

MPLS and Carrier Ethernet

Incumbent and competitive operators have started to provide telecommunications services based on Ethernet. This technology is arising as a real alternative to support both traditional data-based applications such as *Virtual Private Networks* (VPN), and new ones such as Triple Play.

Ethernet has several benefits, namely:

- It improves the flexibility and granularity of legacy TDM-based technologies. Many times, the same Ethernet interface can provide a wide range of bit rates without the need of upgrading network equipment.
- Ethernet is cheaper, more simple and more scalable than ATM and *Frame Relay* (FR). Today, Ethernet scales up to 100 Gb/s, and discussion on Terabit Ethernet is starting.

Furthermore, Ethernet is a well-known technology, and it has been dominant in enterprise networks for many years. However, Ethernet, based on the IEEE standards, has some important drawbacks that limit its roll-out, especially when the extension, number of hosts and type of services grow. This is the reason why, in many cases, Ethernet must be upgraded to carrier-class, to match the basic requirements for a proper telecom service in terms of quality, resilience and OAM (see Figure 1.1).

1.1 ETHERNET AS A MAN / WAN SERVICE

Ethernet has been used by companies for short-range and medium/high-bandwidth connections, typical of LANs. To connect hosts from remote LANs, up to now it has been necessary to provide either FR, ATM or leased lines. This means that the Ethernet data flow must be converted to a different protocol to be sent over the service-provider network and then converted back to Ethernet again. Using Ethernet in MAN and WAN environments would simplify the interface, and there would be no need for total or partial protocol conversions (see Figure 1.2).

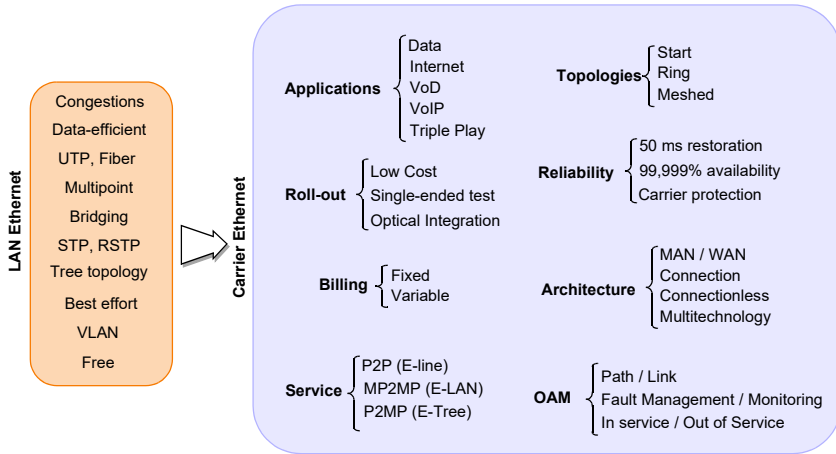


Figure 1.1 The path to Carrier-Class Ethernet.

Currently, the *Metro Ethernet Forum* (MEF), the *Internet Engineering Task Force* (IETF), the *Institute of Electrical and Electronics Engineers* (IEEE), and the *International Telecommunications Union* (ITU) are working to find solutions to enable the deployment of Carrier-Class Ethernet networks, also known as *Metro-Ethernet Networks* (MEN). This includes the definition of generic services, interfaces, deployment alternatives and interworking with current technologies. Carrier-Class Ethernet is not only a low-cost solution to interface with the subscriber network and carry its data across long distances, but it is also part of a converged network for any type of information, including voice, video and data.

1.1.1 Network Architecture

The ideal Metro-Ethernet Network makes use of pure Ethernet technology: Ethernet switches, interfaces and links. But in reality, Ethernet is often used together with other technologies currently available in the metropolitan network environment. Most of these technologies can inter-network with Ethernet, thus extending the range of the network. Next-Generation SDH (NG SDH) nodes can transport Ethernet frames transparently. Additionally, Ethernet can be transported by layer-2 networks, such as FR or ATM.

Today, many service providers are offering Ethernet to their customers simply as a service interface. The technology used to deliver the data is not an issue. In metro-

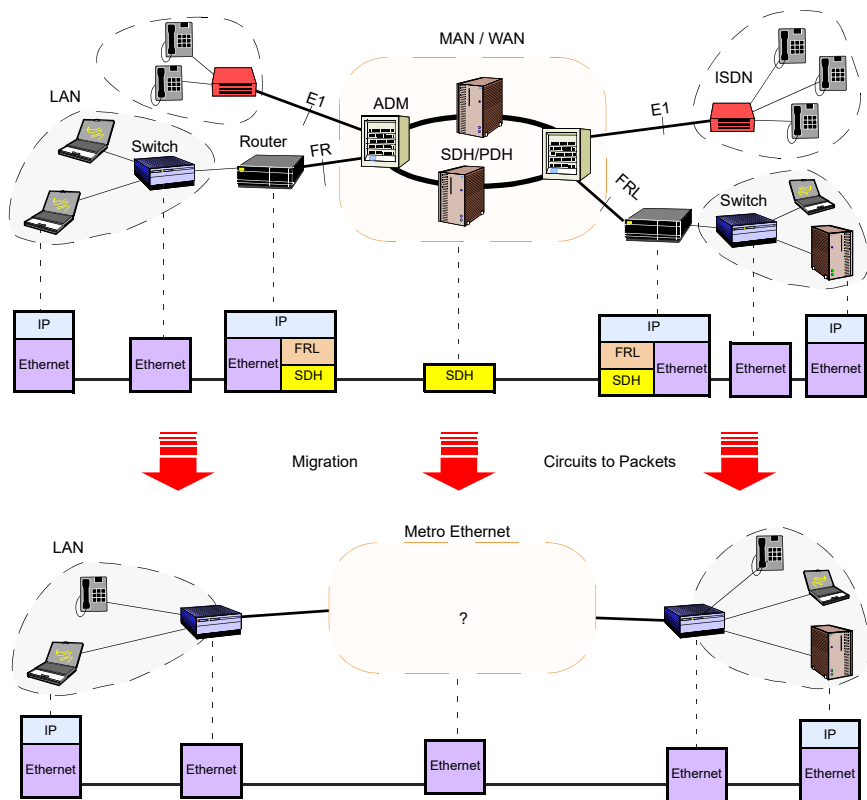


Figure 1.2 Migration to end-to-end Ethernet.

politan networks this technology can be Ethernet or SDH. Inter-city services are almost exclusively transported across SDH.

The interface between the customer premises equipment and the service-provider facilities is called User-to-Network Interface (UNI). The fact that Ethernet is being offered as a service interface makes the definition of the Ethernet UNI very important. In fact, this is one of the main points addressed by standardization organizations. The deployment plans for the UNI include three phases:

1. UNI Type 1 focuses on the Ethernet users of the existing IEEE Ethernet physical and MAC layers.
2. UNI Type 2 requires static service discovery functionality with auto-discovery and OAM capabilities.
3. UNI Type 3 requires a dynamic connection setup such that *Ethernet Virtual*

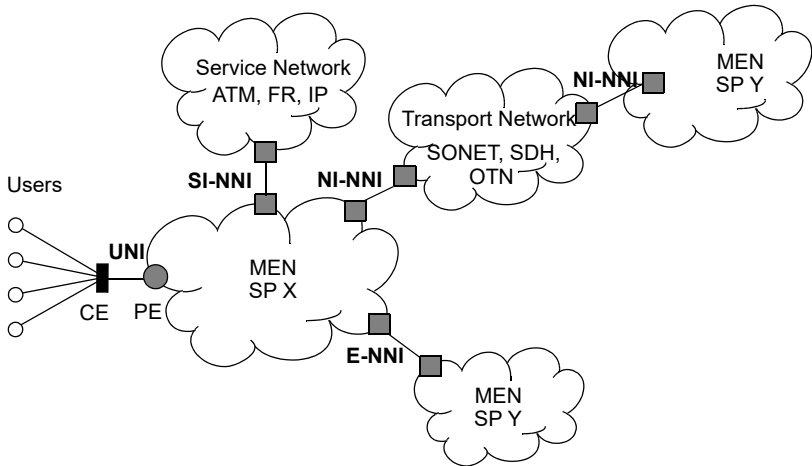


Figure 1.3 UNI and NNI in the MEN.

Connections (EVC) can be set up and / or modified from the customer UNI equipment.

The customer premises equipment that enables access to the MEN can be a router or a switch. This equipment is usually called Customer Edge (CE) equipment. The service provider equipment connected to the UNI, known as Provider Edge (PE), is a switch but deployments with routers are possible as well.

Many other interfaces are still to be defined, including the *Network-to-Network Interface* (NNI) for MEN inter-networking (see Figure 1.3). The network elements of the same MEN are connected by *Internal NNIs* (I-NNI). Two autonomous MENs are connected at an *External NNI* (E-NNI). The inter-networking to a transport network based on SDH, or *Optical Transport Network* (OTN), is done at the *Network Inter-Networking NNI* (NI-NNI). Finally, the connection to a different layer-2 network is established at the *Services Inter-Networking NNI* (SI-NNI).

1.1.2 Ethernet Virtual Connections

An *Ethernet Virtual Connection* (EVC) is defined as an association of two or more UNIs. A point-to-point EVC is limited to two UNIs, but a multipoint-to-multipoint EVC can have two or more UNIs that can be dynamically added or removed.

An EVC can be compared with the *Virtual Circuits* (VC) used by FR and ATM – however, the EVC has multipoint capabilities, whilst VCs are strictly point-to-point. This feature makes it possible to emulate the multicast nature of Ethernet. An EVC

facilitates the transmission of frames between UNIs, but also prevents the transmission of information outside the EVC.

Origin and destination MAC addresses and frame contents remain unchanged in the EVC, which is a major difference compared to routed networks where MAC addresses are modified at each Ethernet segment.

1.1.3 Multiplexing and Bundling

An Ethernet port can support several EVCs simultaneously. This feature, called service multiplexing, improves port utilization by lowering the number of ports per switch. It also makes service activation more simple (see Figure 1.4). Service multiplexing is achieved by using the IEEE 802.1Q *Virtual LAN* (VLAN) ID as a connection identifier.

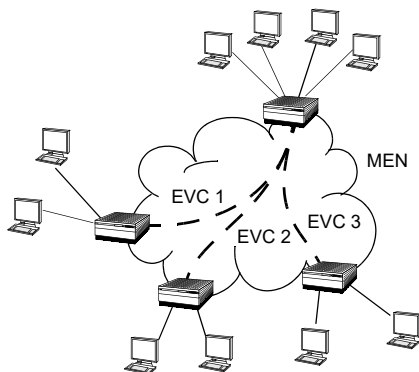


Figure 1.4 EVC Service multiplexing in a single port.

Service multiplexing makes it possible to provide new services without installing new cabling or nodes. This, consequently, reduces capital expenditure.

Bundling occurs when more than one subscriber's VLAN ID is mapped to the same EVC. Bundling is useful when the VLAN tagging scheme must be preserved across the MEN when remote branch offices are going to be connected. A special case of bundling occurs when every VLAN ID is mapped to a single EVC. This is called *all-to-one bundling*.

1.1.4 MEF Generic Service Types

Currently, the MEF has defined three generic service types: *Ethernet Line* (E-Line), *Ethernet LAN* (E-LAN) and *Ethernet Tree* (E-Tree) (see Figure 1.5).

1.1.4.1 E-Line Service Type

The *E-Line service* is a point-to-point EVC with attributes such as *Quality of Service* (QoS) parameters, VLAN tag support, and transparency to layer-2 protocols . The E-Line service can be compared, in some way, with *Permanent VCs* (PVCs) of FR or ATM, but E-Line is more scalable and has more service options.

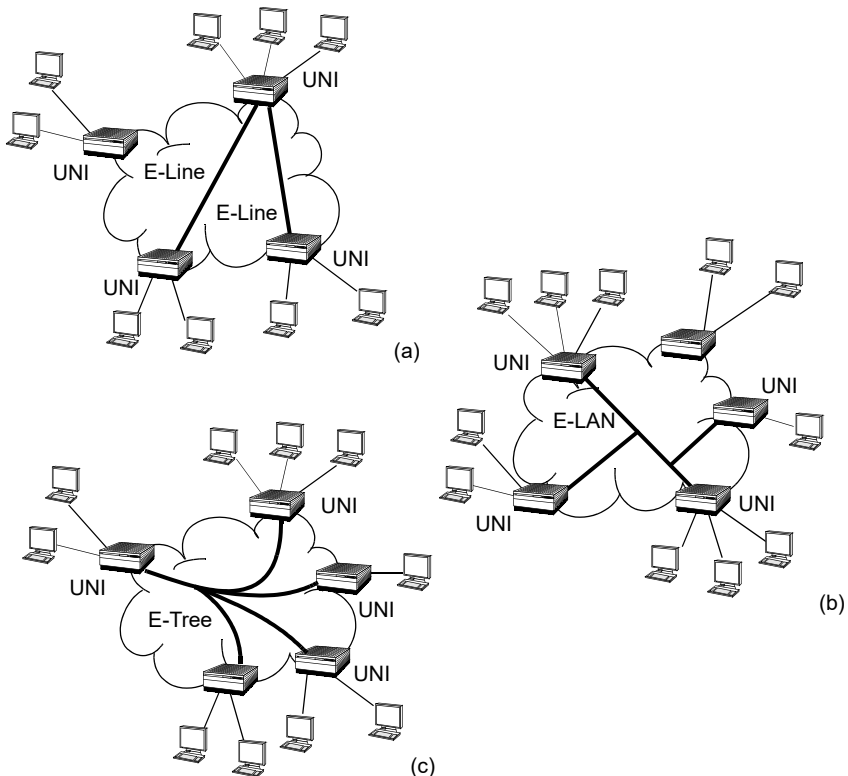


Figure 1.5 (a) The E-Line is understood as a point-to-point virtual circuit (b) The E-LAN service is multipoint to multipoint (c) E-Tree service is point-to-multipoint.

An E-Line service type can be a just simple Ethernet point-to-point with best effort connection, but it can also be a sophisticated TDM private line emulation.

1.1.4.2 E-LAN Service Type

The *E-LAN service* is an important new feature of Carrier-Class Ethernet. It provides a multipoint-to-multipoint data connection (see Figure 1.5). UNIs are allowed to be connected or disconnected from the E-LAN dynamically. The data sent from one UNI is sent to all other UNIs of the same E-LAN in the same way as happens in a classical Ethernet LAN. The E-LAN service offers many advantages over FR and ATM hub-and-spoke architectures that depend on various point-to-point PVCs to implement multicast communications.

The E-LAN can be offered simply as a best-effort service type, but it can also provide a specific QoS. Every UNI is allowed to have its own bandwidth profile. This could be useful when several branch offices are connected to one central office. In this case the *Committed Information Rate* (CIR) in the UNI for every branch office could be 10 Mb/s, and 100 Mb/s for the central office.

1.1.4.3 E-Tree Service Type

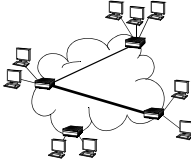
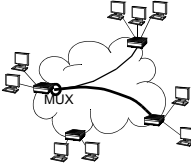
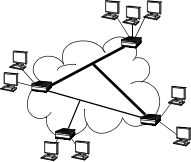
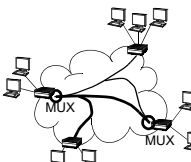
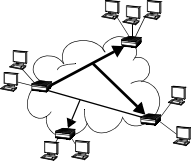
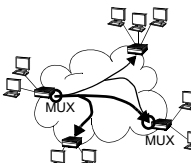
The E-Tree service type is suitable for delivering point-to-multipoint applications like IPTV. E-Tree is based on Ethernet multipoint connections with tree topology. Compared with the E-LAN service family, the E-Tree is different in that E-Tree multipoint connections have one or various well defined root nodes while other nodes remain as the leaves of the tree. Traffic flows from root to leaves but it cannot follow a direct path from leaf to leaf. A single E-Tree service could be replaced by several E-Line services with a *hub-and-spoke* configuration but the E-Tree is simpler and make better use of the network resources.

1.1.5 Connectivity Services

An Ethernet service arises when a generic service type (E-Line, E-LAN or E-Tree) is offered with particular EVC and UNI features. When a port-based service – that is, one single service per port – is provided at the UNI, it is called *Ethernet Private Line* (EPL), *Ethernet Private LAN* (EPLAN) or *Ethernet Private Tree* (EPTree), depending on if it is point-to-point, multipoint-to-multipoint or point-to-multipoint. Multiplexed services are called virtual. An *Ethernet Virtual Private Line* (EVPL) service, an *Ethernet Virtual Private LAN* (EVPLAN) and an *Ethernet Virtual Private Tree* (EVPTree) service can be defined (see Table 1.1).

From the point of view of the customer, the main differences between virtual and non-virtual services are that EPLs, EPLANs and EPTrees provide better frame transparency, and they are subject to more demanding *Service Level Agreement* (SLA) margins than EVPLs, EVPLANs and EVPTrees.

Table 1.1 Ethernet Connectivity Services

EVC to UNI Relationship			
	VLAN-Based Service	Port-Based Service	
Generic Etherservice Type	E-Line - Point to point - Best-effort or guaranteed QoS - Optional multiplexing and bundling	Ethernet Private Line (EPL) 	Ethernet Virtual Private Line (EVPL) 
	E-LAN - Multipoint to multipoint - Best effort or guaranteed QoS - Optional multiplexing and bundling	Ethernet Private LAN (EPLAN) 	Ethernet Virtual Private LAN (EVPLAN) 
	E-Tree - Point to multipoint - Best effort or guaranteed QoS - Optional multiplexing and bundling	Ethernet Private Tree (EPTree) 	Ethernet Virtual Private Tree (EVPTree) 

The meaning of multiplexed services in the case of EVPLs, EVPLANs and EVPTrees needs to be further explained. For example, several E-Line service types may be multiplexed in different SDH timeslots and be still considered EPLs. This is because the *Time Division Multiplexing* (TDM) resource-sharing technique of SDH makes it possible to divide the available bandwidth in such a way that congestion in some timeslots does not affect other timeslots. This way, it is possible to maintain the strong SLA margins typical of EPLs, EPLANs and EPTrees in those timeslots that are not affected by congestion.

EVPLs, EVPLANs and EVPTrees are statistically multiplexed services. They make use of service multiplexing, and thus VLAN IDs are used as EVC identifiers at the UNI.

1.1.5.1 Ethernet Private Lines

The *Ethernet Private Line* (EPL) service is a point-to-point Ethernet service that provides high frame transparency, and it is usually subject to strong SLAs. It can be considered as the Ethernet equivalent of a private line, but it offers the benefit of an Ethernet interface to the customer.

The EPLs make use of all-to-one bundling and subscriber VLAN tag transparency. This allows the customer to easily extend the VLAN architecture between sites at both ends of the MAN/WAN connection. Frame transparency enables typical layer-2 protocols, such as IEEE 802.1q *Spanning Tree Protocol* (STP), to be tunneled through the MAN/WAN.

EPLs are sometimes delivered over dedicated lines, but they can be supplied by means of layer-1 (TDM or lambdas) or layer-2 (MPLS, ATM, FR) multiplexed circuits. Some service providers want to emphasize this, and they talk about dedicated EPLs, if dedicated lines or layer-1 multiplexed circuits are used to deliver the service, or shared EPLs if layer-2 multiplexing is used.

EPLs are the most extended Metro Ethernet services today. They are best suited for critical, real-time applications.

1.1.5.2 Ethernet Virtual Private Lines

The *Ethernet Virtual Private Line* (EVPL) is a point-to-point Ethernet service similar to the EPL, except that service multiplexing is allowed, and it can be opaque to certain types of frames. For example, STP frames can be dropped by the network-side UNI.

Shared resources make it difficult for the EVPL to meet SLAs as precise as those of EPLs. The EVPL is similar to the FR or ATM PVCs. The VLAN ID for EVPLs is the equivalent of the FR *Data Link Connection Identifier* (DLCI) or the ATM *Virtual Circuit Identifier* (VCI) / *Virtual Path Identifier* (VPI).

One application of EVPLs could be a high-performance ISP-to-customer connection.

1.1.5.3 Ethernet Private LANs

Ethernet Private LANs (EPLAN) are multipoint-to-multipoint dedicated Carrier-Class Ethernet services. The EPLAN service is similar to the classic LAN Ethernet service, but over a MAN or a WAN. It is a dedicated service in the sense that Ether-

net traffic belonging to different customers is not mixed within the service-provider network.

Ethernet frames reach their destination thanks to the MAC switching supported by the service-provider network. Broadcasting, as well as multicasting are supported.

EPLAN services make use of all-to-one bundling and subscriber VLAN tag transparency to support the customer's VLAN architecture. Frame transparency is implemented in EPLANs to support LAN protocols across different sites.

1.1.5.4 Ethernet Virtual Private LANs

The *Ethernet Virtual Private LAN* (EVPLAN) is similar to the EPLAN, but EVPLANs are supported by a shared SP architecture instead of a dedicated one. The EVPLAN also has some common points with the EVPL. For example, the VLAN tag is used for service multiplexing, and EVPLANs could be opaque to some LAN protocols, such as the STP.

Both EPLAN and EVPLAN will probably be the most important Carrier-Class Ethernet services in the future. They have many attractive features. The same technology, Ethernet, is used in LAN, MAN and WAN environments. One connection to the service-provider network per site is enough, and EPLAN and EVPLAN offer an interesting alternative to today's layer-3 VPNs. EPLANs and EVPLANs enable the customers to deploy their own IP routers on top of the layer-2 Ethernet VPN.

1.2 ETHERNET DEPLOYMENT ALTERNATIVES

Today's installations use Ethernet on LANs to connect servers and workstations. Data applications and Internet services use WANs to get or to provide access to / from remote sites by means of leased lines, PDH / SDH TDM circuits and ATM / FR PVCs. Routers are the intermediate devices that using IP as a common language can also talk to the LAN and WAN protocols. This has been a very popular solution, but it is not a real end-to-end Ethernet service. This means that MAC frames "die" as soon as IP packets enter on the PDH/SDH domain, and they are created again when they reach the far-end.

This option, which is now often considered as legacy, has been the most popular networking data solution. During the past couple of decades, routing technologies have formed flexible and distributed layer-3 VPNs. Since Ethernet is present in both LANs, why not use Ethernet across the WAN as well?

The first approaches for extending Ethernet over a WAN are based on mixing Ethernet with legacy technologies, for example Ethernet over ATM, as defined in the

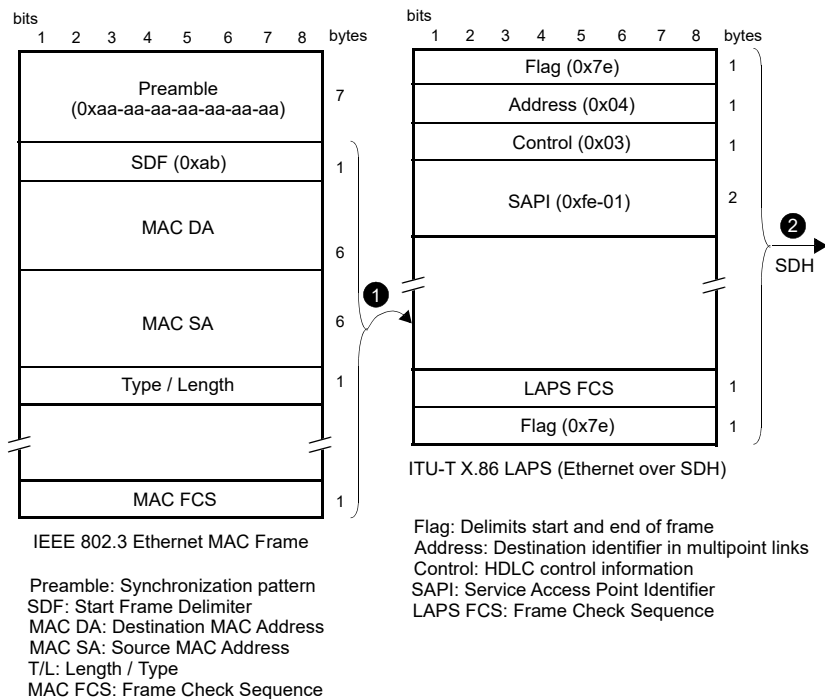


Figure 1.6 Legacy encapsulation for transporting Ethernet over SDH networks. The SDH solution makes use of the of the LAPS encapsulation.

IETF standard RFC-2684, or Ethernet over SDH by means of the *Link Access Procedure - SDH* (LAPS) as per ITU-T Recommendation X.86.

- The LAPS is a genuine Ethernet solution that provides bit rate adaptation and frame delineation. It offers LAN connectivity, allowing switches and hubs to interface directly with classic SDH. But it uses a byte-stuffing technique that makes the length of the frames data-dependent. This solution tunnels the Ethernet frames over SDH TDM timeslots called Virtual Containers (VCs). The Ethernet MAC frames remain passive within the network, and therefore this solution is only useful for simple solutions, such as point-to point dedicated circuits (see Figure 1.6).
- The solution based on Ethernet over ATM is more flexible and attractive for service providers, because it allows to set up point-to-point switched circuits based on ATM PVCs. With this solution, Ethernet frames are tunneled across the ATM network. Switching is based on ATM VPI / VCI fields. The main problems of this architecture are high cost and low efficiency, combined with the poor scalability of ATM (see Figure 1.7).

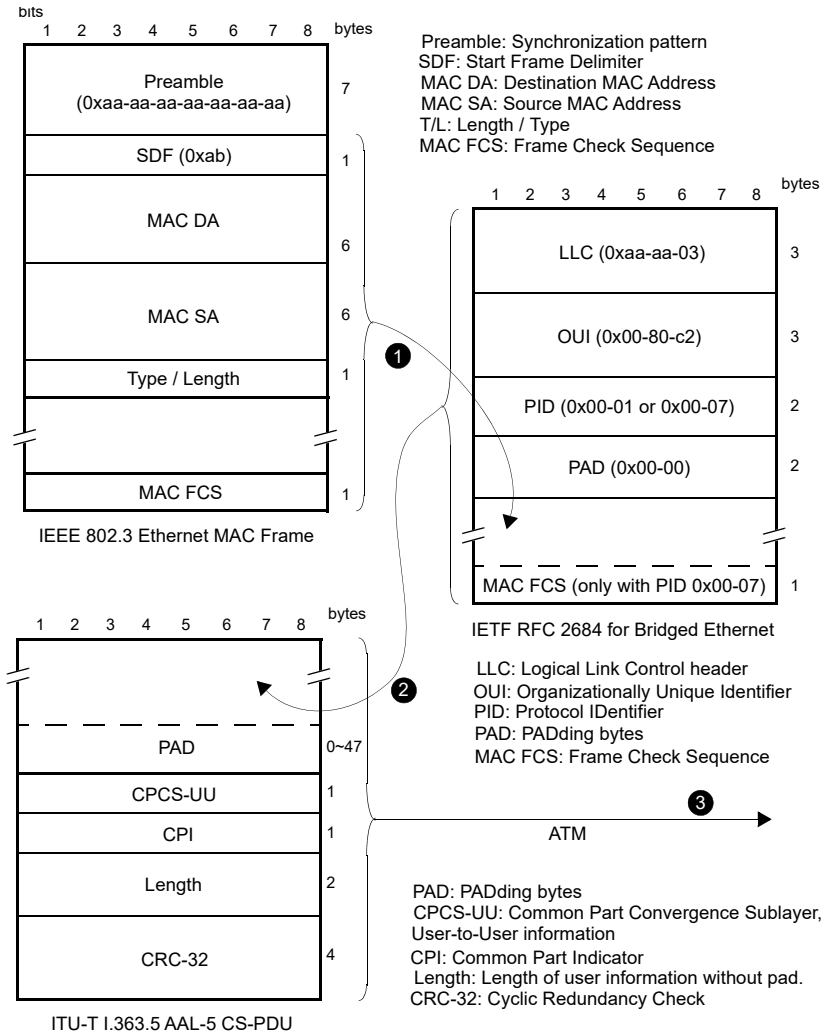


Figure 1.7 Legacy encapsulations for transporting Ethernet over ATM networks. The ATM mapping uses RFC-2684 and AAL-5 encapsulations.

The proposed alternatives, generically known as Carrier-Class Ethernet, replace ATM, FR or other layer-2 switching by Ethernet bridging based on MAC addresses (see Figure 1.8). Several architectures can fulfil the requirements, including dark fiber, WDM, NG-SDH. In principle, all of these architectures are able to support Carrier Ethernet services such as E-Line, E-LAN and E-Tree – however, some are more appropriate than others.

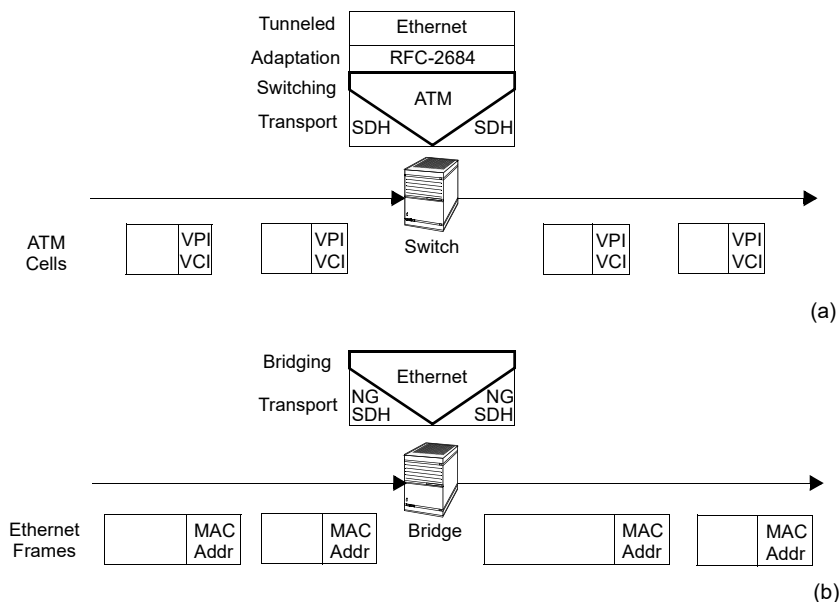


Figure 1.8 How NG-SDH raises the importance of Ethernet in the MAN / WAN. (a) Ethernet traffic is passively transported like any other user data. The ATM layer, specific for the WAN, is used for switching traffic. (b) The ATM layer disappears and the Ethernet layer becomes active. Traffic is now guided to its destination by means of Ethernet bridging.

1.2.1 Optical Ethernet

Ethernet can now be used in metropolitan networks due to the standardization of long-range, high-bandwidth Ethernet interfaces. It can be said that Ethernet bandwidths and ranges are at least of the same order as the bandwidths and ranges provided by classical WAN technologies.

MENs based on optical Ethernet are typical of early implementations. They are built by means of standard IEEE interfaces over dark optical fiber. They are therefore pure Ethernet networks. Multiple homing, link aggregation and VLAN tags can be used in order to increase resilience, bandwidth and traffic segregation. Interworking with the legacy SDH network can be achieved with the help of the WAN Interface Subsystem (WIS). The WIS is part of the WAN PHY specification for 10-Gigabit Ethernet. It provides multigigabit connectivity across SDH and WDM networks as an alternative to the LAN PHY for native-format networks.

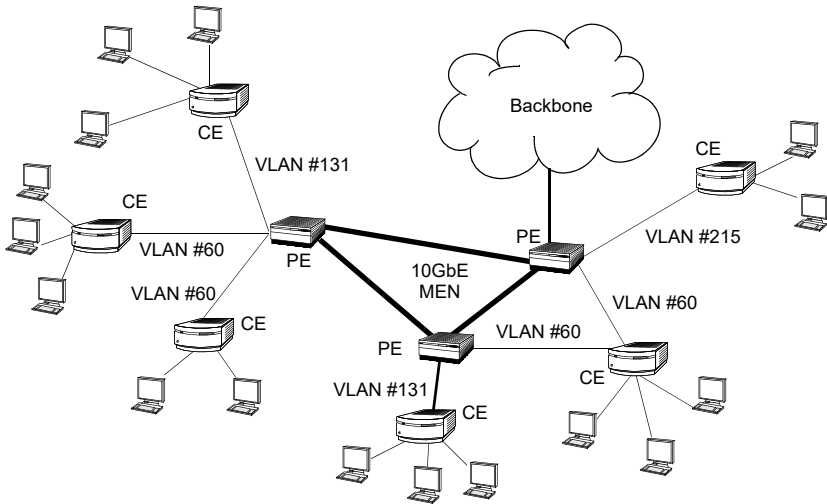


Figure 1.9 Optical Ethernet Carrier network. The trunk MEN links are implemented with optical 10-Gigabit Ethernet. Access links can be based on 1-Gigabit Ethernet or *Ethernet in the First Mile* (EFM), depending on the bandwidth requirements of every placement. Segregation of traffic from different subscribers or work groups is done by using VLAN tagging.

With this simple solution, a competitive operator can take advantage of packet switching, multipoint-to-multipoint applications and quick service roll-out. This option is a cost-effective in those areas where spare dark fiber is available and tree topologies are likely. Despite of its simplicity, pure Ethernet solutions for MEN have big scalability issues. Furthermore, they suffer from insufficient QoS, OAM, and resilience mechanisms.

Optical Ethernet has been often the architecture implemented by new operators to compete with the incumbent ones (see Figure 1.9). This kind of solution is still practical in metropolitan environments with a small number of connected subscribers or subsidiaries. Specifically, the pure Ethernet over dark fiber approach is discouraged for operators who want to provide services to a large number of residential and Small Office/Home Office (SOHO) customers.

1.2.2 Ethernet over WDM

The transport capability of the existing fiber can be multiplied by 16 or more if *Wavelength-Division Multiplexing* (WDM) is used. The resulting wavelengths are distributed to legacy and new technologies such Ethernet that will get individual lambdas while sharing fiber optics. WDM is a good option for core networks serving

very high bandwidth demands from applications like triple play, remote backups or hard disk mirroring. However, cost can be a limiting factor.

One of the inconveniences of this approach is the need to keep track of different and probably incompatible management platforms: one for Ethernet, another one for lambdas carrying SDH or other TDM technologies, and finally a third one for WDM. That makes OAM, traffic engineering and maintenance difficult.

1.2.3 Ethernet over SDH or OTN

Solutions for transporting Ethernet over SDH based on the *Generic Framing Procedure* (GFP), Virtual Concatenation and the *Link Capacity Adjustment Scheme* (LCAS) are generically known as *Ethernet over SDH* (EoS). The idea behind EoS is to substitute the native Ethernet layer 1 by SDH. The Ethernet MAC layer remains untouched to guarantee as much compatibility as possible with the IEEE Ethernet. Due to this, EoS cannot be considered as a true Ethernet technology. However, it is of great importance, because SDH is the *de facto* standard for transport networks. EoS makes it possible to reuse the existing infrastructure by taking advantage of the best of the SDH world, including resilience, long range and extended OAM capabilities.

NG-SDH unifies circuit and packet services under a unique architecture, providing Ethernet with a reliable infrastructure very rich in OAM functions (see Figure 1.10).

The three new elements that have made this migration possible are:

1. *Generic Framing Procedure* (GFP), as specified in Recommendation G.7041, is an encapsulation procedure for transporting packetized data over SDH. In principal, GFP performs bit rate adaptation and mapping into SDH circuits.
2. *Virtual Concatenation* (VCAT), as specified in Recommendation G.707, creates channels of customized bandwidth sizes rather than the fixed bandwidth provision of classic SDH, making transport and bandwidth provision more flexible and efficient.
3. *Link Capacity Adjustment Scheme* (LCAS), as specified in Recommendation G.7042, can modify the bandwidth of the VCAT channels dynamically, by adding or removing bandwidth elements of the channels, also known as members.

Ethernet traffic can be encapsulated in two modes:

1. *Transparent GFP* (GFP-T) is equivalent to a leased line with the bandwidth of the Ethernet bit rate. No delays, but expensive.
2. *Framed GFP* (GFP-F) is more efficient, because it removes interframe gaps

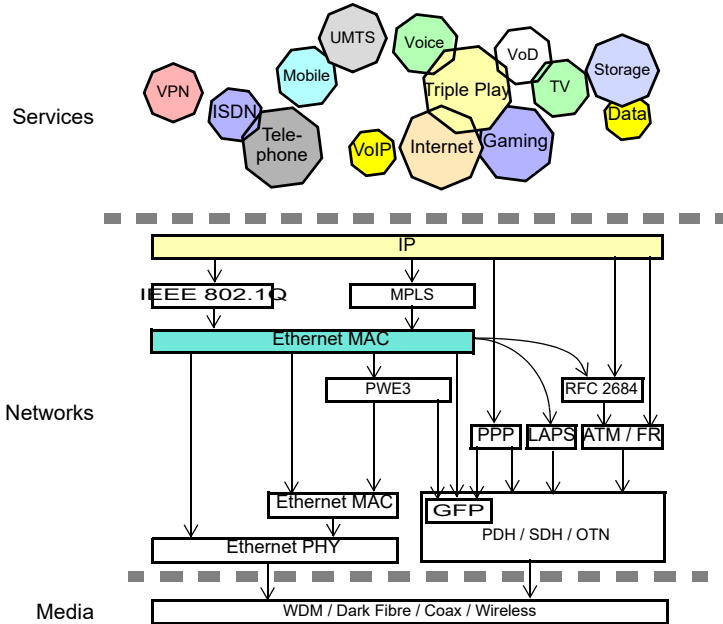


Figure 1.10 Transporting Ethernet and IP over packet- or circuit-switched infrastructures

and unnecessary frame fields. It also allows bandwidth sharing among several traffic flows. With GFP-F, service providers benefit from the statistical multiplexing gain, although subscribers may receive reduced performance when compared to GFP-T. This is due to the use of queues that increase end-to-end delay. Differentiated traffic profiles can be offered to customer signals (see Figure 1.11).

Compared to ATM, the GFP-F encapsulation has at least three critical advantages:

1. It adds very little overhead to the traffic stream. ATM adds 5 overhead bytes for every 53 delivered bytes plus AAL overhead.
2. It carries payloads with variable length, as opposed to ATM that can only carry 48-byte payloads. This makes it necessary to split long packets into small pieces before they are mapped in ATM.
3. It has not been designed as a complete networking layer like ATM – it is just an encapsulation. Specifically, it does not contain VPI / VCI or other equivalent fields for switching traffic. Switching is left to the upper layer, usually Ethernet.

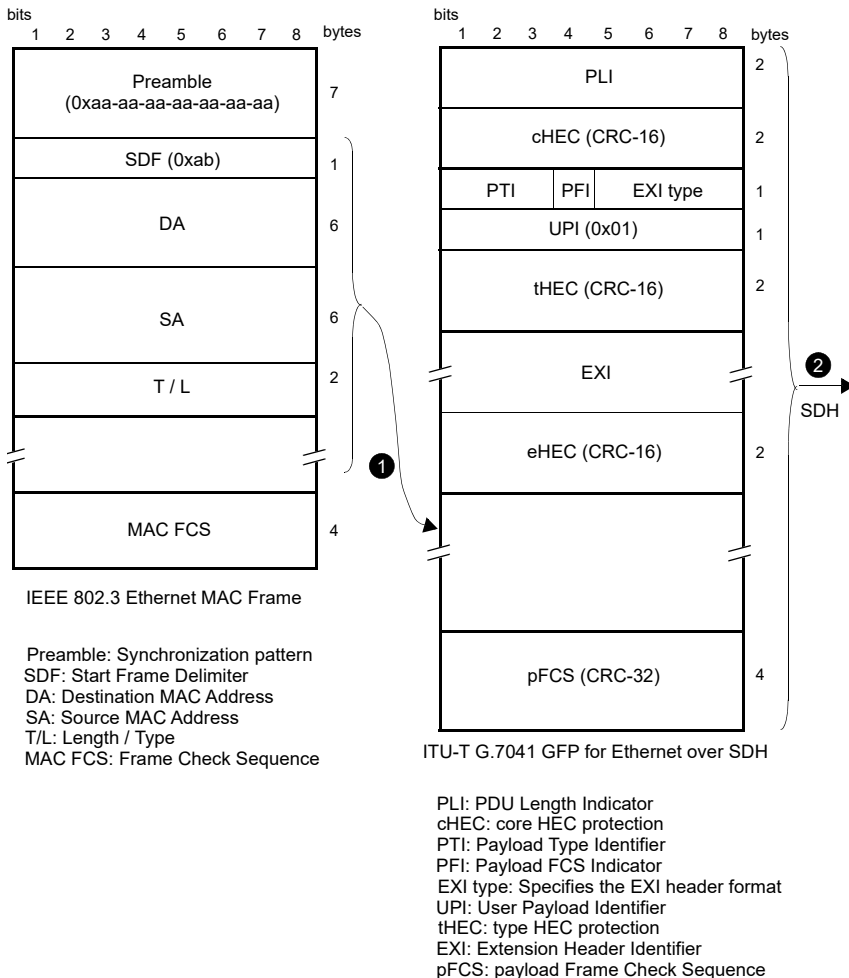
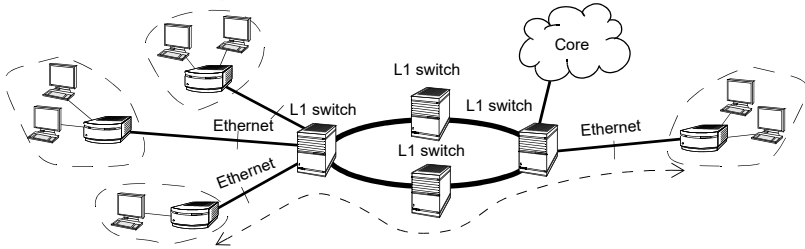


Figure 1.11 The GFP-F mapping for Ethernet makes ATM unnecessary. Now there is no VPI / VCI to switch the traffic, but the Ethernet MAC addresses can be used for similar purposes.

Like LAPS, GFP can be used for tunneling of Ethernet traffic over an SDH path, but the importance of this new mapping is that it allows Ethernet traffic to be active within the WAN. With the help of GFP, SDH network elements are able to bridge MAC frames like any other switch based on the Ethernet physical layer. The features of SDH MAC switches include MAC address learning and flooding of frames with unknown destination MAC address. In a few words, SDH MAC switches enable us to emulate an Ethernet LAN over an SDH network (see Figure 1.12).

NG-SDH - Customer switching



NG-SDH - Network switching

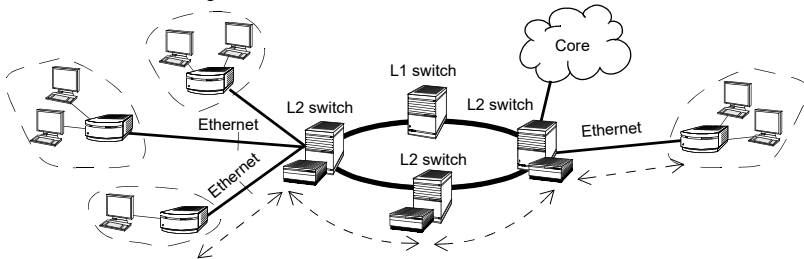


Figure 1.12 Ethernet over NG-SDH. Depending on the requirements, two approaches are possible: a) Customer switching – simple; the transport network is just a link between the customer switches. b) Network switching – more flexible; one step forward toward a more sophisticated service based on MPLS.

Deploying Ethernet in MAN / WAN environments makes it necessary to develop new types SDH *Add / Drop Multiplexers* (ADMs) and *Digital Cross-Connects* (DXC) with layer-2 bridging capabilities (see Figure 1.13):

- Enhanced ADMs are like a traditional ADM, but they include Ethernet interfaces to enable access to new services, and TDM interfaces for legacy services. Many of these network elements add Ethernet bridging capabilities, and some support MPLS and *Resilient Packet Ring* (RPR). New services benefit from the advantages of NG-SDH. New and legacy services are segregated in different SDH TDM timeslots.
- Packet ADMs have a configuration similar to enhanced ADMs: They include TDM and packet interfaces but packet ADM offers common packet-based management for both new and legacy services. The TDM tributaries are converted into packets before being forwarded to the network. *Circuit Emulation over Packet* (CEP) features are needed. MPLS is likely to be the technology in charge of multiplexing new and legacy services together in packet ADMs, due

to the flexibility given by MPLS connections known as *Label-Switched Paths* (LSP). Packet ADMs provide the same advantages as enhanced ADMs, but additionally, the network operator can benefit from increased efficiency and simplified management due to a unified switching paradigm.

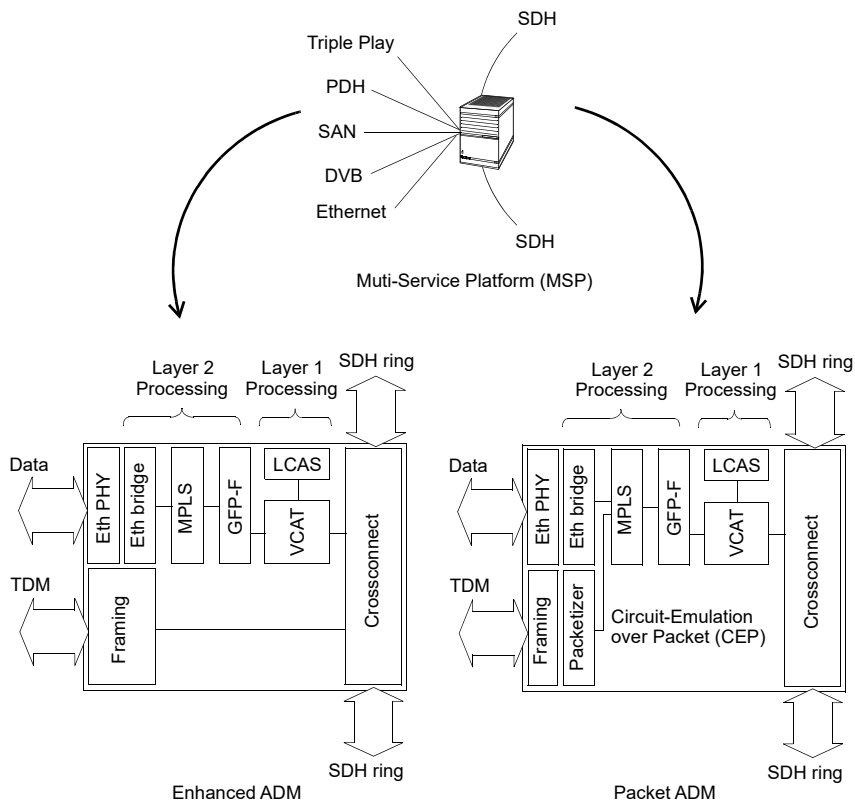


Figure 1.13 New SDH network elements. The Enhanced ADM offers packet and TDM interfaces in the same network element. Packet ADMs offer the same, but over an unified packet-based switching paradigm for all tributaries.

EoS is the technology preferred by incumbent operators, as they already have a large basis of SDH equipment in use. On the other hand, new operators generally prefer Carrier Ethernet directly implemented over optical layers.

1.3 LIMITATIONS OF BRIDGED NETWORKS

Metro network architectures based only on standard Ethernet switches operating over SDH or WDM are like large LANs; with all their advantages and inconvenienc-

es. We know that if these networks are lightly used by a reduced number of subscribers, they operate like LANs. However, when the metropolitan network starts growing, it becomes more and more difficult to keep the quality of service for all customers or it may even be impossible to supply network services to all subscribers due to scalability limitations of Ethernet LAN technology. For this reason, metropolitan network operators require help of some additional technologies or new mechanisms for the metropolitan network. These technologies and mechanisms are related in some way or other with Ethernet. Some solutions adopted by network operators to extend Ethernet to metropolitan networks are based on Multi-Protocol Label Switching (MPLS) or Provider Backbone Bridge with Traffic Engineering (PBB-TE).

1.3.1 Scalability

Ethernet switches, employ flooding to deliver data to their destination. They also depend on dynamic learning of MAC addresses to build their switching tables (IEEE 802.1D). When a switch is requested to send a frame to a host whose localization is unknown, it has to flow the frame to all its ports (with the only exception of the port that received the frame). This operation mode is neither efficient nor secure.

Unlike it happens with IP addresses, MAC addresses are not hierarchical. For this reason, Ethernet switching tables do not scale well and Ethernet switches are not efficient when they operate in very large networks with many potential destinations. This effect is known as MAC table explosion. Another reason of poor efficiency of bridging based on MAC addresses is that all network switches have to learn addresses dynamically for each new host connected to the network.

Using VLANs (IEEE 808.1Q) is a simple fix to these issues. One switch can be split in smaller switches, each belonging to an specific VLAN. VLANs are used to split a single broadcast domain in several smaller domains. In this way it is reduced the amount of broadcast traffic and at the same time security is improves (because frames are not sent to hosts not connected to the VLAN). An extra advantage of IEEE 802.1Q VLANs is that they enable provision of QoS with the help of the three 802.1p user priority bits (VLAN CoS bits).

The amount of available VLAN identifiers (VIDs) is, however, limited to 4,096. Service providers using more than a single VID per customer may exhaust all the available identifiers very quickly. Furthermore, subscribers may also have their own VLANs with the corresponding VIDs. It is interesting to define a solution to enable service providers and subscribers to coordinate their VLANs without the need to add special configuration in their networks.

The solution given for this issue is known as VLAN stacking or Q-in-Q. With this solution, two VLAN tags are used in each Ethernet frame thus increasing the total of

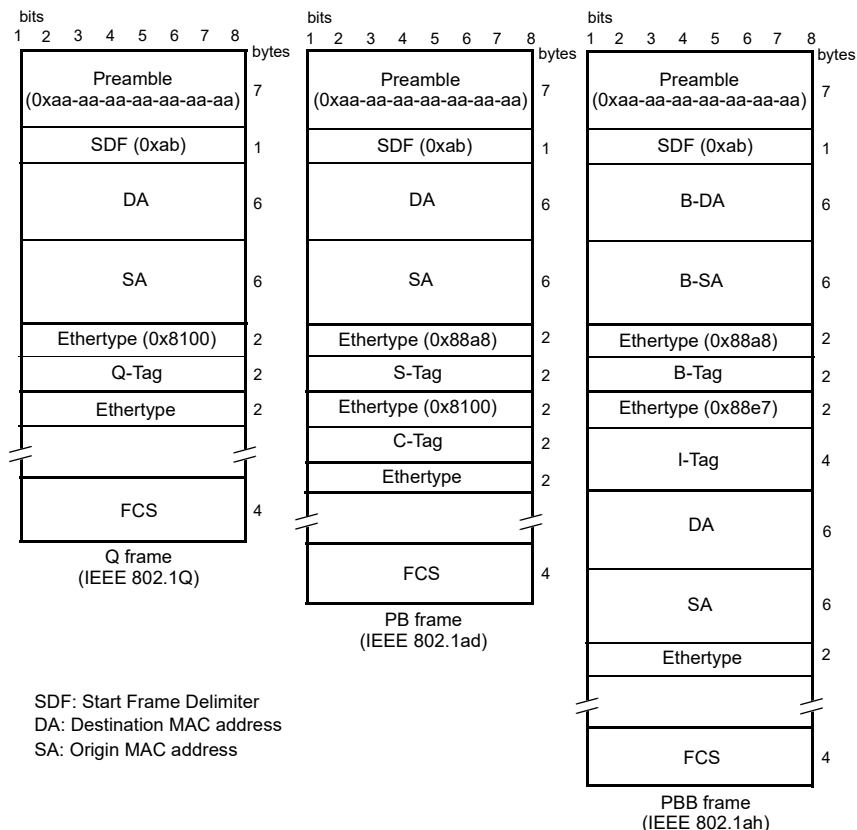


Figure 1.14 Ethernet frame formats for service provider networks. These frame formats offer a more scalable Ethernet.

available VLANs. VLAN stacking is a standard solution defined in IEEE 802.1ad for Ethernet Provider Bridges (PB). An even more powerful solution than VLAN stacking exists. It is the so called MAC address stacking or MAC-in-MAC, defined in standard IEEE 802.1ah for Ethernet Provider Backbone Bridges (PBB). PBB employed in Ethernet switches isolates the subscriber and service provider broadcast domains. This is useful to fight against the MAC table explosion issue (see Figure 1.14).

The Q-in-Q and MAC-in-MAC frame formats are also interesting because they provide a solution to the issue of Ethernet service demarcation. When the service provider network is based on ATM or FR and the subscriber network on Ethernet, it exists a clear border between the networks, but if both networks are Ethernet, things are not clear anymore and it is necessary to answer to questions like:

- Can typical LAN protocols, like the STP, deployed by subscribers in their network, modify or damage the service supplied by the service provider?
- How the traffic generated by one specific customer affects global operation of the service provider MAN / WAN?
- Is there any implication in the service provider network related with continuous host connection or disconnection within the subscriber network?

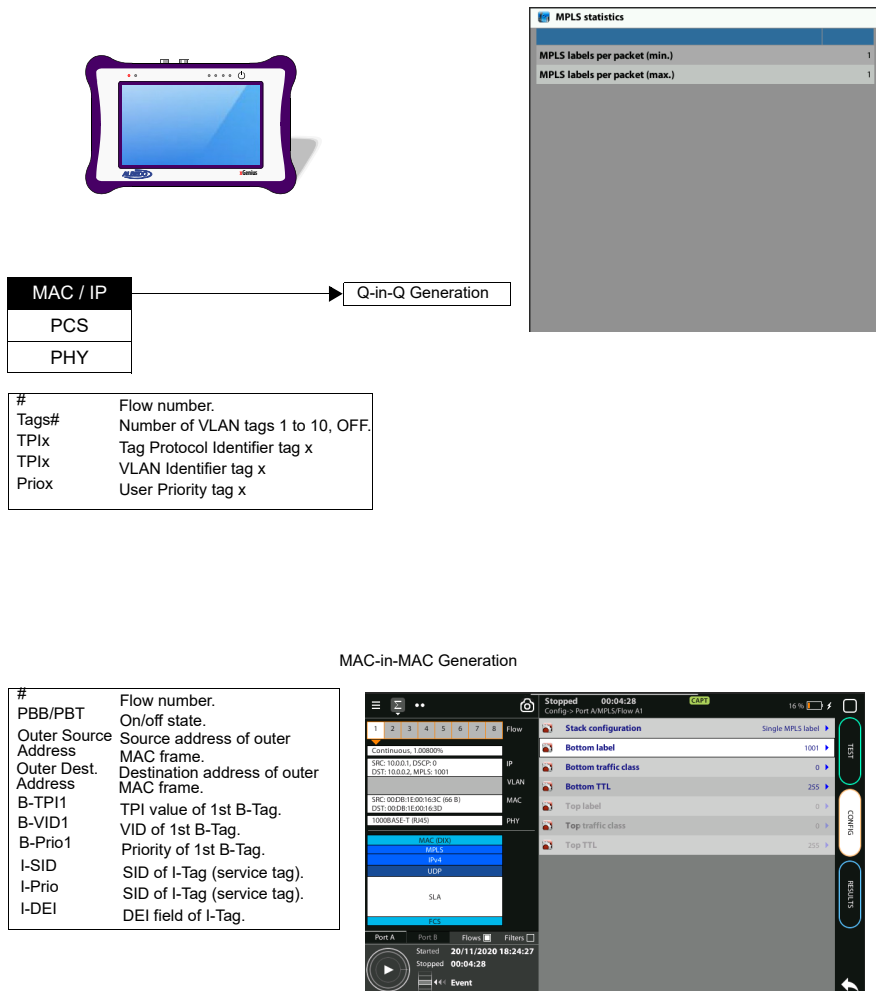
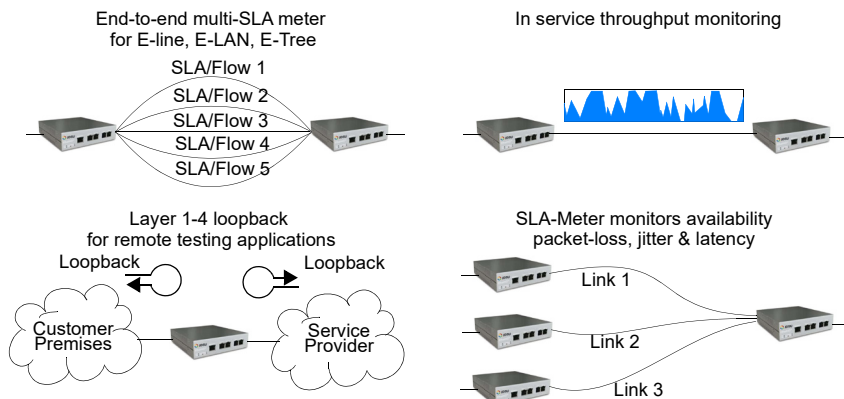


Figure 1.15 Q-in-Q and MAC-in-MAC traffic generation with the ALBEDO xGenius.

There is not a simple answer to these questions when the technology used in the service provider network is native Ethernet. In this case installation of Ethernet demarcation devices to filter and isolate traffic in subscriber and provider networks is almost compulsory. The ability to split MAC addresses and VLANs enabled by MAC-in-MAC and Q-in-Q formats has great value to achieve clear and effective Ethernet service demarcation.

Demarcation	<p>EtherNID</p> <p>Designed to demarc the edge of your network, EtherNIDs offer advanced packet Performance Assurance and service creation directly from custome premises and cell-sites. With a full range of Ethernet rates and interfaces, the comprehensive EtherNID family fits your network from end-to-end.</p>
	<p>MetroNID</p> <p>High-performance MetroNID units provide carrier-grade demarcation within metro and access networks. Designed for cellular hubs, aggregation nodes, and carrier hand-offs. MetroNIDs segment monitor and bridge diverse networks, delivering pervasive OAM and performance monitoring visibility.</p>



- Service Assurance (OAM, Test & Quality Assurance)**
- Ethernet & IP Service Demarcation
 - Layer 1-4 Intelligent Loopback
 - In-Service RFC-2544 Testing
 - Performance Assurance Agent
 - SLA Creation & SLA-Meter
 - End-to-End OAM Overlay
 - Automated RFC-2544 Test Suite
 - Test Set Support

- Service Creation (Networking Functionality)**
- Ethernet Service Mapping
 - Bandwidth Policing
 - Traffic Filtering
 - Zero-Latency Traffic Shaping
 - Multi-Port Aggregation

Figure 1.16 Service demarcation with demarcation devices

1.3.2 Quality of Service

Availability of QoS mechanisms in basic Ethernet is very limited. VLAN-tagged frames enable definition of up to eight different class of service (CoS) labels with different priorities. However, native Ethernet technology is unable to supply services with strong Service Level Agreement (SLA) requirements due to the lack of mechanisms related with network resource management and traffic engineering.

Maybe the most evident solution to this problems is to use one technology known as PBB with Traffic Engineering (PBB-TE). As defined in IEEE 802.1Qay, PBB-TE replaces standard Ethernet bridging within a range or in all MAC addresses by a new switching paradigm more suited to the needs of a service provider operating a large Ethernet network. One possible choice is to use centralized switching table management to allow close control of the available transmission resources (see Figure 1.17).

PBB-TE uses the PBB encapsulation for the data and it simply redefines how some fields and attributes of the PBB frame format are used by the network. Thanks to the PBB format it is possible to keep bridging with flooding and dynamic self learning for the subscriber MAC addresses and migrate to the new switching model the service provider MAC addresses. With some smart but simple changes in how some frame fields are interpreted and small modifications in switch operation, PBB-TE fulfills important objectives: Ethernet becomes a connection oriented technology. Now it is not difficult to allocate network resources on specific Ethernet connections. The consequence is that it becomes much easier to provide strong QoS over these Ethernet connections.

PBB-TE is much more than a solution to improve the QoS capabilities of Ethernet. There are many advantages in PBB-TE. For example, PBB-TE enables operators to perform load balancing over two or more Ethernet connections, route towards different ports traffic from different QoS classes or perform any other simple or complex traffic engineering task. All this would be very difficult to accomplish with basic Ethernet features.

PBB-TE network management and native Ethernet management are very different. PBB-TE management is much closer to the traditional network management of the like of telecom service providers. The main issue here is that some of the procedures and mechanisms to be used are still drafts or they have been recently released.

One of the weak points of PBB-TE that has been often criticized is the lack of support for point-to-multipoint and multipoint-to -multipoint but in this field solutions are also coming.

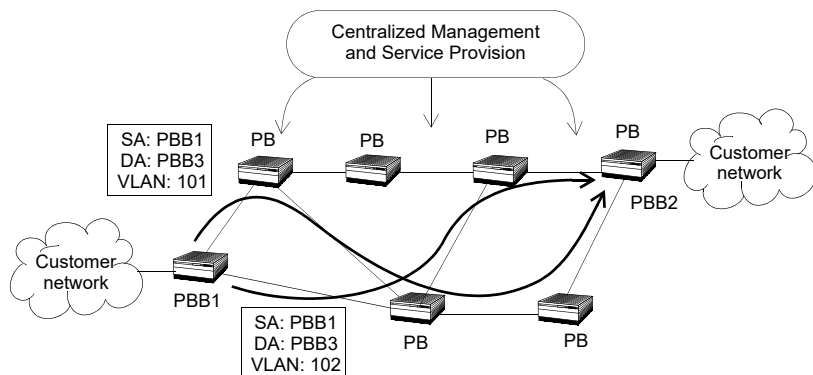


Figure 1.17 Ethernet switched paths based on PBB-TE. Data is encapsulated in MAC-in-MAC frames. Meaning of service provider fields has been altered. For example, the VLAN field now identifies paths with the same origin and destination.

1.3.3 Resiliency and Fault Tolerance

Redundancy is the essential ingredient to achieve fault tolerance in a network. In native Ethernet fault tolerance depends on special protocols like the Spanning Tree Protocol (STP), Rapid Spanning Tree Protocol (RSTP) and Multiple Spanning Tree Protocol (MSTP). These protocols switch to a protection path in a few seconds at worst. This is considered enough in LAN environments but they are often insufficient for massive deployment of IP services in provider networks. In these situations it is required protection switching better than 50 ms. This is the quality level offered by SDH in the 1990s.

Other issue related with the STP is the lack of efficiency in some network topologies. This is the case, for example in rings. STP disables one link and the ring topology becomes a linear topology. The result is a partially used network. STP disables redundant links to build an spanning tree for the network but this is not the best way to use resources. This is specially true when the STP decides to disable an expensive, long reach link in a MAN or a WAN.

Again, the quickest solution to this issue is PBB-TE. This technology makes it possible to preconfigure redundant Ethernet connections and perform protection switching to these connections when a failure is detected.

1.4 MULTI-PROTOCOL LABEL SWITCHING

Multi-Protocol Label Switching (MPLS) is a technology designed to speed up IP packet switching in routers by separating the functions of route selection and packet forwarding into two planes:

- *Control Plane*: This plane manages route learning and selection with the help of traditional routing protocols such as *Open Shortest Path First* (OSPF) or *Intermediate System - Intermediate System* (IS-IS).
- *Forwarding Plane*: This plane switches IP packets, taking as a basis short labels prepended to them. To do this, the forwarding plane needs to maintain a switching table that associates each incoming labeled packet with an output port and a new label.

The traditional IP routers switch packets according to their routing table. This mechanism involves complex operations that slow down switching. Specifically, traditional routers must find the longest network address prefix in the routing table that matches the destination of every IP datagram entering the router.

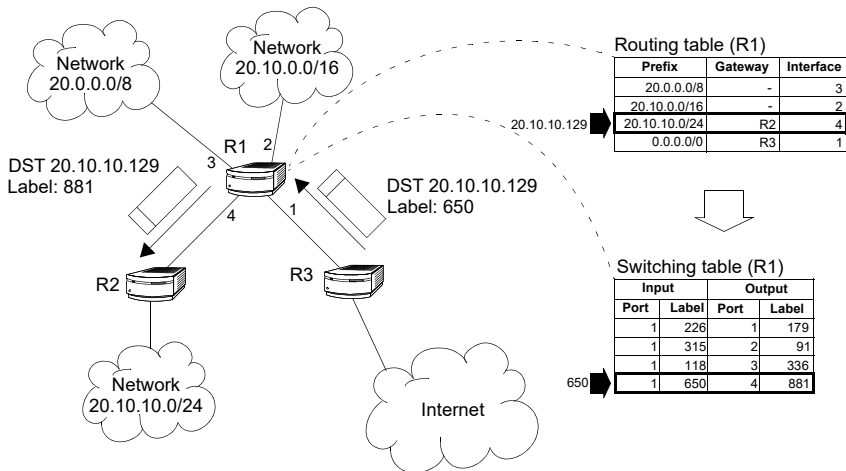


Figure 1.18 Traditional routers have to perform complex operations to resolve the output interface of incoming packets. LSRs resolve the output interface with the help of a simple switching table.

On the other hand, MPLS routers, also known as *Label-Switched Routers* (LSR), use simple, fixed-length label forwarding instead of a variable-length IP network prefix for fast forwarding of packetized data (see Figure 1.18).

MPLS enables the establishment of a special type of virtual circuits called *Label-Switched Paths* (LSP) in IP networks. Thanks to this feature, it is possible to implement resource management mechanisms for providing hard QoS on a per-LSP basis, or to deploy advanced traffic engineering tools that provide the operator with tight control over the path that follows every packet within the network. Both QoS provision and advanced traffic engineering are difficult, if not impossible to solve in traditional IP networks.

To sum up, the separation of two planes allows MPLS to combine the best of two worlds: the flexibility of the IP network to manage big and dynamic topologies automatically, and the efficiency of connection-oriented networks by using preestablished paths to route the traffic in order to reduce packet process on each node.

1.4.1 Labels

When Ethernet is used as the transport infrastructure, it is necessary to add an extra “shim” header between the IEEE 802.3 MAC frames and the IP header to carry the MPLS label. This MPLS header is very short (32 bits), and it has the following fields (see Figure 1.19):

- *Label (20 bits)*: This field contains the MPLS label used for switching traffic.
- *Exp (3 bits)*: This field contains the experimental bits. It was first thought that this field could carry the 3 Type-of-Service (ToS) bits defined for Class of Service (CoS) definition in the IP version 4, but currently, the ToS field is being replaced by 6-bit *Differentiated Services Code Points* (DSCP). This means that only a partial mapping of all the possible DSCPs into the Exp bits is possible.
- *S (1 bit)*: This bit is used to stack MPLS headers. It is set to 0 to show that there is an inner label, otherwise it is set to 1. Label stacking is an important feature of MPLS, because it enables network operators to establish LSP hierarchies.
- *TTL (8 bits)*: This field contains a *Time To Live* value that is decremented by one unit every time the packet traverses an LSR. The packet is discarded if the value reaches 0.

MPLS can be used in SDH transport infrastructures as well. IP routers with SDH interfaces can benefit from the advantages of MPLS like any other IP router. Since the MPLS header must be inserted between layer-2 and layer-3 headers, it was necessary to encapsulate MPLS-labeled frames into Ethernet MAC frames before they are mapped to SDH. However, newer ITU-T recommendations allow direct mapping of MPLS-labeled packets to GFP-F for transport across NG-SDH circuits. This is an important exception of the common frame labeling, because in this case labels are inserted between a layer-1 header (GFP-F) and a layer-3 header (IP). This new mapping improves efficiency of SDH LSRs by eliminating the need of a passive Ethernet encapsulation used only for adaptation (see Figure 1.20).

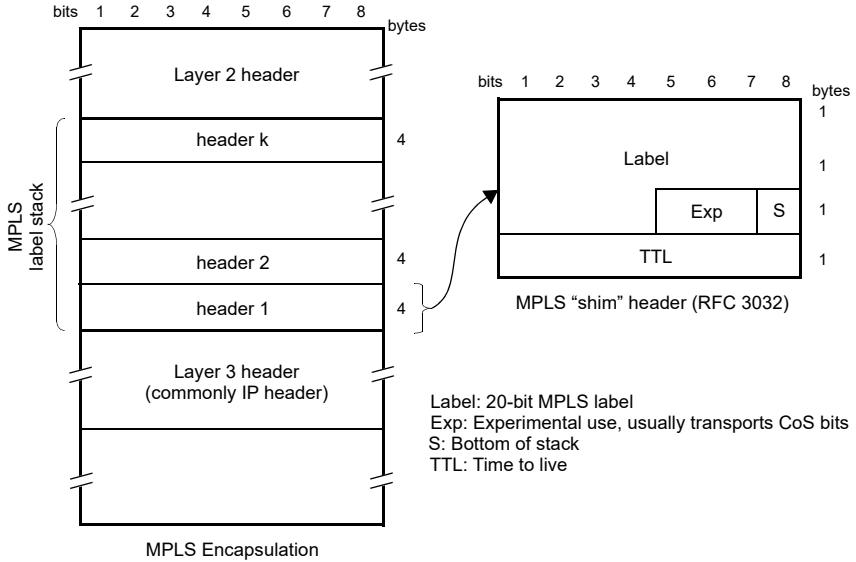


Figure 1.19 MPLS “shim” header format. The label is usually inserted between layer-2 and layer-3 headers.

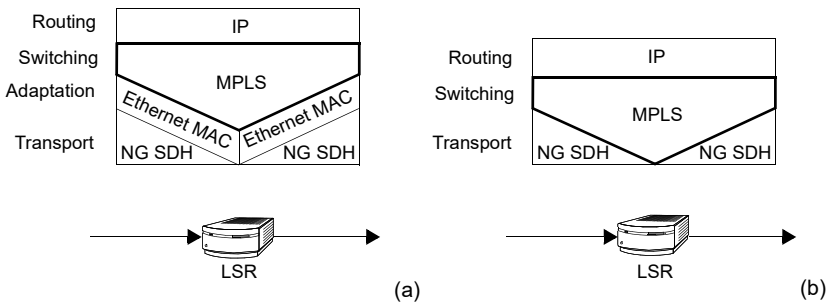


Figure 1.20 Protocol stacks of SDH LSRs. (a) Traditional protocol stack for an SDH LSR: the MPLS header is inserted between layer-2 (Ethernet MAC) and layer-3 (IP) headers. (b) Direct mapping of MPLS over SDH: The MPLS header is mapped between a layer-1 (GFP-F) overhead and layer-3 (IP) overhead without the need of a passive Ethernet encapsulation only used for adaptation.

The MPLS label is sometimes included in the “shim” header inserted between the layer-2 and layer-3 headers, but this is not always true. Almost any header field used for switching can be reinterpreted as an MPLS label. The FR 10-bit *Data Link Con-*

nection Identifier (DLCI) field or the ATM *Virtual Path Identifier* (VPI) and *Virtual Circuit Identifier* (VCI) are two examples of this. The ATM VPI / VCI example is of special importance, because it allows a smooth transition from the ATM-based network core to an IP / MPLS core. An ATM switch can be used as an LSR with the help of relatively simple upgrade.

MPLS has proved to be a technology with incredible flexibility. Timeslot numbers in TDM frames, or even wavelengths in WDM signals can be re-interpreted as MPLS labels as well. This approach opens the door to a new way of managing TDM / WDM networks. The MPLS-based management plane for TDM / WDM networks is compatible with distributed IP routing, and at the same time it benefits from the powerful traffic engineering features of MPLS. This, in fact, forms a new technology and known as Generalized MPLS (GMPLS).

1.4.2 MPLS Forwarding Plane

Whenever a packet enters an MPLS domain, the ingress router, known as ingress *Label Edge Router* (LER), inserts a header that contains a label that will be used by the LSR to route packets to their destination. When the packet reaches the edge where the egress router is, the label is dropped and the packet is delivered to its destination (see Figure 1.21). Only input labels are used for forwarding the packets within the network, while encapsulated addresses like IP or MAC are completely ignored.

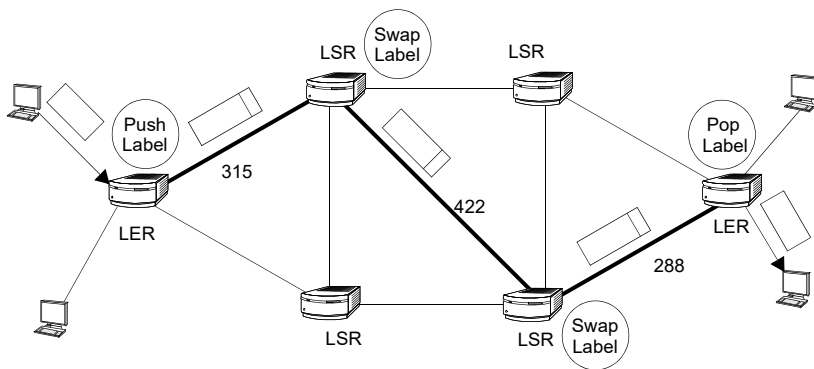


Figure 1.21 Label processing within an MPLS domain. A label is pushed by the ingressing LER, swapped by the intermediate LSR across the LSP, and popped by the egressing LER.

In typical applications, labels are chosen to force the IP packets to follow the same paths they would follow if they were switched with routing tables. This means that the entries in the LSR routing tables must be taken into account when assigning labels to packets and building switching tables.

The set of packets that would receive the same treatment by an LSR (i.e., packets that will be forwarded towards the same destination network) is called *Forwarding Equivalence Class* (FEC). LSRs bind FECs with label / port pairs. For example, all packets that must be delivered to the network 20.10.10.0/24 constitute an FEC that might be bound to the pair (4, 881). All packets directed to that network will be switched to the port 4 with label 881. The treatment that packets will receive on the next hop depends on the selection of the outgoing label. In our example, a packet switched to the port 4 with label 882 will probably never arrive to network 20.10.10.0/24. An LSR may need to request the right label at the next hop to ensure that the packets will receive the desired treatment and that they will be forwarded to the correct destination.

The most common FECs are defined by network address prefixes stored in the routing tables of LSRs. In the routing table, the network prefix determines the outgoing interface for the set of incoming packets matching this prefix. If we wish to emulate the behaviour of a traditional IP router, every network prefix must be bound with a label.

Within the MPLS domain, labels only have a local meaning, which is why the same label can be re-used by different LSRs. For the same packet, the value of the label can be different at every hop, but the path a packet follows in the network is totally determined by the label assigned by the ingressing LER. The sequence of labels [315, 422, 288] defines an LSP route, all packets following the LSP receive the same treatment in terms of bandwidth, delays, or priority enabling specific treatment to each traffic flow like voice, data or video. There are two LSP types (see Figure 1.22):

- *Hop-by-hop LSPs* are computed with routing protocols alone. MPLS networks with only hop-by-hop LSP route packets are like traditional IP networks but with enhanced forwarding performance provided by label switching.
- *Explicit LSPs* are computed by the network administrators to meet specific purposes, and configured either manually in the LSRs, or with the help of the management platform. The path followed by the packets forwarded across explicit LSPs may be different from the paths computed by routing protocols. They can be useful to improve network utilization or select custom paths for certain packets. The ability to provide explicit LSPs converts MPLS into a powerful traffic engineering tool.

An explicit LSP can be strict or loose, depending on how it is established:

- If all the hops that constitute the explicit LSP are specified one by one, the LSP is said to be strictly specified.
- If some but not all the hops that constitute the explicit LSP are specified but some others are left to the decision of the distributed routing algorithms, the LSP is said to be loosely specified.

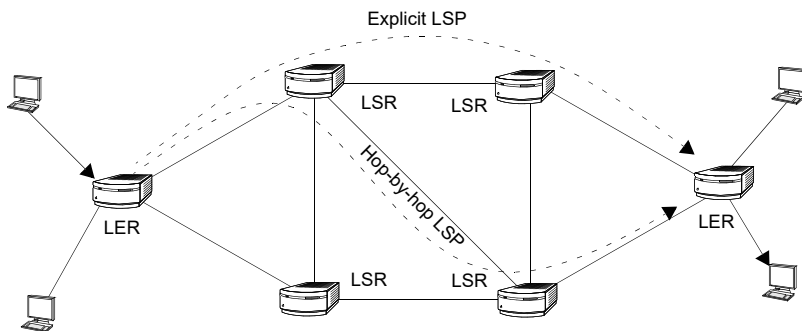


Figure 1.22 A hop-by-hop LSP and an explicit LSP between the same source and destination. The hop-by-hop LSP is computed by the routing protocols running in the LSRs. The explicit route is computed by an external *Network Management System* (NMS).

1.4.3 Label Distribution

The LSR needs to know which label to assign to outgoing packets to make sure they arrive to the correct destination. The obvious way to do this is to configure the switching tables manually in every LSRs. Of course, this approach is not the best possible if there are many LSPs dynamically established and released. To deal with this situation a label distribution protocol is needed.

A label distribution protocol enables an LSR to tell other LSRs the meaning of the labels it is using, as well as the destination of the packets that contain certain labels. In other words, by using a label distribution protocol the LSR can assign labels to FECs.

The RFC 3036 defines the *Label Distribution Protocol* (LDP) that was specifically designed for distributing labels. As MPLS technology evolved, this protocol showed its limitations:

- *It can only manage hop-by-hop LSPs.* It cannot establish explicit LSPs and therefore does not allow traffic engineering in the MPLS network.
- *It cannot reserve resources on a per-LSP basis.* This limits the QoS that can be obtained with LSPs established with LDP.

The basic LDP protocol is extended in RFC 3212 to support these and some other features. The result is known as the *Constraint-based Routed LDP* (CR-LDP). Another different approach is to extend an external protocol to work with MPLS. This is the idea behind the *ReSerVation Protocol with Traffic Engineering extension* (RSVP-TE) as defined in RFC 3209. The original purpose of the RSVP is to allocate

and release resources along traditional IP routes, but it can be easily extended to work with LSPs. The traffic engineering extension allows this protocol to establish both strict and loose explicit LSPs.

1.4.3.1 The Label Distribution Protocol

The LDP enables LSRs to request and share MPLS labels. To do this it uses four different message types.

1. *Discovery messages* announce the presence of LSRs in the network. LSRs send “Hello” messages periodically, to announce their presence to other LSRs. These “Hello” messages are delivered to the 646 UDP port. They can be unicast to a specific LSR or multicast to all routers in the subnetwork.
2. *Session messages* establish, maintain and terminate sessions between LDP peers. To share label to FEC binding information, two LSRs need to establish an LDP session between them. Sessions are transported across the reliable TCP protocol and they directed to port 646.
3. *Advertisement messages* create, modify or delete label mappings for FECs. To exchange advertisement messages, the LSRs must first establish a session.
4. *Notification messages* are used to deliver advisory or error information.

The most important LDP messages are (see Figure 1.23):

- The *Label Request Message*, used by the LSR to request a label to bind with an FEC that is attached to the message. The FEC is commonly specified as a network prefix address.
- The *Label Mapping Message*, distributed by the LSR to inform a remote LSR on which label to use for a specific FEC.

The LSR can request a label for an FEC by using request messages, but it can also deliver labels to FEC bindings without explicit request from other LSRs. The former is an operation mode called *Downstream on Demand*, and the latter is known as *Downstream Unsolicited*. Both modes can be used simultaneously in the same network.

Regarding the behaviour of LSRs when they operate in the Downstream on Demand mode, receiving label request messages, there are two different options:

- *Independent label distribution control*: LSRs are allowed to reply to label requests with label mappings whenever they desire, for example immediately after the request arrives. This mode can be compared to the *Address Resolution Protocol* (ARP) used in LANs to request mappings between destination IP addresses and MAC addresses.

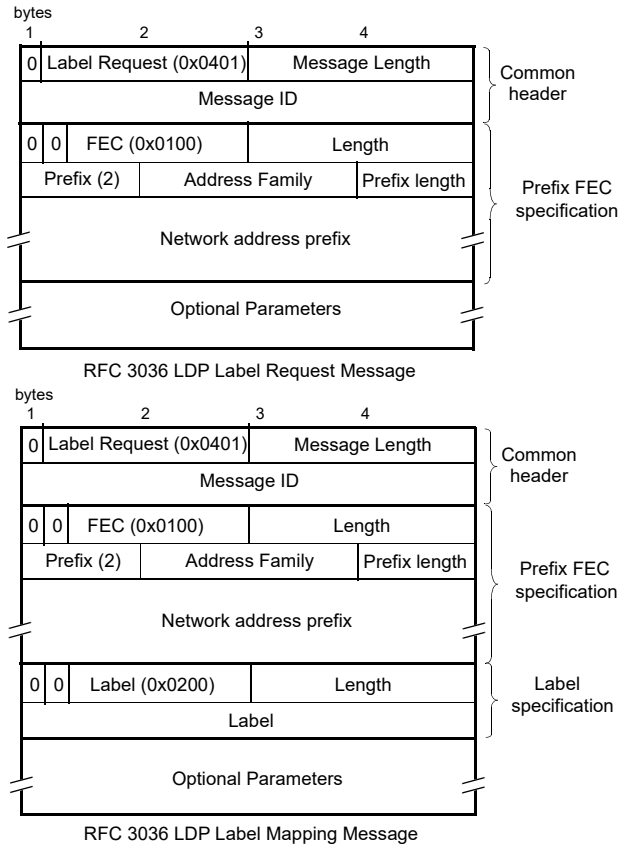


Figure 1.23 Two important LDP messages. The *Label Request Message* requests a label from a remote LSR for binding with an FEC that is attached to the message. The *Label Mapping Message* is used to inform a remote LSR on which label to use for a specific FEC.

- *Ordered label distribution control* (see Figure 1.24): LSRs are not allowed to reply to label requests until they know what to do with the packets belonging to the mapped FEC. In other words, LSRs cannot map an FEC with a label unless they have a label for the FEC, or if they are egress LERs themselves. When an LSR operates in this mode, it propagates the label requests downstream and waits for a reply before replying upstream.

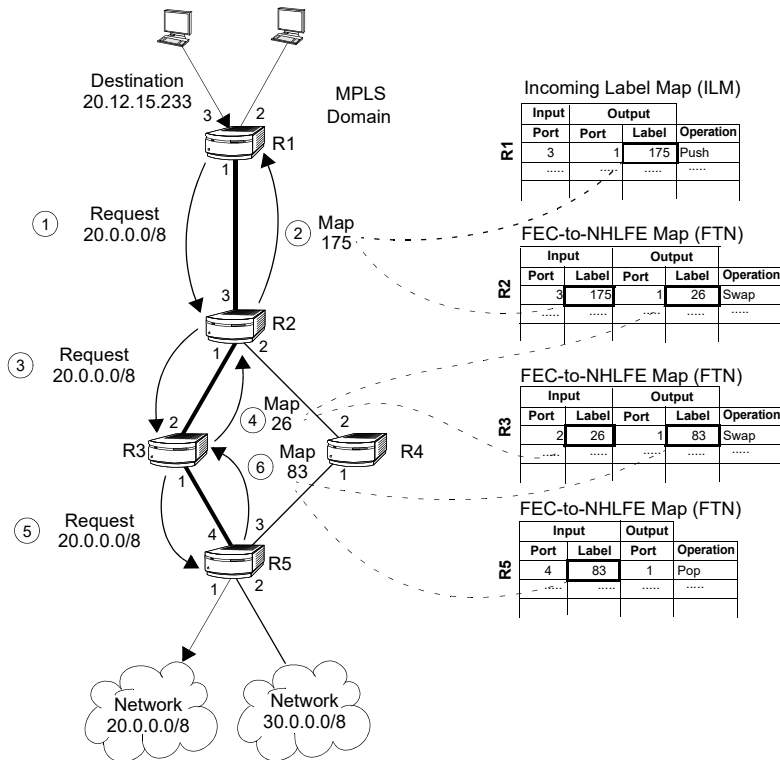


Figure 1.24 LDP in Downstream-on-Demand and independent label distribution mode. LSRs in the LSP generate label requests. The replies they receive from LSRs upstream are used to fill their switching tables.

1.4.4 Martini Encapsulation

In the MPLS network, only the ingress and egress LERs are directly attached to the end-user equipment. This makes them suitable for establishing edge-to-edge sessions to enable communications between remote users. In this network model, the roles of LSRs and LERs would be:

- *LSRs* are in charge of guiding the frame through the MPLS network, using either IP routing protocols or paths that the network administrator has chosen by means of explicit LSPs.
- The *Ingress LER* is in charge of the same tasks as any other LSR, but it also establishes sessions with remote LERs to deliver traffic to the end-user equipment attached to them.

- The *Egress LER* acts as the peer of the ingress LER in the edge-to-edge session, but it does not need to guide the traffic through the MPLS network, because the traffic leaves the network in this node and it is not routed back to it.

There is an elegant way to implement the discussed model without any new overhead or signaling: by using label stacking. This model needs an encapsulation with a two-label stack known as the *Martini encapsulation* (see Figure 1.25):

- The *Tunnel label* is used to guide the frame through the MPLS network. This label is pushed by the ingress LER and popped by the egress LER, but it can also be popped by the penultimate hop in the path, because this LSR makes the last routing decision within the MPLS domain, thus making the Tunnel label unnecessary for the last hop (the egress LER).
- The *VC label* is used by the egress LER to identify client traffic and forward the frames to their destination. The way the traffic reaches end users is a decision taken by the ingress and egress nodes, and it does not involve the internal LSRs. The VC label is therefore pushed by the ingress LSR and popped by the egress LSR.

In the non-hierarchical one-label model, all the routers in the LSP participate in establishing an edge-to-edge session, and all are involved in routing decisions as well. A two-label model involves two types of LSPs. The tunnel LSP may have many hops, but the VC LSP has only two nodes, the ingress and egress LERs. VC LSPs can be interpreted as edge-to-edge sessions that are classified into groups and delivered across the MPLS network within Tunnel LSPs (see Figure 1.26). Tunnel LSPs are established and released independently of the VC LSPs. For example, Tunnel LSPs can be established or modified when new nodes are connected to the network, and VC LSPs could be set up when users wish to communicate between them.

The two-label model makes routing and session management independent of each other. It is not necessary to maintain status information about sessions in the internal LSRs. All these tasks are carried out by LERs. The signaling of the VC LSP is also different from that of the Tunnel LSP. While establishing a Tunnel LSP may require specific QoS or it may depend on administrative policies relying on traffic engineering, VC LSPs are much more simple. This is the reason why label distribution of Tunnel LSPs is carried out with the CR-LDP or the RSVP-TE protocols, but VC LSPs can be managed with the simple LDP.

Although the two-label approach is valid for any MPLS implementation, it has been defined to be used with pseudowires.

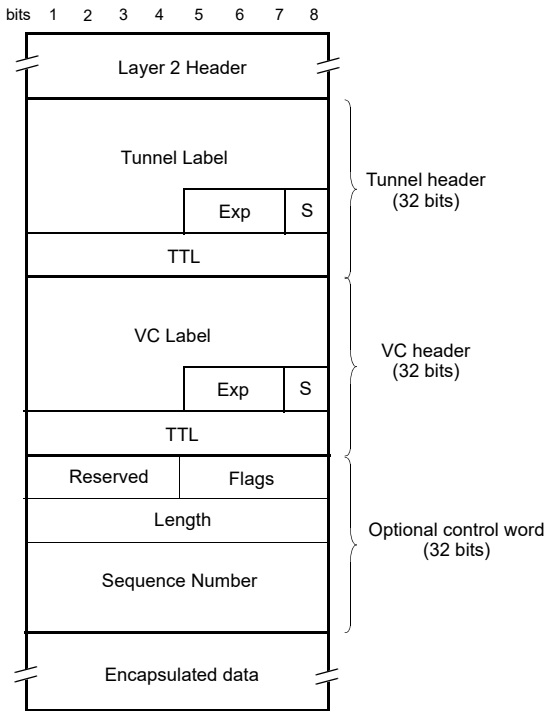


Figure 1.25 Two-label MPLS stack with a Tunnel label and a VC label. The control word may be required when carrying non-IP traffic.

1.4.5 Pseudowires

Pseudowires are entities that carry the essential elements of layer-2 frames or TDM circuits over a packet-switched network with the help of MPLS¹. The standardization of pseudowires is driven by the demand of *Virtual Private Wire Services* (VPWS) that can transport Ethernet, FR, ATM, PPP, SDH, Fiber Channel and other technologies in a very flexible and scalable way. This fact moved the IETF to create the *Pseudowire Edge-to-Edge Emulation* (PWE3) working group that generates standards for encapsulations, signaling, architectures and applications of pseudowires.

1. Although it is possible to implement pseudowires without MPLS, it is used in all the important solutions due to its better performance when compared to other options.

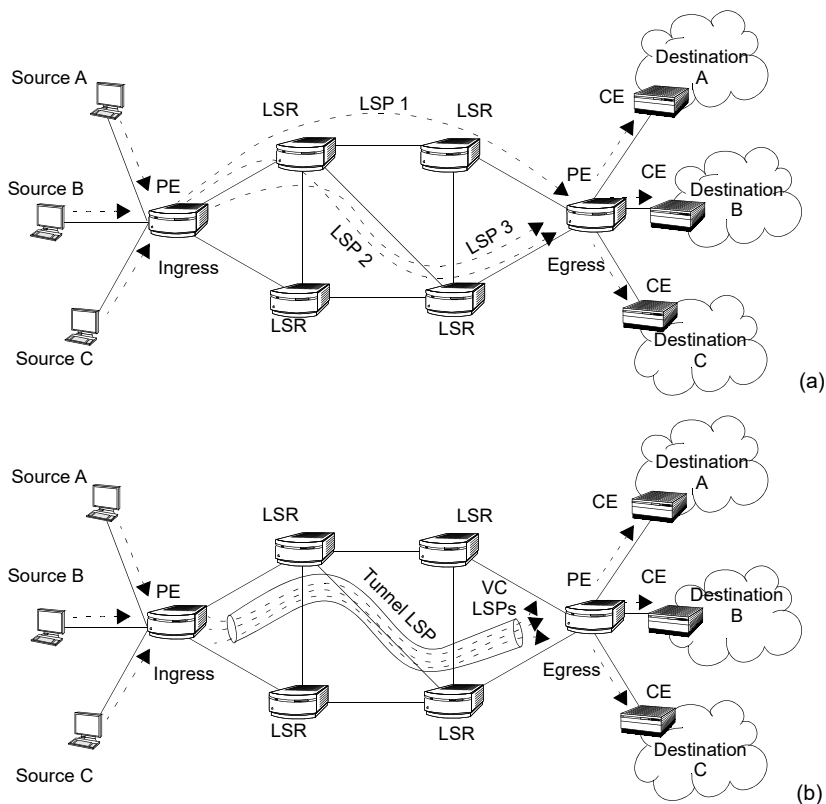


Figure 1.26 (a) One-label approach: the decision to establish routing and edge-to-edge sessions is shared between all the routers. (b) Two-label model: edge-to-edge sessions are tunneled, and internal LSRs are unaware of them.

The concept of pseudowire relies on a simple fact: within the MPLS network, only labels are used to forward the traffic, and any other field located in the payload that could be used for switching is ignored. This means that the data behind the MPLS header could be potentially anything, not limited to an IP datagram. The advanced QoS capabilities of MPLS, including resource management with the RSVP-TE or the CR-LDP protocols make it suitable for transporting traffic subject to tight delay and jitter constraints, including SDH and other technologies based on TDM frames (see Figure 1.27).

It is worth noting that although in MPLS-based pseudowires IP datagrams are replaced by layer-2 or TDM data, IP routing is still an important part of the network. OSPF, IS-IS or other routing protocols are still necessary to find routes in the service provider network when they are not explicitly defined in the LSP setup process. This

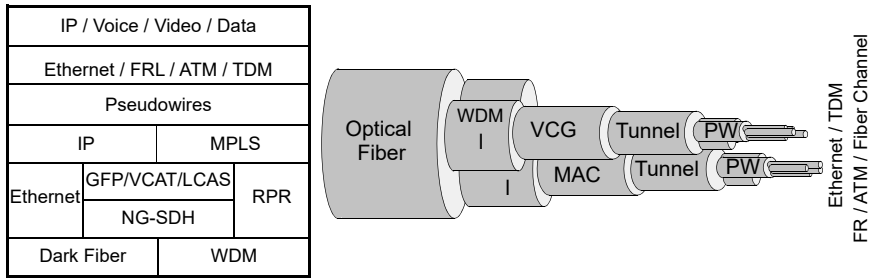


Figure 1.27 Pseudowires can encapsulate and transport ATM, FR, Ethernet, PPP, TDM or Fiber Channel, which is why these protocols do not need a dedicated network unifying the transport in one single network and interface.

means that in the MPLS network carrying pseudowires, IP numbering must be maintained in the network interfaces, because IP routing protocols need IP addresses to work.

Pseudowires are tunneled across the packet-switched network (see Figure 1.28). Any network capable of providing tunnels can be used as a transport infrastructure. By far, MPLS is the most common transport infrastructure for pseudowires, but pure IP networks can be used for the same purpose as well. The MPLS-based pseudowires use LSPs as tunnels, but other tunnels can also be used. Examples are *Generic Routing Encapsulation (GRE)* tunnels or *Layer-2 Tunneling Protocol (L2TP)* tunnels.

Many pseudowires are allowed to be multiplexed in the same tunnel, and therefore it is necessary to identify them. For this reason MPLS architectures need two labels for carrying pseudowires: the first to identify the tunnel and the second to identify the pseudowire. The tunnel / VC double labeling is applied to this case. Here, the VC label becomes the pseudowire identifier, and it is therefore known as the *PseudoWire (PW)* label.

In the traditional MPLS applications, FECs are specified by means of IP addresses or IP network prefixes. Once a label is bound with a network prefix, the network node automatically knows how to forward those packets that carry this label. However, this simple approach does not work with pseudowires, because they carry non-IP data. It is necessary to specify a new way to tell the pseudowire end points how to process the data carried by the pseudowire. This means that new ways of specifying FECs must be defined. Furthermore, each technology may need its own FEC

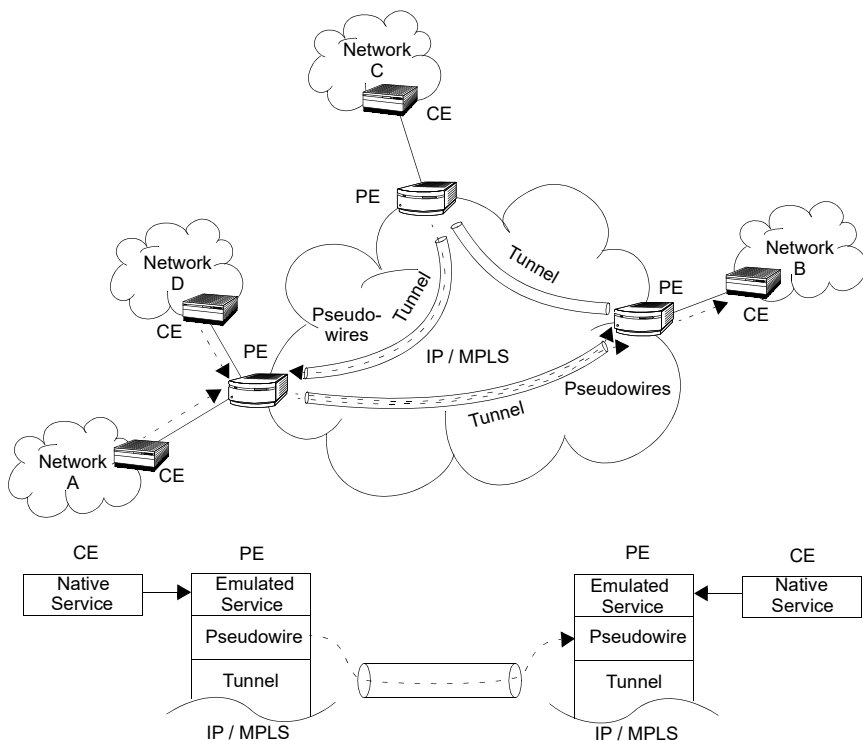
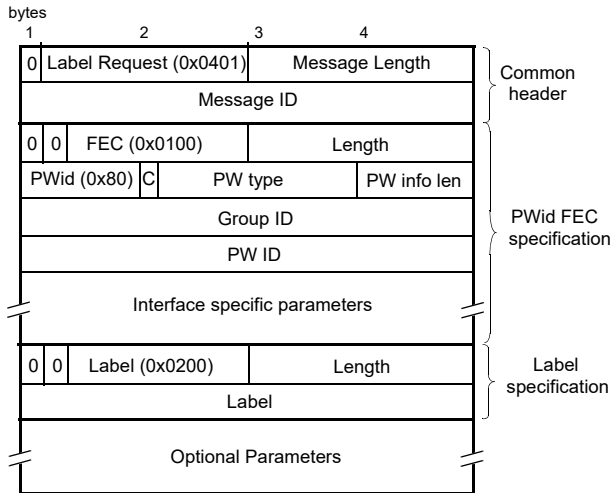


Figure 1.28 Emulation of connectivity services over pseudowires and tunneling across an IP / MPLS network.

specification. For example, forwarding Ethernet frames from or to pseudowires depends on the physical port and the VLAN tag, but this is not necessarily true for ATM or SDH pseudowires. This problem is addressed by extending the LDP protocol to work with pseudowires (see Figure 1.29).

The existing definitions are generalistic and have different interpretations for different types of pseudowire. This is the reason why the new FEC specifications include a 16-bit field for choosing the service emulated over the packet-switched network (see Table 1.2).

Sometimes, it must be ensured that packets are received in the correct order. Other times it is necessary to pad small packets with extra bits, or add technology-specific control bits. To deal with these issues, an extra 32-bit word may be inserted between the PW label and the encapsulated data (see Figure 1.25). The presence of this control word is sometimes required, other times optional, and occasionally not required



RFC 3036 / RFC 4447 LDP PW Label Mapping Message

Figure 1.29 The LDP Label-Mapping message as used to map a pseudowire to an MPLS label. The label-to-pseudowire binding is done by using the PWid FEC element specified in RFC 4447.

at all, depending on the type of pseudowire used. The presence of the control word is signaled in the LDP protocol when the pseudowire is established.

Table 1.2 The existing types of pseudowire

PW type	Description
0x0001	Frame Relay DLCI (Martini mode)
0x0002	ATM AAL5 SDU VCC transport
0x0003	ATM transparent cell transport
0x0004	Ethernet tagged mode
0x0005	Ethernet
0x0006	HDLC
0x0007	PPP
0x0008	SONET / SDH Circuit Emulation Service over MPLS
0x0009	ATM n-to-one VCC cell transport
0x000a	ATM n-to-one VPC cell transport
0x000b	IP Layer2 transport
0x000c	ATM one-to-one VCC cell mode
0x000d	ATM one-to-one VPC cell mode
0x000e	ATM AAL5 PDU VCC transport
0x000f	Frame Relay port mode

Table 1.2 The existing types of pseudowire

PW type	Description
0x0010	SONET / SDH circuit emulation over packet
0x0011	Structure-agnostic E1 over packet
0x0012	Structure-agnostic T1 (DS1) over packet
0x0013	Structure-agnostic E3 over packet
0x0014	Structure-agnostic T3 (DS3) over packet
0x0015	CESoPSN basic mode
0x0016	TDMoIP AAL1 mode
0x0017	CESoPSN TCM with CAS
0x0018	TDMoIP AAL2 mode
0x0019	Frame Relay DLCI

1.4.6 Ethernet Pseudowires

The aim of Ethernet pseudowires is to enable transport of Ethernet frames across a packet-switched network and emulate the essential attributes of Ethernet LANs, such as MAC frame bridging or VLAN filtering across that network.

Standardization of pseudowires enables IP / MPLS networks to transport Ethernet efficiently. The Ethernet pseudowire is perhaps the most important type of pseudowire, because it can be used by network operators to fix some of the scalability, resilience, security and QoS problems of standard Ethernet bridges, thus making it possible to offer a wide range of carrier grade, point-to-point and multipoint-to-multipoint Ethernet services, including EPL, EVPL, EPLAN and EVPLAN.

Provider Edge (PE) routers with Ethernet pseudowires can be understood as network elements with both physical and virtual ports. The physical ports are the attachment circuits where Customer Edge (CE) are connected through standard Ethernet interfaces. The virtual ports are Ethernet pseudowires. Frames are forwarded to physical or virtual ports, depending on their incoming port and VLAN tags. These network elements may also include flooding and learning features to bridge frames to and from physical ports and Ethernet pseudowires, thus making it possible to offer emulated multipoint-to-multipoint LAN services. Many of these PE routers are also able to shape and police Ethernet traffic to limit traffic ingressing in the service provider network.

When a new PE router is connected to the network, it must create tunnels to reach remote PE routers. The remote router addresses may be provided by the network administrators but many PE routers have autodiscovery features. Once the tunnels are established, it is possible to start the pseudowire setup with the help of LDP signaling. LDP mapping signals tell the remote PE routers to which physical port and to which VLANs frames with specified PW labels (see Figure 1.30) will be switched.

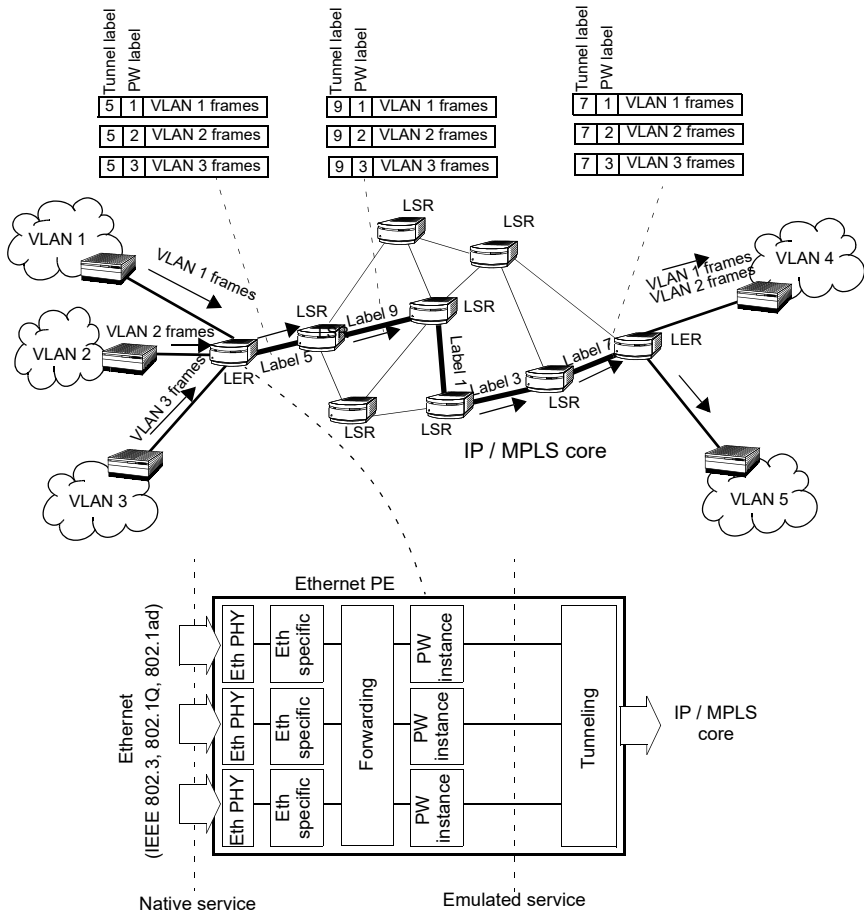


Figure 1.30 Operation of Ethernet pseudowires. The PE router becomes an Ethernet bridge with physical and virtual ports. Physical ports are connected to CEs with standard Ethernet interfaces. Virtual ports are Ethernet pseudowires tunneled across the IP / MPLS core.

The physical attachment circuits of the PE router are standard Ethernet interfaces. Some of them may be trunk links with VLAN-tagged MAC frames, or even double VLAN-tagged Q-in-Q frames. Regarding how VLAN tags are processed, the PE routers have two operation modes:

- *Tagged mode:* The MAC frames contain at least one service-delimiting VLAN tag. Frames with different VLAN IDs may belong to different customers, or if they belong to the same customer, they may require different treatment in the

service provider network. MAC frames with service-delimiting VLAN tags may be forwarded to different pseudowires or mapped to different Exp values for custom QoS treatment.

- *Raw mode*: The MAC frames may contain VLAN tags, but they are not service-delimiting. This means that any VLAN tag is part of the customer VLAN structure and must be transparently passed through the network without processing.

1.4.6.1 Virtual Private LAN Service

The *Virtual Private LAN Service* (VPLS) is a multipoint-to-multipoint service that emulates a bridged LAN across the IP / MPLS core.

VPLS is an important example of a layer-2 *Virtual Private Network* (VPN) service. Unlike more traditional layer-3 VPNs, based on network layer encapsulations and routing, layer-2 VPNs are based on bridging to connect two or more remote locations as if they were connected to the same LAN. Layer-2 VPNs are simple and well suited to business subscribers demanding Ethernet connectivity. VPLS also constitutes a key technology for metropolitan networks. This technology is currently available for network operators who want to provide broadband triple play services to a large number of residential customers.

When running VPLS, the service provider network behaves like a huge Ethernet switch that forwards MAC frames where necessary, learns new MAC addresses dynamically, and performs flooding of MAC frames with unknown destination. In this architecture, PE routers behave like Ethernet bridges that can forward frames both to physical ports and pseudowires.

As with physical wires, bridging loops may also occur in pseudowires. If fact, it is likely that this occurs if the pseudowire topology is not closely controlled, because pseudowires are no more than automatically established LDP sessions. A bridged network cannot work with loops. Fortunately, the STP or any of its variants can be used with pseudowires, as is done with physical wires to avoid them. However, there is another approach recommended by the standards. The most dangerous situation occurs when a PE router relays MAC frames from a pseudowire to a second pseudowire. To avoid pseudowire-to-pseudowire relaying, a direct pseudowire connection must be enabled between each PE router in the network. This implies a full-mesh pseudowire topology (see Figure 1.31). The full-mesh topology is completed with the *split-horizon rule*: It is forbidden to relay a MAC frame from a pseudowire to another one in the same VPLS mesh. Relaying would any way be unnecessary because there is a direct connection with every possible destination.

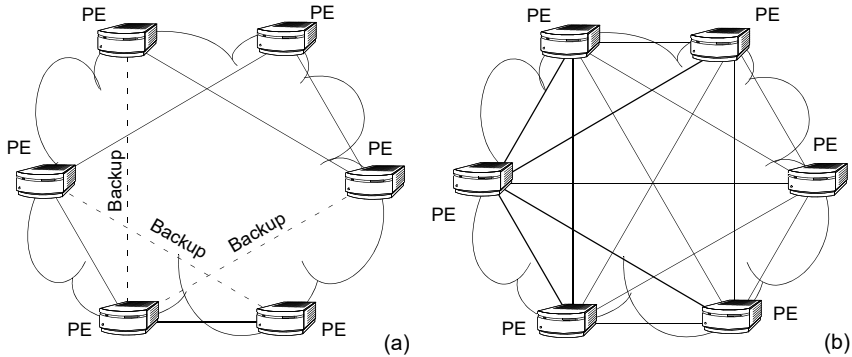


Figure 1.31 Pseudowire topologies in VPLS: (a) Partial mesh with STP. Some of the pseudowires are disabled to avoid loops. (b) Full mesh of pseudowires. The split-horizon rule is applied to avoid bridging loops.

To understand how VPLS works we can think of two end users, S and D, who want to communicate to each other (see Figure 1.32). User S wants to send a MAC frame to user D across a shared network running VPLS.

1. S sends the MAC frame towards D. LAN A is unable to find a local connection to D and finally the frame reaches bridge CE 1 that connects LAN A to a service provider network.
2. Bridge CE 1 forwards S's frame to PE 1 placed at the edge of a VPLS mesh. If PE 1 has not previously learnt S's MAC address, it binds it with the physical port where the frame came from.
3. The PE 1 bridge has not previously learnt the destination address of the MAC frame (D's MAC address), and therefore it floods the frame to all its physical attachment circuits. S's frame reaches LAN B, but D is not connected to it.
4. PE 1 not only performs flooding on its physical ports, but also on the pseudowires. S's frame is thus forwarded to all other PEs in the network by means of direct pseudowire connections across the VPLS mesh.
5. S's frame reaches PE 2 attached to pseudowire PW12. If PE 2 has not previously learnt the received source MAC address, it binds it with pseudowire PW12. In this case, PE 2 does not know where D is, so it flows the MAC frame to all the physical ports and arrives to LAN C, however D is not connected to that LAN. Following the split-horizon rule, the frame is not flooded to other pseudowires.
6. S's frame reaches PE 4. It learns S's MAC address if it is unaware of it. After learning, S's address is bound to pseudowire PW14. In this case PE 4 has previously bounded D's address to pseudowire PW34, and therefore it does not

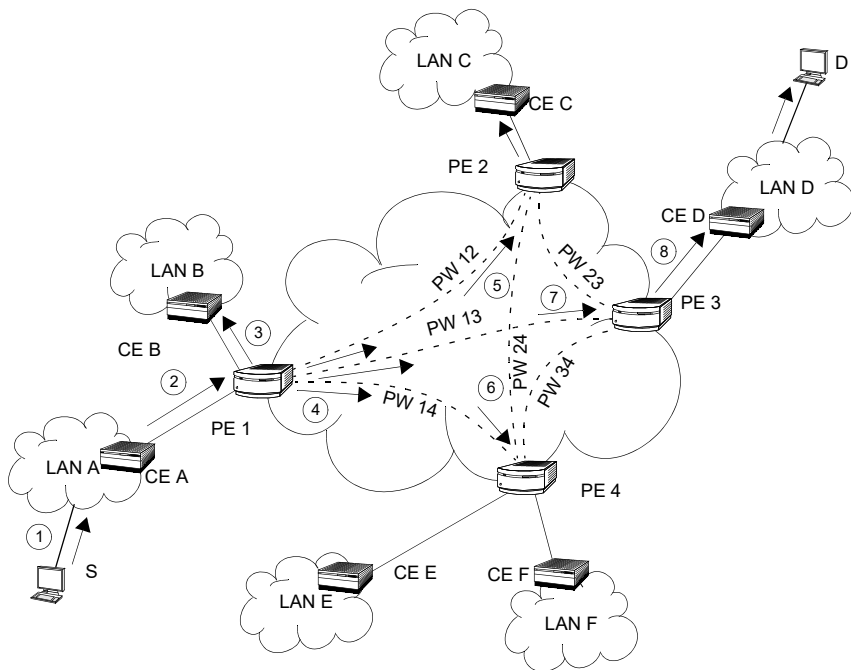


Figure 1.32 Flooding and learning in VPLS serves emulate a LAN broadcast domain.

forward S's frame to LANE or LANF. The frame is not forwarded to pseudowire PW 4 either, because of the split-horizon rule.

7. S's frame reaches PE 3. This router performs the same learning actions as PE 2 and PE 4 if needed, and binds S's MAC address to pseudowire PW13. In this case, PE 3 has previously learnt that D can be reached by one of its physical ports, and therefore it forwards S's frame to it.
8. S's frame reaches CE D that forwards this frame to its final destination.

The previous example deals with a single broadcast domain that appears as a single distributed LAN. But this may not be acceptable when providing services to many customers. Every customer will normally require its own broadcast domain. The natural way to solve this is by means of VLANs. Every subscriber is assigned a service-delimiting VLAN ID. Every VLAN is then mapped to a VPLS instance (i.e., a broadcast domain) with its own pseudowire mesh and learning tables. The link between CE and PE routers is multiplexed, and customers are identified by VLAN tags. This deployment is useful for offering EVPLAN services as defined by the MEF.

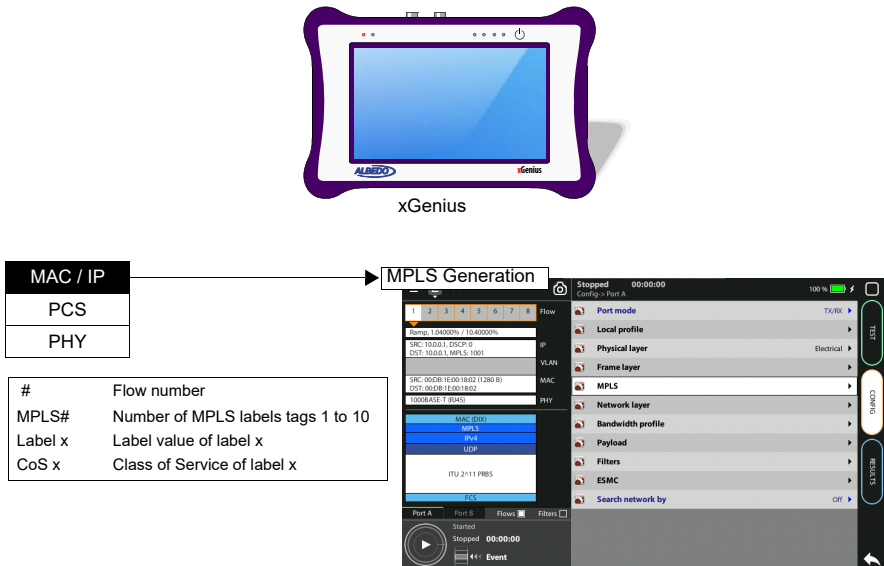


Figure 1.33 Example of MPLS traffic generation with the ALBEDO xGenius

But VLAN tags are not always meaningful for the service provider network. All VLAN tags can be mapped to a single VPLS instance and therefore all of them are part of the same broadcast domain within the service provider network. In this case VLAN-tagged frames are filtered by the subscriber network, but they are leaved unchanged in the service provider network. Different customers can still be assigned to different broadcast domains, but not on a per-VLAN-ID basis. Mapping customers to VPLS instances on a per-physical-port basis is the solution in this case. This second deployment option is compatible with the EPLAN connectivity service definition given by the MEF.

1.4.6.2 Hierarchical VPLS

VPLS has demonstrated to be flexible, reliable and efficient, but it still lacks scalability due excessive packet replication and excessive LDP signaling. The origin of

the problem is on the full meshed pseudowire topology. The total number of pseudowires needed for a network with n PE routers is $n(n-1)/2$. This limits the maximum number of PE routers to about 60 units with current technology.

Hierarchical VPLS (HVPLS) is an attempt to solve this problem by replacing the full meshed topology with a more scalable one. To do this it uses a new type of network element, the *Multi-Tenant Unit* (MTU). In HVPLS, the pseudowire topology is extended from the PE to the MTU. The MTU now performs some of the functions of the PE, such as interacting with the CE and bridging. The main function of the PE is still frame forwarding based on VLAN tags or labels. In some HVPLS architectures, the PE does not implement bridging. The result is a two-tier architecture with a full mesh of pseudowires in the core and non redundant point-to-point links between the PE and the MTU (see Figure 1.34). A full mesh between the MTUs is not required, and this reduces the number of pseudowires. The core network still needs the full mesh, but now the number of PEs can be reduced, because some of their functions have been moved to the access network.

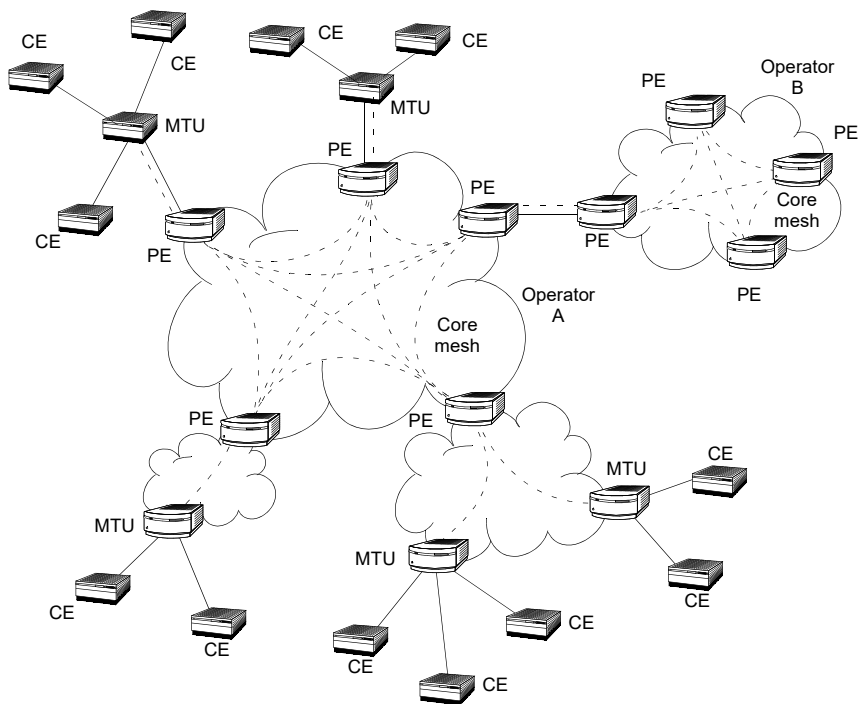


Figure 1.34 In HVPLS, the full mesh of pseudowires is replaced by a two-tier topology with full mesh only in the core and non-redundant point-to-point links in the access.

The MTUs behave like normal bridges. They have one (and only one) active pseudowire connection with the PE per VPLS instance. Flooding, as well as MAC address learning and aging is performed in the pseudowire as if it were a physical wire. The PE operates the same way in an HVPLS as in a flat VPLS, but the PE-MTU pseudowire connection is considered as a physical wire. This means that the split-horizon rule does not apply to this interface.

In practical architectures, the MTUs are not always MPLS routers. Implementations based on IEEE 802.1ad service provider bridges are valid as well. These bridges make use of Q-in-Q encapsulation with two stacked VLAN tags. One of these tags is the service delimiting P-VLAN tag added by the MTU. The P-VLAN designates the customer, and it is used by the PE for mapping the frames to the correct VPLS instance.

HVPLS can be used to extend the simple VPLS to a multioperator environment. In this case, the PE-MTU non-redundant links are replaced by PE-PE links where each PE in the link belongs to a different operator.

The main drawback of the HVPLS architecture is the need for non-redundant MTU-PE pseudowires. A more fault tolerant approach would cause bridging loops. One solution is a multi-homed architecture with only one simultaneous MTU-PE pseudowire active. The STP can help in managing active and backup pseudowires in the multi-homed solution.

1.4.7 MPLS Transport Profile

The transport network must provide aggregation and reliable transmission of large amounts of information. It must be predictable but flexible enough to accept any possible client service or application.

The requirements of the transport network have been fulfilled by various TDM technologies like the *Plesiochronous Digital Hierarchy* (PDH), the *Synchronous Digital Hierarchy* (SDH) / *Synchronous Optical Network* (Sonet) and the *Optical Transport Network* (OTN). More recently MPLS has been proposed as the new transport network technology.

MPLS is different to the SDH / Sonet or the OTN it that is a packet-switching technology. Also in that previous TDM-based transport network technologies are “standalone” in the sense that each of them is all that is needed to build the transport

network, but MPLS requires a server layer acting as the transport infrastructure. We already know that the MPLS transport infrastructure can be either Ethernet or TDM.

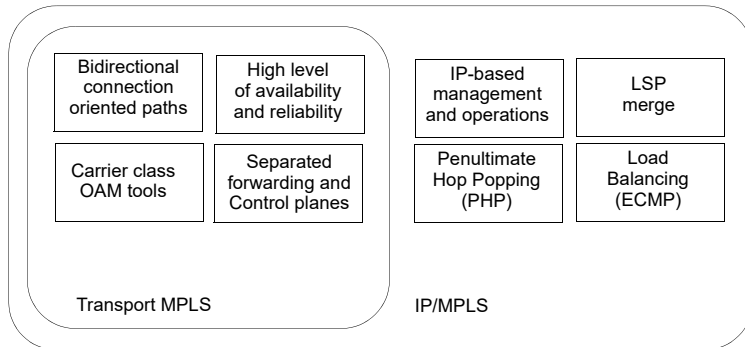


Figure 1.35 MPLS-TP is a strict subset of IP/MPLS. Some IP/MPLS features are left out in MPLS-TP and some other have been defined specifically for MPLS-TP but all them constitute a single MPLS standard.

The IP dependence of MPLS is a more serious issue because transport network operators want a protocol agnostic network. Of course, they may want to transport other applications than IP but also they have their own operation and management mechanisms. These mechanisms are usually centralized in the Network Operations Center (NOC). That means they don't trust the distributed and unpredictable IP management.

The independence of the MPLS layer was addressed for first time by the PWE3 working group when pseudowires were defined. The pseudowire user plane does not require the IP encapsulation. For this reason, some things learnt with pseudowires are also applied to the transport MPLS. However, if MPLS has to be applied to the transport network, it has to be a completely IP-free technology. This is not possible without the introduction of new control and management planes for MPLS.

More specifically, the transport MPLS standards define several types of bidirectional connection-oriented transport paths, protection and restoration mechanisms, comprehensive Operations, Administration, and Maintenance (OAM) functions, and network management procedures free of a dynamic control plane or IP forwarding support. These extensions are defined in a way that makes them applicable also to existing IP/MPLS networks in order to enable the interoperability between both technologies.

In the same way that transport MPLS requires new features, there are some MPLS capabilities available from the beginning that are not needed in transport applica-

tions. Requirements for the MPLS transport network are stated in standard RFC 5654. As defined today, transport MPLS is a strict subset of MPLS, and comprises only those functions that are necessary to meet the requirements of RFC 5654 (see Figure 1.35). These are the major properties of the transport flavour of MPLS not yet mentioned:

- *It is strictly connection-oriented.* If fact, LSPs are always connections within the MPLS domain, but there are some applications where MPLS emulates a connectionless network and provides connectionless services. The most important examples of this, are VPN technologies based on MPLS. VPLS, for example, uses Ethernet pseudowires to emulate a bridged network.
- *Defines bidirectional connections.* TDM transport networks operate exclusively with bidirectional connections. One of the reasons for this is that traditional telephony services are symmetric. There is some interaction between directions of the bidirectional path. For example failures in one direction are reported back using the return path. This mechanism improves reporting capabilities of OAM and makes easier path protection. Transport network operators are interested in keeping this useful properties of TDM transport networks for MPLS-TP but unidirectional point-to-point and point-to-multipoint connections are allowed as well.
- *MPLS-TP is prepared to accommodate any control and management planes.* It is even possible to operate the MPLS-TP network without any control plane and leave static provisioning as the only way to deliver services. As it as been stated, control based on IP routing algorithms and protocols are usually undesirable and thus its usage is discouraged but it is not forbidden. The necessary flexibility to accommodate very different control planes can only be achieved if it is imposed a strict logical separation of the control and management planes from the data plane.
- *Equal-Cost Multi-Path (ECMP) is forbidden in the MPLS transport network.* ECMP is a routing strategy that distributes the traffic directed to one destination over various paths. ECMP improves utilization but it affects network predictability and makes operation more complex.
- *Penultimate Hop Popping (PHP), must be disabled by default on transport LSPs.* With PHP, the top label in the label stack is removed one hop before its destination. PHP is performed in some routers because it reduces the load on the egress LER (more exactly, the load is shared between the egress LER and the penultimate hop). PHP may interfere with end-to-end network procedures like OAM or path protection. It is also a potential problem in IP-less environments.
- *LSP merge is not supported.* LSP merge reuses the same label in different LSPs. It is useful to simplify label management in those situations where traffic

from different LSPs is sent to the same destination. LSP merge hides the traffic source and thus makes more complex network operation and control.

1.4.7.1 MPLS-TP and ITU-T T-MPLS

Discussion on MPLS for transport networks was started by the ITU-T Study Group 15 (SG15) under the acronym of Transport MPLS (T-MPLS). Some recommendations relative to T-MPLS were released in the period from 2005 to 2007 including the ITU-T G.8110.1, G.8112, G.8121, G.8131 and G.8151 addressing different topics like architecture, interfaces or management of the MPLS transport network.

IETF expressed its concern that T-MPLS will break IP/MPLS and cause potentially massive interoperability issues. IETF concern was justified in two points.

- T-MPLS duplicates mechanisms available for IP/MPLS in IETF RFCs. These mechanisms are oriented towards the transport network needs but they are incompatible with IP/MPLS. For example, T-MPLS incorporates new pseudowire types that duplicate existing IETF PWE3 pseudowires.
- T-MPLS and IP/MPLS share the same frame format and forwarding semantics. Particularly, the protocol identifiers are the same for T-MPLS and IP/MPLS. The reserved Ethertypes are 0x8847 (unicast packets) and 0x8848 (multicast packets) for both.

To avoid future interoperability issues, T-MPLS must either use its own Ethertypes or pass through an harmonization process to guarantee compatibility with IETF standards. In February 2008, the ITU-T and IETF agreed to rework T-MPLS to keep compatibility with IETF standards. Based on this agreement, IETF and ITU-T experts started working out the requirements and solutions available for the transport MPLS, now designated the MPLS Transport Profile (MPLS-TP). ITU-T in turn agreed with updating the existing T-MPLS standards based on the MPLS-TP.

To develop MPLS-TP, a Joint Working Team (JWT) was established. The JWT is supported by an IETF Design Team and an Ad Hoc Group on T-MPLS in the ITU-T.

1.4.7.2 MPLS-TP Forwarding Plane

The MPLS-TP data plane or forwarding plane, as defined in RFC 5960, is in agreement with the general MPLS/IP architecture (RFC 3031, RFC 3032).

MPLS-TP accepts IP payloads (that may themselves have MPLS labels) or pseudowire packets. In fact pseudowires are accepted within the own MPLS-TP forwarding architecture. Thanks to this feature, MPLS-TP is suitable for transporting virtually any packet or circuit based technology (see Figure 1.36).

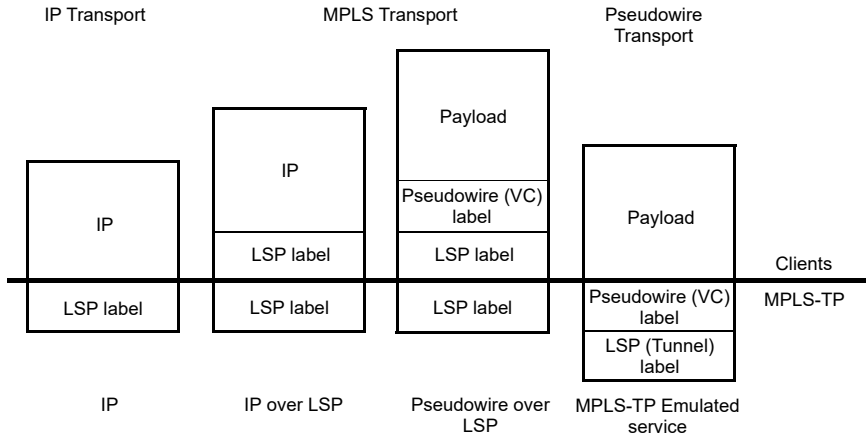


Figure 1.36 MPLS-TP client layers. Any client layer accepted by MPLS is also accepted by MPLS-TP. Pseudowires can be built within the MPLS-TP framework. Thanks to this feature, virtually any client layer can be accepted by MPLS-TP.

MPLS-TP classifies all possible LSPs in four different families that include the traditional ones:

- *Point-to-point unidirectional LSP:* These are equivalent to the LSPs defined for the general MPLS architecture and they operate in the same way.
- *Point-to-point associated bidirectional LSP:* Is a pair of point-to-point unidirectional LSPs configured in opposite directions. These LSPs are regarded as entities providing a single logical bidirectional path.
- *Point-to-point co-routed bidirectional LSP:* Is equivalent to a point-to-point associated bidirectional LSP with the additional requirement that the unidirectional components of the LSP follow the same links and nodes.
- *Point-to-multipoint unidirectional LSP:* This LSP type is equivalent to a point-to-point LSP but it may have more than one egress interface.

Note that multipoint-to-multipoint or multipoint-to-point varieties have been intentionally left out of this classification.

The MPLS-TP uses the LSP stack to define the concept of *section*. This concept has been used by transport network operators for years. Two LSRs define an MPLS-TP section at some MPLS layer if they are adjacent at this layer. MPLS-TP section hierarchy is fundamental for transport network operation and management (see Figure 1.37).

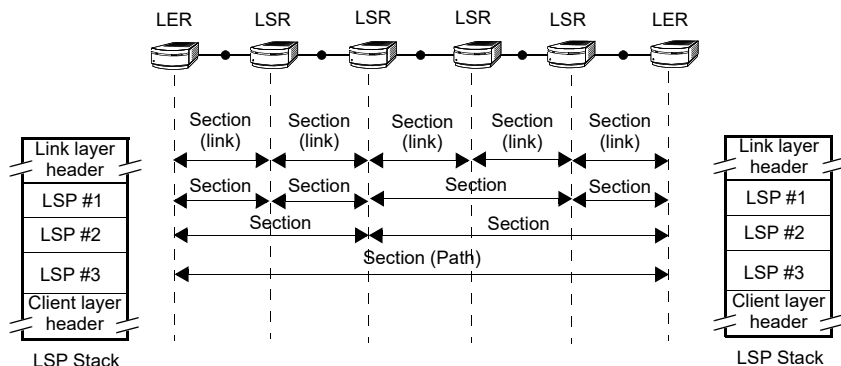


Figure 1.37 MPLS-TP sections are no more than path chunks between adjacent LSRs

1.4.7.3 The Generic Associated Channel

The Generic Associated Channel (GACH) is an extension of the RFC 4385 pseudowire associated channel but it does not need to be associated to a pseudowire. The GACH supports control, management, and OAM traffic associated with MPLS-TP transport entities.

One issue is how to identify and demultiplex user and control traffic transported in the GACH. The MPLS-TP approach to this issue is to reserve one label for the signalling channel. This label is referred as Generic Associated channel Label (GAL) and its value is 13.

To encapsulate the GACH, MPLS-TP adds the GAL to the label stack, immediately after the transport LSP label or labels. This mechanism guarantees that user and control frames will share fate under any circumstance, which is one of the fundamental requirements for control traffic in MPLS-TP networks. If MPLS-TP is using pseudowires, then no innovations are necessary: The pseudowire control word and the pseudowire associated channel are used for user and control traffic demultiplexing (see Figure 1.38). Thanks to this encapsulation for the control information it is possible to extend the same pseudowire over IP/MPLS and MPLS-TP sections without restrictions.

1.4.7.4 MPLS-TE Control and Management Planes

Control and management planes are related but they are separated aspects of the transport network. The control plane decides how to route data plane traffic across the network. If the network is connection-oriented like MPLS-TP the control plane establishes and terminates connections and reserves resources for them based on dif-

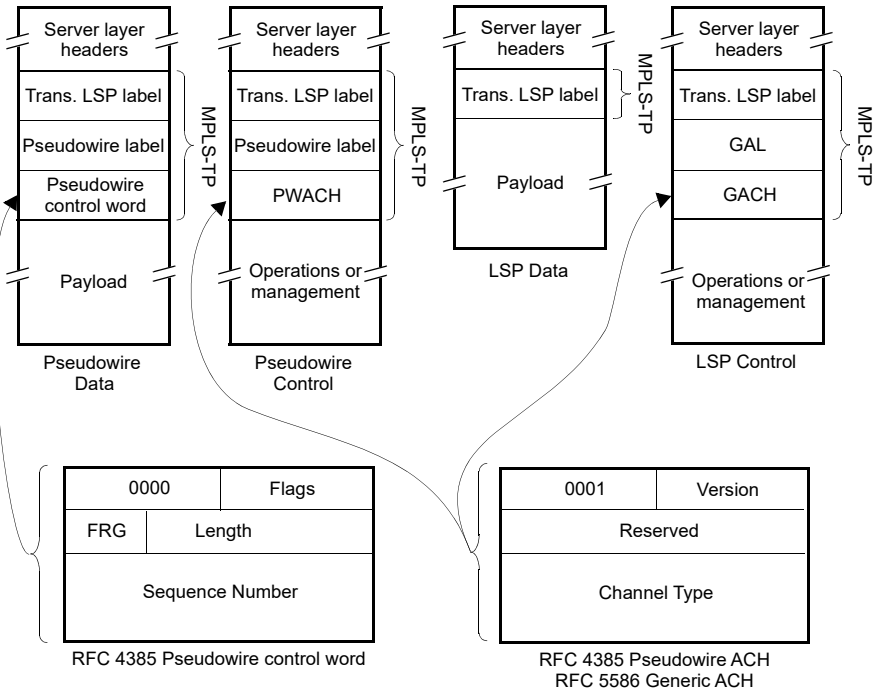


Figure 1.38 MPLS-TP user and GACH PDUs. The packet format for the GACH is inspired in the pseudowire associated channel.

ferent criteria. The role of the management plane is simply to manage the control plane. In fact, the transport network could work without a control plane. This is possible if it is left to the management plane the ability to statically set connections without intervention of any special routing or signalling protocol. Static management of medium sized or large IP networks is very uncommon but carriers are used to operate their transport networks using the management plane. This fits very well in the model of a distributed system controlled from a single central location, the Network Operations Center (NOC). For this reason operation without control plane is optional in MPLS-TP. However, if used, the MPLS-TE control plane must be based on Generalized MPLS (GMPLS) and in the case of transport pseudowires, in the exiting PWE3 control plane.

GMPLS is an automated control plane technology that reinterprets any traffic identifier as a label. In this way, TDM timeslots and WDM fibers and wavelengths are seen as labels. Thanks to this conceptual reinterpretation, MPLS can be extended to virtually any network technology, but GMPLS is specially suited for transport networks such as WDM, SDH and now MPLS-TP.

GMPLS requires generalized label distribution procedures that are not supported by the generic label distribution protocols. Therefore, these protocols have to be extended. The GMPLS versions of CR-LDP and RSVP-TE for GMPLS are the Generalized CR-LDP and the Generalized RSVP-TE.

The second important concept related with GMPLS is traffic engineering. As mentioned, transport networks require a more closer and explicit control of the routing function than standard IP networks. Resource availability, SLA and business plans must be considered for route selection in the transport network, but these routing criteria are not supported by the vanilla version of IP routing protocols.

For this reason, the GMPLS routing function is left to protocols with traffic engineering extensions like OSPF-TE or ISIS-TE. With the help of these protocols, the routing function is in control of manual operators. They monitor the state of the network, route the traffic or provision additional resources to compensate for problems as they arise. Alternatively, these protocols may be driven by automated processes reacting to information fed back.

The last building block for the GMPLS architecture is the *Link Management Protocol* (LMP). The mandatory management capabilities of LMP are control channel management and TE link property correlation. Optionally, LMP may provide physical connectivity verification and fault management.

1.4.8 Hands -on: MPLS-TP Traffic Analysis

The MPLS-TP test option for the xGenius provides the ability to generate and analyze full line rate MPLS-TP data traffic for 10 Mb/s to 10 Gb/s packet transport network links. As a terminate or passive monitor application, it verifies key Service Level Agreement (SLA) Quality of Service (QoS) metrics. It also supports comprehensive MPLS-TP OAM in compliance with both ITU-T pre-standard G.8114 and IETF draft MPLS-TP OAM based on Y.1731. By generating and monitoring OAM messages at pseudowire, label switched path (LSP), or section layer, operating with both label 13 or label 14, proper OAM operation can be verified.

1.4.8.1 Value Proposition

MPLS-TP, an emerging Layer 2 packet-based transport technology is critical to the successful deployment of Carrier Ethernet services driven by high-bandwidth, high-performance applications such as LTE, IP video, and mobile backhaul. As service providers offer and install more packet-based MPLS-TP services for their customers, the ALBEDO xGenius software option provides a cost-effective method for verifying the installation of MPLS-TP services and circuits. The new ALBEDO test suite gives providers confidence that MPLS-TP services are delivered with true carrier-class QoS; with properly functioning end-to-end OAM; as well as protection

switching. By providing both customer data and control plane traffic verification in one easy to use tool, the MPLS-TP test suite saves both installation and troubleshooting time and efforts. Simple to understand pass (green)/fail (red) results as well as detailed traffic and OAM statistics appeal to both expert and novice users.

Table 1.3Feature/Benefit Summary

Feature	Description	Advantage	Benefit
MPLS-TP line rate traffic generation	Configurable MPLS header service provider and customer labels	Flexibility to connect to any point within MPLS-TP network	Proactively verifies correct circuit provisioning before handling live traffic
MPLS-TP SLA/KPI analysis	Reports key metrics of throughput, frame loss, delay, and jitter	Provides repeatable and simple pass/fail results as well as detailed statistics	Ensures that service meets true customer QoS and removes guesswork in troubleshooting
Label 13 or 14 OAM message generation and monitoring	Continuity Check (CC), Loopback (LB), and Alarm Indication Signal (AIS)	Multiple OAM types supported encompassing all network possibilities	Guarantees proper OAM operation with flexible analysis and ubiquitous usage
Simultaneous MPLS-TP customer data and OAM traffic	Real-time OAM analysis with background traffic generation	Emulates true network operation by exposing utilization impact	Comprehensive troubleshooting analysis in one easy to use tool

1.4.8.2 Use Case: End-to-End Traffic and OAM Verification

The xGenius can be used to generate and analyze end-to-end MPLS-TP traffic by connecting a test set to a switch or router port on both the near and far end of the circuit. In this scenario, each test set is configured in terminate mode and is used to transmit test traffic emulating customer data. Detected test traffic can then be analyzed on each test set displaying key traffic metrics such as throughput (bandwidth utilization or CIR), frame loss (FL), round trip delay (FD), and jitter (FDV), as well as MPLS-TP header label information (see Figure 3.39).

In this mode OAM control plane traffic can also be generated and analyzed for OAM verification at turn-up or for troubleshooting scenarios. Link connectivity can be verified using CCM and fault isolation can be identified using loopback/link trace.

Use Case: Passive Monitor Mode

The xGenius can be used to monitor and analyze MPLS-TP traffic by connecting it to a mirror or spare port on a switch or router. In this scenario, the test set is configured in a passive monitor mode and is used to detect live MPLS-TP traffic that is forwarded to this mirror port by the switch. The discovered traffic can then be analyzed on the test set displaying key traffic statistics such as bandwidth utilization, received frame counts, and MPLS-TP header label information including service provider (SP) and customer label ID and priority.

1.4.8.3 Simplified MPLS-TP Setup and Results

User configurable frame header labels are displayed in a clear graphical format for both SP and customer tunnel layers. Filters can be optionally set on the filters tab to further narrow down the detected traffic (see Figure 1.39).

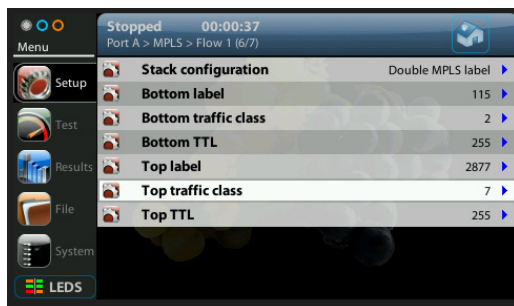


Figure 1.39 The xGenius provides simplified MPLS-TP configuration and result screens

The analyzed traffic can be viewed using tables or graphs, presenting key SLA/KPI measurements and statistics. Errors are instantly revealed and indicated by red warnings, with histograms and absolute time graphs providing essential troubleshooting information.

1.5 QUALITY OF SERVICE

Quality of Service (QoS) is the ability of a network to provide services with predictable performance.

Time Division Multiplexing (TDM) networks are predictable, because performance parameters such as throughput, delay and jitter are constant or nearly constant. Packet-switched networks are much more efficient because of the statistical multiplexing gain, but they have difficulties in controlling the performance parameters. An important goal of next generation packet technologies is to be able to ensure a specific QoS over packet-switched networks.

1.5.1 QoS Control Basics

Packet switched network nodes store the information in queues if the output interface is busy. When data is queuing, the following two points must be taken into account:

1. Packet delay in the queue varies depending on the load in the network.
2. Packets can be discarded if, under high-load conditions, there is no space to store them.

A typical solution to deal with congestion in packet switched networks has been to increase the transmission bandwidth to keep network utilization low. Over provisioning is a good solution when bandwidth is cheap – otherwise it is necessary to find a way to keep delay low and predictable while improving network utilization to the maximum. The current networking technology achieves this by using traffic differentiation and congestion management mechanisms specifically designed for packet switched networks (see Figure 1.40).

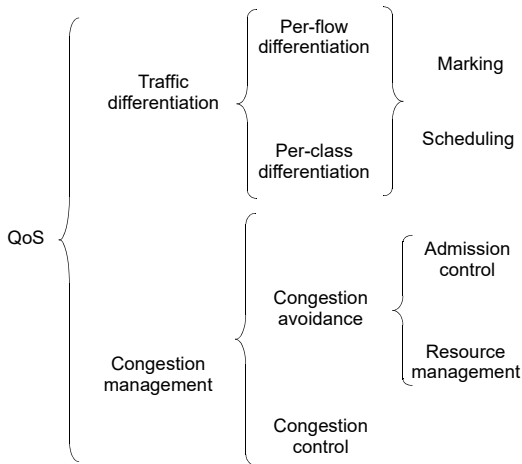


Figure 1.40 It is difficult to achieve good QoS features with one single mechanism. The best way is to mix many elements to get the desired result.

1.5.1.1 Traffic Differentiation

Traffic differentiation separates the bulk traffic load into smaller sets, and treats each set in a customized way. There are two issues related to traffic identification:

1. *Traffic classification.* The traffic is divided into classes or flows. Sometimes it is necessary to explicitly mark the traffic with a *Class-of-Service (CoS)* identifier.
2. *Customized treatment of traffic classes and flows.* Some packets have more privileges than others in network elements. Some may have a higher priority, or there may be resources reserved for their use only.

Traffic differentiation makes it possible to improve performance for certain groups of packets and define new types of services for the packet-switched network.

- *Differentiated services.* We can talk about differentiated services when a part of the traffic is treated ‘better’ than the rest. This way, it is possible to establish some QoS guarantees for the traffic. The QoS defined for differentiated services is also known as soft QoS.
- *Guaranteed services.* Guaranteed services take a step further. They are provided by reserving network resources only for chosen traffic flows. Guaranteed services are more QoS-reliable than differentiated services, but they make efficient bandwidth use difficult. The QoS for guaranteed services is also known as hard QoS.

1.5.1.2 Congestion Management

Congestion is the degradation of network performance due to excessive traffic load. By efficiently managing network resources, it is possible to keep performance with higher loads, but congestion will always occur, sooner or later. So, when delivering services with QoS, one must always deal with congestion, one way or another.

There are two ways to deal with congestion:

1. *Congestion control* is a set of mechanisms to deal with congestion once it has been detected in a switch, router or network. These mechanisms basically consist of discarding elements. The question is: which packets to discard first?
2. *Congestion avoidance* is a set of mechanisms to deal with congestion before it happens. There are two types of congestion avoidance techniques:
 - Admission control operates only at the provider network edge nodes, ensuring that the incoming traffic does not exceed the transmission resources of the network.
 - Resource management is used to allocate and free resources in the packet switched network.

Congestion avoidance, and especially traffic admission, checks the properties of the subscriber traffic entering the provider network. These properties may include the average bit rate allowed in order to enter the network, but other parameters are used as well. For example, a network provider may choose to limit the amount of uploaded or downloaded data. Bandwidth profiles are used to specify the subscriber traffic, and the packets that meet the bandwidth profile are called conforming packets.

There are different types of filters that can help to classify non-conformant packets, and each of them have different effects on the traffic:

- *Policers* are filters that discard all non-conformant packets. Policers are well-suited to those error-tolerant applications that have strict timing constraints, for example VoIP or some interactive video applications (see Figure 1.41b).
- *Shapers* work much the same way as policers, but they do not discard packets. Non-conformant traffic is buffered and delayed until it can be sent without violating the SLA agreement or compromising network resources. Shapers conserve all the information that was sent, but they modify timing, so they may cause problems for real-time and interactive communications (see Figure 1.41c).

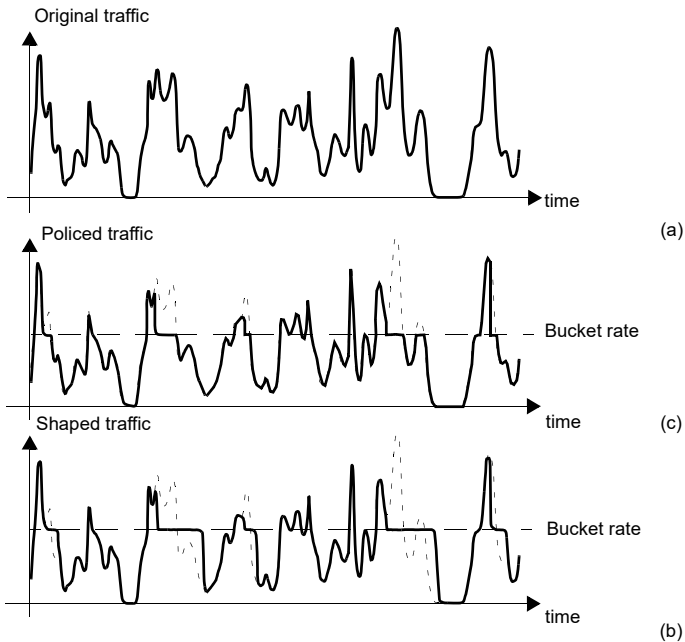


Figure 1.41 Shaping and policing of user traffic. (a) When traffic is shaped, no packets are dropped, but some of them may be delayed. (b) When traffic is policed, it is never delayed, but some packets may be dropped.

- *Markers* can be used to deal with non-conformant packets. Instead of dropping or delaying non-conformant packets, they are delivered with low priority or “best effort”.

There is a contract between the subscriber and the service provider that specifies the QoS, the bandwidth profile, and how to deal with the traffic that falls outside the bandwidth profile. This contract is known as the *Service-Level Agreement (SLA)*.

1.5.2 QoS In Ethernet Networks

Current Metro Ethernet networks are QoS-capable Ethernet network that offers services beyond the classical best-effort LAN Ethernet services. These services can be, for instance, *Time-Division Multiplexing* (TDM) circuit emulation, *Voice over IP* (VoIP) or *Video on Demand* (VoD).

Native Ethernet, however, as a best-effort technology, does not provide customized QoS. To maintain QoS, it is necessary to carry out a number of operations, such as traffic marking, traffic conditioning and congestion avoidance.

1.5.2.1 Bandwidth Profiles

Once Ethernet access has been set up at 10/100/1000/10000 Mb/s, the carrier performs admission control over the customer traffic at the UNI. Admission control for Ethernet services uses bandwidth profiles based on four parameters defined by the MEF:

- *Committed Information Rate* (CIR) — average rate up to which service frames are delivered as per the service performance objectives.
- *Committed Burst Size* (CBS) — maximum number of bytes up to which service frames may be sent as per the service performance objectives without considering the CIR.

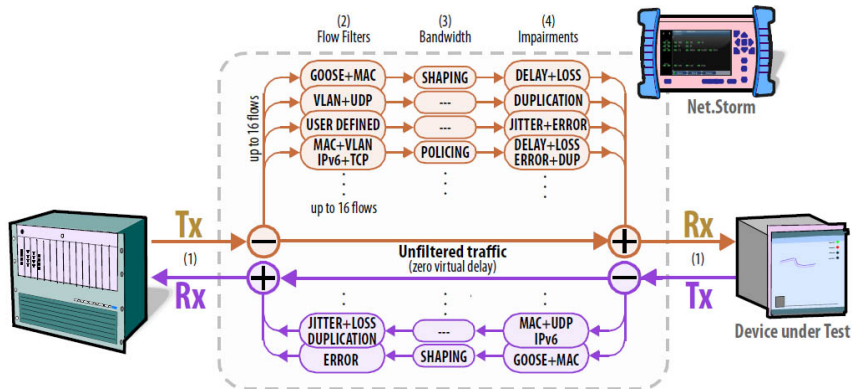


Figure 1.42 Net.Storm simulates links and networks in terms of bandwidth and quality of service. Traffic is separated by user-defined filters into independent flows that receive specific treatment to replicate real-world traffic conditions through impairments and bandwidth limitations for Ethernet impairment emulation and thus are important to check all the QoS mechanisms deployed in these networks.

- *Excess Information Rate (EIR)* — average rate, greater than or equal to the CIR, up to which service frames do not have any performance objectives.
- *Excess Burst Size (EBS)* — the number of bytes up to which service frames are sent (without performance objectives), even if they are out of the EIR threshold.

The MEF specifies a the *Two-rate Three-Color Marker (trTCM)* as the admission control filter for Metro Ethernet (see Figure 1.43). The trTCM is obtained by chaining two simple token bucket policers. Tokens fill the main bucket until they reach the capacity given by the CBS parameter, at a rate given by the CIR parameter. The secondary bucket is filled with tokens with the EIR rate until they reach the capacity given by the EBS parameter.

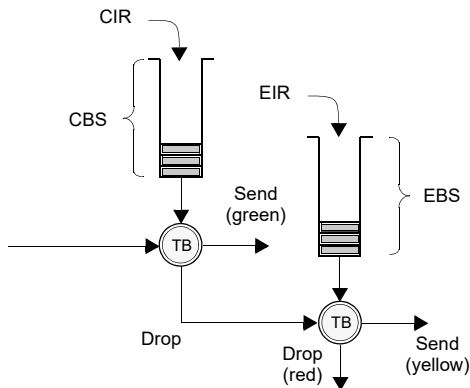


Figure 1.43 Two-rate three-color marker policer

The traffic that passes through the first bucket (*green traffic*) is delivered with the QoS agreed with the service provider, but any traffic that passes through the secondary bucket (*yellow traffic*) is re-classified and delivered as best-effort traffic, or it is given a low priority. Non-conformant traffic (*red traffic*) is dropped.

The 'best effort' classical service can be obtained by simply setting the CIR parameter to zero. The bandwidth profile can be applied per EVC, per UNI, or per the Class-of-Service (CoS) identifier. It is therefore possible to define more than one bandwidth profile simultaneously in the same UNI.

1.5.2.2 Class of Service Labels

IEEE 802.3 Ethernet frames do not have CoS fields, which is why they need to support additional structures.

The IEEE 802.1Q/p tag defines a three-bit CoS field, and it is commonly used to classify traffic. The three-bit CoS field present in IEEE 802.1Q/p frames allows eight levels of priority to be set for each frame. These values range from zero for the lowest priority through to seven for the highest priority (see Figure 1.44).

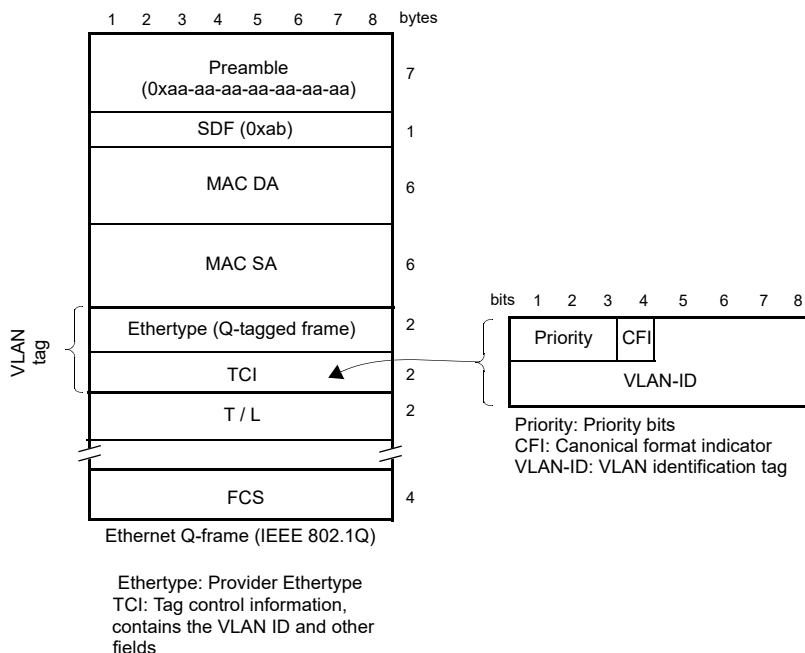


Figure 1.44 The IEEE.802.1Q VLAN frame format enables traffic classification through the three user priority bits.

It is also possible to map the eight possible values of the priority field to *Differentiated Services* (DSs), *Per Hop Behaviors* (PHBs) such as *Expedited Forwarding* (EF) or *Assured Forwarding* (AF) to obtain more sophisticated QoS management.

Sometimes traffic classes are defined on a per VLAN-ID basis rather than by means of CoS marks. To offer a single CoS per physical interface is a different approach.

1.5.2.3 Resource Management

Those technologies that are based on VCs, for example ATM, can potentially provide the same level of service as any other circuit-switched network, while maintaining high flexibility thanks to the ability to perform end-to-end connections (see Figure 1.45). Legacy Ethernet networks are connectionless. The solution is either to

redefine Ethernet or rely on other technologies for resource management. The alternatives currently available are the following:

- *Resource Reservation Protocol (RSVP)*: The RSVP is the most important of all the resource management protocols proposed for IP. It is an important component of the *Integrated Services (IS)* architecture suggested for IP networks. This architecture actually turns IP into a connection-oriented technology. To be efficient, the RSVP needs to be supported by all the network elements, and not only by the end user equipment. Both RSVP and IS call for a new generation of IP routers.

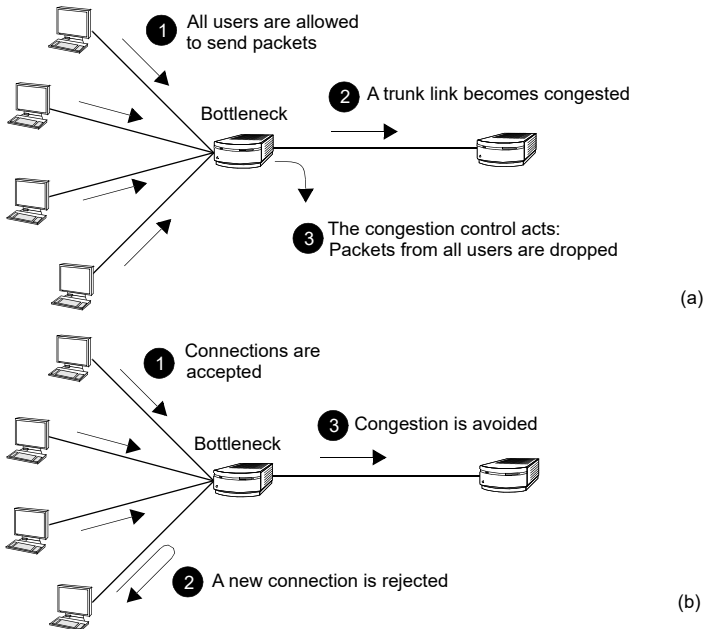


Figure 1.45 How resource management acts: (a) Without resource management, all users experience degradation on their applications whenever there is congestion in the network. (b) If congestion management is used, only some subscribers are not allowed to send data, but the others are not affected.

- *Multiprotocol Label Switching (MPLS)*: MPLS is a switching technology based on labels carried between the layer-2 and layer-3 headers that speed up IP datagram switching. MPLS can be used for QoS provisioning in Ethernet networks. One of the reasons for this is that MPLS supports a special type of connections called *Label-Switched Paths (LSP)*. The LSP setup and tear-down relies on a resource management protocol, usually the *Label Distribution Pro-*

ocol (LDP), but RSVP with the appropriate extension for MPLS can be used as well.

- *Provider Backbone Bridging with Traffic Engineering* (PBB-TE): PBB-TE is a group of improvements that turn Ethernet into a connection-oriented technology by re-interpreting some fields of the MAC frame. With PBB-TE, MAC addresses keep their global meaning. This has good implications for OAM, when compared to technologies based on labels with a local meaning, like ATM or MPLS. Given a source and destination MAC addresses, the route of a PBB-TE virtual circuit is identified by means of VLAN tags. VLAN tags can be reused, and this increases scalability. The *Spanning Tree Protocol* (STP) and IEEE 802.1ad bridging are not used and can be disabled. In PBB-TE, switching tables are not auto-configured by bridging, but set by a control plane separated from the forwarding plane.

1.5.2.4 Hands-on: Checking Ethernet Admission Control

Admission control is a congestion avoidance mechanism that helps operators to control the amount of traffic allowed to enter in their networks. It is the basis of QoS architectures such as Differentiated Services (DS). Most service providers need to deploy admission control mechanisms, if they aim to deliver Ethernet services to their customers in MAN environments. Today, it is possible to configure medium-cost switches and routers to provide admission control in LANs as well. It is important to remember that admission control is applied to the incoming interfaces of network elements, usually in the boundaries of the network, but it is not applied to any of the outgoing interfaces.

LAN operators may be interested in traffic admission, if they are running applications with specific QoS requirements, or when they have users that need differentiated service levels. If QoS-demanding services are to be connected to dedicated, well known physical ports, traffic admission control can be configured on a per port basis in switches or routers. Traffic admission has to be implemented for both QoS-demanding and best-effort services. A good example of this situation is a LAN transporting IP telephony traffic where data is generated in VoIP telephones connected to dedicated outlets in the network. In this case, it is possible to configure custom traffic admission filters for VoIP and data ports. However, a traffic class is not always generated in well-known network connections. When this occurs, applications can still be identified at the IP layer by using differentiated services code points. Most routers (and some switches) have QoS features that enable them to define traffic classes based on DS code points, and treat each traffic class differently. This includes custom admission control filters that depend on the DS code point value.

MAN operators have VLAN tags at their disposal for traffic marking and admission control. They use connection control to isolate customers or applications, and to pre-

vent congestion by limiting the rate of the traffic entering the network. There are three user priority bits within the VLAN tag that make it possible to define CoS marks, but admission control can also be implemented using the VID. A service provider may book one or several VIDs per customer and define specific admission control rules for each VID. Further refinement is possible, if priority bits are used for every VLAN. Of course, a port-based admission control is still available, but VLANs make it more quick, flexible and easy to define and provision services.

Sometimes, users are interested in checking whether the service they have purchased can reach the performance they are expecting. For example, it a customer may wish to test the maximum transmission rate allowed for different services (VPNs, VoIP, Internet access, etc). Service providers may also be interested in running similar tests during installation and troubleshooting. In this section we will see how to check the bandwidth of a connection that is using traffic admission filters. The basic tools to do this are provided by the IETF RFC 2544 that defines test configurations and procedures to check different performance figures for Ethernet devices, links and even entire networks. There are two performance parameters that are of interest for this purpose:

- *Throughput* is the maximum rate at which the Device Under Test (DUT) drops no frames. To test throughput, RFC 2544 compliant testers send a certain number of frames at pre-configured rates through the device under test, and then check the frames that are transmitted through the DUT without errors. The number of frames offered and forwarded is compared, and depending on the result, a new iteration starts and the test is performed again with a different frame rate. After some iterations, the test rate converges to the throughput of the device under test.
- *Back-to-back* tests measure the length of the longest maximum-rate frame burst a device can accept without dropping any frames. To perform this measurement, the RFC 2544 compliant tester sends a burst of frames with minimum interframe gaps to the DUT and counts the number of frames forwarded by this device. If the number of transmitted frames is equal to the number of frames forwarded, the length of the burst is increased and the test is performed again. If the number of forwarded frames is less than the number of frames transmitted, the length of the burst is reduced and the test is performed again. Finally, the burst length converges to the longest possible back-to-back burst.

The RFC 2544 throughput test is used to check the steady-state bandwidth of an Ethernet connection. If the average transmission rate is higher the CIR (or EIR, depending on the admission control filter), frames will be dropped sooner or later. If the transmission rate is constant, and smaller than the CIR or EIR, no frames should be dropped. This makes it possible to measure both CIR and EIR. If the admission control filter implements the trTCM algorithm, it is not possible to measure the CIR with a throughput test, because excess traffic is sent to a cascaded policer rather than

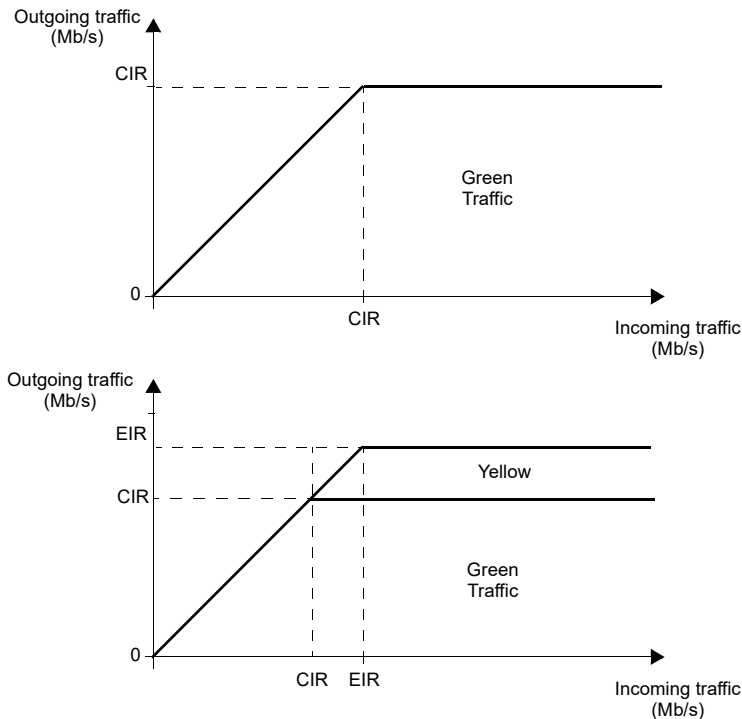


Figure 1.46 The amount of traffic that crosses an admission control filter. Graphics represent steady states, traffic is usually allowed to be greater than the CIR and EIR for short periods of time. (a) The CIR is equal to the EIR, the network guarantees traffic delivery if incoming traffic is smaller than the CIR. (b) The EIR is greater than the CIR. Traffic delivery is guaranteed if the rate is smaller than the CIR. Excess traffic (traffic above the CIR and below the EIR) is delivered as well, but it is marked as

being dropped. To measure the CIR, in this case, a tester that can detect traffic marks is needed. The throughput test also has limited applicability when the access control filter contains shapers, because theoretically these filters never drop frames.

CBS and EBS are admission control parameters related with the dynamic behavior of the filter, and they can only be tested when not in the steady state. To measure CBS (or EBS), the RFC 2544 back-to-back test is used. This test fills the buckets with a fast packet stream, and when the first packet is discarded, the test stops. In a connection with an admission control filter made up of a simple token-bucket policer, the size of the CBS can be measured by using the following formula:

$$\text{CBS} = \text{ICBS} - \text{CIR} \times \text{TCBS}$$

I_{CBS} is the amount of data that has entered the network before the first frame is lost. In other words, it is the result of the back-to-back frame test. T_{CBS} is the time interval between the start of the test and the first frame drop event. It can be derived from I_{CBS} , if frames are injected with constant and deterministic rate in the back-to-back test. CBS is different from I_{CBS} , because some data leaves the policer while the traffic generator attempts to fill it. I_{CBS} accounts for data ingressing in the policer, and $CIRxT_{CBS}$ for data leaving the policer. CBS is the difference between these two.

If the admission control filter implements the trTCM algorithm, it is difficult to determine both CBS and EBS, because non-compliant traffic is sometimes remarked, and remarking events are not valid triggers for the RFC 2544 back-to-back test. However, the CBS formula is still useful as a merit figure for the trTCM and more complex policers. In this case, the result represents the size of a token bucket policer equivalent to the connection admission filter under test.

Testing admission control calls for a traffic generator/analyzer that is able to generate customizable synthetic traffic, and a loopback device of some sort to send the traffic back once it has passed through the DUT. Traffic should not be altered during the return path (from the loopback device to the traffic generator/analyzer), or the result may be affected by other effects. Admission control is applied to incoming interfaces only (not to outgoing interfaces). It is also important to obtain accurate results, so that the DUT can be put out of service to avoid any interference between test traffic and ordinary network traffic.

Test traffic, here, is just standard unicast Ethernet traffic. The source MAC address must be used as the address of the traffic generator/analyzer, and the destination MAC address must be the same as the address of the loopback device. The loopback device must support MAC address swapping, and depending on the DUT, IP address swapping as well. This way, traffic can find its way back to the generator/analyzer without disturbing network operation.

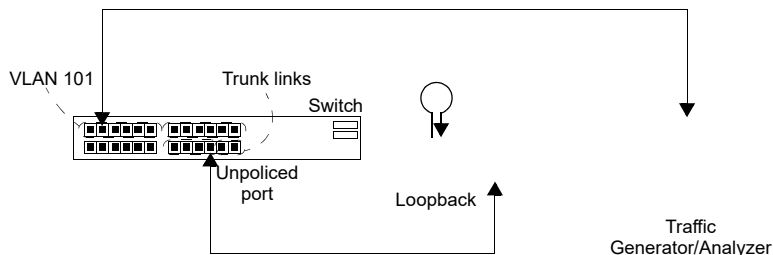


Figure 1.47 In this test, traffic is delivered through an IEEE 802.3 interface to a device connected to an IEEE 802.1Q interface.

In a typical test setup for LAN environments (see Figure 1.47), the traffic generator/analyzer is connected to a user interface (IEEE 802.3) and the loopback device to a trunk interface (IEEE 802.1Q). IP packets encapsulated in Ethernet frames can be delivered through the DUT, and it is even possible to add DS code points to the test traffic, to check how DS classes are processed by the DUT. In MAN setups, VLANs are used to isolate users or services. The traffic generator/analyzer is therefore connected to a trunk IEEE 802.3Q port in the DUT. The loopback is connected to the uplink interface in the DUT. This interface can use a Q-in-Q encapsulation, for example. If the DS code points, the VID or the user priority bits are service-delimiting, the test can be repeated for several field values to check how results vary for different services. Traffic generators with multistream traffic generation and analysis features can check different services at the same time. This gives further insight on the isolation of services based on DS code points, VLANs or user priority bits.

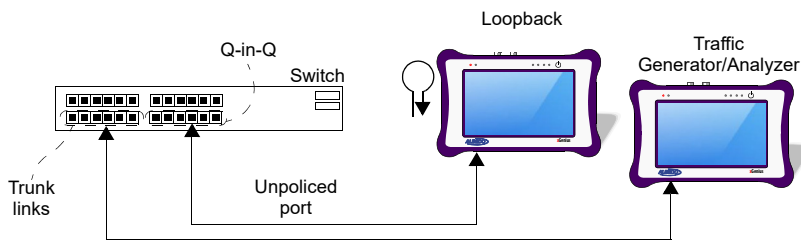


Figure 1.48 In this test, traffic is delivered through an IEEE 802.1Q interface to an IEEE 802.1ad (Q-in-Q). This is a very typical situation in a service provider network.

1.5.3 QoS in IP Networks

Ethernet often relies on other complementary technologies such as IP or MPLS for QoS provision. IP in particular has become a key technology for multiplay networks, and it is quite realistic to think that it will be in charge of QoS provisioning as well. There are two QoS architectures available for IP:

1. The *Integrated Services (IS)* architecture provides QoS to traffic flows. It relies on allocation of resources in network elements with the help of a signaling protocol, the ReSerVation Protocol (RSVP).
2. The *Differentiated Services (DS)* architecture provides QoS to traffic classes. Packets are classified when they enter the network, and they are marked with DS code points. Within the network, they receive custom QoS treatment according to their code points only.

The IS architecture is more complex than the DS architecture, but it potentially provides better performance. One of the most important features of the IS approach is

the ability to provide absolute delay limits to flows. On the other hand, the DS approach does not rely on a signaling protocol to reserve resources, and does not need to store flow status information in every router of the network. Complex operations involving classifying, marking, policing and shaping are carried out by the edge nodes, while intermediate nodes are only involved in simple forwarding operations. The IS architecture is better suited to small or medium-size networks, and the more scalable DS approach to large networks.

1.5.3.1 Class of Service Labels

IP CoS labels are defined either by the ToS labels or the DS code points (see Figure 1.49). The ToS byte forms a part of the IP specification since the beginning, but it has never been extensively used. The original purpose of the ToS bit was to enhance the performance of selected datagrams, to make it better than best-effort transmission QoS. To do this, a four-bit field within the ToS byte is defined, and it includes the requirements that this packet needs to meet (see Table 1.4).

Table 1.4 Meaning of ToS bits.

Binary value	Meaning
1xxx	Minimize delay
x1xx	Maximize throughput
xx1x	Maximize reliability
xxx1	Minimize monetary cost
0000	Normal service

In addition to the four-bit field mentioned before, there is a three-bit precedence field that makes it possible to implement simple priority rules for IP datagrams (see Table 1.5).

Table 1.5 Precedence bits and their meaning

Binary value	Meaning
000	Routine
001	Priority
010	Intermediate
011	Flash
100	Flash override
101	Critic / ECP
110	Internetwork control
111	Network control

The ToS values encode some QoS requirements for the IP datagrams, but the decision on how to deal with these values is left to the network operator. For example, some operators might meet the “Minimize delay” requirement by prioritizing pack-

ets with this mark, but other operators might rather select a special route reserved for high-priority traffic.

This is a major difference between ToS values and DS code points. While the ToS values specify the QoS requirements for the IP traffic, the DS code points request specific services from the network. Defining these services, created by means of different PHBs, is the core of the DS architecture specification.

Although there are some recommendations, most of the PHB encoding by means of DS code points are configurable, and they can be freely chosen by the network administrator. The only constraint for this is the backwards compatibility with the old ToS encodings.

There are some PHBs defined to be used by DS routers. The most basic of them is the *default PHB* that provides basic best-effort service and must be supported by all the routers. The recommended DS code point for the default PHB is 000000. Additionally, the *Assured Forwarding (AF)* PHB has a controlled packet loss, and the *Expedited Forwarding (EF)* PHB has a controlled delay. Other experimental PHBs are the *Less than Best Effort (LBE)* PHB for transporting low-priority background traffic, or the *Alternative Best Effort (ABE)* PHB that provides a cost-effective way to transport interactive applications by making the end-to-end delay shorter, but with higher packet loss.

1.5.4 End-to-End Performance Metrics

The first step in offering QoS is to find a set of parameters to quantify and compare the performance of the network. QoS is provided by the network infrastructure, but experienced by the users. This is the reason why QoS is specified by means of end-to-end parameters. There are at least four critical QoS metrics to define: delay, delay variation, loss and bandwidth.

1.5.4.1 One-way Delay

The end-to-end *one-way delay* experienced by a packet when it crosses a path in a network is the time it takes to deliver the packet from source to destination. This delay is the sum of delays on each link and node crossed by the packet (see Figure 1.50).

The *Round Trip Delay (RTD)*, or latency, is a parameter related to one-way delay. It is the delay of a packet on its way from the source to the destination and back. RTD is easier to evaluate than other delay parameters, because it can be measured from one end with a single device. Packet timestamping is not required, but a marking mechanism of some kind is needed for packet recognition. The best-known RTD tool is Ping. This tool sends *Internet Control Message Protocol (ICMP)* echo request

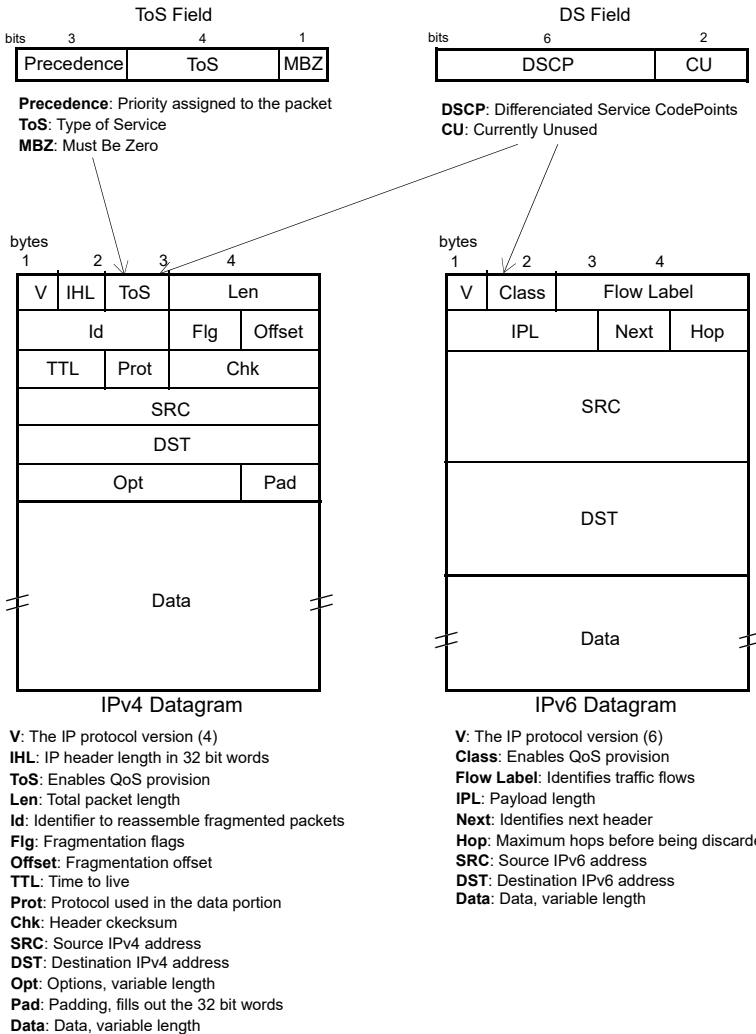


Figure 1.49 IPv4 and IPv6 datagrams and the format of the ToS and DS fields, both related to QoS provisioning.

messages to a remote host, and receive ICMP echo replay messages from the same host.

There are three types of one-way delay:

- *Processing delay* is the time needed by the switch to process a packet.
- *Serialization delay* is the delay between the transmission time of the first and the last bit of a packet. It depends on the size of the packet.

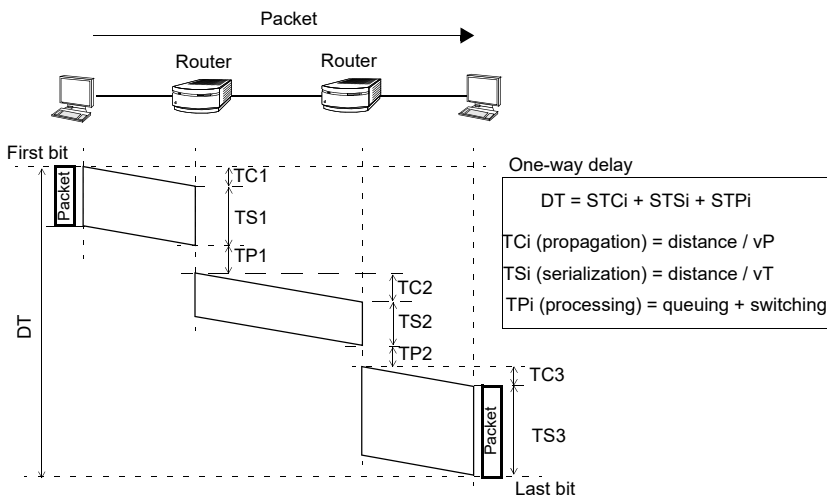


Figure 1.50 One-way delay is the sum of delays on each link and node crossed by a frame.

- *Propagation delay* is the delay between the time the last bit is transmitted at the transmitting node and received at the receiving node. It is constant, and it depends on the physical properties of the transmission channel.

1.5.4.2 One-way Delay Variation

The *one-way delay variation* of two consecutively transmitted packets is the one-way delay experienced by the last transmitted packet, minus the one-way delay of the first packet (see Figure 1.51). The one-way delay variation is sometimes referred to as *packet jitter*.

In packet-switched networks, the main sources of delay variation are: variable queuing times in the intermediate network elements, variable serialization and processing time of packets with variable length, and variable route delay when the network implements load-balancing techniques to improve utilization.

1.5.4.3 Packet Loss

A packet is said to be lost if it does not arrive to its destination. It can be considered that packets that contain errors or arrive too late are also lost.

Packet loss may occur when transmission errors are registered, but the main reason behind these events is network congestion. Intermediate nodes react to high traffic load conditions by dropping packets and thus generating packet loss. Congestion

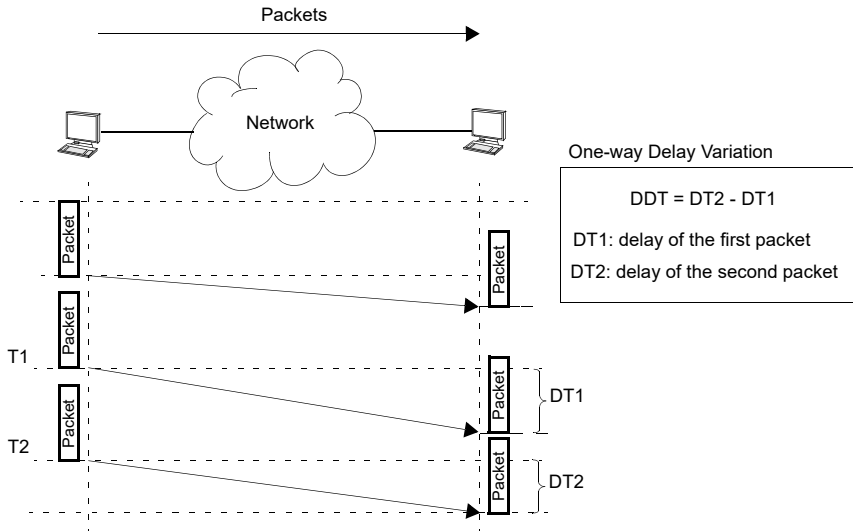


Figure 1.51 One-way delay variation: measurement and impact on data periodicity

tends to group loss events, and this harms voice and video decoders optimized to work with uniformly distributed loss events. Loss distance and loss period are metrics that give information on the distribution of loss events.

- *Loss distance* is the difference in the sequence numbers of two consecutively lost packets, separated or not by received packets.
- *Loss period* is the number of packets in a group where all the packets have been lost.

1.5.4.4 Bandwidth

Bandwidth is a measure of the ability of a link or a network to transfer information during a given period of time. Capacity and available bandwidth can be defined for links, or for entire transmission paths formed by several links. However, for QoS, the most important bandwidth metric is the available end-to-end capacity, because only end-to-end parameters are relevant when evaluating a service.

1.5.4.5 Hands-on: Checking End-to-End Performance

Once devices are interconnected and remote applications accessible, it is time to test performance and resource availability. QoS tests check *frame loss*, *latency* and *jitter*, and in some cases some other parameters as well. Frame loss, latency and jitter are all important, but there are applications that are not sensitive to some of them (see

Table 1.6). For example, VoIP is sensitive to jitter and latency. On the other hand, streamed video and business data are sensitive to frame loss ratio.

Table 1.6 ITU-T Y.1541 Network Performance Objectives.

QoS Classes	Applications	Packet Loss	Delay	Jitter
0	Real-time, jitter-sensitive, highly interactive traffic (VoIP, videoconference)	1×10^{-3}	100 ms	50 ms
1	Real-time, jitter-sensitive, interactive traffic (VoIP, videoconference)	1×10^{-3}	400 ms	50 ms
2	Transaction data, highly interactive traffic (signalling)	1×10^{-3}	100 ms	Unspecified
3	Transaction data, interactive traffic (signalling)	1×10^{-3}	400 ms	Unspecified
4	Low-loss data traffic (short transactions, bulk data, video streaming)	1×10^{-3}	Unspecified	Unspecified
5	Best-effort traffic (traditional IP data)	Unspecified	Unspecified	Unspecified
6	Real-time, jitter-sensitive, highly interactive, low error-tolerant traffic	1×10^{-5}	100 ms	50 ms
7	Real-time, jitter sensitive, interactive, low error-tolerant traffic	1×10^{-5}	400 ms	50 ms

To guarantee the QoS for each application, a number of parameters need to be measured, end-to-end. It is common to measure QoS at the IP layer, because IP is the technology that applications use to be available at end points where QoS tests are performed. However, QoS tests can also be carried out at the Ethernet layer where Ethernet is available.

QoS tests can be made out-of-service by injecting synthetic traffic to the network during installation, bringing-into-service and troubleshooting, but in-service tests are also common when monitoring applications. In fact, continuous or on-demand QoS parameter evaluation is part of the current Operation, Administration and Maintenance (OAM) framework for Ethernet defined in IEEE 802.1ag and ITU-T Y.1731. For both in-service and out-of-service applications, QoS tests need to inject traffic into the network. For in-service applications, care must be taken to avoid damaging user applications with the test traffic.

Even though IETF RFC 2544 tests are defined for testing interconnection devices, they can be used to test end-to-end paths as well. These tests may generate large amounts of traffic and cause congestion. They are therefore best suited for out-of-service tasks. There are RFC 2544 tests for checking latency and frame loss, but frame delay variation must be checked in a different way. RFC 2544 tests are performed as follows:

- The RFC 2544 *latency* test determines the delay inherent in the device or network under test. The initial data rate is based on the results of a previous

throughput test. Time-stamped packets are transmitted, and the time it takes for them to travel through the device or network under test is recorded.

- The RFC 2544 *frame loss* test determines the frame loss ratio across the entire range of input data rates and frame sizes. The test is performed by sending several bit rates, starting with the bit rate that corresponds to 100% of the maximum rate, on the input media. The bit rate is reduced at each iteration.

The RFC 2544 has limited applications in QoS testing due to its inability to provide delay variation results, and because it can only be used for out-of-service measurements. Other, more generic QoS tests are sometimes also performed. These tests include a customizable traffic generator that delivers packets with time stamps and sequence numbers, and a traffic analyzer that computes delay, delay variation and frame loss events.

The traffic generator and the traffic analyzer can be packed in different boxes and connected to different points in the network (see Figure 1.52), if delay variation and frame loss are the only parameters to test. Things are more difficult if delay is measured, because in this case the transmitter and the analyzer must be synchronized. The most obvious solution is to pack the transmitter and the receiver into the same box and use a loopback device at the remote end to send the traffic back to the origin. If this solution is adopted, the generator/analyzer computes the Round Trip Delay (RTD) rather than one-way latency. All round-trip parameters have the same problem: it is difficult to determine the contribution of the forward and backward path to the end result. For RTD, it turns out to be impossible to separate these two without synchronizing all the measurement devices: generator, analyzer and loopback.

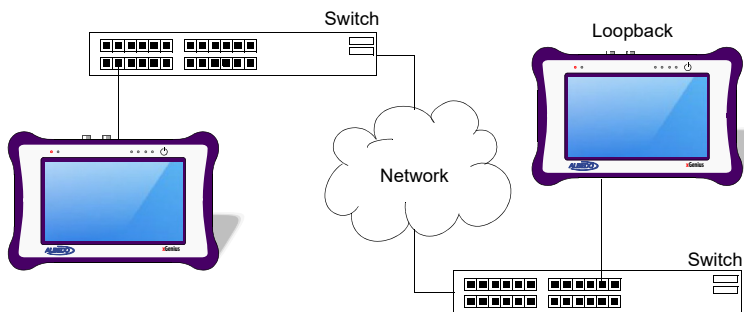


Figure 1.52 Simple QoS test setup. A traffic generator/analyzer and a loopback device are connected to remote devices. The traffic crosses the network in two directions. The traffic generator/analyzer collects statistics on the test traffic.

Compared to the RFC 2544 test, one of the advantages of a test setup where a customizable traffic generator is used is that the latter gives more freedom to define the

bandwidth profile for the test traffic. For example, bursty traffic, ramps, multistream and random bandwidth profiles are now possible. So, this test can obtain results under realistic operation conditions.

When setting up the QoS test, it is necessary to decide how long the test is going to run, what is going to be the traffic profile and how big will the packets be. Some suggestions:

- Installation and bringing-into-service tests have a definite *duration*. Test duration is variable, and it may be different in different situations. ITU-T Recommendation Y.1541 suggests a minimum evaluation interval of 1 minute for delay, delay variation and packet loss evaluation. Monitoring is more focused on tracking events than in obtaining performance figures at the end of the test. This is the reason why monitoring tasks usually have an unspecified duration. Monitoring tests are often run during very long time periods.
- To make decisions on the *bandwidth profile* of the test traffic, it is necessary to previously get information on congestion avoidance for the end-to-end path to be tested. Especially non-conformant traffic may cause high packet loss ratio and delay. In normal situations, constant bit rate is well suited for testing. Bursty traffic or other more complex traffic profiles are only needed for special purposes. It is useful to run the test with different bit rates to check how the QoS figures evolve as the traffic load increases. It is also useful to generate multistream traffic. Different streams can be placed in different traffic classes. Some streams can be used as background traffic replacing real user traffic in out-of-service measurements. Multistream traffic also makes it possible to measure QoS statistics for different traffic classes simultaneously. By increasing traffic load for background streams and checking the evolution of QoS statistics in foreground streams, isolation between traffic classes can be checked. This is another important test that can only be performed with multistream traffic.
- The third decision concerns the packet size to use for the test traffic. Latency, delay variation and loss tend to grow when packet size increases. It is often a clever decision to start testing with big packets. ITU-T Recommendation Y.1541 suggests a packet size of 1500 bytes for QoS testing. In some cases, it may be interesting to check how QoS statistics evolve as packet size changes. If the traffic generator supports multistream traffic, QoS statistics can be collected for different packet sizes of background traffic both in and outside the foreground traffic class. This way, you can check how traffic differentiation protects the QoS of the foreground stream.

Now that the test setup and execution issues are solved, it is important to decide whether the test results can be accepted or not. The IETF defines performance parameters, but it does not provide any limits for them. The DS traffic classes are de-

fined by the IETF to transport services with specific QoS requirements with some performance guarantees. However, operators have to adapt these classes to their own performance objectives. The only international standards organization that provides explicit performance requirements for IP-based applications is ITU, with Recommendation Y.1541 (see Table 1.6). This ITU-T standard defines eight traffic classes numbered from 0 to 7. Classes 6 and 7 are provisional. Classes 1 and 2 are defined for interactive traffic, such as VoIP or videoconferencing. Classes 2 and 3 are designed to transport short transactions sensitive to delay, mainly signalling. Classes 4 and 5 are for data traffic and non-interactive multimedia, such as video streaming. The provisional traffic classes are for interactive traffic with low tolerance to errors and packet loss. High-quality IPTV is well suited to these traffic classes.

The performance limits given in ITU-T Y.1541 have been chosen to enable reliable multiplay service provision in converged IP networks. ITU has collected information on how errors and delay degrade services such as VoIP and IP video. Regarding VoIP, ITU has rated the subjective quality of a VoIP service under different delay and packet loss conditions. Delay variation does not need to be taken into account directly, because VoIP receivers transform delay variation into delay with a de-jittering filter.

Table 1.7
VoIP Service Degradation under Different Transmission Conditions

QoS Class	Network delay	Terminal delay	Total delay	R (no loss)	R (loss 10^{-3})
0	100 ms	50 ms	150 ms	89.5	87.6
0	100 ms	80 ms	180 ms	87.8	87.5
1	150 ms	80 ms	230 ms	81.9	81.5
1	233 ms	80 ms	313 ms	71.1	70.7

The VoIP service benchmarking parameter chosen by the ITU-T is the R-Factor, defined in ITU-T G.107 (the so-called E-model). The R-Factor rates the conversational quality of voice communications on a scale from 0 to 100. The R-Factor should be better than 80, and it should never drop below 70. The ITU-T results (see Table 1.7) show that packet loss is not an issue for VoIP, as long as the packet loss ratio is better than 10^{-3} . This is partly due to the packet concealing algorithms of common VoIP encoders. These algorithms provide packets for the decoder when the actual packets are lost in the network. They cause effects similar to the Forward Error Correction (FEC) mechanisms, but they have been especially designed for VoIP applications. Delay appears to be the most important issue in VoIP. Small packet size, reduced de-jittering filters and high-performance transmission is required to achieve the minimum required QoS. Results show that the value for one-way delay that meets the requirement of 'better than 80' is around 150 ms. Delays of about 300 ms or even more are still acceptable in some circumstances.

Table 1.8
Digital Television Loss/Error Ratio Requirements

Application	One performance hit per 10 days	One performance hit per day	10 Performance hits per day
Contribution (270 Mb/s)	4×10^{-11}	4×10^{-10}	4×10^{-9}
Primary distribution (40 Mb/s)	3×10^{-10}	3×10^{-9}	3×10^{-8}
Access distribution (3 Mb/s)	4×10^{-9}	4×10^{-8}	4×10^{-7}

In video services such as IPTV, quality can be rated in error/loss events per time unit. The amount of degradation that parties are likely to accept depends on the particular video service profile. ITU-T Y.1541 defines three of these profiles:

- *Contribution* services make it possible for a network or its affiliates to exchange content for further use. Sometimes video contents are immediately re-broadcast and other times they are stored to be edited or broadcast later. Contribution video is generally lightly compressed, and it requires a lot of bandwidth for transmission.
- *Primary distribution* services include delivery to head-ends for transmission through cable, satellite or TV. This service generally requires less bandwidth than contribution services.
- *Access distribution* services include delivery to the end user through cable, satellite or copper network. It requires less bandwidth than the primary distribution service.

The packet loss ratio can be calculated for these three service profiles used in transmission channels with different performances. For all of these services, the packet loss ratio required is around 10^{-10} or 10^{-9} (see Table 1.8). There is no Y.1541 traffic class that meets this requirement. Even the provisional low-loss ratio traffic classes (6 and 7) are unable to provide the desired packet loss ratio. This shows the importance of FEC in video transport to correct errors at the destination, at the price of increased overhead during transmission (see Table 1.9).

Table 1.9
Approximate FEC overhead for different channels, necessary to achieve acceptable overhead in video transmission.

	High Performance	Medium Performance	Low performance
Loss Distance	100 packets	50 packets	50 packets
Loss Period	5 packets	5 packets	10 packets
FEC Overhead	5 %	10 %	20 %

1.5.4.6 Hands-on: Accurate Testing of the One-way Delay

The One-way Delay test option enables Cell Site Ethernet backhaul providers, mission-critical government agencies, and financial institutions to measure the delay of

Ethernet, IPv4 and IPv6 traffic that is received from a sender using a highly accurate CDMA receiver. The delay of information transmitted may not be the same as the delay of information received. This can be caused by different paths taken by the traffic across the network or by differences in the way the traffic is buffered or prioritized by devices.

For technicians and engineers needing to install Ethernet or IP circuits, the One-way Delay option saves hours of troubleshooting by detecting asymmetric traffic delays. Accuracies 10 times greater than most common *Service Level Agreements* (SLAs) can be attained, permitting Ethernet network providers to differentiate their offering and allowing network planners to understand the delay tolerances affecting their applications (see Table 1.10).

Table 1.10
Feature/Benefit Summary

Feature	Description	Advantage	Benefit
UTC timestamp	Use CDMA derived time	Both ends use same time of day timestamp in test	Accurate delay calculation based on timestamp
BITS / SETS clock input	Accurate clock from CDMA network	Global clock synchronization between test sets	Reliable timing source for test
CDMA receiver	CDMA receiver provides time & clock	Test any Ethernet network within cell phone range	Test in the widest geographic footprint
Zero configuration	No additional configuration needed	Plug in CDMA receiver and run the test	Minimal training and no special configurations to learn

The measurement of highly-accurate one way delay in an Ethernet/IP backhaul scenario improves application debugging. Even though a device may be at the very edge of the network, asymmetric delays can still occur. In a VoIP application, the greater the delay, the more the devices buffer the information so that speech can be smoothed out. Unfortunately, if delay is unequal, one side of the conversation may sound perfectly clear while the other caller may be constantly talking over the speaker. In a financial environment where the receipt of information is many times acknowledged, differences in one-way delay can create the appearance of some devices receiving the information before others when in fact the problem is a delay in receiving the acknowledgement. With highly-accurate one-way delay measure-

ment, network planners now have the information needed to optimize their networks to improve the quality of service and overall customer satisfaction (see Figure 1.53).

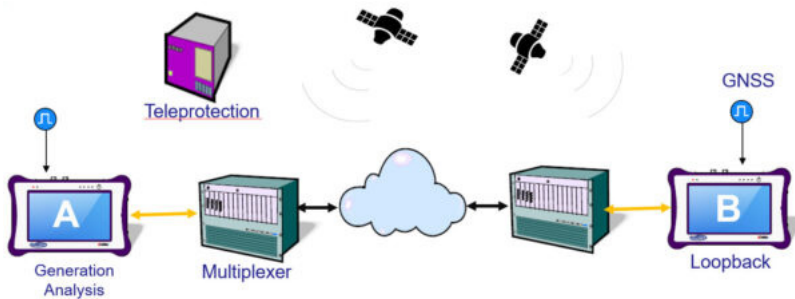


Figure 1.53 The One-way Delay option saves hours of troubleshooting by detecting asymmetric traffic delays

In a mission critical environment, information broadcast to many devices or sent over satellite links may take longer to be acknowledged by some than by others even if it was received at the same time. This delay can be due to varying weather conditions encountered by the different transmitters that are located in different parts of the world. With highly-accurate one way delay measurement, the functioning of mission-critical applications can be improved and overall mission objectives met.

1.5.4.7 Hands-on: Profiling Bandwidth Usage in Ethernet

J-Profiler is a test option for the xGenius Multi-Services Application Module (MSAM) and xGeniusDual Module Carrier (DMC) that provides automatic detection of up to 128 unique traffic streams for 10 Mb/s to 1 Gb/s links. As a passive, hardware-based monitor application, it discovers Ethernet / IP traffic organized by VLAN, MAC, and IP Addresses or Ports and displays the bandwidth utilization of each stream, allowing a view of top talkers.

For technicians who troubleshoot Ethernet and IP circuits, the J-Profiler test option provides valuable insight and simplifies complex issues by illustrating the full network picture and characterizing which customers, services, or applications are consuming bandwidth. By instantly detecting and exposing network utilization in a user-configurable manner, J-Profiler is a beneficial first step to troubleshooting complex networks quickly and efficiently. By providing time-saving high-level visibility, it functions seamlessly as a complementary investigative tool to the in-depth

ALBEDO Capture, Decode and J-Mentor protocol analysis test options (see Table 1.11).

Table 1.11
Feature/Benefit Summary

Feature	Description	Advantage	Benefit
Optional Filter Selections	Specify filters by traffic type, address, VLAN or port	Troubleshoot customer specific traffic flow / service	Eliminate excess information to narrow scope prior to stream discovery
Traffic Stream Classifications	Select stream organization by VLAN, MAC, IP Address, or Ports	Multiple available classifications encompassing ALL network traffic types	Ensures flexible analysis capabilities and ubiquitous usage
Auto-Stream Discovery	Detects and displays up to 128 live traffic streams	Analyze key traffic information and utilization in real time	View top talkers by customers, services, or applications
Customizable Display Table	Sort by any parameter, rearrange information, and hide or display columns	Adjustable viewing options to suit various personal preferences	Provides organized, simplified, and clear picture of complex network traffic streams

The xGenius and Zeus can be used to troubleshoot Ethernet and IP problems by connecting to a mirror or spare test port on a switch or router. In this scenario, the test set is configured in a passive monitor mode and is used to auto-discover live traffic streams up to 1 Gb/s that are forwarded to this mirror port by the switch. The discovered streams can then be analyzed on the test set, displaying key traffic information such as VLAN ID, MAC/IP addresses, and Port Numbers, as well as the bandwidth utilization of each stream.

There is only one set-up configuration required for the J-Profiler application: the selection of desired traffic profile or classification. Filters can be optionally set on the *Filters* tab to further narrow down the discovered streams. The profiled traffic streams are auto-detected and displayed in a customizable table, revealing key network information and bandwidth utilization in real time. Click “Restart” to refresh stream discovery (see Figure 1.54).

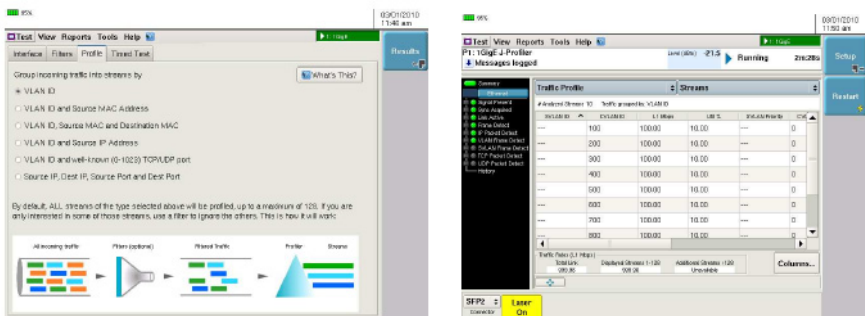


Figure 1.54 There is only one set-up configuration required for the **J-Profiler** application. The profiled traffic streams are auto-detected and displayed

1.6 OPERATION, ADMINISTRATION AND MAINTENANCE

The purpose of Operation, Administration and Maintenance (OAM) is to provide failure detection and management mechanisms and to deliver availability and performance figures to specific points in the network. OAM has traditionally been an important requirement of carrier networks but it was basically missing in legacy Ethernet.

OAM capabilities reside within special entities in network elements like switches and routers but sometimes they are implemented by dedicated devices which may carry different kinds of analysis in the network.

The approach to OAM provided by all modern technologies, including Ethernet and MPLS, but also SDH, OTN and ATM is hierarchical. This enables multiple levels of maintenance to be managed with the same OAM mechanism. Different parties involved in the communication, including service providers, carriers and end users, can have their own OAM domain. Switches or routers belonging to lower level maintenance domain, forward transparently frames from higher domains (see Figure 1.55).

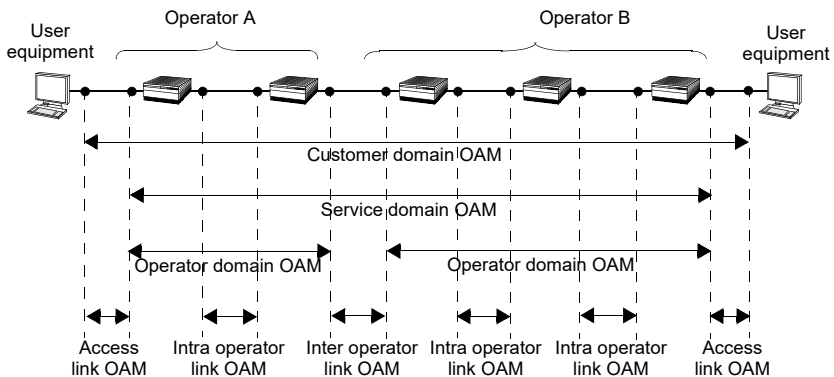


Figure 1.55 Switches forward transparently frames from higher domains.

OAM entities receive different names depending on the functionality they add to the network. This terminology is shared by all OAM standards. Maintenance domains are called Maintenance Entity Groups (MEGs) by the OAM standards. In the same way, other useful terms are MEG EndPoint (MEP) and MEG Intermediate Point (MIP).

1.6.1 Ethernet OAM

OAM is a key feature of Carrier Ethernet. The IEEE, ITU-T and MEF are actively developing a OAM framework for Ethernet:

- The IEEE 802.1 committee approved in December 2007 the IEEE 802.1ag standard for Connectivity Fault Management (CFM) of Ethernet networks. This standard defines the base OAM frame formats, protocol elements and functionalities. Other previously existing IEEE standards related with OAM are the IEEE 802.1ab and 802.3ah. The former defines the Link Layer Discovery Protocol (LLDP) that allows stations to advertise their capabilities with discovery and automatic configuration purposes. The latter is part of the Ethernet in the First Mile (EFM) standard. It provides link OAM capabilities to Ethernet access networks. The link OAM enables access network operators to monitor and troubleshoot the Ethernet link between the customer and network operator equipment. IEEE 802.1ah capabilities include discovery, link monitoring, remote failure indication and remote loopback.
- The ITU-T SG13 released in May 2006 Recommendation Y.1731, that agrees with the procedures and protocols defined by IEEE 802.1ag but extends its functionality. The Recommendation Y.1731 defines failure detection and management as well as performance monitoring procedures.
- The MEF released the standard MEF 17 in April 2007. This standard adapts the OAM specifications to the own MEF framework. The MEF 17 does not define specific OAM mechanisms. It rather defines OAM requirements to enable carrier class operation.

IEEE, ITU-T and MEF are working closely with Ethernet OAM. Terminology used by all three organizations is similar and protocols and procedures defined in the resulting standards are highly compatible. Maybe the most important OAM standard is the ITU-T Y.1731 because it is compatible with IEEE 802.1ag and at the same time it is a superset of it.

OAM frames are encapsulated in standard Ethernet, VLAN, PB (Q-in-Q) or PBB (MAC-in-MAC) frames. Depending on where an OAM frame is analyzed, the encapsulation may change. However, the OAM payload does not change when it is transmitted between MEPs of the same domain. Some of the fields of the OAM payload are common to all OAM procedures and services and some others depend on the particular information being transported by the frame (see Figure 1.56). Common fields are:

- The *MEG Level (MEL)* is a 3-bit field identifies the maintenance domain level associated to the frame. This field helps the destination MEP to recognize OAM frames attached to it:

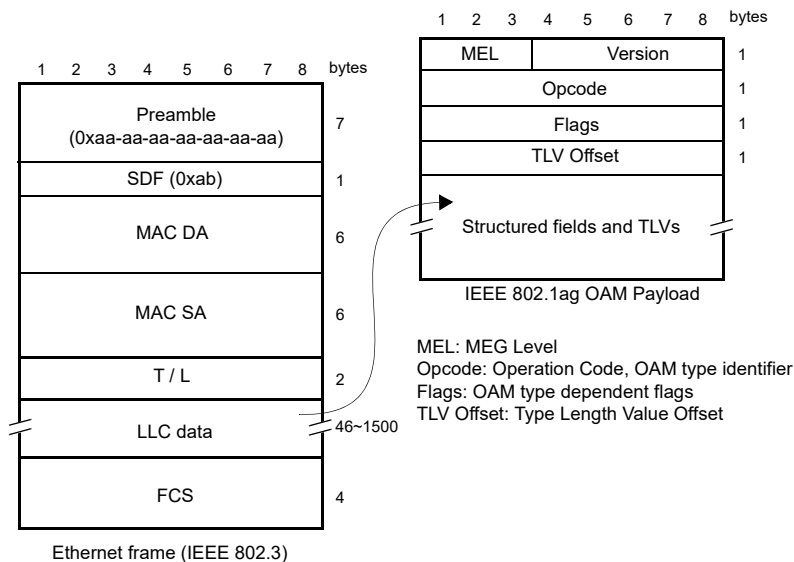


Figure 1.56 Structure of the IEEE 802.1ag PDU and its mapping in IEEE 802.3 Ethernet frames.

- The *Version* (5-bits) identifies the OAM protocol version carried in the current frame. Currently all bits in this field are set to zero.
- The *Opcode* (8-bits) identifies the OAM type carried in the frame. This value is used to decode the remaining content of the OAM payload. Opcodes from 1 to 4 are used by IEEE 802.1ag, Opcodes from 33 to 50 are used by Y.1731. All other Opcodes are currently reserved.
- The *Flags* (8-bits) field contains flags whose meaning is dependent of the OAM type carried by the frame.
- Parameters in Ethernet OAM frames are encoded in Type Length Values (TLVs). The *TLV Offset* (8-bits) field indicates the offset to the first TLV value in the OAM frame relative to the own TLV Offset field. The value of this field depends on the particular OAM type and it can be different even for frames of the same OAM type.

Services provided by OAM can be classified in two different families depending on their purpose:

1. *Fault management* enables detection and notification of different defect conditions.
2. *Performance monitoring* allows measurement of performance parameters such

as packet loss, delay and delay variation.

1.6.1.1 Fault Management

The most important Fault management OAM procedures are defined in IEEE 802.1ag but the standard ITU-T Y.1731 greatly expands the functionality of this IEEE standard (see Table 1.12).

Table 1.12
Ethernet Fault Management OAM functions

Function	Messages	Standards	Description
Continuity check	CCM	ITU-T Y.1731 IEEE 802.1ag	Detects loss of continuity between the endpoints in an OAM domain.
Loopback	LBM, LBR	ITU-T Y.1731 IEEE 802.1ag	Verifies bidirectional connectivity between OAM entities.
Link trace	LTM, LTR	ITU-T Y.1731 IEEE 802.1ag	Computes the path between two OAM entities.
Alarm Indication Signal	AIS	ITU-T Y.1731	Communicates downstream a failure in a server level.
Test Signal	TST	ITU-T Y.1731	Performs one-way diagnostic tests.
Remote Defect Indication	CCN	ITU-T Y.1731 IEEE 802.1ag	Communicates upstream a failure in a server level.
Lock Signal	LCK	ITU-T Y.1731	Signals intentional diagnostic actions.
Automatic Protection Switching	APS	ITU-T Y.1731	Control linear protection switching operations
Maintenance Communications Channel	MCC	ITU-T Y.1731	Provides a communications channel to enable remote maintenance tasks.
Experimental OAM	EXM, EXR	ITU-T Y.1731	Used to try new OAM functionalities
Vendor-specific OAM	VSM, VSR	ITU-T Y.1731	Transports own vendor specific OAM

The Continuity Check Message (CCM) is probably the most important Ethernet OAM message. Its main purpose is detection of Loss Of Continuity (LOC) between two MEPs but also has other functions like communication of the Remote Defect Indication (RDI). The LoopBack Message (LBM) and LoopBack Reply (LBR) are used either to verify bidirectional connectivity of a MEP with a MIP or MEP or to perform in-service or out-of-service diagnostics between two MEPs. In the latter case it may be necessary for the LBM/LBR messages to carry test patterns to enable bit error detection or bandwidth estimation. The Link Trace Message (LTM) and Link Trace Reply (LTR) constitute the basis of the link trace OAM function.

This function is initiated on-demand by a MEP and enables retrieval of adjacency relationships and fault localization. The the link trace function retrieves information about the nodes placed between source and destination in a similar way that the IP trace route function does. Other fault management OAM mechanism is the Ethernet Alarm Indication Signal (AIS). When a MEP detects a connectivity failure in a serving OAM level, it sends an AIS in the next higher OAM level in the direction away from the detected failure to inform to the peer MEPs that the server path has failed and to suppress other redundant alarms at upper levels. The LoCKed (LCK) message

is used to communicate the administrative locking of a OAM level, enabling client MEPs to differentiate between the defect conditions and intentional diagnostic actions at the performed at serving OAM level. The TeST (TST) message carries a test pattern and it is used to perform one-way diagnostics tests. This includes testing of throughput, frame loss, bit errors, etc. These tests can be in-service or out-of-service. Out-of-service tests require previous administrative locking of the MEP to be tested. The Automatic Protection Switching (APS) OAM message is used to control protection switching operation in linear topologies.

The APS payload is defined in ITU-T Y.1731 but applications are included in ITU-T G.8031/Y.1342 for Ethernet linear protection procedures. The Maintenance Communications Channel (MCC), provides a data channel with remote maintenance purposes. The specific contents of this channel is not specified and it is vendor specific. Finally, the OAM protocol can be extended with OAM messages with the EXperimental Message (EXM), EXperimental Reply (EXR), Vendor-Specific Message (VSM) and Vendor-Specific Reply (VSR).

Table 1.13
Ethernet Performance Monitoring OAM functions

Function	Messages	Standards	Description
Dual-ended Frame Loss	CCM	ITU-T Y.1731	Frame loss measurement coordinated from two network nodes
Single-ended Frame Loss	LMM, LMR	ITU-T Y.1731	On-demand frame loss measurement carried out from a single end.
One-way Frame Delay	1DM	ITU-T Y.1731	One-way delay measurement coordinated from two network nodes
Two-way Frame Delay	DMM, DMR	ITU-T Y.1731	One-way delay frame loss measurement carried out from a single end.
Throughput	LBM, LBR, TST	ITU-T Y.1731	Maximum bit rate without frame loss.

1.6.1.2 Performance Monitoring

The ITU-T Recommendation Y.1731 defines network performance parameters along with the necessary OAM functions to compute these parameters in real environments (see Table 1.13). Specifically, the ITU-T Y.1731 defines the following four parameters:

- *Throughput*, is the maximum rate at which no frame is dropped. The TST or the LBM/LBR OAM messages are used to carry out one-way or two-way throughput measurements.
- *Frame loss ratio*, the ratio of service frames lost and the total number of service frames delivered, can be measured in two different ways. Dual-ended measurements use the CCM OAM message while the single-ended measurement uses frame Loss Measurement Message (LMM) and frame Loss Measurement Reply (LMR) OAM payloads. Frame loss ratio test is the result of the

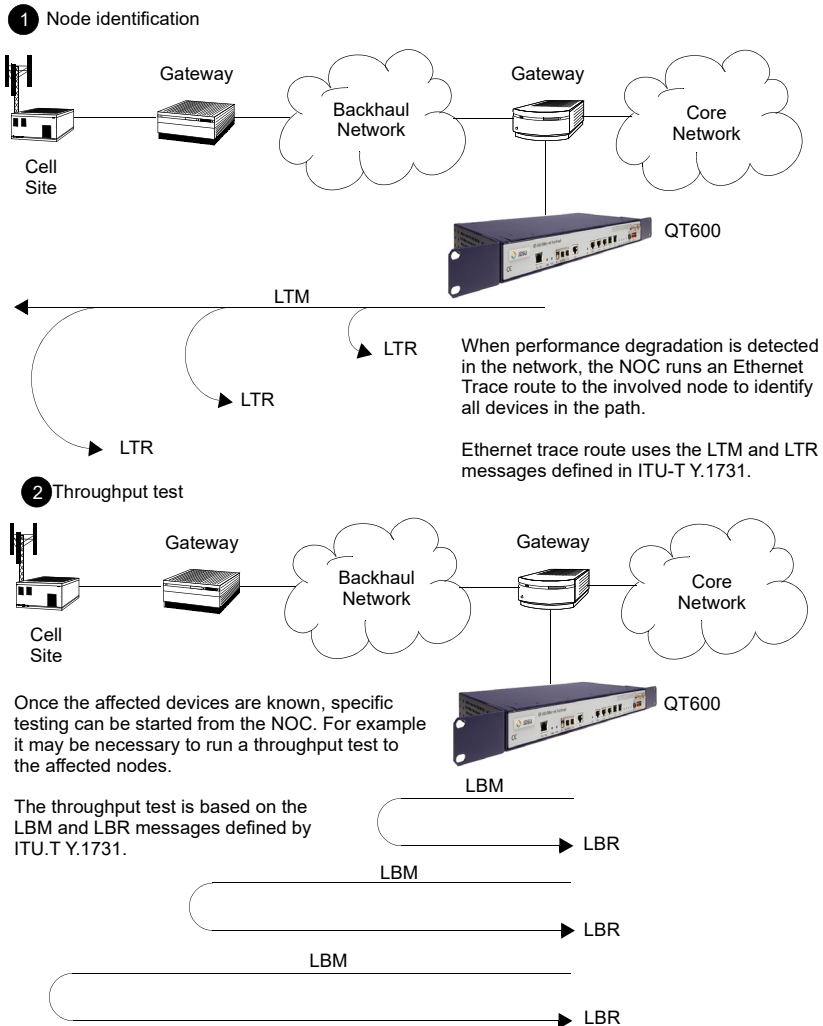


Figure 1.57 Using ITU-T Y.1731 Loopback and Trace route messages with the QT600 Ethernet probe for Ethernet network diagnostics.

exchange of the appropriate counts of transmitted and received frames and correlation of local and far end data in every MEP.

- Frame delay* is computed either via a one-way test or a two-way test. In one-way frame delay measurement, the MEP sends a one-way Delay Measurement (1DM) message with a timestamp and its peer calculates the delay as the difference between the reception time and the timestamp value. This calculation requires clock synchronization in peering MEPs. If clock synchronization is not available, delay can be still calculated as a two-way test. In this case the Delay

Measurement Message (DMM) and the Delay Measurement Replay (DMR) OAM payloads are used. The result is a round trip delay between peering MEPs rather than the one-way delay.

- Unlike frame delay, *frame delay variation* does not require clock synchronization between MEPs. This parameter is computed with the same mechanisms that frame delay. Specifically, frame delay variation is calculated as the difference between two consecutive two-way frame delay measurements.

1.6.2 MPLS OAM

As MPLS is adopted as the main building block of the carrier-class packet switched network and the key enabler of versatile multiplay services, its OAM functionality is being extended to (at least) the same level of any legacy TDM technology.

Evolution of MPLS OAM standards has been quite different to Ethernet OAM. While ITU-T Y.1731 and IEEE 802.1ag basically define the same OAM procedures, the MPLS OAM standards are fragmented and they are sometimes duplicated and incompatible.

Responsibility on MPLS OAM standardization rely on the ITU-T and the IETF. MPLS was introduced as an improvement for IP and thus for the Internet. For this reason, first MPLS standards were generated under the umbrella of the IETF. On the other hand, when carriers adopted the MPLS technology, the ITU-T generated its own MPLS network reference model and a set of recommendations specially suited for carriers and service providers, including OAM recommendations. The existing OAM initiatives are the following

- The RFC 4379 extends the *ping* and *trace route* mechanisms, widely available and popular in IP networks, to MPLS. MPLS ping and trace route are based on a UDP echo request and reply. For this reason these mechanisms are not suitable for protocol agnostic transport networks.
- *Bidirectional Forwarding Detection* (BFD), aim is to provide low-overhead, short-duration detection of failures in the path between MPLS devices. The RFC 5880 BFD mechanism is no more than a simple and flexible *hello* protocol.
- *Virtual Circuit Connectivity Verification* defines Control Channels for pseudowires, Connectivity Verification (CV) procedures, and setup mechanisms compatible with LDP and other pseudowire control protocols.
- The *ITU-T Y.1711*, defines Forward Defect Indication (FDI), Backward Defect Indication (BDI), LSP trail identification and other OAM mechanisms. No further developments of this OAM model are expected in the future. OAM func-

tions defined by this ITU-T recommendation are expected to be migrated to the MPLS-TP OAM framework jointly developed by the IETF and the ITU-T.

- The *MPLS-TP OAM* framework provides exhaustive OAM to MPLS specifically adapted for the transport network. MPLS-TP OAM uses existing procedures (BFD, VCCV, MPLS ping, trace route,...) wherever possible. Extensions or new OAM mechanisms are defined only where necessary

1.6.2.1 MPLS Ping and Trace Route

Ping and trace route are two popular troubleshooting tools for IP networks. Ping checks end-to-end connectivity and trace route provides fault location by means hop-by-hop connection verification through the transmission origin and destination.

The MPLS ping and trace route have similar purpose that their equivalent IP counterparts. However, the MPLS ping and trace route have a message format that is specific of them. Both the MPLS ping and trace route are based on a common MPLS echo request and reply message defined in RFC 4379. The echo request / reply messages are UDP packets with standard IPv4 or IPv6 headers (see Figure 1.58). The UDP port used by the MPLS echo request / reply is the 3503.

To test a particular LSP with the MPLS ping tool, the source sends an MPLS echo request message which carries a FEC specification in its payload through the LSP. The echo request message is captured by intermediate routers in the LSP and they check that the label used in their incoming interface is the one advertised for the FEC specified in the echo request packet payload. This procedure can be used to check the coherence between labels and FEC through the LSP. When the echo request is received by the egress LER, it checks that the FEC specified in the payload can be reached by some router interface and an echo reply message is sent to the source on success.

The MPLS ping sets the MPLS Time To Live (TTL) to 255 but the IP TTL to 1. Furthermore, the destination address of the echo request message is a 127.0.0.0/8 address that belongs to an special subnet that is never expected to be found in a network.

If the LSP under test is broken, then some LSR will receive an exposed (unlabeled) echo request message and the TTL will be decremented to 0 by the LSR. The 127.0.0.0/8 destination address cuts any possibility of this packet to be routed to a wrong destination. An error message is then issued towards the source.

MPLS trace route is similar but in this case, the source generates a sequence of MPLS request messages with increasing TTL values. The MPLS trace route uses an special payload Type, Length, Value (TLV) known as downstream mapping TLV to

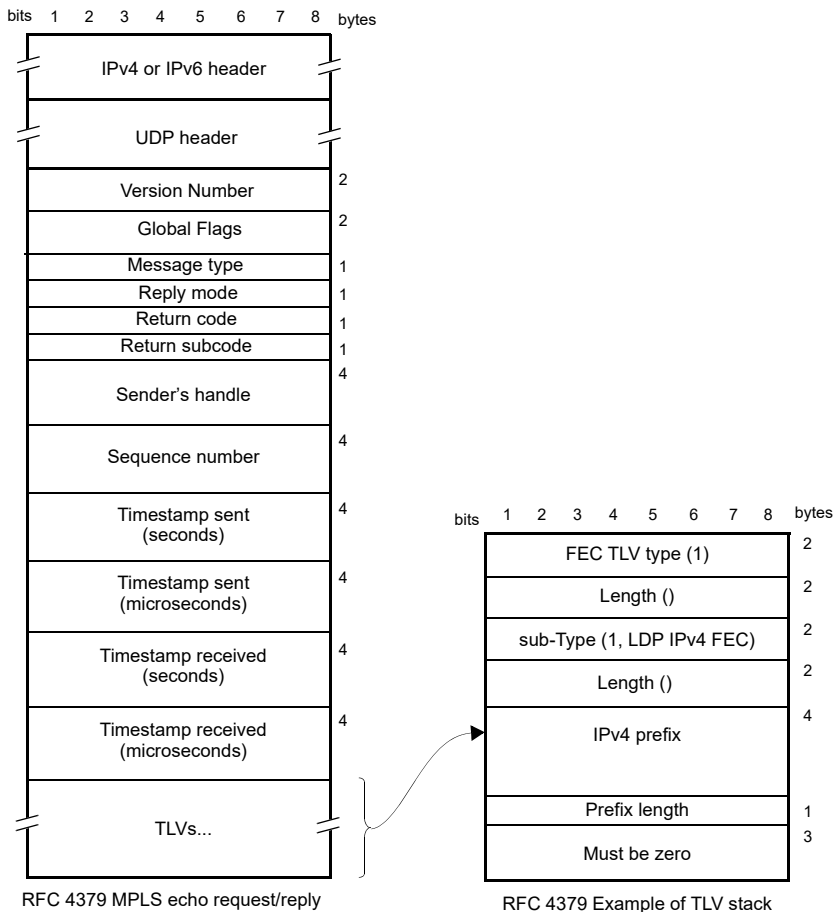


Figure 1.58 The MPLS ping and trace router tools are based on a MPLS echo request and reply messages. These messages are UDP packets encapsulated in IPv4 or IPv6 datagrams and probably labeled by one or more MPLS labels.

discover downstream neighbors through the LSP. In each iteration, one further LSR is discovered. The process continues until the last router in the LSP is reached or some error condition is detected.

MPLS ping and trace route are good replacement of IP ping and trace route in IP/MPLS networks because they provide accurate diagnostics where the native IP tools are unclear. MPLS ping and trace route are however limited because they rely on the IP protocol stack and they cannot be used when IP is not available.

1.6.2.2 Bidirectional Forwarding Detection

The Bidirectional Forwarding Detection (BFD) mechanism, is defined in RFC 5880 as a general purpose *hello* protocol.

The goal BFD is to provide low-overhead, short-duration detection of failures in the path between routers. Additionally, BFD provides a single mechanism that can be used for liveness detection over any media, at any protocol layer. MPLS BFD packets are required to use an UDP encapsulation however. The UDP destination port for BFD sessions is the 3784. UDP BFD packets may use an IPv4 or IPv6 envelope. The destination IP address for such packets is chosen within the 127.0.0.0/8 subnet (IPv4) or the 0:0:0:0:FFFF:7F00/104 range (IPv6).

The BFD mechanism is quite flexible and it supports various operation modes:

- *Asynchronous mode*, in this mode the transmission ends periodically send BFD Control packets to one another. If a number of consecutive packets are not received by the remote system, the transmission path is declared to be down.
- *Demand mode*, in this case, one system may ask the remote system to stop sending BFD Control packets. Transmission is resumed on demand when it is required explicit verification.

The BFD has an *Echo* functionality that can be enabled both in *Asynchronous* and the *Demand* modes. If the *Echo* is enabled an stream of special BFD Echo packets is generated. These packets are looped back to the origin by the remote system using its forwarding path.

The BFD is versatile enough to allow the end systems to negotiate the Control and Echo packet transmission periods with specific protocol functionalities (*Desired Min TX Interval*, *Required Min RX Interval*, *Required Min Echo RX Interval* Control packet fields). BFD sessions can also be multiplexed by using the *My Discriminator* and *Your Discriminator* Control packet fields (see Figure 1.59). Another interesting BFD feature is the ability to authenticate the protocol packets in order to make sure they come from the correct source.

In MPLS environments, BFD can be used to detect a data plane failure in the forwarding path of an MPLS LSP. This functionality is different but complementary to the already mentioned MPLS ping. The MPLS ping checks coherence between the LSPs and their associated FECs and therefore it is useful to verify MPLS control plane failures.

When used in MPLS, BFD packets require an IP envelope and for this reason the BFD mechanism is not available where the IP protocol stack is not present, like for example in the transport network.

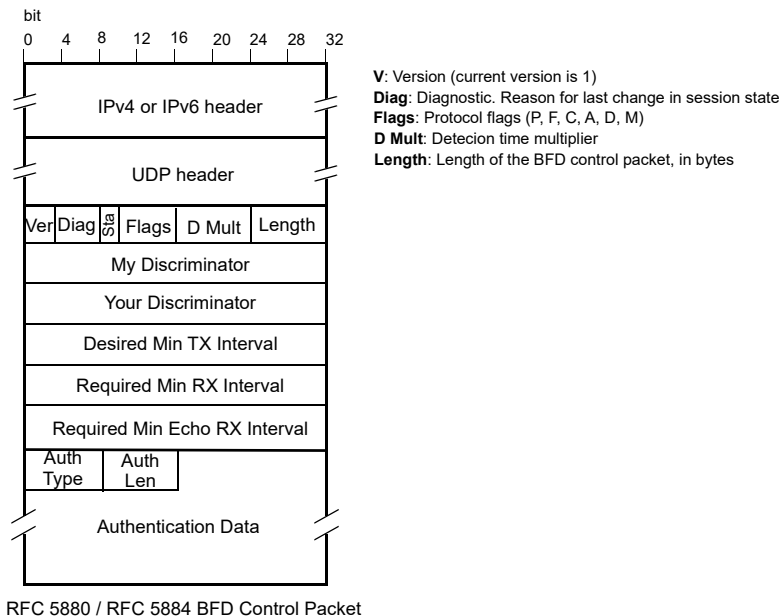


Figure 1.59 BFD control packet format

1.6.2.3 Virtual Circuit Connectivity Verification

RFC 5085 defines the Virtual Circuit Connectivity Verification (VCCV) channel for pseudowires. This channel can be used to supply OAM functionality.

VCCV is enabled and configured during the pseudowire setup process through the LDP or other pseudowire signalling protocol. Due to this particular setup mechanism, VCCV cannot be modified after it has been configured.

The VCCV mechanism relies on a Control Channel (CC) which in turn carry several types of verification procedures defined by the Connectivity Verification (CV). There are several types of CC and CV. RFC 5085 includes extensions for LDP (an other pseudowire signalling protocols) to include VCCV capability information, including combinations of supported CCs and CV types. The currently available VCCV CC types are (see Figure 1.60):

- *In-band VCCV*: User plane and VCCV packets have identical label stacks. But rather than the pseudowire control word, VCCV packets use the PseudoWire Associated Channel (PWACH) as defined in RFC 4385. The control word always starts with 0000 (binary representation) but the PWACH Header

(PWACH) starts with 0001. Recognition and demultiplexing of user and control packets is thus possible.

- Out-of-band VCCV:** In this case the VCCV packets use the special MPLS router alert label which has the reserved value of 1. The MPLS router alert label is pushed in the top of the label stack, after the pseudowire label. Packets containing this special label receive special treatment. They are delivered to the router processor rather than being switched to an outgoing interface. The out-of-band VCCV has the inconvenience that user and control packets may follow different paths if a load balancing mechanism like the Equal Cost Multi-Path (ECMP) is used.

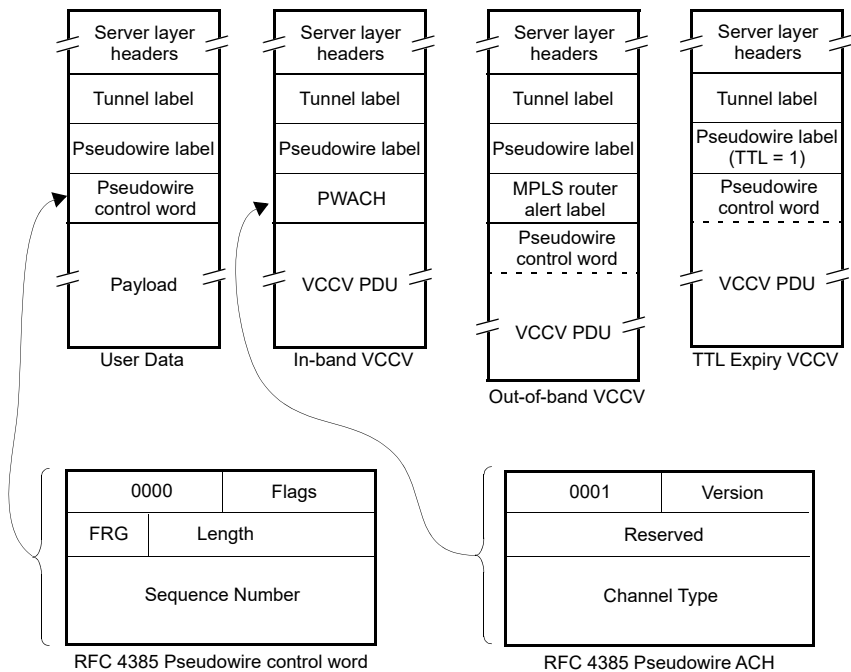


Figure 1.60 In-band and out-of-band multiplexing of pseudowire data plane information and VCCV.

- TTL Expiry VCCV:** This CC type does not require any special header or label. It simply sets the TTL value to 1 in the pseudowire label. In this way, when the control packets reach the destination node, the TTL value is decremented one unit (to 0) and the packets are thus processed by the node rather than being forwarded. Like the out-of-band VCCV, the TTL Expiry has problems dealing with load balancing.

The CV types accepted by VCCV are the ICMP ping and the MPLS ping. BFD is also compatible with VCCV. RFC 5885 defines the VC types for BFD over VCCV with or without IP/UDP encapsulation. The BFD without IP/UDP encapsulation is of special relevance because it is the basis of the Continuity Check (CC) and Connection Verification (CV) mechanisms for MPLS-TP.

1.6.2.4 ITU-T Y.1711

The ITU-T Y.1711 fundamental concept is the *Connection Verification (CV)* flow. LSPs may have an associated CV flow to them for OAM purposes. The ingress LER generates CV packets and these packets are received by the egress LER. If some faulty condition is found in a CV flow by the egress LER then one or more defects will be notified.

The ingress LER generates one CV packet per second. The egress LER waits for three seconds to receive a CV packet. After three seconds, the node declares a loss of CV defect (dLOCV). Even if CV packets are received, they may contain different kinds of errors. For example if packets are received with a frequency above the nominal rate of one packet per second something may be wrong in the network (see Table 1.14).

Some defects (dTTSI_Mismatch, dTTSI_Mismerge) require identification of the LSP and ingress LER. This functionality is provided by the *Trail Termination Source Identifier (TTSI)*. The TTSI contains the 16 byte IPv6 address corresponding to the ingress LER output port (LSR identifier) and a 4 byte tunnel identifier (LSP identifier) (see Figure 1.61).

If a LSR detects some ITU-T Y.1711 defect, then it propagates the information through two OAM flows defined in this recommendation. These MPLS OAM defect notification flows are the *Forward Defect Indication (FDI)*, *Backward Defect Indication (BDI)*. The defect notification flows are copied to all affected MPLS client layers. They are generated with a nominal rate of one packet per second. For the BDI to work, it must exist a return path from the egress LER to the ingress LER.

Some applications require defect detection faster than 3 seconds, the delay required by the CV flow to operate. The most important example of this is protection switching, that is often required to switch to a protection path in less than 50 ms. For this reason, the ITU-T Y.1711 also defines the *Fast Failure Detection (FFD)* flow. The FFD is similar to the CV but the packet generation period is configurable and it is not limited to 1 second. The FFD is much better suited for protection switching than the CV.

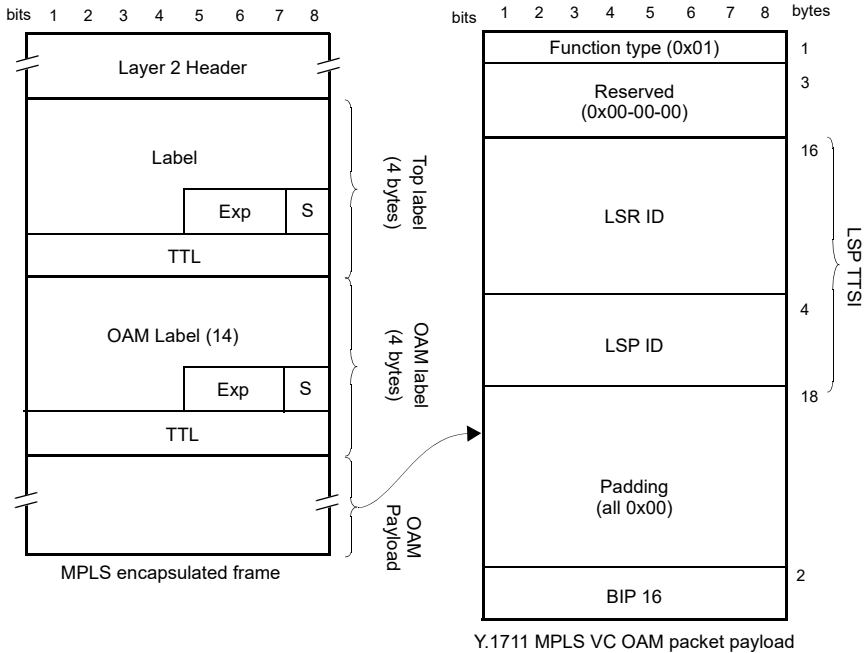


Figure 1.61 The structure of the Y.1711 OAM channel is based on double labeled frames. The label value used for OAM is 14.

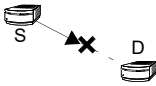
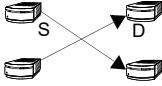
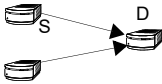
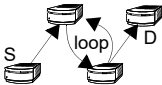
ITU-T Y.1711 OAM are limited in their scope. For example, out-of-service analysis and troubleshooting tools remain undefined within the ITU-T Y.1711 OAM framework. Furthermore, ITU-T Y.1711 OAM services are provided through MPLS label 14 whose usage is deprecated. ITU-T Y.1711 OAM functionality is expected to be redefined under MPLS-TP but now using the new MPLS label 13.

1.6.2.5 MPLS-TP OAM

Extensive OAM is a key requirement for MPLS-TP. In general terms, existing MPLS OAM mechanisms are used wherever possible and extensions or new OAM mechanisms are defined only where necessary.

New MPLS OAM functionality operate in-band on the transport pseudowire or LSP such that they do not depend on any other protocol layer. OAM packets are distinguished from the user data packets using the GAL (label 13), the PWACH and GACH.

Table 1.14
ITU-T Y.1711 MPLS layer defects

Defect	Codepoint	Diagram	Description
dLOCV	0x02-01		No CV packets are received in the LSP. This defect can be caused if the LSP is broken due to a configuration problem, degraded transmission medium or a broken LSR.
dTTSL_Mismatch	0x02-02		Unexpected TTSI found by the egress LER. This is caused by an LSP misconnection.
dTTSL_Mismerge	0x02-03		Both correct and unexpected TTSIs are found within the same LSP and they are detected when the LSP is merged with traffic from unsolicited sources due to a configuration failure.
dExcess	0x02-04		CV packets are detected with rate above the nominal rate of 1 packet/s. Possible reason for this defects are self-mismerge or Denial of Service (DoS) attacks.

MPLS-TP defines a multilevel, hierarchical OAM architecture. It defines several MEP types carrying out OAM tasks at section, end-to-end LSP and pseudowire level (see Figure 1.62). The MPLS-TP OAM framework also provides support for maintenance of arbitrary LSP and pseudowire parts.

MPLS-TP OAM mechanisms are classified in proactive monitoring and on-demand functions.

Proactive monitoring is carried out continuously or it is preconfigured to act on certain events such as alarm signals. Proactive monitoring is usually performed in-service. MPLS-TP proactive monitoring is based on the *Continuity Check (CC)* and *Connectivity Verification (CV)* flows. The former is used to check the availability of the peer MEP, the latter detects unexpected connections caused by LSP mismerges or misconnections.

The MPLS-TP proactive monitoring functions are derived from the VCCV for pseudowires but now the VCCV mechanism is supported also by LSPs with the help of the GACH. The CC and CV OAM payload use BFD packets without IP and UDP envelopes. The BFD requires no modification to operate in MPLS-TP but it has to be profiled to meet the MPLS-TP requirements.

The CV is different to the CC in that the CV requires identification of the source MEP. A globally unique alphanumeric MEP ID is used for this purpose (see Figure 1.63).

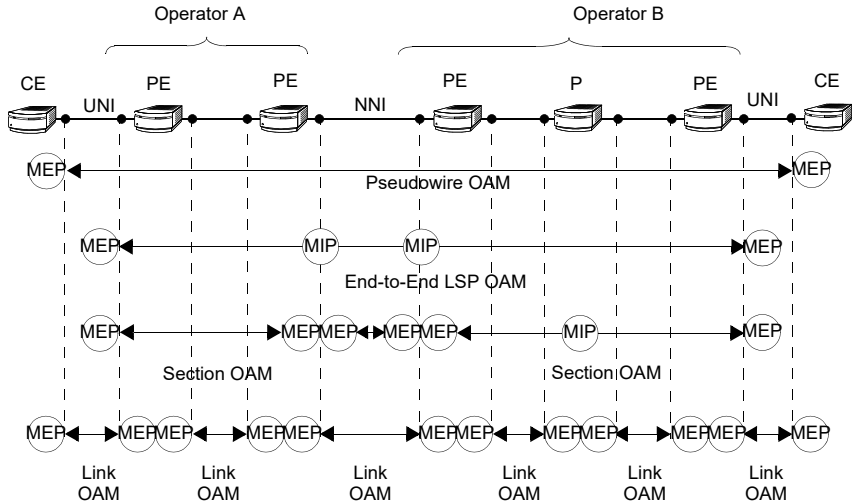


Figure 1.62 The MPLS-TP OAM framework defines different MEPs and MIPs operating at pseudowire, LSP and section levels.

The CC and CV can be used to detect several defects in transport LSPs or pseudowires. Examples of this are the Loss of Continuity (LOC) defect, the Mis-connectivity defect, Period misconfiguration defect and Unexpected encapsulation defect. To share defect information MPLS-TP defines an Alarm Indication Signal (AIS) and a Remote Defect Indication (RDI).

The CC and CV flows are associated with fault management but the MPLS-TP OAM provides also performance management functions. Packet loss is measured by means special packet Loss Measurement (LM) OAM packets and latency measures are assisted by Delay Measurement (DM) OAM packets.

Unlike proactive monitoring tools, on-demand OAM mechanisms are initiated manually and for a limited amount of time, usually for operations such as diagnostics to investigate a defect condition. On demand OAM is also planned for MPLS-TP. In order to meet this requirement, the IETF is working in appropriate extensions of the MPLS ping and trace route for MPLS-TP. These extensions enable the ping and trace route to operate both with and without IP, being the IP-less operation the most interesting one for transport applications.

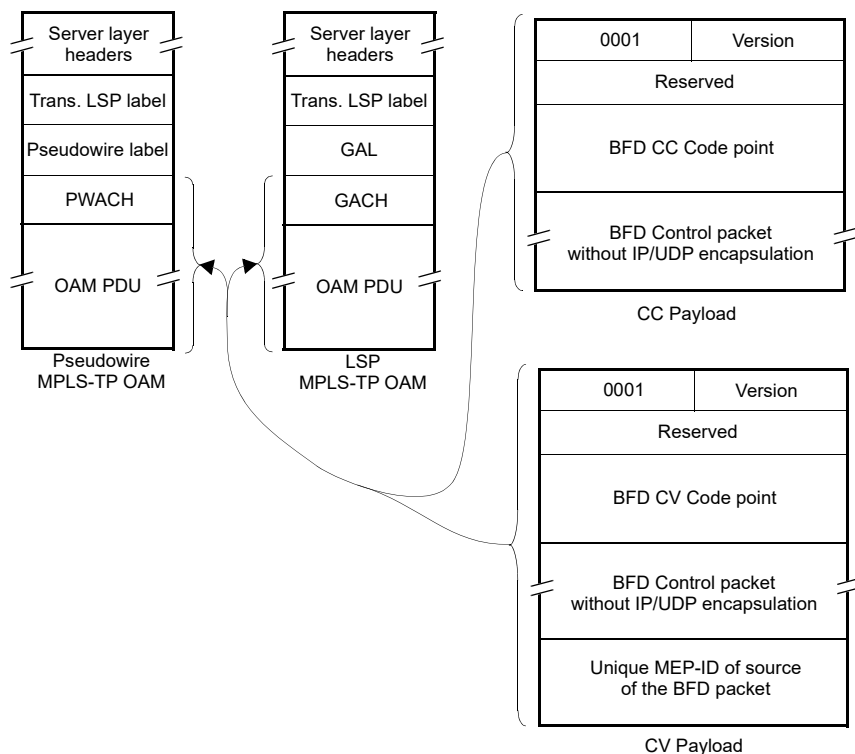


Figure 1.63 Projected MPLS-TP messages for proactive monitoring. These messages constitute the CC and CV flows and they are based on the BFD mechanism.

Selected Bibliography

- [1] IEEE 802.1D-2004, "Media Access Control (MAC) Bridges," June 2004.
- [2] IEEE 802.1Q-2005, "Virtual Bridged Local Area Networks Revision," May 2006.
- [3] IEEE 802.1ad-2005, "Virtual Bridged Local Area Networks Amendment 4: Provider Bridges," May 2006.
- [4] IEEE 802.1ag-2007, "Virtual Bridged Local Area Networks Amendment 5: Connectivity Fault Management," December 2007
- [5] IEEE 802.1ah-2008, "Virtual Bridged Local Area Networks Amendment 7: Provider Backbone Bridges," August 2008.
- [6] IEEE 802.1Qay-2009, "Virtual Bridged Local Area Networks Amendment 10: Provider Backbone Bridge Traffic Engineering," August 2009.

- [7] ITU-T Rec. Y.1540, "Internet protocol data communication service - IP packet transfer and availability performance parameters," November 2007.
- [8] ITU-T Rec. Y.1541, "Network performance objectives for IP-based services," February 2006.
- [9] ITU-T Rec. Y.1711, "Operation & Maintenance mechanism for MPLS networks," February 2004.
- [10] ITU-T Rec. Y.1731, "OAM functions and mechanisms for Ethernet based networks," February 2008
- [11] Allan D., Bragg N., McGuire A., Reid A., "Ethernet as Carrier Transport Infrastructure," *IEEE Communications Magazine*, Feb 2006, pp. 134-140.
- [12] Ryoo J., Song J., Park J., Joo B., "OAM and its Performance Monitoring Mechanisms for Carrier Ethernet Transport Networks," *IEEE Communications Magazine*, March 2008, pp.97-103.
- [13] Rosen E., Viswanathan A., Callon R., "Multiprotocol Label Switching architecture," IETF Request For Comments RFC 3031, January 2001.
- [14] Rosen E., Tappan D., Fedorkow G., Rekhter Y., Farinacci D., Li T., Conta A., "MPLS Label Stack Encoding," IETF Request For Comments RFC 3032, January 2001.
- [15] Andersson L., Doolan P., Feldman N., Fredette A., Thomas B., "LDP Specification," IETF Request For Comments RFC 3036, January 2001.
- [16] Awduche D., Berger L., Gan D., Li T., Srinivasan V., Swallow G., "RSVP-TE: Extensions to RSVP for LSP Tunnels", IETF Request For Comments RFC 3209, December 2001.
- [17] Jamoussi B., Andersson L., Callon R., Dantu R., Wu L., Doolan P., Worster T., Feldman N., Fredette A., Girish M., Gray E., Heinanen J., Kilty T., Malis A., "Constraint-Based LSP Setup using LDP," IETF Request For Comments RFC 3212, January 2002.
- [18] Bryant S., Pate P., "Pseudo Wire Emulation Edge-to-Edge (PWE3) architecture," IETF Request For Comments RFC 3985, March 2005.
- [19] Martini L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)," IETF Request For Comments RFC 4446, April 2006.
- [20] Martini L., Rosen E., El-Aawar N., Smith T., Heron G., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)," IETF Request For Comments RFC 4447, April 2006.
- [21] Martini L., Rosen E., El-Aawar N., Heron G., "Encapsulation Methods for Transport of Ethernet over MPLS Networks," IETF Request For Comments RFC 4448, April 2006.
- [22] Lasserre M., Kompella V., "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling," IETF Request For Comments RFC 4762, January 2007.
- [23] Awduche D. et al., "Overview and Principles of Internet Traffic Engineering," IETF Request For Comments RFC 3272, May 2002.
- [24] Niven-Jenkins B., Brungard D., Betts M., Sprecher N., Ueno S., "Requirements of an MPLS Transport Profile," IETF Request For Comments RFC 5654, September 2009.
- [25] Bocci M., Bryant S., Frost D., Levrau L., Berger L., "A Framework for MPLS in Transport Networks," IETF Request For Comments RFC 5921, July 2010.

- [26] Frost D., Bryant S., Bocci M., "MPLS Transport Profile Data Plane Architecture," IETF Request For Comments RFC 5960, August 2010.
- [27] Bocci M., Vigoureux M., Bryant S., "MPLS Generic Associated Channel," IETF Request for Comments RFC 5586, June 2009.
- [28] Kompella K., Swallow G., "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures," IETF Request For Comments RFC 4379, February 2006.
- [29] Katz D., Ward D., "Bidirectional Forwarding Detection (BFD)," IETF Request For Comments RFC 5880, June 2010.
- [30] Aggarwal R., Kompella K., Nedeau T., Swallow G., "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)," IETF Request For Comments RFC 5884, June 2010.
- [31] Nadeau T., Pignataro C., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires," IETF Request for Comments RFC 5885, December 2007.
- [32] Bai Y., Ito, M.R., "QoS Control for Video and Audio Communication in Conventional and Active Networks: Approaches and Comparison," *IEEE Communications Surveys*, vol. 6, no. 1, first quarter 2004.
- [33] Labrador M.A., Banerjee S., "Packet Dropping Policies for ATM and IP Networks," *IEEE Communications Surveys*, vol. 2, no. 3, third quarter 1999.
- [34] Michaut F., Lepage F., "Application-Oriented Network Metrology: Metrics and Active Measurement Tools," *IEEE Communications Surveys*, vol. 7, no. 2, second quarter 2005.
- [35] Xi Peng Xiao, Telkamp T., Fineberg V., Cheng Chen, Lionel M. Ni, "A Practical Approach for Providing QoS in the Internet Backbone," *IEEE Communications Magazine*, December 2002, pp. 56-62.
- [36] Yang Chen, Chunming Qiao, Hamdi M., Tsang D. H. K., "Proportional Differentiation: A Scalable QoS Approach," *IEEE Communications Magazine*, June 2003, pp. 52-58.
- [37] Adams A., Bu T., Horowitz J., Towsley D., Cáceres R., Duffield N., Lo Presti F., "The Use of End-to-End Multicast Measurements for Characterizing Internal Network Behavior," *IEEE Communications Magazine*, May 2000, pp. 152-158.
- [38] Christin N., Liebeherr J., "A QoS Architecture for Quantitative Service Differentiation," *IEEE Communications Magazine*, June 2003, pp. 38-45.
- [39] Almes et al., "A One-way Packet Loss Metric for IPPM," IETF Request For Comments RFC 2680, September 1999.
- [40] Paxson V., Almes G., Mahdavi J., Mathis M., "Framework for IP Performance Metrics", IETF Request for Comments RFC 2330, May 1998.
- [41] Matrawy A., Lambaradis I., "A Survey of Congestion Control Schemes for Multicast Video Applications," *IEEE Communications Surveys*, vol. 5, no. 2, fourth quarter 2003.
- [42] Tryfonas C., Varma A., "MPEG-2 Transport over ATM Networks," *IEEE Communications Surveys*, vol. 2, no. 4, fourth quarter 1999.
- [43] Vali D., Plakalis S., Kaloxylas A., "A Survey of Internet QoS Signaling," *IEEE Communications Surveys*, vol. 6, no. 4, fourth quarter 2004.

- [44] Marthy L., Edwards C., Hutchison D., "The Internet: A Global Telecommunications Solution?," *IEEE Network Magazine*, July/August 2000, pp. 46-57.
- [45] Xiao X., Ni L. M., "Internet QoS: A Big Picture," *IEEE Network Magazine*, March/April 1999, pp. 8-18.
- [46] White, P. P., "RSVP and Integrated Services in the Internet: A Tutorial," *IEEE Communications Magazine*, May 1997, pp. 100-106.
- [47] Giordano S., Salsano S., Van den Berghe S., Ventre G., Giannakopoulos D., "Advanced QoS Provisioning in IP Networks: The European Premium IP Projects," *IEEE Communications Magazine*, January 2003, pp. 2-8.
- [48] Mase K., "Toward Scalable Admission Control for VoIP Networks," *IEEE Communications Magazine*, July 2004, pp. 42-47.
- [49] Welzl M., Franzens L., Mühlhäuser M., "Scalability and Quality of Service: A Trade-off?," *IEEE Communications Magazine*, June 2003, pp. 32-36.
- [50] Cavendish D., Ohta H., Rakotoranto H., "Operation, Administration, and Maintenance in MPLS Networks," *IEEE Communications Magazine*, October 2004, pp. 91-99.
- [51] Zhang L., Deering S., Estrin D., Shenker S., Zappala D., "RSVP: A New Resource Reservation Protocol," *IEEE Network Magazine*, September 1993, vol. 7, no. 5.
- [52] Braden R., Clark D., Shenker S., "Integrated Services in the Internet Architecture: an Overview," IETF Request For Comments RFC 1633, June 1994.
- [53] Blake S., Black D., Carlson M., Davies E., Wang Z., Weiss W., "An architecture for Differentiated Services", IETF Request for Comments RFC 2475, December 1998.
- [54] Heinanen J., Baker F., Weiss W., Wrockawski J., "Assured Forwarding PHB Group", IETF Request for Comments RFC 2597, June 1999.
- [55] Davie B. et al., "An Expedited Forwarding PHB (Per-Hop Behavior)", IETF Request For Comments RFC 3246, March 2002.
- [56] Braden R., Zhang L., Berson S., Herzog S., Jamin S., "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", IETF Request For Comments RFC 2205, September 1997.
- [57] Wroclawsky J., "The use of RSVP with IETF Integrated Services", IETF Request For Comments RFC 2210, September 1997.
- [58] Shenker S., Wroclawski J., "General Characterization Parameters for Integrated Service Network Elements", IETF Request For Comments RFC 2215, September 1997.

