

Ethernet Service Deployment

A Comprehensive Guide to
Test, Measurement, and Service Assurance

Ethernet Service Deployment

A Comprehensive Guide to
Test, Measurement, and Service Assurance

The information contained in this document is the property of ALBEDO Telecom S.L. and is supplied without liability for errors and omissions.

No part of this document may be reproduced or used except as authorised by contract or other written permission from ALBEDO Telecom S.L. The copyright and all restrictions on reproduction and use apply to all media in which this information may be placed.

© ALBEDO Telecom S.L. 2011
All rights reserved

Issue 1, 10/11

For any query or requirement regarding the *AT-2048 E1 / Datacom tester*, contact with ALBEDO Telecom using the following contact details:

ALBEDO Telecom S.L.
C/ Joan d'Àustria 112

08018 Barcelona - Spain

E-mail: support.telecom@albedo.biz
Telephone: +34 93 221 28 73

Table of Contents

Chapter 1: Gigabit Ethernet Networks	1
1 Gb/s Ethernet	5
1000BASE-X Architecture.....	5
1000BASE-T Architecture	8
10 Gb/s Ethernet.....	9
Optical Transmission	10
10 Gb/s Ethernet over Copper	12
Compatibility with SDH/SONET.....	14
Higher Speed Ethernet.....	16
Multilane Distribution Procedure	17
Physical Media	19
40 Gb/s and 100 Gb/s Ethernet over OTN.....	21
Chapter 2: Switched Ethernet	25
From Shared to Dedicated Media.....	26
Ethernet Bridging	27
Full-Duplex Operation	30
Hands-on: Performance of Ethernet Switches with RFC 2544	32
Virtual LANs	40
Hands-on: Transparency Tests across VLANs.....	42
The Spanning Tree Protocol Family	47
Redundancy and Bridging.....	49
The Classic Spanning Tree Protocol	50
Rapid Spanning Tree Protocol.....	54
Multiple Spanning Tree Protocol	54
The Network Layer	56
The TCP/IP Reference Model.....	57
The Internet Protocol	58
Internet Control Message Protocol	62
Address Resolution Protocol	62
Higher Layers of the TCP/IP Protocol Stack.....	66
Chapter 3: Carrier Ethernet	71
Ethernet as a MAN / WAN Service.....	73
Network Architecture.....	74
Ethernet Virtual Connections	76
Multiplexing and Bundling	77
MEF Generic Service Types.....	78
Connectivity Services.....	80
Ethernet Deployment Alternatives	84

Optical Ethernet.....	86
Ethernet over WDM.....	89
Ethernet over SDH or OTN	90
Limitations of Bridged Networks.....	95
Scalability	96
Quality of Service	99
Resiliency and Fault Tolerance	101
Multi-Protocol Label Switching.....	102
Labels.....	104
MPLS Forwarding Plane	107
Label Distribution	109
Martini Encapsulation.....	114
Pseudowires	116
Ethernet Pseudowires.....	122
MPLS Transport Profile	131
Quality of Service	141
QoS Control Basics	141
QoS In Ethernet Networks.....	145
QoS in IP Networks	157
End-to-End Performance Metrics	159
Operation, Administration and Maintenance.....	171
Ethernet OAM.....	171
MPLS OAM	178
Chapter 4: Ethernet in Access Networks	197
Fiber to the Neighborhood	199
Ethernet over Telephone Copper Pairs	202
Ethernet in Optical Access Networks	203
The Need of an Optical Access Network.....	204
1Gb/s and 10 Gb/s Ethernet PON.....	205
PON Concepts and Alternatives.....	206
EPON Particularities.....	212
Chapter 5: Ethernet Mobile Backhaul Networks.....	221
Towards the “All-IP” Network.....	223
Circuit Emulation Services	225
Transmission of Timing Information	226
Structure Aware vs. Structure Agnostic CES.....	228
Encapsulations for Structure Agnostic CES.....	231
Encapsulations for Structure Aware CES.....	234

Hands-on: MEF 18 and CES Certification	236
Ethernet Synchronization with IEEE 1588	238
Precedents: IP Synchronization with NTP	238
PTP Protocol Details	238
Protocol Encapsulation	241
Synchronous Ethernet	242
Ethernet Synchronization Messaging Channel	244
Appendix A : The OSI Reference Model	249
Seven Layers	250
Physical Layer	251
Data Link Layer	252
Network Layer	254
Transport Layer	254
Session Layer	254
Presentation Layer	255
Application Layer	255
Appendix B : Introduction to Ethernet	257
A Brief History of Ethernet	259
Ethernet and the OSI Reference Model	260
PHY and MAC Layer Independence	261
The Ethernet PHY	263
Legacy Ethernet Interfaces	263
Hands-on: Good Cabling Practices	268
Hands-on: Testing Auto-Negotiation	277
The Ethernet MAC	284
CSMA/CD	285
The Ethernet Frames	289
Hands-on: Determining Support of Jumbo Frames	292
The Ethernet LLC	296
Appendix C : Time Division Multiplexing	299
Deterministic TDM	300
Pulse Code Modulation	300
Channel Coding	301
Multiplexing and Multiple Access	302
Basic Rates: T1 and E1	304
The DS1 Frame	307

SDH/SONET	307
Network Elements.....	307
SDH/SONET Formats and Procedures.....	309
Ethernet over SDH	311
Optical Transport Network.....	312
Interfaces and Payload	312
Forward Error Correction.....	313
Overhead.....	313
Hands-on: Performance of TDM networks	315
In-Service and Out-of-Service Measurements.....	315
Bit Error Rate	315
ITU-T Error Performance Recommendations	317
Appendix D : Timing Methods.....	319
Synchronization Architectures.....	320
Synchronization Network Topologies.....	322
Interconnection of Nodes	324
Synchronization Signals.....	324
Global Positioning System	326
Disturbances in Synchronization Signals	327
Frequency Offset	327
Phase Fluctuation.....	328
Synchronization Models.....	333
Pointers and Timing Compensation.....	335
Pointer Formats and Procedures	335

Gigabit Ethernet Networks

Chapter 1

Gigabit Ethernet Networks

Gigabit Ethernet (GbE) is known to be a good and cost-effective technology for service providers seeking to roll out *Metropolitan* and *Campus Area Networks* (MAN/CAN). *10 Gigabit Ethernet* (10GbE), *40 Gigabit Ethernet* (40GbE) and *100 Gigabit Ethernet* (100GbE) were designed keeping MAN and WAN applications in mind (see Figure 1.1), but it can also be used for some bandwidth-consuming LAN and *Storage Area Network* (SAN) applications. High speed Ethernet opens up opportunities for new technologies such as MPLS-TP, and for applications like triple play.

	Name	Media Type	H/F	Coding	Line	MFS	Range
Ethernet IEEE 802.3a-1	10BASE-2	One 50 Ω thin coaxial cable	H	4B/5B	Manch.	64	185 m
	10BASE-5	One 50 Ω thick coaxial cable	H	4B/5B	Manch.	64	500 m
	10BROAD-36	One 75 Ω coaxial (CATV)	H	4B/5B	Manch.	64	3600 m
	10BASE-T	Two pairs of UTP 3	H/F	4B/5B	Manch.	64	100 m
	10BASE-FP	Two optical 62.5 μ m MMF passive hub	H/F	4B/5B	Manch.	64	1000 m
	10BASE-FL	Two optical 62.5 μ m MMF asyn hub	H/F	4B/5B	Manch.	64	2000 m
	10BASE-FB	Two optical 62.5 μ m MMF sync hub	H/F	4B/5B	Manch.	64	2000 m
Fast Ethernet IEEE 802.3u	100BASE-T4	Four pairs of UTP 3	H/F	8B/6T	MLT3	64	100 m
	100BASE-T2	Two pairs of UTP 3	H/F	PAM5x5	PAM5	64	100 m
	100BASE-TX	Two pairs of UTP 5	H/F	4B/5B	MLT3	64	100 m
	100BASE-TX	Two pairs of STP cables	H/F	4B/5B	MLT3	64	200 m
	100BASE-FX	Two optical 62.5 μ m MMF	H/F	4B/5B	NRZI	64	2 km
	100BASE-FX	Two optical 50 μ m SMF	H/F	4B/5B	NRZI	64	40 km
	Gigabit Ethernet IEEE 802.3z, 802.3ab	1000BASE-CX	Two twinax cables	H/F	8B/10B	NRZ	416
1000BASE-KX		PCB traces with two connectors 100 Ω	H/F	8B/10B	NRZ	416	1 m
1000BASE-T		Four pair UTP Cat. 5	H/F	8B1Q4	4D-PAM5	520	100 m
1000BASE-SX		Two 50 μ m MMF, 850 nm	H/F	8B/10B	NRZ	416	500/750 m
1000BASE-SX		Two 62.5 μ m MMF, 850 nm	H/F	8B/10B	NRZ	416	220/400 m
1000BASE-LX		Two 50 μ m MMF, 1310 nm	H/F	8B/10B	NRZ	416	550/2000 m
1000BASE-LX		Two 62.5 μ m MMF, 1310 nm	H/F	8B/10B	NRZ	416	550/1000 m
1000BASE-LX		Two 8 ~ 10 μ m SMF, 1310 nm	H/F	8B/10B	NRZ	416	5 km
1000BASE-ZX		Two 8 ~ 10 μ m SMF, 1550 nm	H/F	8B/10B	NRZ	416	80 km

Table 1.1

IEEE 802.3 Ethernet versions. List of acronyms: H/F: Half-Duplex and Full-Duplex ability. MFS: Minimum Frame Size in bytes. N/A: Not applicable. MMF: Multimode Fiber. SMF: Single Mode Fiber.

10GEthernet IEEE 802.3ae, 802.3ak, 802.3an, 802.3ap	10GBASE-CX4	Two twinax cables	F	8B/10B	NRZ	N/A	15 m
	10GBASE-KX4	PCB traces with two connectors 100 Ω	F	8B/10B	NRZ	N/A	1 m
	10GBASE-T	Four pair UTP Cat. 6a	F	64B/65B	DSQ 28	N/A	100 m
	10GBASE-SR	Two 50 μ m MMF, 850 nm	F	64B/66B	NRZ	N/A	2 ~ 300 m
	10GBASE-SW	Two 62.5 μ m MMF, 850 nm	F	64B/66B	NRZ	N/A	2 ~ 33 m
	10GBASE-LX4	Two 50 μ m MMF, ~1300 nm, 4 x WDM signal	F	8B/10B	NRZ	N/A	300 m
	10GBASE-LX4	Two 62.5 μ m MMF, ~1300 nm, 4 x WDM signal	F	8B/10B	NRZ	N/A	300 m
	10GBASE-LX4	Two 8 ~ 10 μ m SMF, ~1300 nm, 4 x WDM	F	8B/10B	NRZ	N/A	10 km
	10GBASE-LR	Two 8 ~ 10 μ m SMF, 1300 nm	F	64B/66B	NRZ	N/A	10 km
	10GBASE-LW	Two 8 ~ 10 μ m SMF, 1310 nm	F	64B/66B	NRZ	N/A	10 km
10GBASE-ER	Two 8 ~ 10 μ m SMF, 1550 nm	F	64B/66B	NRZ	N/A	2 ~ 40 km	
10GBASE-EW	Two 8 ~ 10 μ m SMF, 1550 nm	F	64B/66B	NRZ	N/A	2 ~ 40 km	
40G and 100GEthernet IEEE 802.3ba	40GBASE-CR4	4 + 4 twinax cables	F	64B/66B	NRZ	N/A	10 m
	40GBASE-KR4	PCB traces with two connectors 100 Ω	F	64B/66B	NRZ	N/A	1 m
	40GBASE-SR4	4 + 4 OM3 MMF, 850 nm	F	64B/66B	NRZ	N/A	100 m
	40GBASE-LR4	Two 8 ~ 10 μ m SMF, 1310 nm, 4 x WDM signal	F	64B/66B	NRZ	N/A	10 km
	100GBASE-CR10	10 + 10 twinax cables	F	64B/66B	NRZ	N/A	10 m
	100GBASE-SR10	10 + 10 OM3 MMF, 850 nm	F	64B/66B	NRZ	N/A	100 m
	100GBASE-LR4	Two 8 ~ 10 μ m SMF, 1310 nm, 10 x WDM	F	64B/66B	NRZ	N/A	10 km
100GBASE-ER4	Two 8 ~ 10 μ m SMF, 1310 nm, 10 x WDM	F	64B/66B	NRZ	N/A	40 km	

Table 1.1

IEEE 802.3 Ethernet versions. List of acronyms: H/F: Half-Duplex and Full-Duplex ability.
MFS: Minimum Frame Size in bytes. N/A: Not applicable. MMF: Multimode Fiber. SMF: Single Mode Fiber.

Prominence of Ethernet is today stronger than ever due to some recent important industry decisions:

- The ITU-T has defined the *Optical Transport Unit 4 (OTU4)* for the *Optical Transport Network (OTN)*. The OTU4 rate (112 Gb/s) is not a 4x multiplier of the lower speed interface (OTU3, 43 Gb/s). Instead, the OTU4 is prepared from the beginning for 100 Gb/s Ethernet.
- The transport profile for *Multi-Protocol Label Switching (MPLS-TP)* is a replacement of traditional *Time Division Multiplexing (TDM)* transport networks and specifically of the *Synchronous Digital Hierarchy (SDH)* and the *Synchronous Optical Network (Sonet)*. MPLS-TP is not directly Ethernet but can be used to supply or extend Ethernet services like *E-line*, *E-LAN* and *E-Tree*. High speed

Ethernet networks are perfectly suitable as the transport infrastructure required by MPLS-TP.

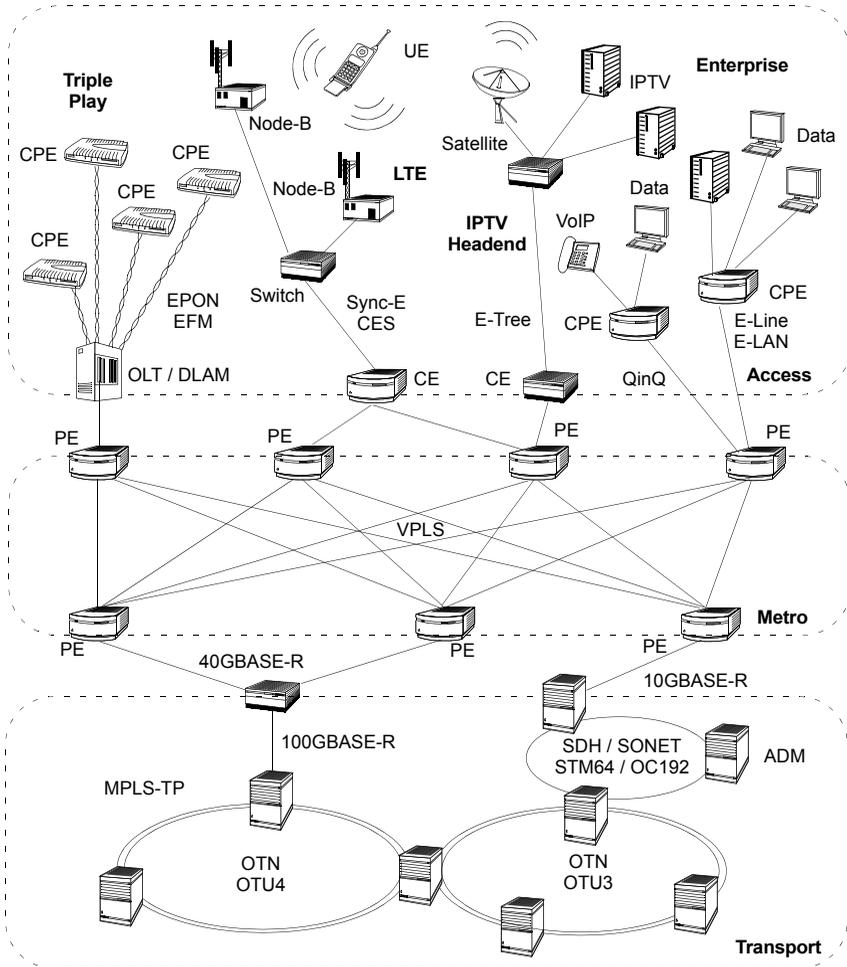


Figure 1.1 Today, Ethernet finds applications in the *Access*, *Metro* and *Transport* networks. It works smoothly with MPLS and IP or it can be used to deliver synchronization.

-
- The 3GPP *Long Term Evolution* (LTE) cellular communications standard defines a multiplay service network, the 'All-IP' network, suitable both for cellular and fixed services. The 'All-IP' network is prepared for delivery of any kind of service. With the LTE and the 'All-IP' network it is completed the migration of cellular network technology to packet-switching and Ethernet.
 - *Synchronous Ethernet* is a tremendously important initiative which aims to use the Ethernet physical layer for transmission of accurate synchronization. Good synchronization is a key requirement of many services and application, particularly for the mobile network. For first time, Ethernet offers a synchronization quality which is at least of the same level of TDM networks.

1 Gb/s Ethernet

The Gigabit Ethernet standards were first released in 1998. The IEEE 802.3 standardization resulted in two primary specifications:

- IEEE 802.3z (1000BASE-X) over optical fiber and STP cable.
- IEEE 802.3ab (1000BASE-T) over Category 5 UTP cable or better.

1 Gb/s Ethernet uses the same formats and protocols as its predecessors, which guarantees integration and smooth migration from earlier versions. For Gigabit Ethernet, the PHY and MAC layers were adapted for faster bit rates and new physical media.

1000BASE-X Architecture

In 1998, the IEEE approved a standard for 1 Gb/s Ethernet over fiber optic cable, IEEE 802.3z. The physical layer used was the ANSI X3.230 Fiber Channel, a technology devoted to high-speed data transfer used by mainframes and servers.

There are three different versions of 1000BASE-X: 1000BASE-CX, 1000BASE-SX and 1000BASE-LX (see Figure 1.2). The first one uses an STP cable, and the second and the third one use optical fiber.

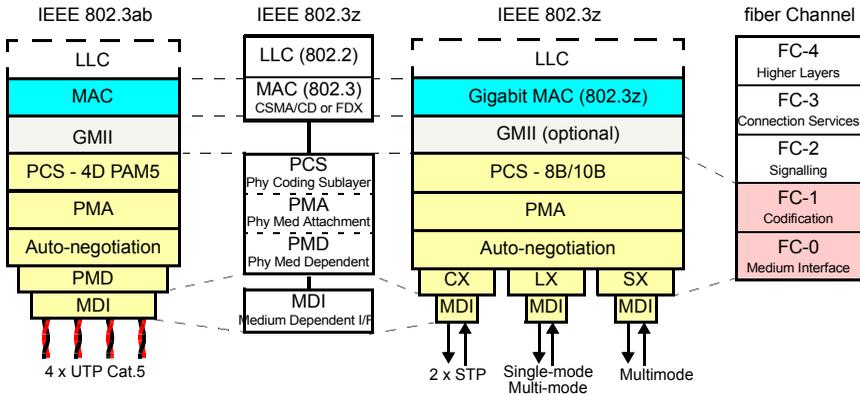


Figure 1.2 Gigabit Ethernet defines several transmission media, specified in the IEEE 802.3z (1000BASE-X) and 802.3ab (1000BASE-T). The first one is based on the existing fiber channel technology and covers three different types of media, and the second one uses the popular UTP cable.

- 1000BASE-CX, designed for short interconnections of network equipment in the wiring closet. This interface is based on copper, easier to handle than fiber. It uses a 150 Ω twinax cable similar to the original IBM Token Ring cabling.
- 1000BASE-SX, a cost-effective interface for short backbones or horizontal cabling. This PHY is based on inexpensive 850 nm photodiodes and MMF. The reach ranges from 220 to 750 m.
- 1000BASE-LX, targeted at longer backbones and vertical cabling. This interface is based on 1310 nm lasers, and runs over an SMF or an MMF. The reach of this PHY is 5000 m for SMF, and between 550 and 1000 m for MMF.

Some manufacturers include the 1000BASE-ZX interface in their equipment. This is a non-standard interface for Gigabit Ethernet that operates on 1550 nm lasers over SMF. This interface can reach

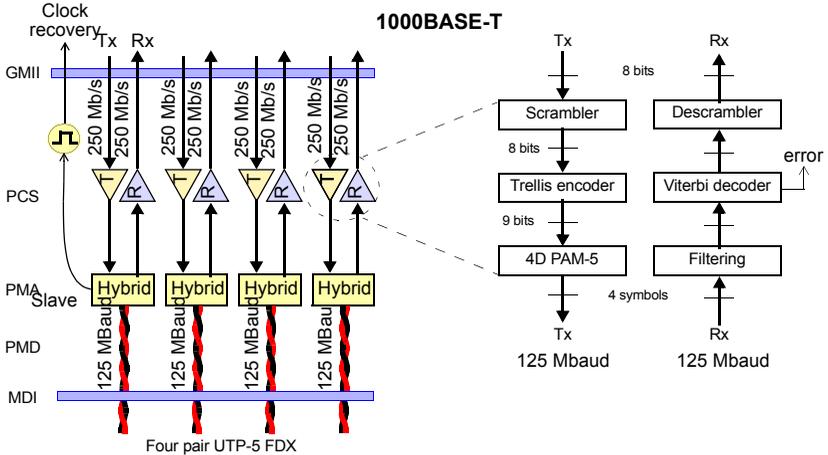


Figure 1.3 1000BASE-T transmits and receives signals simultaneously over the same pairs.

up to 80 km without repeaters, and it is well-suited for MAN and WAN applications.

The 1000BASE-X interface uses 8B/10B encoding followed by a simple *Non-Return to Zero* (NRZ) modulation. When data is ready to be transmitted, each 8-bit data byte is mapped into 10-bit symbols (8B/10B block-coding system) for serial transmission. Additional codes are included for control reasons. The channel rate of 1000BASE-X is 1250 Mb/s, and the data rate is 1000 Mb/s, due to the use of the 8B/10B encoding method.

The 8B/10B encoding method is the basis of the ANSI Fiber Channel standard for high-performance mass-storage devices, and it has properties such as excellent transition density – that is, a high number of transitions from the logic 1 to logic 0 state, which the PLL circuits require to recover the clock. It inherits excellent DC balance – there is no accumulation of DC offset that might cause the DC baseline to wander in the receiver. Furthermore, 8B/10B has

excellent error detection capabilities and provides reliable synchronization and clock recovery.

1000BASE-T Architecture

A twisted-pair version was introduced by the IEEE in 1999 under the name IEEE 802.3ab. The physical layer was specified as UTP Cat. 5 cabling to guarantee easy integration with existing 10BASE-T and 100BASE-T networks. 1000BASE-T over UTP is usually the preferred option for horizontal cabling and desktop connection. This is an alternative to 1000BASE-CX, which is rarely used in practice.

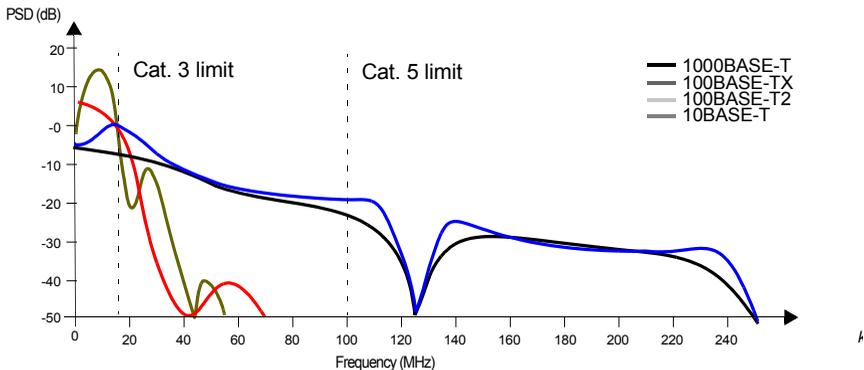


Figure 1.4 Power Spectrum Density (PSD) for 10/100/1000BASE-T electrical technologies.

1000BASE-T operates over Cat. 5 (or better) cabling systems by using all four pairs, sending and receiving a 250 Mb/s data stream over each of the four pairs ($4 \times 250 \text{ Mb/s} = 1 \text{ Gb/s}$) simultaneously (Figure 1.3). Hybrid circuits are used to enable bidirectional transmission and reception over the same pair. These circuits perform sophisticated *Digital Signal Processing* (DSP), filtering, and equalization of the received signal. They also perform echo canceling and remove crosstalk, to compensate for distortion from the UTP wiring.

The 1000BASE-T PHY uses an 8B1Q4 encoding followed by a 4D-PAM5 line modulation to achieve a 250 Mb/s throughput using baseband signaling at 125 MBaud. It achieves a half-duplex data rate of 1 Gb/s at a spectral power density similar to that of 100BASE-TX (Figure 1.4).

10 Gb/s Ethernet

In June 2002 was approved the IEEE 802.3ae standard for transmission of Ethernet at 10 Gb/s. This standard made Ethernet suitable for WAN applications for the first time. In fact, compatibility and convergence with current WAN technologies is one of the most attractive features of 10 GbE. After IEEE 802.3ae was published, other standards completed the original specification: New Ethernet interfaces operating at 10 Gb/s are defined in IEEE 802.3ak, 802.3an and 802.3aq. Some important features of the 10GbE standards are:

- Half duplex has been abandoned and only full-duplex operation is permitted. This makes CSMA/CD unnecessary.
- Copper transmission was discarded in IEEE 802.3ae. However, more recent IEEE standards define copper transmission interfaces for twinax cables, twisted pairs and backplanes operating at 10 Gb/s.
- The 64B/66B coding is widely used by the PCS sublayer of 10GbE interfaces. This code has similar functions as the 8B/10B code used in the 1 Gigabit Ethernet. It makes frame delineation, clock recovery and single or multiple error detection possible at the physical layer. The 64B/66B encoding makes the signaling rate slightly higher than 10 GBaud. Specifically, the signaling rate of 64B/66B signals is 10.3125 GBaud.

Despite all these new features, 10GbE still is Ethernet. This is why it interfaces so easily with lower rate Ethernet, and the deployment of 10GbE is less costly than that of other WAN technologies.

Optical Transmission

The interfaces defined in IEEE 802.3ae are 10GBASE-LX4, 10GBASE-S, 10GBASE-L and 10GBASE-E (see Table 1.2). There are two versions, R and W, of each of the last three. The W version is partially compatible with SDH/SONET interfaces, and thus it is specially suitable for a connection with WAN interfaces (see Figure 1.5).

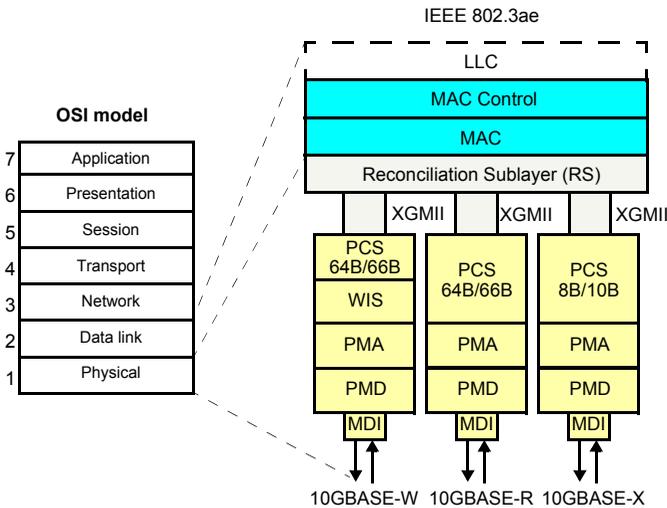


Figure 1.5 Layered model of IEEE 802.3ae 10 Gigabit Ethernet

The 10GbE standard defines short-haul (10GBASE-S), long-haul (10GBASE-L) and very long-haul interfaces (10GBASE-E). The most suitable for WAN applications is the 10GBASE-E. The maximum range of this interface is 40 km, but some manufacturers specify longer distances than the standard.

The 10GBASE-LX4 is another important new interface. It makes use of a low cost version of the *Wavelength Division Multiplexing* (WDM) technology called *Coarse WDM* (CWDM). The CWDM of the

Interface	Fiber	Wavelength	Modal bandwidth	Range
10GBASE-LX4	62.5 μm MMF	~ 1300 nm	500 MHz*km	2 ~ 300 m
	62.5 μm MMF	~ 1300 nm	400 MHz*km	2 ~ 240 m
	50 μm MMF	~ 1300 nm	500 MHz*km	2 ~ 300 m
	10 μm SMF	~ 1300 nm	-	2 ~ 10 km
10GBASE-S	62.5 μm MMF	850 nm	160 MHz*km	2 ~ 26 m
	62.5 μm MMF	850 nm	200 MHz*km	2 ~ 33 m
	50 μm MMF	850 nm	400 MHz*km	2 ~ 66 m
	50 μm MMF	850 nm	500 MHz*km	2 ~ 82 m
	50 μm MMF	850 nm	2000 MHz*km	2 ~ 300 m
10GBASE-L	8 ~ 10 μm SMF	1310 nm	-	2 ~ 10 km
10GBASE-E	8 ~ 10 μm SMF	1550 nm	-	2 ~ 40 km

Table 1.2 Range of Optical 10GbE interfaces.

10GBASE-LX4 interface is used to multiplex four wavelengths near the second optical transmission window (1310 nm). The 10GBASE-LX4 PMD does not use 64B/66B coding. It instead transports 8B/10B, NRZ coded signals operating at 3.125 GBaud in each wavelength.

Lane	Wavelength margin
#1	1269.0 ~ 1282.4 nm
#2	1293.0 ~ 1306.9 nm
#3	1318.0 ~ 1331.4 nm
#4	1342.5 ~ 1355.9 nm

Table 1.3 Wavelength specifications for the 10GBASE-LX4 interface

The maximum range of the 10BASE-LX4 is 300 m over 500 MHz * km, 50 μm MMF. This is not enough for MAN or WAN applications, but it can be very useful to upgrade installed MMF to 10 Gb/s.

10 Gb/s Ethernet over Copper

Despite the initial lack of support copper transmission media in IEEE 802.3ae, later standards like the IEEE 802.3ak, 802.3an and 802.3ap fixed this situation.

IEEE 802.3ak, released in 2004 defines the 10GBASE-CX4, that constitutes the electrical counterpart of the previous 10GBASE-LX4. Like the 10GBASE-LX4, the 10GBASE-CX4 uses 8B/10B coding and transmits information over four parallel lanes. However, the 10GBASE-CX4 operates on twinax cable assemblies of up to 15 m rather than on optical fibre. The 10GBASE-CX4 standard responds to the broad demand for high-speed interconnects within wiring closets, 10GBASE-CX4 also can be used in data centers to aggregate servers.

Perhaps the most relevant standard for copper communication applications at 10GbE is the IEEE 802.3an, released in 2006. This standard defines the 10GBASE-T interface for transmission of 10Gb/s over category 6a cables or better with range of 100 m.

The 10GBASE-T employs some mechanisms already available in 1000BASE-T. However, the 4D-PAM5 line modulation, successfully applied to GbE, does not provide the required performance level on the bandwidth supplied by cabling systems and the transmission rate of 10 Gb/s. Therefore IEEE developed a new coding and modulation scheme.

The 10GBASE-T PHY encodes the data from the upper layers with a 64B/65B code. The 64B/65B blocks are then placed in a Low Density Parity Check (LDPC) frame. The modulation uses an structure known as Double Square 128 (DSQ 128). The DSQ 128 symbols are generated by grouping symbols from 16-level PAM (PAM16) modulated signals in pairs. From the resulting 256 (16x16) point constellation, one half are removed in such a way that the distance from the original matrix is increased in a factor of $\sqrt{2}$ (see Figure 1.6). This provides a 3 dB improvement in the Signal-to-

Noise Ratio (SNR) which leads to a significant improvement of the Bit Error Ratio (BER).

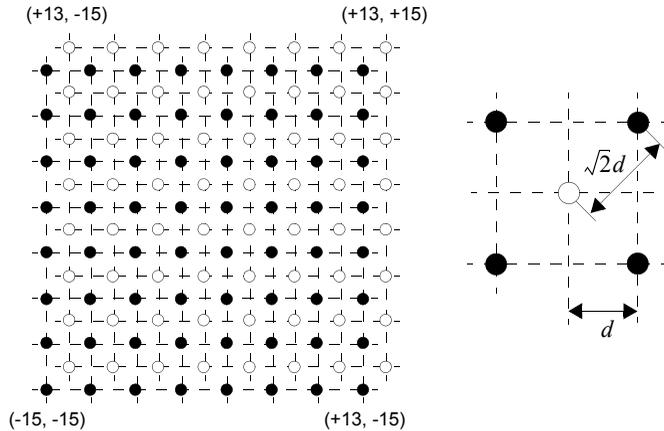


Figure 1.6 The DSQ 128 constellation is built with two PAM-16 symbols. Alternate points the resulting 256 symbol constellation are chosen such as in a chess board. The resulting constellation has 128 symbols that can be encoded by 7-bit words.

Each pair in the 10GBASE-T cable carries a PAM16 signal with a symbol rate of 800 MBaud which requires around 400 MHz of transmission bandwidth. This is not available in Cat. 5 and Cat. 5e or even in Cat. 6 cables used for 1GbE. For this reason, the new cable categories, starting with Cat. 6a have been qualified for transmission of 500 MHz bandwidths or more.

Standard IEEE 802.3ap, published in 2007, defines interfaces for Ethernet backplanes. The 10GBASE-KX4, based on the 10BASE-LX4 and 10GBASE-CX4, delivers 8B/10B encoded signals over four lanes with a signalling rate of 3.125 GBaud per lane. The 10GBASE-KR, similar to the 10GBASE-SR, 10GBASE-LR and 10GBASE-ER, transmits 64B/66B signals over a single serial transmission medium. The signalling rate in this case is 10.3125 GBaud. The IEEE 802.3ap

defines also one interface for backplanes operating at 1 Gb/s, the 1000BASE-KX.

These interfaces are defined to be used in blade servers and routers/switches with upgradable line cards. IEEE 802.3ap implementations are required to operate in an environment comprising up to 1 m of copper printed circuit board.

Compatibility with SDH/SONET

The 10GBASE-W interfaces included in the 10GbE standard can be connected directly to the STM-64/OC-192 ports of the SDH/SONET access equipment. So, an Ethernet switch can be connected to the SDH/SONET network without the need of any special adaptation device. The 10GBASE-W interface makes migration to an Ethernet/IP packet network easier, because it re-uses the existing optical WAN equipment.

Partial compatibility with SDH/SONET is achieved by means of:

- Rate compatibility with STM-64/OC-192. The 10GBASE-W signaling rate is 9.953280 GBaud, the same as for STM-64/OC-192
- Standard SDH/SONET framing and scrambling
- Support of a reduced set of SDH/SONET functions

Tasks such as framing and scrambling are left to a special sublayer of the physical layer called *WAN Interface Sublayer* (WIS). The WIS takes the continuous bit stream from the PCS sublayer, mapping it into an SDH/SONET concatenated container with a self-generated *Path Overhead* (POH). Then it builds an STM-64/OC-192 frame with a fixed pointer value (522 with concatenation indication), an internally generated *Multiplexer Section Overhead* (MSOH) and a *Regenerator Section Overhead* (RSOH). Finally, it scrambles the SDH/

SONET frame and sends the resulting bit stream to the lower sublayer (see Figure 1.7).

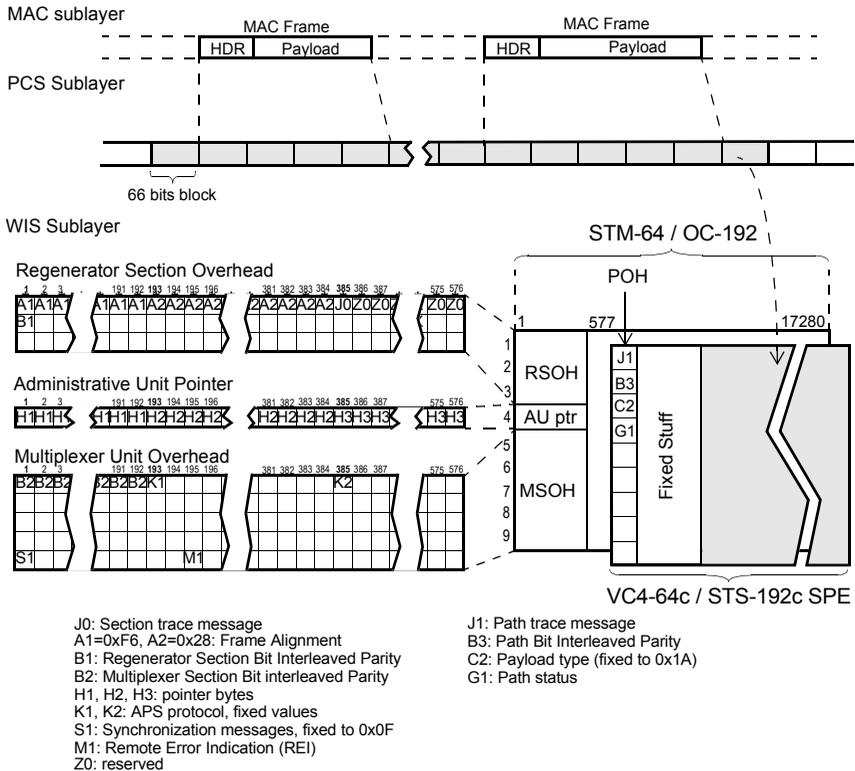


Figure 1.7 Mapping of 64B/66B data into a simplified STM-64 / OC-192 frame performed by the WIS in the 10GBASE-W interfaces

The available bandwidth for mapping the 64B/66B-coded Ethernet signal corresponds to the capacity of the VC4-64c / STS-192c container, 9.58464 Gb/s. This number is different from the bit rate

generated by the 10GBASE-R interface (10.3125 Gb/s). However, direct switching between interfaces with WIS and without WIS is theoretically possible, the same way as Ethernet at 1 Gb/s can be switched to 100 Mb/s Ethernet.

Not every function defined for the SDH/SONET equipment is supported by the WIS. Many of the functions of the overhead bytes are either partially supported or not supported at all. *Bit Interleaved Parity* (BIP) code generation and analysis, a minimum set of SDH/SONET alarms and path trace messages are supported. On the other hand, SDH synchronization is not supported. This means that it is not necessary to synchronize the Ethernet equipment with a central clock in the same way that is done with SDH/SONET devices. Therefore, Ethernet equipment continues being asynchronous. Low-rate container multiplexing and protection switching are not supported either. The transmitter does not need a pointer processor, but it is needed at the receiver end, because pointer adjustments may occur within the SDH/SONET network, and the receiver still needs to be able to demap the tributary signal.

The Physical Media Dependent (PMD) layer for the 10GBASE-W interfaces is the same as for the 10GBASE-R interfaces. The SDH/SONET specification is not followed at this level. The objective is to be able to manufacture inexpensive Ethernet equipment competitive with other WAN technologies.

Higher Speed Ethernet

In 2006, the IEEE 802.3 working group formed the Higher Speed Study Group (HSSG). The result of the HSSG efforts is the standard IEEE 802.3ba, ratified on June of 2010. The HSSG determined that two new rates were needed: 40 Gb/s for server and computing applications and 100 Gb/s mainly for service provider networks.

Like all the lower rate Ethernet standards, the new interfaces preserve the MAC frame format defined in all other IEEE Ethernet

standards. The Physical Coding Sublayer (PCS) adopted in all the new interfaces include 64B/66B coding already used in the 10 Gb/s Ethernet PCS. The IEEE 802.3ba standard also defines new PMD sublayers for SMF, MMF and copper media.

An important new feature of 40Gb/s and 100Gb/s Ethernet is that, unlike all other Ethernet versions (or at least most of them), Higher Speed Ethernet has been designed for parallel data transmission. That means that, the 40 Gb/s or 100 Gb/s data flow is split over parallel physical paths or WDM wavelengths.

Multilane Distribution Procedure

All current PMDs for 100 Gb/s and 40 Gb/s are based on parallel transmission. Many different configurations are supported and future development of optical technologies are considered by the standard. Furthermore, the protocol stack is defined in such a way that new media and data distribution across paths in the PMD does not require to redefine the upper level protocol layers. The mechanism defined in IEEE 802.3ba to ensure flexible data distribution over the physical transmission media is the Multilane Distribution (MLD) procedure.

The MLD scheme implemented in the PCS is fundamentally based on a striping / regrouping the 66-bit blocks resulting from the 64B/66B encoding following a round robin algorithm. Each of the groups resulting from the MLD striping is called a lane. All bits from a lane are always transmitted over the same physical transmission medium. Periodically, unique alignment blocks are added to each lane to allow deskew in the receiver.

With the help of the MLD methodology, the encoding, scrambling and deskew functions can all be implemented in a CMOS circuit, which is expected to reside on the host device. Minimal processing

of the data bits happens in the high speed electronics embedded with an optical module.

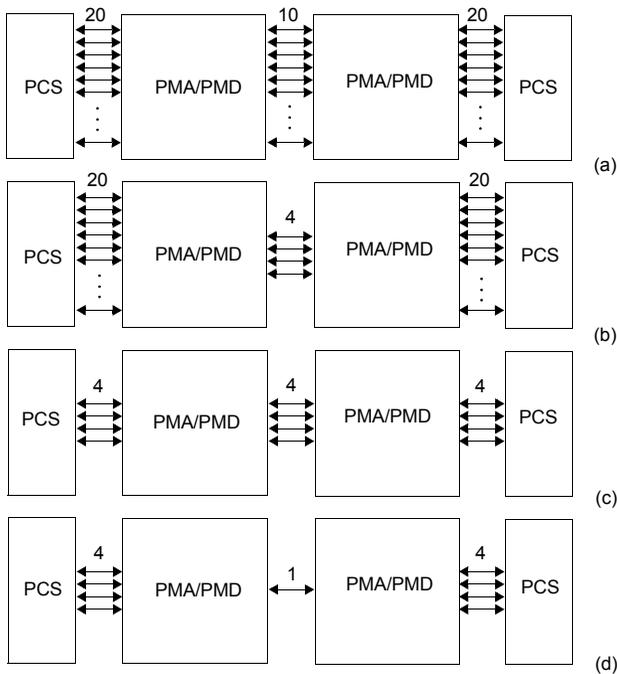


Figure 1.8 Some lane configurations: (a) Configuration for 100 Gb/s Ethernet operating over MMF or copper. (b) Configuration for 100 Gb/s Ethernet over SMF. (c) Configuration for all 40 Gb/s currently defined interfaces. (d) 40G/bs over a single optical interface (configuration not included in IEEE 802.3ba).

The PCS generates a fixed number of lanes (for a fixed rate, either 100 Gb/s or 40 Gb/s) that is multiplexed before transmission by the PMA sublayer to the number required by the PMD. For 100 Gigabit Ethernet, the PCS generates 20 lanes. The number of electrical or optical interface widths supportable in this architecture is equivalent to the number of factors of the total PCS lanes. Therefore, 20 PCS lanes support interface widths of 1, 2, 4, 5, 10 and

20 channels or wavelengths. For 40 Gigabit Ethernet 4 PCS lanes support interface widths of 1, 2, and 4 channels or wavelengths. However, not every possible configuration is found in the interfaces defined by the IEEE. Currently, there interfaces with 10 and 4 physical lanes (see Figure 1.8).

Physical Media

IEEE 802.3ba extends the Ethernet rate over backplanes and copper cable assemblies to 40 Gb/s and 100 Gb/s. The 40GBASE-KR4 extends to 40 Gb/s the rate in backplanes. This interface employs four lanes of 64B/66B encoded signals. Like all other backplane Ethernet interfaces, the reach of the 40GBASE-KR4 is 1 m. Currently, there is not a backplane Ethernet interface operating at 100 Gb/s.

40 Gb/s Interface	100 Gb/s Interface	Physical Media	Range
40GBASE-KR4	-	SMB traces	1 m
40GBASE-CR4	100GBASE-CR10	Twinax cable	10 m
40GBASE-SR4	100GBASE-SR10	OC3 MMF	100 m
40GBASE-LR4	100GBASE-LR4	SMF	10 km
-	100GBASE-ER4	SMF	40 km

Table 1.4 40 Gb/s and 100 Gb/s Ethernet interfaces summary

The 40GBASE-CR4 and 100GBASE-CR10 are designed to operate in the same environment that the 10GBASE-CX4. These interfaces use four or ten twinax cable assemblies of up to 10 m in each direction as a transmission medium. The 40GBASE-CR4 and 100GBASE-CR10 include the same 64B/66B PCS than all other 40 Gb/s and 100 Gb/s interfaces. The result is that the signalling rate for them is 10.3125 GBaud. This is also true for other 40 Gb/s and 100 Gb/s interfaces based on 10 Gb/s physical lanes, including the ones for backplanes.

Regarding the optical interfaces, there are two kinds of them. The 40GBASE-SR4 and 100GBASE-SR10 operate over MMF in short haul links. On the other hand, 40GBASE-LR4, 100GBASE-LR4 and

100GBASE-ER4 are designed for SMF transmission with the help of WDM technology. The 40GBASE-SR4 and 100GBASE-SR10 are more suited for data center applications due to the limited reach and the dependency on cheaper MMF and low cost optical transceivers. The 40GBASE-LR4, 100GBASE-LR4 and 100GBASE-ER4 could be used for service provider network aggregation (see Table 1.4).

The short haul interfaces employ four or ten OC3 MMF in each transmission direction, depending on the nominal bit rate. These interfaces operate in the 850 nm window.

Lane	Center wavelengths	Wavelength ranges
#1	1271 nm	1264.5 ~ 1277.5 nm
#2	1291 nm	1284.5 ~ 1297.5 nm
#3	1311 nm	1304.5 ~ 1317.5 nm
#4	1331 nm	1324.5 ~ 1337.5 nm

Table 1.5 Wavelength specifications employed in 40 Gb/s WDM interfaces

The long haul interfaces have four physical lanes and they use WDM. That means that they require only one SMF per transmission direction. In the development of the 100GBASE-ER4 specification the receiver was assumed to include optical amplification such as a Semiconductor Optical Amplifier (SOA) to compensate for the optical loss budget of 40 km of single mode fiber

Lane	Center frequencies	Center wavelengths	Wavelength ranges
#1	231.4 THz	1295.56 nm	1294.53 ~ 1296.59 nm
#2	230.6 THz	1300.05 nm	1299.02 ~ 1301.09 nm
#3	229.8 THz	1304.58 nm	1303.54 ~ 1305.63 nm
#4	229.0 THz	1309.14 nm	1308.09 ~ 1310.19 nm

Table 1.6 Wavelength specifications employed in 100 Gb/s WDM interfaces

The WDM interfaces use either a 10.3125 GBaud signalling rate (40GBASE-LR4) or a 25.78125 GBaud (100GBASE-LR4 and 100GBASE-ER4). The 40 Gb/s interfaces use the wavelength grid

defined in ITU-T G.694.2 for Coarse WDM (CWDM) (see Table 1.5) while the 100 Gb/s uses the so called LAN WDM grid which is based on the ITU-T G.694.1 for Dense WDM (DWDM) (see Table 1.6).

All 40 Gb/s and 100Gb/s interfaces require many transmitters and receivers to work over parallel physical lanes. To allow high density connections in network devices it is of paramount importance to develop new electrical and optical transceiver modules enabled for multilane transmission. The most important of these are the QSFP for 40 Gb/s applications, the CXP specialised in 100 Gb/s MMF applications and CFP, which supports all 40 Gb/s and 100 Gb/s interfaces defined in IEEE 802.3ba.

Finally, as higher speed electrical and optical components are developed it is expected that Ethernet will continue its development to become cheaper and more flexible. It is likely that the next steps will be the incorporation of IEEE 802.3az capabilities to reduce energy consumption of network equipment, development of 25 Gb/s signalling for 100 GbE over backplanes and copper cable applications, and definition of serial signalling specifications for 40 Gb/s and 100 Gb/s, with the later probably using some kind of phase modulation at the optical layer like the Differential Quadrature Phase Shift Keying (DQPSK).

40 Gb/s and 100 Gb/s Ethernet over OTN

Transport of 10 Gb/s over the *Optical Transport Network* (OTN) requires either the definition of new *Optical Data Units* (ODUs) and *Optical Transport Units* (OTUs) or the introduction of semi-transparent mappings for the existing OTN rates. The reason is that a 64B/66B encoded 10GBASE-R signal (10.3125 Gb/s) does not fit in a ODU2 (capacity of 9.99528 Gb/s). However, semi-transparent mapping do not provide enough transparency for critical applications and the introduction of the overclocked OTU2e and OTU3e adds complexity to the whole OTN hierarchy.

To accommodate the 100 Gb/s Ethernet interfaces, the OTN defines the Optical Transport Unit 4 (OTU4) operating at 112 Gb/s (see Figure 1.9). This OTN rate is specifically defined for 100 Gb/s Ethernet (the ODU4 capacity is 104.356 Gb/s and the 100 Gb/s Ethernet line rate is 103.125 Gb/s). The OTU4 makes the interconnection of Ethernet and OTN at 100 Gb/s very straightforward.

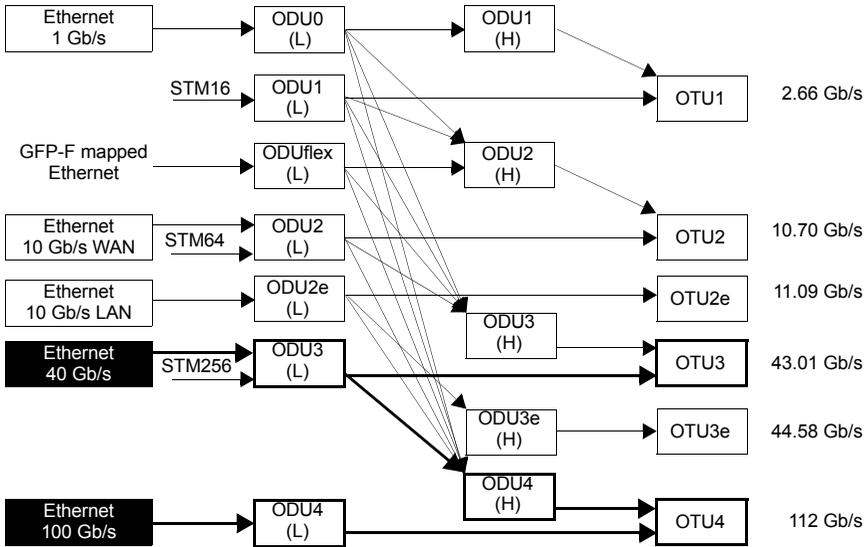


Figure 1.9 Transport of Gigabit Ethernet signals over OTN. The 40 Gb/s and 100 Gb/s are carried over the OTU3 (43 Gb/s) and OTU4 (112 Gb/s)

At first sight, the situation with the 40 Gb/s Ethernet looks similar that for the 10 Gb/s interfaces. The line rate of 40 Gb/s Ethernet is 41.25 Gb/s but the capacity provided by the ODU3 is only 40.15 Gb/s. However, a simple transcoding procedure reduces the Ethernet line rate to 40.12 Gb/s. This signal is suitable for mapping into the

ODU3. No new OTN rates and interfaces are therefore required. Transcoding of 64B/66B encoded 40 Gb/s Ethernet signals to the more efficient 512B/513B code is described in ITU-T G.709 Annex B.

Selected Bibliography

- [1] IEEE 802.3-2008, "Part 3: Carrier sense multiple access with collision detection (CSMA/CD) Access Method and Physical Layer Specifications", December 2008.
 - [2] ITU-T Rec. G.709/Y.1331, "Interfaces for the Optical Transport Network (OTN)", December 2009.
 - [3] Rich Seifert, *Gigabit Ethernet Technology and Applications for High/Speed LANs*, Addison Wesley Oct 1999
 - [4] Kevin L. Paton, *Gigabit Ethernet Test Challenges*, Oct 2001 Test and Measurement World Magazine.
 - [5] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.
 - [6] Millán R., Esfandiari S., "40 y 100 Gigabit Ethernet," BIT Magazine, no. 179, February-March 2010, pp. 46-48.
 - [7] Cole C., Huebner B., Johnson J., "Photonic Integration for High-Volume, Low-Cost Applications," *IEEE Communications Magazine*, March 2009.
-



Switched Ethernet

Ethernet has evolved dramatically since the first card was shipped demonstrating a great ability to answer to new challenges as well as to the growing business needs.

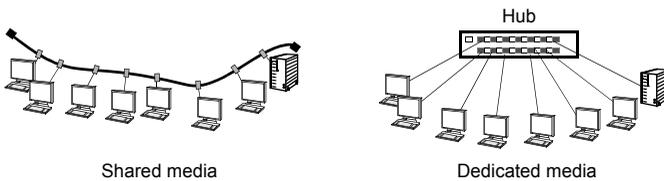


Figure 2.1 First implementations of Ethernet used shared transmission media that evolved to dedicated media when hubs were introduced

Ethernet was originally defined as a shared-media technology, where all the stations had to compete to get access to the common transmission medium. However, this limitation no longer exists, as new versions have been developed where stations do not need to compete for resources such as media or transmission capabilities.

From Shared to Dedicated Media

Sharing media means to share not only bandwidth but also problems. A simple discontinuity in a 10BASE-2 or 10BASE-5 cable could mean that all the attached devices are unavailable. 10BASE-T addressed this problem by dedicating a cable to each station and connecting all of them to a hub (Figure 2.1). The first hubs were only central points of the cabling system, but soon enough intelligence was added to detect anomalies and to disconnect faulty stations.

One of the fundamental issues of CSMA/CD within a shared-bandwidth medium is that the more traffic there is on the network, the more collisions there will be. That is, when the use of the network increases, the number of collisions increases as well, and the network could become unmanageable or even collapse (Figure 2.2). To solve this problem, Ethernet switches and bridges were introduced. These devices forward MAC frames by means of MAC addresses, and make better use of transmission resources than hubs based on frame broadcasting.

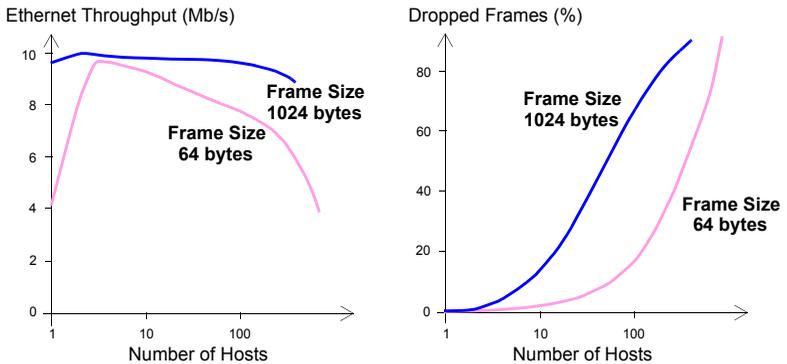


Figure 2.2 Shared Ethernet collapses as the number of hosts increases. Optimum performance is obtained for three or four stations.

To cut down on the number of collisions, the first step is to use segmentation by means of bridges. Switches subdivide the network into multiple collision domains. This reduces the number of stations competing for the same resource. The second step is to dedicate one segment to those stations that have high bandwidth requirements. The final step is to configure a network that is totally switched (Figure 2.3). In this type of networks, each station has its own collision domain. Collisions are thus impossible, and each station gets to use the whole bandwidth (so, it is not shared). Those

Ethernet networks where each station has its own collision domain are called micro-segmented networks.

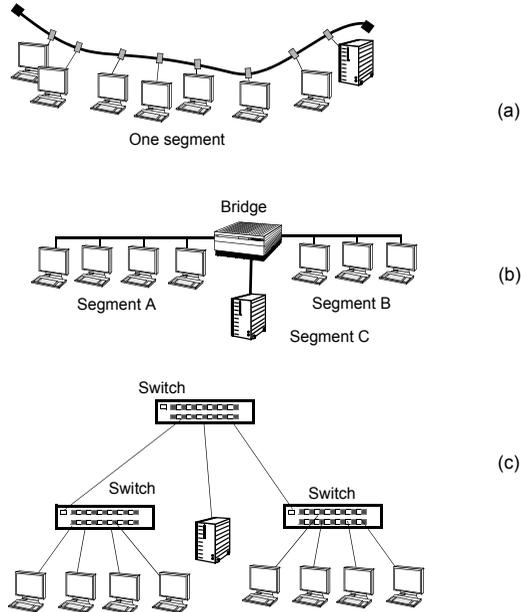


Figure 2.3 Segmentation and switching. (a) Shared Ethernet with bus topology. (b) Segmented Ethernet. (c) Microsegmented Ethernet.

Ethernet Bridging

Bridging is a forwarding technique for packet-switched networks that makes no assumptions about where in a network a particular station is located. Instead, broadcasting is used to locate unknown devices. Once a device has been located, its location is recorded in a switching table to preclude the need for further broadcasting. Ethernet switches and bridges use bridging as specified in standard IEEE 802.1D.

Both switches and bridges contain a table that is used to switch Ethernet frames to the right output interface. These tables store pairs of destination addresses and associated output interfaces. When a frame enters the switch in a specific interface, the destination MAC address is checked:

- 1.If the address is found in the switching table, the frame is delivered to the associated output interface.
- 2.If the address is not found, the frame is broadcast to all the output interfaces except the incoming one.

The source MAC address is also checked for every incoming frame:

- 1.If the address is not found in the switching table, it is stored in the table and associated to the incoming interface. This prevents broadcasting when the same address is found in the destination field of other frames.
- 2.If the address is found in the switching table, no action is needed.

An Ethernet bridge can be considered an Ethernet switch with only two ports. Historically, bridges appeared before switches.

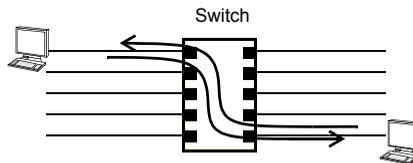


Figure 2.4 Segmentation reduces the probability of collision; full-duplex and switching removes it completely.

In micro-segmented networks, each station is connected to a switch port. Switching eliminates the possibility of collisions, making CSMA/CD unnecessary (see Figure 2.4). More precisely:

- The carrier-sense protocol is not needed, because the media is never busy, as there is a link dedicated for each transmitter/receiver couple.
- The collision detection protocol is not needed either, because collisions never happen, and no jamming signals are needed.

Full-Duplex Operation

To guarantee access to the media, it is important that simultaneous transmission from and reception by the same station occurs without any interference. The classic *half-duplex* (HDX) operation of Ethernet can be replaced by *full-duplex* (FDX) operation (see Figure 2.5). A station connected to a 100BASE-T interface can transmit at 100 Mb/s bit rate and receive data simultaneously at 100 Mb/s

Furthermore, in switched Ethernet networks, distance limitations are removed. Note that in shared networks, distance and frame size were restricted to allow stations to detect collisions while transmitting. In FDX systems the distance between stations

depends on the characteristics of the media and the quality of the transmitters; predefined limits do not apply.

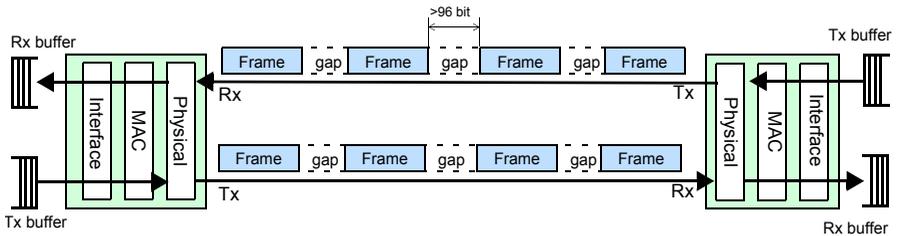


Figure 2.5 FDX operation enables two-way transmission simultaneously without contention, collisions, extension bits or retransmissions. The only restriction is that a gap must be allowed between two consecutive frames. FDX also requests flow control, which is transmitted by the receiver to request that the transmitter temporarily stops transmitting.

One side-effect of full-duplex occurs when a transmitter that is constantly sending packets may cause the receiver buffer to overflow. To avoid this, a *pause protocol* was defined. It is a

mechanism whereby a congested receiver can ask the transmitter to stop transmission.

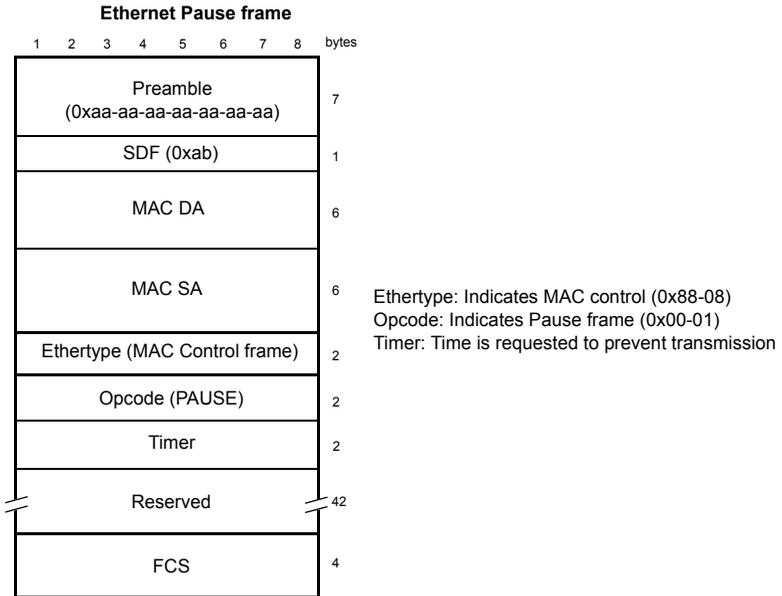


Figure 2.6 Pause frame, used for the flow control protocol. The unit of pause time equals to 512 bits. If pause time is 0, transmission should be stopped.

This protocol is based on a short packet known as a Pause frame (see Figure 2.6). The pause frame contains a timer value, expressed as a multiple of 512-bit times; this specifies for how long the transmitter should remain silent. If the receiver becomes uncongested before this time has passed, it may send a second pause frame with a value of zero to resume the transmission. The pause protocol operates only on point-to-point links and cannot be forwarded through bridges, switches or routers.

Gigabit Ethernet introduces the concept of Asymmetric Flow Control (AFC), which lets a device indicate that it may send pause frames, but declines to respond to them. If the link partner is willing to co-operate, pause frames will flow in only one direction on the link.

There are full-duplex operation modes for all the important interfaces that operate at 10, 100 and 1000 Mb/s. Gigabit Ethernet can run in either half-duplex or full-duplex mode. While this is true in theory, nearly all the demand for GbE is for full duplex. In spite of this, it was necessary to increase the slot time to 512 bytes, to make sure that CSMA/CD works correctly. However, if GbE is only used in full-duplex mode, CSMA/CD can effectively be removed.

CSMA/CD would impose too restrictive operation to 10GbE and higher bit rate devices. Therefore there is not half-duplex operation for them. They always run with full-duplex mode.

Hands-on: Performance of Ethernet Switches with RFC 2544

The RFC 2544 is an IETF standard that describes benchmarking tests for network devices. Vendors can use these tests to measure and outline the performance characteristics of their Ethernet switching equipment. As these tests follow standard procedures, they also make it easier for customers to make sense of the glitzy marketing-speak employed by most vendors.

The tests described in the document aim to evaluate how a device would act in a real situation. The RFC 2544 describes six out-of-service tests, which means that real traffic must be stopped and the tester will generate specific frames to evaluate throughput, latency, frame loss rate, burst tolerance, overload conditions recovery and reset recovery. The document also describes specific formats for reporting the results of these tests.

Test Conditions

All Ethernet tests should be run consistently without changing the configuration of the device, and without running a specific protocol or feature. The DUT should include the normal routing update intervals and keep frequency alive.

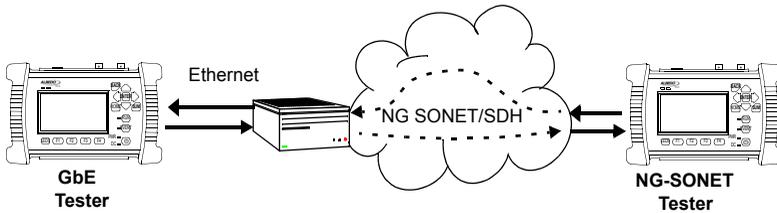


Figure 2.7 WAN performance may be tested by setting up two identical devices connected by the appropriate short-haul versions of the WAN modems. Performance is then measured between a LAN interface on one DUT, and a LAN interface on the other DUT.

Which Tester to Use

A tester with both transmitting and receiving ports is recommended for these tests. The tester must include sequence numbers in the frames it transmits, so that it can check that all frames transmitted are also received back.

The RFC 2544 can be used to test Layer 2 and Layer 3 devices. For Layer 3 testing, IP packets need to be configured, including parameters such as mask and subnetworks that can be understood by routers. MAC frames must always be programmed, including parameters such as frame size, bit rate, or traffic profile.

Traffic Used in the Test

- *Traffic pattern* – The traffic on a real network is not constant, but occurs in bursts. The RFC 2544 suggests that the tests should be carried out using constant traffic and with test conditions traffic, i.e., repeated bursts of frames, the frames within the bursts separated by the minimum inter-frame gap.

-
- *Protocol addresses* – The simplest way to perform these tests is to use a single stream of data. Networks in the real world do not have just one stream of data. The RFC 2544 suggests that after the tests have been run in this way, they should be re-run using a random destination address. For routers the RFC 2544 suggests that the addresses used should be random, and evenly distributed over a range of 256 networks. For bridges the range should be uniformly distributed over the full MAC range.
 - *Maximum frame rate* – When testing on a LAN, the maximum frame rate for the medium and frame size being used should be used for the test. When testing on a WAN, a rate greater than the maximum theoretical rate for the medium and frame size should be used.
 - *Frame sizes* – The RFC 2544 recommends that the tests are carried out at a range of frame sizes - 64, 128, 256, 512, 1024, 1280, 1518 bits. This covers the range of frame sizes that are typically transmitted.
 - *Frame formats* – The format of the frames of TCP/IP over Ethernet are specified in appendix C of the RFC.

Test Duration

These tests are designed to measure how a device will perform under continuous operation. The test time must be a compromise between this and the time available to complete a test suite. The RFC recommends that the duration of each trial should be at least 60 seconds.

RFC 2544 was designed for laboratory testing of equipment, which is why the tests as described may take several days to complete. This duration is unlikely to be possible or necessary when testing a network in the field. The time taken for the test can be reduced by selecting the tests to be run, and by reducing the number of repetitions.

Test Setup

The aim of this set of tests is to evaluate the performance of equipment in real-world situations. The RFC 2544 states that all the protocols supported by the device must be enabled when testing, and the equipment must be set up according to the instructions supplied to the user. The only changes allowed between tests are those needed to perform the different tests. It is not acceptable, for example, to change the size of the frame-handling buffer between tests of frame-handling rates.

Regarding the test reports, the RFC recommends that, in addition to the results, the following should be included in test reports:

- DUT setup: which functions are disabled, which ones used
- DUT software version
- Frame formats
- Filter setups

Running the Trial

The RFC defines a test as being made up of multiple trials. Each trial gives a piece of data, for example the loss rate at a particular input frame rate. The following procedure describes the steps for a single trial:

1. If the device you are testing is a router, send the routing update to the input port and wait two seconds.
 2. Send the trial frames to the output port.
 3. Run the trial.
 4. Wait for two seconds to receive all the data back.
 5. Wait at least five seconds before starting the next trial.
-

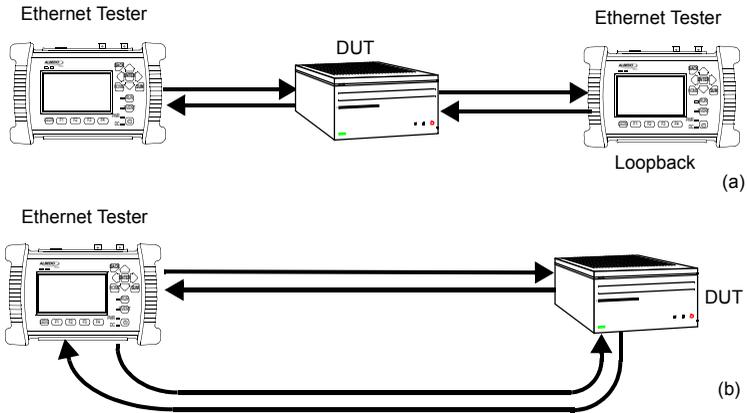


Figure 2.8 RFC 2455 performance evaluation of a multiport Ethernet: (a) Test is carried out with the help of two testers. (b) If the tester has two testing ports, all testing can be finished with a single test

Throughput

The aim of a *throughput test* is to determine the maximum number of frames per second that the device can process and forward without dropping or losing any. The procedure is as follows:

1. Send a certain number of frames at a specific rate through the DUT and count the frames transmitted by the DUT.
2. If the count of transmitted frames is equal to the count of received frames, increase the throughput and re-run the test.
3. Re-run the test until fewer frames are transmitted than received by the DUT.

The throughput is the fastest rate at which the count of test frames transmitted by the DUT is equal to the number of test frames sent to it by the test equipment.

Results must be expressed in frames per second, or alternatively in bits or bytes per second.

The statement of performance must include the following information:

- The measured maximum frame rate
- The size of the frame used
- The theoretical limit of the media for that frame size
- The type of protocol used in the test.

Latency

This test determines the latency inherent in the DUT. The initial data rate is based on the results of the throughput test. Typically, packets are time stamped, and the time they take to travel through the DUT is measured.

You must first measure the throughput for the DUT at each of the defined frame sizes, and send a stream of frames through the DUT at the determined throughput rate to a specific destination. The duration of the stream should be at least 120 seconds. After 60 seconds, an identifying tag should be included in one frame.

The time at which this frame is completely transmitted is recorded, and this will be timestamp A. The receiver of the test equipment must recognize the tag information in the frame stream and record the reception time of the tagged frame. This will be timestamp B.

The latency is the difference between timestamp B and timestamp A, according to the definition found in RFC 1242.

The test report must state which definition of latency, from RFC 1242, was used for this test. The latency results should be reported as a table, with a row for each tested frame size. There should be columns for the frame size, the rate at which the latency

test was run for that frame size, for the media types tested, and for the latency values for each type of data stream tested.

Frame Loss Ratio

The aim of this test is to determine the frame loss ratio throughout the entire range of input data rates and frame sizes. The procedure is the following:

1. Send a certain number of frames at a specific rate through the DUT, counting the frames transmitted.
The first trial should be run for the frame rate that is 100% of the maximum rate for the frame size on the input media.
The frame loss rate at each point is calculated as follows:
$$((input_count - output_count) * 100) / input_count$$
2. Repeat the procedure for the rate that corresponds to 90% of the maximum rate used, and then for 80% of this rate.
3. Continue this sequence (at reducing 10% intervals) until there are two consecutive trials where no frames are lost. The maximum granularity of the trials may be 10% of the maximum rate, although you may want to define a finer granularity.

Back-to-Back Frames

A back-to-back frame test determines the *node buffer capacity* by sending bursts of traffic at the highest theoretical rate, and then measuring the longest burst where no packets are dropped. This is done to check the speed at which a DUT recovers from an overload condition, and the procedure is as follows:

1. Send a burst of frames with minimum inter-frame gaps to the DUT, and count the number of frames forwarded.
 2. If the count of transmitted frames is equal to the number of frames forwarded, increase the length of the burst and re-run the
-

test. If the number of forwarded frames is less than the number transmitted, reduce the length of the burst and re-run the test.

The back-to-back value is the number of frames in the longest burst that the DUT can handle without losing any frames. The trial length must be at least 2 seconds, and it should be repeated at least 50 times with the average of the recorded values being reported.

The back-to-back results should be reported as a table, with a row for each of the tested frame sizes. There should be columns for the frame size and for the resultant average frame count for each type of data stream tested. The standard deviation for each measurement may also be reported.

System Recovery

This test determines the node speed at which a DUT recovers from an overload condition. The procedure is as follows:

1. Measure the throughput for a DUT at each of the listed frame sizes.
2. Send a stream of frames at a rate that is 110% of the recorded throughput rate or the maximum rate for the media, whichever is lower, for at least 60 seconds.
3. At Timestamp A, reduce the frame rate to 50% of the above rate and record the time of the last frame lost (Timestamp B). The system recovery time is calculated by subtracting Timestamp B from Timestamp A. The test must be repeated a number of times, and the average of the recorded values is reported.

The system recovery results should be reported as a table, with a row for each of the tested frame sizes. There should be columns for the frame size, the frame rate used as the throughput rate for each type of data stream tested, and for the measured recovery time for each type of data stream tested.

Reset

This test is intended to characterize the speed at which a DUT recovers from a device or software reset. The procedure is as follows:

1. Measure the throughput for the DUT for the minimum frame size on the media used in the testing.
2. Send a continuous stream of frames at the determined throughput rate for the minimum-sized frames.
3. Reset the DUT.
4. Monitor the output until frames begin to be forwarded, and record the time that the last frame (Timestamp A) of the initial stream and the first frame of the new stream (Timestamp B) are received.

A power-interruption reset test is performed as described above, except that instead of resetting, the power to the DUT should be interrupted for 10 seconds.

This test should only be run using frames addressed to networks directly connected to the DUT, so that there is no requirement to delay until a routing update is received. The reset value is calculated by subtracting Timestamp A from Timestamp B. Hardware and software resets, as well as a power interruption should be tested.

Virtual LANs

Virtual LAN (VLAN) is a local-area network that is logically segmented on an organizational basis, by functions, project teams or applications, rather than on a physical or a geographical basis. The network can be reconfigured through software, instead of physically unplugging and moving devices or wires. Stations are connected by switches and routers (see Figure 2.9). VLANs are an

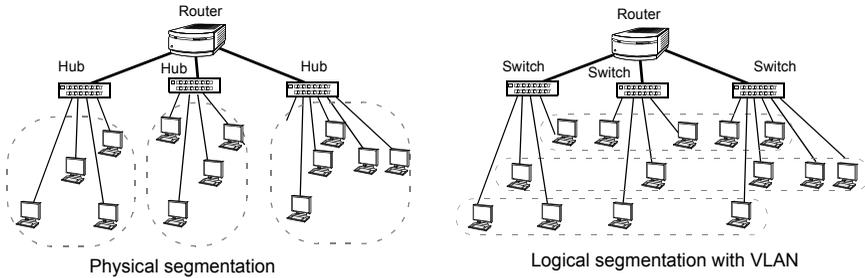


Figure 2.9 Virtual LAN vs. Segmented LAN.

important contribution to scalable Ethernet networks, because they limit broadcast traffic inherent to the bridging mechanism. Large amounts of broadcast traffic may damage performance and even collapse network equipment, which is why it must be controlled (see Figure 2.10).

Every virtual switch remains isolated and can only be communicated to other virtual switch by a layer-3 device. Ports from different physical switches can be attached to the same VLAN, and distant stations separated by thousands of kilometers could be part of the same virtual segment. The switch knows how to process traffic from different VLANs, because each Ethernet frame transmitted between switches has a special label that carries a VLAN IDentifier (VID). The format of VLAN labels(see Figure 2.11) is defined in the IEEE 802.1Q standard.

VLANs are created to provide the segmentation of services regardless of the physical configuration of the network. VLANs include address scalability, security and network management. Routers in VLAN topologies are very important, because they

provide broadcast filtering, addressing and traffic flow management.

Hands-on: Transparency Tests across VLANs

Ethernet networks with VLANs use two fields to forward frames: the destination MAC address and the VID. These networks must also have layer-3 routing mechanisms to make different VLANs communicate. As a result, the fields the network has at its disposal for forwarding packets are: Destination MAC address, Destination IP address and VID.

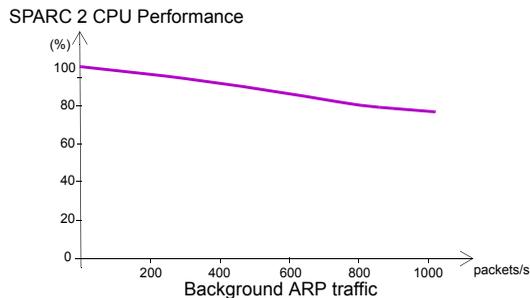
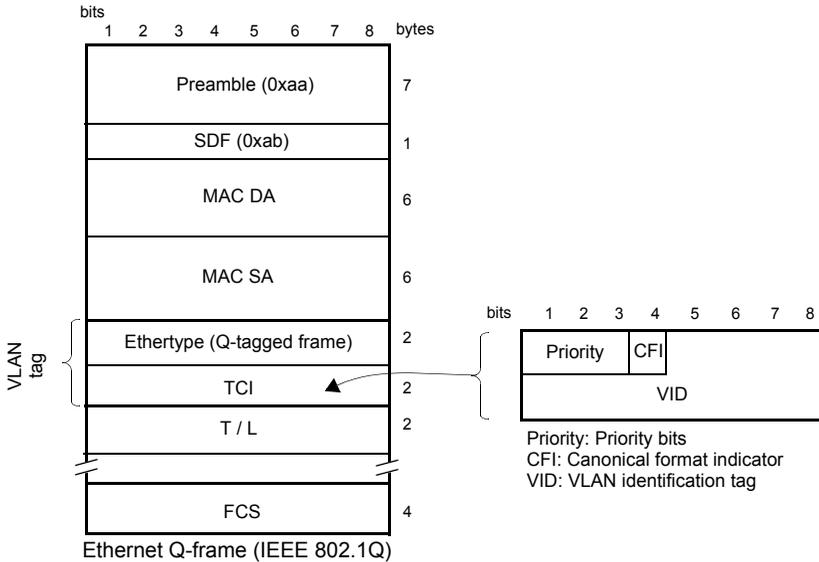


Figure 2.10 Loss of performance of a workstation due to broadcast Ethernet frames transporting ARP data.

How does the network coordinate forwarding using these fields?

VLAN configuration affects IP addressing. VLANs form separate broadcast domains. Those network interfaces that are connected to the same domain must belong to the same IP network, and they must have the same IP network prefix. On the other hand, those network interfaces that are connected to different VLANs must have different IP network prefix. In fact, this is one of the advantages of VLANs: They allow end users to keep the same IP addressing regardless of the physical port they are using. For

example, a network made up of three VLANs, 101, 102 and 103, could be configured with three different IP network prefixes like 192.168.101.0/24, 192.168.102.0/24 and 192.168.16.103.0/24.



Ethertype: Ethertype for VLAN frames (0x8100)
TCI: Tag control information, contains the VLAN ID and other fields

Figure 2.11 Ethernet frame with 802.1Q VLAN field structure.

When a user in a particular VLAN decides to send information to other user in the same VLAN, the network simply forwards the frames by using bridging. It may be necessary to use the ARP protocol to get the destination MAC address before transmission, but the IP layer is not used in any other way during the communication process. A switch may use flooding, if the destination MAC address is not listed in the local switching table.

Flooding occurs directly in the untagged ports of the switch, but frames are also forwarded to trunk links to enable communication with those users who are connected to other switches in the same VLAN. In this case, the corresponding VID is added to the original MAC frame.

Communication between VLANs takes place when one user sends packets with a destination IP address associated to a different VLAN. The data transmitted goes through the source VLAN and arrives to a VLAN default gateway or router. The default gateway then removes VLAN tags, if necessary, and decides whether the packet is to be forwarded to a new VLAN by using a routing table. Finally, the router adds the necessary VLAN tags with the correct VID, and forwards the frame to the correct interface (see Figure 2.12). Things are similar when a user needs to send information to an external network, but in this case, the router may need to add an entirely new layer-2 encapsulation, such as PPP, to the packets, rather than remove and add VLAN tags.

Routing between VLANs can be carried out more efficiently, if the router is connected to a trunk interface. With this architecture, only one physical link between one switch and the router is used to interconnect VLANs. This connection scheme requires a router that supports the IEEE 802.1Q interface. This network interface must be configured in the router with as many IP addresses as VLANs to interconnect. Each IP address must be associated with a specific VLAN. For this interface, IP addressing must be compatible with the VLAN addressing of the network (see Figure 2.13). Some switches with layer-3 features can route traffic between VLANs without using an external router. These switches have their own routing tables, and they use the destination IP address to choose the outgoing VLAN.

All networks and network segments with VLANs should be tested before bringing them into service. Especially, it is very important to

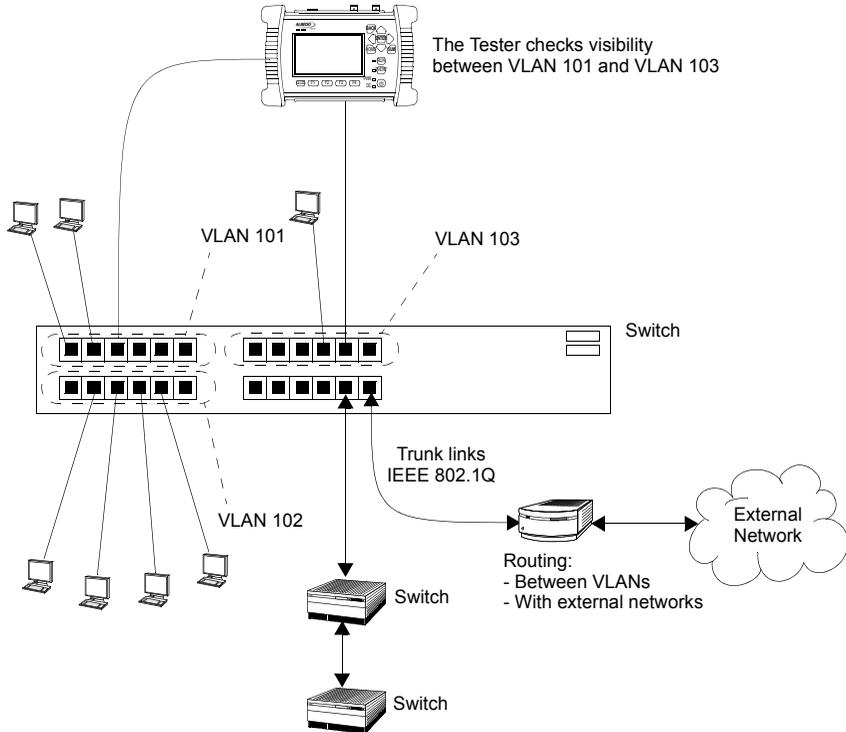


Figure 2.12 A switch with two trunks carrying VLAN-tagged frames. One of the links connects VLAN users from different switches across the network. The other one is used for connection between VLANs and with external networks. All frames that need to change their VID must pass through this router.

test isolation and routing between VLANs. Isolation can be tested in two different ways:

- From a user interface. A traffic generator is connected to an interface associated with the VLAN under test, and broadcast traffic is generated (destination MAC address `ff:ff:ff:ff:ff:ff`). A traffic analyzer is connected to those network interfaces that are associated

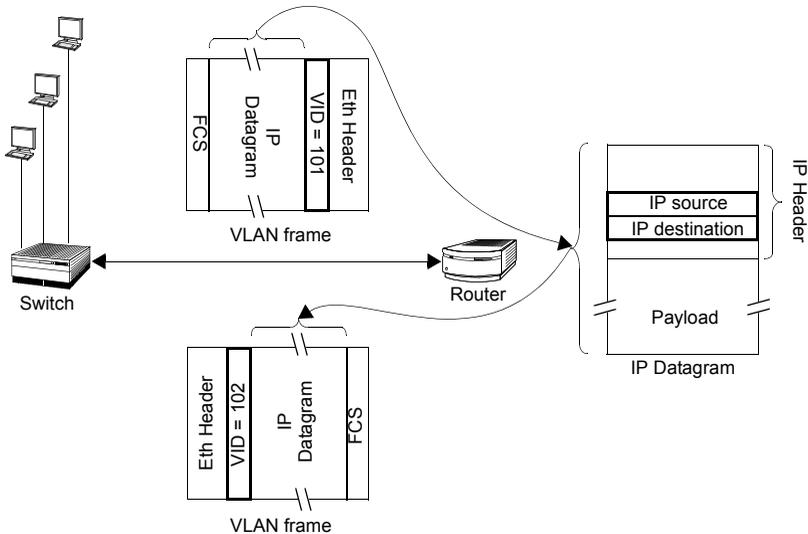


Figure 2.13 Routing between VLANs. Traffic is delivered to a router. The router checks the destination IP address and replaces the source VLAN tag with the destination VLAN tag. The network then forwards the traffic to the destination by bridging.

with the same and different VLANs. The test is performed to check that broadcast traffic is flooded only to those interfaces that are attached to the same VLAN.

- From a trunk interface. A traffic generator is connected to a trunk interface. Broadcast traffic is generated from the trunk interface, and the test is carried out to check that traffic is received in the VLAN expected. All VLANs can be tested at the same time with a traffic generator that has multistream generation capabilities. Several simultaneous broadcast traffic flows with different VLAN tags are then generated. The analyzers used will check that every VLAN receives the corresponding traffic flow.

To test routing between VLANs, it is necessary to use a traffic generator with IP generation capabilities. The test traffic must be addressed to an IP address in the remote VLAN to be tested. If the network has been set up correctly, the traffic will reach the router. The router then removes and adds the correct VLAN tags, and finally, the traffic will reach the traffic analyzer at the destination VLAN. There is no need to configure any of the destination MAC addresses, if the traffic generator can generate ARP requests, and if both the router and the traffic analyzer are able to answer to them. Detailed analysis of the traffic in the network should detect this ARP traffic and the test traffic as well. The inter-VLAN test traffic is unicast. This means that it will not be received in several ports at the same time. Especially, if the source and destination VLANs are in the same switch, the test traffic will not be switched to any trunk interfaces other than the one connected to the router. It will probably be possible to detect some broadcast ARP traffic in trunk links while the test is being carried out.

The Spanning Tree Protocol Family

The first Ethernet networks were implemented with a *coaxial bus topology*. Individual coaxial segments could be interconnected with repeaters, as long as multiple paths did not exist between any two stations.

During the 1980s, bridges and routers reduced the number of stations per segment to split traffic in a more efficient way, according to the user requirements. Separating traffic by departments, users, servers or any other criteria reduces collisions while increasing aggregated network performance.

Since the early 1990s, the network configuration of choice has been the *star-connected topology*. The center of the star is a hub, a switch or a router, and all connections are point-to-point links from the

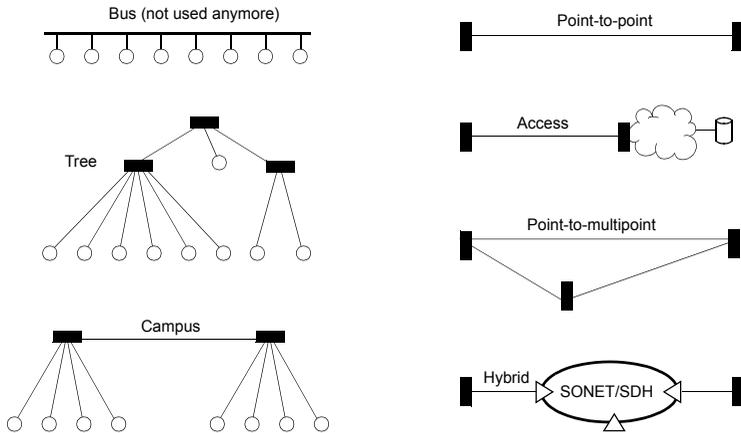


Figure 2.14 *Topology* is a very general term, and only very simple networks can be classified into one topology only. Ethernet networks could be a combination of several trees interconnected to other trees and remote services as well, using different solutions such as point-to-point, multipoint or hybrid.

center to the station. This topology has proven to be the most flexible and easy to manage in LAN networks.

New high-speed Ethernet versions have gained increasing acceptance since the year 2000, competing for the campus and metropolitan markets where *point-to-point*, *ring*, *meshed* physical topologies are common (see Figure 2.14).

However, logical topology of pure Ethernet networks based on bridging must meet some conditions in order to guarantee proper operation. If not the physical topology, at least the logical topology must be free of loops. Otherwise, the network could become unusable.

Redundancy and Bridging

We say that there is path redundancy between one selected origin and destination in a network if there is more than one physical path (not containing loops) between origin and destination. Some networks like routed IP networks, can be configured to use path redundancy as an advantage. Load balancing splits traffic between all or some of the available paths to a destination to improve efficiency and network usage. When a particular path becomes unavailable traffic can be routed to a backup path thus improving network resiliency.

When it comes to switched Ethernet networks, advantages of physical redundancy are not so clear. One of the reasons is that there is not an equivalent of load balancing for bridged networks. In fact, redundancy, can completely shut down a bridged network when is not configured in the right way.

The first thing to be noticed is that in switched Ethernet networks multiple copies of the same frame can be received at destination when the network contains loops (i. e. redundancy) (see Figure 2.15). This should not be a big problem because upper layers will probably recognise and drop duplicated packets by means sequence numbers or an equivalent mechanism. Duplicated packets waste bandwidth however.

The second issue linked with loops is switching table coherence and stability. Ethernet switches use the frames they receive to build their switching table. However, it can be easily seen that if the network contains loops there are not guarantees regarding the port from which frames from a selected origin will ingress in the switch. Therefore, the switch cannot locate the source and it is unable to switch frames directed towards this source.

Broadcast storms constitute the third issue related with loops in bridged networks. Broadcast frames are forwarded to all the switch ports (within the same VLAN) with the exception of the ingress

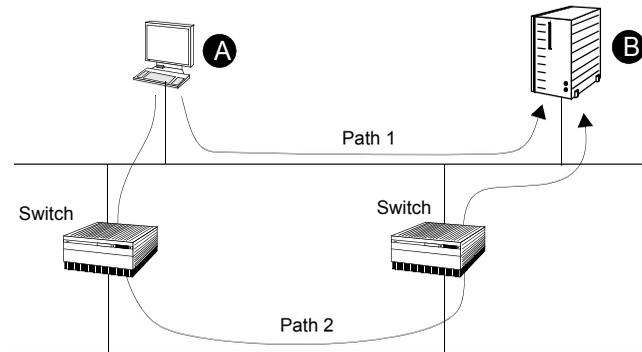


Figure 2.15 There is a loop in the network illustrated in the figure. As a result, data has to different ways to flow from A to B.

port. If the network contains loops a broadcast frame may reach a port of a switch that previously broadcasted the same frame. The result is that a single broadcast packet is multiplied without control until completely exhausting the available bandwidth. If this happens all other communication will be very difficult if not impossible.

The Classic Spanning Tree Protocol

The Spanning Tree protocol (STP) was invented to remove loops in a bridged network and avoid frame multiplication, switching table instability and broadcast storms.

The STP is the tool network administrators have to allow physical redundancy but restrain the possibility of any logical loop by means a tree topology. The STP protects the network and provides resiliency.

The first version of the STP was released by Digital Equipment Corporation (DEC) in 1985. This protocol was later reviewed by the

IEEE 802.1 committee and published in the IEEE 802.1D standard in 1990. DEC STP and IEEE STP are not compatible and they cannot be used in the same network.

Operation Principle

In order to compute a spanning tree, the STP first selects one switch in the network to become the tree root. All other switches will be connected to the root switch by a single path while potential alternative paths will be disabled by the protocol.

A single port per switch is chosen to become the switch root port (for all switches but the root). All traffic directed to the root switch will be directed through the root port. A single switch per network segment is selected as the designated switch for that segment. All traffic in the network segment directed to the root switch will use the designated switch for that segment. The idea is that if just one switch per segment forwards traffic to the root bridge loops will be broken (see Figure 2.16).

Once the root switch, the root ports and the designated switches have been decided, a single path can be found joining each switch and the root switch. In other words, all switches will be connected and there will be no loops in the network topology.

Bridging PDUs

To compute the root switch, designated switches and root ports, switches in an Ethernet network need to share information about topology, link weights, and other miscellaneous information. This is done with the help of the so called Bridging Protocol Data Units (BPDUs). BPDUs are generated only by switches and are directed to the STP MAC multicast address defined as: 01:80:C2:00:00:00. BPDUs are sent every 2 seconds by default but network administrators may change this time at their will.

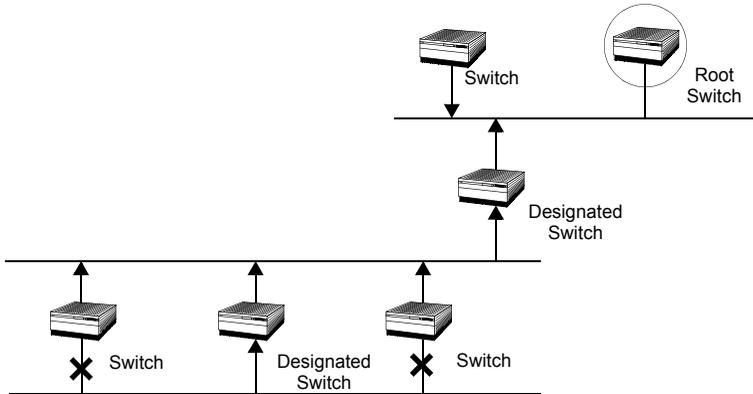


Figure 2.16 Spanning tree computed with the STP protocol. An arrow is selected in root ports and a cross in non-designated ports.

One important data field carried by BPDUs is the Bridge IDentifier (BID). Each switch has its own BID. The STP chooses the root switch as the switch with smaller BID. The BID is a composite value between the switch MAC address and a 2-byte value configurable by network administrators. This value can be used to control the switch who becomes the root switch when the STP is run.

When the root bridge has been selected, each non-root switch chooses the port (root port) it will be used to forward all the traffic directed to the root switch. To do that, switches use the concept of weight. A number called weight is assigned to each segment based on its bandwidth: The higher the bandwidth is, the lower the weight. The BPDUs sent by a switch in every port contain the cumulative cost towards the root. This information is then used to decide the best choice to interconnect the switch to the root.

However, this alone, does not guarantee a loop free topology. The last operation performed by the STP involves selection one and only one bridge per network segment as the responsible for forwarding frames towards the root bridge (segment designated switch). Selection of designated bridges is a weight based decision like it is selection of root ports.

Ports different to the root and designated ports become non-designated ports. Normally, non-designated ports are in a blocked status and they do not forward traffic.

STP and Network Resiliency

The STP protects the network against broadcast storms and switching table instability but it also provides resiliency in front of failures.

Under normal operation, the root switch generates BPDUs periodically (usually every 2 seconds) and all other switches propagate these BPDUs when they receive them.

The time the last BPDU was received in every port is recorded by switches. After a defined aging time (usually 20 seconds) without receiving any BPDU a loss of connectivity with the root switch is declared and a new spanning tree is computed to restore communications. To enable calculation of a new spanning tree, the blocked ports go to a listening status. While they are in this status they send and receive BPDUs and with the information they get from the network they choose new root and designated ports. Once the new tree has been computed ports go to a learning status before start forwarding frames. During the learning status they build their switching tables. Usually one port stay 15 seconds in a listening status plus 15 more seconds in learning status. If this is added to the 20 seconds aging timer we get 50 seconds as the worst case protection switching time.

Rapid Spanning Tree Protocol

While a protection switching time around 50 seconds can be considered a quite good figure in some Enterprise networks this is clearly not enough when a highly available service is required.

The Rapid STP or RSTP was defined to speed up the protection switching time in bridged networks. This protocol was published in standard IEEE 802.1w as a supplement of IEEE 802.1D-1998 and it is currently included in IEEE 802.1D-2004. The RSTP is backwards compatible with the older STP. However, if STP and RSTP are deployed in the same network all the RSTP benefits will be lost.

The basis of the RSTP is the definition of new mechanisms to enable direct transition between the blocked and the forwarding port status without going through the listening and learning status.

Ports where stations are connected are defined as edge ports. Theoretically, an edge port can never belong to a loop. For this reason, edge ports are directly put in a forwarding status. Ports connecting one switch to another switch may potentially belong to a loop. In this case switches exchange explicit signalling messages to prevent loops. This signalling involves commands to put a port in forwarding status directly from status blocked without going through the learning and listening status as it happens with the classic STP protocol. In this way, the RSTP saves the time STP wastes learning and listening (usually 30 seconds).

A secondary RSTP improvement comes from the fact that the time to declare loss of connectivity has been shortened to three 'hello' periods (6 seconds usually). That means that the overall protection switching time for the RSTP is around 6 seconds. Much better than the STP but still far from the standard for WAN: 50 ms.

Multiple Spanning Tree Protocol

While the STP and the RSTP are useful to solve issues related with

loops and provides some kind of resiliency but they fail to maximize network usage because they block transmission over not designated ports.

An obvious solution is to run as many STP instances as VLANs you have in the network but this is very resource consuming in networks with many VLANs. The IEEE 802.1s standard defines the Multiple Spanning Tree Protocol (MSTP) which has a different approach.

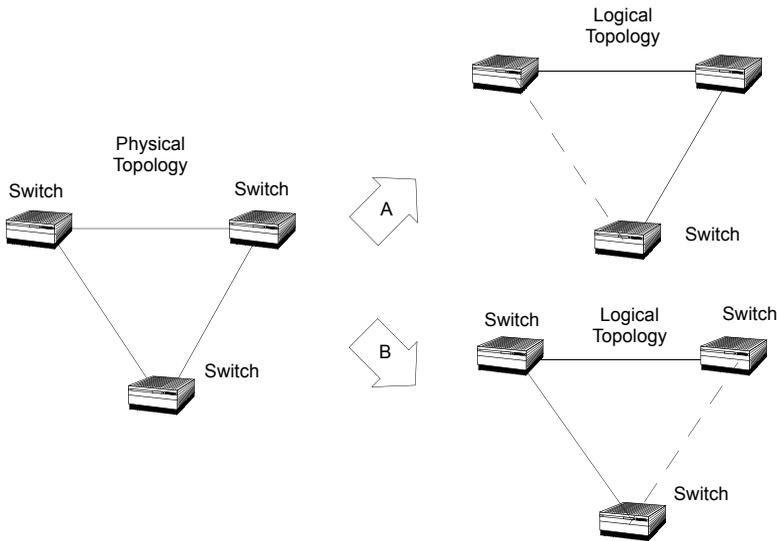


Figure 2.17 .Physical and logical topologies in a bridged network running the MSTP protocol.

The MSTP builds several logical topologies over a single physical topology by generating several spanning trees and after that mapping existing VLANs to the trees to improve network efficiency (see Figure 2.17). Note that this is very different to computing an spanning tree per each VLAN. With the help of MSTP VLANs reuse

the same spanning tree. As a result, switch resources are not wasted.

MSTP does not need to send a BPDU for each spanning tree. Instead of that, a special spanning tree instance called Internal Spanning Tree (IST) is designated to carry signaling for all. In the default configuration, all VLANs are mapped to the IST and the MSTP becomes the RSTP but other configurations are possible. Many alternative spanning trees can coexist with the IST, these new entities are known as Multiple Spanning Tree Instances (MSTIs).

Each MSTI may assign different priorities to switches, may have different link costs, port priorities and thus end up with it's own logical topology. For each MSTIs, this information is carried in specially dedicated fields within the IST BPDUs.

The Network Layer

The Network Layer, or Layer 3, provides end-to-end connectivity between stations that can use heterogeneous underlying technologies (see Figure 2.18) but are not necessarily attached to the same network. Routers are devices that are designed to manage Layer 3 protocols and data forwarding based on routing tables.

Despite their similarities, such as the ability of matching addresses to output interfaces, routing and switching tables have fundamental differences:

- Layer 2 addresses are unstructured and simple to use, and are intended for a short or medium number of destinations. Switching tables, that manage layer 2 host addresses, are built in a learning process based on previous transmissions.
 - Layer 3 addresses are hierarchical to facilitate complex and efficient addressing for a large quantity of destinations. Routing ta-
-

bles manage layer 3 host and network addresses. They are configured manually or by the routing protocols.

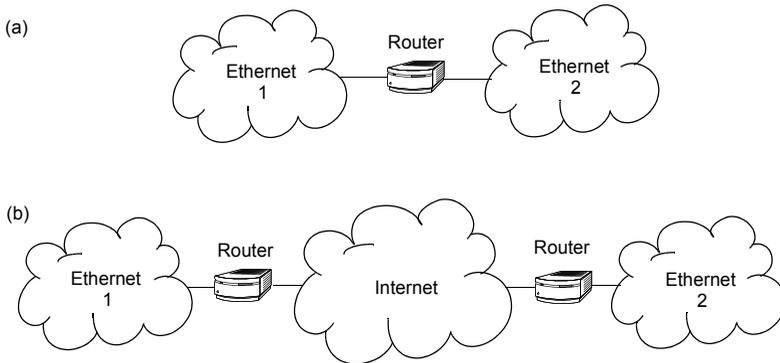


Figure 2.18 Interworking with routers. (a) A router is used to forward data from an Ethernet network to a second Ethernet network. (b) A chain of routers delivers traffic from an Ethernet network to a second Ethernet network through the Internet.

The TCP/IP Reference Model

The TCP/IP reference model and protocol suite is the most popular architecture to interconnect heterogeneous networks. They have made possible the Internet and most of the intranets since are equally well suited for both local and wide area networks. TCP/IP certainly has similarities to the OSI model but is not necessarily compliant with it (see Figure 2.19).

The complete suite of TCP/IP protocols is structured in four layers, that correspond from the data link layer up to the application. Encapsulations, protocols and data formats are used to interchange data between the layers.

All the TCP/IP architecture relies in a very simple fact: There is a single Layer 3 protocol, the Internet Protocol (IP). However, this

basic premise is not true anymore due to the introduction of the IP version 6 (IPv6) protocol. Currently, there are two coexisting network protocol versions, the new IPv6 and the legacy IPv4. IPv6 provides a 64-bit address space in front of the 32-bit IPv4 addresses. Due to the larger address space, the IPv6 will fix the IPv4 address scarcity problem but coexistence of IPv4 and IPv6 (at least for some time) is a major challenge for the IP and the Internet.

OSI model	TCP/IP model	TCP/IP protocols
Application	Application	DNS, FTP, HTTP POP3, SMTP, RTP, Telnet, SNMP, SIP
Presentation		
Session		
Transport	Transport	TCP, UDP
Network	Internet	IPv4, IPv6
Data Link	Data Link	ARP, PPP
Physical	Physical	SDH, OTN, Ethernet

Figure 2.19 TCP/IP reference model and its relationship with the OSI model.

The Internet Protocol

The *Internet Protocol* (IP) is the most popular Layer3 protocol. It was conceived by the U.S. *Department of Defense* (DoD) during the cold war to facilitate communication between dissimilar computer systems and is a reliable technology. IP interconnects public or private autonomous systems providing a connectionless service.

There are two IP protocol versions (IPv4 and IPv6). IPv4 addresses are defined as a subset of the IPv6 addressing space but IPv4 and IPv6 can be regarded as different and incompatible network protocols in any other sense (see Figure 2.20). Currently, IPv4 is still the dominant version but this will change in the future.

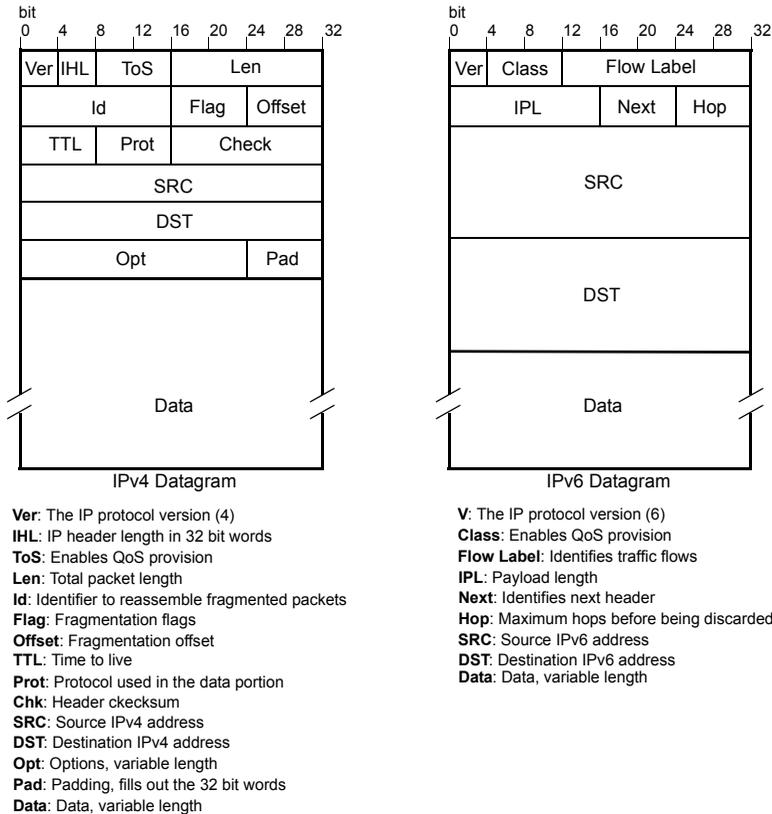


Figure 2.20 IPv4 and IPv6 datagram formats. IPv4 is based in 32 bit addresses. IPv6 uses an extended address set based on 64 bit addresses.

Addresses and Networks

The addressing scheme of version 4 of IP is based on fixed length 32 bit addresses commonly written in decimal dotted format (see Figure 2.21). For example, 62.22.33.1 is a valid IP address in dotted decimal representation.

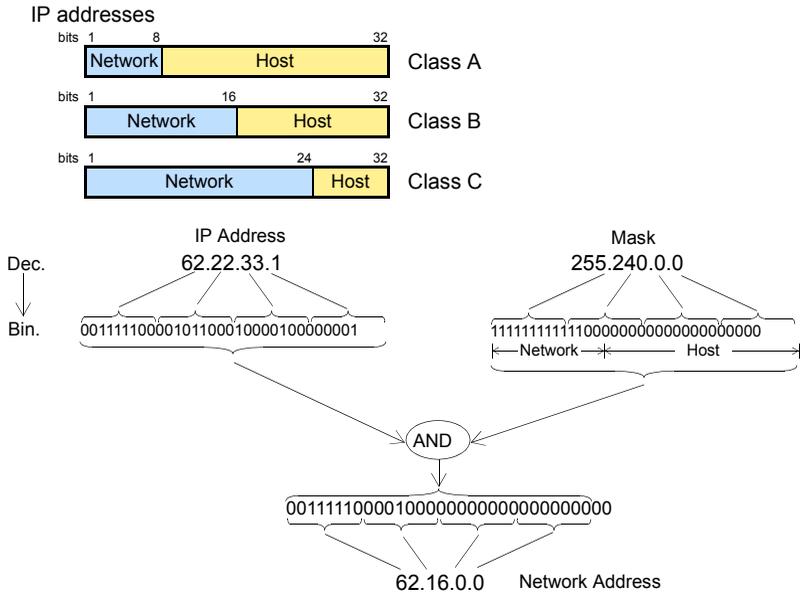


Figure 2.21 Relationship between the IP address, the network mask, the network address and conversion between the binary and the dotted decimal representations. The 32-bits mask has binary 1s in all bits specifying the network field and 0s in the host.

Each 32-bit address is divided into two fields:

1. *Network field*, assigned by the Internet Network Information Center (InterNIC) and is used to identify a network.
2. *Host field*, assigned by the Network Administrator and is used to identify a host on a network.

The size of each field varies depending on the type of address (see Figure 2.21), so it is necessary to use a mask to obtain Network and Host identifiers. The Mask must be supplied and stored in the routing tables because the routing tables are used for assessing each field, of every, IP address.

IPv4 addresses are organized into five different classes: A, B, C, D, and E being high-order bits to indicate the class (see Table 2.1). Networks can also be divided into subnetworks to be managed by local administrators to make IP addressing more efficient and flexible.

<i>Address class</i>	<i>First byte (binary)</i>	<i>Address range (decimal)</i>	<i>Number</i>
Class A	0xxxxxxx	0.0.0.0 ~ 127.255.255.255	2,147,483,648
Class B	10xxxxxx	128.0.0.0 ~ 191.255.255.255	1,073,741,824
Class C	110xxxxx	192.0.0.0 ~ 223.255.255.255	536,870,912
Class D	1110xxxx	224.0.0.0 ~ 239.255.255.255	268,435,456
Class E	1111xxxx	240.0.0.0 ~ 255.255.255.255	268,435,456

Table 2.1
Internet address classes

IP Routing

The forwarding mechanism used by IP networks is known as *routing*. IP routing is connectionless because IP datagrams do not follow a preestablished path. Instead, IP routers compute the path for each individual datagram. The resulting path may or may not be the same for all them. Each IP router has one or several routing tables to indicate the next hop to jump. Router involvement in the routing process is limited to forwarding packets based on routing tables. These tables are built based on information provided by specialised counting protocols such as Intermediate System - Intermediate System (IS-IS) protocol, Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), or Border Gateway Protocol (BGP).

Routing is very different in nature to the Ethernet bridging. Ethernet bridges are cheaper and faster than routers but on the other hand, routing is more scalable and technology agnostic. Global communications through the Internet would not be possible without the routing paradigm.

IP addressing in Ethernet Networks

Hosts within large Ethernet networks are commonly identified by their IP addresses, rather than their MAC addresses. This fact not only means that hosts are independent of their MAC addresses, it also means that a network host can be attached to different Layer 2 and Layer 1 technologies whenever they keep a common Layer 3 scheme.

Internet Control Message Protocol

IP networks not monitor whether the packets get to final destination, nor does IP provide for error reporting when routing anomalies occur. This task is executed by the ICMP protocol.

The *Internet Control Message Protocol (ICMP)* is a network layer Internet protocol that provides mechanisms to report errors and other information regarding IP packet processing back to the source. It is used for error reporting and analysis, transferring messages from routers and stations, and for reporting network configuration and performance problems.

ICMP generates several kinds of useful messages, including *Destination Unreachable*, *Echo Request* and *Echo Reply*, *Redirect*, *Time Exceeded*, and *Router Advertisement* and *Router Solicitation*.

The ICMP functionality includes: Report network errors, Congestion indication, Troubleshooting assistance, Announce packet time-outs when TTL field is set to zero.

Address Resolution Protocol

Imagine that a source host is willing to send a data packet to a destination host but only has its IP address. To get the destination MAC address the source has to broadcast an *Address Resolution Protocol (ARP)* packet which contains the IP address of the destination host and then wait for a response that contains the MAC address. RFC 826 describes the ARP.

ARP in a segment

First case (see Figure 2.22). Imagine a source and a destination host are attached to the same Ethernet network:

1. Host J wants to send information to host K but only knows its IP address.
2. Host J broadcasts an ARP packet.
3. Host K responds to the request, the rest of the hosts ignore it.
4. Host J receives the response and matches the MAC and IP addresses of host K.
5. Data transmission from host J to K can now start.

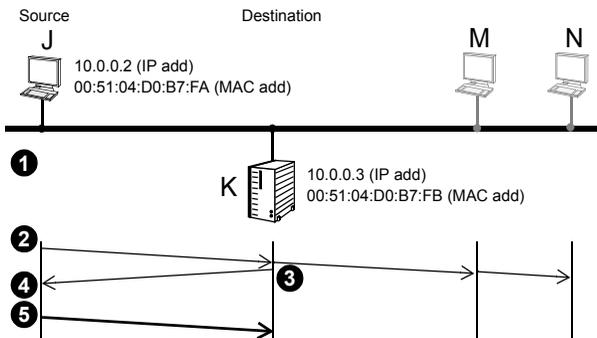


Figure 2.22 ARP operation when source and destination hosts are both in the same Ethernet network.

ARP in different LAN segments

Second case (see Figure 2.23). Imagine source and destination hosts attached to different segments connected by a router in a LAN. The router is configured as ARP proxy:

1. Host J wants to send information to host S but only knows its IP address.
2. Host J broadcasts an ARP packet containing the destination IP address.
3. The router receives the ARP and reads the network field of the destination IP address. The router finds out that the K host is in the other segment and immediately the router responds to the ARP with its own router MAC address.
4. Host J receives the response and matches the MAC address of the router to the IP address of host S.
5. Host J starts sending IP data to the destination using the router MAC address.
6. The router forwards IP data packets to host S through the outgoing interface indicated by its routing table.

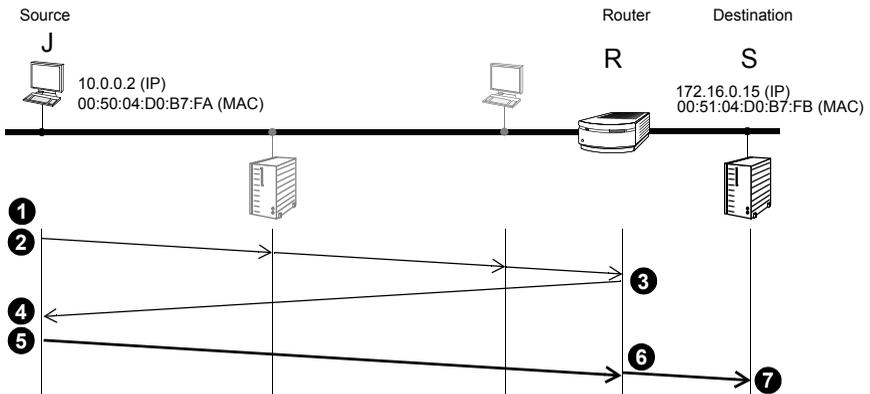


Figure 2.23 ARP operation when source and destination hosts are in different Ethernet networks and the information is forwarded between them by a router.

ARP in heterogeneous connections

Third case (see Figure 2.24). Imagine that the source and destination hosts are attached to different technologies, that is, Ethernet and ADSL, so consequently there is no continuity of MAC frames. A common solution is the use of layer 2 encapsulation, like the Point to Point Protocol (PPP).

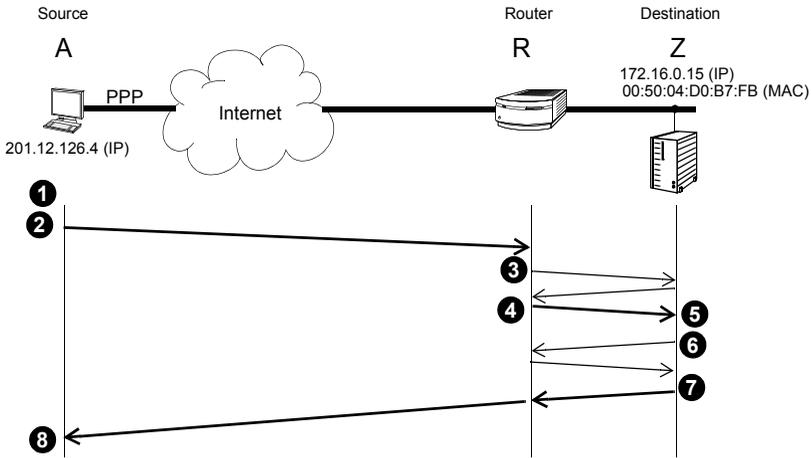


Figure 2.24 Access to a remote host placed behind a NAT firewall with a PPP connection to the Internet.

Routers connecting LANs to the Internet must facilitate the internal hosts external connectivity and protect against unauthorized access. *Network Address Translation* (NAT) is the mechanism to share the scarce public addresses used on the Internet. NAT swaps private addresses for public addresses of the outgoing packets. It does the opposite with incoming packets coming from the Internet.

Let us trace a sample of a remote access to an Ethernet host behind a NAT firewall (see Figure 2.24):

-
1. Host A wants to send information to host Z but only knows its IP address.
 2. Host A sends data using the destination IP address, however ARP is not necessary because it is using PPP which is a point to point protocol.
 3. The router R receives the packet. If the firewall rules grant access to the Z host then the packet can pass. However if the router does not know the destination MAC address, then it is necessary to perform an ARP operation.
 4. Once the router obtains the MAC address of host Z it can forward the packet that was received from host A. Before host A sends the packet, the NAT swaps the addresses.
 5. The packet is finally delivered to host Z. It has the source IP address of host A, the source MAC address of router R, the private destination IP address of host Z, and destination MAC addresses of host Z as well.
 6. Before sending packets back, it is necessary to find out the MAC address of host A using an ARP request. The router R, knows that host A is in a remote network so it responds with its own MAC address.
 7. The firewall must grant the traffic entry before the NAT swaps the private source IP address for a public one. Finally the data packet progresses to the Ethernet network.
 8. The data is routed through the Internet and arrives at host A.

Higher Layers of the TCP/IP Protocol Stack

In the TCP/IP reference model, the user applications are mapped over the IP datagrams, but there is still something needed between the applications and the network. This new ingredient is the *Transport Layer* or *Layer 4*. Unlike Layers 1, 2 and 3, the Layer 4 does not exist in the network, just in the end user equipment. The most important Layer 4 protocols are the *Transmission Control Protocol*

(TCP) and *User Datagram Protocol* (UDP) (see Figure 2.25). Each of them has different advantages and drawbacks.

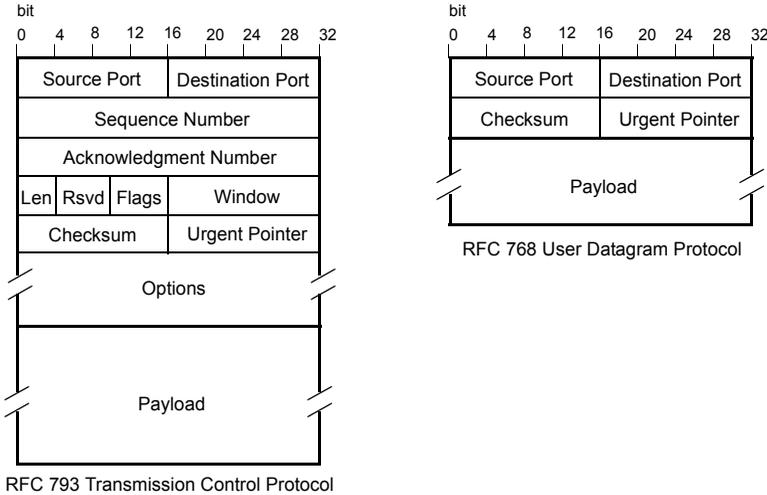


Figure 2.25 TCP and UDP segment structure.

Basically, Layer 4 is in charge of three different tasks within the TCP/IP reference model:

- *Reliable communications* between the transmitter and the receiver. This category includes error control and recovery and also flow control. Of course, applications may define their own reliability mechanisms but they can also delegate this function to a Layer 4 protocol. In this category, TCP and UDP are very different. TCP provides error recovery and flow control but UDP only has basic error detection capabilities through a checksum field and a sequence number.
- *Connection oriented services*. Some applications, like for example VoIP telephone calls require some degree of persistent connections. Again, applications may define their own specific connections (incidentally, this is usually the case for VoIP) or they can rely

on a Layer 4 protocol. TCP is connection oriented and UDP is not. Layer 4 connection oriented services provided by TCP are very different to Layer 1, 2 or 3 connection oriented services. TCP connections are just endpoint associations. The network is unaware of them and therefore the network itself cannot carry out any operation based on Layer 4 associations.

- *Application multiplexing.* This is probably the most important functionality added by Layer 4 protocols because it cannot be delegated to the applications. IP provides communications to hosts but not to applications. This task is left to TCP and UDP. In this case both protocols are identical. Applications use Layer 4 ports to send and receive data. A port is simply a 16-bit identifier. Associations of IP addresses and ports constitute virtual communications channels called sockets. Applications speak one each other through sockets.

Transmission Control Protocol (TCP)

The TCP provides reliable transmission of data in an IP environment by means of connection-oriented service. Streams are identified by sequence numbers to acknowledge those bytes that have been received correctly. TCP offers end-to-end flow control, full-duplex operation, multiplexing features that allow several conversations to be multiplexed over a single connection.

User Datagram Protocol (UDP)

UDP is a connectionless transport-layer protocol. Unlike the TCP, UDP does not offer reliability, flow-control, or error-recovery functions. UDP offers simplicity as headers are shorter and functionality is limited compared with TCP. The consequence is less overhead, more efficiency, and higher bit rates.

UDP is useful in situations where the reliability mechanisms of TCP are unnecessary, for instance, does it make sense to recover a bit error during a VoIP conversation?

UDP is the transport protocol for several well-known application-layer protocols, including NFS, SNMP, DNS, TFTP.

Application-Layer Protocols

The Internet protocol suite also specifies common application-layer protocols including:

- File Transfer Protocol (FTP): used to transfer files
 - Simple Network-Management Protocol (SNMP): reports network conditions
 - Telnet, Secure SHell (SSH): terminal emulation
 - X Windows: distributed windowing and graphics system used for X terminals and UNIX workstations
 - Network File System (NFS): Distributed
 - Remote Simple Mail Transfer Protocol (SMTP): electronic mail
 - Domain Name System (DNS): Translates domain names into addresses
-

Selected Bibliography

- [1] IEEE 802.3-2008, "Part 3: Carrier sense multiple access with collision detection (CSMA/CD) Access Method and Physical Layer Specifications," December 2008.
 - [2] IEEE 802.1D-2004, "Media Access Control (MAC) Bridges," June 2004.
 - [3] IEEE 802.1Q-2005, "Virtual Bridged Local Area Networks Revision," May 2006.
 - [4] S. Armyros, "On the Behaviour of Ethernet: Are Existing Analytic Models Adequate?," Technical Report CSRI-259, Computer Systems Research Institute, University of Toronto, February 1992.
 - [5] Rich Seifert, *Gigabit Ethernet Technology and Applications for High/Speed LANs*, Addison Wesley Oct 1999.
 - [6] William Stallings, *Data and Computer Communications*, Prentice Hall, 1997.
 - [7] Kevin L. Paton, "Gigabit Ethernet Test Challenges," Test and Measurement World Magazine, Oct 2001.
 - [8] Mandeville R., "Benchmarking Terminology for LAN Switching Devices," IETF Request For Comments RFC 2285, February 1998
 - [9] Bradner S., McQuaid J., "Benchmarking Methodology for Network Interconnect Devices," IETF Request For Comments RFC 2544, March 1999.
 - [10] Bradner S., "Benchmarking terminology for network interconnection devices," IETF Request For Comments RFC 1242, July 1991.
 - [11] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.
 - [12] R. Metcalfe and D. Boggs, "Ethernet: Distributed packet switching for local computer networks," *CACM*, Vol. 19, N. 7, July 1976.
-

Carrier Ethernet

Chapter 3

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.

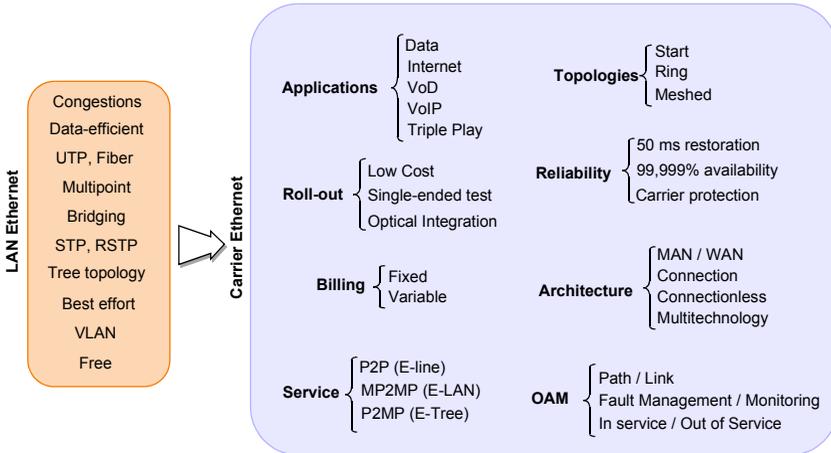


Figure 3.1 The path to Carrier-Class Ethernet.

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 3.1).

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 3.2).

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.

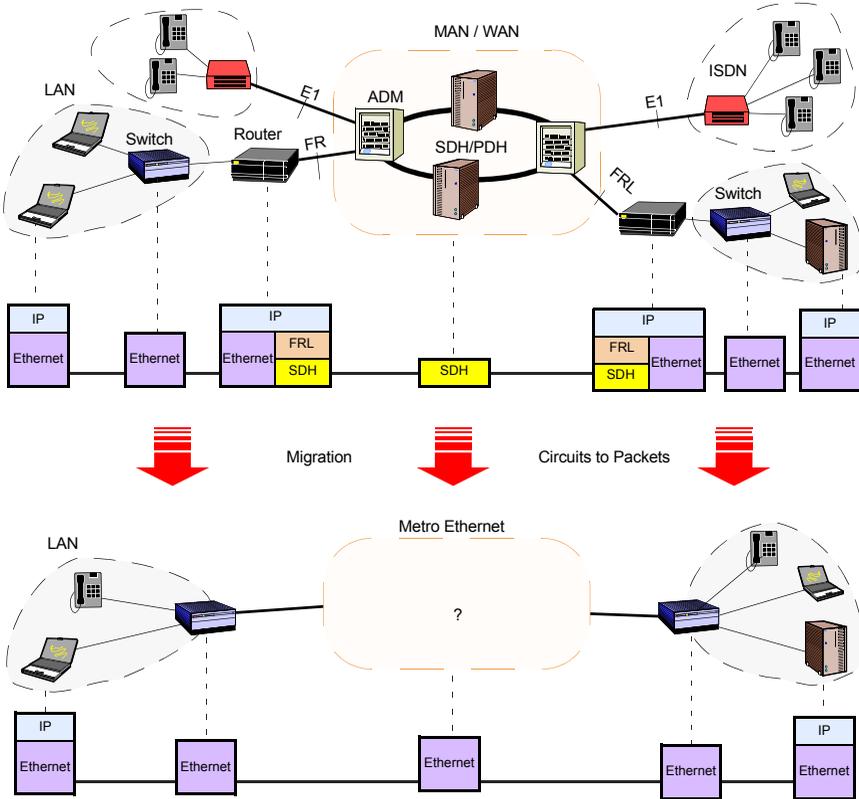


Figure 3.2 Migration to end-to-end Ethernet.

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

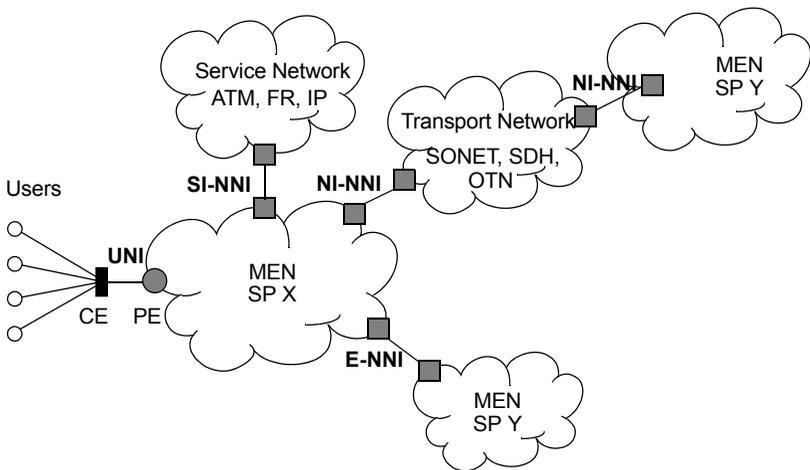


Figure 3.3 UNI and NNI in the MEN.

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 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 (Figure 3.3). 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).

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.

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 3.4). Service multiplexing is achieved by using the IEEE 802.1Q *Virtual LAN* (VLAN) ID as a connection identifier.

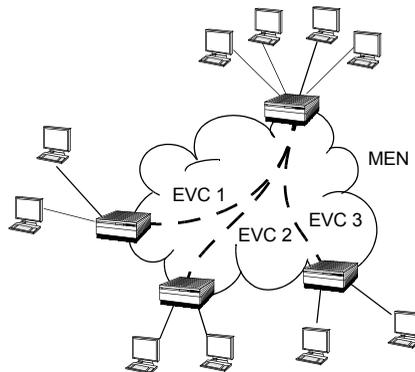


Figure 3.4 EVC Service multiplexing in a single port.

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*.

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 3.5).

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.

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.

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 3.5). 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

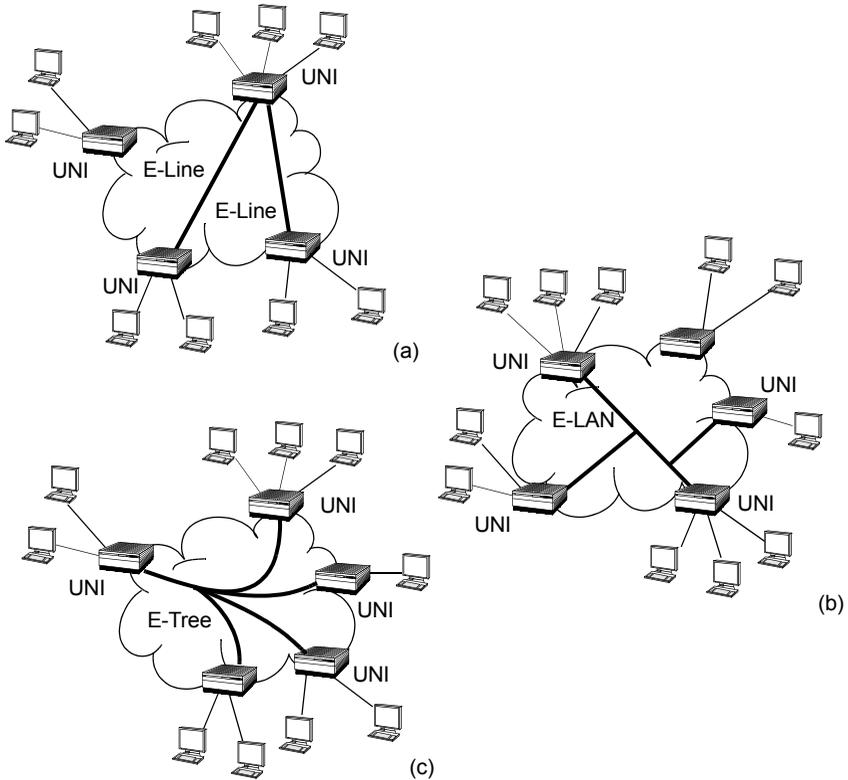


Figure 3.5 (a) The E-Line is understood as a point-to-point virtual circuit (b) The E-LAN service is multipoint to multipoint (c) E-Tree service is point-to-multipoint.

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.

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.

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 3.1).

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.

Generic Etherservice Type

EVC to UNI Relationship

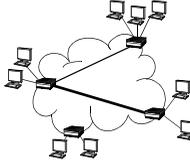
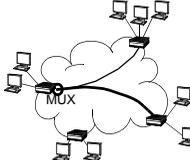
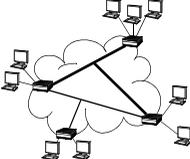
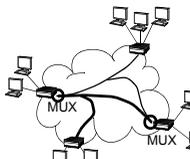
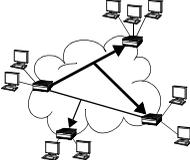
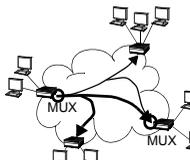
	VLAN-Based Service - Service Multiplexing - Shared Bandwidth	Port-Based Service - No Service Multiplexing - Dedicated Bandwidth
E-Line - Point to point - Best-effort or guaranteed QoS - Optional multiplexing and bundling	Ethernet Private Line (EPL) 	Ethernet Virtual Private Line (EVPL) 
E-LAN - Multipoint 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) 

Table 3.1 Ethernet Connectivity Services

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.

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.

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.

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 Ethernet 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.

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.

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 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
-

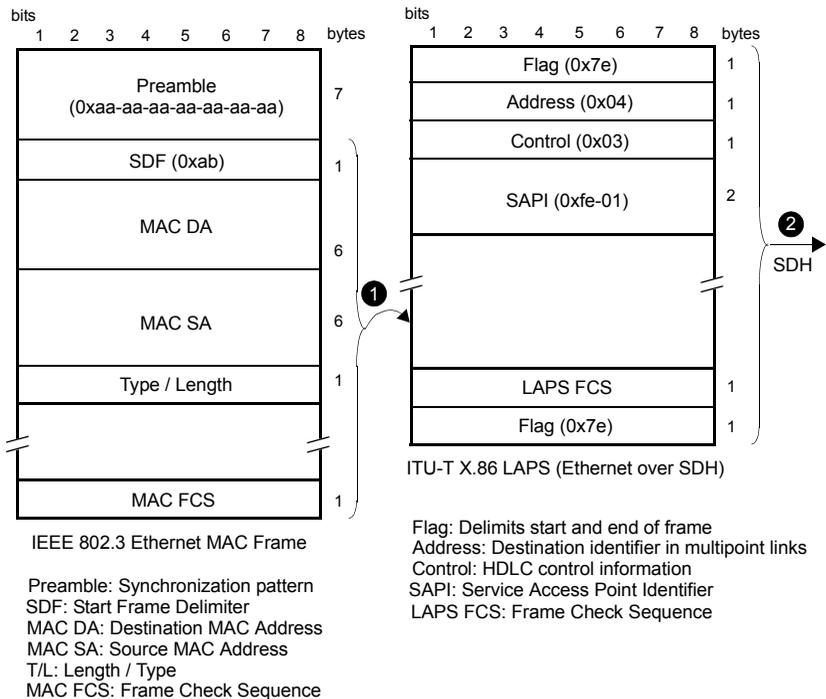


Figure 3.6 Legacy encapsulation for transporting Ethernet over SDH networks. The SDH solution makes use of the of the LAPS encapsulation.

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 3.6).

- 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 3.7).

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 3.8). 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.

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.

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

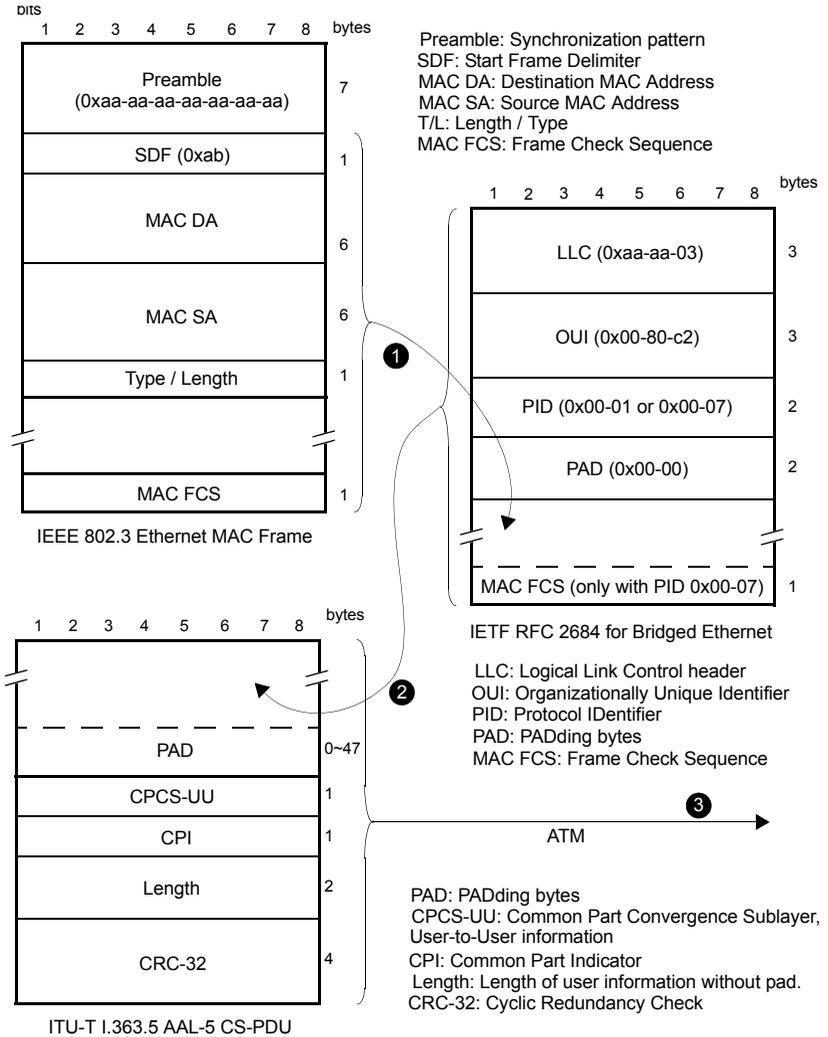


Figure 3.7 Legacy encapsulations for transporting Ethernet over ATM networks. The ATM mapping uses RFC-2684 and AAL-5 encapsulations.

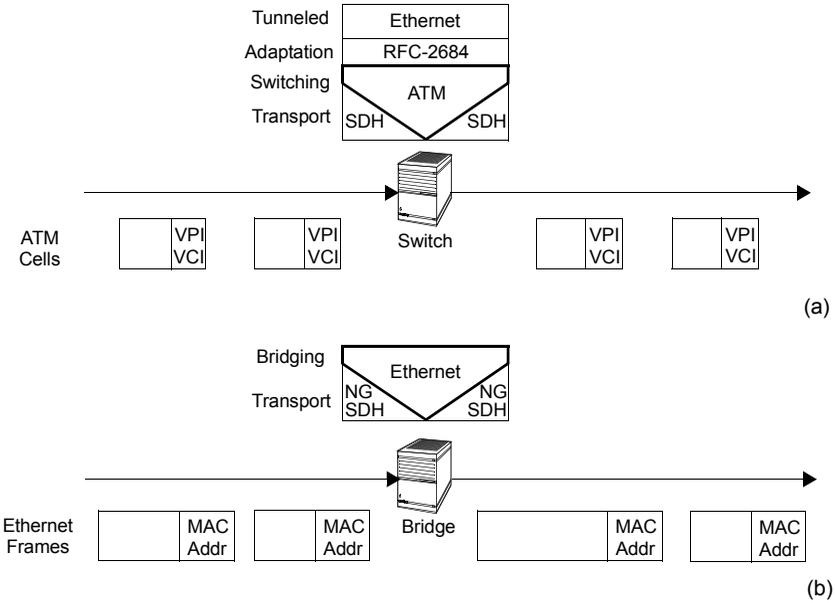


Figure 3.8 How NG-SDH raises the importance of Ethernet in the MAN / WAN. (a) Ethernet traffic is passively transported like any other user data. The ATM layer, specific for the WAN, is used for switching traffic. (b) The ATM layer disappears and the Ethernet layer becomes active. Traffic is now guided to its destination by means of Ethernet bridging.

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 (Figure 3.9). 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

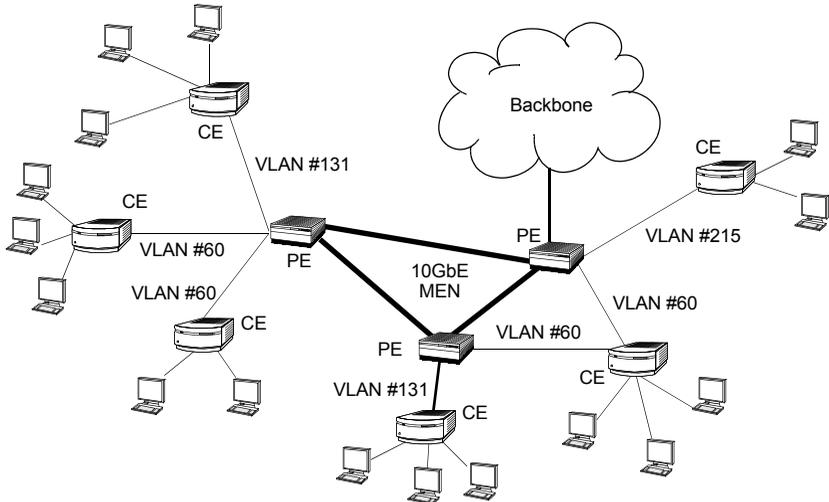


Figure 3.9 Optical Ethernet Carrier network. The trunk MEN links are implemented with optical 10-Gigabit Ethernet. Access links can be based on 1-Gigabit Ethernet or *Ethernet in the First Mile (EFM)*, depending on the bandwidth requirements of every placement. Segregation of traffic from different subscribers or work groups is done by using VLAN tagging.

discouraged for operators who want to provide services to a large number of residential and Small Office/Home Office (SOHO) customers.

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.

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 3.10).

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.

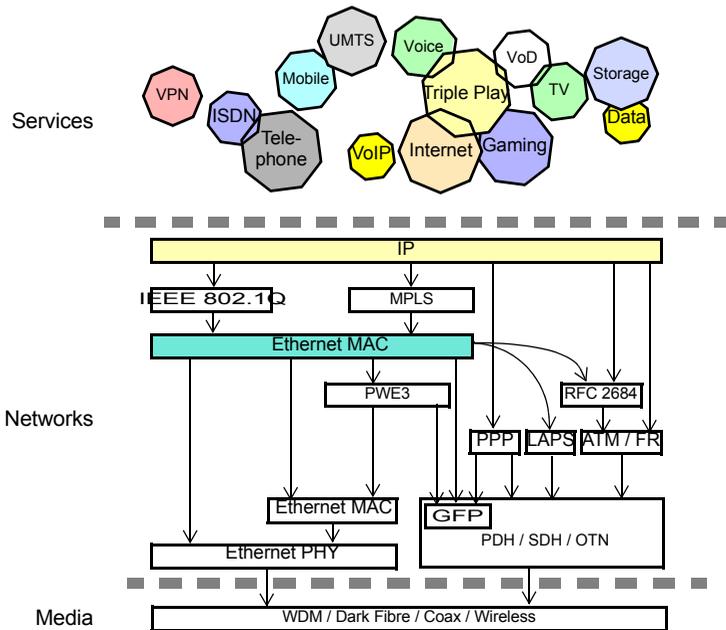


Figure 3.10 Transporting Ethernet and IP over packet- or circuit-switched infrastructures

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 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 3.11).

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.

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 3.12).

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 3.13):

- Enhanced ADMs are like a traditional ADM, but they include Ethernet interfaces to enable access to new services, and TDM in-
-

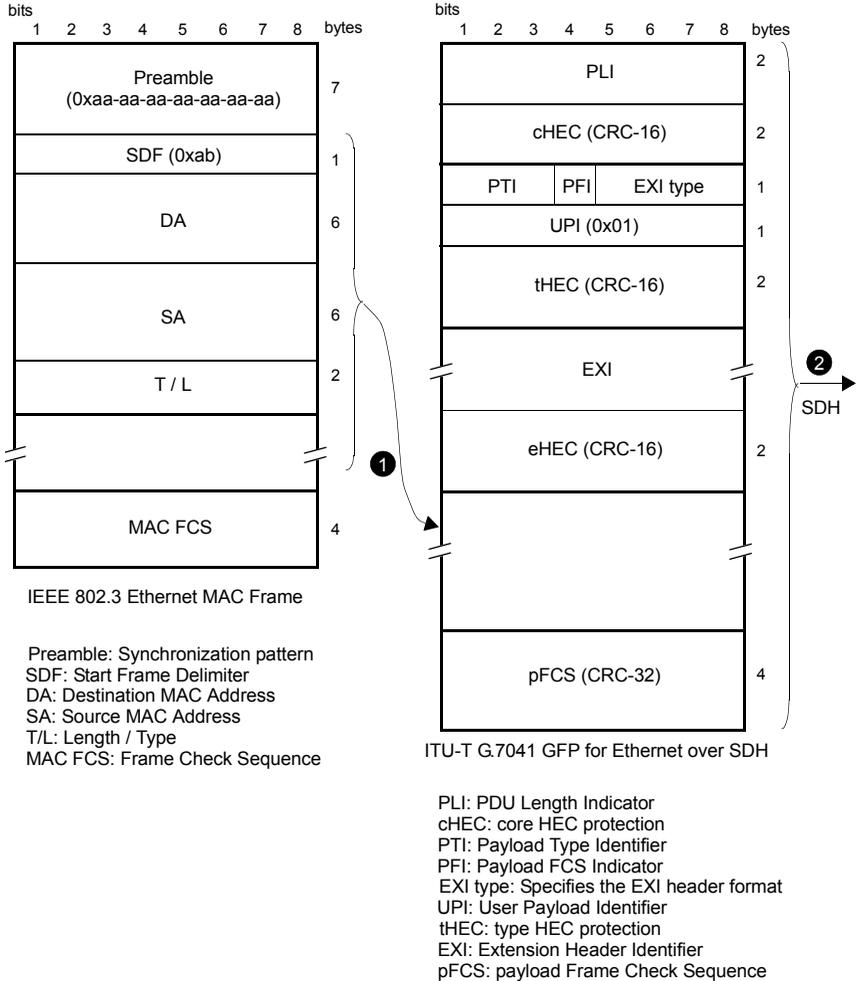


Figure 3.11 The GFP-F mapping for Ethernet makes ATM unnecessary. Now there is no VPI / VCI to switch the traffic, but the Ethernet MAC addresses can be used for similar purposes.

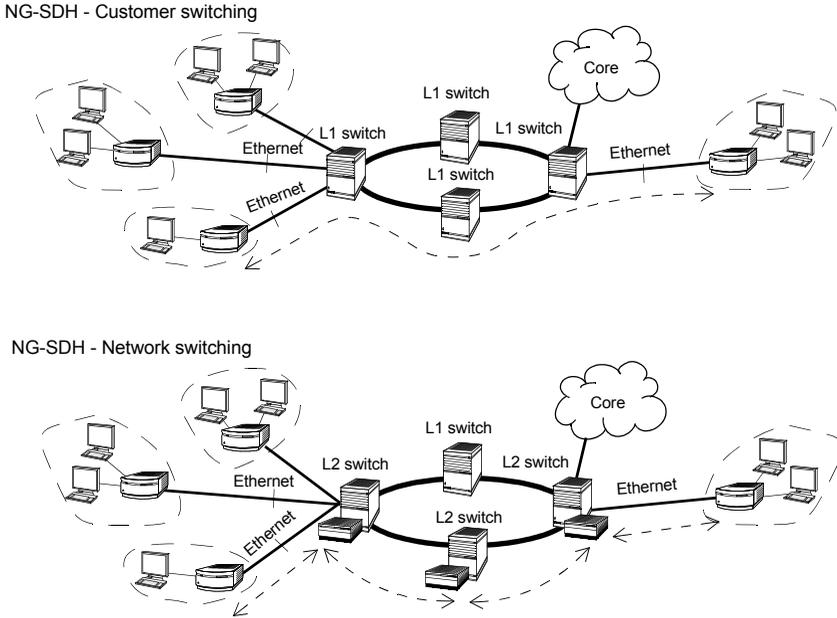


Figure 3.12 Ethernet over NG-SDH. Depending on the requirements, two approaches are possible: a) Customer switching – simple; the transport network is just a link between the customer switches. b) Network switching – more flexible; one step forward toward a more sophisticated service based on MPLS.

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.

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.

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 inconveniences. 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).

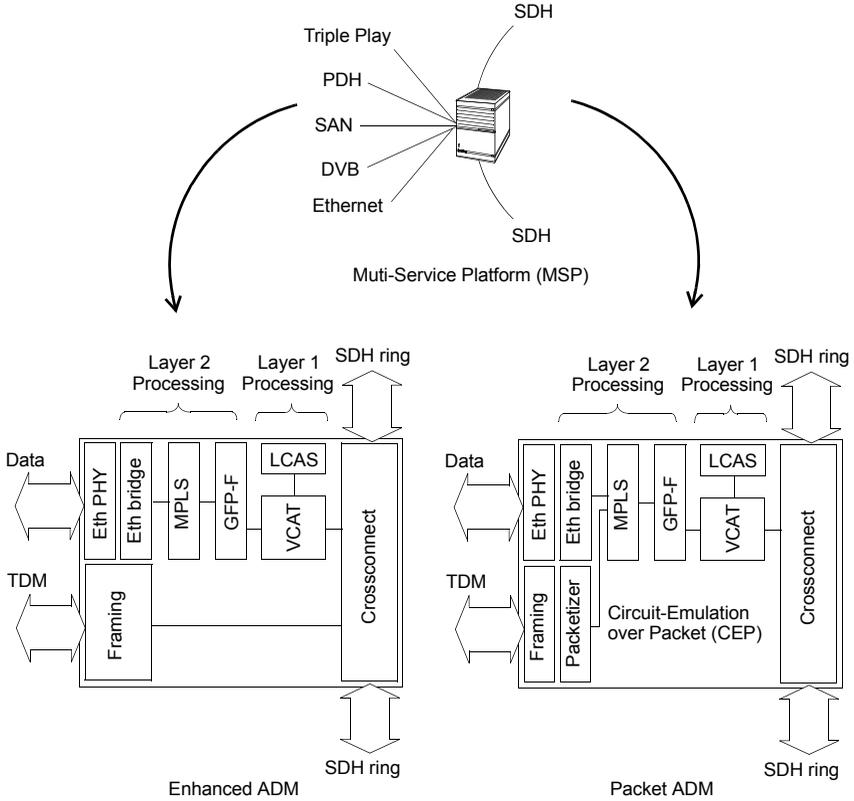


Figure 3.13 New SDH network elements. The Enhanced ADM offers packet and TDM interfaces in the same network element. Packet ADMs offer the same, but over an unified packet-based switching paradigm for all tributaries.

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 available VIDs. VLAN stacking is an standard solution defined in IEEE 802.1ad for Ethernet Provider

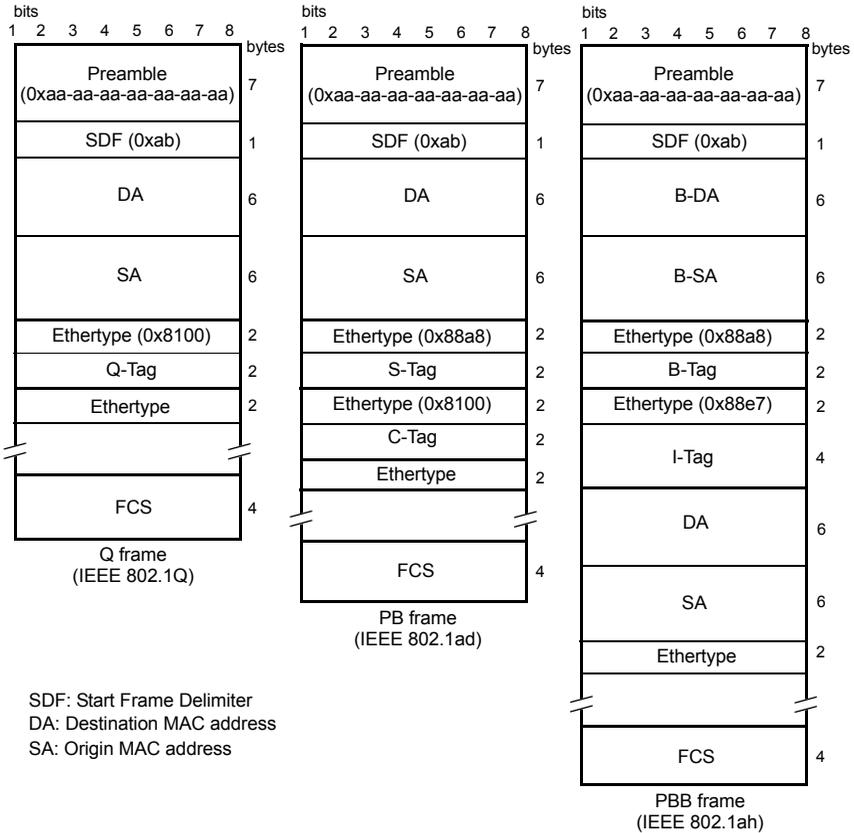


Figure 3.14 Ethernet frame formats for service provider networks. These frame formats offer a more scalable Ethernet.

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 3.14).

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?

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.

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 3.15).

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

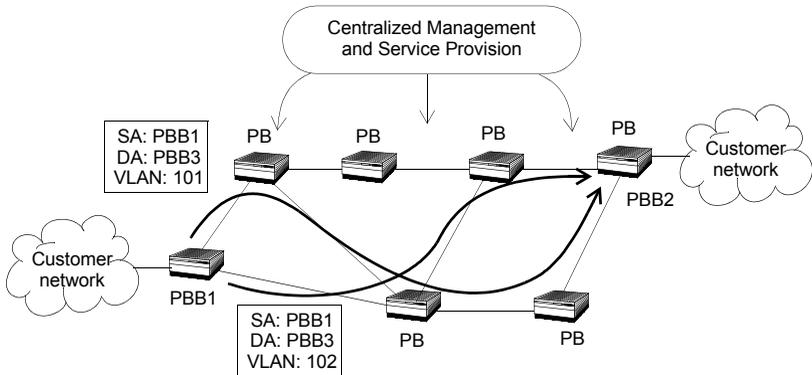


Figure 3.15 Ethernet switched paths based on PBB-TE. Data is encapsulated in MAC-in-MAC frames. Meaning of service provider fields has been altered. For example, the VLAN field now identifies paths with the same origin and destination.

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.

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.

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.

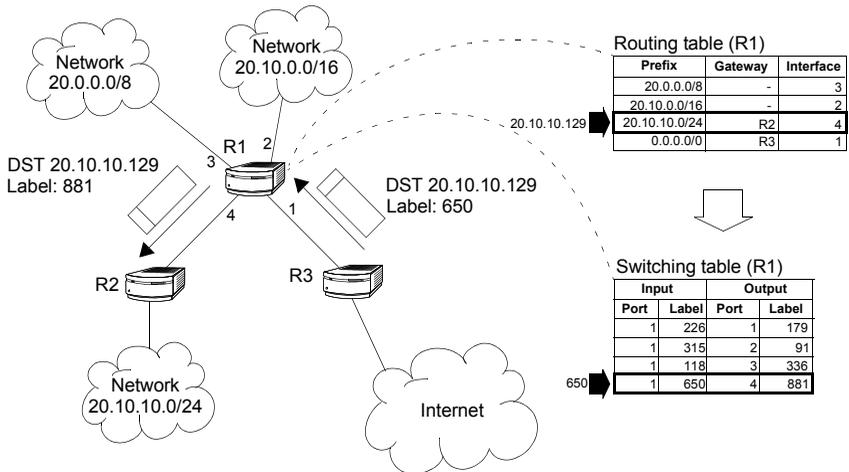


Figure 3.16 Traditional routers have to perform complex operations to resolve the output interface of incoming packets. LSRs resolve the output interface with the help of a simple switching table.

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 3.16).

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.

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 3.17):

- *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

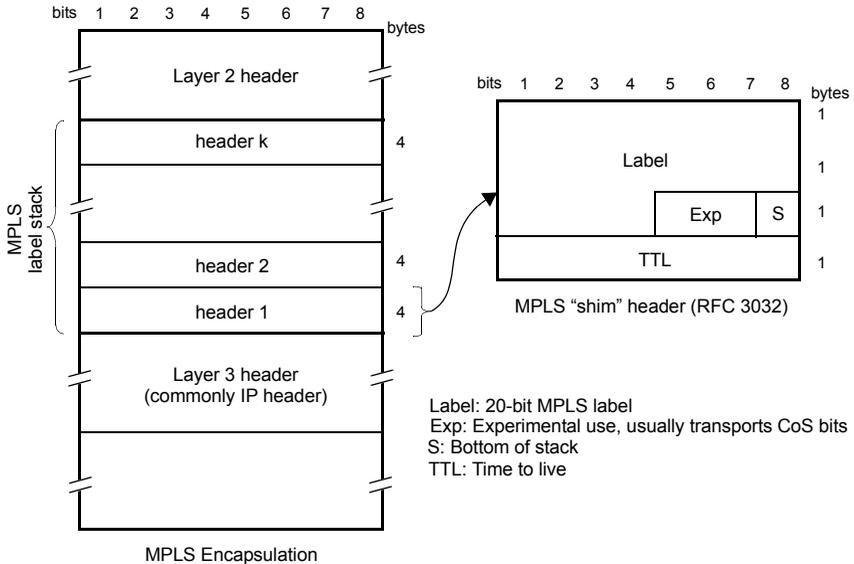


Figure 3.17 MPLS "shim" header format. The label is usually inserted between layer-2 and layer-3 headers.

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 3.18).

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

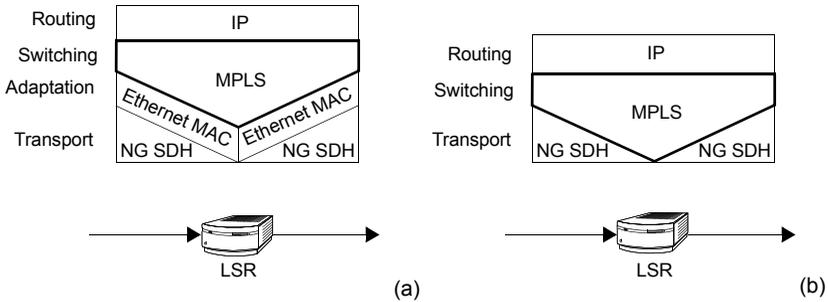


Figure 3.18 Protocol stacks of SDH LSRs. (a) Traditional protocol stack for an SDH LSR: the MPLS header is inserted between layer-2 (Ethernet MAC) and layer-3 (IP) headers. (b) Direct mapping of MPLS over SDH: The MPLS header is mapped between a layer-1 (GFP-F) overhead and layer-3 (IP) overhead without the need of a passive Ethernet encapsulation only used for adaptation.

an MPLS label. The FR 10-bit *Data Link Connection 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).

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 3.19). Only input labels are used for forwarding the packets within the network, while encapsulated addresses like IP or MAC are completely ignored.

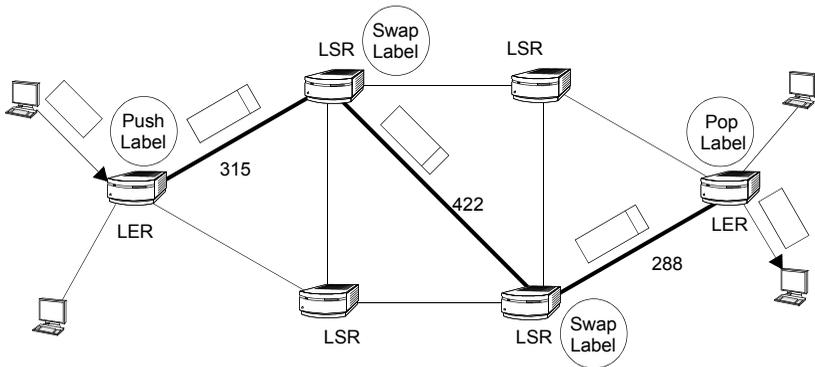


Figure 3.19 Label processing within an MPLS domain. A label is pushed by the ingressing LER, swapped by the intermediate LSR across the LSP, and popped by the egressing LER.

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 3.20):

- *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.

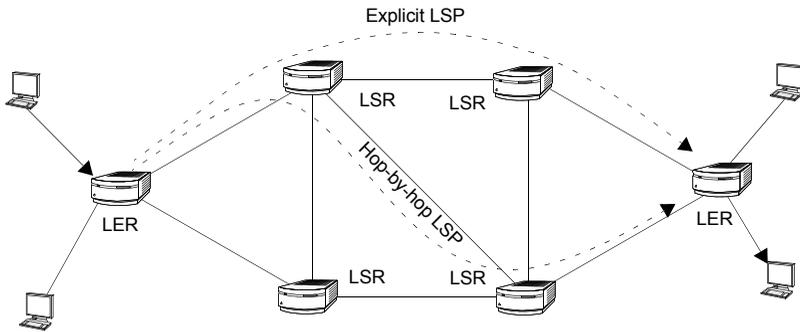


Figure 3.20 A hop-by-hop LSP and an explicit LSP between the same source and destination. The hop-by-hop LSP is computed by the routing protocols running in the LSRs. The explicit route is computed by an external *Network Management System* (NMS).

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.

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.

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 unicasted to a specific LSR or multicasted 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 3.21):

- 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 ex-
-

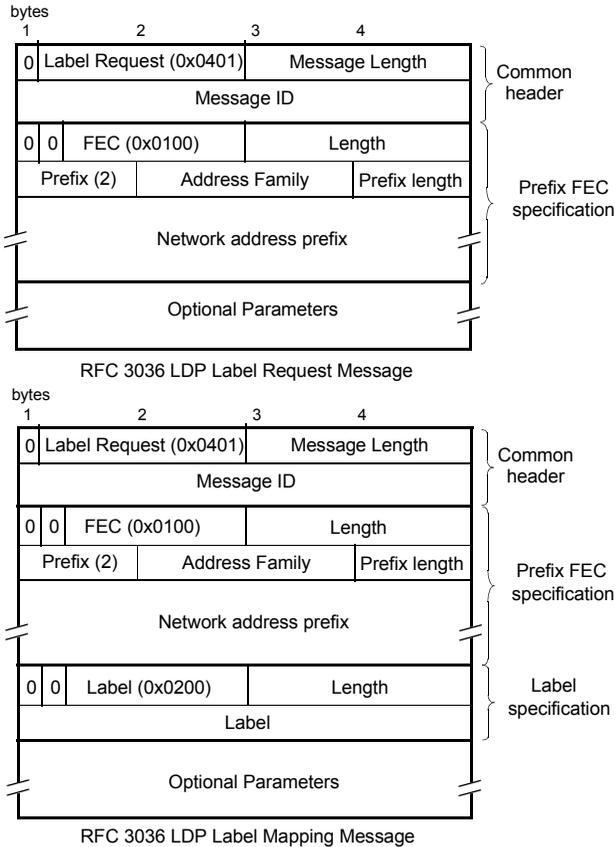


Figure 3.21 Two important LDP messages. The *Label Request Message* requests a label from a remote LSR for binding with an FEC that is attached to the message. The *Label Mapping Message* is used to inform a remote LSR on which label to use for a specific FEC.

ample immediately after the request arrives. This mode can be compared to the *Address Resolution Protocol* (ARP) used in LANs

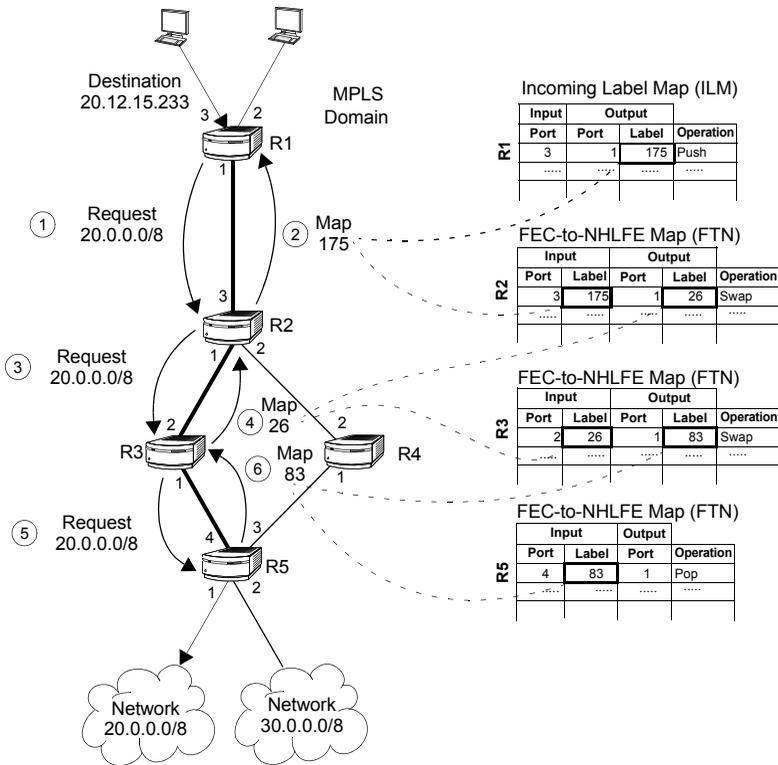


Figure 3.22 LDP in Downstream-on-Demand and independent label distribution mode. LSRs in the LSP generate label requests. The replies they receive from LSRs upstream are used to fill their switching tables.

to request mappings between destination IP addresses and MAC addresses.

- *Ordered label distribution control* (see Figure 3.22): 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.

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 3.23):

- 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.

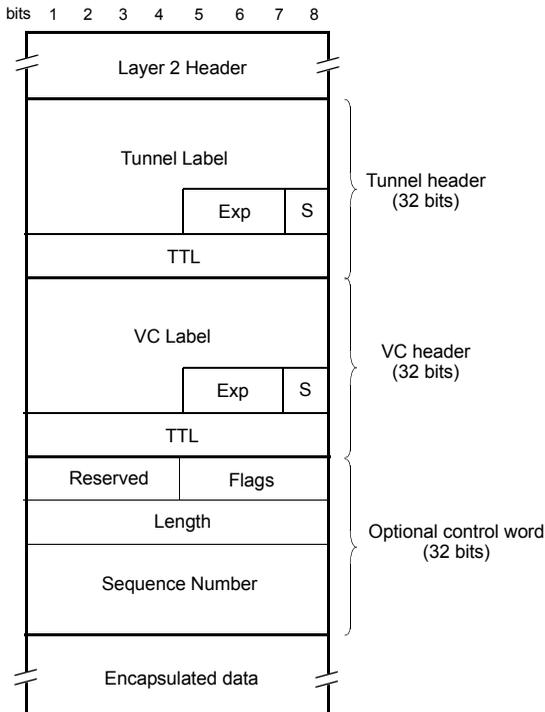


Figure 3.23 Two-label MPLS stack with a Tunnel label and a VC label. The control word may be required when carrying non-IP traffic.

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 3.24). 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.

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

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.

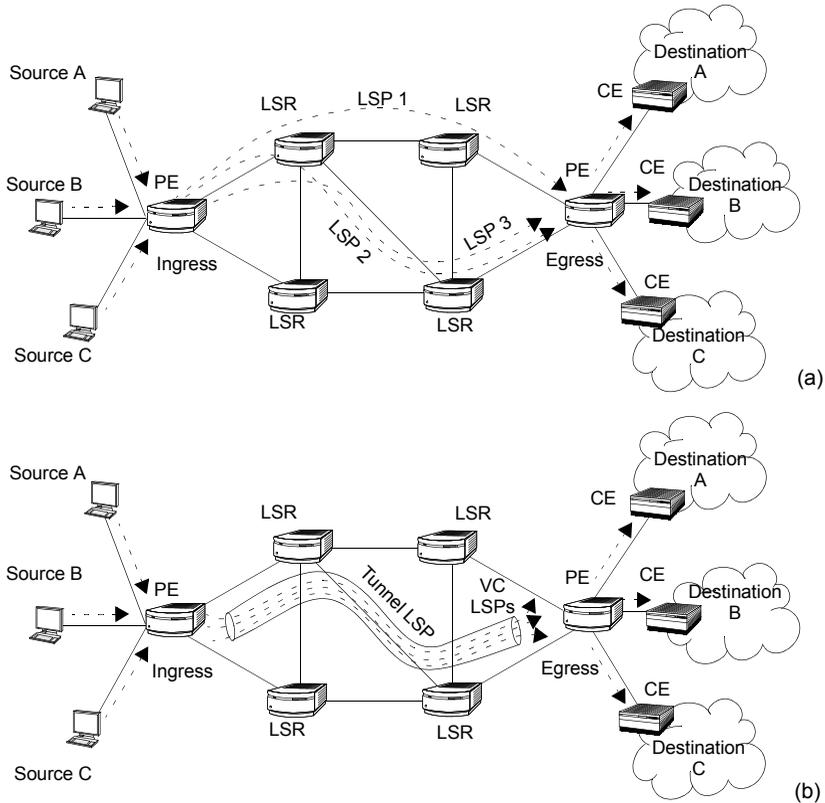


Figure 3.24 (a) One-label approach: the decision to establish routing and edge-to-edge sessions is shared between all the routers. (b) Two-label model: edge-to-edge sessions are tunneled, and internal LSRs are unaware of them.

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.

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 3.25).

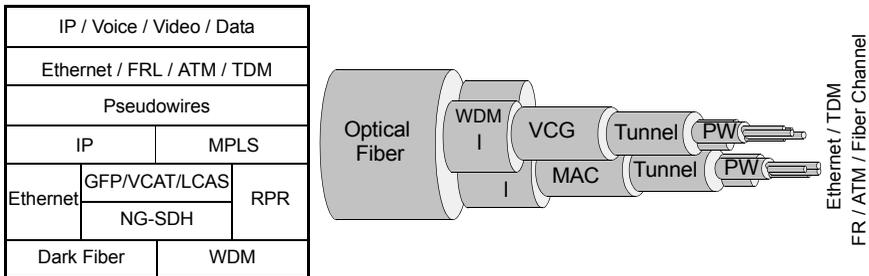


Figure 3.25 Pseudowires can encapsulate and transport ATM, FR, Ethernet, PPP, TDM or Fiber Channel, which is why these protocols do not need a dedicated network unifying the transport in one single network and interface.

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 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 3.26). 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 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 3.27).

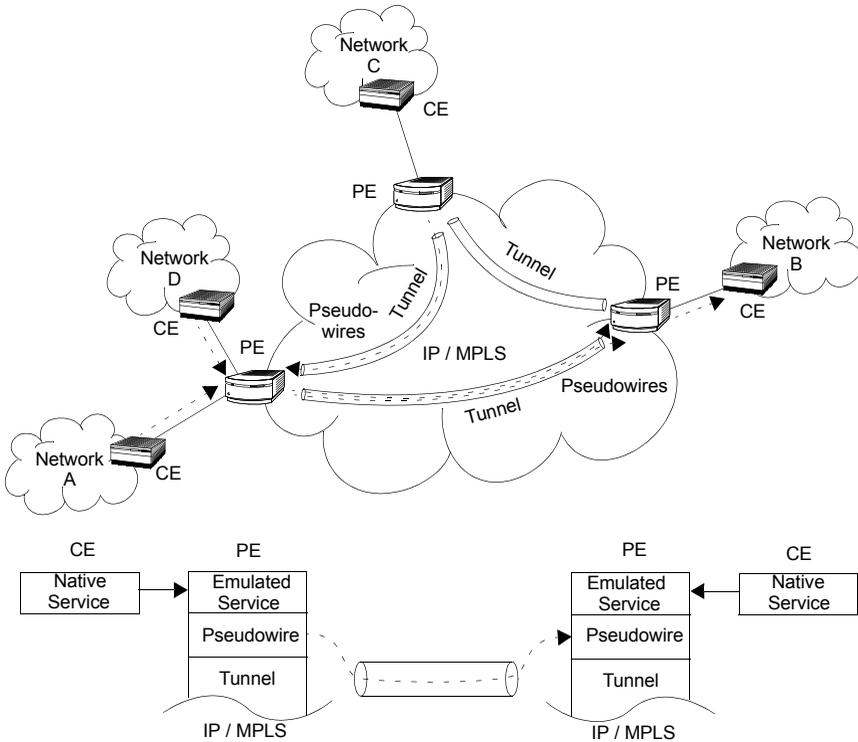


Figure 3.26 Emulation of connectivity services over pseudowires and tunneling across an IP / MPLS network.

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 3.2).

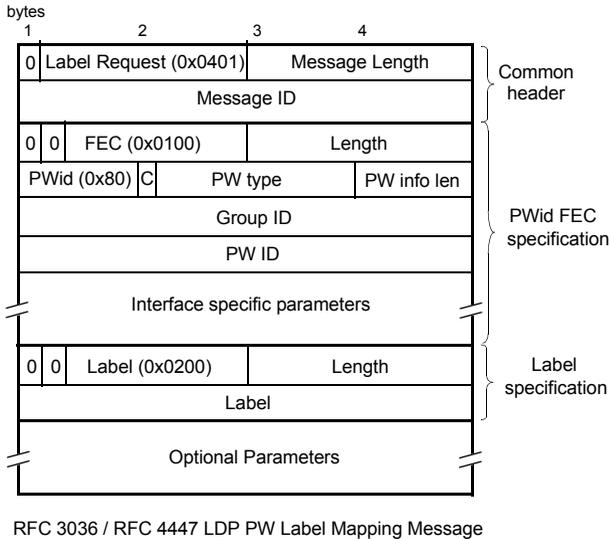


Figure 3.27 The LDP Label-Mapping message as used to map a pseudowire to an MPLS label. The label-to-pseudowire binding is done by using the PWid FEC element specified in RFC 4447.

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 3.23). The presence of this control word is sometimes required, other times optional, and occasionally not required 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.

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
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

Table 3.2 The existing types of pseudowire

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 3.28) will be switched.

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 dif-
-

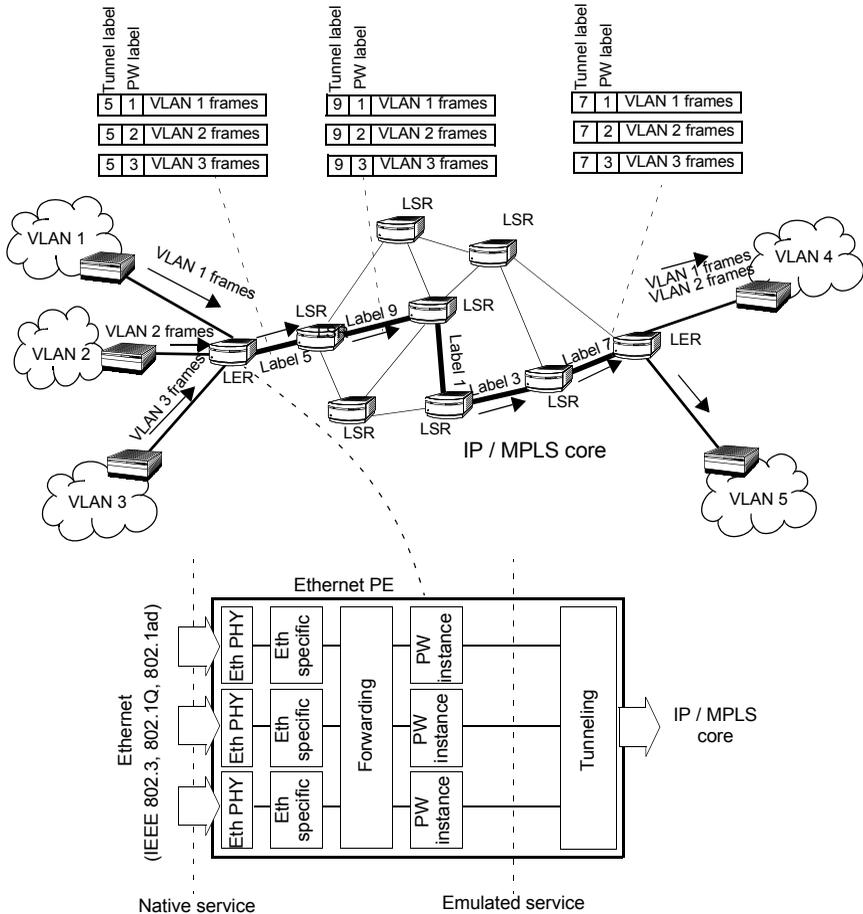


Figure 3.28 Operation of Ethernet pseudowires. The PE router becomes an Ethernet bridge with physical and virtual ports. Physical ports are connected to CEs with standard Ethernet interfaces. Virtual ports are Ethernet pseudowires tunneled across the IP / MPLS core.

ferent 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.

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. In fact, it is likely that this occurs if the pseudowire topology is not closely controlled, because pseudowires are no

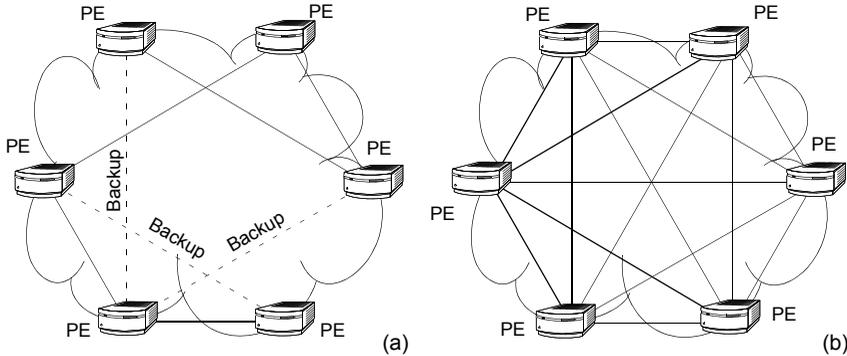


Figure 3.29 Pseudowire topologies in VPLS: (a) Partial mesh with STP. Some of the pseudowires are disabled to avoid loops. (b) Full mesh of pseudowires. The split-horizon rule is applied to avoid bridging loops.

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 3.29). 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 VLPS mesh. Relaying would anyway be unnecessary because there is a direct connection with every possible destination.

To understand how VPLS works we can think of two end users, S and D, who want to communicate to each other (see Figure 3.30). User S wants to send a MAC frame to user D across a shared network running VPLS.

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 forward S's frame to LAN E or LAN F. 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 CED 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.

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.

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 3.31). 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.

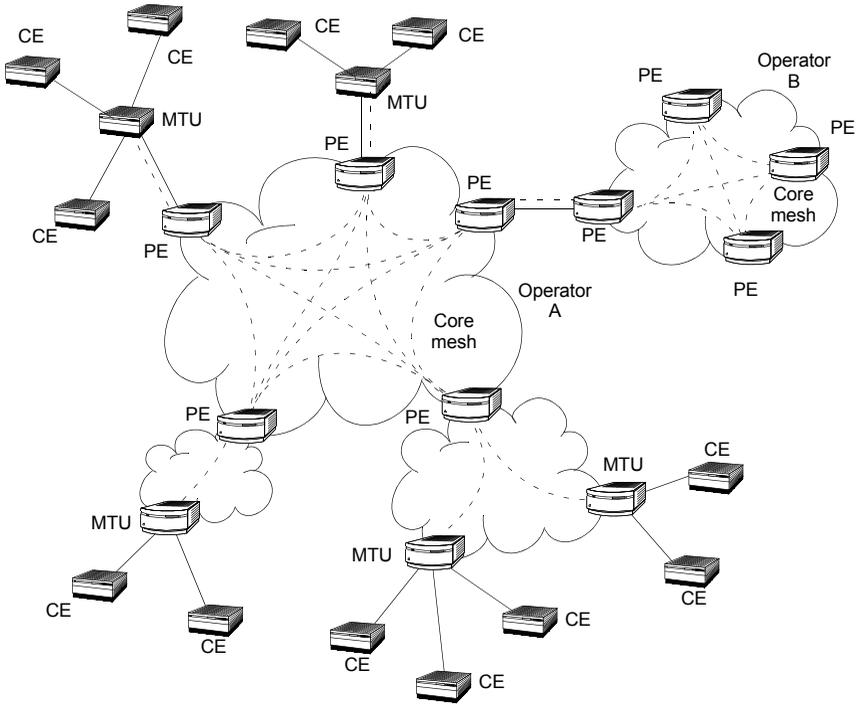


Figure 3.31 In HVPLS, the full mesh of pseudowires is replaced by a two-tier topology with full mesh only in the core and non-redundant point-to-point links in the access.

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.

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.

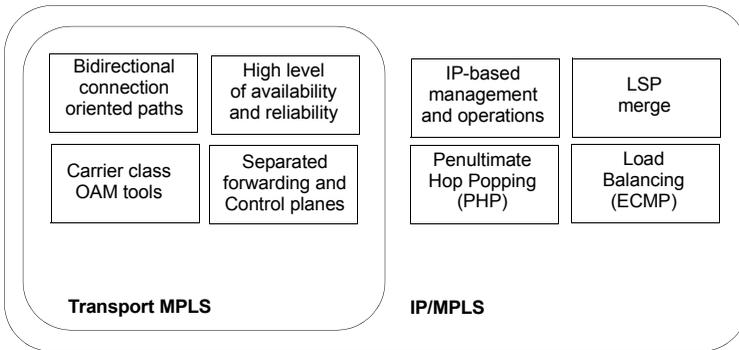


Figure 3.32 MPLS-TP is a strict subset of IP/MPLS. Some IP/MPLS features are left out in MPLS-TP and some other have been defined specifically for MPLS-TP but all them constitute a single MPLS standard.

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 the 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 applications. 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 3.32). These are the major properties of the transport flavour of MPLS not yet mentioned:

- *It is strictly connection-oriented.* In 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 opera-
-

tors 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.
-

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.

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 3.33).

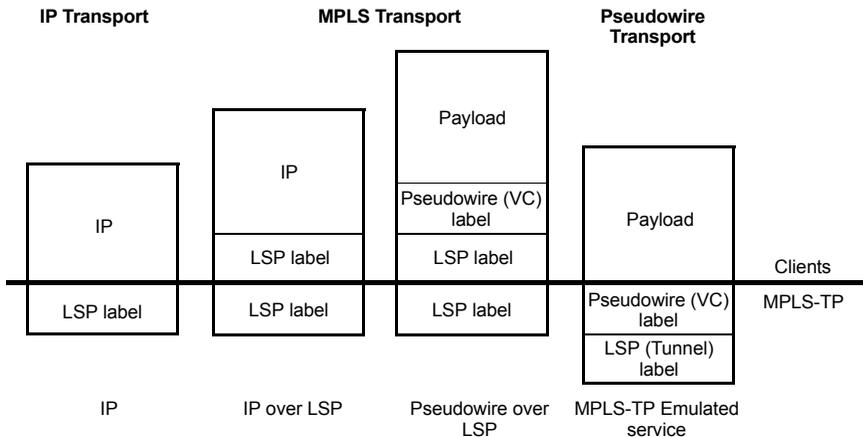


Figure 3.33 MPLS-TP client layers. Any client layer accepted by MPLS is also accepted by MPLS-TP. Pseudowires can be built within the MPLS-TP framework. Thanks to this feature, virtually any client layer can be accepted by MPLS-TP.

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.

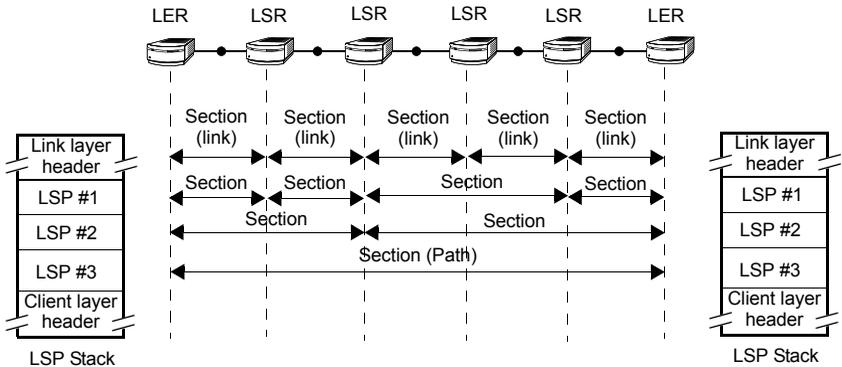


Figure 3.34 MPLS-TP sections are no more than path chunks between adjacent LSRs

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 3.34).

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.

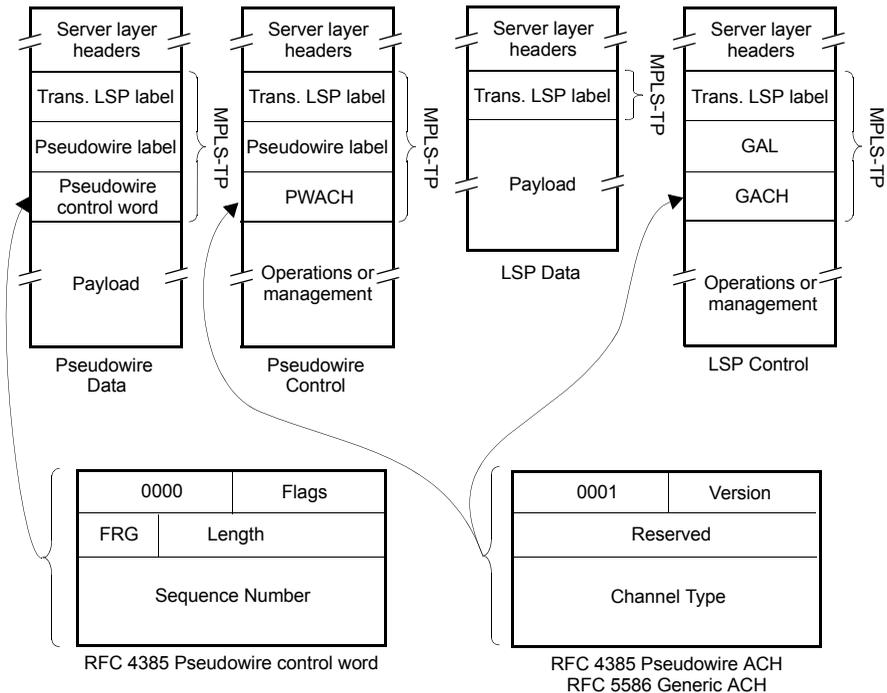


Figure 3.35 MPLS-TP user and GACH PDUs. The packet format for the GACH is inspired in the pseudowire associated channel.

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 to 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 3.35). 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.

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 different 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.

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.

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 3.36).

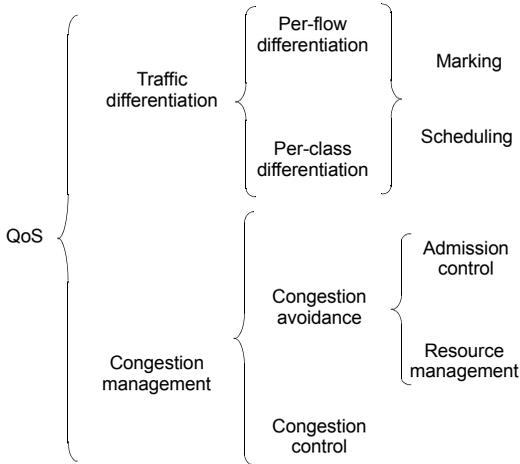


Figure 3.36 It is difficult to achieve good QoS features with one single mechanism. The best way is to mix many elements to get the desired result.

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.

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 3.37b).
- *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 3.37c).
- *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)*.

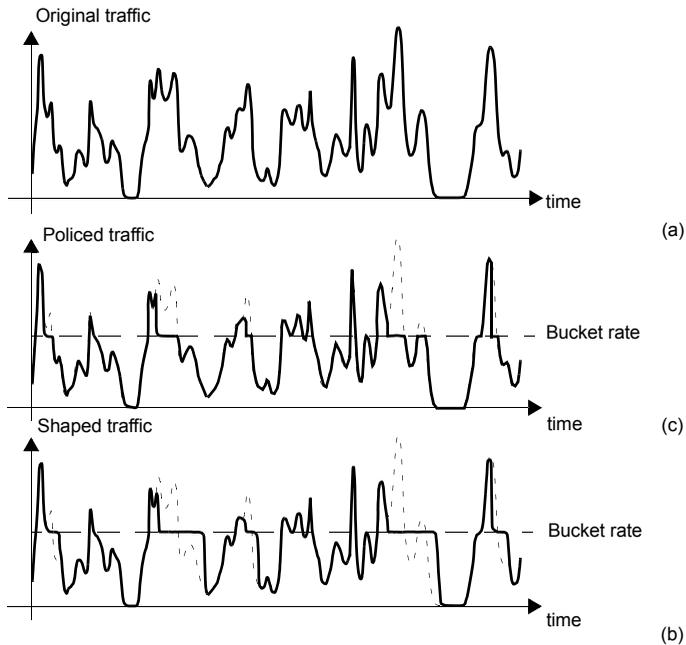


Figure 3.37 Shaping and policing of user traffic. (a) When traffic is shaped, no packets are dropped, but some of them may be delayed. (b) When traffic is policed, it is never delayed, but some packets may be dropped.

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.

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.
- *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 3.38). 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.

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.

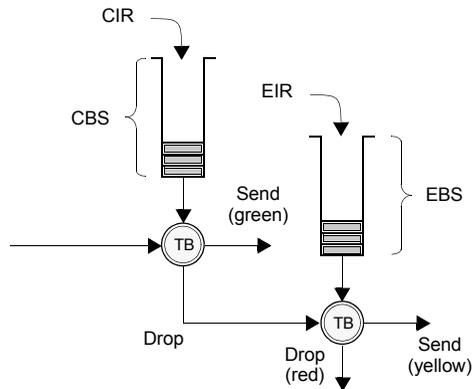


Figure 3.38 Two-rate three-color marker policer

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.

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 3.39).

It is also possible to map the eight possible values of the priority field to *Differentiated Services (DSs)*, *Per Hop Behaviors (PHBs)* such

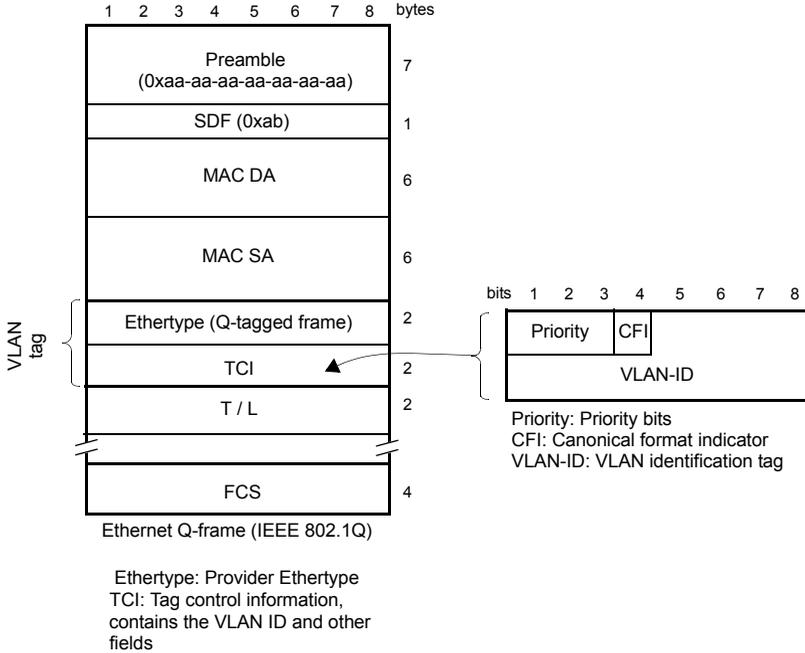


Figure 3.39 The IEEE.802.1Q VLAN frame format enables traffic classification through the three user priority bits.

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.

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 3.40). 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.
 - *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 Protocol (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-config-
-

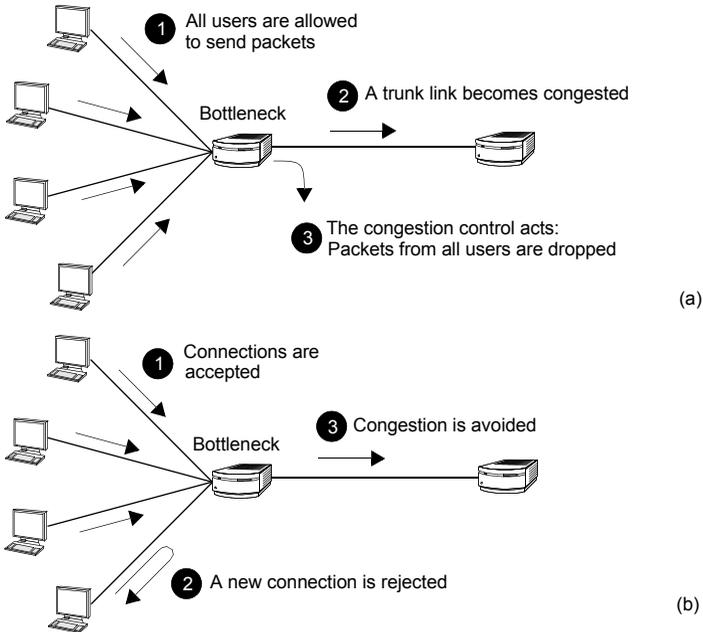


Figure 3.40 How resource management acts: (a) Without resource management, all users experience degradation on their applications whenever there is congestion in the network. (b) If congestion management is used, only some subscribers are not allowed to send data, but the others are not affected.

ured by bridging, but set by a control plane separated from the forwarding plane.

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 prevent 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.

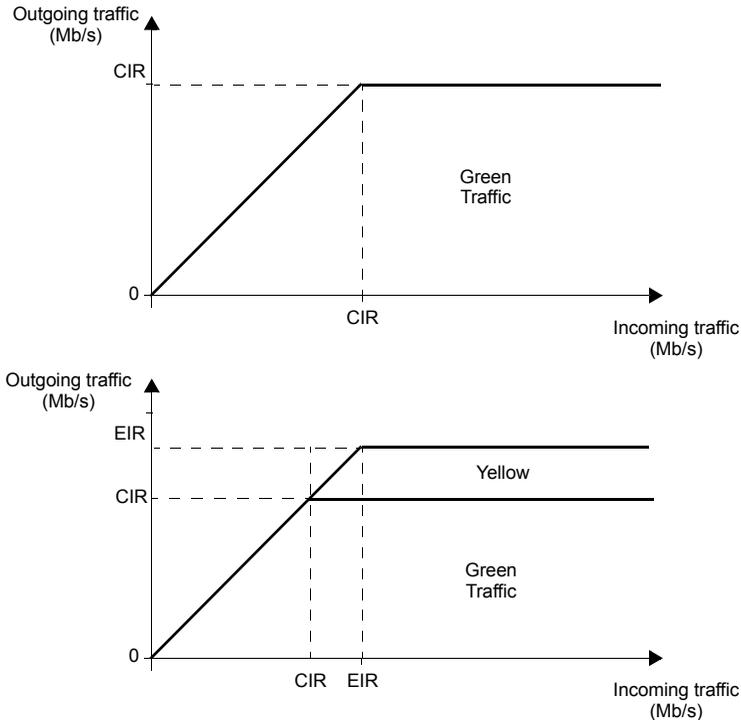


Figure 3.41 The amount of traffic that crosses an admission control filter. Graphics represent steady states, traffic is usually allowed to be greater than the CIR and EIR for short periods of time. (a) The CIR is equal to the EIR, the network guarantees traffic delivery if incoming traffic is smaller 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 low priority and usually discarded first if congestion occurs.

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 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:

$$CBS = ICBS - CIR \times T_{CBS}$$

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 $CIR \times T_{CBS}$ 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.

In a typical test setup for LAN environments (see Figure 3.42), 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

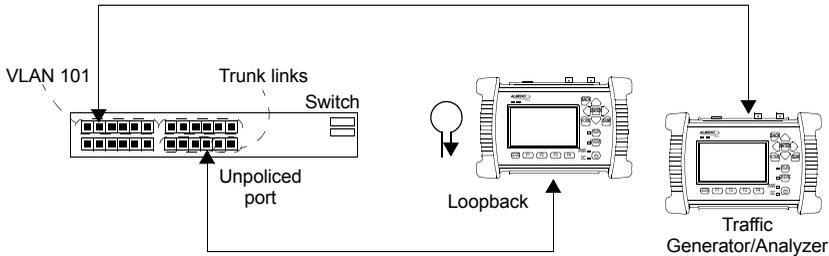


Figure 3.42 In this test, traffic is delivered through an IEEE 802.3 interface to a device connected to an IEEE 802.1Q interface.

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, VIDs or user priority bits.

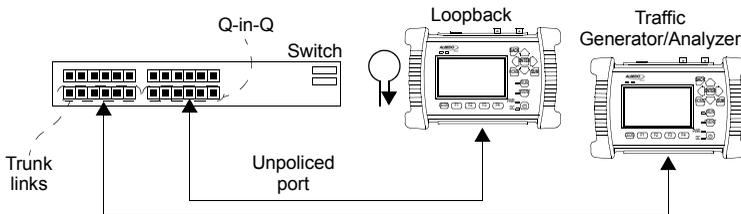


Figure 3.43 In this test, traffic is delivered through an IEEE 802.1Q interface to an IEEE 802.1ad (Q-in-Q). This is a very typical situation in a service provider network.

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.

Class of Service Labels

IP CoS labels are defined either by the ToS labels or the DS code points (see Figure 3.44). 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 3.3).

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

Table 3.3 Meaning of ToS bits.

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 3.4).

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

Table 3.4 Precedence bits and their meaning

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 packets 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.

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.

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

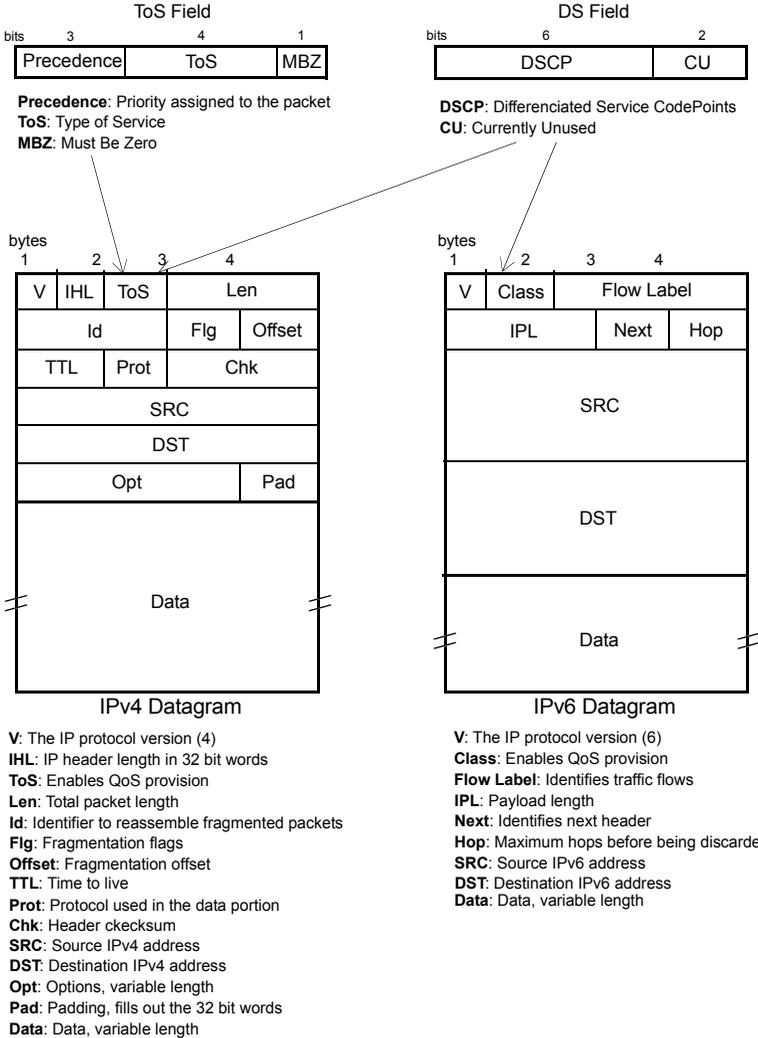


Figure 3.44 IPv4 and IPv6 datagrams and the format of the ToS and DS fields, both related to QoS provisioning.

from source to destination. This delay is the sum of delays on each link and node crossed by the packet (Figure 3.45).

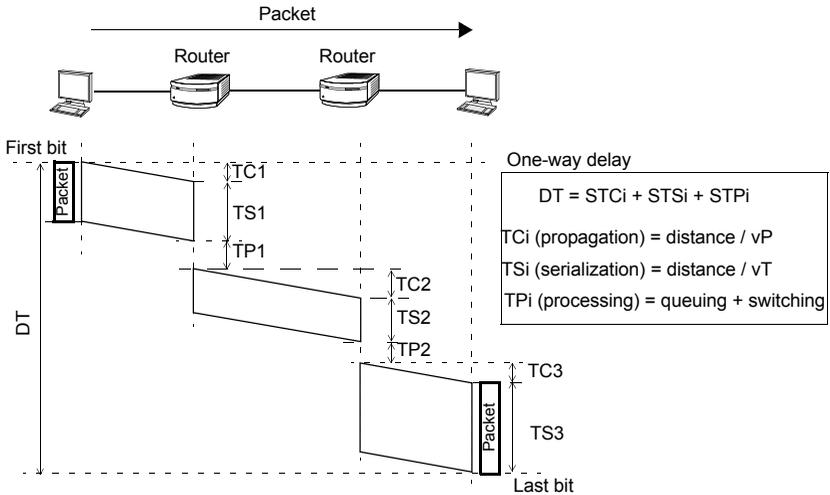


Figure 3.45 One-way delay is the sum of delays on each link and node crossed by a frame.

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 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.
- *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.

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 3.46). 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.

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 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.

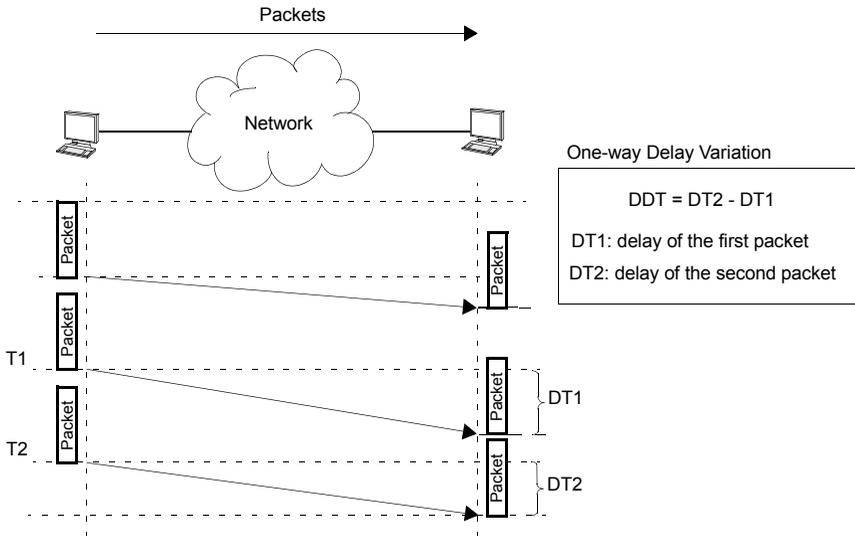


Figure 3.46 One-way delay variation: measurement and impact on data periodicity

- *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.

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.

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

QoS Class	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

Table 3.5 ITU-T Y.1541 Network Performance Objectives.

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 3.47), 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.

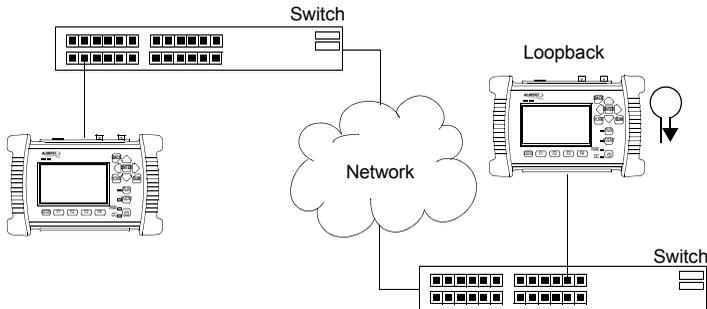


Figure 3.47 Simple QoS test setup. A traffic generator/analyzer and a loopback device are connected to remote devices. The traffic crosses the network in two directions. The traffic generator/analyzer collects statistics on the test traffic.

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 defined 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 3.5). 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.

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

Table 3.6
VoIP Service Degradation under Different Transmission Conditions

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 3.6) 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.

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}

Table 3.7
Digital Television Loss/Error Ratio Requirements

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 3.7). 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 3.8).

	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 %

Table 3.8

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

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 3.48).

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).

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
-

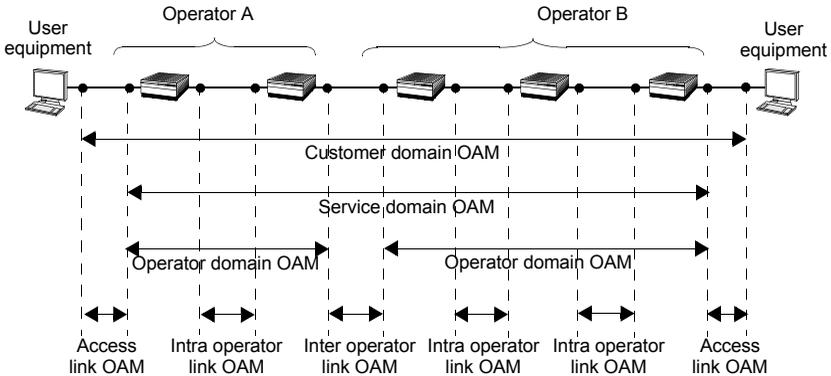


Figure 3.48 Switches forward transparently frames from higher domains.

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 3.49). 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:
 - 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.
-

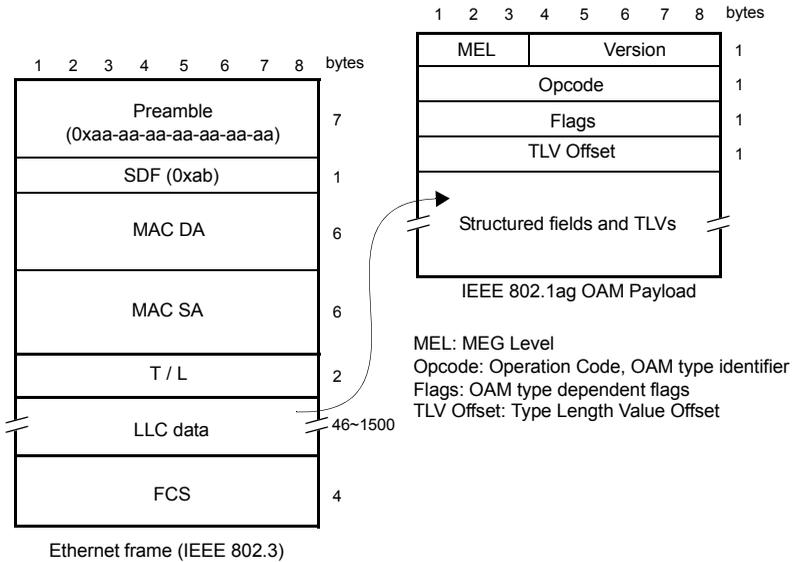


Figure 3.49 Structure of the IEEE 802.1ag PDU and its mapping in IEEE 802.3 Ethernet frames.

- 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:

- Fault management* enables detection and notification of different defect conditions.
- Performance monitoring* allows measurement of performance parameters such as packet loss, delay and delay variation.

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 3.9).

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

Table 3.9
Ethernet Fault Management OAM functions

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 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).

Performance Monitoring

The ITU-T Recommendation Y.1731 defines network performance parameters along with the necessary OAM functions to compute

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.

Table 3.10
Ethernet Performance Monitoring OAM functions

these parameters in real environments (see Table 3.10). 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 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 Reply (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.

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 be-
-

tween 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 functions 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

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 3.50). The UDP port used by the MPLS echo request / reply is the 3503.

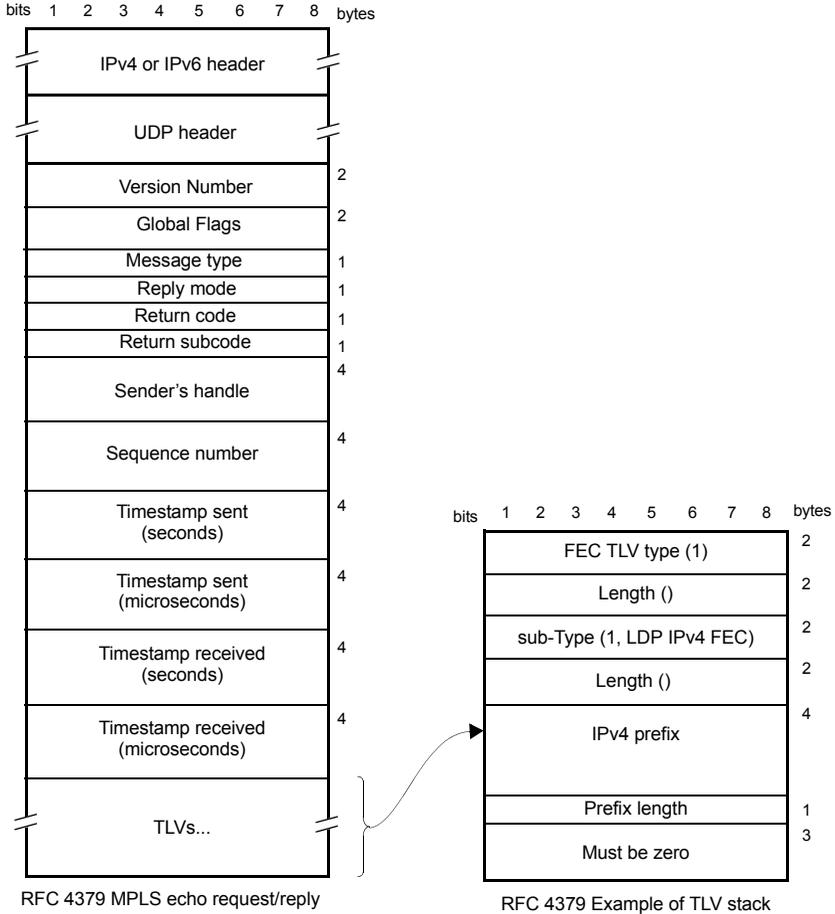


Figure 3.50 The MPLS ping and trace router tools are based on a MPLS echo request and reply messages. These messages are UDP packets encapsulated in IPv4 or IPv6 datagrams and probably labeled by one or more MPLS labels.

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 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.

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 IPv4 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 3.51). Another interesting BFD feature is the ability to authenticate the protocol packets in order to make sure they come from the correct source.

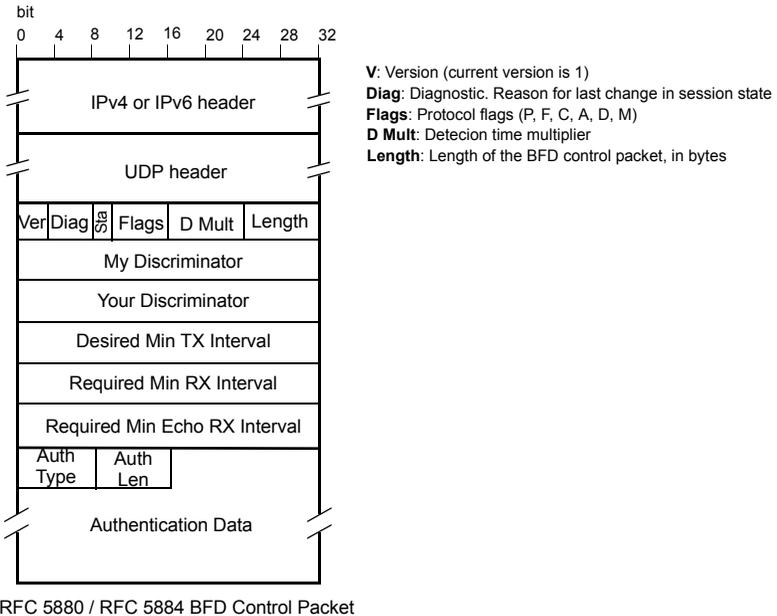


Figure 3.51 BFD control packet format

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.

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 (and 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 3.52):

- *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 con-
-

control packets may follow different paths if a load balancing mechanism like the Equal Cost Multi-Path (ECMP) is used.

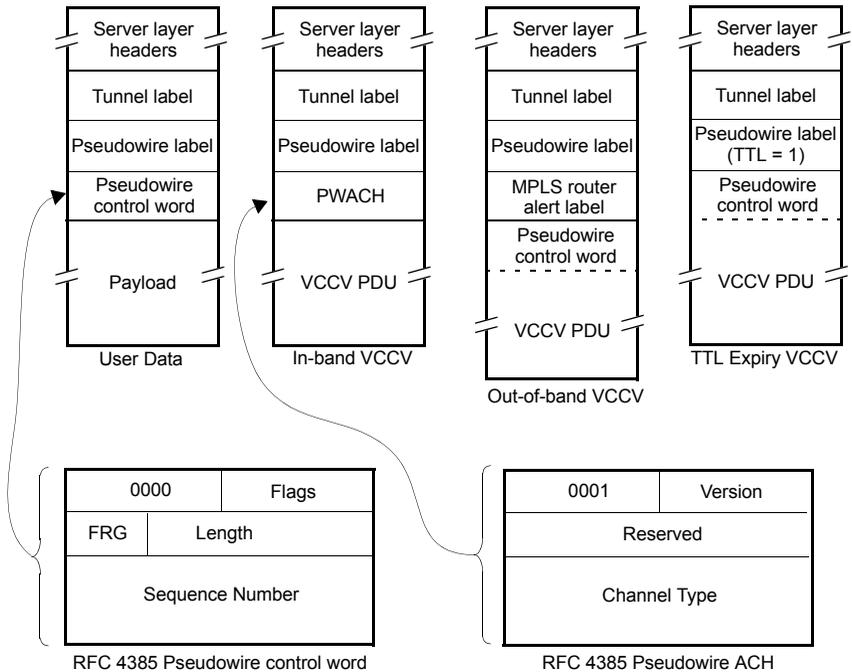


Figure 3.52 In-band and out-of-band multiplexing of pseudowire data plane information and VCCV.

- **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.

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 3.11).

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 3.53).

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.

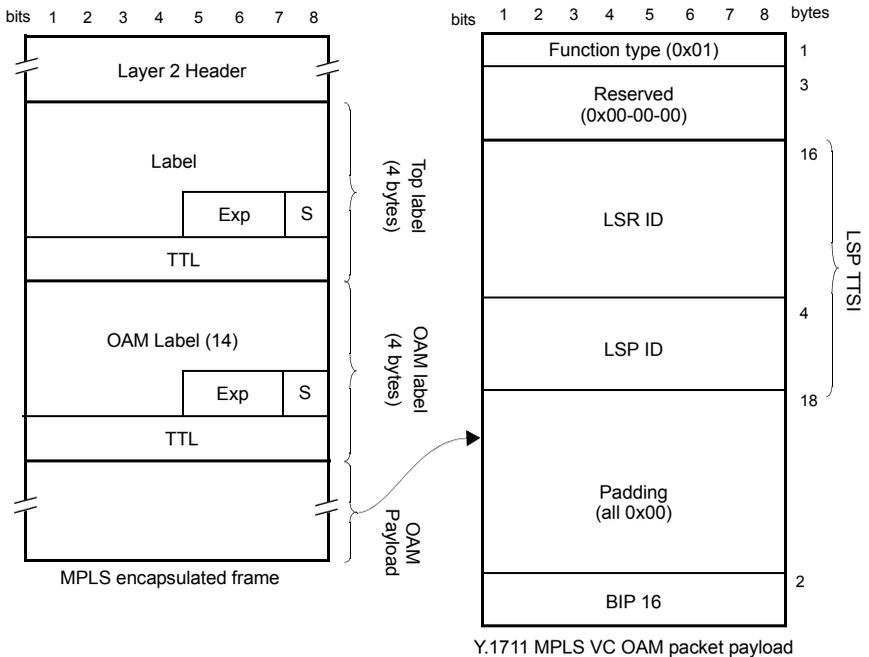


Figure 3.53 The structure of the Y.1711 OAM channel is based on double labeled frames. The label value used for OAM is 14.

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.

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.

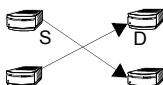
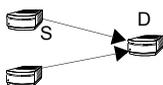
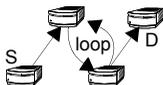
Defect	Codepoint	Diagram	Description
dLOCV	0x02-01		No CV packets are received in the LSP. This defect can be caused if the LSP is broken due to a configuration problem, degraded transmission medium or a broken LSR.
dTTSI_Mismatch	0x02-02		Unexpected TTSI found by the egress LER. This is caused by an LSP misconnection.
dTTSI_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-mismerging or Denial of Service (DoS) attacks.

Table 3.11
ITU-T Y.1711 MPLS layer defects

MPLS-TP OAM

Extensive OAM is a key requirement for MPLS-TP. Definition of OAM mechanisms for MPLS-TP is a work in progress within the IETF. 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.

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 3.54). The MPLS-TP OAM framework also provides support for maintenance of arbitrary LSP and pseudowire parts.

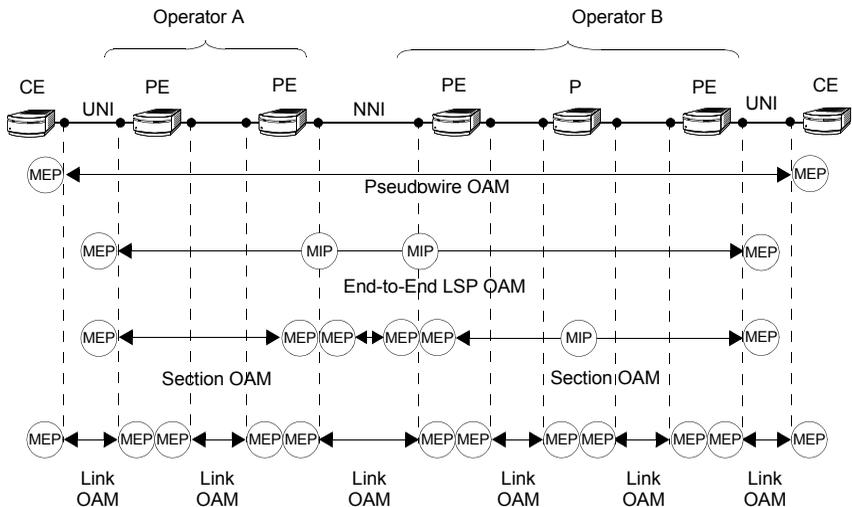


Figure 3.54 The MPLS-TP OAM framework defines different MEPs and MIPs operating at pseudowire, LSP and section levels.

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 3.55).

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

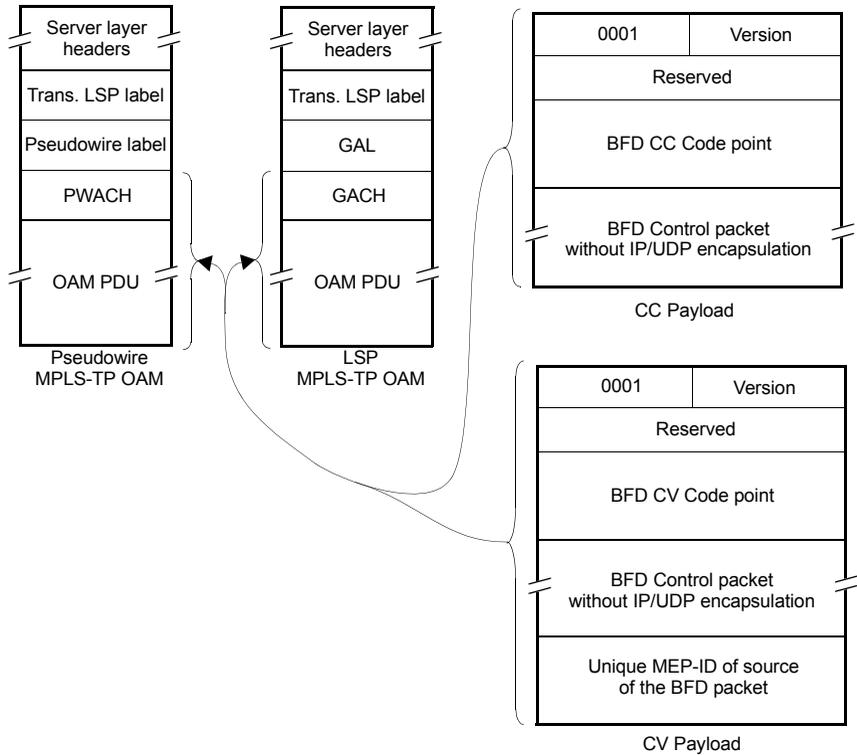


Figure 3.55 Projected MPLS-TP messages for proactive monitoring. These messages constitute the CC and CV flows and they are based on the BFD mechanism.

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.

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. Freddete 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., Wroclawsky J., "General Characterization Parameters for Integrated Service Network Elements", IETF Request For Comments RFC 2215, September 1997.
-

Ethernet in Access Networks

Chapter 4

Ethernet in Access Networks

The standard IEEE 802.3ah for *Ethernet in the First Mile* (EFM) was released with the aim of extending Ethernet to the local loop for both residential and business customers.

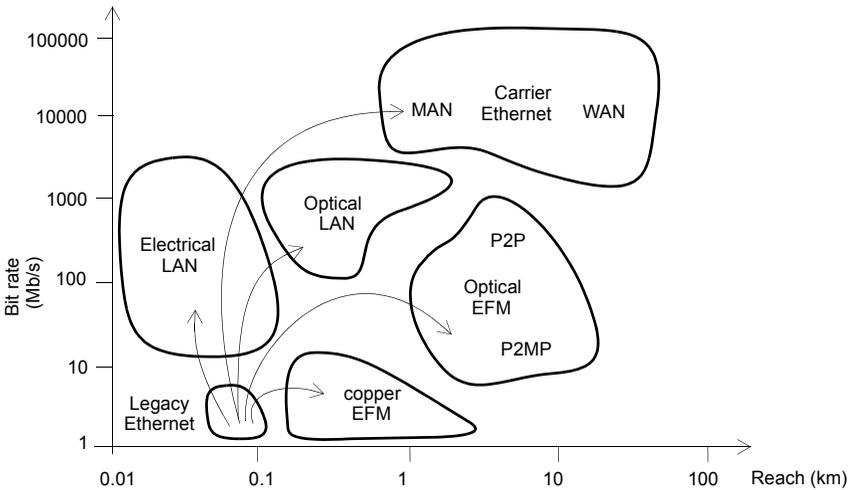


Figure 4.1 Ethernet applications and EFM

EFM interfaces provide low and medium speeds when compared with the available LAN or WAN standards (see Figure 4.1). The new interfaces, however, are optimized to be profitable in the existing and newly installed provider access networks. The copper EFM takes advantage of DSL technology for telephone copper pairs, and optical EFM is suitable for both PON networks and active Ethernet (see Table 4.1).

Interface	Medium	Wavelength (nm)	Rate (Mb/s)	Reach (km)
100BASE-LX10	Two SMFs	1310	100	10
100BASE-BX10	One SMFs	1310 (US), 1550 (DS)	100	10
1000BASE-LX10	Two SMF	1310	1000	10
1000BASE-LX10	Two MMF	1310	1000	0.55
1000BASE-BX10	One SMF	1310 (US), 1490 (DS)	1000	10
1000BASE-PX10	One SMF PON	1310 (US), 1490 (DS)	1000	10
1000BASE-PX20	One SMF PON	1310 (US), 1490 (DS)	1000	20
10GBASE-PR10	One SMF PON	1270 (US), 1577 (DS)	10000	10
10GBASE-PR20	One SMF PON	1270 (US), 1577 (DS)	10000	20
10GBASE-PR30	One SMF PON	1270 (US), 1577 (DS)	10000	30
10/1GBASE-PRX10	One SMF PON	1310 (US), 1577 (DS)	1G/10G	10
10/1GBASE-PRX20	One SMF PON	1310 (US), 1577 (DS)	1G/10G	20
10/1GBASE-PRX30	One SMF PON	1310 (US), 1577 (DS)	1G/10G	30
10PASS-TS	One or more telephone pairs	-	10	0.75
2BASE-TL	One or more telephone pairs	-	2	2.7

Table 4.1
IEEE 802.3ah and 802.3av Interface Summary

Fiber to the Neighborhood

Deployment of bandwidth demanding applications like IPTV is pushing network operators to upgrade their copper based access infrastructure. This is the reason why some operators have already started to deploy new access networks based on optical fiber. However, only a few of these deployments offer *Fiber To The Home* (FTTH). Most of them are (depending on where the optical link is terminated) *Fiber To The Building* (FTTB), *Fiber To The Cabinet* (FTTCab), etc.

Currently, there are many different options for FTTx. Electrical links can be built with DSL or Ethernet. The ITU-T Recommendation G.993.1 defines the *Very-high-bit-rate DSL* (VDSL), a DSL type designed for FTTCab and FTTB architectures. The VDSL technology offers downstream bit rates around 50 Mb/s within the range of

300 meters. VDSL has been improved in the new ITU-T Recommendation G.993.2. This new technology is known as VDSL2, and it delivers symmetrical 100 Mb/s bit rate within the range of 300 m. In FTTB architectures, the access network operator may choose to deploy Ethernet over *Unshielded Twisted Pair* (UTP) cable, if cable lengths are shorter than 100 m. The IEEE 802.3 100BASE-T and 1000BASE-T are likely to be the chosen interfaces. 100BASE-T offers 100 Mb/s of symmetrical bit rate, and 1000BASE-T 1 Gb/s of symmetrical bit rate. The range is limited to 100 m for both.

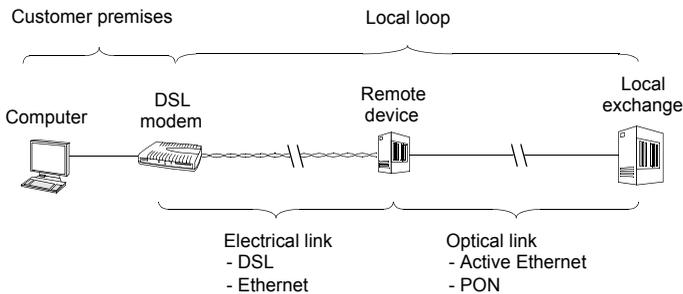


Figure 4.2 FTTx architecture for the local loop

Active Ethernet and *Passive Optical Network* (PON) are the main options for the optical portion of the local loop (see Figure 4.3):

- Active Ethernet is made up of point-to-point fiber links between the local exchange and the customer premises. This means that large quantities of optical fiber must be used in the local loop, and this is expensive. However, the use of dedicated fiber links guarantees maximum bandwidth. To reduce the amount of fiber, an Ethernet switch can be installed close to the subscriber, and it acts as a concentrator. Between the switch and the local ex-

change, it is enough to install a single optical link, or maybe two for redundancy.

- PON has been proposed to avoid installing active elements, such as Ethernet concentrators, in the local loop. Active elements are replaced by simple passive optical splitters, giving as a result a point-to-multipoint topology. PON can be used to offer gigabit-level bandwidth to subscribers. This technology is considered more cost effective than active Ethernet, and at the same time it is well suited for applications like TV that can be overlapped with data on a different wavelength. The main drawback is the need for complex shared-media access mechanisms to avoid collisions between the traffic of different subscribers.

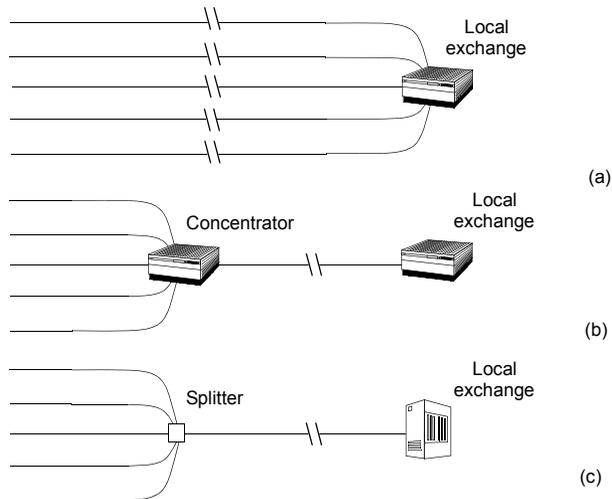


Figure 4.3 Optical fiber installation in the local loop: (a) The point-to-point topology needs a large amount of fiber. (b) With active Ethernet, less fiber is needed, because a switch can be placed close to the subscribers. (c) The PON solution replaces the switch with an inexpensive and passive optical splitter.

Ethernet over Telephone Copper Pairs

The EFM standard defines two interfaces for Ethernet transmission over telephone copper pairs:

- The 2BASE-TL interface is best suited to long-haul applications. It provides a symmetric, full-duplex 2-Mb/s Ethernet transmission channel with a nominal reach of 2.7 km. It is based on SHDSL as per ITU-T G.991.2. The 2BASE-TL interface is optimized for local exchange applications.
- The 10PASS-TS interface is intended for short-haul applications. It offers a symmetric, full-duplex 10 Mb/s transmission with a nominal reach of 750 m. It is based on the VDSL (ANSI T1.424) technology and optimized for deep fiber roll-outs like FTTB or FTTCab. It can be combined with EPON or active Ethernet to offer a simple bridged access network. The 10PASS-TS interface is compatible with baseband transmission of analogue voice.

The 10PASS-TS and 2BASE-TL are mostly based in existing technology, such as SHDSL and VDSL, mainly for the following reasons:

- Extensive DSL deployments exist and have existed for the past 10 years or so. DSL is a well-known technology, and network operators have a lot of experience with it.
- DSL has proven to be efficient, cost-effective and easy to deploy.
- National-level spectrum compatibility standards make it difficult to introduce signals with new spectrum shapes.

One of the challenges of Ethernet over copper is the lack of a strict definition of what is understood by a voice-grade copper pair. The reason for this is that telephone cabling started in the 19th century, much before any telecommunication regulations. Most of the current telephone pairs fall into the TIA / EIA categories 1 and 3. Unlike other Ethernet standards, 2BASE-TL and 10PASS-TS are not specified for a transmission media of known features, and therefore

the performance of these interfaces remains largely unpredictable in untested cables.

One of the few changes introduced by the IEEE in the DSL specifications was the encapsulation defined for Ethernet. The original ITU-T encapsulation was based on an HDLC framing but Copper EFM uses the new 64/65-octet encapsulation.

Another important feature of the EFM interface for copper is the *bonding function*. This feature is useful in providing Ethernet services over copper without the severe distance limitations.

Ethernet in Optical Access Networks

Optical EFM interfaces provide better performance than copper EFM in terms of reach and bit rate, but they require optical fiber. These interfaces have been especially developed for deep-fiber rollouts based on *Point-to-Point* (P2P) and *Point-to-MultiPoint* (P2MP) architectures.

In the case of the P2MP architecture, the EFM interface offers EPON. For P2P, the EFM adapts the available Ethernet interfaces so that they operate in the access network. For example, bidirectional interfaces take advantage of the WDM technology to duplex the upstream and downstream in a single fiber. This makes it unnecessary to install two fibers per customer. Extended temperature operation is another improvement important for external plant operation.

The most extended EPON interfaces provide 1 Gb/s symmetrical capacity, typically to be shared by 16 subscribers. This means that the minimum guaranteed bandwidth in FTTH is around 60 Mb/s per subscriber, but depending on the network load, it could increase up to several hundreds of megabits per second. The P2P interfaces for active Ethernet roll-outs provide 100 Mb/s per customer in FTTH. Gigabit interfaces also exist, but these are

typically used for backhaul in fiber-to-the-neighborhood applications, or they may be combined with copper in FTTB deployments.

The Need of an Optical Access Network

Copper access networks alone cannot meet the challenges of future broadband applications such as HDTV that may need up to 100 Mb/s. Copper technologies like DSL cannot provide long range and high transmission rate simultaneously (see Figure 4.4).

DSL depends on the telephone wires on which it operates, and this means that this technology has some limitations. DSL signals have to suffer many impairments in a transmission channel that was not originally designed to carry them. Two of these limitations are critical:

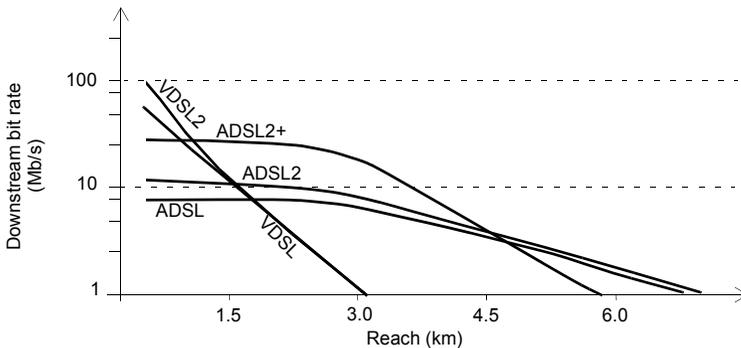


Figure 4.4 Approximate reach achieved with different DSL technologies

1. *Attenuation* is caused by progressive loss of the electrical energy of the DSL signal in the transmission line. Attenuation is higher in longer loops, and it also depends on the frequency of the signal being transmitted. The higher the frequency band used for trans-

mission, the more attenuation the signal will suffer.

2. *Crosstalk* is the electromagnetic coupling between transmission lines that are close to one another. In the access network, copper pairs are grouped into binders. One binder may contain dozens or even hundreds of copper pairs, and this is why they are vulnerable to crosstalk.

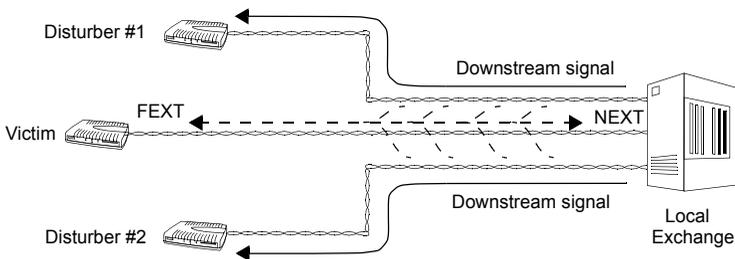


Figure 4.5 Crosstalk between copper pairs. Signals from disturbing lines are coupled to the victim line, damaging communication.

Crosstalk control has special relevance. After local loop unbundling took place, the number of signals in the loop started to increase, and crosstalk from some local loop signals could potentially damage other operator's service. Due to this, the spectral compatibility between copper access technologies had to be studied, and new (national) regulation had to be developed to control the management of the copper loop spectrum.

1Gb/s and 10 Gb/s Ethernet PON

The Ethernet PON or EPON is the IEEE alternative for PON. The first version of EPON was released in 2004. EPON is based on Ethernet, the most successful networking technology specified by the IEEE. In fact, EPON is part of the EFM initiative that attempts to extend

Ethernet to the local loop. EPON is a direct competitor of the GPON technology defined by the ITU-T.

The IEEE 802.3ah defines two alternative interfaces for EPON, known as 1000BASE-PX10 and 1000BASE-PX20. The former has a minimum range of 10 km and the latter 20 km. The typical number of ONUs in an EPON is 16, but alternative splitting ratios are also possible. There is a trade-off between range and splitting ratio, because optical loss increases with both distance and split count. This means that more ONUs can be served if the distance between the ONU and the OLT is shorter.

The IEEE 802.3av, released in September 2009, extends the speed of EPON to 10 Gb/s. The IEEE 802.3av standard has good interoperability features with the IEEE 802.3ah GPON due to a smart choice of the transmission wavelengths and an asymmetric 1 Gb/s - 10 Gb/s operation mode.

PON Concepts and Alternatives

The *Passive Optical Network* (PON) is an optical technology for the access network, based only on passive elements such as splitters. In a PON, the transmission medium is shared, and traffic from different stations is multiplexed. Optical transmission increases transmission bandwidth and range dramatically when compared to some copper pair technologies such as DSL. Furthermore, due to the use of simple and inexpensive transmission elements and shared medium, a PON is a cost-effective solution for the optical access network.

The logical deployment alternative enabling optical communications in the local loop is to replace the copper links by optical fiber links, but this requires a lot of fiber. Installing Ethernet switches acting as traffic concentrators near the customer premises requires less fiber, but massive installation of Ethernet switches has the same inconveniences as remote DSLAMs: suitable placement and power supply must be provided. This is one of the reasons why

PON, based only on passive elements that do not need feeding, is a very attractive solution.

Preliminary works on the PON technology date back to the late 1980s, but the first important achievement regarding its standardization did not arrive until 1995. This year, the *Full-Service Access Network* (FSAN) was formed and presented a system specification for *ATM PON* (APON). Later, in 1997, the ITU-T released Recommendation G.983.1 based on the FSAN specification. The APON is known today as *Broadband PON* (BPON) to emphasize that although ATM-based, any broadband service can be provided with this technology.

Since the release of Recommendations G.994.x for *Gigabit PON* (GPON) in 2003, APON / BPON is considered a legacy technology. GPON has been specified with the help of the FSAN, and it provides multigigabit bandwidths at lower costs than BPON, while achieving more efficiency transporting packetized data with the new lightweight *GPON Encapsulation Mode* (GEM). The GEM is based on a concept similar to the *Generic Framing Procedure* (GFP), a successful encapsulation for mapping packets in SDH networks.

	APON / BPON	GPON / XGPON	EPON
Downstream rates (Mb/s)	155, 622	1244, 2488, 9953	1000, 10000
Upstream rates (Mb/s)	155, 622, 2488	155, 622, 1244, 2488	1000, 10000
Range (km)	20	20	20
Encapsulation	ATM	GEM / ATM	Ethernet

Table 4.2
PON Technology Comparison

An alternative approach is the *Ethernet PON* (EPON), released in 2004 as a part of the IEEE 802.3ah standard for Ethernet in access networks. The main innovation of EPON is that it encapsulates data in Ethernet MAC frames for transmission. Today, EPON has become a strong competitor for GPON, and there are supporters and deployments for both technologies (see Table 4.2).

Architecture and Operation Fundamentals

The physical properties of passive optical splitters make the distribution of optical signals with PON different from other technologies with a shared access to the transmission medium. Ports in optical splitters do not all have the same properties, and thus the network elements connected to them are different:

- The *Optical Line Termination* (OLT) is connected to the uplink port of the optical splitter. Any signal transmitted from the OLT is broadcast to all the other ports of the splitter.
- The *Optical Network Unit* (ONU) is connected to the ordinary ports of the optical splitter. When signals transmitted from the ONU arrive to the splitter, they are retransmitted to the uplink towards the OLT, but not to other ordinary ports where other ONUs could be connected. This makes direct communication between ONUs impossible.

The OLT constitutes the network side of the PON, and it usually resides in the local exchange. The ONUs are the user side. They can be placed in the customer premises in FTTH roll-outs, but they can also be deployed in cabinets, basements of buildings or other locations close to the subscribers. In cases where the ONU is not directly available to the subscribers, the signal is delivered to them by means of other technologies such as DSL or Ethernet. The ONU in FTTH is sometimes referred to as *Optical Network Termination* (ONT). ONUs serving several end users are known as Multi-Dwelling Units (MDUs).

Signals from two or more ONUs transmitting simultaneously will collide in the uplink, and the OLT will be unable to separate them, unless a bandwidth-sharing mechanism is implemented. WDM appears to be the most natural way to share the transmission media for PON, but it would require either installing tunable lasers in the ONUs or having many different classes of ONUs for transmitting at different wavelengths. The high cost of the first

solution and the complexity of the second one make WDM-PON unfeasible today, but attractive in the future.

Transmission in current PONs is based on TDM rather than on WDM. TDM allows for a single downstream wavelength, but it relies on complex shared-media access algorithms. These algorithms take into account that only communications from an ONU to an OLT, but not between ONUs are possible, and therefore they assign to the OLT controller functions. The OLT decides which ONUs are allowed to transmit, when they are allowed to do it, and how much data are they allowed to transmit upstream. The decisions made by the OLT must avoid collision even in the case of propagation delays, and at the same time they must grant fair bandwidth sharing and high network usage. All the transceivers in the PON must be synchronized to a common time reference in order to work properly. The OLT is the network element that is usually in charge of distributing synchronization.

The downstream of a PON is dedicated, and thus no bandwidth sharing mechanisms need to be implemented. However, the downstream link is a broadcast channel, and information transmitted by the OLT is received by all ONUs even if this information is not addressed to all of them. This has some privacy implications and makes it necessary to encrypt private downstream data.

Advantages

PON offers increased bandwidth and range when compared to DSL. It is also more cost-effective and easier to maintain than active Ethernet. It also has several other advantages, namely:

- PONs are highly transparent, as the optical distribution network only contains layer-1 devices. Virtually any type of service can be built over PONs, either packet, TDM or wavelength-based, or even analogue. Transparency eases migration to new technologies without the need to replace network elements. For example, mi-
-

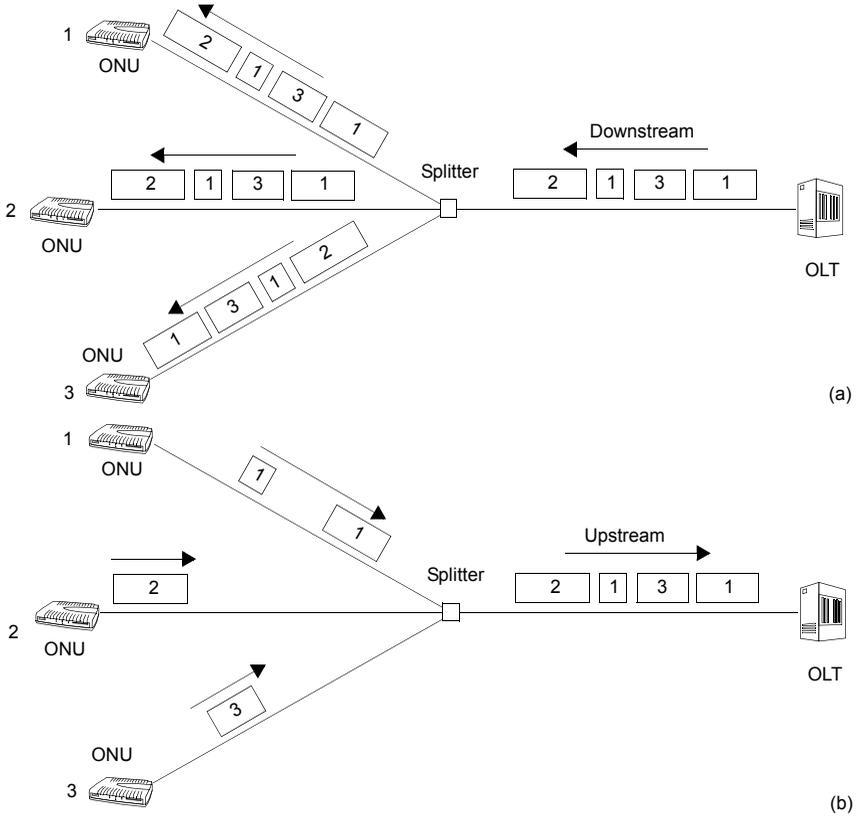


Figure 4.6 Transmission medium sharing a PON: (a) The downstream signal is broadcast to all the ONUs. (b) The upstream signal is point-to-point. The section between the splitter and the OLT is shared between all the ONUs.

gration to WDM PON would require replacing end equipment, but not the optical distribution network.

- The PON point-to-multipoint architecture in the downstream makes it easy to offer broadcast services such as TV. Broadcast

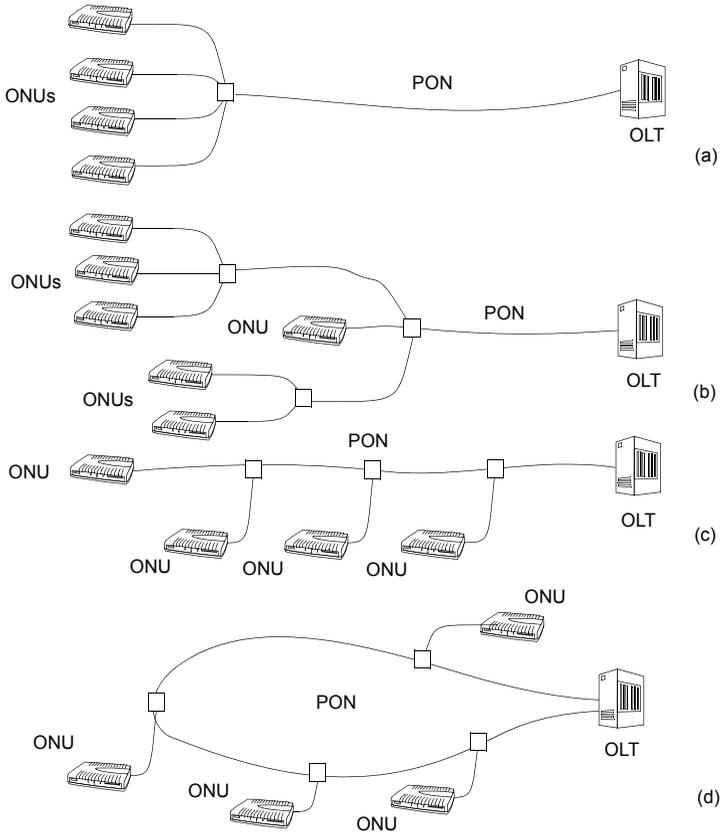


Figure 4.7 Different PON topologies: (a) star (b) tree (c) bus (d) ring.

services can be provided in a dedicated wavelength separated from unicast and multicast data services.

- There are many topologies compatible with the PON technology beyond the basic star topology. Various 1:N passive splitters can be chained, allowing for a tree topology. Using 1:2 tap couplers

enables bus and ring topologies. Furthermore, basic topologies can be easily extended to redundant topologies offering resiliency when facing service shortages (see Figure 4.7).

On the other hand, using PONs has some inconveniences as well. The most important drawbacks are reduced range and bandwidth when compared to active Ethernet, due to the attenuation introduced by the splitters and the effect of sharing resources.

EPON Particularities

The main goal of the EPON physical interfaces is to provide an access point where to connect MAC entities capable of transmitting standard IEEE 802.3 MAC frames. PON networks are a mixture of a dedicated and shared medium and EPON emulates point-to-point links over this medium. To do that, it extends the traditional Ethernet physical layer by defining:

- A scheduling protocol called *Multi-Point Control Protocol* (MPCP) that distributes transmission time among the ONUs to avoid upstream traffic collisions.
- Tags known as *Logical Link Identifiers* (LLID) that define point-to-point associations between the ONU and the OLT at physical level.

As a result, the EPON is compatible with most of the advantages provided by switched Ethernet networks like IEEE 802.1D bridging or VLANs. These features can be provided by the ONUs and OLTs themselves. Furthermore, the EPON defines other features that are not native in traditional Ethernet networks. For example, *Forward Error Correction* (FEC) is defined to increase range and splitting ratio.

Physical Layer

The EPON upstream and downstream are duplexed in a single SMF fiber. For IEEE 802.3ah interfaces, the upstream is transmitted at a

nominal wavelength of 1310 nm, and the downstream at 1490 nm. This allows for the EPON to coexist with other services, such as broadcast video or private WDM transmitted in the 1550 nm window. The signal is encoded with the same 8B/10B code that was specified by most of the Gigabit Ethernet interfaces operating at 1 Gb/s. This means that the signaling rate for 1 Gb/s EPON is 1.25 Gbaud. The 10 Gb/s signals use the 64B/66B PCS. That, results to a 10.3125 Gbaud signalling rate.

The upstream and downstream rates for 1 Gb/s EPON are always 1Gb/s, but 10 Gb/s EPON defines both 1 Gb/s (10/1GBASE-PRX10, 10/1GBASE-PRX20 and 10/1GBASE-PRX30 interfaces) and 10 Gb/s (10GBASE-PR10, 10GBASE-PR20 and 10GBASE-PRX30) upstream rates. Simultaneous support of two bit rates in 10 Gb/s EPON makes this standard more complex than 1 Gb/s EPON. The IEEE gives strong relevance to smooth compatibility of both bit rates within the same access network.

All 1 Gb/s upstream use the 1310 nm wavelength for the 10 Gb/s EPON upstream signal. This allows the OLT to use the same receiver for all 1 Gb/s signals even in cases where the downstream is a 10 Gb/s signal. Up streams based on the 10 Gb/s rate use a wavelength close to the 1 Gb/s window. In fact both regions overlap. OLTs supporting 1 Gb/s and 10 Gb/s operation, are required to implement dual rate burst-mode reception of optical signals. This is one of the most challenging issues for simultaneous operation of 1 Gb/s and 10 Gb/s EPON.

Multiplexing of downstream signals is accomplished with WDM. All 10 Gb/s are mapped the 1574-1580 nm window. Older 1 Gb/s upstream use the 1490 nm for the same purpose. The appropriate optical receiver for each wavelength is required in 1 Gb/s or 10 Gb/s ONUs.

Other new feature of 10 Gb/s EPON is the availability of new interfaces with extended power budget to enable higher splitting

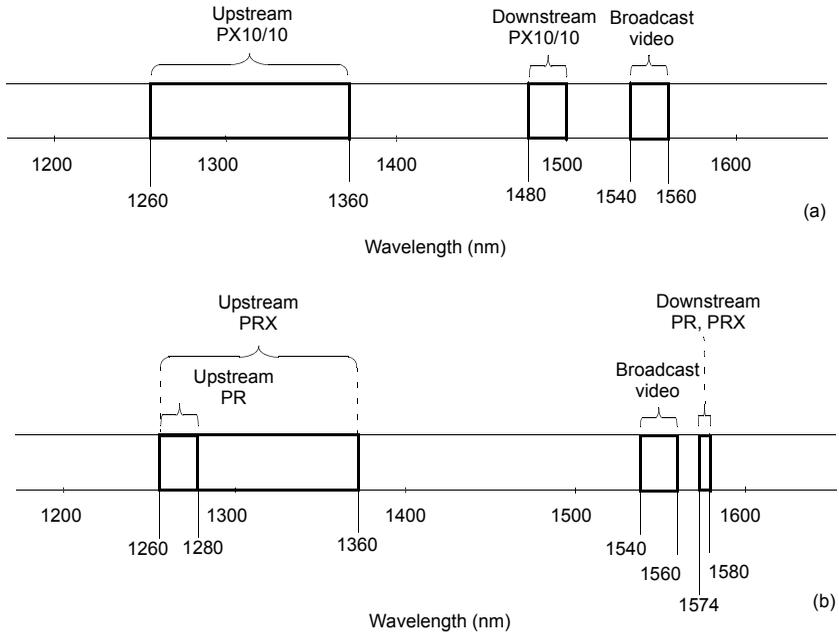


Figure 4.8 Optical window allocations for 1 Gb/s and 10 Gb/s EPON: (a) Allocation for 1 Gb/s EPON (b) Allocation for 10 Gb/s.

ratio. These interfaces are termed as PR30 (10 Gb/s symmetric) and PRX30 (1 Gb/s - 10 Gb/s asymmetric). The extended reach 10 Gb/s EPON has stricter channel insertion loss requirements (≤ 29 dB, 20 km, 32 users) than 20 km interfaces (≤ 24 dB, 30 km, 16 users) and 10 km interfaces (≤ 20 dB, 30 km, 16 users). In order to increase operation range of PR30 and PRX30, the IEEE 802.3av mandates a Forward Error Correction (FEC) based on a Reed-Solomon (255, 239) code which reates 16 redundancy bits for each 239 data bits. The Reed-Solomon coding scheme can be used in PR10, PR20, PRX10 and PRX20 interfaces as well but it is not mandatory. on

When the FEC mechanism is taken into account, the effective data rate of a 10 Gb/s EPON link is reduced to approximately 8.7 Gb/s.

The Multi-Point Control Protocol

The *Multi-Point Control Protocol* or MPCP is a signaling protocol for EPON, and its main function is to allow the OLT to manage the downstream bandwidth assigned to the ONUs. This protocol can perform other functions as well, namely:

- Enable the ONUs to request upstream bandwidth for transmission, and the OLT to assign this bandwidth in a way that collisions do not occur and network utilization is optimized.
- Allow parameter negotiation through the EPON network.
- Enable ranging by monitoring the *Round Trip Delay* (RTD) between ONUs and OLT. This feature is important for correctly scheduling upstream transmissions.
- Support ONU autodiscovery and registration.

The MPCP is implemented as an extension of the MAC control protocol and therefore MPCP messages are carried over standard Ethernet frames with the Type/Length field set to 0x88-08. There are five MPCP messages currently defined.

- GATE – grants access to the upstream bandwidth for the ONUs for certain periods of time.
- REPORT – used by the ONUs to report local information to the OLT. This information is used by the OLT to decide how the upstream bandwidth is distributed.
- REGISTER, REGISTER_REQUEST and REGISTER_ACK – used for registering ONUs in the network.

The IEEE standards define the protocol for scheduling bandwidth, but equipment manufacturers select the actual scheduling algorithm.

Logical Link Identifiers

Logical Link Identifiers or LLIDs are physical layer link identifiers defined to enable 802.1D bridging over an EPON (see Figure 4.9).

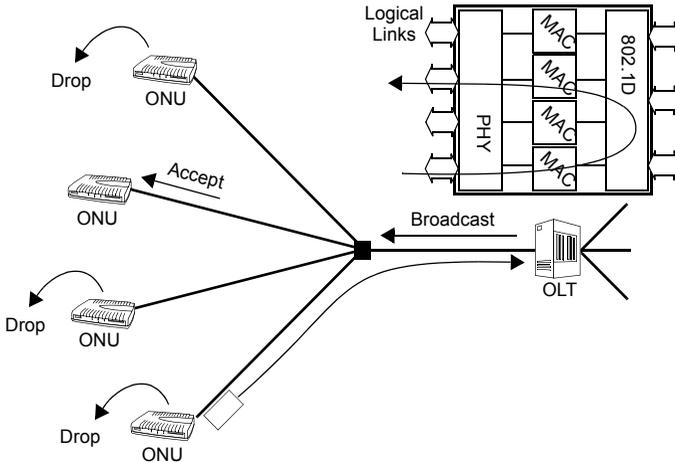


Figure 4.9 ONU-to-ONU bridging would not be possible without LLIDs. ONU-to-OLT associations defined by the LLIDs can be considered as point-to-point logical links. An 802.1D bridge can then perform learning and forwarding operations on the logical links.

The LLID is delivered in EPON Ethernet frames as a 16-bit field that replaces the two last bytes of the frame preamble (see Figure 4.10). This field is added when a frame is transmitted by an EPON interface and transparently removed when received before being processed by the MAC layer.

LLIDs define point-to-point associations or logical links between the ONU and the OLT. Link identifiers are dynamically assigned when ONUs are registered in OLTs as a part of the initialization process. ONUs and OLTs choose the LLID to put in the delivered frames depending on the logical link they wish to use. Point-to-

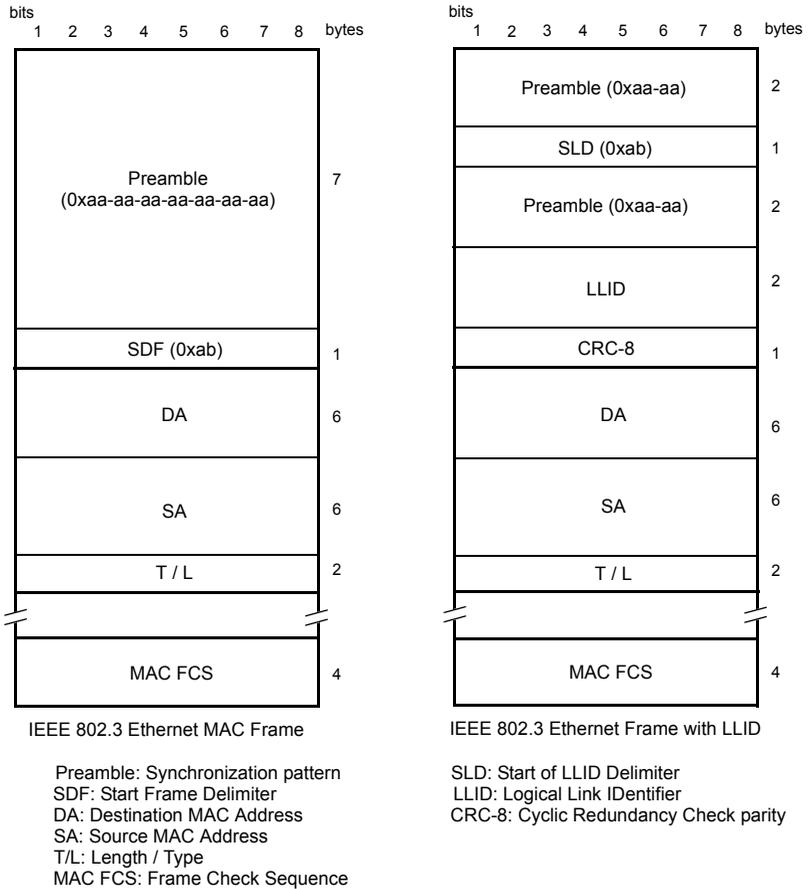


Figure 4.10 The preamble of an Ethernet frame carries the LLID, the SLD that helps processing the modified frame, and a CRC that detects errors in these new fields.

point emulation is achieved by following simple filtering rule: If a frame is received by an ONU or OLT with an LLID matching a known link identifier, it is forwarded to the right MAC entity that processes it. Otherwise, the frame is discarded. ONUs need to support a single LLID. They mark outgoing frames with the LLID assigned to them, and they accept frames marked with this LLID. The OLTs are more complex: they need one LLID per connected ONU.

The point-to-point link emulation is the primary operation mode for EPONs, but they may optionally support a shared LAN emulation mode. It is also possible to take advantage of the broadcast nature of the downstream by defining a special channel called *Single Copy Broadcast* (SCB) channel. Frames sent by the SCB channel are accepted by all the ONUs.

Selected Bibliography

- [1] IEEE 802.3-2008, "Part 3: Carrier sense multiple access with collision detection (CSMA/CD) Access Method and Physical Layer Specifications," December 2008.
 - [2] ITU-T Rec. G.984.1, "Gigabit-capable passive optical networks (GPON): General characteristics," March 2008.
 - [3] ITU-T Rec. G.984.2, "Gigabit-capable Passive Optical Networks (GPON): Physical Media Dependent (PMD) layer specification," March 2003.
 - [4] ITU-T Rec. G.984.3, "Gigabit-capable Passive Optical Networks (G-PON): Transmission convergence layer specification," March 2008.
 - [5] ITU-T Rec. G.984.4, "Gigabit-capable Passive Optical Networks (G-PON): ONT management and control interface specification," February 2008.
 - [6] ITU-T Rec. G.984.5, "Gigabit-capable Passive Optical Networks (G-PON): Enhancement band," September 2007.
 - [7] ITU-T Rec. G.984.6, "Gigabit-capable passive optical networks (GPON): Reach extension," March 2008.
 - [8] ITU-T Rec. G.987, "10-Gigabit-capable passive optical network (XG-PON) systems: Definitions, abbreviations, and acronyms," January 2010.
-

-
- [9] ITU-T Rec. G.987.1, "10-Gigabit-capable passive optical networks (XG-PON): General requirements," January 2010.
 - [10] ITU-T Rec. G.987.2, "10-Gigabit-capable passive optical networks (XG-PON): Physical media dependent (PMD) layer specification," January 2010.
 - [11] ITU-T Rec. G.992.3, "Asymmetrical digital subscriber line transceivers 2 (ADSL2)," January 2005.
 - [12] ITU-T Rec. G.993.1, "Very high speed digital subscriber line," June 2004.
 - [13] ITU-T Rec. G.998.1, "ATM-based multi-pair bonding," January 2005.
 - [14] ITU-T Rec. G.998.2, "Ethernet-based multi-pair bonding," January 2005.
 - [15] Sargento S., Valadas R., Gonçalves J., Sousa H., "IP-Based Access Networks for Broadband Multimedia Services," *IEEE Communications Magazine*, February 2003, pp. 146-154.
 - [16] Kerpez K., "DSL Spectrum Management Standard," *IEEE Communications Magazine*, November 2002, pp. 116-123.
 - [17] Kerpez K., Waring D., Galli S., Dixon J., Madon P., "Advanced DSL Management," *IEEE Communications Magazine*, September 2003, pp. 116-123.
 - [18] Kramer G., Pesavento G., "Ethernet Passive Optical Network (EPON): Building a Next-Generation Optical Access Network," *IEEE Communications Magazine*, February 2002, pp. 66-73.
 - [19] Kramer G., Mukherjee B., Pesavento G., "IPACT: A Dynamic Protocol for an Ethernet PON (EPON)," *IEEE Communications Magazine*, February 2002, pp. 74-80.
 - [20] Effenberger F., Ichibangase H., Yamashita H., "Advances in Broadband Passive Optical Networking Technology," *IEEE Communications Magazine*, December 2001, pp. 118-124.
 - [21] Maeda Y., Okada K., Faulkner D., "FSAN OAN-WG and Future Issues for Broadband Optical Access Networks," *IEEE Communications Magazine*, December 2001, pp. 126-132.
 - [22] Ueda H., Okada K., Ford B., Mahony G., Homung S., Faulkner D., Abiven J., Durel S., Ballart R., Erikson J., "Deployment Status and Common Technical Specifications for a B-PON System," *IEEE Communications Magazine*, December 2001, pp. 134-141.
 - [23] Pesavento G., "Ethernet Passive Optical Network (EPON) architecture for broadband access," *Optical Networks Magazine*, January/February 2003.
 - [24] Eriksson P., Odenhammar B., "VDSL2: Next important broadband technology," *Ericsson Review*, No. 1, 2006, pp. 36-47.
-

Ethernet Mobile Backhaul Networks

Chapter 5

Ethernet Mobile Backhaul Networks

For old Telcos, 64 kb/s was equivalent to one voice circuit and in some way, to one single user. The result is that in the old telephony business, network capacity (and revenue) was tightly bound with the subscriber population.

All this changed when the telecommunications operators developed their data and multiplay services based on Digital Subscriber Loop (DSL) access and later in optical technology. The most evident example is perhaps IPTV. While one TV channel may require hundreds the bandwidth of voice, the service provider is unable to charge their customers one hundred times more than it did in the past. This effect is what expertise call traffic/revenue decoupling and is on the basis of the recent evolution of service provider networks.

If you cannot charge your subscribers as per the bandwidth they use in multiplay applications like IPTV, there is only one way to protect your business profitability: you have to optimize your network so that the cost per bit (CAPEX and OPEX) becomes smaller. The strategy telecom operators have found to get maximum profitability of their networks is packet switching depending on three related technologies: IP, Ethernet and MPLS.

Now, that mobile operators are offering wideband data services and multiplay as well, they are following the same path than fixed operators a few years ago. As expected, the solution for them is migration to packet-switched technology their backhaul networks. Much better if they can use the same network available for delivering fixed services to residential and businesses.

In fact, almost everything has been used for mobile backhaul: E1/DS1, SDH/SONET, DSL, DOCSIS, PON, Microwave/WiMAX. However, Ethernet and related technologies like pseudowires are gaining momentum and they are becoming really popular.

Towards the “All-IP” Network

Traditional mobile backhaul is TDM based. First data services for the mobile network were based on circuit switching and TDM as well. Examples are the Circuit-Switched Data (CSD) and High-Speed CSD (HSCSD) for the Global System for Mobile communications (GSM). These offer a few tens of kb/s at most.

Of course, the data services are the best candidates for migration to packet switching. For GSM, the first packet switched service was the General Packet Radio Service (GPRS). This service requires new special nodes to be installed in the network: the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN).

With time, much better services than GPRS such as High-Speed Packet Access (HSPA) are becoming widely available. As prominence of packet based services increases, it becomes more natural to implement packet based backhaul and it is here where it arises the question of what to do with legacy services depending on TDM. This is the reason because the ITU-T, IETF, MEF and other organizations have released standards for Circuit Emulation Services (CES) over packet switched networks. Based on this strategy, legacy TDM is transported with the help of some CES technology while data grow is handled by native packet switching based on Ethernet and IP.

Requirements for the packet switched backhaul network are pretty much the same that for any other carrier network: scalability, Quality of Service (QoS), advanced Operation, Administration and Maintenance (OAM) mechanisms, high resiliency and traffic engineering features. MEF Ethernet service classes (E-Line, E-LAN, E-Tree) are flexible enough for mobile backhaul applications.

Even when all services have been migrated to packet switched networks there is still one important issue left: cellular base stations need an accurate timing (usually better than 1 ppm) reference to

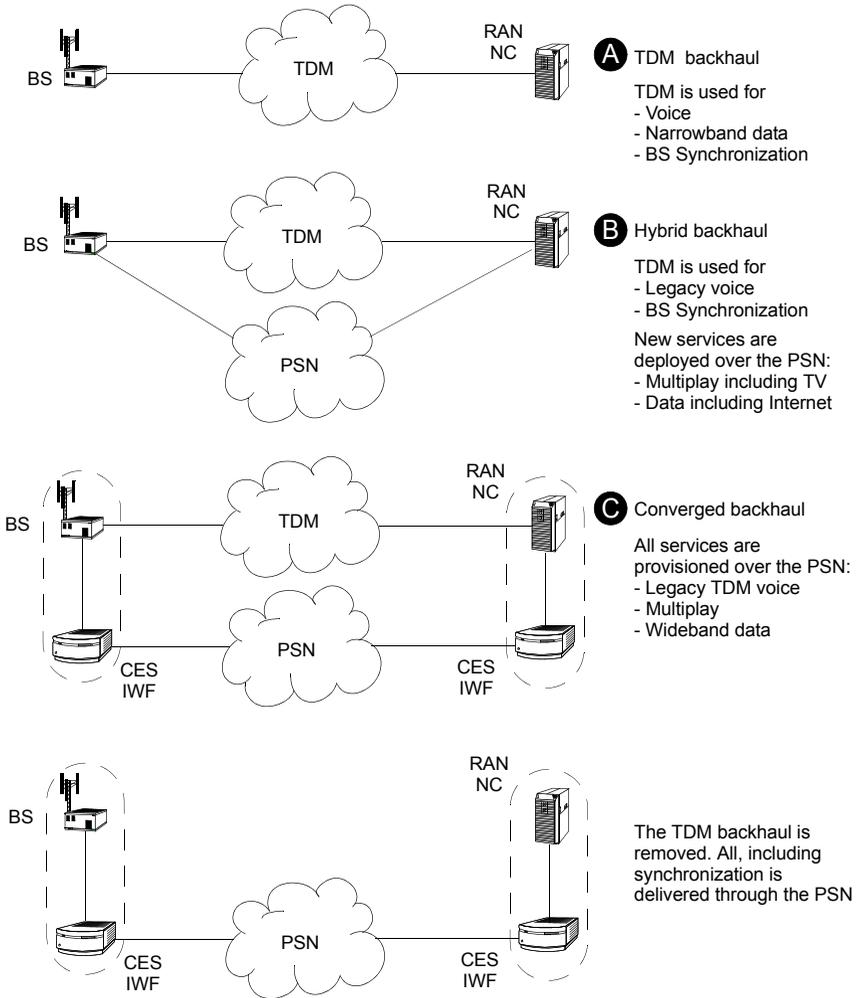


Figure 5.1 Migration to all-IP/Ethernet backhaul from a pure TDM mobile backhaul network.

make efficient use of the radio spectrum and enable handover operation between cell sites. This accurate clock is most likely delivered over TDM signals. There are different alternatives to migrate the TDM base station synchronization in Ethernet/IP backhaul scenarios. Base stations may get their timing from a specialised network separated from the mobile network like for example the Global Positioning System (GPS) satellite network. However, mobile operators most likely will be interested in propagating synchronization over the mobile backhaul network. In case of Ethernet backhaul, operators may use IEEE 1588 or Synchronous Ethernet to accomplish this (see Figure 5.1).

Circuit Emulation Services

With the proliferation of Ethernet and IP in MAN and WAN, transport of TDM services over packet switched infrastructures has become a key topic.

Circuit Emulation Services (CES) over packet switched networks provide the functions essential to emulate the service offered by a circuit. It therefore offers a transmission service for those applications that generate information as constant bit flows; for example, uncompressed telephony or video signals, and so on. The functions to carry out this type of service are the following:

- Data transmission at a constant, usually low/medium rate between the source and the destination.
- Transfer of timing information between source and destination.
- Structured data transmission between originating and destination users.
- Failure recovery mechanisms or at least tools for reporting errors to the network management system.

All these issues related with CES are discussed in the following sections.

Transmission of Timing Information

An important requirement for CES is the capability of recovering original TDM timing at the destination. There are two methods for transporting timing information. *Adaptive clock recovery*, does not require explicit timestamps attached to data packets but it offers limited performance, on the other hand *differential clock recovery* offers much better performance but it requires timestamps and a common synchronization source in the origin and destination.

Adaptive Clock Recovery Method

This is the simpler of the two methods. In this method, what is transported is not really explicit source clock information. The source clock is recovered at the destination, as the average of data received at the destination is an indication of the generating frequency at the originating point. This average rate softens the effects of packet delay variation. These effects may distort the estimated originating clock frequency, as they may easily cause oscillation in the packets received at short intervals.

To obtain the mean value of packets received per unit of time, and in this way estimate the clock frequency with which the information must be delivered to the destination user, a buffer must be used and its level monitored. The information blocks received are stored in a buffer. The buffer level is monitored continuously, and it is used to control the phase-locked loop (PLL) that generates the clock in the destination. This clock marks the reading pace of the buffer, to deliver the data to the destination. This is to maintain the buffer level environment at a mean level, so that if the level of information increases, it indicates that the generation rate has also increased at the originating point. That is why this must be done by the destination clock. On the other hand, if the level decreases, the clock frequency will decrease as well, generating information at a lower rate (see Figure 5.2).

If, for instance, adaptive clock recovery is used to emulate a point-to-point 2,048 kb/s circuit, the signal obtained in the destination will comply with the requirements associated with the jitter of the signal. However, it is not possible to guarantee that the recovered signal will meet the wander requirements explained in recommendations ITU-T G.823 and G.824.

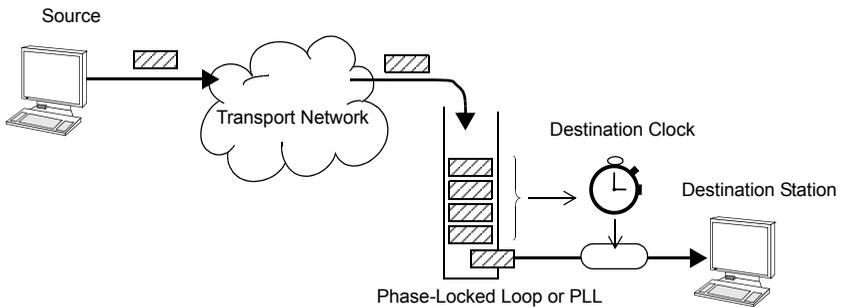


Figure 5.2 Synchronization by adaptive clock. The destination clock is synchronized according to the level of the input buffer. It is intended that this level remains half full.

Differential Clock Recovery Method

In this method, synchronization information is transmitted between two users. This method can only be used if the same base clock is used at both ends. This common clock may be provided by a GPS receiver or by any other means.

At both ends, a clock is used that is derived from the clock provided by the synchronization network. At the originating point, this clock is compared to that of the data to be transmitted. The relation between these two clocks is formed by a fixed nominal part and a variable residual part, due to frequency variations in both clocks. It is the value of this residual part that is transmitted across the network. At the destination point, this value will be used to modify

the local derived clock, and in this way obtain the same clock as at the originating point. This clock will be used to deliver the signal to the destination with the same clock as at the originating point (see Figure 5.3).

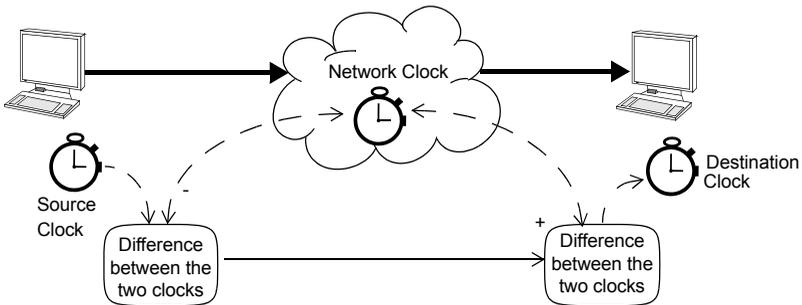


Figure 5.3 Synchronous residual time stamp (SRTS) synchronization. The difference is transported as information.

Structure Aware vs. Structure Agnostic CES

The IETF, ITU-T and MEF have published different standards regarding CES. Operators have thus many choices at their disposal to map their TDM data over packet switched networks. The most important decision is probably between structure aware CES or structure agnostic CES. The former considers the TDM signal a continuous and uniform bit stream and the later requires knowledge about the TDM frame structure enveloping and carrying the user data.

Structure agnostic CES is ideal for transport of truly unstructured (unframed) TDM, but is also suitable for transport of structured TDM when there is no need to protect structure integrity nor interpret or manipulate individual channels during transport. Structure agnostic CES is best suited in packet switched networks

with negligible packet loss, and for applications that do not require discrimination between channels nor intervention in TDM signalling.

Structure aware transport should be reserved for deployments where some advanced interaction between the packet switched network and the TDM network is desirable or required. When structure-aware TDM transport is employed, it is possible to explicitly safeguard TDM structure during transport over the packet switched network. Sometimes, the structure aware CES transmission system may strip the Frame Alignment Sequence (FAS) and other frame fields before mapping the information in packets and regenerate these fields at the receiving ends

Precedents: TDM over ATM using AAL1

ATM provides different ATM Adaptation Layers (AALs) to satisfy the transmission needs of different services and traffic types.

The AAL1 layer receives user information as a continuous flow and groups it into blocks of 47 bytes. This feature makes AAL1 suitable for transporting TDM data.

To each user data block is added a header byte with the following fields (see Figure 5.4):

- The *Convergence Sublayer Indication* (CSI): The first bit that indicates that there is information available on the convergence sub-layer.
 - The *sequence number* (SN): A sequence number of three bits, incrementing for each transmitted block. Sequence hops in the destination make it possible to detect lost cells. Its range is from 0 to 7, and, logically, this does not enable detecting big bursts of lost cells.
 - *Cyclic Redundancy Check* (CRC): 3-bit CRC of the previous fields (CSI and SN).
-

- Parity (P): Provides an even-number parity bit over the seven previous bits (CSI, SN, and CRC).

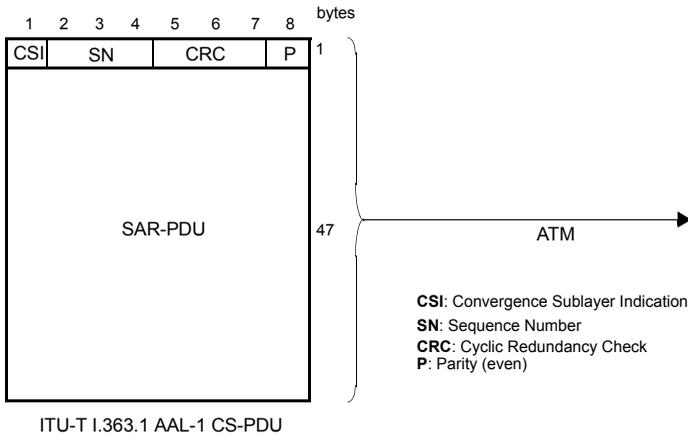


Figure 5.4 AAL1 structure.

Structured Data Transmission

The AAL1 layer offers an *Structured Data Transmission* (SDT) service. This service delivers information about the structure of the data transmitted. This information consists of a field situated in the second byte of the even-numbered blocks. The field contains a pointer indicating the byte of the current block or the following block of 48 bytes that contains the beginning of the transmitted data structure. The first bit of this field indicates if the field is actually transporting a valid pointer value, and the following seven bits made up the pointer value. Note that these seven bits make it possible to address up to 93 positions that are enough to indicate the 46 bytes of data of the current block, plus 47 bytes of data of the next odd-numbered block. This service enables the receiver to rapidly recover the synchronization of the transmitted data, even if there was a significant burst of lost cells.

Encapsulations for Structure Agnostic CES

As defined in RFC 4553, SAToP is a structure-agnostic protocol for transporting TDM using pseudowires. SAToP treats the TDM input as an arbitrary bit stream, completely disregarding any structure that may exist in the TDM flow.

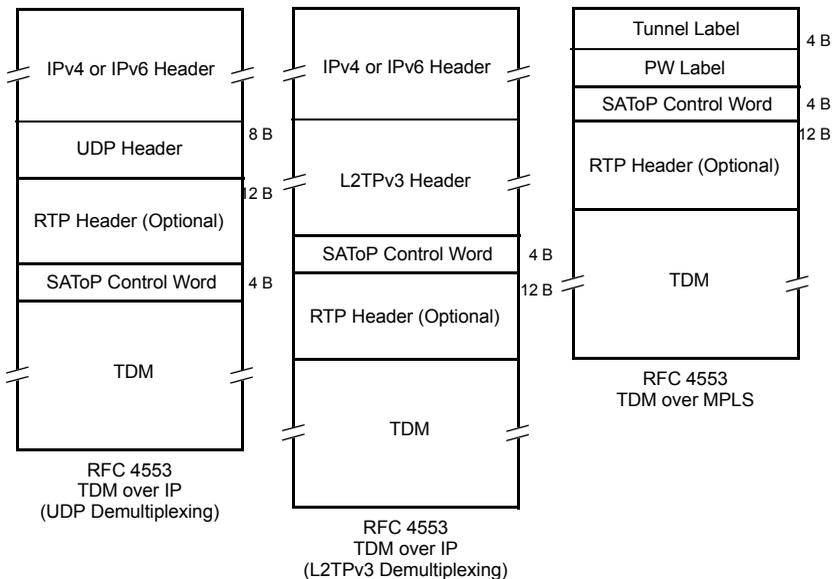


Figure 5.5 SAToP encapsulations for different packet switched networks, including IPv4, IPv6 and MPLS.

The SAToP describes encapsulations for transporting TDM over IPv4, IPv6 and MPLS (see Figure 5.5). Encapsulations for all three packet transport networks have similar structure:

1. First, there is the packet switched network header that is either an standard IPv4 header, an IPv6 header or a MPLS label.

2. It follows a demultiplexing header that enables transport of many SAToP entities over the same network infrastructure. For IPv4/IPv6 both UDP and L2TPv3 are accepted demultiplexing headers. For MPLS packet networks, an MPLS label is employed for demultiplexing. The MPLS transport label and the MPLS demultiplexing label made up the familiar two-label stack of MPLS pseudowires.
3. After the demultiplexing header there is the SAToP control word. This is a two byte overhead which includes specific control information for the SAToP service.
4. An optional RTP header can be used for transporting timestamps from the origin to the destination. These timestamps may help recovering the original timing when combined with differential clock recovery. The position of the RTP header depends on the actual encapsulation. It is inserted before the SAToP control word in UDP demultiplexing and after the SAToP control word otherwise.

Packets belonging to a given SAToP service carry a fixed number of bytes filled with TDM data. The RFC 4553 recommends payload sizes for different bit rates but payload size is negotiated during the connection setup and its value may be different to this value. The payload size is same for both transmission directions, and remains unchanged for the service lifetime.

Bit Rate	Payload length	Standard body
Synchronous Serial	64 bytes	ITU-T
E1	256 bytes	ITU-T, IETF
T1	192 bytes	ITU-T, IETF
E3, T3	1024 bytes	ITU-T, IETF

Table 5.1
Payload length recommended by different standard bodies for different TDM interfaces

The only overhead that is specifically related with structure agnostic TDM transport is the SAToP control word. This is the functionality added by the control word:

- Detection of packet loss or misordering.
- Differentiation between problems related with the packet switched network and the TDM circuit.
- Bandwidth conservation by not transferring invalid TDM data.
- Signaling of faults detected at the pseudowire egress to the pseudowire ingress.

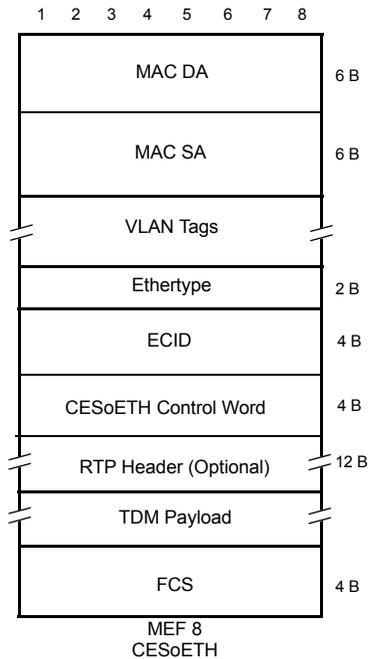


Figure 5.6 CESoETH frame encapsulation as defined in MEF 8 standard.

IETF SAToP is closely related with the ITU-T Y.1413. In one sense, the ITU-T Recommendation has wider coverage than the RFC 4553: It supports structure agnostic and structure aware transmission, it has more transmission rates, and it provides more details about implementation. On the other hand, ITU-T Y.1413 only defines encapsulations for MPLS. However, ITU-T has Recommendation Y.1453 with similar purpose that Y.1513 but for IP packet transport networks.

The third standards body involved in CES standardization is the MEF. MEF has released MEF 8 which includes encapsulations for transporting TDM data over Ethernet frames. Both structure aware and structure agnostic modes are supported. Frame structure is similar to the encapsulations defined in RFC 4553 and ITU-T Y.1413 but the packet transport network header is now the IEEE 802.3 header. There is also a demultiplexing field known as the Emulated Circuit IDentifier (ECID) with a structure similar to an MPLS label, including a 20 bit multiplexing field. The control word defined in MEF 8 is similar to the one used by IETF and ITU-T standards (see Figure 5.6).

Encapsulations for Structure Aware CES

The IETF has two different standards for structure aware CES:

- CESoPSN, defined in RFC 5086 is based on similar principles that SAToP but it includes the extensions needed to encode the frame structured carried by the TDM signal.
- TDMoIP is defined in RFC 5087 and accomplishes the same objective that CESoPSN but it takes the AAL1 and AAL2 frame structures from ATM to efficiently encode TDM signals.

Encapsulations for SAToP and CESoPSN are basically the same but there are slight differences in the control word (see Figure 5.7). In CESoPSN combinations of the *L* and *M* control word fields are used to identify alarms or different kinds of data and signalling packets.

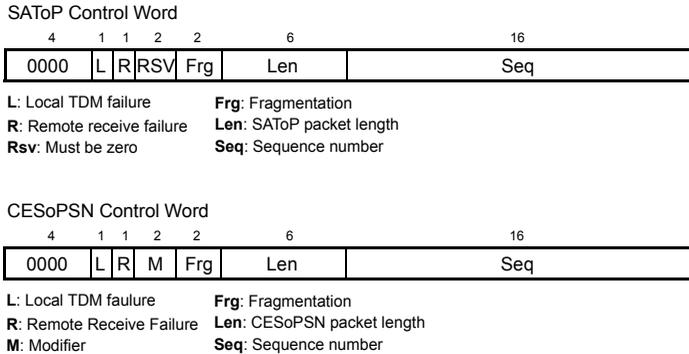


Figure 5.7 Control word for SAToP and CESoPSN packet formats.

Despite of its name TDMoIP, operate over several types of PSN, including UDP over IPv4 or IPv6, MPLS, and Ethernet. Encapsulations for TDMoIP are very similar to the ones defined in the CESoPSN standard but here, the payload structure is entirely different.

For unstructured TDM, or structured but unchannelized TDM constant-rate adaptation is needed. In such cases TDMoIP uses structure-indication to emulate the native TDM circuit, and the adaptation is known as *circuit emulation*. For channelized TDM wherein the individual channels (corresponding to *loops* in telephony terminology) are frequently inactive, bandwidth may be conserved by transporting only active channels. This results in variable-rate real-time traffic known as *loop emulation*.

TDMoIP uses constant-rate AAL1 for circuit emulation, while variable-rate AAL2 is employed for loop emulation. Furthermore, there is a third mode for efficient transport of HDLC based Common Channel Signaling (CCS) carried in TDM channels. AAL is a natural

choice for TDM emulation. Although originally developed to adapt various types of application to ATM, the mechanisms are general solutions to the problem of transporting constant or variable-rate real-time streams over a packet network. These mechanisms are mature and well understood, and implementations are readily available and interworking with legacy networks tends to be simpler.

Hands-on: MEF 18 and CES Certification

The MEF 18 standard defines acceptance tests for CESoETH devices. However, the MEF 18 can be easily extended for other CES encapsulations like the ones defined by the IETF and ITU-T.

The MEF 18 test case is made up of 18 conformance tests. These tests verify the following features of the CES Inter-Working Function (IWF):

- They check that the ECID and the control word fields structure follows standard MEF 8 and that their usage is in accordance with the mentioned standard (tests 1, 2, 3 and 4).
 - They check that both MEF 8 structure agnostic and structure aware is supported (tests 5, 13, 14, 15, 16). These tests verify that usage all frame fields, including MAC addresses, ECID identifiers and the control word meets the requirements of MEF 8.
 - It is verified (test 6) that the received signal meets the jitter and wander requirements stated in ITU-T G.823 (E1 and E3) and G.824 (DS1 and DS3).
 - They document behaviour in front of lost, duplicated and out of order frames (tests 7, 8, 9, 10).
 - They stress the DUT with a signal with rate slightly under or over the nominal and check behaviour in front of buffer underun and overun events (test 11).
-

- They verify that the DUT does not change the Facility Data Link (DFL) messages in DS1 signals with Extended Super Frame (ESF) frame format (test 12).

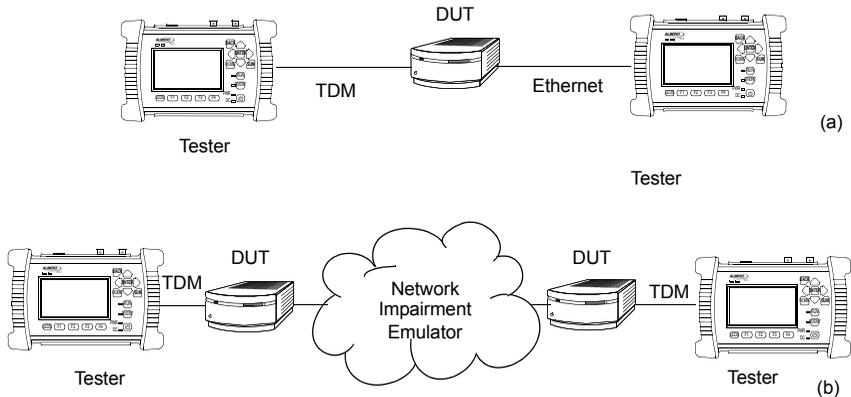


Figure 5.8 MEF 18 test beds for acceptance of CES. (a) This test bed is appropriate for testing encapsulation of TDM data. (b) Test bed for testing end to end performance of CES.

The MEF 18 defines the test beds to be used with the acceptance test suite. Specifically, two test setups are defined (see Figure 5.8):

1. *Testing of the isolated IWF*: To test the IWF TDM and packet interfaces are required. Both the tester for the TDM and for the packet interface must be capable of generating and receiving test signals and the code the
2. *End-to-end CES conformance / performance testing*: This test bed consists of equipment for generating and receiving TDM services, two identical devices to be tested, and equipment representing an Ethernet network. Some tests also require the ability to impair the stream by delaying, reordering, deleting or duplicating frames

Ethernet Synchronization with IEEE 1588

Precision Time Protocol (PTP), included in IEEE standard 1588 was originally designed to provide timing for critical industrial automation. With the 2008 version of this standard (IEEE 1588v2), PTP overcomes effects of latency and jitter through chains of Ethernet switches, providing accuracy in the nanosecond range.

Precedents: IP Synchronization with NTP

The Network Time Protocol (NTP), is one of the oldest protocols still in use and it is available in two levers: the full version and Simple NTP (SNTP), a subset of NTP.

The latest version of NTP, version 4 (NTPv4) can usually maintain time to within 1-20 ms using traditional software-interrupt based solutions over the public Internet and can achieve accuracies of microseconds or better in LANs under ideal conditions. NTP has been the most common and arguably the most popular synchronization solution, because it performs well over LANs and WANs and at the same time it is inexpensive, requiring very little hardware.

NTP should be able to deliver accuracy of 1-2 ms on a LAN and 1-20 ms on a WAN, it is far from guaranteed network-wide largely because of variable delays added by switches and routers.

PTP Protocol Details

PTP only requires a central Grandmaster clock and low-cost PTP slave clocks sites. Master and slave network devices are kept synchronized by the transmission of timestamps sent within the PTP messages.

Depending on how many ports has a network clock, it is referred by the IEEE 1588 standard as a Ordinary Clock (single port device) or a Boundary Clock (multi port device). The version 2 standard also

defines the concept of Transparent Clocks that improve timing accuracy when the protocol is run in network paths which contain intermediate switches (see Table 5.2).

Device	Description
Ordinary Clock	A single port device that can be a master or slave clock.
Boundary Clock	A multi port device that can be a master or slave clock.
End-to-end Transparent Clock	A multi port device that is not a master or slave clock but a bridge between the two. Forwards and corrects all PTP messages.
Peer-to-peer Transparent Clock	A multi port device that is not a master or slave clock but a bridge between the two. Forwards and corrects Sync and Follow-up messages only.
Management Node	A device that configures and monitors clocks.

Table 5.2
IEEE 1588v2 Device Description

The normal execution of the PTP has two phases:

1. *Master-Slave hierarchy establishment.* Ordinary and boundary clocks decide which port has the master or slave role in each link with the help of the Best Master Clock (BMC) algorithm. Data from the remote end of a path are provided *Announce* message.
2. *Clock synchronization.* Slave clocks may have a positive or negative offsets when compared with their masters and latency from masters to slaves is also unknown. PTP devices start a procedure to compute latencies and offsets. These parameters will be used to adjust timing in slave devices.

Once the master and slave hierarchies have been established, by observing the clock property information contained in *Announce* messages sent by PTP devices, the synchronization process starts (see Figure 5.9).

Before synchronization between the master and the slave clock has been achieved, it may exist an offset between both clocks. This offset is computed with the help of the *Sync* message. *Sync* messages are sent periodically (usually once every few seconds) by

the master to upgrade offset information in the slave. *Sync* messages may carry an accurate timestamp indicating the departure time of the own message but this requires expensive timestamping hardware which may not be available. To avoid expensive hardware *Follow_Up* messages can be used. *Follow_Up* messages carry timestamps for a previous *Sync* message allowing a more relaxed timestamping procedure and cheaper hardware. *Sync* procedure is based on multicast to enable a more efficient message transmission and processing.

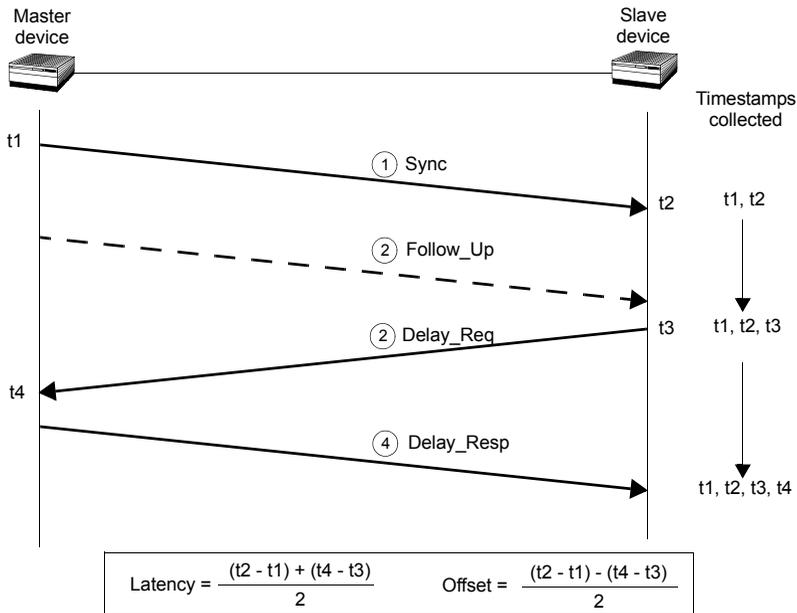


Figure 5.9 Delay Request-Response mechanism used by the PTP. The basic parameters of Latency and Offset are computed from the t1, t2, t3 and t4 timestamps.

The *Sync* mechanism, however, does not take into account propagation time of *Sync* messages through the network. For this reason, the slave requests a latency measurement with a *Delay Req*

message. Masters reply to a *DelayReq* with *Delay_Resp* message. The timestamps the slave gets from the Delay Request-Response mechanism are used to correct *Sync* with a more accurate time estimation.

The most difficult challenge of PTP is operation through chains of Ethernet switches. Most switches store packets in local memory while the MAC address table is searched and the cyclic redundancy field of the packet is checked before it is sent out on the appropriate port/s. This process introduces variations in the time latency of packet forwarding and damages accuracy of the PTP protocol. Version 1 of the PTP protocol deal with this problem by implementing Boundary Clocks within the switches. Version 2 uses the more advanced concept of Peer-to-Peer Transparent Clock to deal with the same problem.

Transparent clocks do not participate in the master-slave hierarchy but they process PTP messages by adding special correction fields within the message with their own estimations of packet residence times in the device and propagation delays from remote peers (which can be also a Peer-to-Peer Transparent Clocks). Network paths where Peer-to-Peer Transparent Clocks are employed do not need the Delay Request-Response mechanism. This mechanism is replaced by a peer-to-peer path correction mechanism based on the *Pdelay_Req* and *Pdelay_Resp* and *Pdelay_Resp_Follow_Up* messages.

Protocol Encapsulation

PTP messages can be carried over a large family of protocols including IPv4, IPv6, IEEE 802.3 Ethernet, DeviceNET, ControlNET and IEC 61158 Type 10. The most important encapsulations are the IP and Ethernet variations (see Figure 5.10).

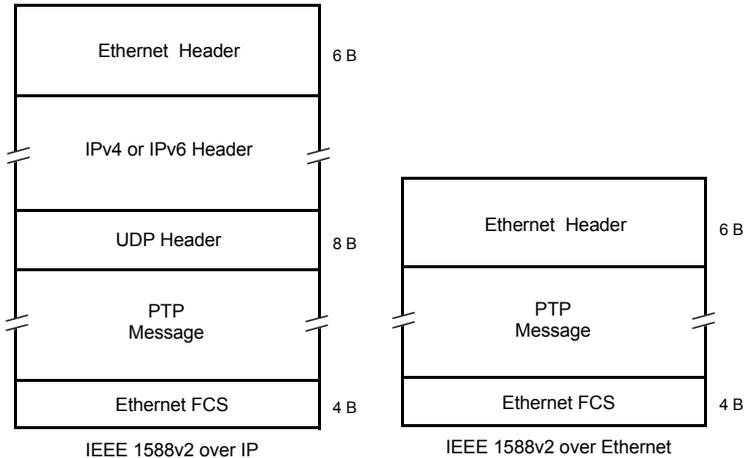


Figure 5.10 IP and Ethernet encapsulations for PTP messages.

Synchronous Ethernet

Synchronous Ethernet is an ITU-T standard that provides mechanisms to transfer frequency over the Ethernet physical layer, which can then be made traceable to an external source such as a network clock. As such, the Ethernet link may be used and considered part of the synchronization network (see Figure 5.11).

The proposal to specify the transport of a reference clock over Ethernet links was brought by operators to ITU-T Study Group 15 in September 2004. The aim of Synchronous Ethernet is to avoid changes to the existing IEEE Ethernet, but to extend it working within its protocol definitions.

Despite being an IEEE standard, Ethernet architecture has been described in ITU-T G.8010 as a network made up of an ETH layer and

a ETY layer. Put in simple terms, the ETY layer corresponds the physical layer as defined in IEEE 802.3, while the ETH layer represents the pure packet layer. Ethernet MAC frames at the ETH layer are carried as a client of the ETY layer. In OSI terminology, ETY is layer 1, ETH layer 2. Synchronous Ethernet is based on the ITU-T G.8010 description of the Ethernet architecture.

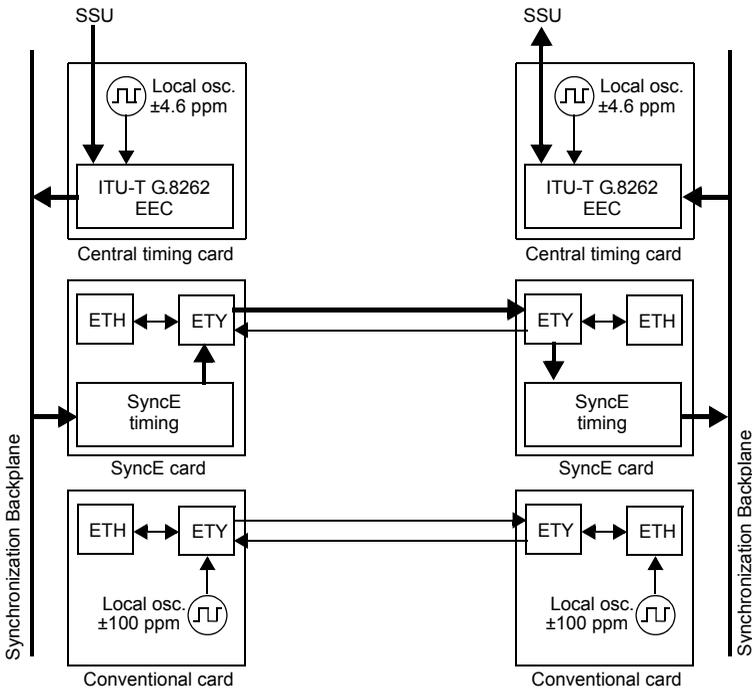


Figure 5.11 Synchronous Ethernet Architecture and comparison with conventional Ethernet

A key issue in Synchronous is the definition of the mechanisms necessary to achieve interworking between SDH and Synchronous Ethernet equipment in order for them to have a single

synchronization network to manage. These mechanisms and procedures are found fundamentally in three different recommendations: ITU-T G.8261, G.8262 and G.8264. The aspects covered there include the following:

- Extension of the synchronization network to include Ethernet as a building block (ITU-T G.8261). This enables Synchronous Ethernet network equipment to be connected to the same synchronization network that SDH. Synchronization for SDH can be transported over Ethernet and the opposite is also true.
- The ITU-T G.8262 defines Synchronous Ethernet clocks compatible with SDH clocks. Synchronous Ethernet clocks are based on ITU-T G.813 clocks and they are defined in terms of accuracy, noise transfer, holdover performance, noise tolerance, and noise generation. These clocks are referred as Ethernet Equipment Slave clocks. While the IEEE 802.3 standard specifies Ethernet clocks to be within ± 100 ppm. EECs accuracy is within ± 4.6 ppm. Additionally, by timing the Ethernet clock, PRC traceability of the interface is achievable.
- ITU-T G.8264 extends the usability of the ITU-T G.707 Synchronization Status Message (SSM) by Synchronous Ethernet equipment. The SSM contain an indication of the quality level of the clock that is driving the synchronization chain. The Ethernet Synchronization Message Channel (ESMC) is used for propagation of the SSM through the Synchronous Ethernet network.

Ethernet Synchronization Messaging Channel

In SDH, the SSM provides traceability of synchronization signals and it is therefore required to extend the SSM functionality to Synchronous Ethernet to achieve full interoperability with SDH equipment.

In SDH, the SSM message is carried in fixed locations within the SDH frame. However, in Ethernet there is no equivalent of a fixed frame. The mechanisms needed to transport the SSM over Synchronous

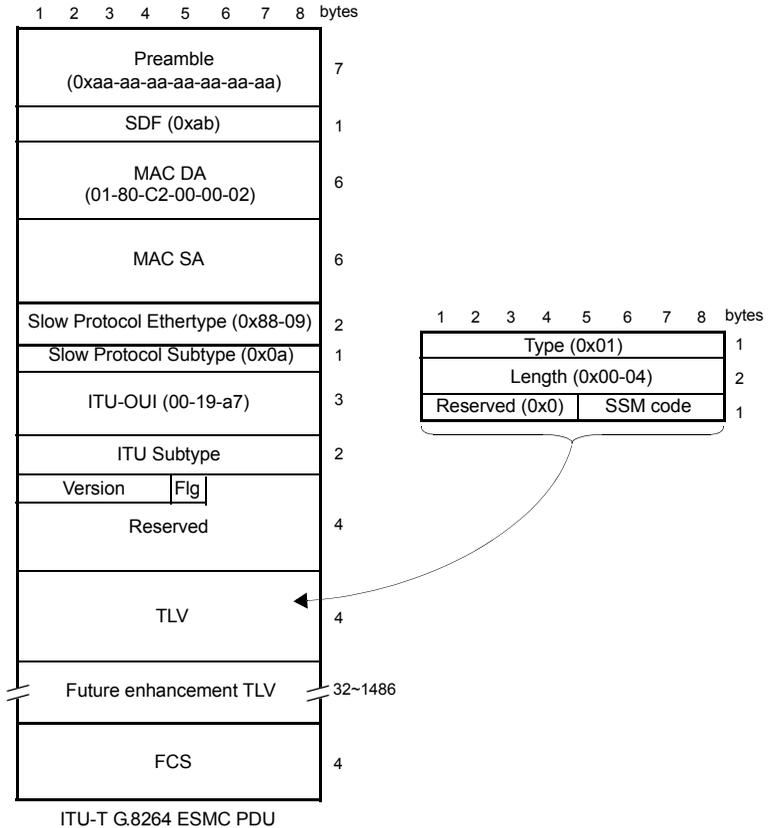


Figure 5.12 Ethernet Synchronization Message Channel (ESMC) protocol data unit.

Ethernet are defined by the ITU-T in G.8264 in cooperation with IEEE. More specifically, the ESMC, defined by the ITU-T is based on the Organization Specific Slow Protocol (OSSP), currently specified in IEEE 802.3ay.

The ITU-T G.8264 defines a background or *heart-beat* message to provide a continuous indication of the clock quality level. However, event type messages with a new SSM quality level are generated immediately.

The ESMC protocol is composed of the standard Ethernet header for a slow protocol, an ITU-T specific header, a flag field, and a type length value (TLV) structure (see Figure 5.12). The SSM encoded within the TLV is a four-bit field whose meaning is described in ITU-T G.781.

Selected Bibliography

- [1] Sargento S., Valadas R., Gonçalves J., Sousa H., "IP-Based Access Networks for Broad-band Multimedia Services," *IEEE Communications Magazine*, February 2003, pp. 146-154.
 - [2] Ferrant J., Gilson M., Jobert S., Mayer M., Ouellette M., Montini L., Rodrigues S., Ruffini S., "Synchronous Ethernet: A Method to Transport Synchronization," *IEEE Communications Magazine*, September 2008, pp. 126-134.
 - [3] Vainshtein A., Stein Y.J., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)," IETF Request For Comments RFC 4553, Jun. 2006.
 - [4] Stein Y(J)., Shashoua R., Insler R., Anavi M., "Time Division Multiplexing over IP (TDMoIP)," IETF Request For Comments RFC 5087, Dec. 2007.
 - [5] Vainshtein A., Sasson I., Metz E., Frost T., Pate P., "Structure-Aware Time Division Multiplexed (TDM) Circuit Emulation Service over Packet Switched Network (CESoPSN)," IETF Request For Comments RFC 5086, Dec. 2007.
 - [6] ITU-T Rec. Y.1413, "TDM-MPLS network interworking - User plane interworking," March 2004.
 - [7] ITU-T Rec. Y.1453, "TDM-IP interworking - User plane interworking," March 2006.
 - [8] ITU-T Rec. G.8261, "Timing and Synchronization Aspects in Packet Networks," February 2008.
 - [9] ITU-T Rec. G.8262, "Timing Characteristics of Synchronous Ethernet Equipment Slave Clock (EEC)," August 2007.
 - [10] ITU-T Rec. G.8264, "Distribution of Timing Through Packet Networks," February 2008.
-

-
- [11] Metro Ethernet Forum Technical Specification MEF 8, "Implementation Agreement for the Emulation of PDH Circuits over Metro Ethernet Networks," October 2004.
 - [12] Metro Ethernet Forum Technical Specification MEF 18, "Abstract Test Suite for Circuit Emulation Services over Ethernet based on MEF 8," May 2007.
-

The OSI Reference Model

Appendix A

The OSI Reference Model

The Open Systems Interconnect Reference Model (ISO 7498) defines a seven-layer communication architecture with physical transport at the lower layer and application protocols at the upper layers to standardize communications across computer networks.

The origin of the OSI model is the attempt to formulate a reference to coordinate the development of new standards that will permit network based applications to cooperate and interchange information independently of their manufacturer or the underlying transport technologies.

The model refers to the structure, protocols, and data formats required to guarantee an efficient, reliable and transparent service capable to support a wide range of applications based on data, voice and video. The model defines generic building blocks to assist the design of architectures to minimize the difficulties to interconnect systems through heterogeneous networks in terms of technology and performance.

Seven Layers

The OSI model describes a framework of seven layers to be followed by compliant implementations. The model tells which protocols should be implemented at each layer and which rules and data structures should be used to communicate with peers layers.

Every layer has to be implemented independently to make easier modifications or substitution whenever is required. Therefore layers will be able to be used by applications without the necessity of any particular agreement, whenever implementations support interfaces, formats and protocols described in the model.

Higher layers 7-6-5 are application-oriented to support end user applications and are closely related with the operating system and the computer. While lower layers 3-2-1 are network-oriented layers

responsibilities for the data transport therefore are closely related with the architectures and technology of the network. Upper layers interact with final users, while lower layers transform the messages sent and received by the application into data packets capable to travel across the network until reaching the far end destination.

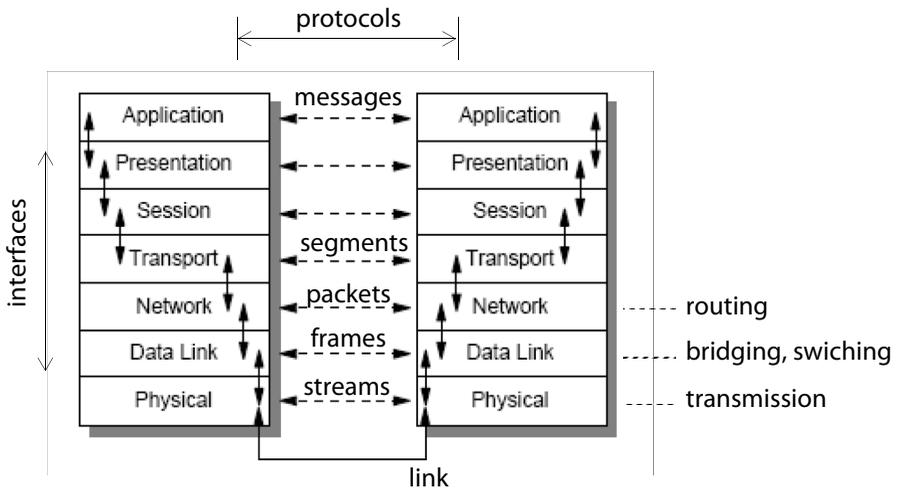


Figure 0.1 In the OSI model each layer communicates conceptually with its peer layer at the remote system but in practice message units are passed to the immediate lower layer. In other words messages are passed vertically from layer to layer, but under the logical point of view, each layer communicates directly with its peer layer on the remote system.

In essence we can see the OSI model as a logical structure where every layer provides a set of services to the above layer, while using those services offered by the layer immediately below.

Physical Layer

Layer 1 provides the mechanical and electrical interface to the transmission media. Whatever the nature of the link (electrical, optical or wireless) this layer has to transmit the bit stream between

two separated nodes. Then it is concern of the physical features of the communications channel in terms of attenuation, signal to noise, optical power, error, alarms, synchronization, codification, framing, etc. These parameters are often determined by the nature of the link that could be electrical, optical or wireless.

The physical layer determines the performance and the quality of the communication, because all the new messages interchanged between applications are finally conducted through the link that connects the nodes that are part of the network.

OSI model	
Application	Gateways
Presentation	Gateways
Session	Gateways
Transport	Gateways
Network	Routers
Data Link	Bridges, Switches
Physical	Hubs, Multiplexers, Regenerators, Connectors, Cables

Figure 0.2 Communication devices anmd the OSI model

Datal Link Layer

Layer 2 is responsible of the identification of adjacent nodes, and then facilitate a reliable transport across the physical link. Functions include framing, topology awareness, error detection, and flow control over a single transmission media.

There is a large variety of implementation however we can distinguish two approaches for LAN and WAN.

LAN architectures

The most popular Data Link Layer is Ethernet that identifies two sublayers:

- LLC: common interface to layer 3 can provide connection services but connectionless -or datagrams- is widely used.
- MAC: manage each physical layer therefore is different depending on the type of electrical, optical, or wireless link

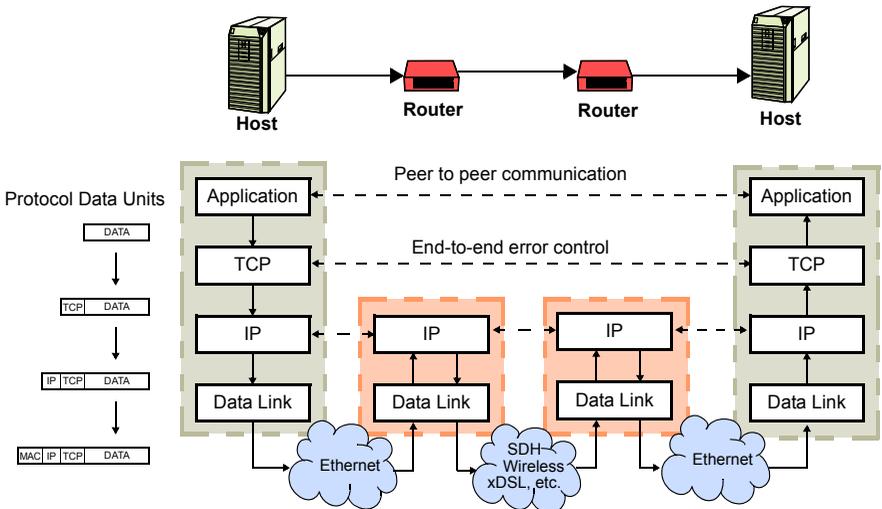


Figure 0.3 The Internet protocol enable hosts to communicate over heterogeneous networks, providing application connectivity specifying how protocol data units should be formatted, addressed, encapsulated, routed and delivered to the right destination.

WAN architectures

- Connection orientated, used mainly in WAN require the establishment of a connection before sending frames. This service has flow control function to control the transmission rate. Samples of SDLC, HDLC, PPP, LAPB, LAPD.

Network Layer

Layer 3 is responsible of the addressing facilitating the establishment of logical paths that messages will transit from source to destination. To make it possible logical addresses must be translated into physical addresses to define a logical connection across the network independently of its underlying technology.

In packet networks routing occurs at this level, being routers responsible of packet forwarding to destination using predetermined criteria also known as *routing protocol*.

Transport Layer

Layer 4 is responsible data is successfully sent and received between two systems transparently providing data transfer to the upper layers. To achieve this layer 4 may use tools as flow control, segmentation/desegmentation, and error control.

Despite not strictly conforming to the OSI model Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) are good samples of transport strategies can be followed at this layer.

- TCP is a connection oriented service. It can chop messages smaller segments that should be reassembled at destination, it also handles message flow control, and error free communication between nodes. It is typically used for data applications such as web navigation and data transactions.
- UDP is a connectionless service intended for those applications do not require the establishment of a virtual connection or error recovery, but require low overhead or are very sensitive to delays when the most important is a fast transmission. i.e. Voice Traffic.

Session Layer

Layer 5 main responsibility is to set and release communication channels between two computers, or two applications, over the

network. Session layer coordinates the dialogue between them providing tools to control the data-exchange process between the programs running in each machine.

Presentation Layer

Layer 6 provides independence from data representation by means to applications using different formats, syntax and semantics to represent information. Then presentation layer translates and maps the information in a meaningful way. Samples are coding conversion (ie. ASCII to/from EBCDIC), data compression schemes (GIF/MPEG/ZIP), and common data encryption algorithms.

These functions ensure that information sent from the application layer of one system would be readable by the application layer of another system.

Application Layer

- Layer 7th provide functionalities such as identification, availability and compatibility. This layer may also manage interfaces with final users and computer applications. Functionalities include applications such as FTP, HTTP, VoIP and IPTV.
-

Introduction to Ethernet

Appendix B

Introduction to Ethernet

The term *Ethernet* does not refer to one technology only, but to a family of technologies for local, metropolitan and access networks covered by the IEEE 802.3 standard. The best-known Ethernet technologies operate at 10 Mb/s; Fast Ethernet at 100 Mb/s, Gigabit Ethernet at 1000 Mb/s, 10-Gigabit Ethernet at 10 Gb/s and Higher Speed Ethernet at 40 Gb/s and 100 Gb/s.

Since *Local Area Networks* (LAN) were first defined 30 years ago, many technologies have been developed for this important market segment. Some time ago, names such as Token Ring, Token Bus, DQDB, FDDI, LATM, 100VG and AnyLAN were in everybody's mouth. However, Ethernet has outlived them all, becoming the standard technology used in nearly all LAN installations.

Even though the performance of Ethernet was quite limited in the beginning, a number of reasons made Ethernet a winner, including low cost, simplicity, flexibility and scalability. However, the most important factor was *technological convergence*, because it guaranteed smooth interworking without the need for any specialized gateways. After all, a network is the *means* to connect computers, not the goal, so Ethernet finally received the support it needed to be universally accepted by manufactures, users and service providers.

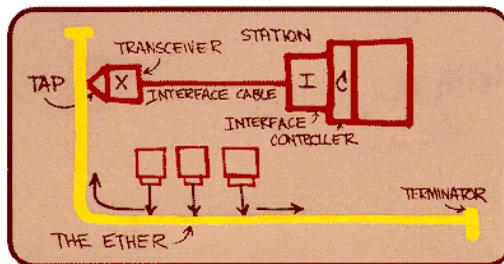


Figure 0.1 A drawing of the first Ethernet system by Bob Metcalfe

Furthermore, the IEEE 802.11 standards for *Wireless LAN* (WLAN) applications are important for emerging technologies. They are not exactly the same as Ethernet: the MAC frame format is different from the IEEE 802.3 Ethernet, and *Carrier Sense Multiple Access / Collision Avoidance* (CSMA/CA) is used instead of *Carrier Sense Multiple Access / Collision Detection* (CSMA/CD). However, the IEEE 802.11 standard is highly interoperable with Ethernet, and it is used frequently as a wireless extension for wired Ethernet networks.

A Brief History of Ethernet

There is a network that has always been considered as the predecessor of Ethernet: *ALOHA*, developed in the late 1960s by Norm Abramson at the University of Hawaii. ALOHA was a digital radio network designed to transmit independent packets of information between the Hawaiian islands (Figure 0.2).

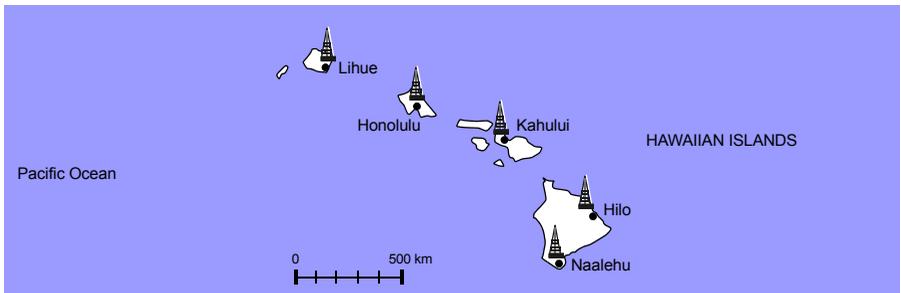


Figure 0.2 ALOHA, a pre-Ethernet network, was developed in the 1960s to transmit data between the Hawaiian islands.

The first real Ethernet was designed in 1973 by Bob Metcalfe in Xerox Corporation's Palo Alto laboratory (Figure 0.1). This first

version was able to operate at 3 Mb/s over a shared coaxial cable, using CSMA/CD. This is a simple, distributed, media access algorithm that operates without the need of a central device for host transmission coordination.

In 1980 a consortium formed by Digital, Intel and Xerox (known as the DIX cartel) developed the 10 Mb/s Ethernet. Finally, in 1983, the IEEE standards board approved the first IEEE 802.3 standard, which was based on the DIX Ethernet and at the same time is the basis of all current Ethernet standards.

Ethernet and the OSI Reference Model

The existing IEEE Ethernet standards define the physical medium, connectors, signals, procedures and protocols needed to connect devices. The functionality defined by the IEEE corresponds to layers 1 (physical) and 2 (data link) in the *Open Systems Interconnection* (OSI) model.

The physical layer (PHY) is defined in the IEEE standard 802.3. This standard includes different versions of Ethernet, operating at rates up to 100 Gb/s over an electrical or optical transmission medium. The Ethernet PHY is fully independent from upper layers, and sometimes it is implemented by using separate equipment. Due to the diversity of Ethernet as a transmission medium, many different architectures can be used for the physical layer. Each physical interface uses encoding and modulation specially designed for optimum performance in the transmission medium used.

The Ethernet data link layer is formed by two sublayers; the *Media Access Control* (MAC) sublayer and the *Logical Link Control* (LLC) sublayer:

- The MAC sublayer describes how a station schedules, transmits and receives data in a shared or dedicated media environment. It generates source and destination addresses to identify the two
-

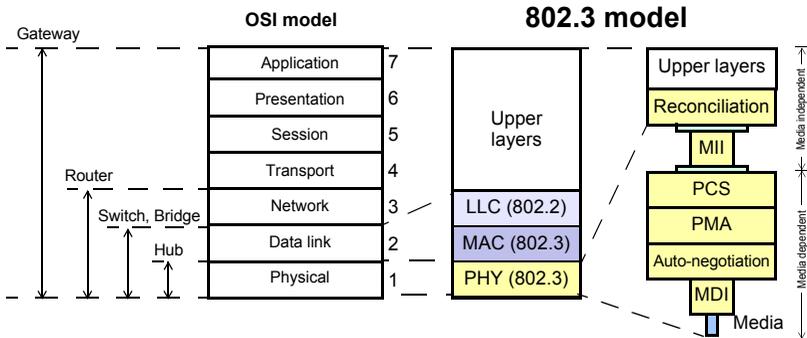


Figure 0.3 Ethernet layers vs. OSI model. Some layers are optional, depending on the version.

ends of the communication process. It also ensures reliable transmission across a link that may be shared, synchronizes data transmission, recognizes errors and controls the data flow.

- The LLC sublayer enables higher layers to ‘talk’ to the hardware-specific MAC layer through a common interface. The Ethernet LLC is shared with other IEEE-based technologies such as Token Ring, and even with other non-IEEE technologies like the *Fiber Distributed Data Interface* (FDDI). The LLC sublayer that they all use is defined in the IEEE 802.2 standard.

PHY and MAC Layer Independence

One of the aims of Ethernet has been to provide media-independence by separating controllers and transceivers, both functionally and physically:

1. *Controllers* hold the common functionalities, such as MAC protocol and interfaces with higher layers.
2. *Transceivers* are specific for each type of media, and they include functions such as codification or traffic functions.

Attachment-Unit Interface

When 10BASE-5, the first commercial Ethernet solution was manufactured, it could only be operated over thick coaxial cable. The evolution towards multiple physical media started with the introduction of the *Attachment Unit Interface* (AUI), developed for rates of 10 Mb/s. The intention was to avoid the difficulty of routing thick and inflexible coaxial cable to each station (Figure 0.4). The AUI is used for coaxial implementations of Ethernet, including 10BASE-5 (Thicknet) and 10BASE-2 (Thinnet). With the advent of 10BASE-T, it became more common to include the physical and MAC layers 'in the same box', and the use of an external AUI started to decline.

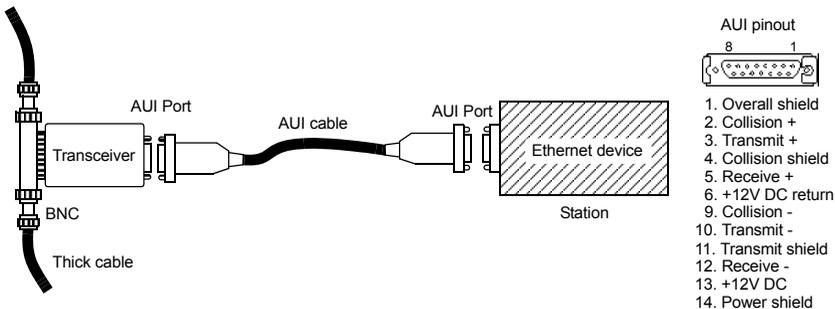


Figure 0.4 The AUI is a little bit more than a connection cable between the Ethernet card and the transceiver.

The AUI connector is a 15-pin DA-15, and the AUI cable can be used for distances of up to 50 m. The AUI includes four types of signals: Transmit data, Receive data, Collision presence and Power.

Medium-Independent Interfaces

The *Medium-Independent Interface* (MII) is the equivalent of the AUI for Fast Ethernet (100 Mb/s). It connects the block by implementing the 100 Mb/s MAC to an Ethernet transceiver. This interface was designed to guarantee the use of Fast Ethernet by different applications; for example, desktop equipment would use Unshielded Twisted Pair (UTP), whereas fiber would be used in backbones. The MII can connect two chips on the same printed circuit board, or two physically different devices by using a pluggable connector.

There are extensions to the MII interface for higher rate Ethernet versions.

The Ethernet PHY

Ethernet has adopted many different transmission media, including coaxial cables, Unshielded Twisted Pair (UTP), Shielded Twisted Pair (STP), Multimode Fiber (MMF) and Single Mode Fiber (SMF) in order to meet the changing market needs.

Generally, the new versions of Ethernet can be used with traditional physical media to enable smooth migration. In some cases, to speed up the development and time-to-market, Ethernet has also adopted some physical layers that were actually designed for other technologies, for example Fiber Channel.

Legacy Ethernet Interfaces

The first Ethernet networks were based on coaxial cable and bus topologies, but many of them were upgraded to UTP cables and star topologies in the 1990s, because these are easier to handle and less expensive.

Optical fiber was introduced so that it could be used where electrical cable cannot; for vertical cabling of LANs, for campus

network backbones, and for environments with high levels of interference.

Ethernet over Coaxial

The original PHY included in the IEEE 802.3 standard is known today as 10BASE-5 or Thicknet. The first Thicknet implementations date back to the early 1970s. This Ethernet version uses a thick coaxial cable with a diameter of 10 mm to transmit 10 Mb/s signals. However, the average system throughput is limited to a few megabits per second, due to the limitations imposed by the multiple access mechanism and some other factors. The coaxial cable has to be terminated with a $50\ \Omega$ resistor to avoid reflections. This system makes it possible to connect up to 100 stations to the same cable segment following a bus topology.

The size of a Thicknet is limited to 2500 m due to the limitations imposed by the multiple access protocol (see Paragraph). The maximum 2500 m long Thicknet is formed by five 500-meter segments separated by four repeaters, but stations can only be attached to three of them (this is known as the 5-4-3 rule). 10BASE-5 uses a simple Manchester code to transmit data (see Figure 0.5).

The thick coaxial cable used in 10BASE-5 Ethernet networks is difficult to install and handle. This is the reason why the 10BASE-2 interface was defined. 10BASE-2 Ethernet networks, or Thinnets, use a thin RG-58 coaxial cable. Thinnet can only reach 185 m per segment, as opposed to the 500-m reach of 10BASE-5. The 5-4-3 rule is still valid, however. The maximum number of stations connected per segment is 30.

Ethernet over UTP

In 1990, the IEEE adopted the 10BASE-T interface in the IEEE 802.3i standard for Ethernet transmission over Category-3 (or better) UTP cables. A traditional UTP cable is formed by four twisted pairs connected to 8-pin RJ-45 connectors. UTP cables commonly use

0.5 mm (24 AWG) wires. 10BASE-T needs only one pair for data transmission and one for reception. The other two pairs are not used.

The maximum length of a 10BASE-T segment is 100 m. The 5-4-3 rule (five UTP segments and four repeaters) still applies, but the limitation of three segments does not make sense for this interface, because the bus topology is replaced by a star configuration.

Like its predecessors, 10BASE-T uses the Manchester code, but now the signal is predistorted to improve transmission over the new medium. 10BASE-T is also the first interface that implements the link integrity feature that makes installation and troubleshooting easier: it sends periodical 'heartbeat pulses' that enable remote stations to recognize physical connection with other devices in the network.

Coaxial cable offers a better performance than UTP, but the 10BASE-T system benefits from structured cabling based on central repeaters and a star-shaped, hierarchical wiring topology. The new topology is superior to the point-to-point, unstructured and single-failure-point bus topology of coaxial cable networks. Ethernet over UTP has become a real success. Today, coaxial cable has virtually disappeared from the local area network.

The 100BASE-T family of interfaces, also known as Fast Ethernet, was specified in May 1995, and it is a 100 Mb/s extension to 10BASE-T. It keeps the same MAC layer as the 10 Mb/s Ethernet, including the frame structure, but it defines a new PHY. There are three different PHY specifications for electrical Fast Ethernet and one more for Fast Ethernet over optical fiber. The electrical Fast Ethernet interfaces are the following:

- 100BASE-TX – requires two pairs of Cat. 5 UTP. This means that 100BASE-TX may not operate if 10BASE-T cabling is used.
-

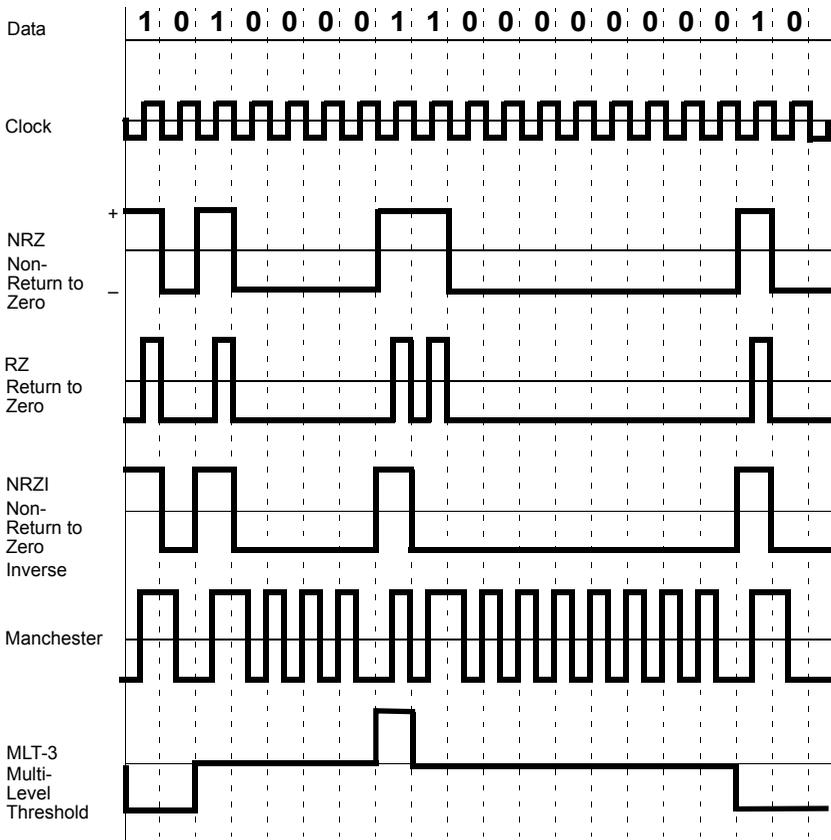


Figure 0.5 Line encoding technologies. Depending on the media, a "+" signal corresponds to high voltage on copper or high intensity on optical fiber, and a "-" signal to low voltage or low intensity. PAM5 uses 5 levels (-2, -1, 0, 1, 2), several pairs (two in 100BASE-T and four in 1000BASE-T), and a complex encoding rule to generate the symbols transmitted in parallel over each pair.

-
- 100BASE-T4 – needs four pairs of Cat. 3 UTP cables. This interface can be used when Cat. 5 cabling is not available; when upgrading old 10BASE-T installations, for instance. However, this interface has never been widely used.
 - 100BASE-T2 – calls for two pairs of Cat. 3 UTP cables. This interface was specified about one year later than the other 10BASE-T interfaces, and it has not been used in commercial devices.

Almost every electrical 100 Mb/s Ethernet link is based on the 100BASE-T interface. This PHY is based on the FDDI physical layer. It encodes the data stream with the 4B/5B encoding method and uses the *MultiLevel Threshold-3* (MLT-3) line code for signal transmission (see Figure 0.5).

The 100BASE-TX interface has the same advantages as the 10BASE-T, but better performance in terms of bandwidth. Both line rates can be used in the same network by using switches. In this case, 10 Mb/s links can be used for connections to workstations, while 100 Mb/s offers inexpensive bandwidth for connections to servers.

Ethernet over Optical Fiber

The first fiber Ethernet standard was the 10BASE-F standard, released in 1993. These interfaces use duplex *MultiMode Fiber* (MMF) as the transmission medium to transmit infrared light. In fact, 10BASE-F refers to a family of three interfaces: 10BASE-FL, 10BASE-FP and 10BASE-FB. The 10BASE-FL (*L* for Link) is meant for connecting stations, repeaters and switches, 10BASE-FB (*B* for Backbone) is for backbone repeaters, and 10BASE-FP (*P* for Passive) is for use in passive central repeaters. The most important 10BASE-F interface is the 10BASE-FL that is based on and is backwards-compatible with the *Fiber Optic Inter-Repeater Link* (FOIRL) specification.

As well as 10BASE-F, the 100BASE-FX interface uses duplex MMF. Although not standard, some vendors have supplied 100BASE-FX

over *Single-Mode Fiber (SMF)*. In this case, the PHY is based on the FDDI physical layer.

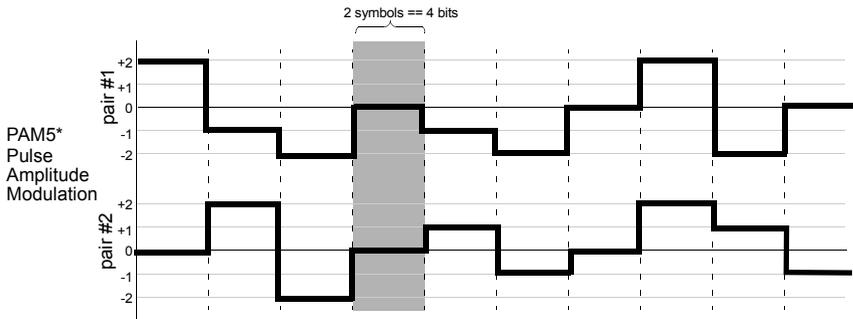


Figure 0.6 Line encoding technologies. Depending on the media, a “+” signal corresponds to high voltage on copper or high intensity on optical fiber, and a “-” signal to low voltage or low intensity. PAM5 uses 5 levels (-2, -1, 0, 1, 2), several pairs (two in 100BASE-T and four in 1000BASE-T), and a complex encoding rule to generate the symbols transmitted in parallel over each pair.

Hands-on: Good Cabling Practices

Twisted pairs have successfully replaced coaxial cable in LAN applications, due to cost-effectiveness, ease of installation and the ability to build simple and flexible cabling systems.

Cabling for data applications has been addressed in many different standards and recommendations, the most important ones being EIA/TIA-568-B (USA), CENELEC EN 50173 (Europe) and ISO/IEC 11801 (international). These standards define cabling types,

distances, connectors, architectures, performance and testing practices.

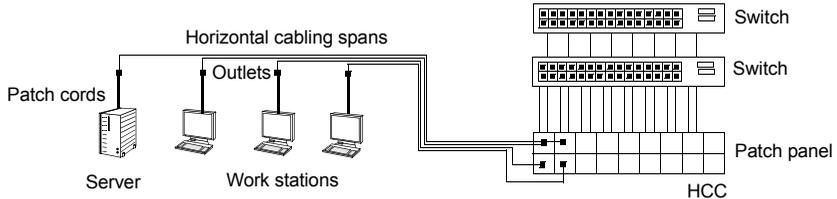


Figure 0.7 Simple, structured cabling layout. All stations are connected to a central location in a star topology.

The current cabling standards specify an extended star topology for data networks. In the simplest case, user equipment and servers are all connected to a central location known as Horizontal Cross-Connect (HCC) by means of so-called horizontal cabling spans. In the HCC users are connected to services and to each other, and most likely to an external network as well (see Figure 1.7). Horizontal spans are terminated in outlets usually installed in the walls of the building. User equipment is connected to the outlets by using patch cords. This cabling model is clearly not enough to interconnect more than just a couple of users. Larger networks are structured hierarchically. There is at least one HCC on every floor of the building. One of the floors hosts an equipment room with the Main Cross-Connect (MCC) that interconnects floors between them by using backbone cable spans (see Figure 1.8). In campus networks, there is one more hierarchical level. Each building in the

network has one Intermediate Cross-Connect (ICC), and the MCC interconnects ICCs from different buildings.

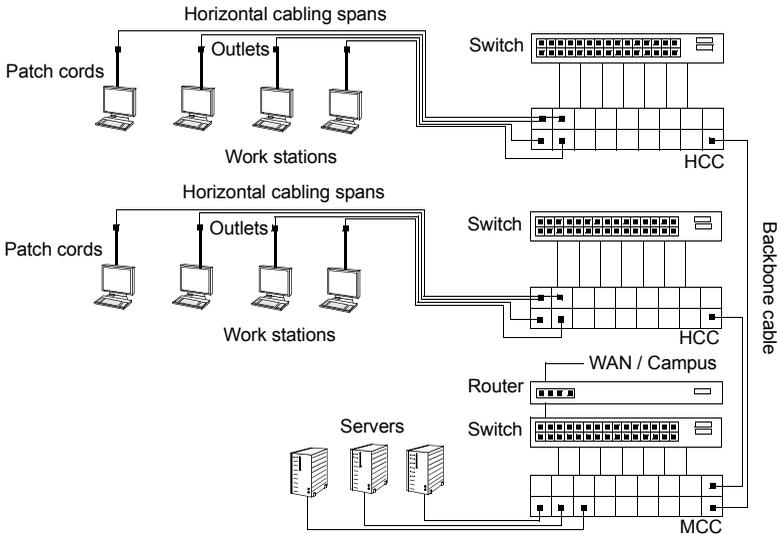


Figure 0.8 Typical cabling layout, including both horizontal and vertical cable spans. The network is structured with an extended star topology centered in the MCC.

Network operators can use different types of optical and electrical cable. Copper pair is just one of the choices. UTP is generally used for horizontal cabling and patch cords. Only under special circumstances UTP is replaced by screened cable or optical fiber. Things are different in backbone cable spans. In this case, MMF is the preferred choice for new installations, but UTP is also found in old installations. For interconnections between buildings, either MMF or SMF is used. The use of electrical cables is discouraged, due to their bad performance in terms of range and bandwidth, but also to avoid potential earthing/grounding problems.

If twisted pair is chosen, one must still decide which type of cable should be used (see Table 0.1). Even in the case of UTP cables, LAN operators can choose from different types. Choosing the correct cable category is one of the most important decisions (Table 0.2). The performance of twisted pairs depends on the cable category used, and this limits the type of applications that can be used in a particular network. All new LAN installations should use at least Cat 5e cable (or Cat D in ISO terminology) to support Ethernet operating at 1 Gb/s. Ethernet at 1 Gb/s should also work in existing networks using Cat 5 cables, but not in those that use Cat 3. All these cable categories can be implemented with UTP. Screened versions of these cables are used where for some reason UTP cannot be used. For example, in environments where high electromagnetic interferences are expected.

Newer categories 6a and 7 are designed with 10 Gb/s applications in mind. Category 6a can be built with UTP but versions with shielded cables are also very common. Category 7 is built with cable known as S/STP that includes one metallic shield wrapped around the pairs and also an individual shielding jacket for each pair. 10G/s can also be deployed over existing cat. 6 cables but range is reduced to 55 m. For this reason, these deployments are not recommended.

Cables of different categories are different in many ways. Performance-wise, what really makes them different is crosstalk (see Table 0.3). For example Cat 6 (ISO Cat E) causes less crosstalk (and is much more tolerant to crosstalk) than Cat 5. This is the main reason why Cat 6 is much better for high-speed data transmission. The reason why some UTP cables have better performance than others is narrow twisting. Twisting keeps the electromagnetic field generated by charges and currents within a smaller area. The narrower the twisting is, the better the electromagnetic fields are confined, and as a result, the pair has better crosstalk performance.

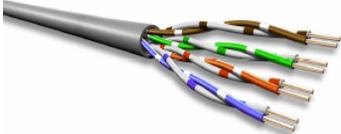
Cable type	Comments
Unshielded Twisted Pair (UTP) 	The most common (and low-cost) twisted pair. A UTP cable usually contains four different copper pairs. However, it is also manufactured in bundles of 25 pairs and even more.
Foil Twisted Pair (FTP) 	The foil-screened version of the twisted pair is known as FTP. In this type of cable, a metallic shield is wrapped around all pairs. FTP is thicker, more expensive and more difficult to handle than UTP. The shield must be earthed/grounded, otherwise performance might be even worse than with UTP. However, if used with care, FTP has better performance than UTP.
Shielded Twisted Pair (STP) 	STP is a cable where individual pairs are shielded from each other. The shield protects signals from being damaged by crosstalk or external interferences. STP has the same disadvantages as FTP, but it offers better performance.

Table 0.1 Common twisted pair types for data networks.

Even UTP cables within the same category are not exactly the same. There are UTP cables with flexible stranded core and with solid core. Solid core cable is usually cheaper, and offers better electrical performance. However, it is not very flexible and it is difficult to terminate. This is why solid core cable is generally used for inside-wall wiring. On the other hand, flexible stranded cable is well-suited for patch cords. The type of cable to be used for different installations also depends on fire safety standards, local building codes and sometimes on other regulations as well.

There are several factors that can cause that these cables do not to meet the corresponding standard; for example bending, stress, and so on. Cables must be handled with care and compliance with regulations must be tested after installation. ISO and TIA standards

specify the parameters to test. Wire map and crosstalk must always be tested. Other standard tests are cable length, insertion loss, return loss, propagation delay and delay skew.

EIA/TIA Category	Bandwidth	Common Application
1	-	Telephony, ISDN BRI
2	4 MHz	4 Mb/s Token Ring
3	16 MHz	Telephony, 10BASE-T, 100BASE-T4 (four wires)
4	20 MHz	16 Mb/s Token Ring
5	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mb/s ATM
5e	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mb/s ATM
6	250 MHz	1000BASE-T (four wires), 10GBASE-T (55 m maximum)
6a	500 MHz	10GBASE-T
7	600 MHz	10GBASE-T
7a	1000 MHz	-

Table 0.2 Twisted Pair Categories.

Twisted pair cable for Ethernet LAN applications generally comes in groups of four pairs (8 wires). Only two of the four pairs carry information (10 and 100 Mb/s). However, all four pairs are used in 1 Gb/s and 10Gb/s operation. There are many possible interconnections that would work, but only two of them are standard. These are known as T-568A and T-568B wire maps (Figure 0.9). All four pairs are always connected, but for 10 and 100 Mb/s operation only pairs 2 and 3 are used. Cables may have poor performance or may not work at all if wires are not connected properly (see Figure 0.10).

Category	Cat 6	Cat 5e	Cat 5
DC resistance (20°C, 100 m) (max.)	9.5 Ω	9.5 Ω	9.5 Ω
DC resistance unbalance (max.)	2%	2%	5%
Mutual capacitance (100 m) (max.)	5.6 nF	5.6 nF	5.6 nF

Table 0.3 Typical specifications of commercially available UTP cables.

Category	Cat 6	Cat 5e	Cat 5
Worst-case cable skew (100 m)	25 ns	22 ns	45 ns
Nominal velocity of propagation	73%	75%	70%
Loss (20/100/250 MHz)(100 m) (max.)	8/18/30 dB	9/20/33 dB	9/21/- dB
PSNEXT (20/100/250 MHz)(100m)(min)	65/57/48 dB	58/47/41 dB	50/39/- dB

Table 0.3 Typical specifications of commercially available UTP cables.

Once wiring has been verified, crosstalk can be checked. Testers provide numeric results for crosstalk in dB, a pass/fail indication, or both. There are different types of values related to crosstalk:

- *Near-End Crosstalk* (NEXT) is the relationship between the power transmitted by the disturbing line and the power received by the victim line at the same end where the signal is inserted. It does not depend on the length of the line.
- *Far-End Crosstalk* (FEXT) is the relationship between the power transmitted by the disturbing line and that received by the victim line at the end opposite to where the disturbing signal is inserted. It depends on the line length.

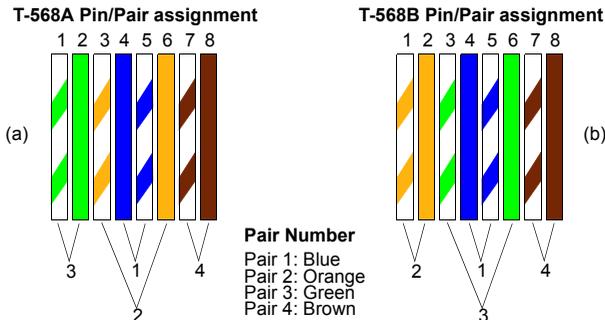


Figure 0.9 T-568A and T-568B wiring standards.

-
- *Equal Level Far-End Crosstalk* (ELFEXT) is defined as the relationship between the power received by the disturbing line and the FEXT power received by the victim pair.
 - *Attenuation to Crosstalk loss Ratio* (ACR) is the difference in dB between the NEXT level registered in the victim pair and the attenuation level in the disturbing line.

Furthermore, if all the disturbing lines are taken into account rather than one single disturbing line, we can also consider the following: Power Sum NEXT (PSNEXT), Power Sum FEXT (PSFEXT), Power Sum ELFEXT (PSELFEXT) and Power Sum ACR (PSACR).

Standards require the measurement of many of these parameters, and provide pass/fail masks for them. NEXT is easier to measure than other types of crosstalk, because it can be measured from a single cable end. Twisting faults located close to the testers are detected with the NEXT test, but faults located far away from this end may remain unnoticed.

Another issue addressed by cabling standards is how and where to connect the tester(s) (see Figure 0.11). This is why channels and links are defined, and test result thresholds are given for each one:

- *Channel*, comprises the cabling link between two Ethernet equipment such as a workstation and a switch or a server. A channel contains fixed elements such as fixed UTP cable and patch panels, and replaceable elements like patch cables, cross-connection cables and equipment cords. Channels may also contain intermediate interconnection elements such as consolidation points. In order to test a particular channel, the Ethernet equipment is replaced by the test equipment at both ends. All cabling and interconnection elements are left as installed for normal network operation.
 - *Links* are defined to ensure that all fixed cabling components meet the standards. Replaceable elements such as patch cords are not taken into account, because they are normally installed
-

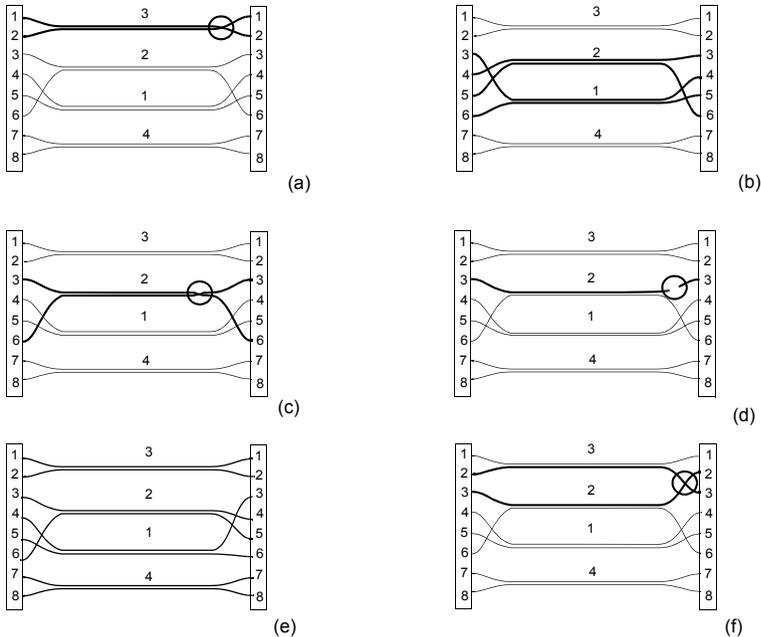


Figure 0.10 Common cabling faults. (a) Inverted cable, polarity is inverted in both ends of the same pair. (b) Pairs 1 and 2 are split. (c) There is a short circuit in pair 2. (d) There is an open circuit in pair 2. (e) Swapped pair, connections are OK but wires of the same connection are twisted in different pairs. (f) Miswired cable.

when all the fixed elements are in place, and they may be changed several times during the life of the network cabling system. When testing a link, cross-connections in patch panels are not tested. The tester is connected directly to the in the termination of the permanent cable in the panel. At the other end, another tester is connected to the outlet. In this setup, a horizontal link can use up to 90 m of fixed cable. These links may contain a consolidation point. Standard patch cords are usually replaced by

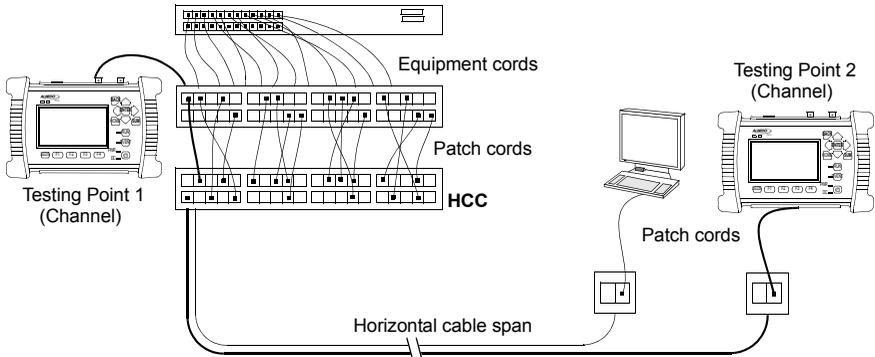


Figure 0.11 Connecting test equipment to the network to check compliance with cabling standards.

specially manufactured high-performance patch cords to minimize the effects of any potential limitations of these elements.

Depending on the measurement, two testers are required: one at each end of the link or channel under test. There are some other tests that call for a line terminated with an open or short circuit, or nominal line impedance. And some measurements require an active Ethernet link.

Hands-on: Testing Auto-Negotiation

The wide variety of Ethernet versions with different physical media, bit rates and protocols have meant that the ability to install a connection without human intervention has become important.

The purpose of *auto-negotiation* is to find a way for two linked stations to communicate with each other, regardless of the Ethernet version that is implemented. Auto-negotiation is performed during link initiation, using the following procedure:

- *Inform* the far end on the Ethernet version and options implemented.
- *Acknowledge* features that both stations share, and *reject* those that are not shared.
- *Configure* each station for highest-level mode of operation that both can support.

Auto-Negotiation for Twisted-Pair Media

Auto-negotiation uses unframed pulses to advertise and respond to optional capabilities. Once both sides have agreed on a common configuration, a logical link is established. The type of station is identified, and the station with most features must reduce its capabilities according to a list of priority resolution criteria (see Table 0.4).

Priority	Type
highest	1000BASE-T full-duplex
.	1000BASE-T
.	100BASE-T2 full-duplex
.	100BASE-TX full-duplex
.	100BASE-T2
.	100BASE-T4
.	100BASE-TX
.	10BASE-T full-duplex
lowest	10BASE-T

Table 0.4 Priority resolution

Auto-negotiation pulses are grouped in bursts known as *Fast Link Pulse* (FLP) bursts. These bursts replace the older *Normal Link Pulses* (NLPs) that were first defined for 10BASE-T signals, to enable hosts to inform remote peers on their availability. One NLP is generated every 16 ms if the transmitter station is not busy. FLP bursts follow the same timing as NLPs, but FLP bursts include from 17 to 33 pulses rather than a single pulse. In an FLP burst, positions 1, 3, 5, 7,

9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33 are always filled by a pulse. All the other 16-bit positions may be filled by a pulse or not, depending on the information transmitted. With this encoding, stations can transmit auto-negotiation information in 16-bit words (see Figure 0.12).

Unlike ordinary Ethernet frames, FLP bursts are made up of unipolar pulses that have two possible values: 0 V or +1 V. Thanks to this feature, the receiver can distinguish FLP bursts from ordinary Ethernet data pulses. FLP bursts have also been designed to be backward compatible with older network interfaces that do not support auto-negotiation

Auto-negotiation was first defined for 10/100 Mb/s interfaces operating over twisted pair, and was later extended to 1 Gb/s and 10 Gb/s. Today, 10BASE-T, 100BASE-T, 1000BASE-T and 10GBASE-T are all compatible.

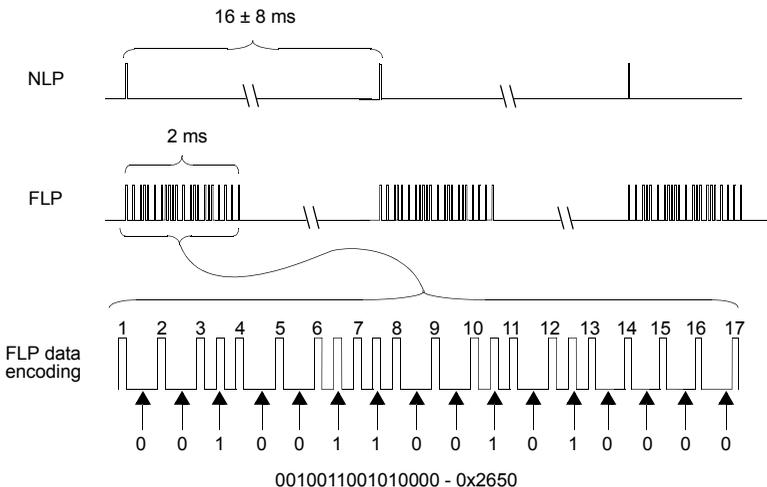


Figure 0.12 NLP pulse train and FLP bursts for 10/100/1000BASE-T auto-negotiation.

Auto-Negotiation in Optical Transmission Media

The various fiber optic Ethernet standards (10BASE-F, 100BASE-FX and 1000BASE-X) use different wavelengths of optical signaling, which makes it impossible to come up with an auto-negotiation signaling system that would work across all three.

Instead, only the 1000BASE-X fiber optic media system has a specification for auto-negotiation. 1000BASE-X auto-negotiation is used to determine if half-duplex or full-duplex mode is used. Flow control and remote fault indications are also decided.

The 1000BASE-X Auto-Negotiation standard is defined in Clause 37 of the IEEE 802.3 standard. Auto-negotiation over optical interfaces uses 16-bit words, but it cannot be based on FLP bursts. Instead, they use reserved combinations of 8B/10B codes used in 1000BASE-X interfaces. A message containing all negotiable parameters is interchanged between the two stations connected through a link.

Verification of the Ethernet Auto-negotiation

Misconfigured auto-negotiation has been reported to be responsible for many problems with Ethernet. This makes auto-negotiation a very relevant topic. Operators must make sure that their Ethernet network interfaces support auto-negotiation. Sometimes, auto-negotiation problems make data exchange impossible, other times these problems cause severe service degradation but not total loss of connectivity. A good example of this is the so-called duplex mismatch. This occurs when one device configured for full-duplex operation is connected to a half-duplex device. The full-duplex device receives and transmits data simultaneously, but the half duplex device diagnoses all data received during transmission as collisions and it will attempt to retransmit the affected data. Some collisions may be detected as late collisions. In this case, the Ethernet network interface will not attempt retransmission, and error recovery will be left to upper protocol layers. This results in very poor performance.

Auto-negotiation is done by exchanging 16-bit words. The main word is the Auto-negotiation base page. In this word, network interfaces choose the operating interface (10BASE-T, 100BASE-T, 100BASE-T4), full/half duplex operation and flow control configuration). However, basic auto-negotiation can be extended by using more 16-bit words. *Next page* is used to indicate that there are additional features to negotiate. A common application for Next page is auto-negotiation for 100BASE-T2 and 1000BASE-T. When configuring a 100BASE-T2 interface, Next page is followed by a third message carried by an *unformatted page*. When configuring 1000BASE-T, Next page is followed by two unformatted pages (see Figure 1.27). Unformatted pages are used to exchange extra parameters between peers and negotiate additional features such as full/half duplex operation of gigabit links, port type (single-port device, multiport device) and synchronization master/slave roles. Master/slave roles can be manually configured by the user. If the user does not configure anything, and if one side is a multiport device, its clock is used as the master, but if both sides have a similar device, a randomly-generated seed is used to decide which station is master.

Optical configuration is simpler than electrical configuration, because there are less parameters to use. Especially, 1000BASE-X auto-negotiation enables the configuration of flow control and full/half duplex features of gigabit optical links by means of a single auto-negotiation base page (see Figure 0.14).

As mentioned before, one of the problems with auto-negotiation is that duplex mismatch may occur. Duplex mismatch can be detected by carrying out a performance test through a chain of switches and routers. For this test, a traffic generator/analyzer is used at one end of the chain, and a loopback device at the other end. The generator/analyzer injects test traffic into the network, and the loopback device sends the traffic back to the originator for analysis. If there is a duplex mismatch in the interconnection between two switches, performance is greatly affected. The reason

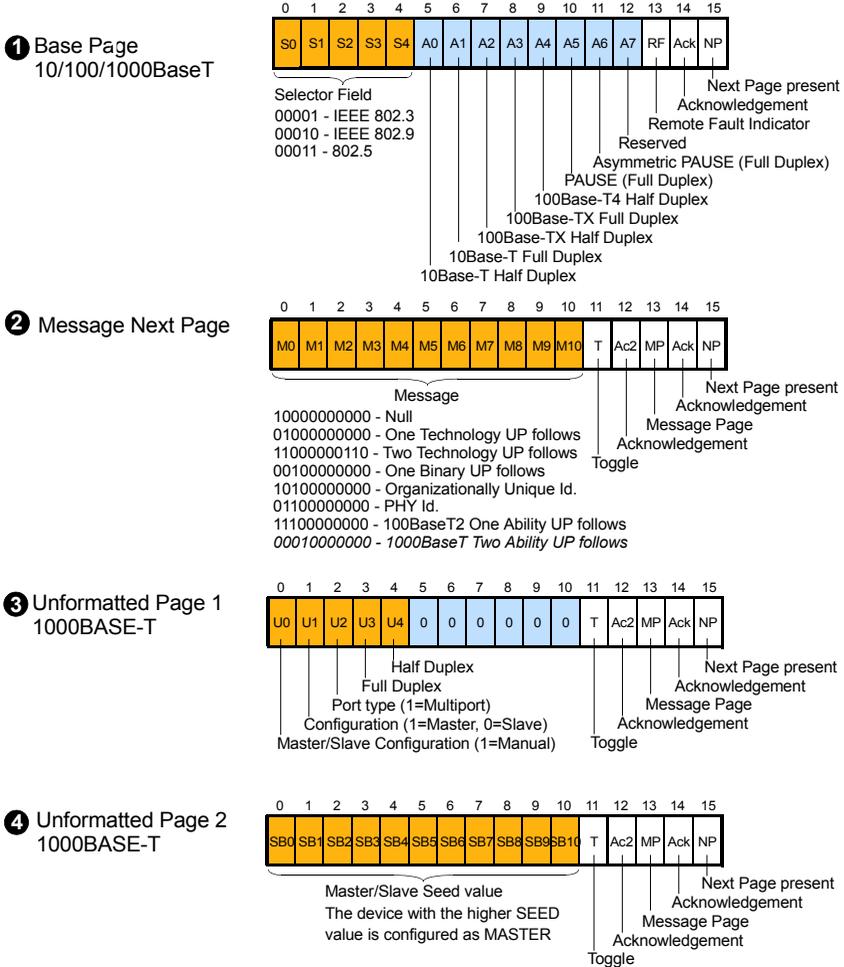


Figure 0.13 Twisted-pair Ethernet auto-negotiation protocol. Auto-negotiation for 10/100 Mb/s requires exchange of one single message (a auto-negotiation base page). To communicate their ability to transmit and receive 1000BASE-T signals, a device must exchange four auto-negotiation messages (a base page, a next page and two unformatted pages) with its remote peer.

is that the half-duplex switch detects part of the incoming traffic as a collision and attempts to retransmit data. Retransmissions result in transmission fragments received by the full duplex interfaces connected to the misconfigured switch. The peer discards the received data due to the invalid CRC. The line is overloaded by the transmission fragments that are finally discarded by the full duplex interface.

Auto-Negotiation Base Page 1000BASE-X

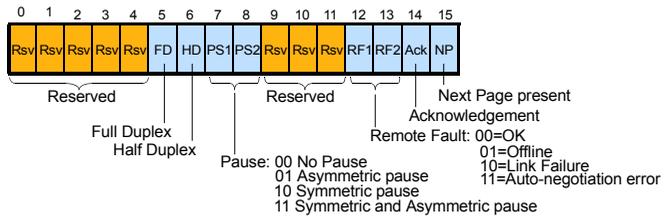


Figure 0.14 1000BASE-X auto-negotiation page description.

If there is reason to suspect that a switch or any other network equipment has a problem with auto-negotiation, tests are needed to confirm correct operation. To test auto-negotiation, it is necessary to use test equipment that is able to configure the preferred and forced values of the parameters to negotiate. The tester can be a dedicated test instrument, but one can also use a switch or some other device with configurable network interfaces. This test setup is used to check the ability of the DUT to negotiate a specific set of parameters. It is even possible to check the effects of a duplex mismatch by configuring the tester for duplex operation if it is connected to a half-duplex network interface.

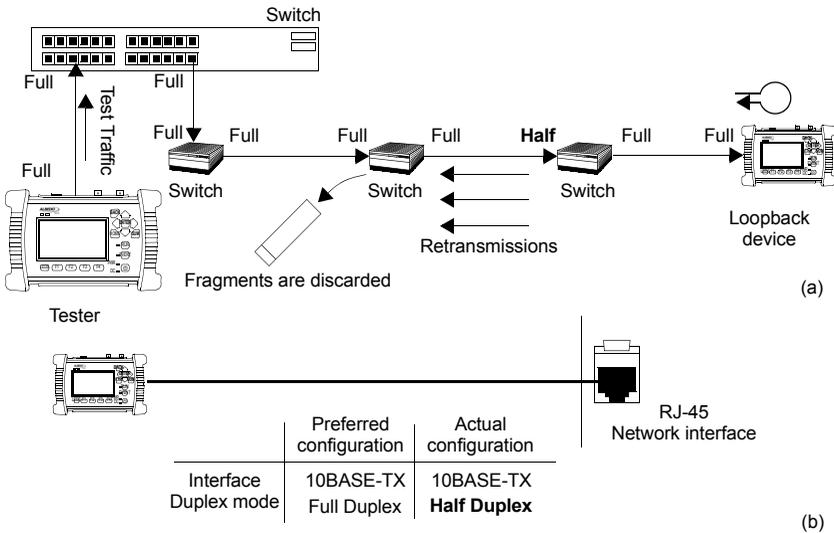


Figure 0.15 Tests related with auto-negotiation. (a) Loopback test to check traffic loss due to a duplex mismatch. (b) Comparison of preferred and actual transmission values in a network interface with unknown features.

The Ethernet MAC

The shared Ethernet medium access protocol is based on the ALOHA mechanism. In ALOHA, stations share the transmission medium by using a simple multiple-access protocol:

1. Any station can transmit a packet at any time, indicating the destination address.

-
2. Once the packet has been sent, the transmitter keeps waiting for the acknowledgment (ACK) from the receiver.
 3. Stations are always listening and reading the destination address of all packets. If a packet received matches the station's address, the station verifies that the *Cyclic Redundancy Check (CRC)* of the packet is correct before answering with a short ACK packet to the transmitter.
 4. If after certain time the ACK is not received by the transmitter, due to a bad CRC or for any other reason, the packet is resent.

The time the transmitter waits for the ACK must be at least twice the latency of the network. This is to allow time for the packet to reach the most distant destination, and then for the ACK to reach the transmitter.

One of the most common CRC errors occurs when two or more stations try to transmit at the same time. This causes interference, making it impossible for any packet to be received. This situation is known as a *collision*.

Collisions mean that the maximum theoretical efficiency of ALOHA-like systems is about 18%. In an improved version, known as Slotted ALOHA, synchronized stations dividing transmit time into windows. To reduce the probability of collisions, stations could only start a transmission at specific times. This increases the maximum efficiency to 36%.

The poor performance of ALOHA-like systems drove the development of CSMA/CD to provide a more efficient *Medium Access Control (MAC)* protocol that would minimize the impact that collisions have on efficiency.

CSMA/CD

The first part of this protocol, the CSMA, forces any station wishing to transmit to Listen to the channel to check if another transmission

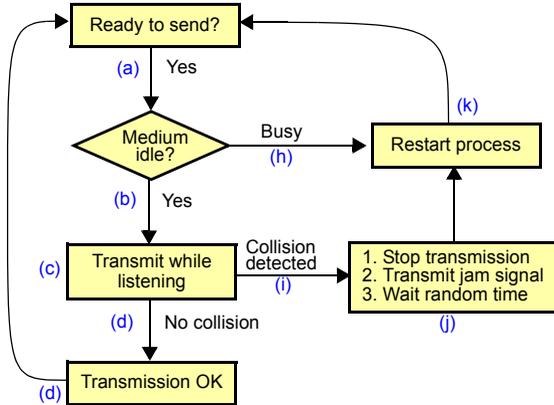


Figure 0.16 CSMA/CD flow chart operation in half-duplex.

is in progress. But, despite the precautions of the CSMA, two or more stations may still attempt to transmit at about the same time, which is when a collision will occur.

Collisions cannot be avoided completely, but their effect can be minimized by reducing the duration of the collision. An important improvement can be made if the station *continues listening to the channel* while transmitting. It will then be able to stop the transmission immediately after a collision is detected (Collision Detection, CD). A collision enforcement *jam signal* is sent, to tell all the stations that a collision has happened (see Figure 0.17). This completes the CSMA/CD protocol. The detailed procedure is the following:

1. Listen to the channel. If a frame is ready to be sent to another station, first check if another transmission is in progress.

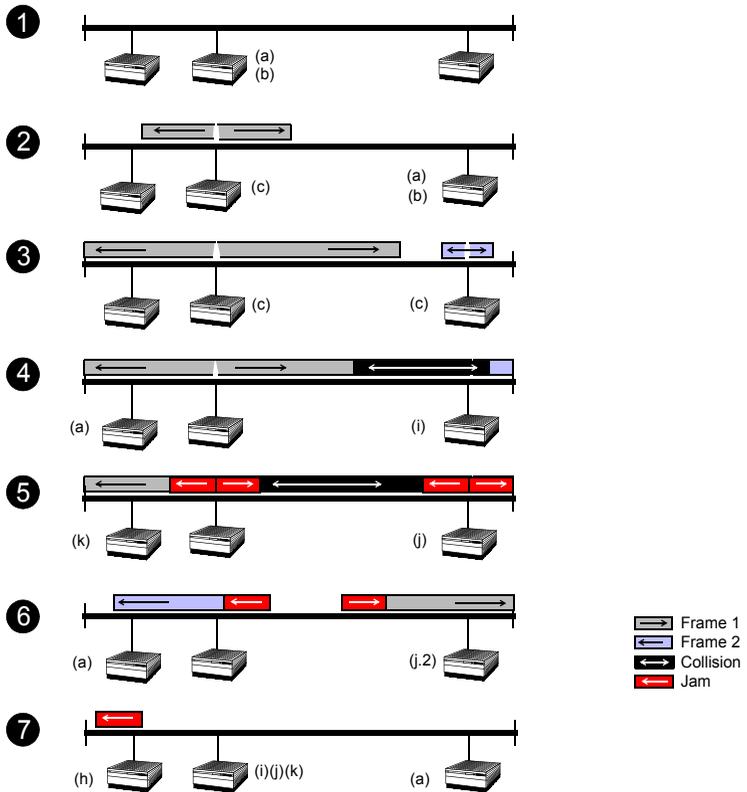


Figure 0.17 Collisions in CSMA/CD.

2. If the channel is idle for a certain minimum period of time, called the *InterFrame Gap* (IFG), start transmission.
If the channel is busy, go to step 1 and start again.
3. If there is no collision, the receiving station checks the CRC value.
If this value is correct, it is delivered to the higher layer. If it is not

- correct, the whole frame is discarded and the process must be restarted from step 1.
4. If a collision is detected during transmission, stop the transmission of the frame and transmit a jamming signal to notify all stations that a collision has occurred. All the stations must stop transmitting.
 5. All the stations must wait during a random time, different for each one, before trying a new transmission. Go to 1.

Collisions

Collisions are the result of a propagation delay between stations, and therefore they are a normal part of the operation of half-duplex Ethernet. However, if there is a large number of collisions, network efficiency is severely affected. We can also see that if the transmitter detects the collision before sending the last byte, this reduces the negative effects. To make this possible, frames have to be long *enough* to completely fill the medium. Then, if a collision occurs, the transmitter will detect the collision and restart the process, rather than waiting for an ACK that never arrives.

Parameter	10 and 100 Mb/s	1000 Mb/s
Slot Time	512 bit times	4096 bit times
Minimum Inter-frame gap	96 bit times	96 bit times
Maximum attempts	16	16
Back-off limit	10	10
Size of jam signal	32 bits	32 bits
Maximum frame size	1518 bytes (12144 bits)	1518 bytes (12144 bits)
Minimum frame size	512 bits (64 bytes)	512 bits (64 bytes)

Table 0.5 Ethernet Timing parameters (half-duplex operation)

To completely fill the medium, frames must have a minimum size to compensate for propagation delays and other types of delays before they reach the edge of the network. Ethernet Transmitters always wait during a certain number of slot times (integer numbers

only) before retransmitting the data again when they detect a collision. In 10 Mb/s and 100 Mb/s, the slot time matches the *Minimum Frame Size* (MFS). GbE networks operating in half-duplex mode are faster and the time slot for them is longer. This will make sure that all stations in the network detect collisions on time (see Table 0.5). When full-duplex versions of Ethernet are used, collisions are avoided, which is why the concept of slot times does not apply.

In some exceptional cases, a late collision may occur after transmitter has sent the last byte. In this case, the CSMA/CD layer is not aware that a collision has occurred, and hence it will not try to resend the packet. Higher-layer protocols will therefore need to resend the packet

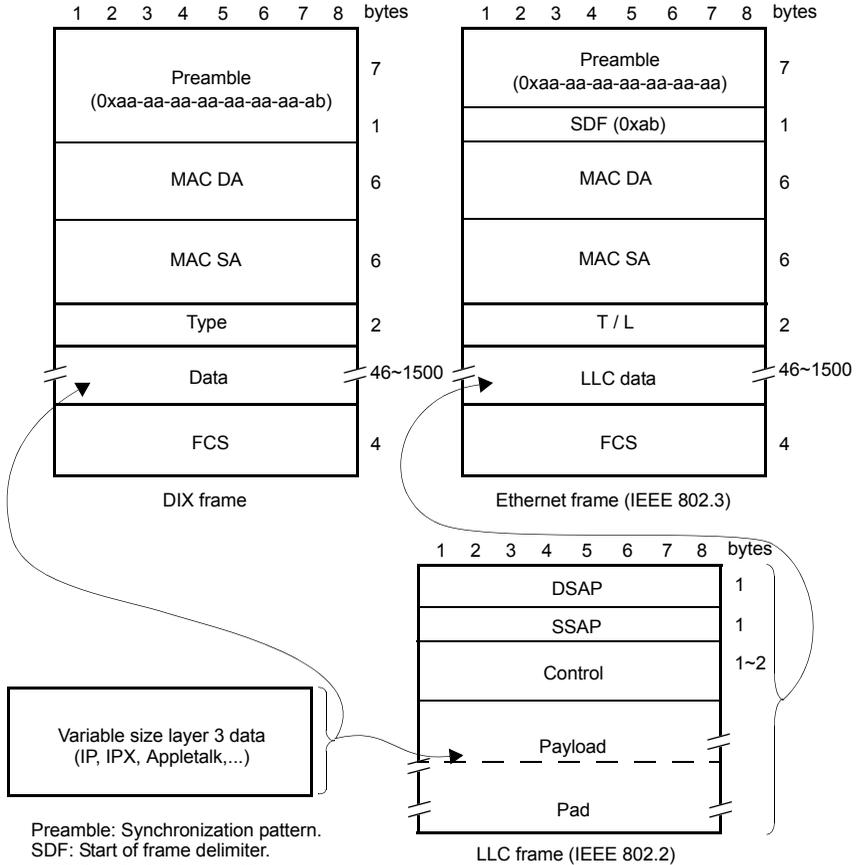
The Collision Enforcement Jam Signal

Transmitters in a shared Ethernet network replace the original signal by a 32-bit jam signal when they detect a collision. If the collision was detected during the 64-bit preamble, the preamble is still sent out, but the 32-bit jam signal is appended to it, so that a minimum of 96 bits is transmitted.

The jam signal ensures that all stations are aware of a collision that has occurred. Repeaters must reinforce the detection of a collision by retransmitting the same collision signal on all ports. This way, all devices connected to the ports are aware of the collision (see Figure 0.17).

The Ethernet Frames

The DIX frame was the first format adopted by the DIX cartel. In 1983, when the IEEE released the first 802.3 standard, the *Start Frame Delimiter* (SFD) field was defined, and this was little more than just a name change. More important was the Length field, since this allows management of the padding operation at the MAC layer, rather than passing this function to higher protocol layers.



Preamble: Synchronization pattern.
 SDF: Start of frame delimiter.
 MAC DA: Destination MAC Address.
 MAC SA: Source MAC Address.
 Type: Indicates the nature of the client protocol (IP, IPX, Appletalk,...).
 T/L: Number of bytes of the LLC data if less than 0x0600, otherwise indicates the payload type.
 FCS: Frame Check Sequence CRC code based on all the fields except the preamble and SDF.

DSAP: Destination Service Access Point
 SSAP: Source Service Access Point
 Control: Miscellaneous LLC control information.
 PAD: Bytes added to ensure a minimum frame size of 64 bytes.

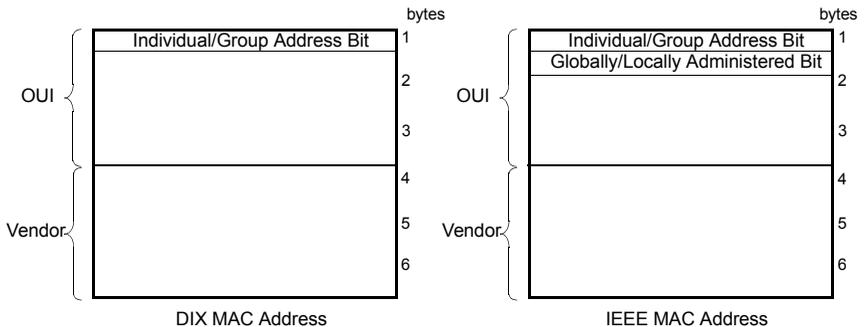
Figure 0.18 The basic 802.3 MAC frame format.

In 1997 the IEEE accepted the use of both Type and Length interpretations of the field that had previously been Type in DIX frames and Length in IEEE 802.3 (1983) frames (Figure 0.18).

Frame Fields

The structure of an IEEE 802.3 'Ethernet' frame is the following:

- *Preamble*, a sequence of 7 bytes, each set to '10101010'. Used to synchronize the receiver before actual data is sent.
- *Start Frame Delimiter (SFD)*, One byte of alternating 1s and 0s, the same as the preamble, except that the last two bits are 1. This is an indication to the receiver that anything following the last two 1s is useful and must be read into the network adapter's memory buffer for processing.



OUI: Organizationally Unique Identifier, IEEE-administered code

Vendor: Administered by the manufacturer.

Individual/Group Address bit: Indicates whether the address is unicast (0) or multicast. Broadcast MAC address is ff-ff-ff-ff-ff-ff.

Globally/Locally Administered bit: Indicates whether the frame is globally unique (0) or locally administered (1).

Figure 0.19 The 24-bit block administered by the IEEE is known as the *Organizationally Unique Identifier (OUI)*. A vendor obtains an OUI number and has another 24-bit block to build up to 2 exp 24 Ethernet devices.

-
- *Destination (MAC) Address, Source (MAC) Address (DA, SA)*, There are three types of addresses: a) *unique*, 48-bit address assigned to each adaptor, each manufacturer gets their own range; b) *broadcast*: all 1s, which means that all the receivers must process the frame; c) *multicast*: first bit is 1 to refer to a group of stations (Figure 0.19).
 - *Type*, A descriptor of the client protocol being transported (IP, IPX, AppleTalk, etc).
 - *Length*, The size of the data field, not including any pad field added to obtain minimum frame size. The maximum size is 1518 bytes (preamble and SDF are not included).
 - *Logical Link Control (LLC)*, The payload, can contain from 48 up to 1500 bytes of data.
 - *Pad*, All frames must be at least 64 bytes long. If the frame is smaller, it contains a pad field to reach the necessary 64 bytes.
 - *Cyclic Redundancy Check (CRC)*, the value of this field is used to check if the frame has been received successfully, or if the contents have been corrupted.

Hands-on: Determining Support of Jumbo Frames

Ethernet frames are usually assumed to have a maximum size of 1518 bytes, accounting for a 1500-byte payload, a 14-byte header (source and destination addresses, type/length byte) and a 4-byte trailer (FCS). However, things are not always as easy as they seem. In fact, frames longer than 1518 bytes do exist. Frames carrying *Virtual LAN (VLAN) Q-tags* are longer. Single Q-tagged IEEE 802.1Q frames have a *Maximum Transport Unit (MTU)* of 1522 bytes and double Q-tagged IEEE 802.1ad frames have an MTU of 1526 bytes. Frames carrying MPLS labels may also be longer than 1518 bytes. All these frames are accepted by the standards and should also be accepted by switches and routers built according to these standards. There are also examples of proprietary frame formats with MTU longer than 1518. A good example is the Cisco Inter-Switch Link (ISL)

frame, an encapsulation used to tag frames in trunk links (see Figure 0.20).

Frames with an MTU that is only slightly larger than 1518 bytes, like Q-tagged frames, are known as baby giant frames but jumbo frames are often six times longer than the longest regular Ethernet frame. This is still far from the 64-KB limit for IPv4 packets, and IPv6 allows packets as long as 4 GB. However, you do not usually see frames much longer than 9 KB, because the 32-byte CRC protection included in the Ethernet trailer is not effective with very long frames. To support these frames, it would be necessary to modify the entire frame structure.

Those who support 9000-byte long frames claim that while Ethernet is now 1 000 times faster than the original 10 Mb/s, the MTU still remains the same. Increasing the MTU (at least for high speeds) increases throughput and reduces processing load in network nodes and end-user equipment. The inconvenience of long frame sizes is that they make interactive communications difficult. Long frames suffer from increased delay, which is why they are not suitable for audio and video applications, such as *Voice over IP* (VoIP) or IPTV. Long frames may also damage the *Quality of Service* (QoS) of short frames when they are all queued together in the same buffer in a network node. Jumbo frames may not be an issue in a backbone operating at 10 Gb/s, but even in this case, data may have been aggregated from slower interfaces such as 100 Mb/s Fast Ethernet. The main applications for jumbo frames are therefore data applications, and, more precisely, *Storage Area Networks* (SAN).

While many baby giant frames are standard, jumbo frames are still proprietary, so many network equipment manufacturers do not support them. Some switches and routers make it possible to configure the MTU from the console. As a result, there is no global agreement on the MTU of Ethernet frames. If MTU values are

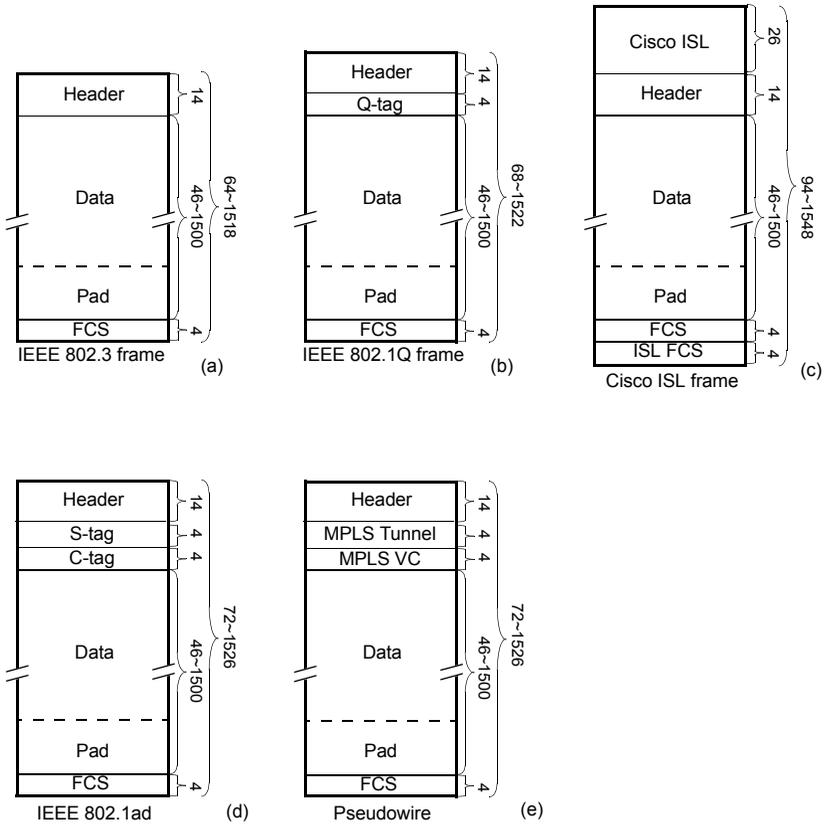


Figure 0.20 (a) The standard Ethernet frame has an MTU of 1518 bytes, (b) Q-tagged VLAN frames have an MTU of 1522 bytes, (c) Cisco Inter-Switch Link (ISL) encapsulated frames have an MTU of 1548 bytes, (d) Q-in-Q Ethernet frames have an MTP of 1526 bytes, (e) MPLS pseudowires built over Ethernet infrastructure have an MTU of 1526 bytes.

different from 1518 bytes, they are usually 9216, 9192, 9180, 9176 or 4470 bytes.

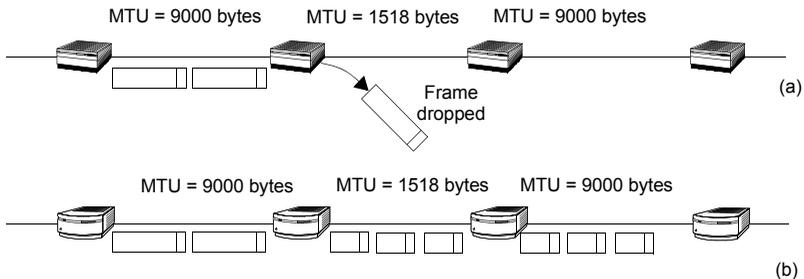


Figure 0.21 (a) Switched path, frames exceeding the S2 MTU are dropped. (b) Routed path, packets are fragmented.

Routers and switches that support jumbo frames can forward this type of data without any problems. Sometimes, IP routers that support packet fragmentation may need to fragment jumbo frames to match the Ethernet MTU configuration. Some routers and switches may fail to forward jumbo frames. If this is the case, these frames may be dropped (see Figure 0.21). Sometimes an *Internet Group Message Protocol* (IGMP) message is generated to inform the transmitter about the event, but this does not always happen. The problems caused depend on the application that is using the problematic link. These problems may be intermittent, and they sometimes depend on the destination. But a closer look often reveals that this behavior can be reproduced if the original conditions are recreated.

Sending and receiving jumbo frames may cause problems, if they are not properly supported by the network. Testing the MTU in critical paths and checking the support for jumbo frames may help (see Figure 0.22). A tester that can check the frame size and compute a histogram can be used to analyze the traffic that passes

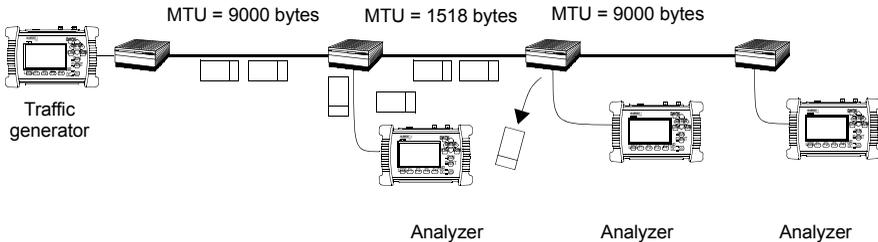


Figure 0.22 (a) Switched path; frames exceeding the S2 MTU are dropped. (b) Routed path; packets are fragmented.

through the network element. The test can be performed in service (without disconnecting users) if the network element has port mirroring capabilities or if the tester supports 'through' or monitoring mode. The traffic to analyze can be either real or non-real. In general, more detailed results can be obtained by using synthetic traffic generated in a tester with traffic injection capabilities. It is also possible to carry out an in-service test with non-real traffic, because the traffic used to analyze the MTU is small and unlikely to cause congestion or damage other services. Tests with traffic injection do not need port mirroring or 'through'/monitor operation. The only requirement is to use the traffic analyzer's MAC address as the destination MAC address. The non-real traffic will then reach the analyzer through a switched or routed path.

The Ethernet LLC

The mission of the *Logical Link Control* (LLC) is to make Ethernet appear to be a point-to-point network, regardless of whether the MAC layer is using a shared or dedicated transmission medium.

The LLC can provide three types of services:

1. *Unacknowledged connectionless service*, which is a simple datagram service just for sending and receiving frames. Higher layers take care of flow and error control.
2. *Acknowledged connectionless service*, where received frames are verified and ACK is sent even if the connection has not been set up.
3. *Connection-oriented service*, which establishes a virtual circuit between two stations.

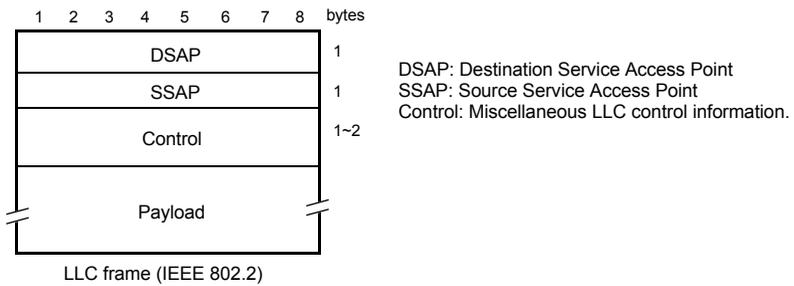


Figure 0.23 Logical Link Control (LLC) format.

Destination Service Access Point (DSAP) and *Source Service Access Point (SSAP)* use one-byte fields assigned by the IEEE to identify the location of the memory buffer on source and destination devices where the data from the frame should be stored.

The control field is either 1 or 2 bytes long, depending on which service is specified in the DSAP and SSAP fields. For example, if the value is 3, which indicates an 'unnumbered format' frame, this means that the LLC uses an unacknowledged, connectionless service.

Selected Bibliography

- [1] IEEE 802.3-2008, "Part 3: Carrier sense multiple access with collision detection (CSMA/CD) Access Method and Physical Layer Specifications", December 2008.
 - [2] Rich Seifert, *Gigabit Ethernet Technology and Applications for High/Speed LANs*, Addison Wesley Oct 1999
 - [3] William Stallings, *Data and Computer Communications*, Prentice Hall, 1997.
 - [4] Kevin L. Paton, *Gigabit Ethernet Test Challenges*, Oct 2001 Test and Measurement World Magazine.
 - [5] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.
 - [6] R. Metcalfe and D. Boggs, "Ethernet: Distributed packet switching for local computer networks", *Communications of the ACM*, vol. 19, no. 7, July 1976, pp. 395-403.
 - [7] 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.
-

Time Division Multiplexing

Appendix C

Time Division Multiplexing

Deterministic TDM

Pulse Code Modulation

PCM involves three phases: sampling, encoding, and quantization:

1. In sampling, values are taken from the analog signal every $1/f_s$.
2. Quantization assigns these samples a value by approximation.
3. Encoding provides the binary value of each quantified sample.

In a telephone channel the bandwidth is set at 4 kHz. Then the sampling frequency must be $f_s \geq 2 \cdot 4,000 = 8,000$ Hz; equivalent to a sample period of $T = 1/8,000 = 125\mu\text{s}$.

In order to codify 256 levels, 8 bits are needed then (v) is:

$$v = 8,000 \text{ samples/s} \times 8 \text{ bits/sample} = 64\text{Kbps}$$

This bit rate is the subprimary level of transmission networks.

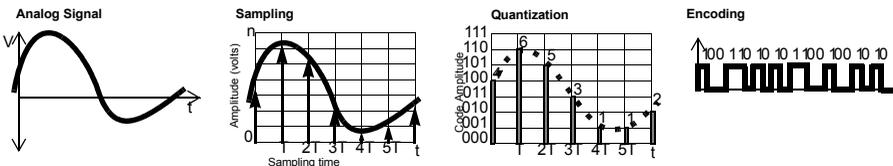


Figure 0.1 Signal digitalization process: sampling, quantization, and encoding.

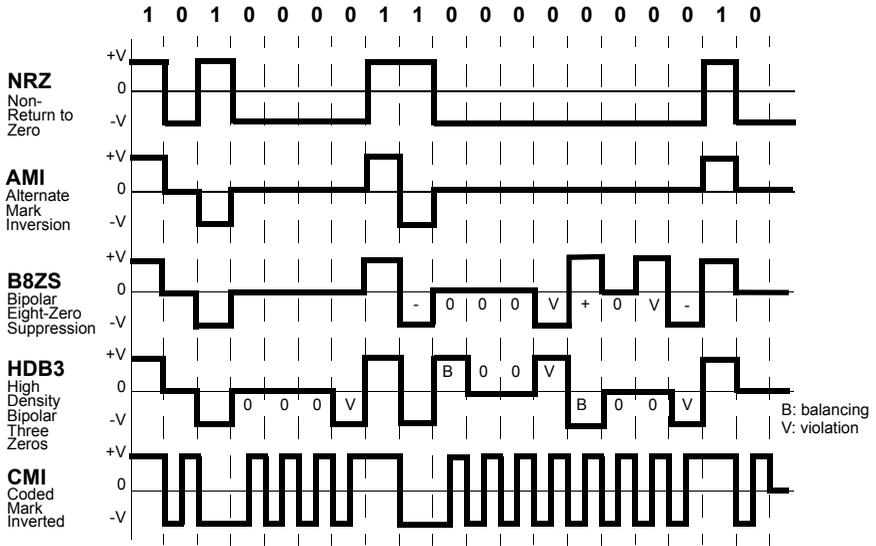


Figure 0.2 Line encoding technologies. AMI and HDB3 are usual in electrical signals, while CMI is often used in optical signals.

Channel Coding

This is the process to converted binary data into signal elements that can cross the transmission medium. The most widely used are:

Non-return to zero

This is a simple method consisting of assigning the bit “1” to the positive value of the signal amplitude (voltage), and the bit “0” to the negative value (Figure 0.2).

Alternate mark inversion

This transmission code, in which a “0” bit is transmitted as a null voltage and the “1” bits are represented alternately as positive and negative voltage.

Bit eight-zero suppression

Bit eight-zero suppression (B8ZS) is a line code in which bipolar violations are deliberately inserted if the user data contains a string of eight or more consecutive zeros.

High-density bipolar three zeroes

High-density bipolar three zeroes (HDB3) is similar to B8ZS, but limits the maximum number of transmitted consecutive zeros to three (Figure 0.2).

Coded mark inverted

The *coded mark inverted* (CMI) code, also based on AMI, is used instead of HDB3 at high transmission rates, because of the greater simplicity of CMI coding and decoding circuits compared to the HDB3 for these rates.

Multiplexing and Multiple Access

Multiplexing is the process by which several signals from share a channel with greater capacity (Figure 0.3). When the sharing is carried out with respect to a remote resource, this is referred to as multiple access. Some multiplexing technologies are:

1. *Frequency division multiplexing* (FDM)
 2. *Time-division multiplexing* (TDM): Assigns all the transport capacity sequentially to each of the channels.
 3. *Code-division multiplexing access* (CDMA): Multiple signals in the same frequency separated by codes.
-

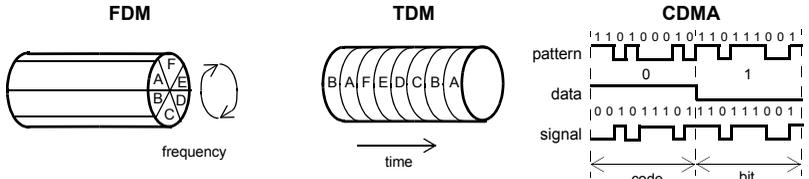


Figure 0.3 FDM is used by WDM, TV, and GSM while TDMA is used in Ethernet, PDH, SONET/SDH. CDMA is used in some wireless networks.

PDH and T-Carrier

The combination of PCM and TDM made up the digital communications systems including PDH and T-Carrier

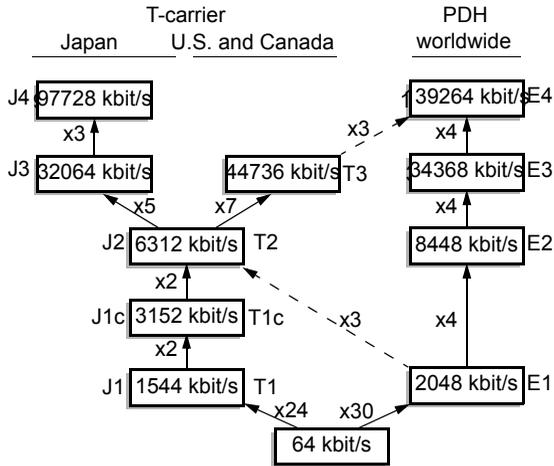


Figure 0.4 The PDH and T-carrier hierarchies, starting at the common 64-kbit/s channel.

Basic Rates: T1 and E1

T1 permits the TDM multiplexing of 24 digital 64 kbit/s channels into one signal (Figure 0.4). A synchronization bit is added to the time slots, then the aggregate transmission rate is:

$$(24_{\text{channels}} \times 8_{\text{bit/channel}} + 1_{\text{bit}}) / 125\mu\text{s} = 1,544\text{Mbps}$$

125 μs is the sampling period

E1 is the European TDM multiplexing scheme which has 32 channels of 64 kbit/s (Figure 0.4), resulting in a rate equals to:

$$(32_{\text{channels}} \times 8_{\text{bit/channel}}) / 125\mu\text{s} = 2,048\text{Mbps}$$

The E1 Frame

The E1 frame defines a cyclical set of 32 time slots of 8 bits. The time slot 0 is devoted to transmission management and time slot 16 for signaling; the rest were assigned originally for voice/data transport (Figure 0.5).

Key characteristics of the E1 frame:

- Frame Alignment, is an indication showing when the first interval of each frame begins. This way, the bytes received in each slot are assigned to the correct channel.
-

- NFAS - Alarm Bit, sent to the transmitter when a device detects either a power failure or a failure of the coder/decoder. The *remote alarm indication* (RAI) is sent in the NFAS of the return frames, with bit 3 being set to "1."
- Signaling can be either Channel, information generated by the users either signaling information.
- 2-Mbit/s frame is used to transmit the signaling information.

The Plesiochronous Digital Hierarchy

Based on the E1 signal, the ITU-T defined a hierarchy of plesiochronous signals that enables signals to be transported at rates of up to 140 Mbit/s. This section describes the characteristics of this hierarchy and the mechanism for dealing with fluctuations in respect to the nominal values of these rates, which are produced as a consequence of the tolerances of the system.

Standard	Binary Rate	Size	Frame/s	Code	Amplitude	Attenuation
G.704/732	2,048 kbit/s \pm 50 ppm	256 bits	8,000	HDB3	2.37-3.00V	6 dB
G.742	8,448 kbit/s \pm 30 ppm	848 bits	9,962.2	HDB3	2.37V	6 dB
G.751	34,368 kbit/s \pm 20 ppm	1536 bits	22,375.0	HDB3	1.00V	12 dB
G.751	139,264 kbit/s \pm 15 ppm	2928 bits	47,562.8	CMI	1.00V	12 dB

Table 0.1 The PDH hierarchy, with four levels from 2 to 140 Mbit/s.

The T-Carrier Hierarchy

As is the case of the PDH, the T-carrier higher levels multiplexing is carried out bit by bit (unlike the multiplexing of 64-kbit/s channels in a DS1 frame, which is byte by byte), thus making it impossible to identify the lower level frames inside a higher level frame. Recovering the tributary frames requires the signal to be fully demultiplexed.

The DS1 Frame

The DS1 frame is made up of 24 byte-interleaved DS0s, the 64-kbit/s channels of eight bits, plus one framing bit that indicates the beginning of the DS1 frame. The DS0 channels are synchronous with each other, and are then time division multiplexed in the DS1 frame. Depending on the application, the DS1 frames are grouped in *superframe* (SF), 12 consecutive DS1 frames, and *Extended Superframe* (ESF), 24 consecutive frames (Figure 0.6). Depending on the application, the DS1 signal is coded in AMI or in B8ZS.

Frame bit

The F-bit delimits the beginning of the frame and has different meanings. Using ESF, the F-bit sequence has a pattern for synchronization, but if SF is used, then there is a synchronization pattern, CRC control, and a data link control channel of 4 kbit/s.

SDH/SONET

Synchronous digital hierarchy (SDH) and *Synchronous optical network* (SONET) are universal standard that defines a common and reliable architecture for transporting telecommunications services. Functional Architecture

Network Elements

SDH/SONET systems make use of a limited number of *network elements* (NEs) within which all the installations are fitted.

- *Regenerators* (REGs) supervise the received data and restore the signal's physical characteristics, including shape and synchronization.
 - *Add and drop multiplexers* (ADMs) can insert or extract data into or from the traffic.
-

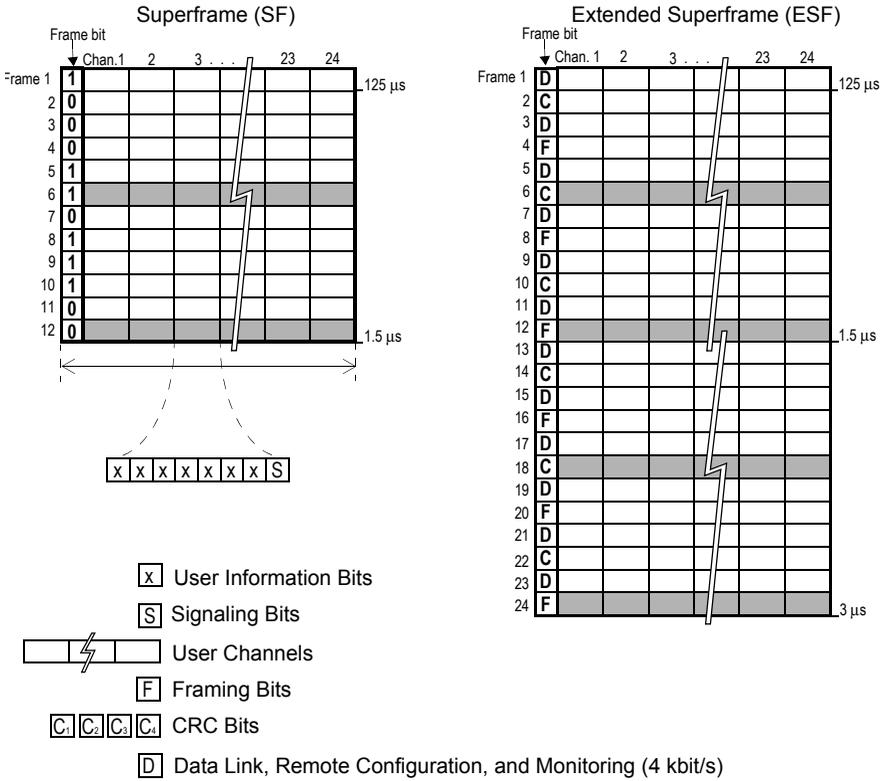


Figure 0.6 The T1 frame and superframe. Depending on the application, the frame bit has different interpretations.

- *Digital cross-connects (DXCs)* configure semipermanent connections to switch traffic between separate networks.

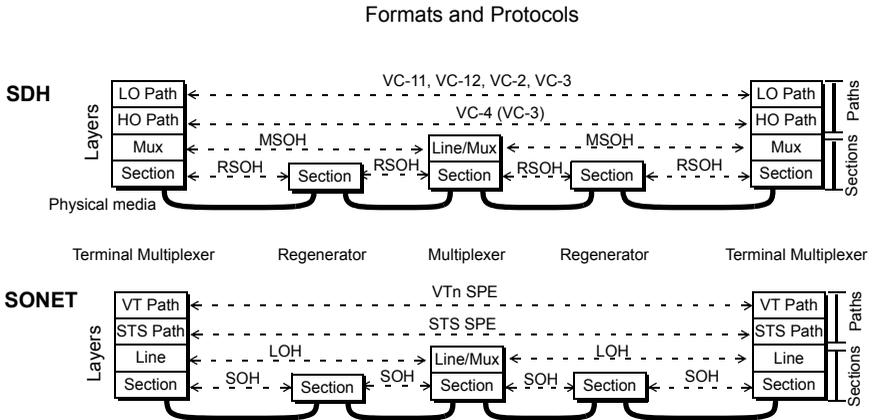


Figure 0.7 SDH and SONET standards define a layered client/server model that can be divided into up to four layers in order to manage transmission services.

Multiplexers provide great flexibility for building Linear and Ring topologies which functionalities have been divided among several layers that manage specific overheads, formats, and protocols.

SDH/SONET Formats and Procedures

SDH defines a set of structures to transport adapted payloads over physical transmission networks (ITU-T Rec. G.707).

- *Mapping*: A procedure by which tributaries are adapted into virtual containers at the boundary of an SDH network.
- *Stuffing*: This is a mapping procedure to adapt the bit rate of client data streams into standardized, fixed-size containers.
- *Multiplexing*: A procedure by which multiple lower-order signals are adapted into a higher-order path signal.

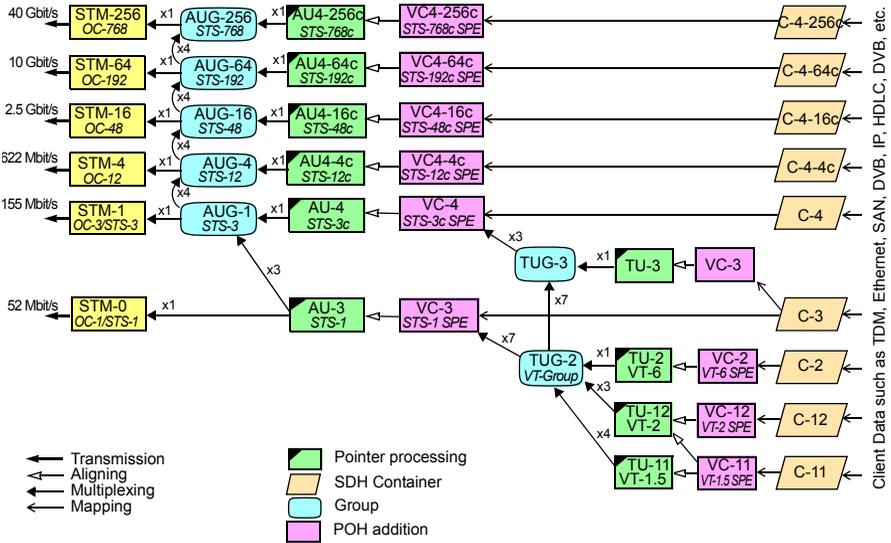


Figure 0.8 SDH and SONET Multiplexing map.

- **Overhead addition:** This procedure is to attach information bytes to a data signal for internal routing and management.
- **Aligning:** A procedure by which a pointer is incorporated into a tributary unit (TU) or an administrative unit (AU).

Multiplexing Map

A multiplexing map is a road map that shows how to transport and multiplex a number of services in STM/OC frames (Figure 0.8).

- The client tributary (PDH, T-carrier, ATM, IP, Ethernet, etc.) needs to be mapped into a C-n container, and a POH added to form a VC-n, or a VT for SONET.
- The VC/VT is aligned with a pointer to match the transport signal rate. Pointers together with VCs form TUs or AUs.

-
- A multiplexing process is the next step, whereby TUG-*n* and AUG-*n* groups are created.
 - When it comes to TUGs, they are multiplexed again to fill up a VC, *synchronous payload envelope* (SPE) in SONET, and a new alignment operation is performed.
 - Finally, an *administrative unit group* (AUG) is placed into the STM/OC transport frame.

Ethernet over SDH

Ethernet has become the standard technology for *local area networks* (LANs). It is cheap, easy to use, well-known, and always in constant evolution toward higher rates. Now it is also being considered as a good technology for access and metro networks, but carriers still need SDH to route high volumes of Ethernet traffic to get long haul. There are several schemes:

- *Ethernet over LAPS*: defined in ITU-T X.86. This is an HDLC family protocol, including performance monitoring, remote fault indication, and flow control.
 - *Generic framing procedure (GFP)*: defined in ITU-T Rec. G.7041. This is a protocol for mapping any type of data link service, including Ethernet, *resilient packet ring* (RPR), and *digital video broadcasting* (DVB).
 - *Virtual concatenation*: defined in ITU-T Rec. G.707, providing quite a lot of flexibility and high compatibility with legacy SDH.
 - *Link capacity adjustment scheme (LCAS)*: defined in ITU-T Rec. G.7042. This dynamically allocates/deallocates new bandwidth to match Ethernet requirements in a flexible and efficient way. It calls for virtual concatenation.
-

Optical Transport Network

At the beginning of the new Milenium data traffic continued growing strongly compared to voice traffic. SONET/SDH, optimized for traditional voice required to be complemented to addapt more efficiently new data centric applications, then new protocols such as Virtual Concatenation (VCAT), Link Capacity Adjustment Scheme (LCAS) and Generic Framing Procedure (GFP) permitted a more flexible transport of Ethernet / IP applications.

G.709 Interface	Line Rate (Gbps)	Corresponding SONET/SDH Rate	Line Rate (Gbps)
OTU1	2.666	OC-48/STM-16	2.488
OTU2	10.709	OC-192/STM-64	9.953
OTU3	43.018	OC-768/STM-256	39.813

Table 1 G.709 line rates and matching SONET/SDH interfaces

The transport core spawned the creation of the Optical Transport Network (ONT) described in the ITU-T rec. G.872 while interfaces are described in the rec. G.709 that includes OAM functions and Ford ward Error Correction (FEC).

Interfaces and Payload

OTN bit rates have been derived from SONET/SDH that have been increased to allocate new overheads and FEC while SONET/SDH and Gigabit Ethernet are the payload. The OTN frame has three parts: overhead, payload and FEC.

OTN defines a hierarchical transport architecture in the sense that lower OTU bit rates are transported into higher interfaces. For instance OTU3 can transport four OTU2 with minor changes on their overheads, while individual FEC are stripped and a new FEC is recalculated.

Several transport combinations are possible:

- SONET/SDH clients directly with a few stuffing bytes

- GFP can be mapped directly into the payload facilitating the transport of Ethernet / IP protocols.
- WDM signals using simple wavelength identification

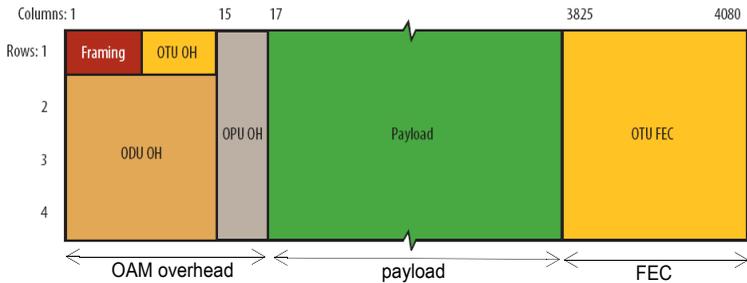


Figure 0.9 G.709 OTN frame is made up of three parts.

Forward Error Correction

FEC permits to know about the quality of the transport producing a redundant code which is part of the OTN frame. At the reception end is possible to correct certain level of transmission errors.

Data is split in 16 streams that include redundancy codes which are capable to correct up to 8 errored bytes per stream. The protocol creates blocks of 255 bytes (1 overhead, 238 info, 16 parity) which finally are interleaved -byte after byte- to create the OTN frame. This strategy reduces the sensitivity to burst of errors enabling the correction of up to 128 consecutive bytes.

Overhead

The architecture of the OTN transport network determines the overheads which are divided in three parts OTU, ODU and OPU.

- *OTU (Optical Transport Unit)* provides supervisory functions and conditions the signal for transport between optical channel ter-

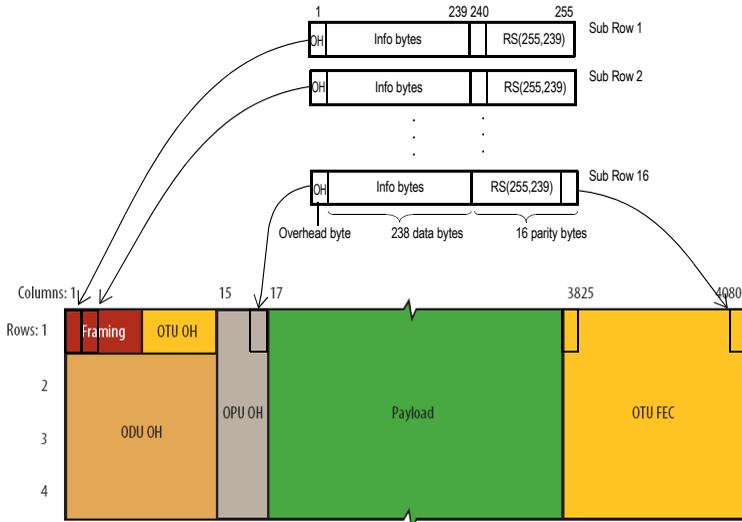


Figure 0.10 OTN frame is created after interleaving 16 data streams. The RS(255, 239) algorithm computes 16 parity bytes per 239 subrow bytes, which combined become the OTU FEC.

minations points where re-timing, reshaping, and regeneration occurs.

- *ODU (Optical Data Unit)* provides end-to-end path supervision and supports tandem connection monitoring
- *OPU (Optical Payload Unit)* provides adapting of the client signals for transport over an optical channel.

Overhead permits the management, transport control, quality monitoring, faults and alarms of the OTN network and the client signals.

Transport can either be synchronous, if OTN and client use the same clock, and asynchronous if do not. In the last case OTN allows movements of the payload.

Hands-on: Performance of TDM networks

In transmission error performance measurements at the physical layer is a major factor in determining the quality of TDM networks.

In-Service and Out-of-Service Measurements

An *out-of-service* (OOS) measurement can be defined as one in which the part of the network on which the measurement is to be performed is disconnected from the rest of the network, and therefore does not provide a service to the user. In contrast, an *in-service* (IS) measurement is one that can be performed without disconnecting the part of the network to be measured from the rest of the network, and the measurement is therefore performed alongside the provision of normal service to the user.

From the above definitions it is clear that for OOS, the traffic of information is simulated (a generator is required), whereas in ISM, the traffic is live (only an analyzer is used). The quantification of the BER forms the basis of OOS, whereas other mechanisms are used for ISM, as will be seen in later sections.

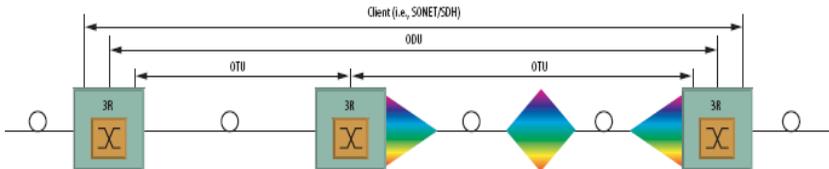


Figure 0.11 OTU, ODU and OPU overhead are interchanged between the termination points of the OTN network.

Bit Error Rate

The BER is the relation between the number of errored bits received and the total number of bits received by an analyzer (Figure 0.12). To obtain this figure, a bit error rate meter is required which must

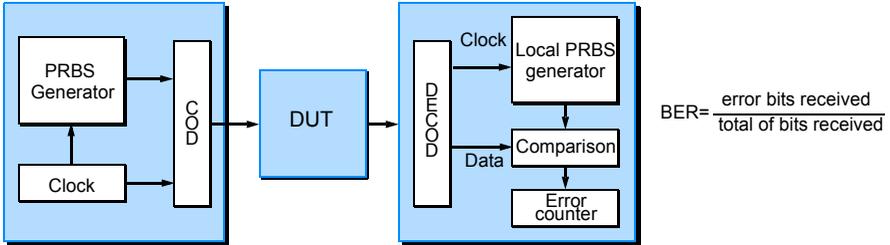


Figure 0.12 Basic diagram for measuring bit error rate (BER).

generate a test signal using its generation section. This signal is a PRBS that is in accordance with the rate of the interface being measured in line with ITU-T Rec. O.150 to O.153 (up to 140 Mbit/s) for test structures in PDH, and O.181 for test structures in SDH. This sequence is sent via the device or system being tested (DUT).

At the receiving end, a bit rate meter receives the signal sent, extracts its clock, and generates the PRBS locally in accordance with this clock. Both sequences (the one received and the one generated locally) are compared bit by bit to determine the number of erroneous bits received. Since the length of the sequence is known, the total number of bits is also known, and the bit error rate can therefore be established.

Out-of-service Measurements

OOS require the live traffic in a link to be replaced by a known test signal, normally a PRBS, and the correct reception of this signal is then checked at the remote end of the communication (see Figure 5.5). This correct reception is quantified by measuring the BER of the test signal used. These tests are intrusive, that is, they interrupt service if they are applied to networks in operation, but they provide exact measurements since the test signal received is checked bit by bit.

In-Service Measurements

When an OOS measurement is made, it causes an interruption in the use of the section or path being measured, thus diminishing, albeit temporarily, the network capacity. This is bad for the operators, who must guarantee not only the quality of the network, but also its availability to the user. The increase in leased lines has accentuated this need. For this reason, preventive maintenance must go hand in hand with the availability of the transmission links. Preventive maintenance is therefore based on ISM; that is, measurements that do not interrupt network traffic.

ISMs are based on checking anomalies in fixed or permitted bit patterns in the live traffic made up of the user data flow (for instance, the FAS) or in the checksum in predefined data blocks. Some of the measurements are applicable to paths, since the parameters are not restarted in an intermediate network interface. Others are only useful at line or section level. These measurements allow for long-term network performance monitoring and preventive maintenance without interrupting user traffic (see Figure 5.10).

ITU-T Error Performance Recommendations

Measurement of error performance form the basis of operation of a number of test set. Error performance is a critical component of transmission quality in digital networks, the ITU-T has published a number of recommendations laying down error performance parameters and objectives. Some of the most relevant include G.821, G.826, G.828, G.829, G.8201, I.356, and the M.21xx series. A set of new recommendations addressing the packet networks have were published, they are Y.1540, Y.1541, Y.1560 and Y.1561.

- **G.821**, originally defined to measure the quality of digital circuits at nx64kbit/s, below 2Mbit/s, and under OOS conditions
-

-
- **G.826**, suitable for constant bit rates equal or higher than 1.5Mbit/s, it measures end-to-end error performance parameters and objectives for international digital paths.
 - **G828**, error performance parameters and objectives for international, constant bit rate synchronous paths.
 - **G.829**, error performance events for SDH multiplex and regeneration sections
 - **G.8201**, error performance parameters and objectives for multi-operator international paths within the OTN
 - **I.356**, performance of ATM cell transfer
 - **M.2100**, performance limits for bringing-into-service and maintenance of international multioperator PDH paths and connections.
 - **M.2101**, performance limits for bringing-into-service and maintenance of international multioperator SDH paths and multiplex sections.
 - **M.2110**, bringing-into-service and maintenance of international multioperator paths, sections and transmission systems.
 - **M.2120**, international multioperator paths, sections, and transmission systems fault detection and localization procedures.

The following family of recommendations are intended for statistical multiplexing networks:

- **Y.1540**, Internet protocol data communications service, IP packet transfer and availability performance parameters.
 - **Y.1541**, network performance objectives for IP-based services.
 - **Y.1560**, parameters for TCP connection performance in the presence of middle box.
 - **Y.1561**, performance and availability parameters for MPLS networks.
-

Timing Methods

Appendix D

Timing Methods

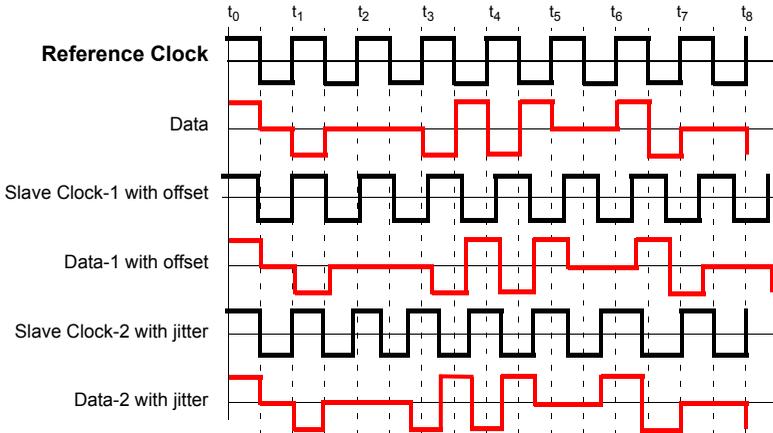


Figure 0.1 A master clock that marks the significant instances for data transmission. Clocks 1 and 2 are badly synchronized, and the data transmitted with these references is also affected by the same phase error.

Synchronization enable the frequency and phase of the network clocks to remain within the specified limits. PDH digital networks are nearly synchronous, plesiochronous indeed, and do not require a common clock. However, synchronization is essential in SDH and SONET networks. A poor *synchronization* causes impairments in systems and services. Some services, like telephony, tolerate a deficient synchronization, but digital TV, VoIP are more sensitive..

Synchronization Architectures

Synchronous networks use *hierarchical synchronization* then a master clock is distributed, making the rest of the clocks slaves:

1. A *master clock*, at the top of the three is a ultra-high quality clock that spread out tthe signal.

2. *slave clocks*, receive the master clock signal and distribute it to all the elements of their node.
3. *Network Elements clocks*, which finish the branches of the tree by taking up the lowest levels of the synchronization chain.
4. *Links*, responsible for transporting the clock signal. .

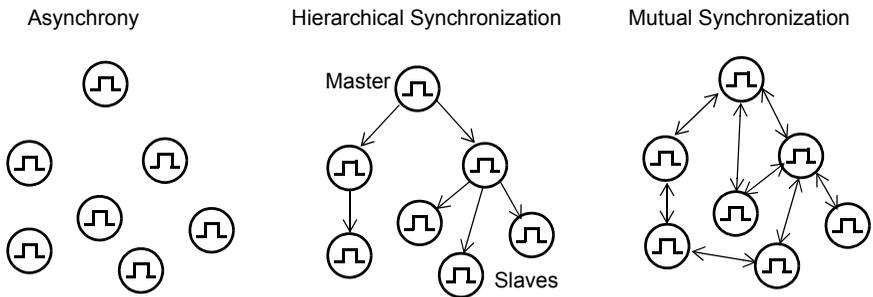


Figure 0.2 Classes of synchronization architectures.

The pure hierarchical synchronization architecture can be modified in several ways to improve network operation. *Mutual synchronization* is based on cooperation between nodes to choose the best possible clock. There can be several master clocks, or even a cooperative synchronization network, besides a synchronization protocol between nodes.

Those networks where different nodes can use a clock of their own, and correct operation of the whole depends on the quality of each individual clock, are called *asynchronous*. Asynchronous operation can only be used if the quality of the node clocks is good enough, or if the transmission rate is reduced. The operation of a network (that may be asynchronous in the sense described above or not) is classified as *plesiochronous* if the equipment clocks are constrained within margins narrow enough to allow simple bit stuffing.

General requirements for today's SONET and SDH networks are that any NE must have at least two reference clocks, of higher or similar quality than the clock itself. All the NEs must be able to generate their own synchronization signal in case they lose their external reference. If such is the case, it is said that the NE is in *holdover*.

A synchronization signal must be filtered and regenerated by all the nodes that receive it, since it degrades when it passes through the transmission path, as we will see later.

Type	Performance
Cesium	From 10^{-11} up to 10^{-13}
Hydrogen	From 10^{-11} up to 10^{-13}
GPS	Usually 10^{-12}
Rubidium	From 10^{-9} up to 10^{-10}
Crystal	From 10^{-5} up to 10^{-9}

Table 0.1
Clock performance.

Synchronization Network Topologies

The synchronization and transport networks are partially mixed, since some NEs both transmit data and distribute clock signals to other NEs.

The most common topologies are:

1. *Tree*: This is a basic topology that relies on a master clock whose reference is distributed to the rest of the slave clocks. It has two weak points: it depends on only one clock, and the signals gradually degrade (Figure 0.5).
2. *Ring*: Basically, this is a tree topology that uses SDH/SONET ring configurations to propagate the synchronization signal. The ring topology offers a way to make a tree secure, but care must be taken to avoid the formation of synchronizing loops.

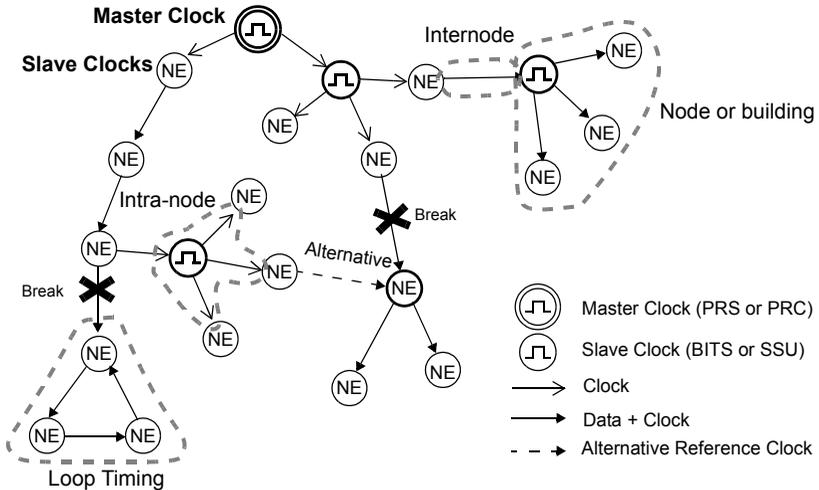


Figure 0.3 Synchronization network topology for SONET and SDH. This figure does not show links that are for transport only.

3. *Distributed*: Nodes make widespread use of many primary clocks. The complete synchronization network is formed by two or more islands; each of them depending on a different primary clock. To be rigorous, such a network is asynchronous, but thanks to the high accuracy of the clocks commonly used as a primary clock, the network operates in a very similar way to a completely synchronous network.
4. *Meshed*: In this topology, nodes form interconnections between each other, in order to have redundancy in case of failure. However, synchronization loops occur easily and should be avoided.

Synchronization networks do not usually have only one topology, but rather a combination of all of them. Duplication and security involving more than one master clock, and the existence of some

kind of synchronization management protocol, are important features of modern networks. The aim is to minimize the problems associated with signal transport, and to avoid depending on only one clock in case of failure. As a result, we get an extremely precise, redundant, and solid synchronization network.

Interconnection of Nodes

There are two basic ways to distribute synchronization across the whole network:

- *Intranode*, which is a high-quality slave clock known as either *synchronization supply unit (SSU)* or *building integrated timing supply (BITS)*. These are responsible for distributing synchronization to NEs situated inside the node.
- *Internode*, where the synchronization signal is sent to another node by a link specifically dedicated to this purpose, or by an STM-*n*/OC-*m* signal.

Synchronization Signals

There are several signals suitable for transporting synchronization:

- Analog, of 1,544 and 2,048 kHz;

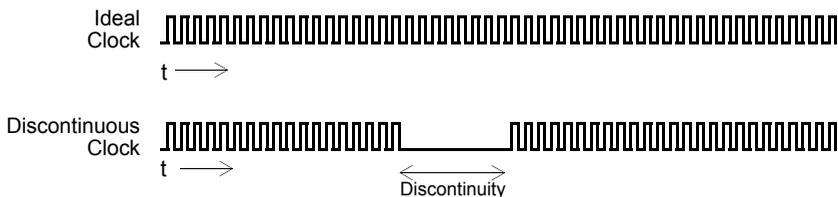


Figure 0.4 A pure clock signal is continuous, as, for example, the one provided by an atomic clock. A discontinuous signal in its turn could be a signal delivered by a T1 circuit transported in SONET.

-
- Digital, of 1,544 and 2,048 kbit/s;
 - STM- n /OC- m line codes, from which one of the above-mentioned signals is derived, by means of a specialized circuit.

In any case, it is extremely important for the clock signal to be continuous. In other words, its mean frequency should never be less than its fundamental frequency (Figure 0.4).

Clock transfer across T-carrier/PDH networks

These types of networks are very suitable for transmitting synchronization signals, as the multiplexing and demultiplexing processes are bit oriented (not byte oriented like SONET and SDH), and justification is performed by removing or adding single bits. As a result, T1 and E1 signals are transmitted almost without being affected by justification jitter, mapping or overhead-originated discontinuities. This characteristic is known as *timing transparency*.

There is only one thing to be careful with, and that is to not let T1 and E1 signals cross any part of SONET or SDH, as they would be affected by phase fluctuation due to mapping processes, excessive overhead, and pointer movements. In short, T1 or E1 would no longer be suitable for synchronization.

Clock transfer across SDH/SONET links

To transport a clock reference across SDH/SONET, a line signal is to be used instead of the tributaries transported, as explained before. The clock derived from an STM- n /OC- m interface is only affected by wander due to temperature and environmental reasons. However, care must be taken with the number of NEs to be chained together, as all the NEs regenerate the STM- n /OC- m signal with their own clock and, even if they were well synchronized, they would still cause small, accumulative phase errors.

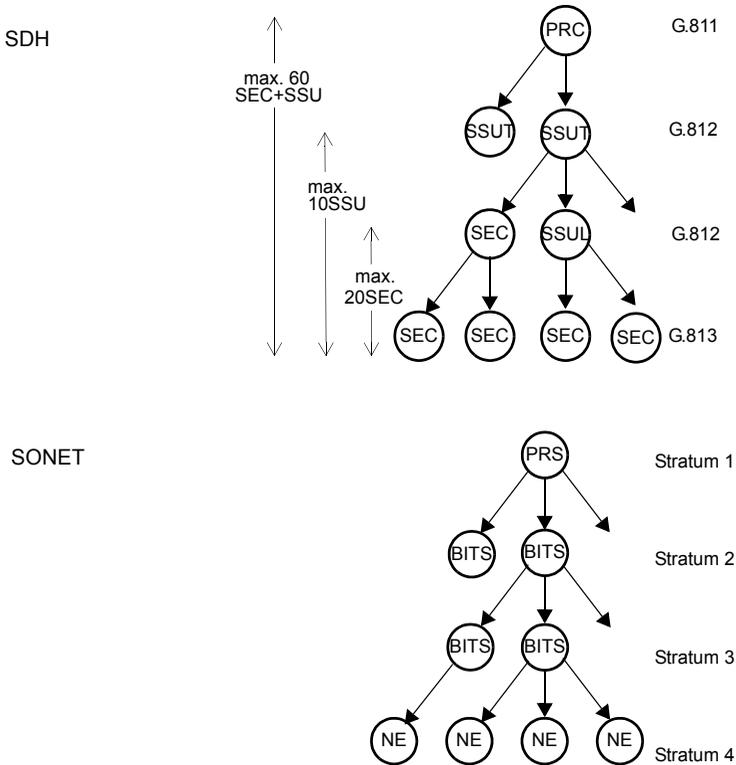


Figure 0.5 Synchronization network model for SONET and SDH. Stratum 3 has the minimum quality required for synchronizing an NE. In SDH the figures indicate the maximum number of clocks that can be chained together by one signal.

Global Positioning System

The *global positioning system* (GPS) is a constellation of 24 satellites that belongs to the U.S. Department of Defense. The GPS receivers

can calculate, with extreme precision, their terrestrial position and the universal time from where they extract the synchronization signal. The GPS meets the performance required from a primary clock. However, the GPS system might get interfered with intentionally, and the U.S. Department of Defense reserves the right to deliberately degrade its performance for tactical reasons.

Disturbances in Synchronization Signals

Since synchronization signals are distributed, degradation in the form of jitter and wander accumulate. At the same time they are affected by different phenomena that cause phase errors, frequency offset, or even the complete loss of the reference clock. Care must be taken to avoid degradation in the form of slips and bit errors by filtering and an adequate synchronization distribution architecture (Figure 0.6).

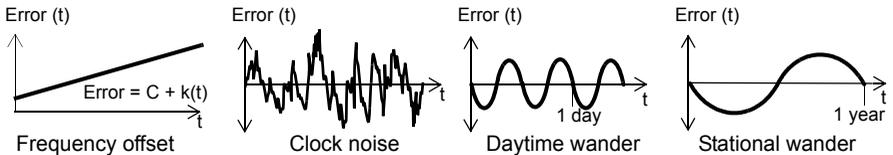


Figure 0.6 Sources of phase variation.

Frequency Offset

Frequency offset is an undesired effect that occurs during the interconnection of networks or services whose clocks are not synchronized. There are several situations where frequency deviations occur (Figure 0.7):

- On the boundary between two synchronized networks with different primary reference clocks;
- When tributaries are inserted into a network by non-synchronized ADMs;
- When, in a synchronization network, a slave clock becomes disconnected from its master clock and enters holdover mode.

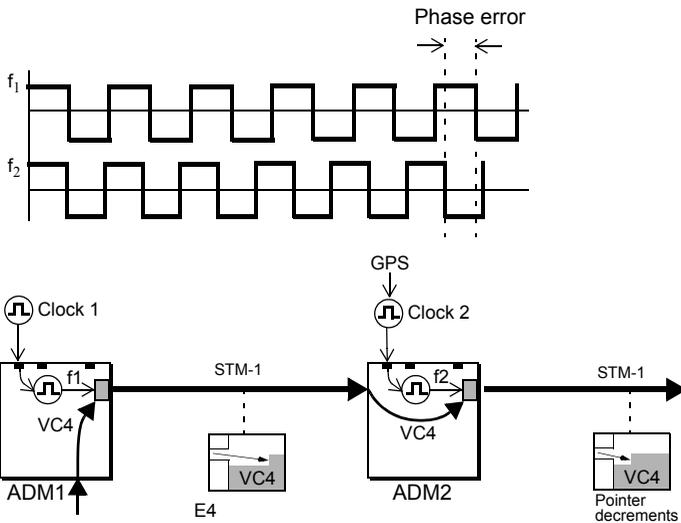


Figure 0.7 Comparison of two reference signals that synchronize two SDH multiplexers. Periodical pointer adjustment occurs due to the frequency offset there is between the two signals.

Phase Fluctuation

In terms of time, the phase of a signal can be defined as the function that provides the position of any significant instant of this signal. It must be noticed that a time reference is necessary for any phase measurement, because only a phase relative to a reference

clock can be defined. A significant instant is defined arbitrarily; it may for instance be a trailing edge or a leading edge, if the clock signal is a square wave (Figure 0.8).

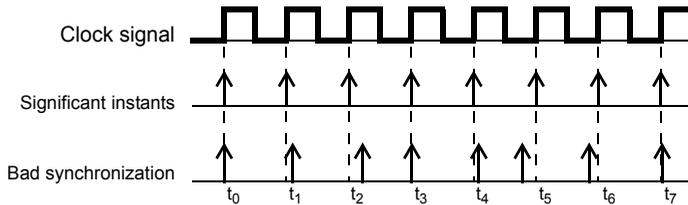


Figure 0.8 Phase error of a signal in relation to its ideal frequency.

Here, when we talk about a phase, we think of it as being related to clock signals. Every digital signal has an associated clock signal to determine, on reception, the instants when to read the value of the bits that this signal is made up of. The clock recovery on reception circuits reads the bit values of a signal correctly when there is no phase fluctuation, or when there is very little. Nevertheless, when the clock recovery circuitry cannot track these fluctuations (absorb them), the sampling instants of the clock obtained from the signal may not coincide with the correct instants, producing bit errors.

When phase fluctuation is fast, this is called jitter. In the case of slow phase fluctuations, known as wander, the previously described effect does not occur.

Phase fluctuation has a number of causes. Some of these are due to imperfections in the physical elements that make up transmission networks, whereas others result from the design of the digital systems in these networks.

Jitter

Jitter is defined as short-term variations of the significant instants of a digital signal from their reference positions in time, ITU-T Rec. G.810 (Figure 0.9). In other words, it is a phase oscillation with a frequency higher than 10 Hz. Jitter causes sampling errors and provokes slips in the *phase-locked loops* (PLL) buffers (Figure 0.10). There are a great many causes, including the following:

Jitter in regenerators

As they travel along line systems, signals go through a radio-electrical, electrical, or optical process to regenerate the signals. But clock recovery in regenerators depends on the bit pattern transported by the signal, and the quality of the recovered clock becomes degraded if transitions in the pattern are distributed heterogeneously, or if the transition rate is too low. This effect can be countered by means of scrambling, which is used to destroy correlation of the user-generated bit sequence. The most commonly used line codes add extra transitions in the pattern, to allow proper clock recovery at the receiving end.

Moreover, this type of jitter is accumulative, which means that it increases together with the increase in the number of repeaters looked at.

Wander

Wander is defined as long-term variations of the significant instants of a digital signal from their reference positions in time (ITU-T Rec. G.810). Strictly speaking, wander is defined as the phase error comprised in the frequency band between 0 and 10 Hz of the spectrum of the phase variation. Wander is difficult to filter when crossing the *phase-locked loops* (PLLs) of the SSUs, since they hardly attenuate phase variations below 0.1 Hz. This is because slow phase variations get compensated with pointer adjustments in SDH/

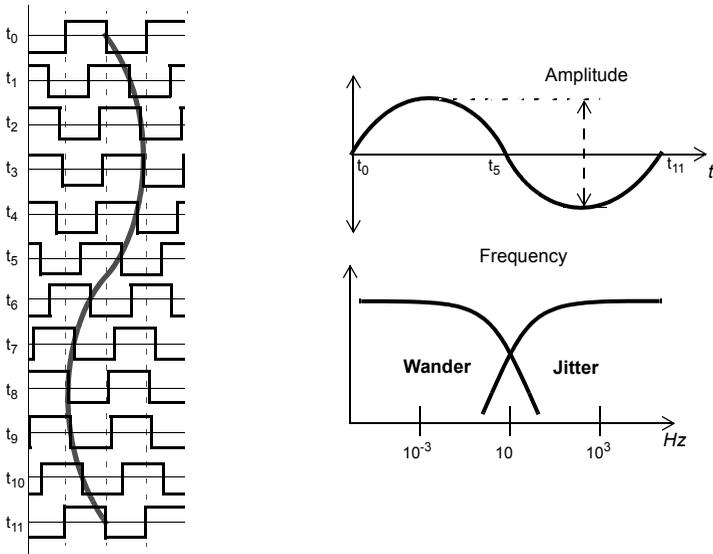


Figure 0.9 A phase fluctuation of a signal is an oscillating movement with an amplitude and a frequency. If this frequency is more than 10 Hz, it is known as jitter, and when it is less than that, it is called wander.

SONET networks, which is one of the main causes of jitter (Figure 0.9).

Wander brings about problems in a very subtle way in a chained sequence of events. First, it causes pointer adjustments, which are then reflected in other parts of the network in the form of jitter. This in its turn ends up provoking slips in the output buffers of the transported tributary.

The following are the most typical causes of wander:

Changes in temperature

Variations between daytime and nighttime temperature, and seasonal temperature changes have three physical effects on transmission media:

- There are variations in the propagation rate of electrical, electromagnetic or optical signals.
- There is variation of length, when the medium used is a cable (electrical or optical), due to changes between daytime and nighttime or winter and summer.
- There is different clock behavior when temperature changes occur.

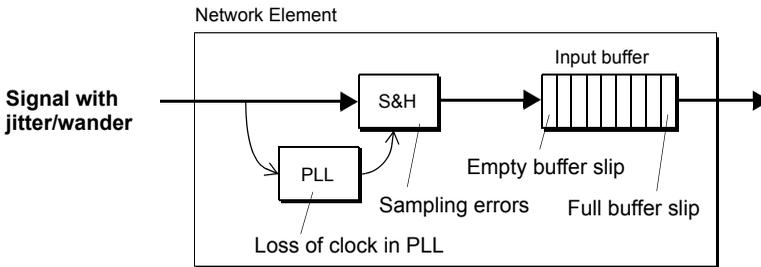


Figure 0.10 Jitter and wander affect every stage of data recovery, producing a number of sampling errors, clock, losses, and overflow.

Clock performance

Clocks are classified according to their average performance in accuracy and offset. The type of resonant oscillator circuit used in

the clock source and the design of its general circuitry both add noise, and this results in wander.

Stratum	Identifier	Accuracy	Drift
1	ST1	1×10^{-10}	2.523/year
2	ST2	1.6×10^{-8}	11.06/day
3	ST3	4.6×10^{-6}	132.48/hour
4	ST4	3.2×10^{-5}	15.36/minute

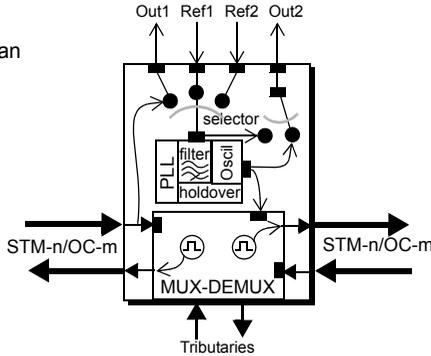
Table 0.2
Stratum timing accuracy.

Synchronization Models

In SDH/SONET networks, there are at least four ways to synchronize the add and drop multiplexers (ADM) and *digital cross connects* (DXC) (Figure 0.11):

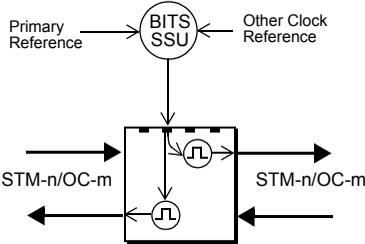
1. *External timing*: The NE obtains its signal from a BITS or *stand-alone synchronization equipment* (SASE). This is a typical way to synchronize, and the NE usually also has an extra reference signal for emergency situations.
2. *Line timing*: The NE obtains its clock by deriving it from one of the STM-*n*/OC-*m* input signals. This is used very much in ADM, when no BITS or SASE clock is available. There is also a special case, known as *loop timing*, where only one STM-*n*/OC-*m* interface is available.
3. *Through timing*: This mode is typical for those ADMs that have two bidirectional STM-*n*/OC-*m* interfaces, where the Tx outputs of one interface are synchronized with the Rx inputs of the opposite interface.
4. *Internal timing*: In this mode, the internal clock of the NE is used to synchronize the STM-*n*/OC-*m* outputs. It may be a temporary

Typical model of an ADM multiplexer

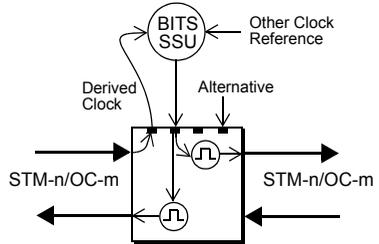


Ref2: Derived Clock
 Out2: Primary Reference
 Out1: Alternative Reference
 Ref1: Clock Output
 PLL: Phase-Locked Loop
 Filter: Low Pass Filter Clock Output
 Oscil: Internal Oscillator

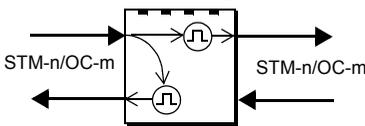
External Timing



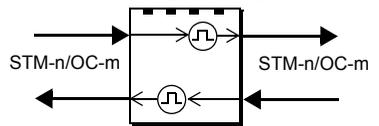
Line-external Timing



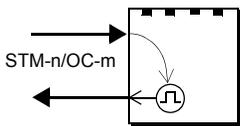
Line Timing



Through Timing



Loop Timing



Internal Timing

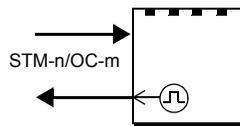


Figure 0.11 Synchronization models of SDH/SONET network elements.

holdover stage after losing the synchronization signal, or it may be a simple line configuration where no other clock is available.

Pointers and Timing Compensation

SDH supports two types of timing mismatches: asynchronous tributaries, and time variations of NE clocks. Justification bits are used to compensate differences with tributaries during the mapping operation. Pointer adjustments are necessary to compensate slight clock differences of the synchronous equipment (basically ADM and DXC).

Pointer Formats and Procedures

Although pointers have different names (AU-4, AU-3, TU-3, TU-2, TU-1, STS ptr or VT ptr), they all share the same format and procedures (Figure 0.12):

- Two bytes allocate the pointer (H1-H2 or V1-V2) that indicates the first byte of the payload (Shown in Table 0.3).
- The pointer value 0 indicates that the payload starts after the last H3 or V3 byte.
- Each pointer has its valid range of values.
- The offset is calculated by multiplying n times the pointer value, and n depends on the payload size.

SDH	Payload	SONET	Payload	Allocation	Range	Hops	Justification
AU-4	VC-4	STS-3 ptr	STS-3c	H1, H2	0 - 782	3 bytes	3 bytes
AU-3	VC-3	STS-1 ptr	STS-1	H1, H2	0 - 782	3 bytes	1 bytes
TU-3	VC-3	—	—	H1, H2	0 - 764	1 byte	1 byte
TU-2	VC-2	VT-6 ptr	VT-6	V1, V2	0 - 427	1 byte	1 byte
TU-12	VC-12	VT-2 ptr	VT-2	V1, V2	0 - 139	1 byte	1 byte
TU-11	VC-2	VT-15 ptr	VT-15	V1, V2	0 - 103	1 byte	1 byte

Table 0.3
SDH and SONET pointers.

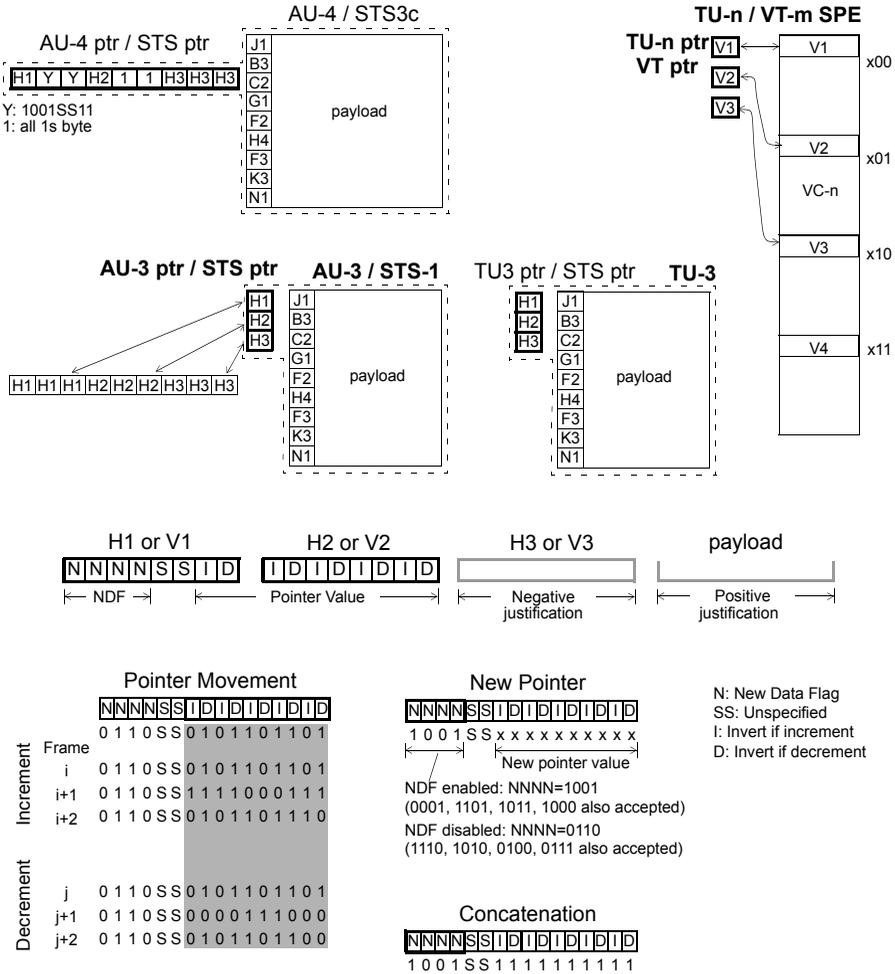


Figure 0.12 Pointer formats, codification, and procedures.

Pointer Generation

In normal operation, pointers are located at fixed positions, and the *new data flag* (NDF) is 0110. However, sometimes it is necessary to change the pointer value, in which case the following rules apply:

- *Minimum time period*: The minimum time period between two consecutive pointer changes is 500 μ s.
 - *Pointer increment*: If a positive justification is required, the pointer value is sent with the I-bits inverted. The new pointer value is the previous value, incremented by one. If the pointer is H1-H2, the position of the payload is shifted three bytes forward, and void bytes are left after H3. If it is V1-V2, the payload is shifted one byte forward, and a void byte is left after V3.
 - *Pointer decrement*: If a negative justification is required, the pointer value is sent with the D-bits inverted. In this case, the new pointer value is the previous value decremented by one. If the pointer is H1-H2, the position of the payload is shifted three bytes backwards, and H3 provides spare bytes. If the pointer is V1-V2, the payload is shifted one byte backwards, and either V3 provides the spare byte.
 - *New pointer*: If the VC-*n* alignment changes for any reason, and it cannot be tracked by pointer increments or decrements, then a new pointer value is sent, and the NDF is set to 1001 to reflect the new value.
-

