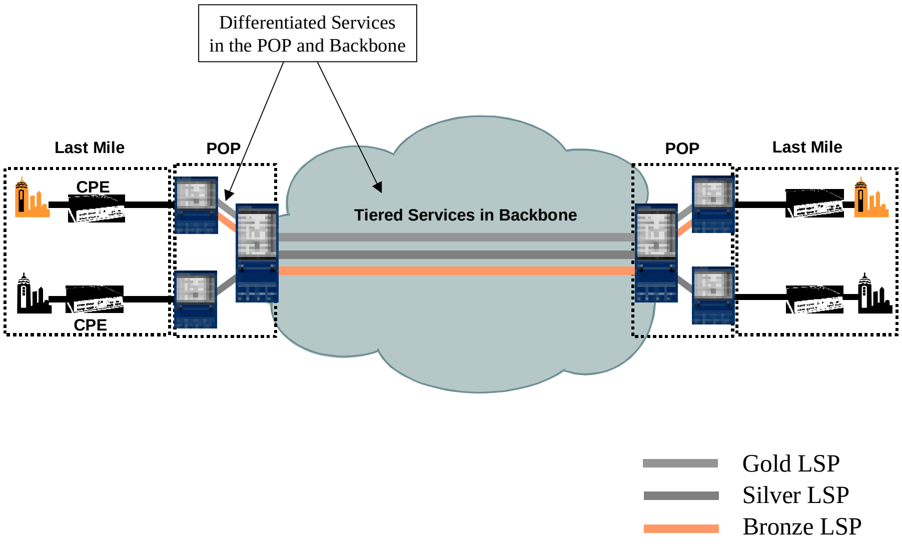

**Figure 6: MPLS Based Quality of Service**

### Load Balancing

Traffic can be load balanced across multiple LSPs to provide active redundancy and potentially higher throughput. It is expected that the diverse LSPs that have been established have identical traffic characteristics. In general, traffic is load balanced at the edge of a network. It is also possible to perform load-balancing functions within the MPLS network.

### Resiliency

Backup LSPs can be used when connectivity of the primary path is lost. Backup LSPs are pre-established for layer 2 traffic since they can not be dynamically computed for layer 2 traffic. Traffic can be switched back to the primary path when it is restored or stay on the backup path.

An alternative is to enable the fast reroute option. Fast reroute allows detour LSPs to be established around all points of failure. If a link or node fails, the traffic is immediately switched to the detour path. The ingress LER gets notified such that a backup path can be used if necessary.

Failures are detected via RSVP or LDP hellos or via routing topology updates. For fast convergence, RSVP hello timers can be set to expire very quickly such that failures can be detected in 50 msec.

## Provisioning MPLS based TLS Services

As discussed in the “MPLS packet flow overview” section, customer tunnels are established via LDP. A unique identifier is assigned to each customer, and each LER is configured with the customers that they serve and with the different customer sites that are part of the same network. LDP Hello messages are sent from each LER to each of the configured remote sites. After establishing a hello adjacency between two LERs, an LDP session is established. A label mapping message is sent to the upstream LER (LDP downstream unsolicited mode) for each customer configured. This message contains a virtual circuit FEC field for which a label is being advertised. This field is encoded with the unique customer identifier and the type of traffic that the LSP will carry (Ethernet for instance). The peer LER now knows which VC LSP the frame needs to follow to get to the specified remote site.

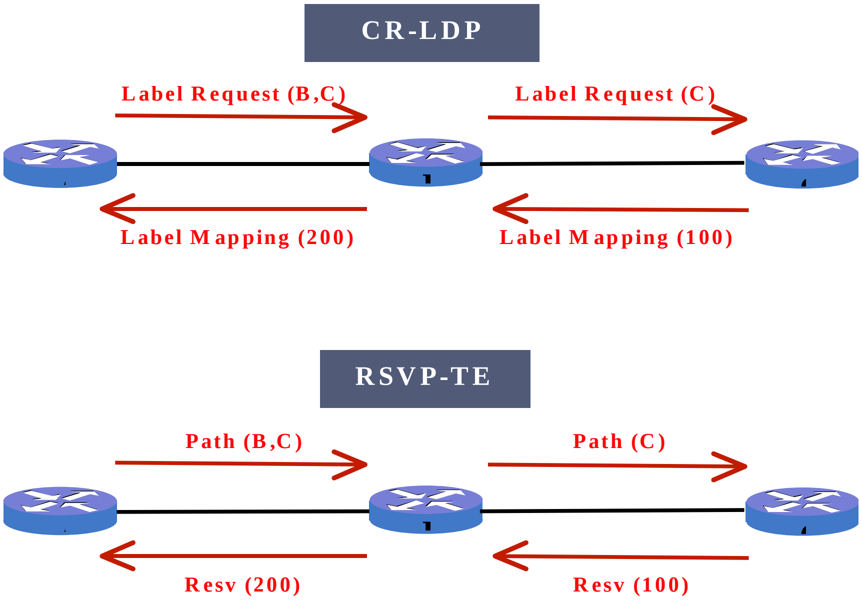

**Figure 7: MPLS Signaling**

Core tunnels are typically established via the most common signaling protocol, RSVP-TE, but could also be signaled via CR-LDP. Each core router at the edge of the RSVP domain is configured with the IP address of each egress router of the RSVP domain. The full RSVP path is configured for traffic engineered LSPs. The bandwidth of each LSP is also configured such that resources will be reserved in each hop along the path. Each router in the RSVP domain is also configured to tunnel LDP messages such that LDP sessions get established end to end. An RSVP path message is then issued from each edge LSR request that an LSP be set up, as shown in figure 7. An RSVP Resv message is then sent back to the originator via the exact same path, along which resources get reserved.

## Conclusion

With MPLS based Transparent LAN Services, service providers can now offer scalable, secured and guaranteed connectivity between customer locations. Dynamic provisioning of hierarchical tunnels greatly simplifies manageability, reducing the headaches associated with managing large-scale VPN deployments. MPLS QoS capabilities deliver consistent and guaranteed performance for critical applications and with the use of virtual circuits, MPLS delivers a high level of security.