Gigabit Ethernet Roll-Out

Understanding How it Works for an Efficient Service

ALBEDO

José M. Caballero
Francisco J. Hens

Gigabit Ethernet roll-out

José M. Caballero
Francisco J. Hens

ALBEDO Telecom
www.albedotelecom.com

ISBN 84-609-5042-5

Juan de Austria, 112
Barcelona - 08018 - Spain

Printed by Printulibro Intergrup, S.A.

Legal Deposit B-20.437-2005

P/N: 400665

Reprinted with corrections: *21 November 2023*

Preface

For about 200,000 years, Neanderthal man inhabited and ruled the chilly forests and steppes of Eurasia. The Neanderthal society was gregarious and hierarchical, and it was formed by scattered tribes that brought up their children and took care of the wounded, the sick, and the elderly. The world view of these early human beings already showed signs of capacity for abstract and synthetic thinking, as they practiced rites and decorated their bodies with necklaces, paintings, and earrings. Thanks to more than 2 million years of human evolution, they had their tools and techniques to make fire and procure and store food. They were also able to tan leather to make clothes and to protect their feet. But their weapons were what converted them into the most extraordinary predators in the food chain.

However, some 40,000 years ago, a new hominid species of African origin started to compete for space with *Homo Neanderthalensis*. The newcomers were slightly different physically; their skin was darker and they were taller, although less muscular. This made them physically less prepared for the cold climate. They did not seem to be more intelligent either; at least if we look at the size of their skull, which was about 10% smaller than that of Neanderthal man. And if this is not enough, the children of this new species took twice as long to grow up; in this way forcing their parents to have fewer descendants. Evolution had made their reproductive period shorter, so that they could feed and take care of their descendants. In spite of all this, after a relatively short period of coexistence, *Homo Neanderthalensis* mysteriously disappeared. Perhaps they were just simply wiped away by their competitors, or killed by the new viruses coming from the south, or maybe they just disappeared because they were unable to adapt themselves to the rapid changes between the Ice Ages. We do not know, but the newcomers known as Cromagnon man became dominants.

So, what was the key difference between Cromagnon and Neanderthal men?

Yes, *communication*. The unusual form of the larynx and the gullet of the new species, also known as *Homo Sapiens,* enabled them to generate and modulate sophisticated sounds. Neanderthal men did not have this capacity, without which it is impossible to create a human language. This theory explains how the hominids moved from waiting for genetic changes to using communication as the vital survival tool. This proved to be more useful than the slow biological evolution in adapting the hominids to their environment. However, acquiring language skills and preparing the brain for learning is a slow process that in this case made new generations mature more slowly, and parents had to spend more years taking care of their immature descendants. Despite this, and other physiological difficulties, the new human beings who, as you probably already know, are nothing less than us, took over in a relatively short period of time, and ended up populating most of the planet.

The second significant milestone in the history of communications was the discovery of writing, probably the most important intellectual tool ever discovered by man. Writing enables us to store information and transmit it between two distant points and even between generations, without distorting or losing the message. There is evidence of earlier attempts, although the first effective form of writing was developed by the Sumerians about 5,000 years ago. The Sumerians lived in city-states on the banks of the Tigris and Euphrates rivers, where such activities as agriculture, cattle raising, craft work, metallurgy, and construction flourished in an extraordinary way. Writing was born in the heart of these urban societies as a means to increase commerce, and solve both legal and social problems. Originally, the Sumerian codes were iconographic, whereby each sign was an icon resembling the object it represented. This way, it was possible to sell or buy a herd of 53 lambs, for instance, or legally divide a property of 180 *ikus* of surface between heirs. When numbers were later developed, this was a huge step forward, as it was no longer necessary to repeat the same icon a number of times. But what really made a change was the invention of the phonetic writing system. Now it was possible to describe battles or the position of stars, or write down laws, such as Hammurabi's Code of Law.

This was the start of our civilization.

For thousands of years, writing was done by hand on clay, stones, papyrus, or leather, until in 1450 the workshop of Gutenberg started to mechanically produce what became the first books. Fifty years later, the few books kept in monasteries and palaces were transformed into more than 10 million volumes. It was finally possible to store and produce a large amount of information at lower cost than before, and without changing the original contents. Knowledge, literature, and science were no longer tools of power for a small elite of scribes, priests, and courtiers. This way, by having the medium to broadcast information to thousands of recipients, the printing industry had an important role in marking the end of the dark Middle Ages.

Some centuries had to pass before electricity was managed in such a way that the first telephone patented by Alexander Graham Bell in 1876 could be developed. A few years later, around 1900, the first radio transmissions took place, and television appeared in the 1930s. Without underestimating television, there is something that makes the telephone special, in that it enables direct interpersonal communication at a distance. We can even see the telephone as an extension of our larynx and ears, while radio and television are one-way media, where the receiver can only connect and disconnect, the same way as you can close this book, but not modify its contents. This difference is notable, and it explains why both radio and TV tend to be desired, controlled and even manipulated by political, economical or religious power, while private telephone conversations offer more liberty and independence. The telephone is by definition a tool where the contents, the language, and the recipient can be decided by the users themselves.

Finally, we arrive at the mid-1990s, when the Internet became an important medium for mass communication. It combines two fundamental inventions: writing and telecommunications. Writing can be very precise and it enables us to store information, crossing the time barrier; while telecommunications overcomes space barriers. It is so efficient that many times we prefer to send electronic messages even within the same office, although it would be easier just to have a short conversation. But the Internet is a lot more than an efficient two-way communication medium. It is also a way to access the immense "universal library," with an impact that can only be compared to the Alexandria Library 2,000 years ago. In the Internet, we have millions of documents with information that can be reached from any part of the world in just a few seconds.

The third generation wireless networks will also bring some changes in the near future, by improving the human-machine relationship. Our mobile telephones will become terminals with Internet access or radio and video broadcast, accessible from anywhere in the world with reasonable costs. During the next years, our written works, both in the office and at home, will depend less and less on paper. There will be a need for new devices to substitute for paper. These devices should be connectable, autonomous, light, easy to handle, and shock-proof. With a high resolution we could read our newspaper in the train or read a book in the garden as comfortably as before, but saving the cost of cutting tons of wood and using chemical substances to make paper.

In other words, human evolution and globalization is a communication matter as old as mankind. Therefore human interaction and multiculturalism are accelerated whenever a new communication milestone is reached, such as the larynx, the art of writing, the printing press, the telephone, radio, television, or the Internet.

Pepe Caballero

Maidenhead, England

Table of Contents

Chapter 1

Ethernet and Gigabit 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 Mbit/s; Fast Ethernet at 100 Mbit/s, Gigabit Ethernet at 1000 Mbit/s and 10-Gigabit Ethernet at 10 Gbit/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 provides.

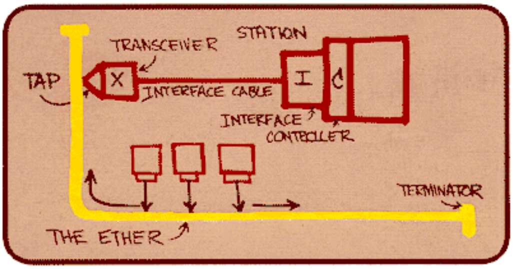


Figure 1.1 A drawing of the first Ethernet system by Bob Metcalfe

Gigabit Ethernet (GbE) is known to be a good and cost-effective technology for enterprises seeking to roll out *Metropolitan* and *Campus Area Networks* (MAN/CAN). GbE perfectly adapts to enterprise data applications, often based on

Ethernet as well. *10 Gigabit Ethernet* (10GbE) was designed keeping MAN and WAN applications in mind, but it can also be used for some bandwidth-consuming LAN applications. GbE and 10 GbE open up opportunities for new services such as MPLS, and for applications like triple play.

The new Gigabit topologies rely on switches that connect stations using dedicated, and generally also full-duplex optical links. The most important exception is 1000BASE-T, which was designed to provide a migration path for existing *Unshielded Twisted Pair* (UTP) cable Ethernet installations.

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 slightly 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. The future of wireless Ethernet is promising, with rates up to 54 Mbit/s today, and convergence with 3G mobile networks expected in the near future.

1.1 A Brief History of Ethernet

There is a network that has always been considered as the predecessor of Ethernet: *ALOHAnet*, 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 (see Figure 1.2).

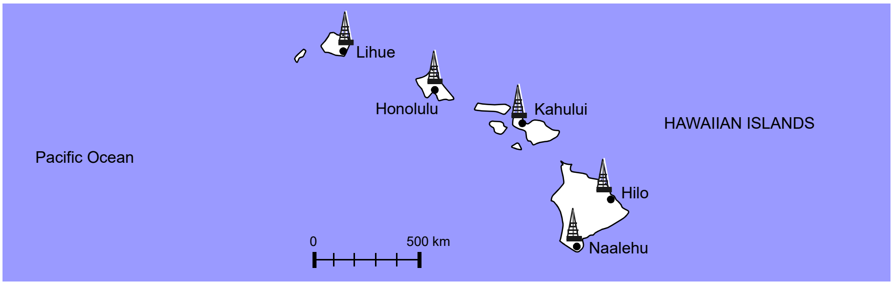


Figure 1.2 ALOHA, a pre-Ethernet network, was developed in the 1960s to transmit data between the Hawaiian islands.

The first real Ethernet was designed in 1973 by Bob Metcalfe in Xerox Corporation's Palo Alto laboratory (see Figure 1.1). This first version was able to operate at 3 Mbit/s over a shared coaxial cable, using CSMA/CD. This was a simple algorithm (see Paragraph 1.4) that improved efficiency by up to 80%, depending on the network configuration and traffic load.

In 1980 a consortium formed by Digital, Intel and Xerox (known as the DIX cartel) developed the 10 Mbit/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.

1.2 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 10 Gbit/s over coaxial cable, copper pairs and optical fiber. 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.

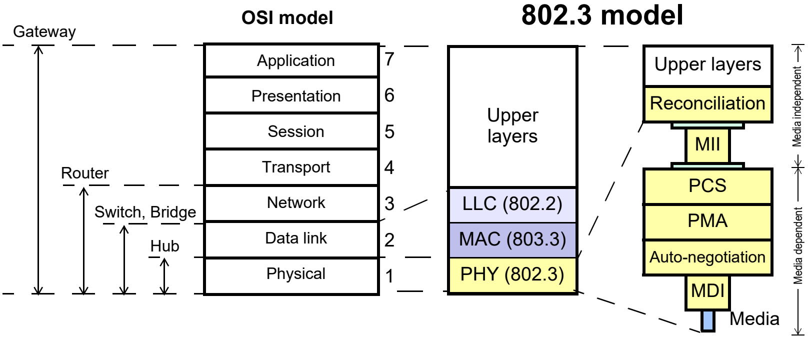


Figure 1.3 Ethernet layers vs. OSI model. Some layers are optional, depending on the version.

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-media environment. It generates source and destination addresses to identify the two 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 IEEE 802.3 standard does not specify the Ethernet MAC layer. Other IEEE-based technologies require special MAC layers that are specified in other standards. Some examples are: Token Bus (IEEE 802.4), Token Ring (IEEE 802.5) and *Resilient Packet Ring* (RPR, IEEE 802.17).
- 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.

1.2.1 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.

1.2.1.1 Attachment-Unit Interface

When 10BASE-5, the first commercial Ethernet solution was manufactured, it could only be operated over thick coaxial cable (see Table 1.1). The evolution towards multiple physical media started with the introduction of the *Attachment Unit Interface* (AUI), developed for rates of 10 Mbit/s. The intention was to avoid the difficulty of routing thick and inflexible coaxial cable to each station (see Figure 1.4). 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.

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.

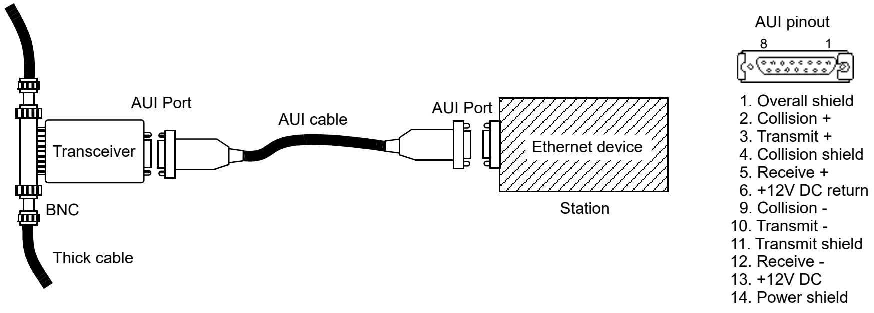


Figure 1.4 The AUI is a little bit more than a connection cable between the Ethernet card and the transceiver.

1.2.1.2 Medium-Independent Interfaces

The *Medium-Independent Interface* (MII) is the equivalent of the AUI for Fast Ethernet (100 Mbit/s). It connects the block by implementing the 100 Mbit/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 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.

The are extensions to the MII interface for 1GbE and 10GbE. The former is known as Gigabit MII (GMII), and the latter as eXtended GMII (XGMII).

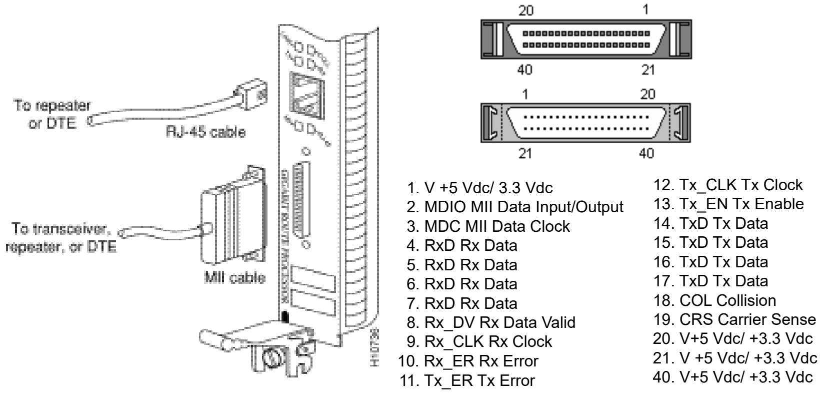


Figure 1.5 MII used for Fast Ethernet at 100 Mbit/s. The pinout provides four groups of signals: power signals, management signals such as clock, transmit/receive signals at one fourth of the data rate, and control signals such as CS and CD.

1.3 The Ethernet PHY

Ethernet has adopted many different transmission media, including coaxial cables, UTP, STP and multimode/monomode fiber, in order to meet the changing market needs (see Table 1.1).

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.

Standard	*Name*	*Media Type*	*H/F*	*(En) coding*	*Line*	*MFS bytes*	*Network Size*
Ethernet IEEE 802.3a-t (clauses1-20) AUI	10BASE-2	One 50 Ω thin coaxial cable	H	4B/5B	Manchester	64	<185 m
	10BASE-5	One 50 Ω thick coaxial cable	H	4B/5B	Manchester	64	<500 m
	10BROAD-36	One 75 Ω coaxial (CATV)	H	4B/5B	Manchester	64	<3600 m
	10BASE-T	Two pairs of UTP 3 (or better)	H/F	4B/5B	Manchester	64	<100 m
	10BASE-FP	Two optical 62.5 µm MMF passive hub	H/F	4B/5B	Manchester	64	<1000 m
	10BASE-FL	Two optical 62.5 µm MMF asyn hub	H/F	4B/5B	Manchester	64	2000 m
	10BASE-FB	Two optical 62.5 µm MMF sync hub	H/F	4B/5B	Manchester	64	<2000 m
Fast Ethernet IEEE 802.3u (clauses 21-29) MII	100BASE-T4	Four pairs of UTP 3 (or better)	H/F	8B/6T	MLT3	64	<100 m
	100BASE-T2	Two pairs of UTP 3 (or better)	H/F	PAM5x5	PAM5	64	<100 m
	100BASE-TX	Two pairs of UTP 5 (or better)	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 µm MMF	H/F	4B/5B	NRZI	64	2 km
	100BASE-FX	Two optical 50 µm SMF	H/F	4B/5B	NRZI	64	40 km
Gigabit Ethernet IEEE 802.3z/ab (clauses 34-42) GMII	1000BASE-CX	Two pairs 150 Ω STP (twinax)	H/F	8B/10B	NRZ	416	25 m
	1000BASE-T	Four pair UTP 5 (or better)	H/F	8B1Q4	4D-PAM5	520	<100 m
	1000BASE-SX	Two 50 µm MMF, 850 nm	H/F	8B/10B	NRZ	416	500/750 m
	1000BASE-SX	Two 62.5 µm MMF, 850 nm	H/F	8B/10B	NRZ	416	220/400 m
	1000BASE-LX	Two 50 µm MMF, 1310 nm	H/F	8B/10B	NRZ	416	550/2000 m
	1000BASE-LX	Two 62.5 µm MMF, 1310 nm	H/F	8B/10B	NRZ	416	550/1000 m
	1000BASE-LX	Two 8 ~ 10 µm SMF,1310 nm	H/F	8B/10B	NRZ	416	5 km
	1000BASE-ZX	Two 8 ~ 10 µm SMF, 1550 nm	H/F	8B/10B	NRZ	416	80 km
10GEthernet IEEE 802.3ae (clause 48-53) XGMII	10GBASE-SR	Two 50 µm MMF, 850 nm	F	64B/66B	NRZ	N/A	2 ~ 300 m
	10GBASE-SW	Two 62.5 µm MMF, 850 nm	F	64B/66B	NRZ	N/A	2 ~ 33 m
	10GBASE-LX4	Two 50 µm MMF, 4 x DWM signal	F	8B/10B	NRZ	N/A	300 m
	10GBASE-LX4	Two 62.5 µm MMF, 4 x DWM signal	F	8B/10B	NRZ	N/A	300 m
	10GBASE-LX4	Two 8 ~ 10 µm SMF, 1310 nm, 4 x DWM signal	F	8B/10B	NRZ	N/A	10 km
	10GBASE-LR	Two 8 ~ 10 µm SMF, 1310 nm	F	64B/66B	NRZ	N/A	10 km
	10GBASE-LW	Two 8 ~ 10 µm SMF, 1310 nm	F	64B/66B	NRZ	N/A	10 km
	10GBASE-ER	Two 8 ~ 10 µm SMF, 1550 nm	F	64B/66B	NRZ	N/A	2 ~ 40 km
	10GBASE-EW	Two 8 ~ 10 µm SMF, 1550 nm	F	64B/66B	NRZ	N/A	2 ~ 40 km

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 channels.

1.3.1 Ethernet at 100 Mbit/s and Less

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.

1.3.1.1 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 Mbit/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 W 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 1.4.1). 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 1.6).

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.

1.3.1.2 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.

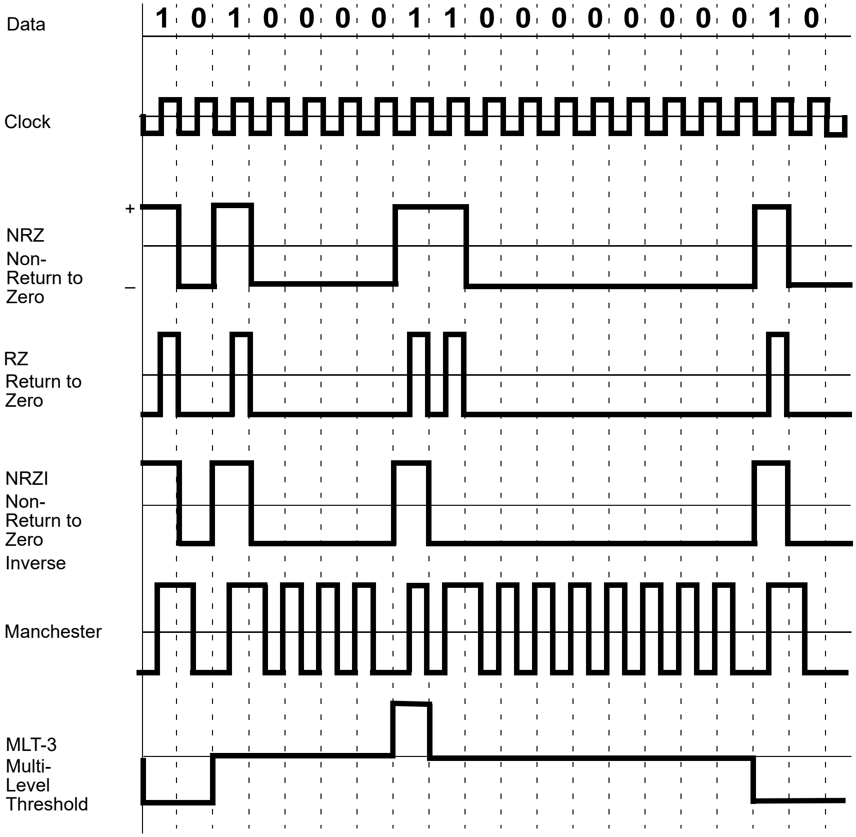


Figure 1.6 Line encoding technologies. Depending on the media, a "+" signal corresponds to high voltage on copper or high intensity on optical fiber, and a "–" signal to low voltage or low intensity. PAM5 uses 5 levels (-2, -1, 0, 1, 2), several pairs (two in 100BASE-T and four in 1000BASE-T), and a complex encoding rule to generate the symbols transmitted in parallel over each pair.

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 almost 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-Mbit/s extension to 10BASE-T. It keeps the same MAC layer as the 10Mbit/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 (see Paragraph 1.3.1.3). The electrical Fast Ethernet interfaces are the following:

- 100BASE-TX – requires two pairs of Category-5 UTP or Type-1 shielded pair cables. This means that 100BASE-TX may not operate if 10BASE-T cabling is used.
- 100BASE-T4 – needs four pairs of Category-3 UTP cables. This interface can be used when Category-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 Category-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 Mbit/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 1.6).

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 Mbit/s links can be used for connections to workstations, while 100 Mbit/s offers inexpensive bandwidth for connections to servers.

1.3.1.3 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 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 are selling 100BASE-FX over *Single-Mode Fiber* (SMF). In this case, the PHY is based on the FDDI physical layer.

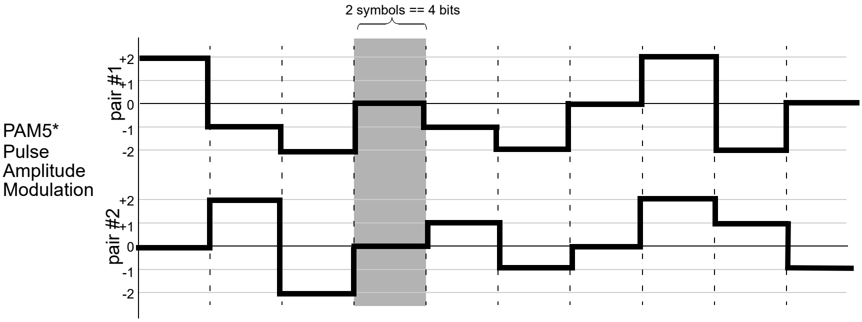


Figure 1.7 Line encoding technologies. Depending on the media, a "+" signal corresponds to high voltage on copper or high intensity on optical fiber, and a "–" signal to low voltage or low intensity. PAM5 uses 5 levels (-2, -1, 0, 1, 2), several pairs (two in 100BASE-T and four in 1000BASE-T), and a complex encoding rule to generate the symbols transmitted in parallel over each pair.

1.3.2 Gigabit 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

Gigabit 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.

1.3.2.1 1000BASE-X Architecture

In 1998, the IEEE approved a standard for Gigabit 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 1.8). The first one uses an STP cable, and the second and the third one use optical fiber.

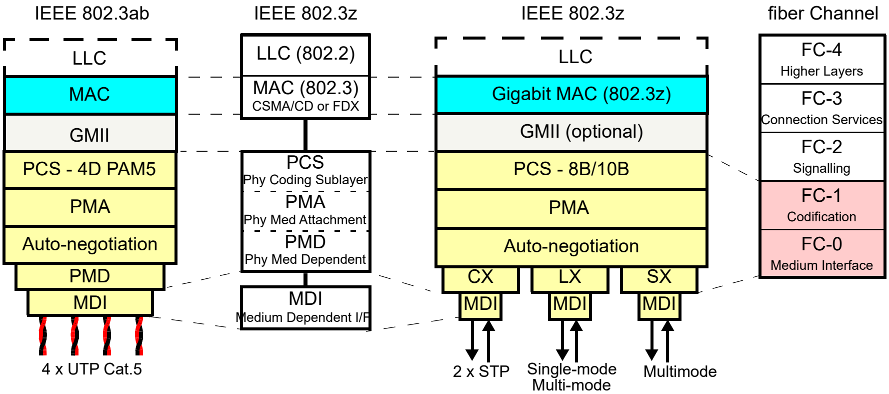


Figure 1.8 Gigabit Ethernet defines several transmission media, specified in the IEEE 802.3z (1000BASE-X) and 802.3ab (1000BASE-T). The first one is based on the existing fiber channel technology and covers three different types of media, and the second one uses the popular UTP cable.

- 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 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 Mbit/s, and the data rate is 1000 Mbit/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.

1.3.2.2 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.

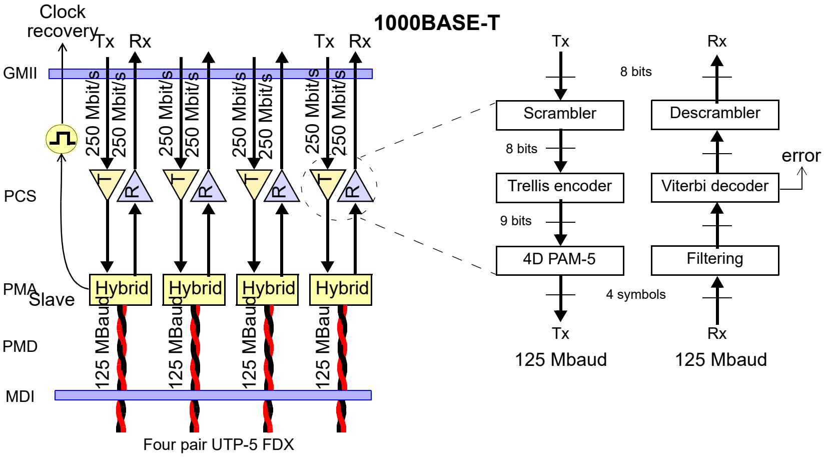


Figure 1.9 1000BASE-T transmits and receives signals simultaneously over the same pairs.

1000BASE-T operates over Cat. 5 (or better) cabling systems by using all four pairs, sending and receiving a 250 Mbit/s data stream over each of the four pairs (4 x 250 Mbit/s = 1 Gbit/s) simultaneously (see Figure 1.9). 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 Mbit/s throughput using baseband signaling at 125 MBaud. It achieves a half-duplex data rate of 1 Gbit/s at a spectral power density similar to that of 100BASE-TX (see Figure 1.10).

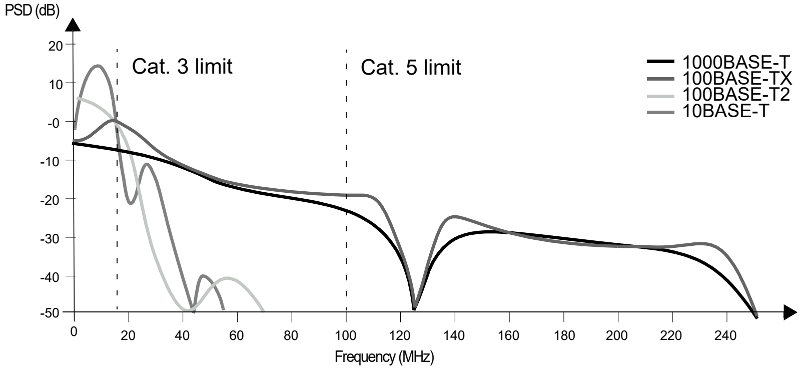
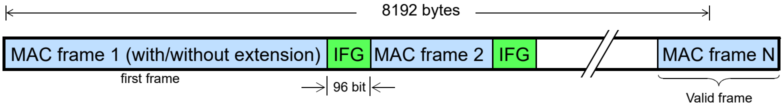


Figure 1.10 *Power Spectral Density* (PSD) for 10/100/1000BASE-T electrical technologies.

1.3.2.3 Frame Bursting in Gigabit Ethernet

Transmission in a shared media can be very inefficient, especially when sending small frames with padding. To help this, the frame bursting feature of Gigabit Ethernet enables a device to send over 8000 bytes in a burst (see Figure 1.11).



IFG: Interframe Gap
Max. burst length: 8192 bytes, last frame still valid
Inter-frame gap: 96 bits

Figure 1.11 Frame bursting. In Gigabit Ethernet a station can send multiple frames to make the HDX mode more efficient. Only the first frame requests the extension.

The first frame is sent in a normal manner, then, without dropping the carrier, the second one is sent and so on, up to the limit allowed. Each frame is separated by a gap without data, keeping the carrier sense active, to prevent other stations from starting a transmission.

1.3.3 10 Gigabit Ethernet

In June 2002, the IEEE 802.3ae standard for transmission of Ethernet at 10 Gbit/s was approved. 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 interesting points of 10 Gigabit Ethernet technologies.

In addition to compatibility with WAN, there are two new important points in 10GbE.

- Half duplex has been abandoned and only full-duplex operation is permitted. This makes CSMA/CD unnecessary (see Chapter 2).
- Copper transmission has also been discarded, and the IEEE 802.3ae only defines optical interfaces.

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.

1.3.3.1 Architecture

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 Paragraph 1.3.3.2).

Table 1.2
Range of 10GbE interfaces

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 MMF	~ 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

The 10GbE standard defines a long-haul interface (10GBASE-E) that can be used for packet-based MAN/WAN applications. 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 10GBASE-LX4 interface is used to multiplex four wavelengths near the second optical transmission window (1310 nm). Each wavelength transports a 8B/10B, NRZ coded signal operating at 3.125 GBaud

Table 1.3
Wavelength specifications for the 10GBASE-LX4 interface

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

The maximum range of the 10BASE-LX4 is 300 m over 160 MHz/km, 62.5 μm MMF. This is not enough for MAN or WAN applications, but it can be very useful to upgrade installed MM fiber to 10 Gbit/s.

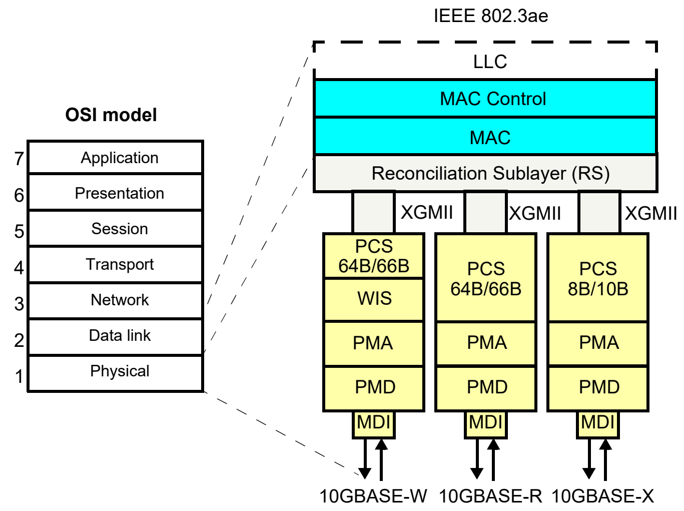


Figure 1.12 Layered model of IEEE 802.3ae 10 Gigabit Ethernet

The 64B/66B coding is used by the PCS sublayer of the 10GBASE-R and 10GBASE-W interfaces. This code has similar functions as the 8B/10B code used in the 1 Gigabit Ethernet and the 10GBASE-LX4 interface. It makes frame delineation, clock recovery and single or multiple error detection possible at the physical layer (see Figure 1.12).

The use of 64B/66B coding means that the signaling rate of 10GBASE-R is not exactly 10 GBaud. The exact symbol rate of 10GBASE-R is 10.3125 GBaud.

1.3.3.2 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 1.13).

The available bandwidth for mapping the 64B/66B-coded Ethernet signal corresponds to the capacity of the VC4-64c / STS-192c container, 9.58464 Gbit/s. This number is different from the bit rate generated by the 10GBASE-R interface (10.3125 Gbit/s). However, direct switching between interfaces with WIS and without WIS is theoretically possible, the same way as Ethernet at 1 Gbit/s can be switched to 100 Mbit/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.

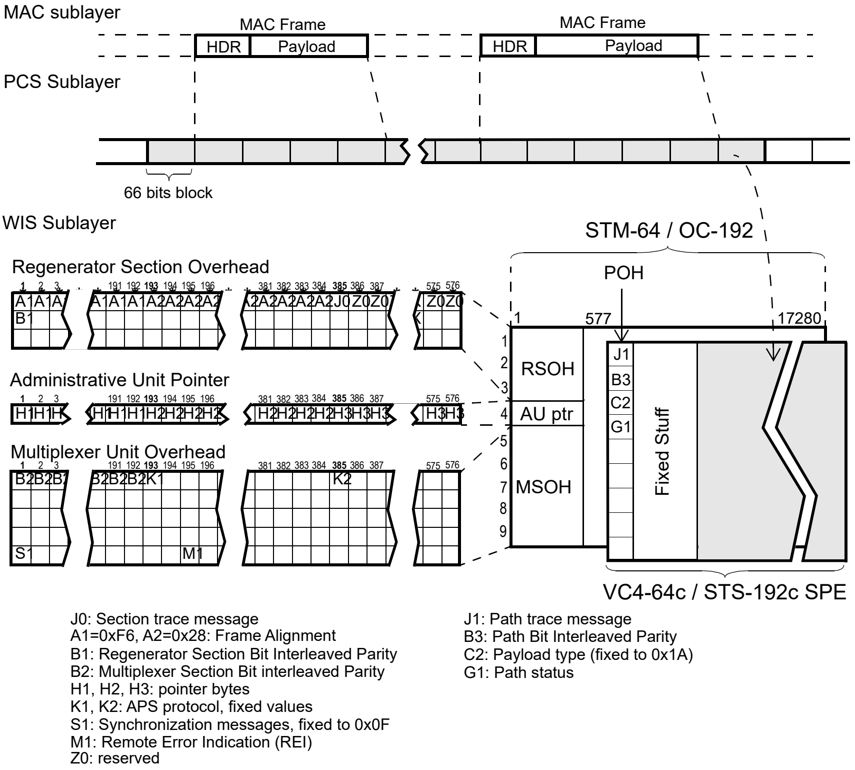

MAC sublayer
MAC Frame
HDR
Payload
MAC Frame
HDR
Payload
PCS Sublayer
66 bits block
WIS Sublayer
STM-64 / OC-192
Regenerator Section Overhead
A1 A1 A1 A1 A1 A2 A2 A2 A2 A2 A2 A2 A2 J0 Z0 Z0 Z0 Z0
B1
Administrative Unit Pointer
H1 H1 H1 H1 H1 H1 H2 H2 H2 H2 H2 H2 H2 H3 H3 H3 H3 H3
Multiplexer Unit Overhead
B2 B2 B2 B2 B2 B2 K1
K2
S1
M1
1 2 3 191 192 193 194 195 196 381 382 383 384 385 386 387 575 576
POH
1
577
17280
1 2 3 4 5 6 7 8 9
RSOH
AU ptr
MSOH
J1
B3
C2
G1
Fixed Stuff
VC4-64c / STS-192c SPE
J0: Section trace message
A1=0xF6, A2=0x28: Frame Alignment
B1: Regenerator Section Bit Interleaved Parity
B2: Multiplexer Section Bit interleaved Parity
H1, H2, H3: pointer bytes
K1, K2: APS protocol, fixed values
S1: Synchronization messages, fixed to 0x0F
M1: Remote Error Indication (REI)
Z0: reserved
J1: Path trace message
B3: Path Bit Interleaved Parity
C2: Payload type (fixed to 0x1A)
G1: Path status

Figure 1.13 Mapping of 64B/66B data into a simplified STM-64 / OC-192 frame performed by the WIS in the 10GBASE-W interfaces

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 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.

1.3.4 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 is specified as an option for 10BASE-T, 100BASE-TX, and 100BASE-T4, but it is a *requirement* for 100BASE-T2 and 1000BASE-T implementations.

1.3.4.1 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 1.4).

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 stations 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 1.14).

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.

Table 1.4
Priority resolution

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

Auto-negotiation was first defined for 10/100 Mbit/s interfaces operating over twisted pair, and was later extended to 1000 Mbit/s. Today, 10BASE-T, 100BASE-T and 1000BASE-T are all compatible, and there are low-cost, highly scalable Ethernet cards available that support data transmission at all three rates.

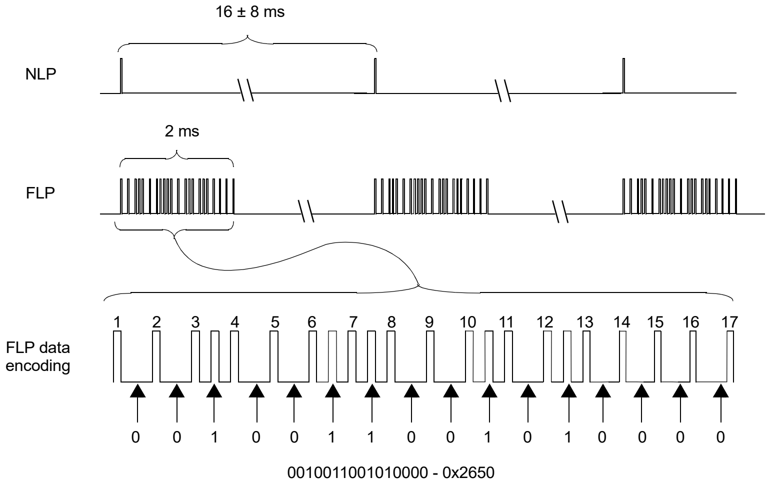


Figure 1.14 NLP pulse train and FLP bursts for 10/100/1000BASE-T auto-negotiation.

1.3.4.2 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.

1.4 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.

1.4.1 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 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 1.15). 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.
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.

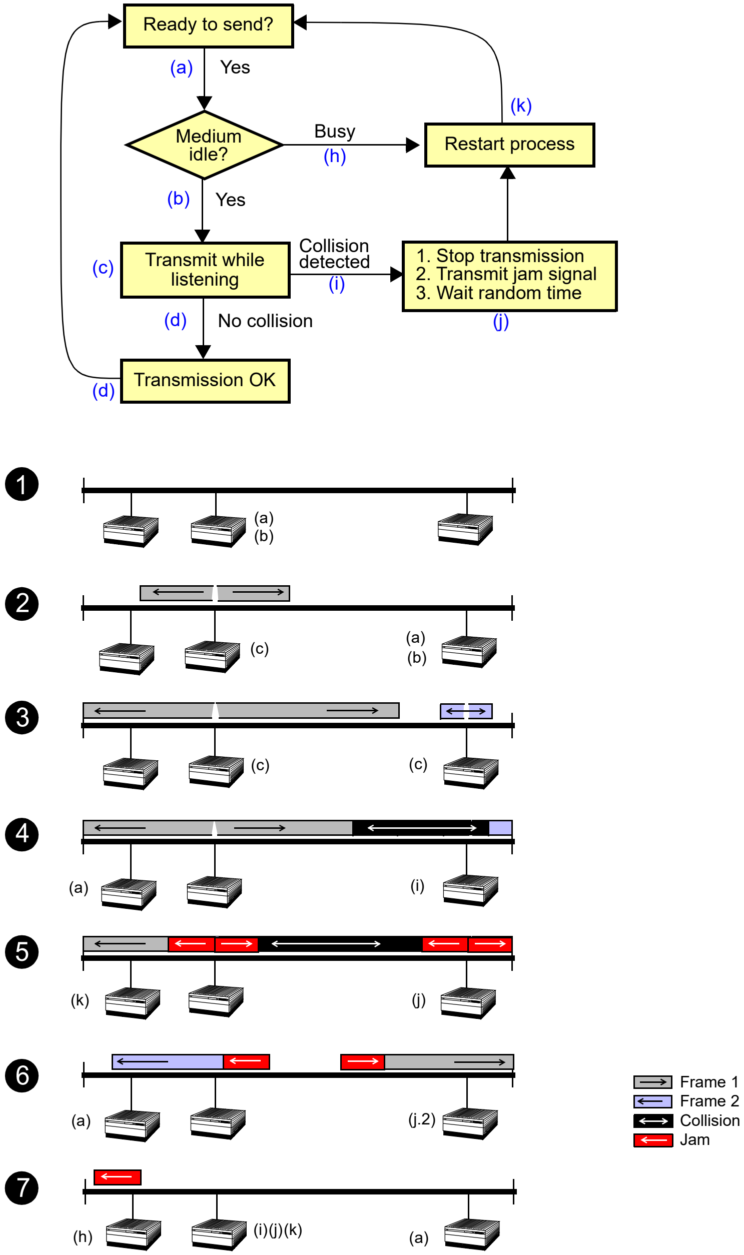


Figure 1.15 CSMA/CD flow chart operation in half-duplex.

1.4.1.1 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 (see Paragraph 1.4.1). 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.

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 Mbit/s and 100 Mbit/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 1.5). When full-duplex versions of Ethernet are used, collisions are avoided, which is why the concept of slot times does not apply.

Table 1.5
Ethernet Timing parameters (half-duplex operation)

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)

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

1.4.1.2 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 1.15).

1.4.2 The Ethernet Frames

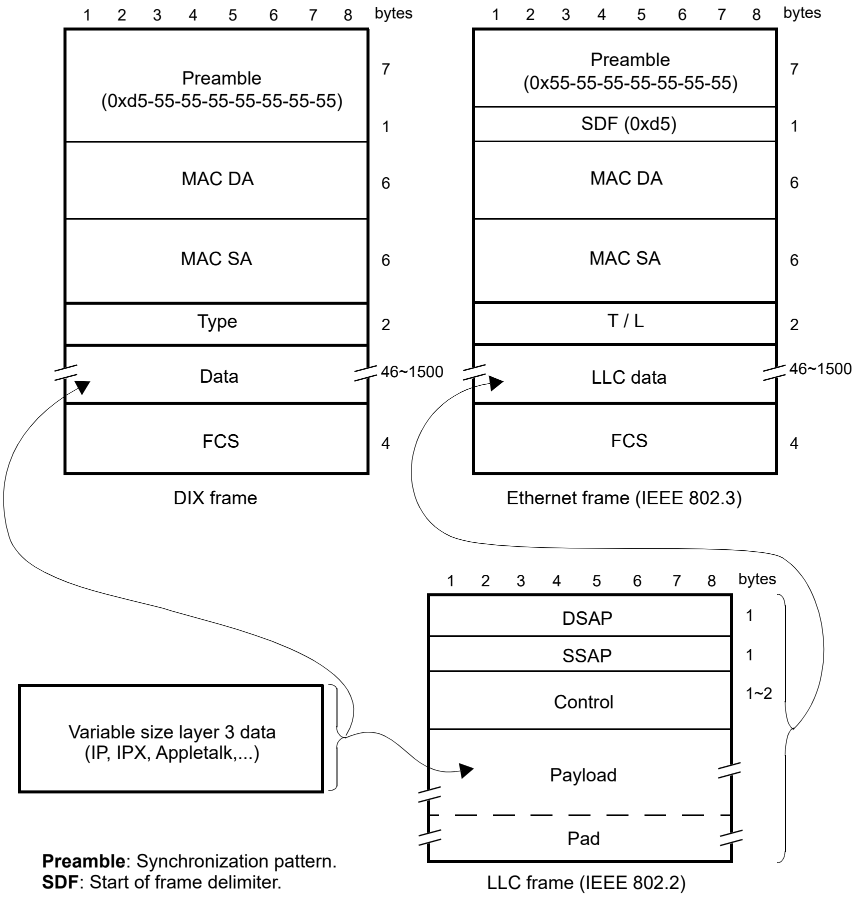
The DIX frame was the first format adopted by the DIX cartel. In 1983, when the IEEE released the first 802.3 standard, the *Start Frame Delimiter* (SFD) field was defined, and this was little more than just a name change. More important was the Length field, since this allows management of the padding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

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 (see Figure 1.16).

1.4.2.1 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.
- *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 (see Figure 1.17).
- *Type***,** A descriptor of the client protocol being transported (IP, IPX, AppleTalk, etc).

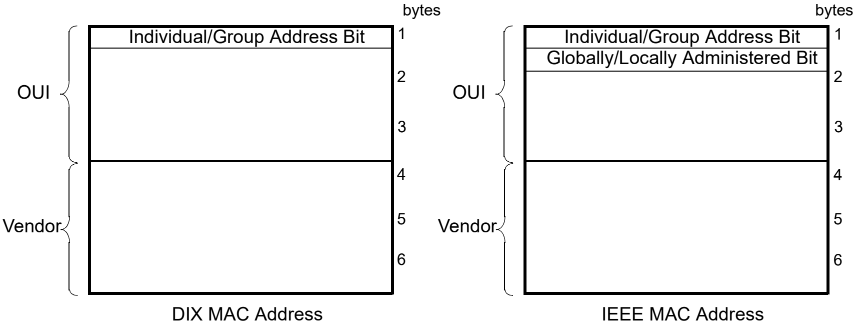


Preamble: Synchronization pattern.
SDF: Start of frame delimiter.
MAC DA: Destination MAC Address.
MAC SA: Source MAC Address.
Type: Indicates the nature of the client protocol (IP, IPX, Appletalk,...).
T/L: Number of bytes of the LLC data if less than 0x0600, otherwise indicates the payload type.
FCS: 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 1.16 The basic 802.3 MAC frame format.

- *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).


bytes
Individual/Group Address Bit
1
2
3
4
5
6
OUI
Vendor
DIX MAC Address
bytes
Individual/Group Address Bit
Globally/Locally Administered Bit
1
2
3
4
5
6
OUI
Vendor
IEEE MAC Address

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 1.17 The 24-bit block administrated 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.

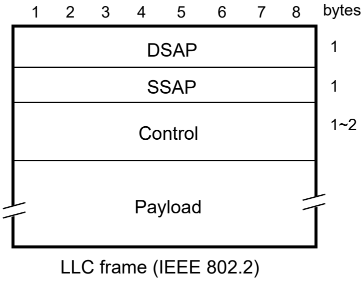
- *Logical Link Control* (LLC)**,** The payload, can contain from 48 up to 1500 bytes of data.
- *Pad***,** All frames must be at least 64 bytes long (see Paragraph 1.3). If the frame is smaller, it contains a pad field to reach the necessary 64 bytes.
- *Cyclic Redundancy Check* (CRC), the value of this field is used to check if the frame has been received successfully, or if the contents have been corrupted.

1.5 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.



DSAP: Destination Service Access Point
SSAP: Source Service Access Point
Control: Miscellaneous LLC control information.

Figure 1.18 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.

1.6 The Network Layer

The Network Layer, or Layer 3, provides end-to-end connectivity between stations that can use heterogeneous underlying technologies (see Figure 1.19) 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 tables manage layer 3 host and network addresses. They are configured manually or by the routing protocols.

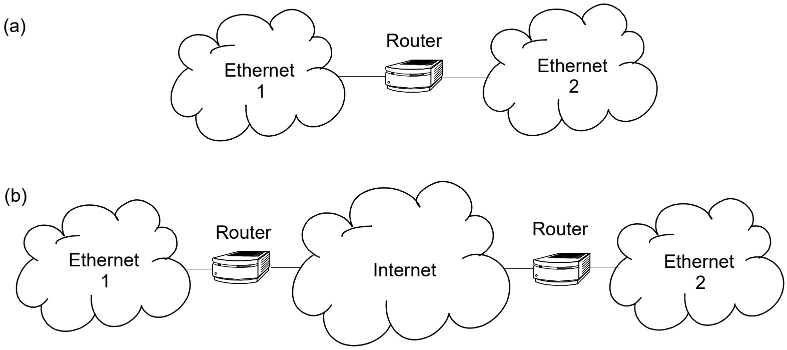


Figure 1.19 Internetworking with routers. (a) A router is used to forward data from an Ethernet network to a second Ethernet network. (b) A chain of routers delivers traffic from an Ethernet network to a second Ethernet network through the Internet.

1.6.1 The Internet Protocol

The *Internet Protocol* (IP) is the most popular open-system 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.

1.6.1.1 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 1.20). For example, 62.22.33.1 is a valid IP address in dotted decimal representation.

Each 32-bit address is divided into two fields:

1. *Network field*, assigned by the Internet Network Information Center 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 1.20), 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.

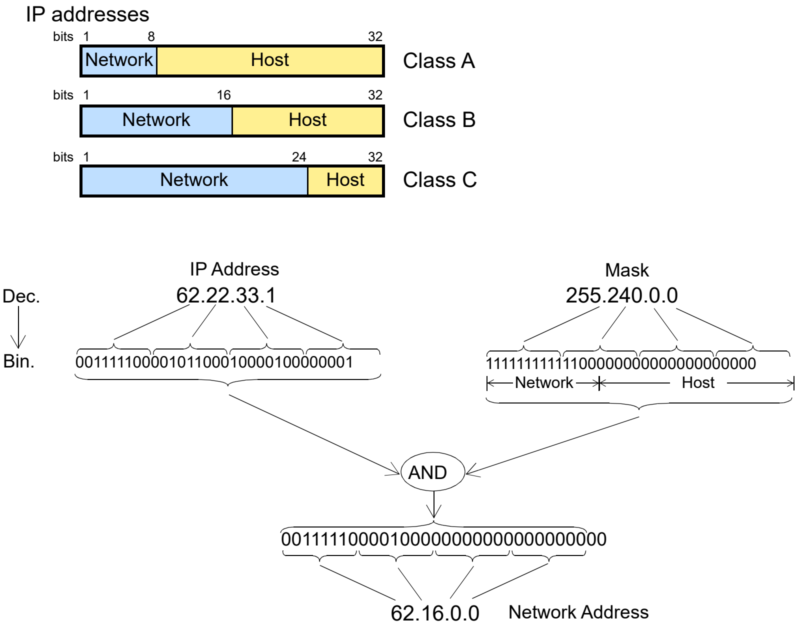


Figure 1.20 Relationship between the IP address, the network mask, the network address and conversion between the binary and the dotted decimal representations. The 32-bits mask has binary 1s in all bits specifying the network field and 0s in the host.

1.6.1.2 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.

1.6.2 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. RFC826 describes the ARP.

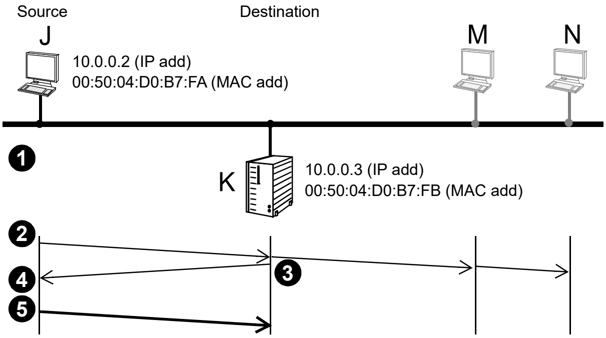


Figure 1.21 ARP operation when source and destination hosts are both in the same Ethernet network.

1.6.2.1 ARP in a segment

First case (see Figure 1.21). 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.

1.6.2.2 ARP in different LAN segments

Second case (see Figure 1.22). Imagine source and destination hosts attached to different segments connected by a router in a LAN:

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.

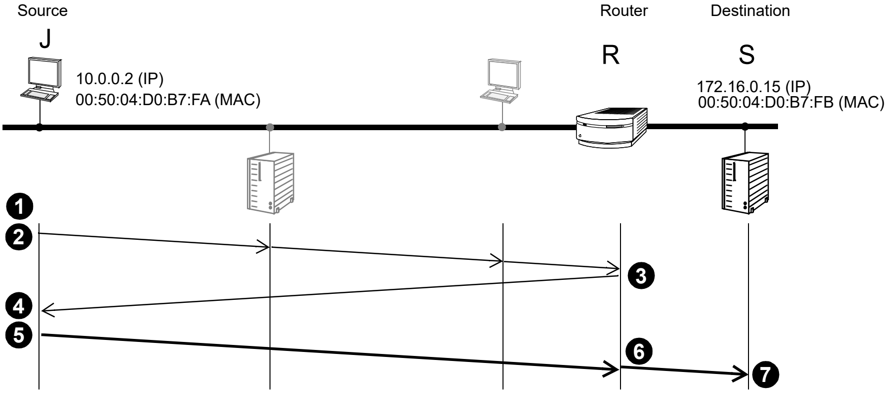


Figure 1.22 ARP operation when source and destination hosts are in different Ethernet networks and the information is forwarded between them by a router.

1.6.2.3 ARP in heterogeneous connections

Third case (see Figure 1.23). 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).

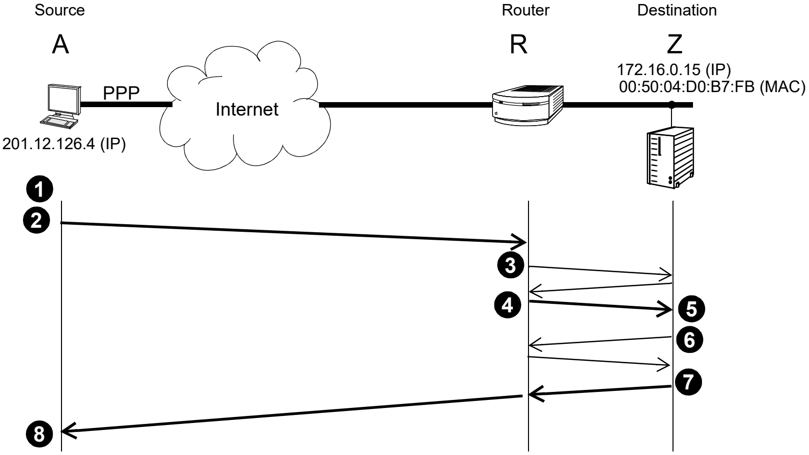


Figure 1.23 Access to a remote host placed behind a NAT firewall with a PPP connection to the Internet.

Routers connecting LANs to the Internet must facilitate the internal host's, 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 1.23):

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.

1.7 Installing Gigabit Ethernet

Before installing Gigabit Ethernet, it is essential to know the equipment offered, and the place in the market that each equipment occupies. The following is a basic view:

- *1000BASE-T*: Using UTP Cat. 5 cabling guarantees compatibility with 10/100BASE-T installations, which means that it is easy to integrate and migrate to.
- *1000BASE-CX*: Designed to run on STP cabling, but completely overshadowed by 1000BASE-T, this is not supported any more by manufacturers.
- *1000BASE-SX*: Short-wavelength laser transmitted over multimode fiber, good for campus and metro.

- *1000BASE-LX*: Long-wavelength laser transmitted over single/multimode fiber, developed for the most demanding applications in terms of distance and quality.

1.7.1 Some Things to Consider when Migrating

Many requirements and strategic considerations must be taken into account when upgrading the existing networks to Gigabit Ethernet. While networking equipment can be removed easily, horizontal cabling can be very difficult and expensive to replace.

1.7.1.1 Horizontal Cabling

For horizontal cabling, or cabling for desktop users, the choice is copper for 10/100/1000 Ethernet. 1000BASE-T will run on Cat. 5 cables (or better) for distances of up to 100 meters, and this bit rate is enough for most users for the next couple of years. Existing Cat. 5 links should be able to support all current Ethernet rates from 10 Mbit/s to 1000 Mbit/s, although they should be *tested* to make sure that they can support gigabit rates.

Unless there are security or *Electromagnetic Interference* (EMI) concerns about copper cabling, or stations are more than 100 meters apart, there is no compelling business reason to deploy fiber to the desktop. Optical cabling is more expensive, and cannot power network-attached devices. In contrast, IP phones, LAN web cams, and other devices can be powered by using copper cabling.

1.7.1.2 Vertical Cabling

For vertical cabling, the best choice today is to deploy a mixture of multimode and single-mode devices and cabling, depending on the requirements and the budget available (see Figure 1.24). By using a mixed network, network managers achieve backward compatibility with conventional 100BASE-FX, which uses L*ight Emitting Diodes* (LEDs) and will not run on single-mode fiber. Backward-compatibility can also be achieved with 1000BASE-SX, which is cheaper than 1000BASE-LX. Single-mode fiber will always be more expensive than multimode fiber.

1.7.1.3 Core and Long Haul

For core and long-haul installation, single-mode fiber is the best option. It can support all backbone lengths up to 10 000 meters at 1000 Mbit/s, but it will also be capable of supporting backbone use at 10-gigabit data rates in the future.

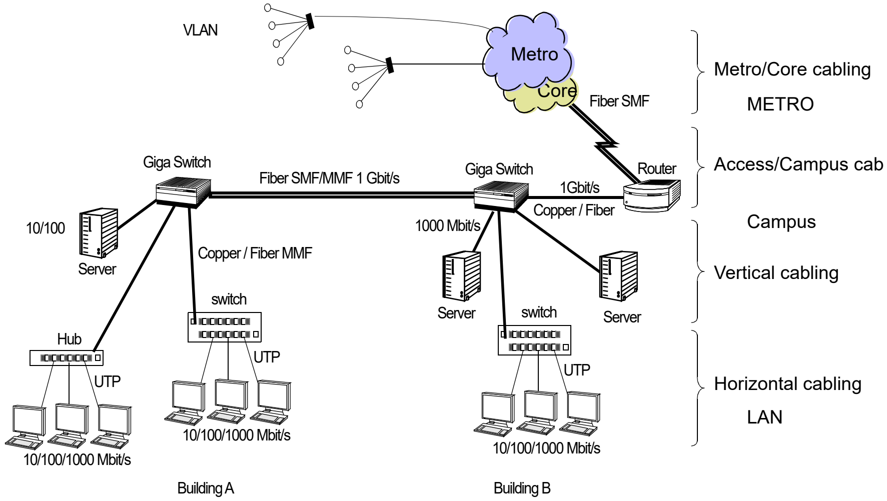


Figure 1.24 Migration to Gigabit Ethernet, cabling and bit rates.

1.7.2 Gigabit Interface Converter

The *Gigabit Interface Converter* (GIBIC) modules have become the most commonly used interface for Gigabit Ethernet, providing hot-swappable modules that fully support each physical layer. There are many manufacturers, and the equipment is small and pluggable in the standard slot.

Selected Bibliography

[1] IEEE Std 802.3™-2005, “Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications”, 9 December 2005.

[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] RFC 2544, *Benchmarking Methodology for Network Interconnect Devices*, S. Bradner and J. McQuaid, March 1999

[6] RFC 1242 - *Benchmarking terminology for network interconnection devices,* S. Bradner July 1991

[7] RFC 2285, *Benchmarking Terminology for LAN Switching Devices*, R. Mandeville Network Lab-

oratories February 1998

[8] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.

[9] 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.

Chapter 2

Switched Ethernet

Since the first Ethernet card was manufactured, the technology has evolved continuously, showing a great ability to adapt to new technologies as well as to the growing business needs.

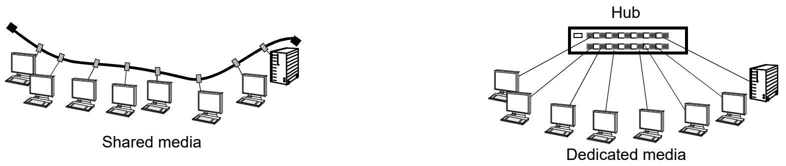


Figure 2.1 From shared media to dedicated media.

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 transmission resources.

2.1 The Evolution 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* (see 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 stations that were causing problems.

One of the fundamental problems 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 (see Figure 2.2). To solve this problem, Ethernet switches and bridges were introduced (see Paragraph 2.2). These devices forward MAC frames by means of MAC addresses, and make better use of transmission resources than hubs based on frame broadcasting.

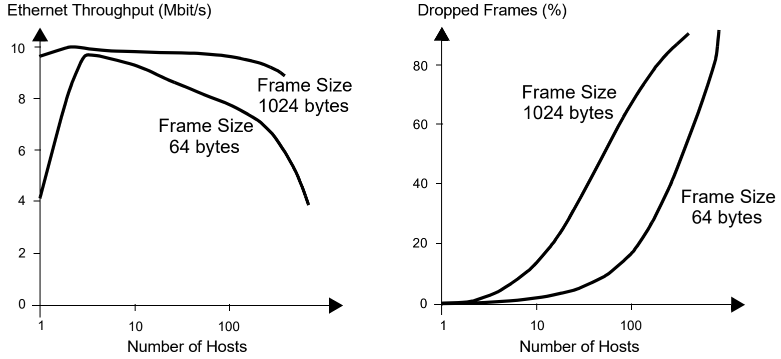


Figure 2.2 Shared Ethernet collapses as the number of hosts increases. Optimum performance is obtained for three or four stations.

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 (see 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.

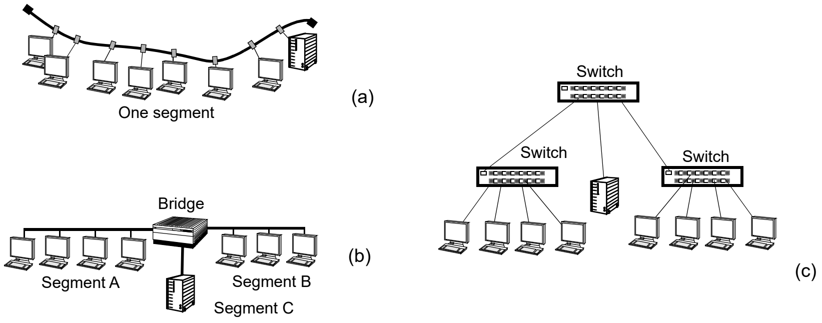


Figure 2.3 Segmentation and switching. (a) Shared Ethernet with bus topology. (b) Segmentec Ethernet. (c) Microsegmented Ethernet.

2.2 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.

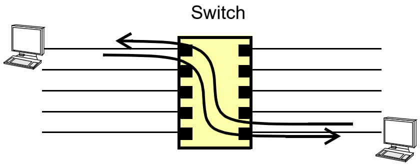


Figure 2.4 Segmentation reduces the probability of collision; full-duplex and switching removes it completely.

In micro-segmented networks, each station is connected to a switch port. Switching eliminates the possibility of collisions, making CSMA/CD (see Paragraph 1.4.1) unnecessary. 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.

2.3 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. A station connected to a 100BASE-T interface can transmit at 100 Mbit/s bit rate and receive data simultaneously at 100 Mbit/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.

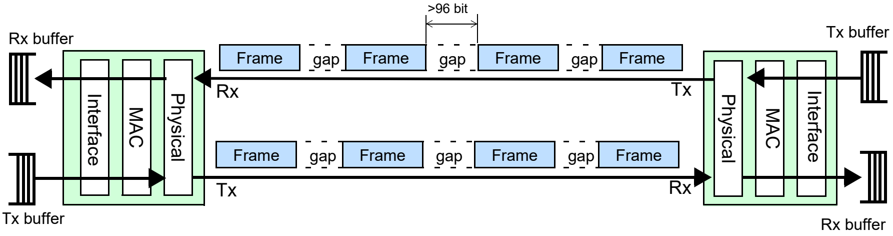
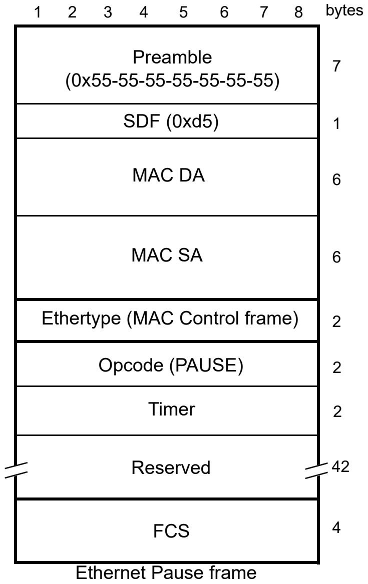


Figure 2.5 FDX operation enables two-way transmission simultaneously without contention, collisions, extension bits or retransmissions. The only restriction is that a gap must be allowed between two consecutive frames. FDX also requests flow control, which is transmitted by the receiver to request that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

One side-effect of FDX occurs when a transmitter that is constantly sending packets causes 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.

This protocol is based on a short packet known as a Pause frame (see Figure 2.5). 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 uncongestioned 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.



Ethertype: Indicates MAC control (0x88-08)
Opcode: Indicates Pause frame (0x00-01)
Timer: Time is requested to prevent 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.

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 Mbit/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 (see Paragraph 1.4.1.1). 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 devices. Therefore there is not half-duplex operation for them. They always run with full-duplex mode.

2.4 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 2.7). VLANs are an 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 2.8).

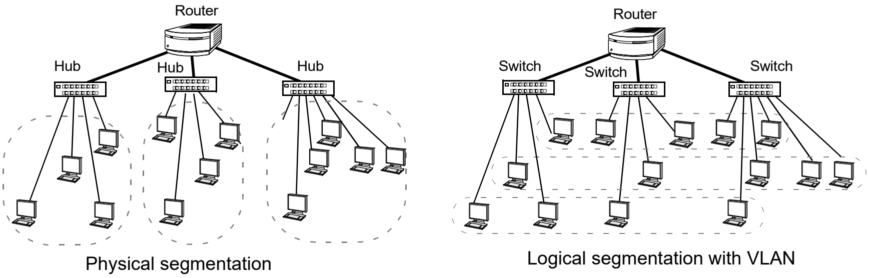


Figure 2.7 Virtual LAN vs. Segmented LAN.

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 2.9) is defined in the IEEE 802.1Q standard.

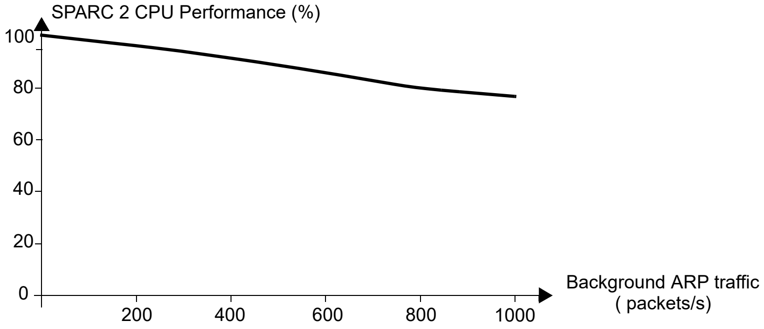
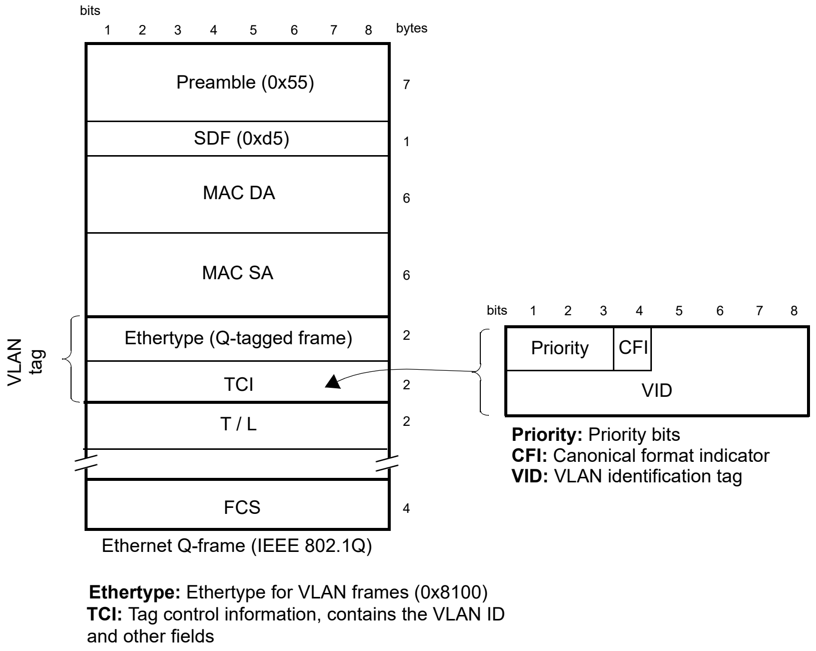


Figure 2.8 Loss of performance of a workstation due to broadcast Ethernet frames transportin ARP data.

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.



Priority: Priority bits
CFI: Canonical format indicator
VID: VLAN identification tag

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

Figure 2.9 Ethernet frame with 802.1Q VLAN field structure.

2.5 TOPOLOGIES

The first Ethernet networks were implemented with a *coaxial bus structure*, up to 100 stations per segment. Individual 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 logical 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 center to the station. This topology has proven to be the most flexible and easy to manage in LAN networks, and it is independent of the technology and the physical medium being used.

New high-speed versions have gained increasing acceptance since the year 2000, competing for the campus and metropolitan markets where *point-to-point*, *ring*, and even *meshed* topologies are common. The adoption of fiber optics has been key to increasing distance and bit rate. A new standard for the local loop has been approved (IEEE P802.3ah, June 2004), so that Ethernet can compete for broadband access where twisted copper pair is the common physical layer.

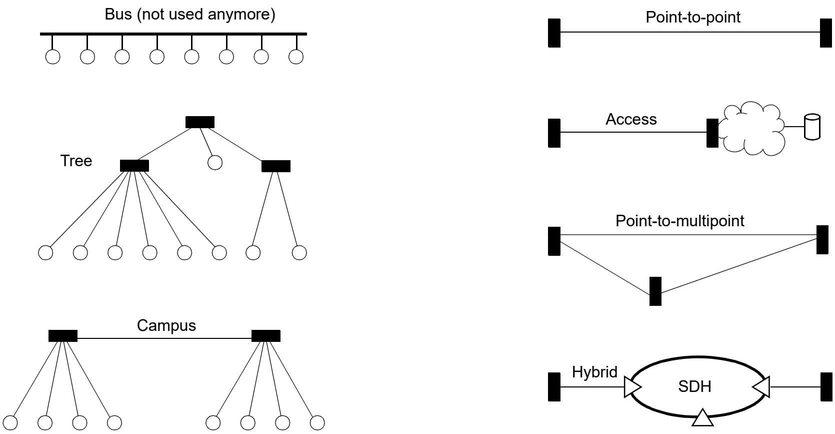


Figure 2.10 *Topology* is a very general term, and only very simple networks can be classified into one topology only. Ethernet networks could be a combination of several trees interconnected to other trees and remote services as well, using different solutions such as point-to-point, multipoint or hybrid.

2.6 Gigabit and 10 Gigabit Applications

Just a few years ago Ethernet was just a successful LAN technology. Standardization of long-haul optical 1 Gbit/s and 10 Gbit/s interfaces makes it possible to use pure Ethernet solutions in MAN and WAN environments as well as in LAN.

2.6.1 LAN Applications

In LAN, the new Ethernet interfaces are used where legacy Fast Ethernet interface has become a bottleneck. Some of these LAN applications are:

- Aggregation of multiple lower-rate segments in switch/server links
- High-traffic interswitch trunk links
- Interconnection of servers in clusters of servers

- LAN extension between buildings in campus applications, or communication between remote LANs

Two gigabit interfaces have been especially designed to reuse the installed LAN infrastructure, 1000BASE-T and 10GBASE-LX4. The former can reuse the Cat 5 UTP cable with the same distance limitations as Fast Ethernet, and the latter is suitable to take advantage of the installed MMF over short distance (300 m) links. The 10GBASE-LX4 can also be used for long-distance (up to 10 km) applications, if 10 μm SMF is used instead of MMF.

2.6.2 MAN Applications

Metropolitan Area Networks are perhaps the most attractive environment for new Ethernet technologies. The use of 10GbE in MAN is rather similar to the use of GbE, but 10GbE makes it possible to have some hierarchy within the network. Interfaces working at 10 Gbit/s can be used for the backbone, and 1 Gbit/s can be used for access links.

Under certain circumstances, legacy SDH/PDH access and metropolitan networks can be replaced by Ethernet. Especially in cases like VPN, where the network only transports IP data generated at Ethernet stations. This way, the metro *service provider* (SP) can benefit from technological simplification and reductions in maintenance costs.

The case of SDH network installations transporting circuits being used by Frame Relay (FR) or Integrated Digital Services Network (ISDN) is quite different. In this case it is probably better to migrate to the Ethernet packet friendly *Next Generation SDH*, while partially substituting the access network with Ethernet over multimode or monomode fiber.

Metro service providers have already started to install gigabit services with dark fiber as a fundamental transport facility. This could be a cost-effective alternative to the traditional SDH/SONET network. It is difficult to scale this solution beyond dedicated point-to-point circuits due to the broadcast nature of Ethernet traffic. A simple service provider network based only on MAC switching of traffic from various customers would be difficult to manage, and it would also cause security problems.

An interesting alternative to the use of dark fiber for gigabit and multigigabit services is Ethernet over dark wavelength when a WDM-based network is available in the MAN. WDM networks are transparent to the upper transmission layer, and virtually any technology, including Ethernet, can be deployed over them. The problem of this solution, again, is the lack of scalability and granularity of the bit rate: it would be expensive to provide a complete wavelength, if its transmission bandwidth is not going to be used (see Figure 2.11).

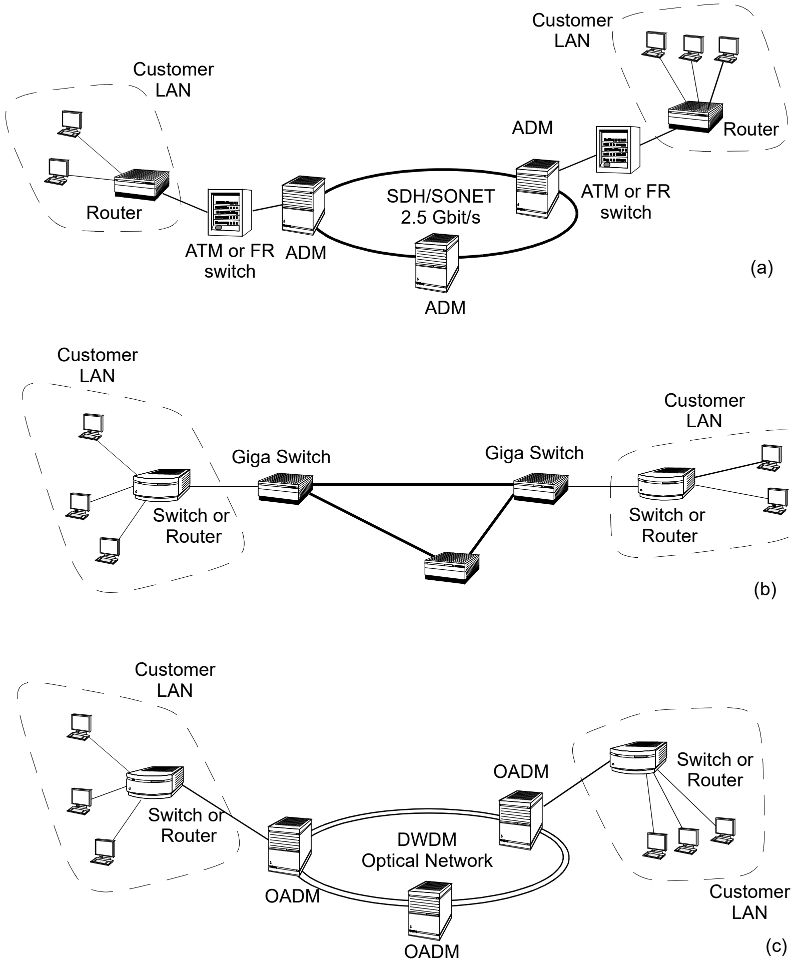


Figure 2.11 10GbE applications for MAN: (a) Classical MAN based on SDH/SONET. Ethernet is not used in the SP network. (b) Ethernet over dark fiber solution. (c) Ethernet over dark wavelength solution.

The *Metro Ethernet Forum* (MEF), the *Internet Engineering Task Force* (IETF) and the ITU-T are working to find the solutions to allow the deployment of Carrier Class Ethernet that will make it possible to define bandwidth profiles similar to the existing Frame Relay and ATM access services. This includes the definition of generic services and interfaces. Carrier Class Ethernet is not only a low-cost solution to interface with the subscriber network and carry its data across long distances, but a true convergent network for any type of information, including voice, video and data.

2.6.3 WAN Applications

While Ethernet is expected to be competitive with traditional transport technologies (FR, ATM, SDH/SONET) in MAN environments, in wide area networks Ethernet will continue to be a complementary transmission technology for some time (see Figure 2.12).

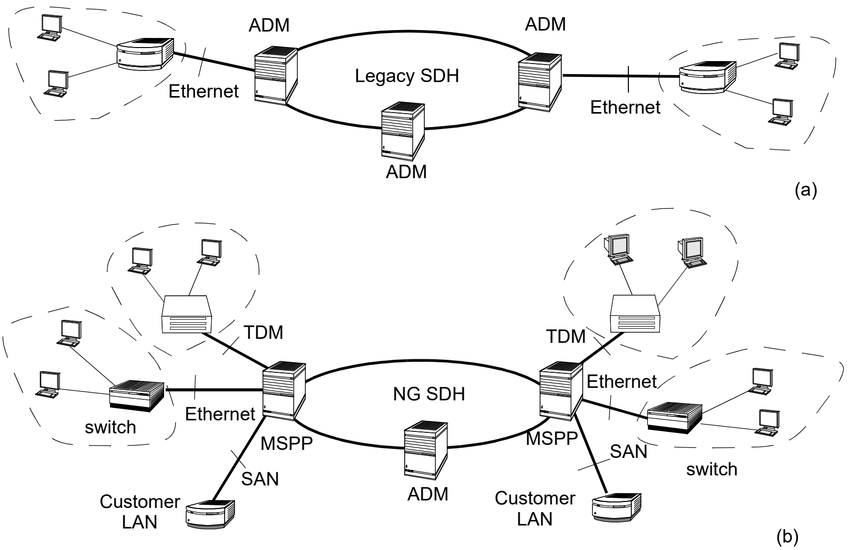


Figure 2.12 Coexistence of Ethernet with legacy transport technologies: (a) WIS interfaces are used for easy internetworking with legacy SDH equipment. (b) NG-SDH equipment allows to use the existing infrastructure for transporting either TDM or the new packet-based services.

WANs are completely dominated by SDH/SONET and ATM equipment. Big investments in these technologies and the lack of maturity of Ethernet for WAN makes it impossible to quickly migrate to an Ethernet-based native packet core network.

The 10GBASE-W allows easy internetworking with SDH/SONET equipment (see Paragraph 1.3.3.2). The SDH/SONET network has developed, and nowadays it is packet aware. This *Next Generation SDH* is an interesting alternative for packet transport over the backbone optical network.

However, it is expected that with time, Ethernet or other packet-based technologies will play a more important role in the WAN. Some experts think that moving the existing SDH/SONET functions to the optical layer will make TDM transport unnecessary, and SDH/SONET will disappear from the protocol stack.

The evolution of the optical layer will probably lead to a *Generalized MultiLabel Switching* (GMPLS) managed optical *Dense WDM* (DWDM) network. Ethernet will be mapped directly over different lambdas, and it will perform similar functions as ATM today: service multiplexing, layer-2 switching and QoS provisioning.

2.7 The Future of Ethernet

Manufacturers have recently been working on a higher-speed version of the *Synchronous Optical Network* (SONET), boosting its capacity from 10 Gbit/s to 40 Gbit/s, which may have an impact on the future of Ethernet. One group of manufacturers wants to piggyback on this work and develop a 40 Gbit/s version of Ethernet based largely on the STM-256/OC-768 specification.

However, other groups want to maintain the ‘multiple of 10’ strategy, which would see 100 Gbit/s Ethernet as the next logical step. There is also a suggestion that vendors are more interested in putting 10 Gigabit Ethernet into the local telephone exchanges in order to obtain better returns, than investing in higher-speed Ethernet. Faster Ethernet definitely has a future, but its placement and time scale are very uncertain at the moment.

Selected Bibliography

[1] IEEE Std 802.3™-2005, “Part 3: Carrier sense multiple access with collision detection

[2] 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.

[3] Rich Seifert, *Gigabit Ethernet Technology and Applications for High/Speed LANs*, Addison Wesley Oct 1999.

[4] William Stallings, *Data and Computer Communications*, Prentice Hall, 1997.

[5] Kevin L. Paton, *Gigabit Ethernet Test Challenges*, Oct 2001 Test and Measurement World Magazine.

[6] RFC 2544, *Benchmarking Methodology for Network Interconnect Devices*, S. Bradner and J. McQuaid, March 1999.

[7] RFC 1242 - *Benchmarking terminology for network interconnection devices,* S. Bradner July 1991.

[8] RFC 2285, *Benchmarking Terminology for LAN Switching Devices*, R. Mandeville Network Laboratories February 1998.

[9] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.

[10] 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.

Chapter 3

IP Multicasting

In the 1990's, the most important Internet application was the *World Wide Web* (WWW), a unicast client/server application. For the WWW, an application-layer protocol called *Hypertext Transfer Protocol* (HTTP) is used by the client to request information from a specialized server. This server then sends HTTP answers to the client. Another popular application for the Internet is e-mail, where one sender delivers mail messages to the mailboxes of one or more recipients using an application-layer protocol, the *Simple Mail Transfer Protocol* (SMTP). The e-mail architecture is quite different from the WWW. E-mail is also a client/server application, but it uses multicasting. In this case, the problem of multicast delivery is solved by SMTP. From the point of view of the network, the paths to the different recipients of an e-mail are seen as different routes by the intermediate routers.

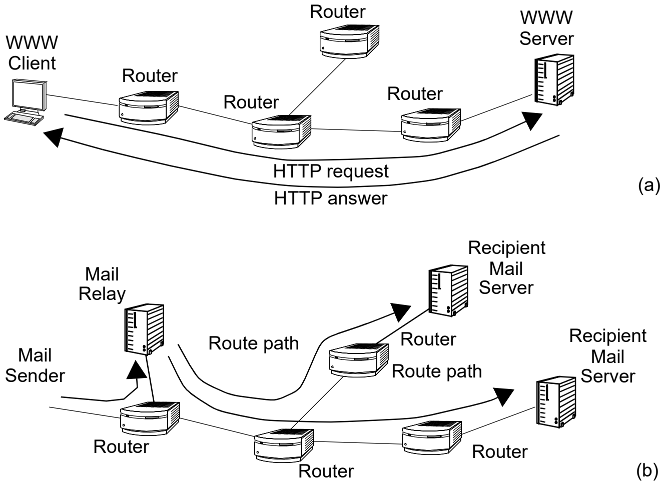


Figure 3.1 Service model for IP data applications: (a) The WWW service model: Unicast and client/server-based. (b) The e-mail service model: Multicast application; multicast delivery is solved at the application layer.

Today, the convergence of circuit-switched and packet-based technologies is making it possible to deliver voice, video and data using one single IP network. Many of the new multimedia applications, such as IPTV or VoIP are multicast, but the multicast service model of the e-mail application is not enough for these applications. It does not have the scalability and simplicity needed to deliver real-time voice or video to hundreds or maybe thousands of receivers. To make an IP network support triple play, it is necessary to implement multicasting in the network layer of the protocol stack. This means that multicast delivery must be supported by intermediate routers as well, not only by terminal equipment.

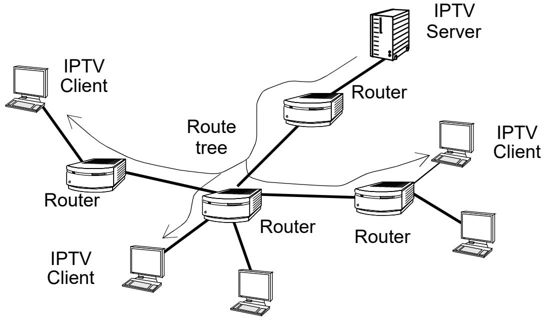


Figure 3.2 Service model for the IPTV application. Multicast delivery is solved at the IP layer and not at the application layer of the protocol stack.

The delivery of voice and video makes it necessary to install multicast routing protocols in IP routers. The aim of these protocols is to find routing trees to send multicast data to the destination; in other words, choose the correct outgoing interface for incoming multicast packets. The second problem to be solved is how to manage multicast groups. This means implementing a mechanism to add and remove recipients dynamically. This is done by using multicast addresses and the *Internet Group Management Protocol* (IGMP).

3.1 IP Multicast Groups and Their Management

In the Internet and in any IP network, multicasting means the transmission of IP datagrams to groups of hosts called multicast groups. Every multicast group is identified by a single IP address. Most of the IP addresses are used to identify hosts in the network, but some of them are reserved not for hosts but for groups of hosts. The multicast IP addresses are Class D addresses, which means that they all start with "1110". Expressed as decimals, these addresses fall in the range from 224.0.0.0. to 239.255.255.255.

Table 3.1
Internet address classes

Address class	*First byte (binary)*	*Address range (decimal)*	*Number*
Class A	0xxxxxxx	0.0.0.0 ~ 127.255.255.255	2,147,483,648
Class B	10xxxxxx	128.0.0.0 ~ 191.255.555.555	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

The *Internet Assigned Numbers Authority* (IANA) is in charge of the administration of the Internet addressing space, including multicast addresses. Some addresses within the Class D range are reserved by the IANA for specific applications, and others are dynamically assigned to transient multicast groups.

Those hosts that need to send data to a multicast group simply put the right multicast address in the destination address field of the IP datagrams sent to this group. There are not many requirements for the transmitter, but IP multicasting must be supported by the network, and the IGMP must be supported by the receiver in order to join or leave multicast groups. Supporting multicasting in IP networks is not mandatory; there are several compliance levels that range from no compliance at all to full compliance.

Table 3.2
Some IP multicast addresses reserved by the IANA

Address	*Receivers attached to the group*
224.0.0.1	All Systems on this Subnet
224.0.0.2	All Routers on this Subnet
224.0.0.5	All OSPF Routers
224.0.0.6	Designated OSPF Routers
224.0.0.9	RIP2 Routers
224.0.0.12	DHCP Server / Relay Agent
224.0.1.1	Network Time Protocol

There is no obligation for the members of a multicast group to be in the same network. However, in a very simple IP multicasting model, all the members of the group are in the same Ethernet network with the sender. In this case, the IGMP is not needed, and it is possible to take advantage of the broadcast nature of Ethernet.

If not all the members of the multicast group are in the same network, a *multicast agent* is needed. This is normally a router that delivers the data to the correct destination. The routers in the multicast IP network must use a multicast routing protocol to be able to perform this task.

3.1.1 Multicasting in Ethernet Networks

The IANA controls the block of Ethernet MAC addresses between 01-00-5e-00-00-00 and 01-00-5e-7f-ff-ff. When transmitting IP multicast data in an Ethernet network, the destination address can be automatically mapped into a multicast Ethernet address. The 28 least significant bits of the IP address are copied to the IANA-managed Ethernet address block explained above. For example, the multicast IP address 239.255.0.1 would be mapped into 01-00-5e-7f-00-01 for Ethernet.

The information arrives to all the hosts in the network, and those belonging to the correct group can recognize the multicast Ethernet address and process the correct frames. Frames arriving to those hosts who do not belong to the group are discarded.

Note that in the mapping process, five bits of the IP address are lost. This means that there are 32 IP multicast addresses with the same Ethernet address. This does not have to be a problem, if the use of IP multicast addresses is planned carefully.

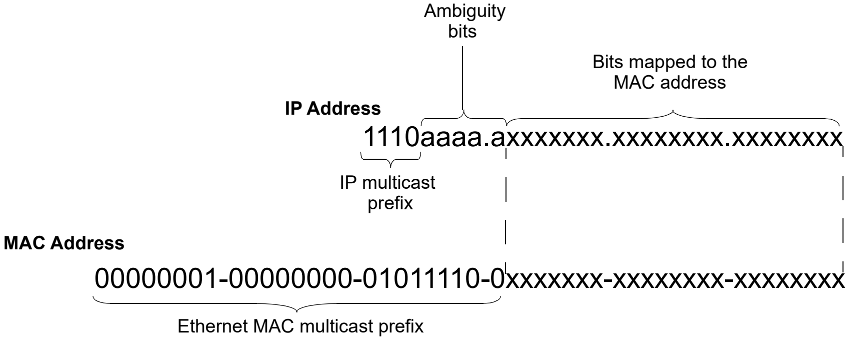

Ambiguity bits
Bits mapped to the MAC address
IP Address
1110aaaa.axxxxxxx.xxxxxxxx.xxxxxxxx
IP multicast prefix
MAC Address
00000001-00000000-01011110-0xxxxxxx-xxxxxxxx-xxxxxxxx
Ethernet MAC multicast prefix

Figure 3.3 Mapping of IP multicast addresses into Ethernet multicast MAC addresses

Ethernet multicasting works in a similar way in broadcast Ethernet networks with shared transmission media and in switched networks with dedicated media. The reason for this is that multicast traffic is broadcast by switches by default. In other words, switched Ethernet behaves the same way as broadcast Ethernet for multicast traffic. This could be a problem, especially in large networks, because multicast Ethernet traffic is directed to hosts that do not request this traffic. In bridged carrier-class networks this may have security implications as well.

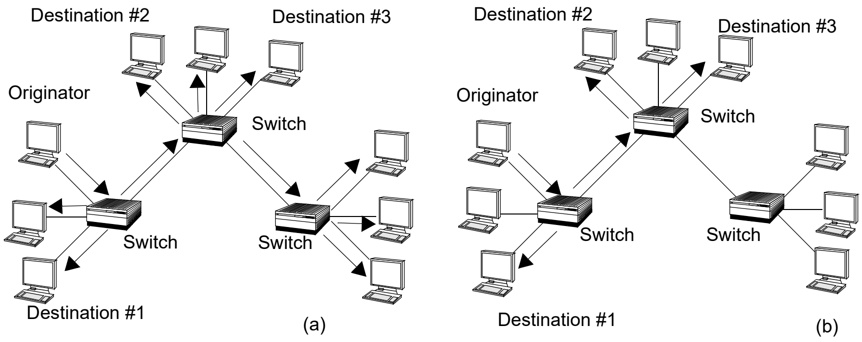


Figure 3.4 (a) Switches not implementing IGMP snooping. (b) Switches implementing IGMP snooping.

It is possible to avoid broadcasting of multicast Ethernet addresses, if IGMP snooping is implemented in the switches. IGMP snooping means that layer-2 switches can check (snoop) layer- 3 IGMP messages. This way, the switches know which are the active multicast groups and who belongs to which group, and they can forward multicast data to the correct ports.

3.1.2 Multicasting and the Internet Group Management Protocol

The IGMP is the network protocol that allows hosts to join or leave multicast groups, and therefore it must be implemented by all hosts that wish to receive multicast information.

An important component of the IP multicasting architecture is the multicast agent. This entity is in charge of:

- *Group management*: The multicast agent grants or denies multicast group membership to hosts. It finds and assigns IP addresses and access keys to the groups and removes inactive hosts from them.
- *Multicast routing*: The multicast agent forwards multicast traffic to remote networks. To do this, it associates IP addresses to sets of interfaces. It is not necessary to maintain a full list of members (and their IP addresses) for every group; routing can be performed just by knowing the network interfaces where the information needs to be forwarded to.

IGMP messages are transmitted between hosts and multicast agents that establish a client/server relationship. Hosts are clients and multicast agents are servers. IGMP packets have the same format, defined by the IETF to include the following fields:

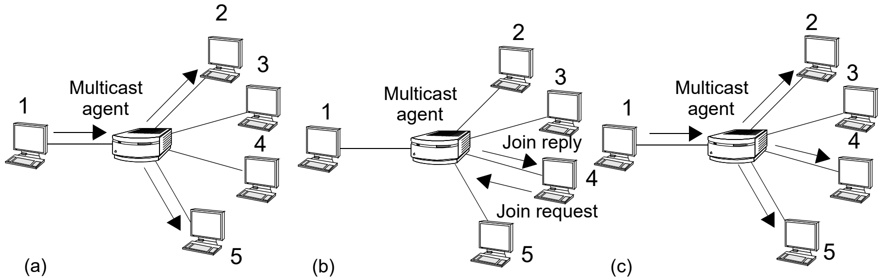
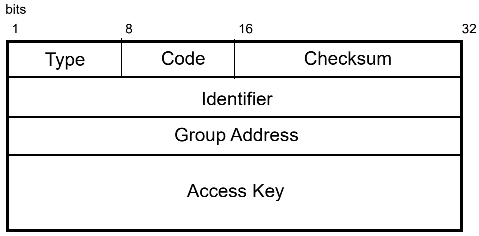


Figure 3.5 Joining a multicast group with IGMP: (a) The multicast agent routes traffic only to the members of a multicast group (hosts 2 and 5). (b) Host 4 requests joining the multicast group, and the multicast agent grants membership. (c) Now, the multicast agent routes the traffic to hosts 2, 3 and 5.

- *Type*: Describes the purpose of the IGMP packet. IGMP messages can be used to join or leave groups, create new multicast groups, or confirm membership to a particular group.
- *Code*: This field is only meaningful when creating a new group. It shows whether the group is going to be public (value 0) or private (value 1). Public groups can be joined freely, but private groups are protected with an access key. In IGMP replies sent by multicast agents, the Code field shows whether the request is granted, denied or pending.
- *Checksum*: This field is for error detection purposes in IGMP messages. It is the one's complement of the one's complement sum of the IGMP message. The Checksum field itself is considered to be zero for this calculation.



Type	Message
0x01	Create Group Request
0x02	Create Group Reply
0x03	Join Group Request
0x04	Join Group Reply
0x05	Leave Group Request
0x06	Leave Group Reply
0x07	Confirm Group Request
0x08	Confirm Group Reply

Figure 3.6 IGMP message format

- *Identifier*: It is useful to distinguish between the different request messages arriving from the same host. This is why different IGMP messages from the same host have a different Identifier field. When the multicast agent sends a reply, it uses the same identifier as the request message.

- *Group Address*: The group address is assigned by the multicast agent; thus, when creating a new group, the address is sent to the host in the Create Group Reply message. When managing group membership, the IP address of the group to be joined or left is sent in the join/leave request message.
- *Access key*: This field is only meaningful when managing private groups. The access key for private groups is assigned by the multicast agent in the Create Group Reply message. The host needs to supply this key in the request message in order to join or leave a private group.

Selected Bibliography

[1] Ramalho M., "Multicast Routing Protocols: A Survey and Taxonomy," *IEEE Communications Surveys*, vol. 3, no. 1, first quarter 2000.

[2] Atwood J.W., "A Classification of Reliable Multicast Protocols," *IEEE Network Magazine*, May/June 2004, pp. 24-34.

[3] Sahasrabuddhe L. H., Mukherjee B., "Multicast Routing Algorithms and Protocols: A Tutorial," *IEEE Network Magazine*, January/February 2000, pp. 90-102.

[4] Bo Li, Jiangchuan Liu, "Multirate Video Multicast over the Internet: An Overview," *IEEE Network Magazine*, January/February 2003, pp. 24-29.

[5] Gossain H., De Morais Cordeiro C., Agrawal D. P., "Multicast: Wired to Wireless," *IEEE Communications Magazine*, June 2002, pp. 116-123.

[6] Striegel A., Maniwaran G., "A Survey of QoS Multicasting Issues," *IEEE Communications Magazine*, June 2002, pp. 82-87.

[7] Smijanic A., "Scheduling of Multicast Traffic in HIgh-Capacity Packet Switches," *IEEE Communications Magazine*, November 2002, pp. 72-77.

[8] Dutta A., Chennikara J., Wai Chen, Altintas O., "Multicasting Streaming Media to Mobile Users," *IEEE Communications Magazine*, October 2003, pp. 2-10.

[9] Mir N. F., "A Survey of Data Multicast Techniques, Architectures, and Algorithms" , *IEEE Communications Magazine*, September 2001, pp. 164-170.

[10] Maxemchuk N. F., "Reliable Multicast with Delay Guarantees," *IEEE Communications Magazine*, September 2002, pp 96-102.

[11] Shapiro J. K., Towsley D., Kurose J. "Optimization-Based Congestion Control for Multicast Communications," *IEEE Communications Magazine*, September 2002, pp. 90-95.

[12] Deering S. E., "Host Extensions for IP Multicasting", IETF Request For Comments RFC 1112, August 1989.

[13] Fenner W., "Internet Group Management Protocol, Version 2", IETF Request For Comments RFC 1112, November 1997.

[14] Cain B., Deering S., Kouvelas I., Fenner B., Thyagarajan A., "Internet Group Management Protocol, Version 3", IETF Request For Comments RFC 3376, October 2002.

Chapter 4

Ethernet 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.

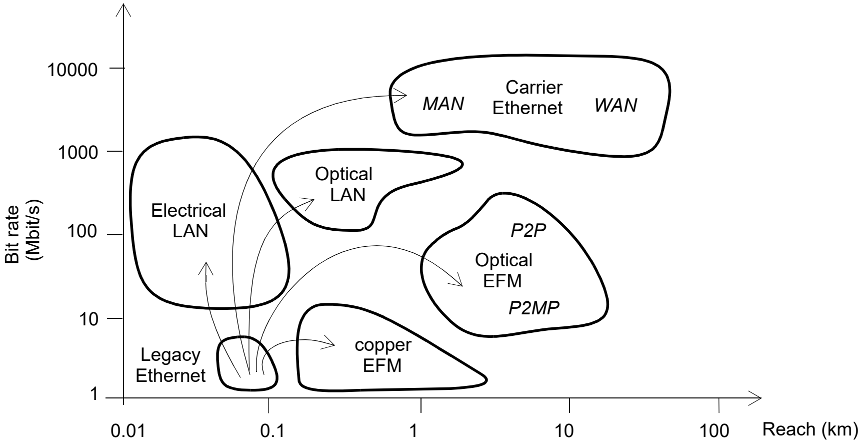


Figure 4.1 Ethernet applications and EFM

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 available for both PON networks and active Ethernet (see Table 4.1).

4.1 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

Table 4.1
EFM Interface Summary

Interface	*Medium*	*Wavelength (nm)*	*Rate (Mbit/s)*	*Reach (km)*
100BASE-LX10	Two single-mode fibers	1310	100	10
100BASE-BX10	One single-mode fiber	1310 (US), 1550 (DS)	100	10
1000BASE-LX10	Two single-mode fiber	1310	1000	10
1000BASE-LX10	Two multimode fiber	1310	1000	0.55
1000BASE-BX10	One single-mode fiber	1310 (US), 1490 (DS)	1000	10
1000BASE-PX10	One single-mode fiber PON	1310 (US), 1490 (DS)	1000	10
1000BASE-PX20	One single-mode fiber PON	1310 (US), 1490 (DS)	1000	20
10PASS-TS	One or more telephone pairs	-	10	0.75
2BASE-TL	One or more telephone pairs	-	2	2.7

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.

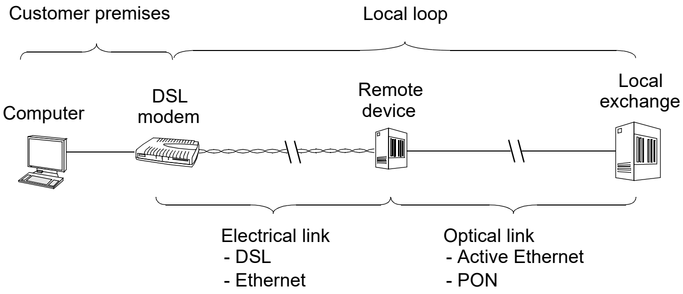


Figure 4.2 FTTx architecture for the local loop

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 Mbit/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 Mbit/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 Mbit/s of symmetrical bit rate, and 1000BASE-T 1 Gbit/s of symmetrical bit rate.The range is limited to 100 m for both.

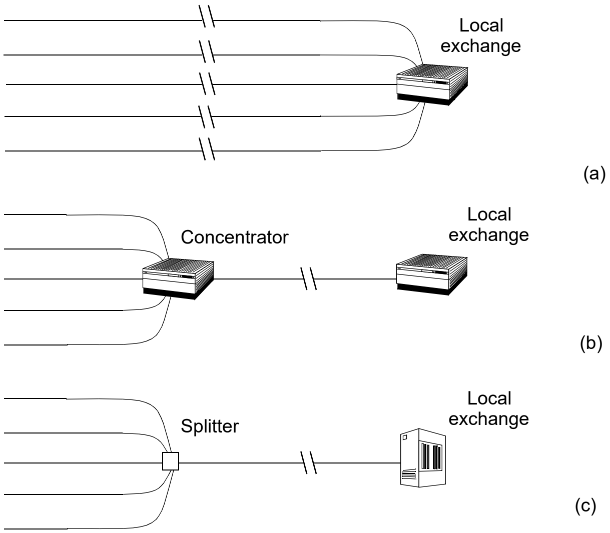


Figure 4.3 Optical fiber installation in the local loop: (a) The point-to-point topology needs a large amount of fiber. (b) With active Ethernet, less fiber is needed, because a switch can be placed close to the subscribers. (c) The PON solution replaces the switch with an inexpensive and passive optical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

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 exchange, 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.

4.2 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-Mbit/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 Mbit/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 analog 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.

Table 4.2
Copper Pair Categories

Category	*Bandwidth*	*Common Application*
1	-	Telephony, ISDN BRI
2	4 MHz	4 Mbit/s Token Ring
3	16 MHz	Telephony, 10BASE-T, 100BASE-T4 (four wires)
4	20 MHz	16 Mbit/s Token Ring
5	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mbit/s ATM
5e	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mbit/s ATM
6	250 MHz	1000BASE-T (four wires)

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 (see

Table 4.2). 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.

4.3 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 (see Paragraph 4.3.2). 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 existing EPONs provide 1 Gbit/s symmetrical capacity, typically to be shared by 16 subscribers. This means that the minimum guaranteed bandwidth in FTTH is around 60 Mbit/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 Mbit/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.

4.3.1 The Need of an Optical Access Network

Cooper access networks alone cannot meet the challenges of future broadband applications such as HDTV that may need up to 100 Mbit/s. Cooper 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:

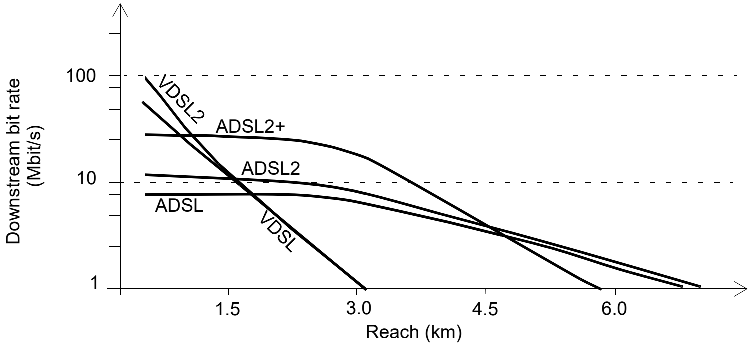


Figure 4.4 Approximate reach achieved with different DSL technologies

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 transmission, the more attenuation the signal will suffer.

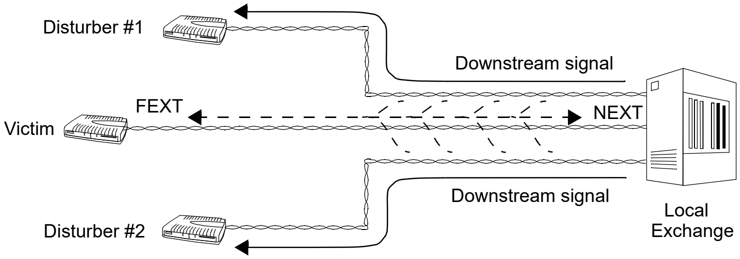


Figure 4.5 Crosstalk between copper pairs. Signals from disturbing lines are coupled to the victim line, damaging communication.

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.

Crosstalk control has special relevance. After 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.

4.3.2 Ethernet PON

The Ethernet PON or EPON is the IEEE alternative for PON. The first version of EPON was released in 2004, which makes this technology the latest PON version to appear at the time of writing. 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.

There are 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.

All the currently defined EPON interfaces are for transmission at 1 Gbit/s, but a 10-Gbit/s EPON standard is expected to be available soon. The EPON upstream and downstream are duplexed in a single SMF fiber. 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 DWDM 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 Gbit/s. This means that the signaling rate for 1 Gbit/s EPON is 1.25 GBd.

4.3.2.1 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.

Table 4.3
PON Technology Comparison

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

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 4.3).

4.3.2.2 Operation

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:

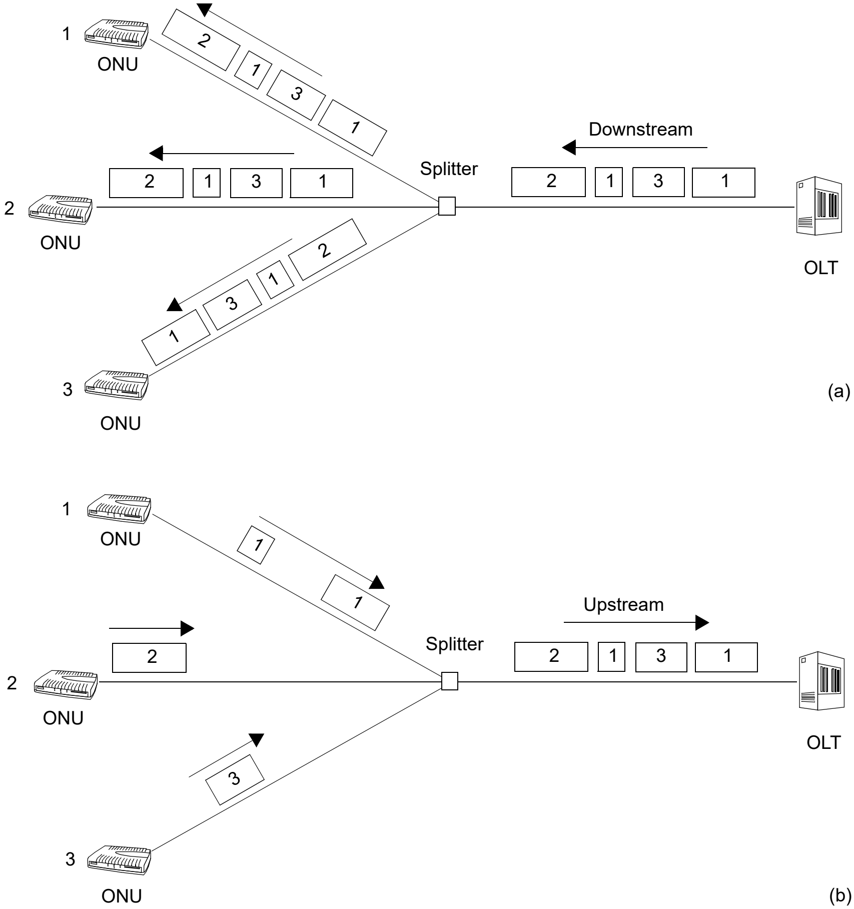


Figure 4.6 Transmission medium sharing a PON: (a) The downstream signal is broadcast to all the ONUs. (b) The upstream signal is point-to point. The section between the splitter and the OLT is shared between all the ONUs.

- 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 ONT 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).

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.

4.3.2.3 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:

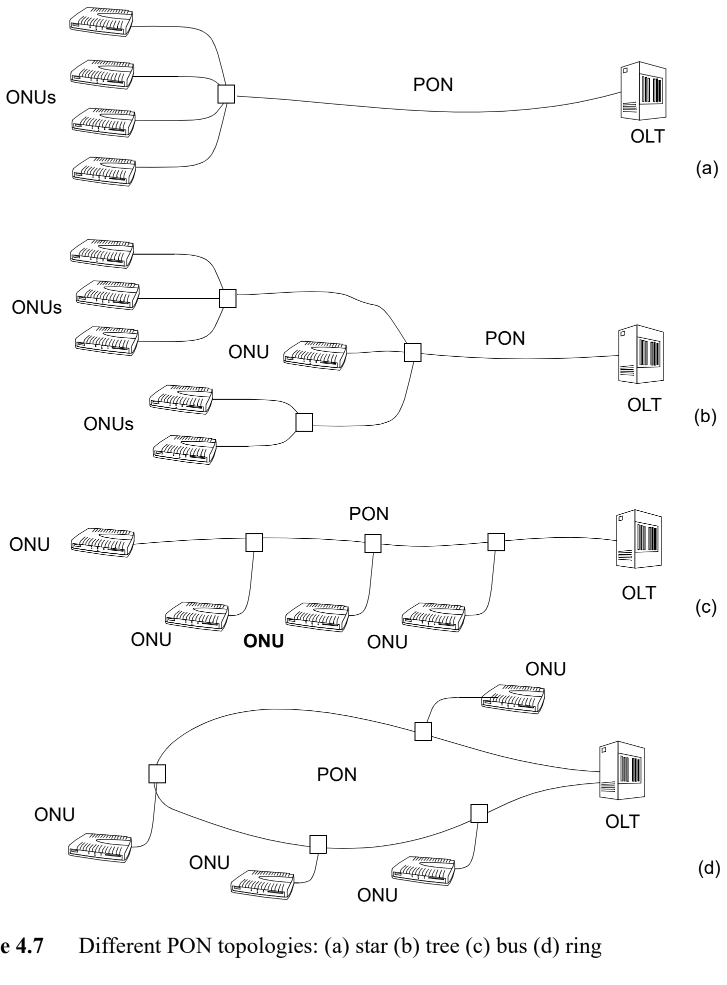


Figure 4.7 Different PON topologies: (a) star (b) tree (c) bus (d) ring

- 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 analog. Transparency eases migration to new technologies without the need to replace network elements. For example, migration 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 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.

4.3.2.4 EPON Particularities

The main goal of the 1000BASE-PX 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.

4.3.2.5 The Multipoint Control Protocol

The *Multi-Point Control Protocol* or MPCP is a signaling protocol for EPON, and it's 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 REGUISTER_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.

4.3.2.6 Logical Link Identifiers

Logical Link Identifiers or LLIDs are physical layer link identifiers defined to enable 802.1D bridging over an EPON (see Figure 4.8). 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 4.9). 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-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

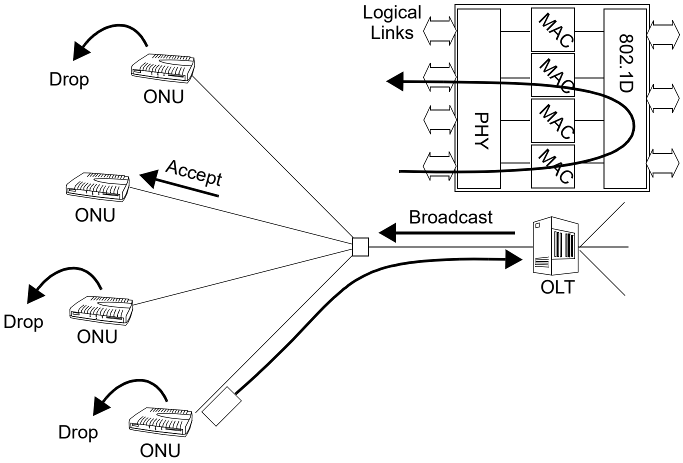


Figure 4.8 ONU-to-ONU bridging would not be possible without LLIDs. ONU-to-OLT associations defined by the LLIDs can be considered as point-to-point logical links. An 802.1D bridge can then perform learning and forwarding operations on the logical links.

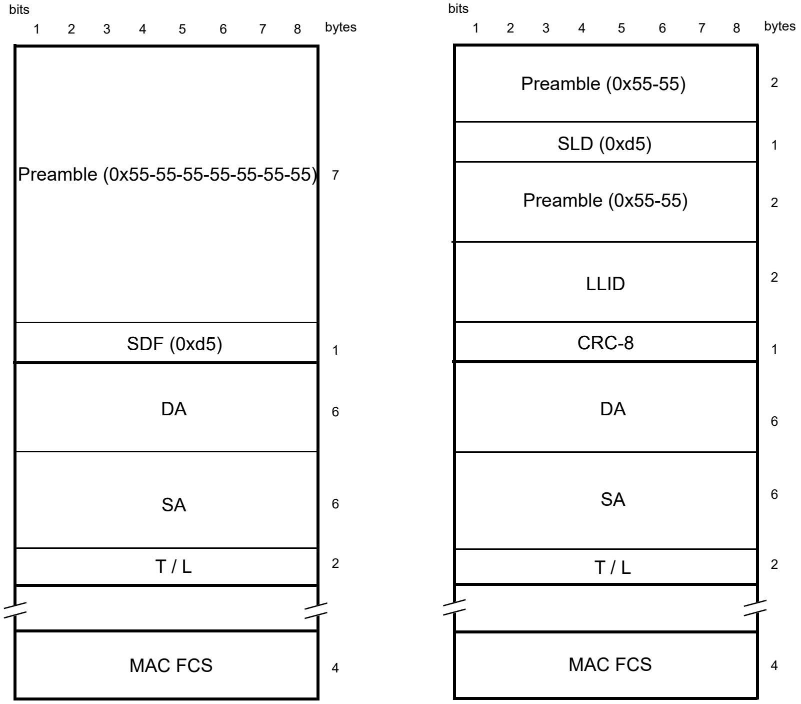
support a single LLID. They mark outgoing frames with the LLID assigned to them, and they accept frames marked with this LLID. The OLTs are more complex: they need one LLID per connected ONU.

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 SBC channel are accepted by all the ONUs.

4.4 Ethernet OAM

A major improvement provided by the IEEE 802.3ah is the definition of link *Operation, Administration and Maintenance* (OAM) services. The link OAM enables access network operators to monitor and troubleshoot the Ethernet link between the customer and network operator equipment. This new type of OAM complements OAM signalling at the service level defined in IEEE 802.1ag. The difference between IEEE 802.3ah and IEEE 802.1ag is that while the former works at the link level, the latter has been designed for OAM end-to-end signaling.

The functions of the link OAM protocol can be summarized as follows:



IEEE 802.3 Ethernet MAC Frame

IEEE 802.3 Ethernet Frame with LLID

Preamble: Synchronization pattern
SDF: Start Frame Delimiter
DA: Destination MAC Address
SA: Source MAC Address
T/L: Length / Type
MAC FCS: Frame Check Sequence

SLD: Start of LLID Delimiter
LLID: Logical Link IDentifier
CRC-8: Cyclic Redundancy Check parity

Figure 4.9 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.

- *Discovery* – Identifies the devices at each end of the link, along with their OAM capabilities.
- *Link Monitoring* – Detects and indicates link faults, providing statistics on the registered errors.
- *Remote Failure Indication* – Reports a failure condition detected by the remote peer of a given switch, such as loss of signal in one direction of the link, an unrecoverable error, etc.
- *Remote Loopback* – Puts the remote peer of a given switch in loopback mode. When a switch is operating in loopback mode, it returns all the traffic it re-

ceives back to the origin. The remote loopback mode is very useful for testing purposes.

Selected Bibliography

[1] Sargento S., Valadas R., Gonçalves J., Sousa H., "IP-Based Access Networks for Broadband Multimedia Services," *IEEE Communications Magazine*, February 2003, pp. 146-154.

[2] Kerpez K., "DSL Spectrum Management Standard," *IEEE Communications Magazine*, November 2002, pp. 116-123.

[3] Kerpez K., Waring D., Galli S., Dixon J., Madon P., "Advanced DSL Management", IEEE Communications Magazine, September 2003, pp. 116-123.

[4] Kramer G., Pesavento G., "Ethernet Passive Optical Network (EPON): Building a Next-Generation Optical Access Network," IEEE Communications Magazine, February 2002, pp. 66-73.

[5] Kramer G., Mukherjee B., Pesavento G., "IPACT: A Dynamic Protocol for an Ethernet PON (EPON)," IEEE Communications Magazine, February 2002, pp. 74-80.

[6] Effenberger F., Ichibangase H., Yamashita H., "Advances in Broadband Passive Optical Networking Technology," IEEE Communications Magazine, December 2001, pp. 118-124.

[7] 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.

[8] 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.

[9] Pesavento G., "Ethernet Passive Optical Network (EPON) architecture for broadband access," Optical Networks Magazine, January/February 2003.

[10] Eriksson P., Odenhammar B, "VDSL2: Next important broadband technology," Ericsson Review, No. 1, 2006, pp. 36-47.

[11] ITU-T Recommendation G.992.3, "Asymmetrical digital subscriber line transceivers 2 (ADSL2)", January 2005.

[12] ITU-T Recommendation G.993.1, "Very high speed digital subscriber line", June 2004.

[13] ITU-T Recommendation G.998.1, "ATM-based multi-pair bonding", January 2005.

[14] ITU-T Recommendation G.998.2, "Ethernet-based multi-pair bonding", January 2005.

Chapter 5

Carrier-Class 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 10 Gbit/s, and soon it will arrive to 100 Gbit/s.

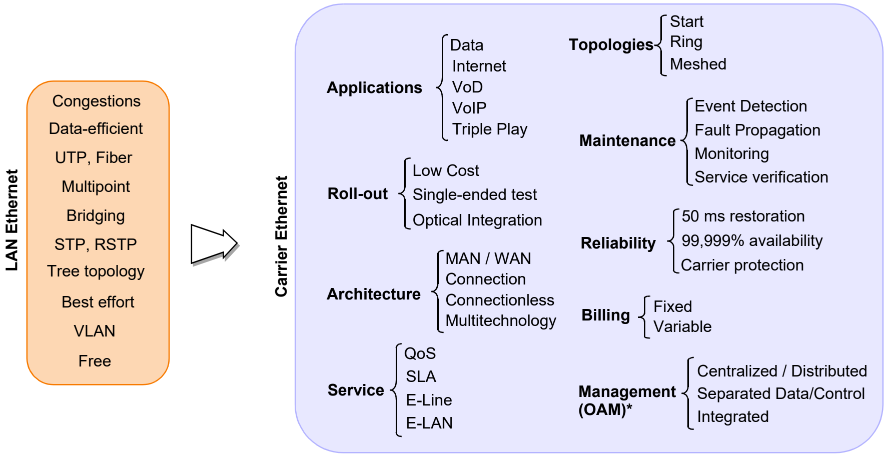


Figure 5.1 The path to Carrier-Class Ethernet.

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 5.1).

5.1 Ethernet Services

The *Metro Ethernet Forum* (MEF) has defined Ethernet services along with their specific requirements for Metropolitan environments. Currently, the MEF has defined two generic service types: Ethernet Line (E-Line) and Ethernet LAN (E-LAN).

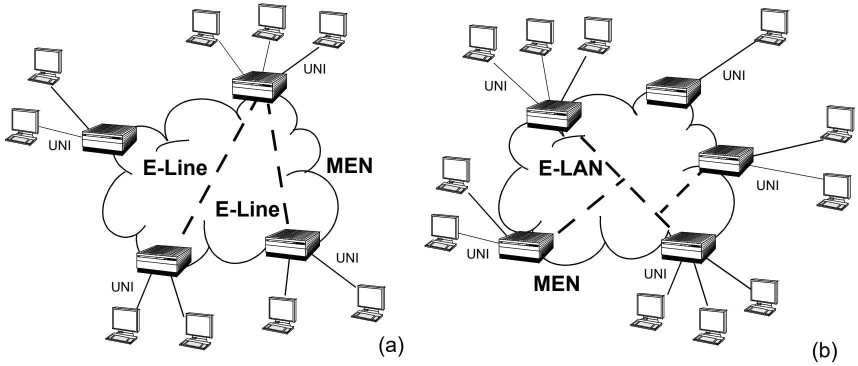


Figure 5.2 (a) The E-Line is understood as a point-to-point virtual circuit
(b) The E-LAN service is multipoint to multipoint

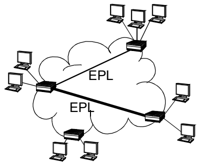
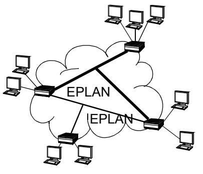
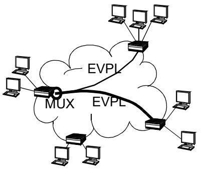
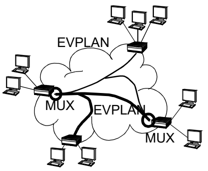
5.1.1 E-Line Service Type

The E-Line is a point-to-point service with attributes such as *Quality of Service* (QoS) parameters, VLAN tag support and transparency to layer-2 protocols (see Figure 5.2). 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 just a simple point-to-point ‘best-effort’ Ethernet connection, but it can also be a sophisticated TDM private line emulation. The MEF E-Lines are often marketed as *Ethernet Private Lines* (EPL) or *Ethernet Virtual Private Lines* (EVPL):

- The 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. 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.
- The 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. The EVPL is similar to the FR or ATM PVCs. The VLAN IDentifier (VID) for EVPLs is the equivalent of the FR *Data Link Connection Identifier* (DLCI) or the ATM *Virtual Circuit Identifier* (VCI) / *Virtual Path Identifier* (VPI).

Table 5.1
Ethernet Connectivity Services

		Generic Etherservice Type	
		E-Line - Point to point - Best-effort or guaranteed QoS - Optional multiplexing and bundling	E-LAN - Multipoint to multipoint - Best effort or guaranteed QoS - Optional multiplexing and bundling
EVC to UNI Relationship	Port-Based Service - No Service Multiplexing - Dedicated Bandwidth	Ethernet Private Line (EPL) 	Ethernet Private LAN (EPLAN) 
	VLAN-Based Service - Service Multiplexing - Shared Bandwidth	Ethernet Virtual Private Line (EVPL) 	Ethernet Virtual Private LAN (EVPLAN) 

5.1.2 E-LAN Service Type

The E-LAN service provides a multipoint-to-multipoint data connection (see Figure 5.2). UNIs can be connected or disconnected from the E-LAN dynamically.

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. The MEF E-LANs are often marketed as *Ethernet Private LAN* (EPLAN) or *Ethernet Virtual Private LAN* (EVPLAN). The former is a dedicated multipoint-to-multipoint service, and the latter is a shared service.

5.2 End-to-End Ethernet

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. Next-Generation SDH (NG SDH) nodes can transport Ethernet frames transparently (see Paragraph 5.2.3). Additionally, Ethernet can be transported by layer-2 networks, such as FR or ATM.

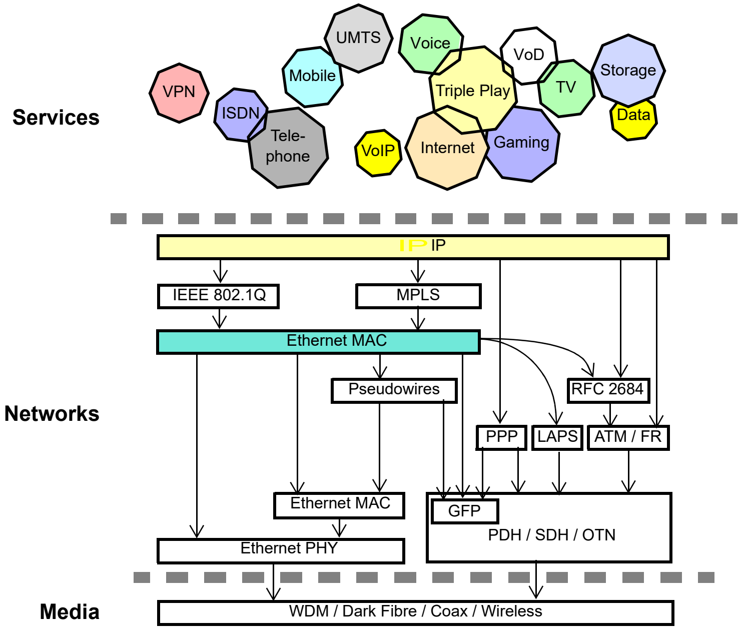


Figure 5.3 Transporting Ethernet and IP over packet- or circuit-switched infrastructures.

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.

5.2.1 Optical Ethernet

Ethernet can now be used in metropolitan networks due to the recent standardization of new long-range, high-bandwidth Ethernet interfaces. It can be said that Ethernet bandwidths and ranges are now 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. Despite of its simplicity, pure Ethernet solutions for MEN have big scalability problems. Furthermore, they suffer from insufficient QoS, OAM, and resilience mechanisms (see Paragraph 5.3).

5.2.2 Ethernet over WDM

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

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.

5.2.3 Ethernet over Next Generation SDH

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 operation.

NG-SDH unifies circuit and packet services under a unique architecture, providing Ethernet with a reliable infrastructure very rich in OAM functions (see Figure 5.3).

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.

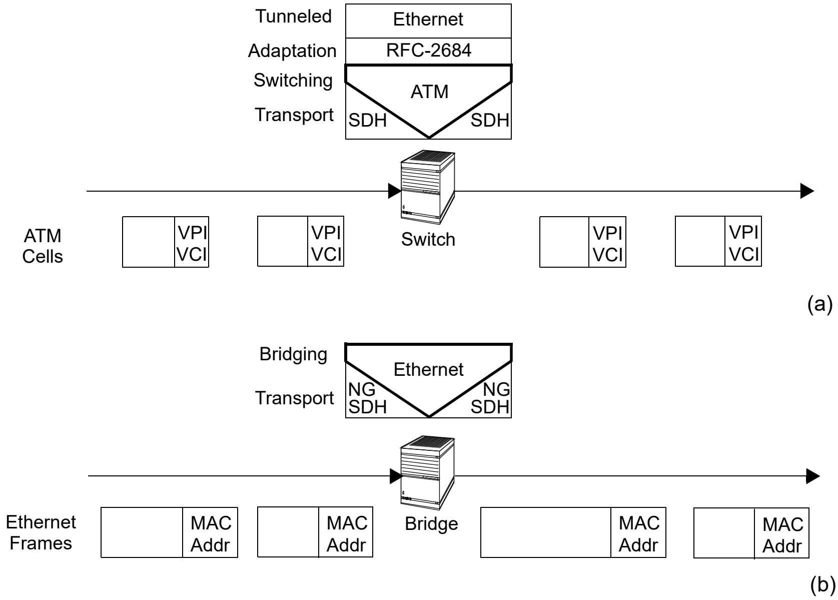


Figure 5.4 How NG-SDH raises the importance of Ethernet in the MAN / WAN. (a) Ethernet traffic is passively transported like any other user data. The ATM layer, specific for the WAN, is used for switching traffic. (b) The ATM layer disappears and the Ethernet layer becomes active. Traffic is now guided to its destination by means of Ethernet bridging.

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.

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.

5.3 Limitations of Bridged Networks

Metro Ethernet architectures, based only on native Ethernet switches and any combination of dark fiber, SDH and WDM, are like a big LAN – they have the same advantages and similar disadvantages. We know that in low traffic conditions and with a limited number of hosts, this network works very well. But as soon as the installation begins to grow, aspects like scalability, quality of service, topologies and protection tend to fall down. That is why Metropolitan Operators of Ethernet networks must rely on other technologies like MPLS to overpass most of these native Ethernet limitations.

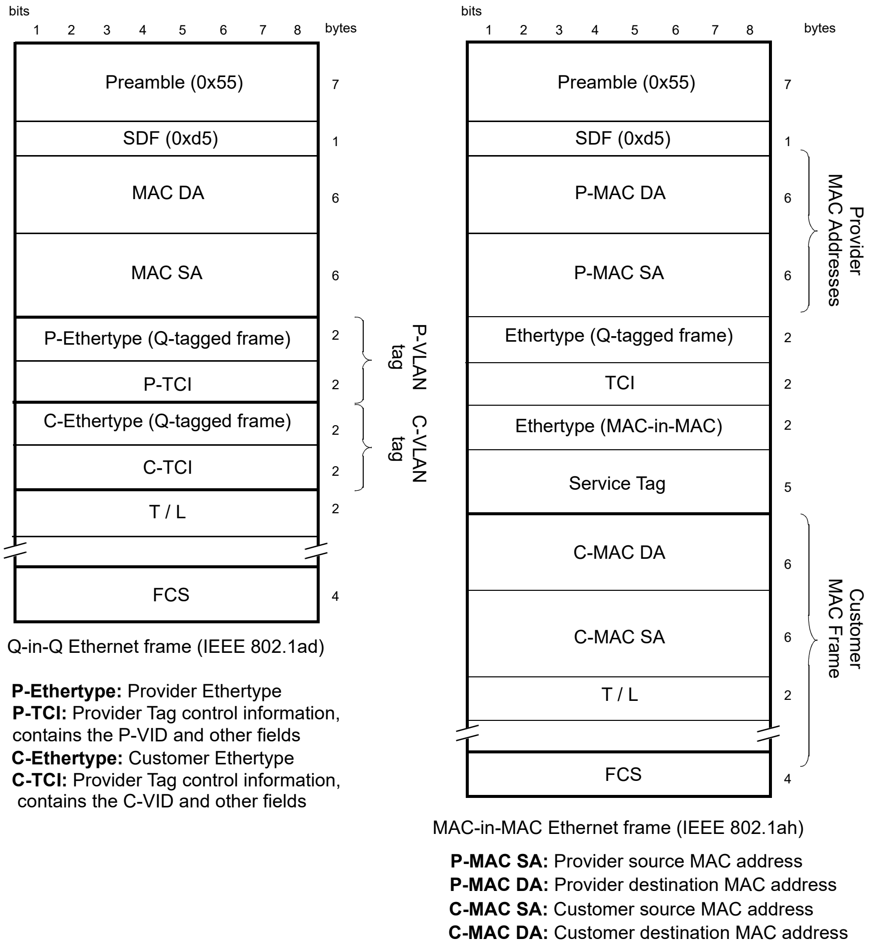
5.3.1 Scalability

Ethernet switches use promiscuous broadcasting to learn addresses constantly (IEEE 802.1D). When a request to forward a frame to an unknown address arrives, the switch has to flood the frame to all the ports, waiting for a response to know where the address is. This way is not very efficient, nor secure.

MAC addresses are not hierarchical, and the switching table does not scale well, slowing down the performance of switches when there is a large number of client hosts (this is known as the MAC switching table explosion problem). Furthermore, all the switches in the network have to constantly learn the MAC addresses of new stations connected to the network.

VLANs (IEEE 802.1Q) offer an easy solution to some of the problems mentioned before. A switch can be divided into smaller virtual switches, each of them belonging to a different VLAN. VLANs are used to split one big broadcast domain into several smaller domains, thus improving security and reducing broadcast traffic.

Another advantage of 802.1Q VLANs is the ability to offer QoS by means of the three 802.1p bits (VLAN CoS bits).



Q-in-Q Ethernet frame (IEEE 802.1ad)

P-Ethertype: Provider Ethertype
P-TCI: Provider Tag control information, contains the P-VID and other fields
C-Ethertype: Customer Ethertype
C-TCI: Provider Tag control information, contains the C-VID and other fields

MAC-in-MAC Ethernet frame (IEEE 802.1ah)

P-MAC SA: Provider source MAC address
P-MAC DA: Provider destination MAC address
C-MAC SA: Customer source MAC address
C-MAC DA: Customer destination MAC address

Figure 5.5 New frame formats for provider Ethernet bridges. These frames offer a more scalable Ethernet for carrier networks.

However, the number of VLAN identifiers is limited to 4,096. This limits the number of subscribers a service provider can have in the same network. Furthermore, subscribers may have their own VLAN structure, and it is desirable to support the customer and the provider VLAN structure simultaneously.

A solution is to stack VLAN tags (Q-in-Q solution) to obtain a larger number of VIDs as is specified in the IEEE 802.1ad standard for provider Ethernet bridges. *MAC Address Stacking* (MAS) is a more powerful solution, designed to deal with MAC switching table explosions (see Figure 5.5). The customer equipment is not supposed to understand the Q-in-Q and MAS frame formats. This means that the service provider VLAN labels and MAC addresses are added by the first hop and removed by the last hop in the carrier network.

5.3.2 Protection

The *Spanning Tree Protocol* (STP) and the *Rapid Spanning Tree Protocol* (RSTP) work well enough in LAN and also in data services. But they do not meet the challenge when it comes to the mass rollout of IP services, or the 50 ms restoration time paradigm of carrier-class services.

Other issue related with the STP is that it is inefficient for some network topologies. This is the case with rings which are used only partially; they are transformed into trees, and unused links are left just for protection.

5.3.3 Demarcation

Another problem is network demarcation, when the same technology is everywhere without a clear border between customer and provider installations. It is necessary to find a solution to questions such as:

- Can the typical LAN protocols, like STP, deployed by subscribers in their networks, influence the overall service-provider network operation?
- How will uncontrolled traffic generated by some subscribes affect the global MAN/WAN operation?
- How will the continuous connection and disconnection of stations in the subscriber network affect the service provider network?

These problems cannot be solved without important enhancements in the classic Ethernet technology, and the installation of demarcation devices needed to isolate and filter the traffic between networks.

5.3.4 Quality of Service

Up to eight levels of priority can be set up to VLAN-tagged Ethernet frames (IEEE 802.1Q/p) to manage different traffic classes – however, native Ethernet is not really a QoS-enabled technology. In fact, Ethernet is unable to provide hard QoS, because it doesn’t have resource management and traffic engineering tools.

5.4 Multi-Protocol Label Switching

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

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.

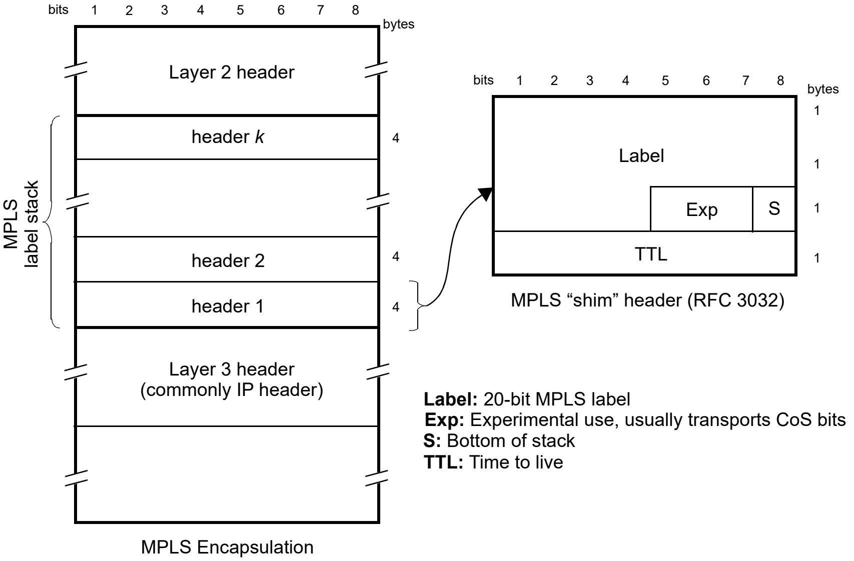
Any IP router can potentially improve its performance with MPLS regardless its particular physical interfaces, being SDH and Ethernet the most important examples. This enables network operators to deploy MPLS based services over any network infrastructure.

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.

5.4.1 Labels

When Ethernet is used as the transport infrastructure, it is necessary to add an extra “shim” header between the IEEE 802.3 MAC frames and the IP header to carry the MPLS label. This MPLS header is very short (32 bits), and it has the following fields (see Figure 5.6):

- *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 traffic differentiation 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.



Label: 20-bit MPLS label
Exp: Experimental use, usually transports CoS bits
S: Bottom of stack
TTL: Time to live

Figure 5.6 MPLS “shim” header format. The label is usually inserted between the layer-2 and layer-3 headers.

5.4.2 MPLS Forwarding Plane

Whenever a packet enters an MPLS domain, the ingress router, known as ingress Label Edge Router (LER), inserts a header that contains a label that will be used by the LSR to route packets to their destination. When the packet reaches the edge where the egress router is, the label is dropped and the packet is delivered to its destination (see Figure 5.7). Only input labels are used for forwarding the packets within the network, while encapsulated addresses like IP or MAC are completely ignored.

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 defines an LSP route:

5.4.3 Label Distribution

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.

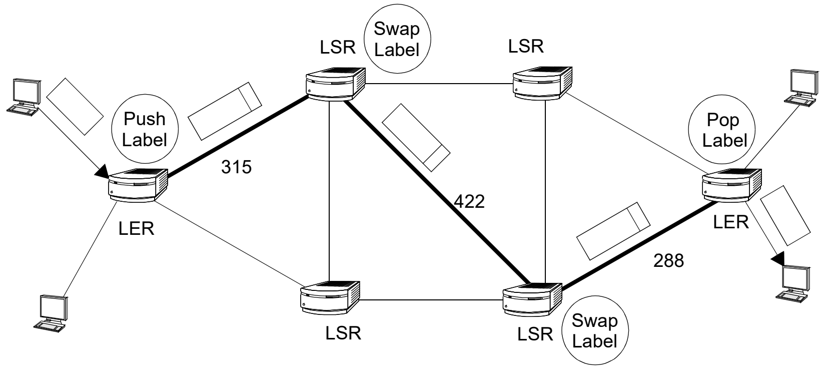


Figure 5.7 Label processing within an MPLS domain. A label is pushed by the ingressing LER, swapped by the intermediate LSR across the LSP, and popped by the egressing LER.

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.

5.4.4 Pseudowires

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:*

- 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.
- The *VC label* is used by the egress LER to identify client traffic and forward the frames to their destination. The way the traffic reaches end users is a decision taken by the ingress and egress nodes, and it does not involve the internal LSRs. The VC label is therefore pushed by the ingress LSR and popped by the egress LSR.

In the non-hierarchical one-label model, all the routers in the LSP participate in establishing an edge-to-edge session, and all are involved in routing decisions as well. A two-label model involves two types of LSPs. The tunnel LSP may have many hops, but the VC LSP has only two nodes, the ingress and egress LERs. VC LSPs can be interpreted as edge-to-edge sessions that are classified into groups and delivered across the MPLS network within Tunnel LSPs (see Figure 5.8). 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.

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

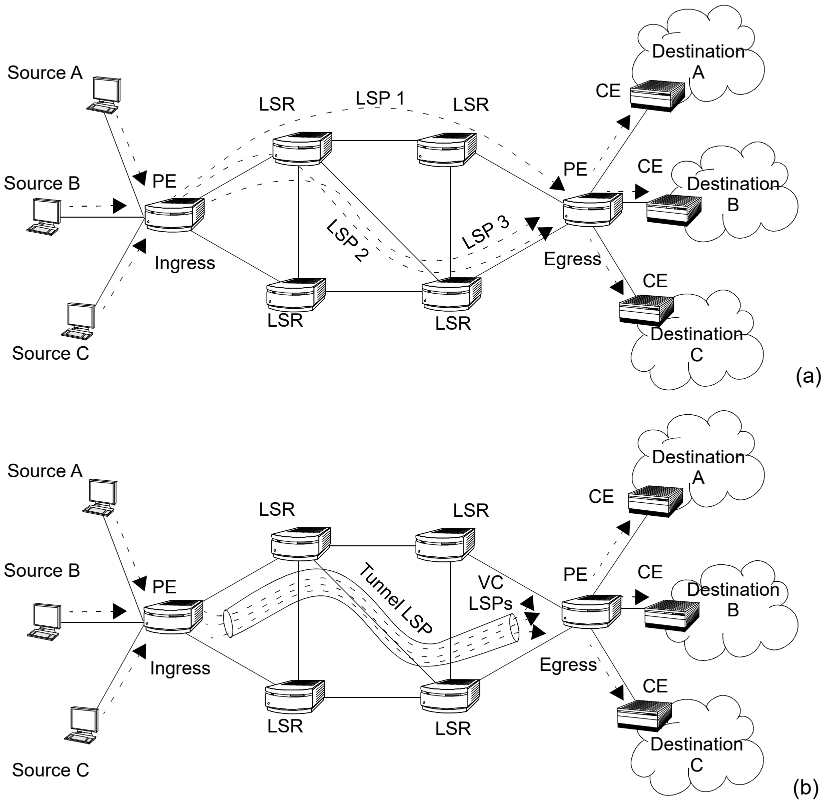


Figure 5.8 (a) One-label approach: the decision to establish routing and edge-to-edge sessions is shared between all the routers. (b) Two-label model: edge-to-edge sessions are tunneled, and internal LSRs are unaware of them.

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 provider 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.

When a new PE router is connected to the network, it must create tunnels to reach remote PE routers with the help of the RSVP-TE or the CR-LPD protocols. 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.

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 VIDs may belong to different customers, or if they belong to the same customer, they may require different treatment in the service provider network. MAC frames with service-delimiting VLAN tags may be forwarded to different pseudowires or mapped to different Exp values for custom QoS treatment.
- *Raw mode*: The MAC frames may contain VLAN tags, but they are not service-delimiting. This means that any VLAN tag is part of the customer VLAN structure and must be transparently passed through the network without processing.

5.4.4.1 Virtual Private LAN Service

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

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.

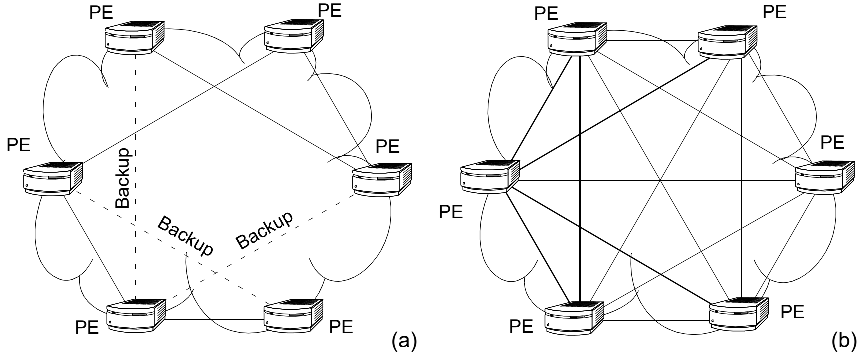


Figure 5.9 Pseudowire topologies in VPLS: (a) Partial mesh with STP. Some of the pseudowires are disabled to avoid loops. (b) Full mesh of pseudowires. The split-horizon rule is applied to avoid bridging loops.

As with physical wires, bridging loops may also occur in pseudowires. If fact, it is likely that this occurs if the pseudowire topology is not closely controlled, because pseudowires are no more than automatically established LDP sessions. A bridged network cannot work with loops. Fortunately, the STP or any of its variants can be used with pseudowires, as is done with physical wires to avoid them. However, there is another approach recommended by the standards. The most dangerous situation occurs when a PE router relays MAC frames from a pseudowire to a second pseudowire. To avoid pseudowire-to-pseudowire relaying, a direct pseudowire connection must be enabled between each PE router in the network. This implies a full-mesh pseudowire topology (see Figure 5.9). 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 any way 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 5.10). User S wants to send a MAC frame to user D across a shared network running VPLS.

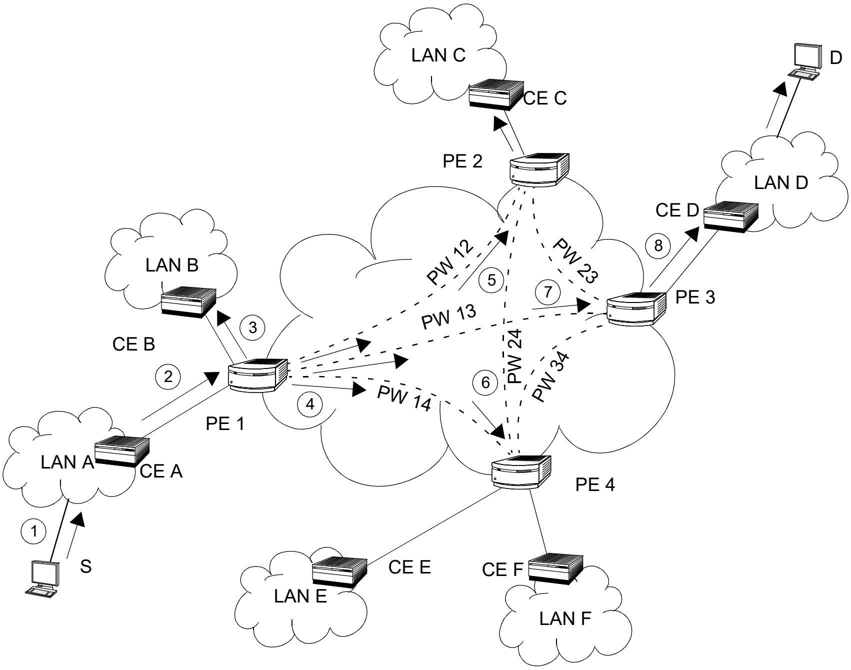


Figure 5.10 Flooding and learning in VPLS serves emulate a LAN broadcast domain

1. S sends the MAC frame towards D. LAN A is unable to find a local connection

to D and finally the frame reaches bridge CE 1 that connects LAN A to a service provider network.

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

The previous example deals with a single broadcast domain that appears as a single distributed LAN. But this may not be acceptable when providing services to many customers. Every customer will normally require its own broadcast domain. The natural way to solve this is by means of VLANs. Every subscriber is assigned a service-delimiting VID. 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-VID 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.

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 5.11). 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.

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.

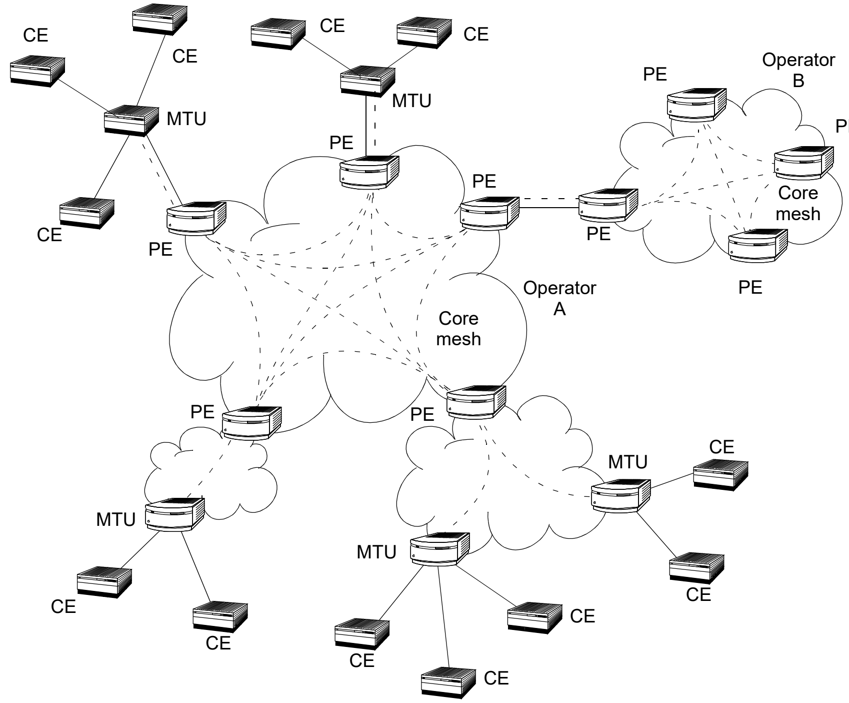


Figure 5.11 In HVPLS, the full mesh of pseudowires is replaced by a two-tier topology with full mesh only in the core and non-redundant point-to-point links in the access.

The 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.

Selected Bibliography

[1] Allan D., Bragg N., McGuire A., Reid A., "Ethernet as Carrier Transport Infrastructure," IEEE Communications Magazine, Feb 2006, pp. 134-140.

[2] Rosen E., Viswanathan A., Callon R., "Multiprotocol Label Switching architecture", IETF Request For Comments RFC 3031, January 2001.

[3] 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.

[4] Andersson L., Doolan P., Feldman N., Fredette A., Thomas B., "LDP Specification", IETF

Request For Comments RFC 3036, January 2001.

[5] 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.

[6] 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.

[7] Bryant S., Pate P., "Pseudo Wire Emulation Edge-to-Edge (PWE3) architecture", IETF Request For Comments RFC 3985, March 2005.

[8] Martini L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", IETF Request For Comments RFC 4446, April 2006.

[9] 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.

[10] 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.

[11] Lasserre M., Kompella V., "Virtual Private LAN Services Using LDP", IETF Internet Draft Document draft-ietf-l2vpn-vpls-ldp, June 2006.

[12] Awduche D. et al., "Overview and Principles of Internet Traffic Engineering," IETF Request For Comments RFC 3272, May 2002.

Chapter 6

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.

6.1 QoS Control Basics

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

Figure 6.1 Testing with Ether.Genius

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. Overprovisioning 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.

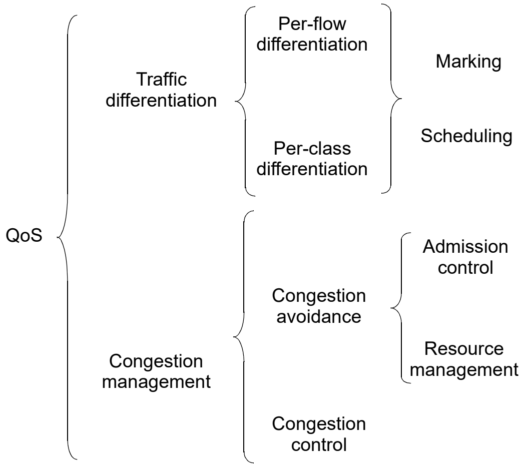


Figure 6.2 It is difficult to achieve good QoS features with one single mechanism. The best way is to mix many elements to get the desired result.

6.1.1 Traffic Differentiation

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

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

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

- *Differentiated services*. We can talk about differentiated services when a part of the traffic is treated 'better' than the rest. This way, it is possible to establish some QoS guarantees for the traffic. The QoS defined for differentiated services is also known as soft QoS.

- *Guaranteed services.* Guaranteed services take a step further. They are provided by reserving network resources only for chosen traffic flows. Guaranteed services are more QoS-reliable than differentiated services, but they make efficient bandwidth use difficult. The QoS for guaranteed services is also known as hard QoS.

6.1.2 Congestion Management

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

There are two ways to deal with congestion:

1. *Congestion control* is a set of mechanisms to deal with congestion once it has been detected in a switch, router or network. These mechanisms basically consist of discarding elements. The question is: which packets to discard first?
2. *Congestion avoidance* is a set of mechanisms to deal with congestion before it happens. There are two types of congestion avoidance techniques:

   - Admission control operates only at the provider network edge nodes, ensuring that the incoming traffic does not exceed the transmission resources of the network.

   - Resource management is used to allocate and free resources in the packet swictched 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.

- *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.
- *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".

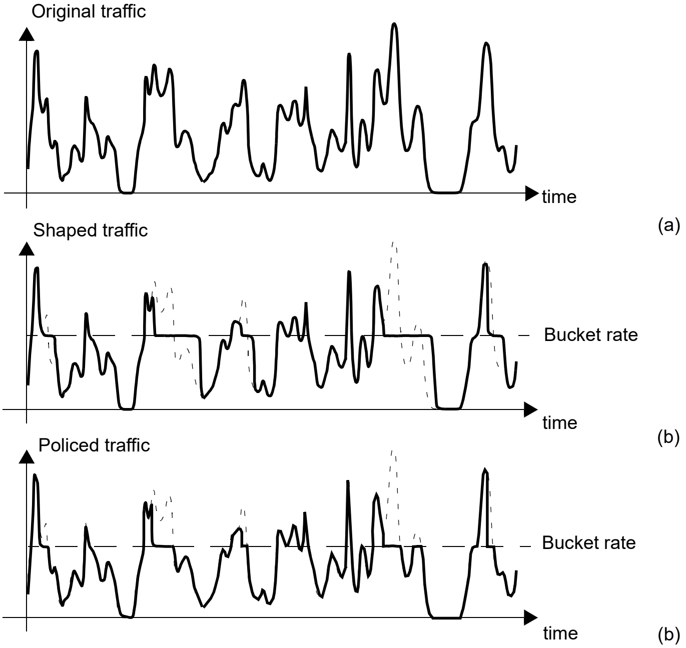


Figure 6.3 Shaping and policing of user traffic. (a) When traffic is shaped, no packets are dropped, but some of them may be delayed. (b) When traffic is policed, it is never delayed, but some packets may be dropped.

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).

6.2 THE FAILURE OF ATM

ATM was the first packet technology that was able to provide custom QoS to subscribers, but most players in the telecommunications market are now abandoning ATM and basing their most innovative QoS solutions on IP and Ethernet. ATM failed for several technical and commercial reasons:

- *ATM is more expensive than other technologies, such as Ethernet or IP*. While ATM has become a technology mainly for the backbone, Ethernet has evolved for years as a *Local Area Network* (LAN) technology, and IP has been used almost everywhere in data applications. The result is that the number of deployed ATM switches is much smaller than the number of deployed Ethernet and IP based equipment, and these technologies have become cheap and widely available.
- *Insufficient scalability*. ATM interfaces at 155 Mbit/s are common, but equipment for rates higher than this is much more uncommon and expensive.
- *Low efficiency and high complexity*. The advanced features of ATM rely on complex signalling protocols and a reduced "user data to overhead" ratio. This makes ATM a powerful but difficult and inefficient technology.

When the deployment of ATM begun, IP and Ethernet were already being used. In fact, the most important application for ATM is to transport IP traffic. Finally, it seems that IP and Ethernet will replace ATM. To make this possible, Ethernet has evolved to be a carrier-class technology that offers high availability, scalability and advanced *Operation, Administration and Maintenance* (OAM) functions.

6.3 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, which makes it possible to implement services such as *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.

6.3.1 Bandwidth Profiles

Once Ethernet access has been set up at 10/100/1000/10000 Mbit/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. The trTCM is obtained by chaining two simple token bucket policers. Tokens fill the main bucket until they reach the capacity given by the CBS parameter, at a rate given by the CIR parameter. The secondary bucket is filled with tokens with the EIR rate until they reach the capacity given by the EBS parameter.

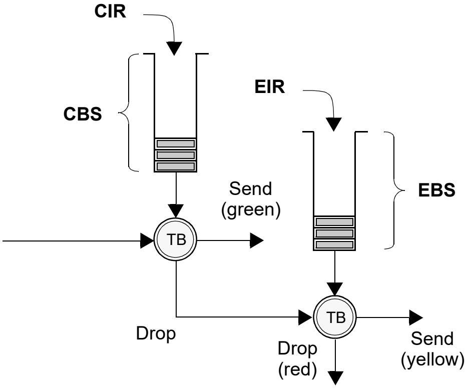


Figure 6.4 Two- rate three-color marker policer

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

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

6.3.2 Class of Service Labels

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

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

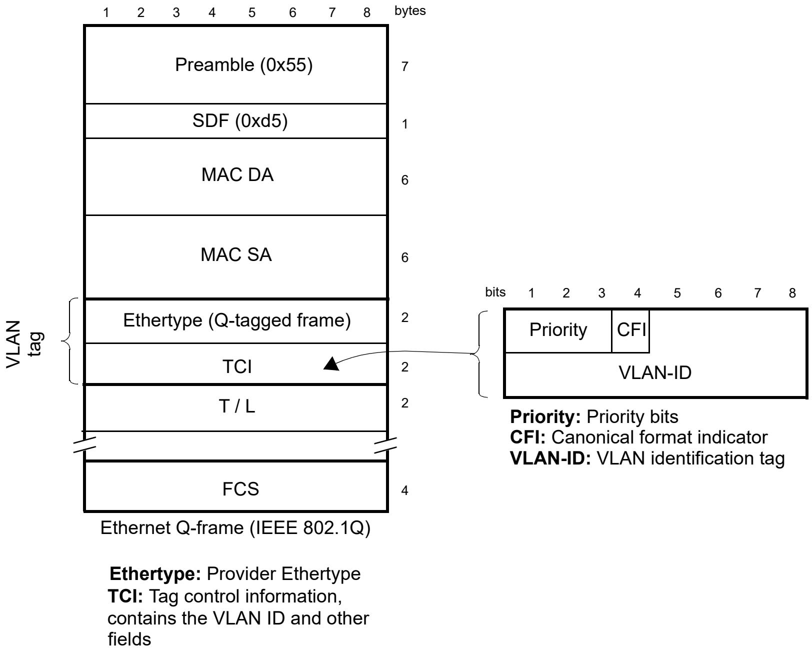

1 2 3 4 5 6 7 8 bytes
Preamble (0x55) 7
SDF (0xd5) 1
MAC DA 6
MAC SA 6
Ethertype (Q-tagged frame) 2
TCI 2
T / L 2
FCS 4
VLAN tag
Ethernet Q-frame (IEEE 802.1Q)
Ethertype: Provider Ethertype
TCI: Tag control information, contains the VLAN ID and other fields
bits 1 2 3 4 5 6 7 8
Priority
CFI
VLAN-ID
Priority: Priority bits
CFI: Canonical format indicator
VLAN-ID: VLAN identification tag

Figure 6.5 New frame formats for provider Ethernet bridges. These frames offer a more scalable Ethernet for carrier networks.

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

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

6.3.3 Resource Management

Those technologies that are based on VCs, for example ATM, can potentially provide the same level of service as any other circuit-switched network, while maintaining high flexibility thanks to the ability to perform end-to-end connections. 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 Transport* (PBT): PBT is a group of improvements that turn Ethernet into a connection-oriented technology by re-interpreting some fields of the MAC frame. With PBT, 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 PBT 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 PBT, switching tables are not autoconfigured by bridging, but set by a control plane separated from the forwarding plane.

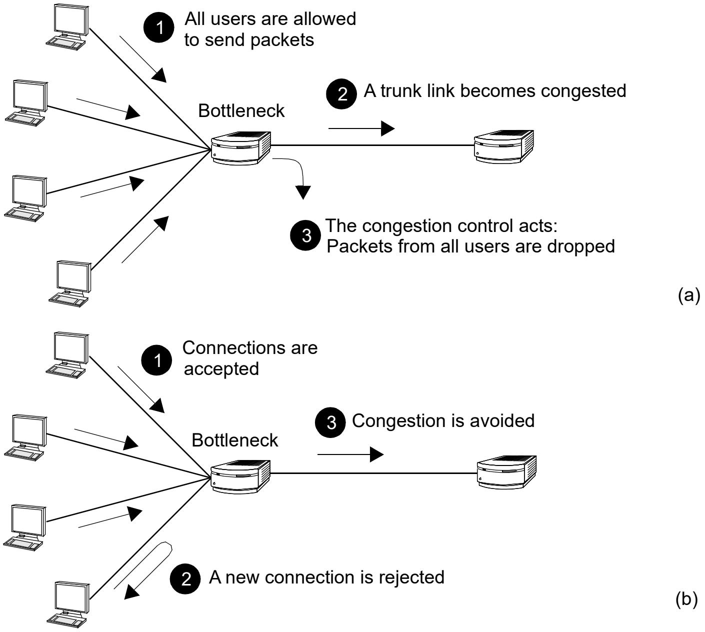

1 All users are allowed to send packets
Bottleneck
2 A trunk link becomes congested
3 The congestion control acts: Packets from all users are dropped
(a)
1 Connections are accepted
Bottleneck
3 Congestion is avoided
2 A new connection is rejected
(b)

Figure 6.6 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.

6.4 IP QoS

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.

6.4.1 Class of Service Labels

IP CoS labels are defined either by the ToS labels or the DS code points. 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 6.1).

Table 6.1
Meaning of ToS bits.

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

In addition to the four-bit field mentioned before, there is a three-bit precedence field that makes it possible to implement simple priority rules for IP datagrams (see Table 6.2).

Table 6.2
Precedence bits and their meaning

Binary value	*Meaning*
000	Routine
001	Priority
010	Intermediate
011	Flash

Table 6.2
Precedence bits and their meaning

Binary value	*Meaning*
100	Flash override
101	Critic / ECP
110	Internetwork control
111	Network control

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

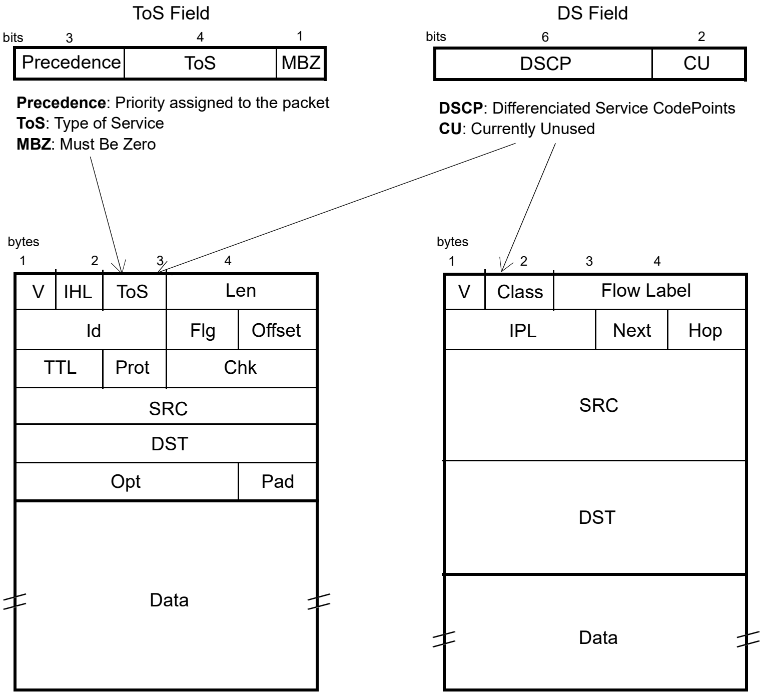
Although there are some recommendations, most of the PHB encoding by means of DS code points are configurable, and they can be freely chosen by the network administrator. The only constraint for this is the backwards compatibility with the old ToS encodings.

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.

6.5 End-to-End Performance Parameters

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



Precedence: Priority assigned to the packet
ToS: Type of Service
MBZ: Must Be Zero

DSCP: Differenciated Service CodePoints
CU: Currently Unused

IPv4 Datagram

V: The IP protocol version (4)
IHL: IP header length in 32 bit words
ToS: Enables QoS provision
Len: Total packet length
Id: Identifier to reassemble fragmented packets
Flg: Fragmentation flags
Offset: Fragmentation offset
TTL: Time to live
Prot: Protocol used in the data portion
Chk: Header ckecksum
SRC: Source IPv4 address
DST: Destination IPv4 address
Opt: Options, variable length
Pad: Padding, fills out the 32 bit words
Data: Data, variable length

IPv6 Datagram

V: The IP protocol version (6)
Class: Enables QoS provision
Flow Label: Identifies traffic flows
IPL: Payload length
Next: Identifies next header
Hop: Maximum hops before being discarded
SRC: Source IPv6 address
DST: Destination IPv6 address
Data: Data, variable length

Figure 6.7 IPv4 and IPv6 datagrams and the format of the ToS and DS fields, both related to QoS provisioning.

- Delay variation
- Loss
- Bandwidth

6.5.1 One-way Delay

The end-to-end *one-way delay* experienced by a packet when it crosses a path in a network is the time it takes to deliver the packet from source to destination. This delay is the sum of delays on each link and node crossed by the packet (see Figure 6.8).

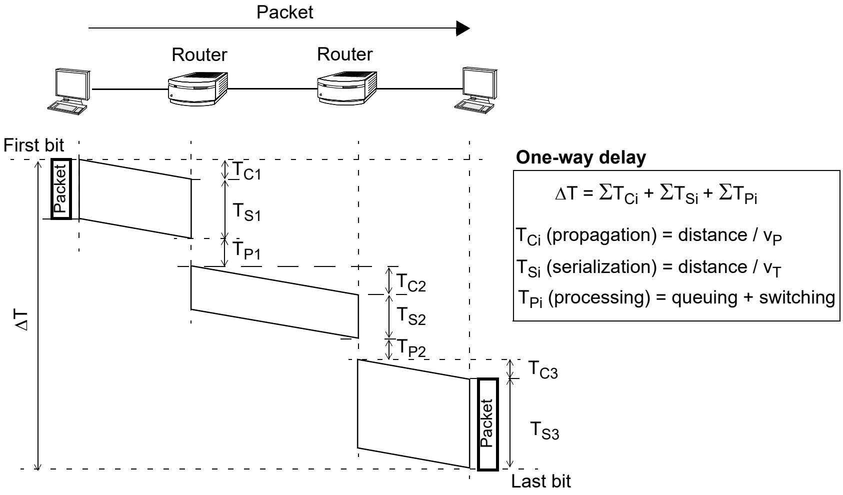

Packet
Router
Router
First bit
Packet
T_{C1}
T_{S1}
T_{P1}
T_{C2}
T_{S2}
T_{P2}
T_{C3}
T_{S3}
Packet
ΔT
Last bit
One-way delay
$\Delta T = \Sigma T_{Ci} + \Sigma T_{Si} + \Sigma T_{Pi}$
T_{Ci} (propagation) = distance / v_P
T_{Si} (serialization) = distance / v_T
T_{Pi} (processing) = queuing + switching

Figure 6.8 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.

6.5.2 One-way Delay Variation

The *one-way delay variation* of two consecutively transmitted packets is the one-way delay experienced by the last transmitted packet, minus the one-way delay of the first packet (see Figure 6.9). The one-way delay variation is sometimes referred to as *packet jitter*.

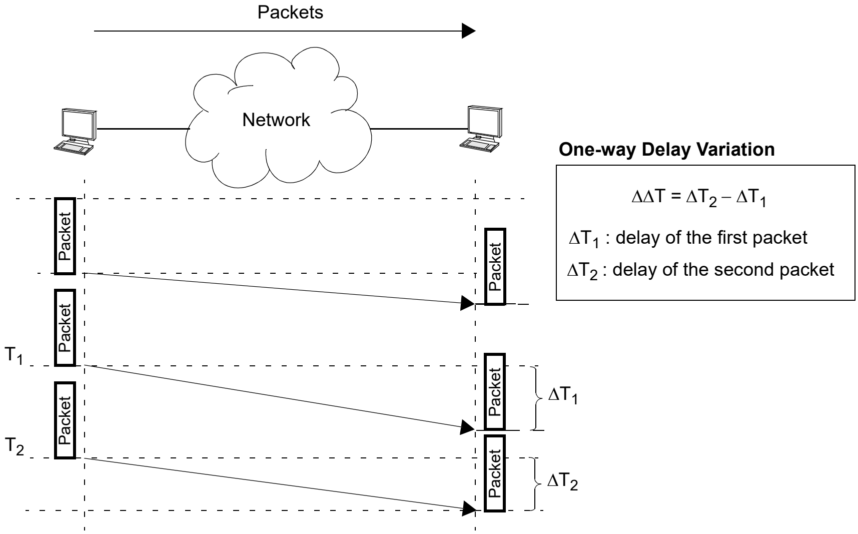

Packets
Network
One-way Delay Variation
$\Delta\Delta T = \Delta T_2 - \Delta T_1$
ΔT_1 : delay of the first packet
ΔT_2 : delay of the second packet
Packet
Packet
Packet
Packet
Packet
Packet
T_1
T_2
ΔT_1
ΔT_2

Figure 6.9 One-way delay variation: measurement and impact on data periodicity

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.

6.5.3 Packet Loss

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

Packet loss may occur when transmission errors are registered, but the main reason behind these events is network congestion. Intermediate nodes react to high traffic load conditions by dropping packets and thus generating packet loss. Congestion tends to group loss events, and this harms voice and video decoders optimized to work with uniformly distributed loss events. Loss distance and loss period are metrics that give information on the distribution of loss events.

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

6.5.4 Bandwidth

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

Selected Bibliography

[1] 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.

[2] Labrador M.A., Banerjee S., "Packet Dropping Policies for ATM and IP Networks," *IEEE Communications Surveys*, vol. 2, no. 3, third quarter 1999.

[3] Michaut F., Lepage F., "Application-Oriented Network Metrology: Metrics and Active Measurement Tools," *IEEE Communications Surveys*, vol. 7, no. 2, second quarter 2005.

[4] 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.

[5] Yang Chen, Chunming Qiao, Hamdi M., Tsang D. H. K., "Proportional Differentiation:A Scalable QoS Approach," *IEEE Communications Magazine*, June 2003, pp. 52-58.

[6] 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.

[7] Christin N., Liebeherr J., "A QoS Architecture for Quantitative Service Differentiation," IEEE Communications Magazine, June 2003, pp. 38-45.

[8] Almes et al., "A One-way Packet Loss Metric for IPPM," IETF Request For Comments RFC 2680, September 1999.

[9] Paxson V., Almes G., Mahdavi J., Mathis M., "Framework for IP Performance Metrics", IETF Request for Comments RFC 2330, May 1998.

[10] Matrawy A., Lambaradis I., "A Survey of Congestion Control Schemes for Multicast Video Applications," *IEEE Communications Surveys*, vol. 5, no. 2, fourth quarter 2003.

[11] Tryfonas C., Varma A., "MPEG-2 Transport over ATM Networks," *IEEE Communications Surveys*, vol. 2, no. 4, fourth quarter 1999.

[12] Vali D., Plakalis S., Kaloxylos A., "A Survey of Internet QoS Signaling," *IEEE Communications*

Surveys, vol. 6, no. 4, fourth quarter 2004.

[13] Marthy L., Edwards C., Hutchison D., “The Internet: A Global Telecommunications Solution?,” *IEEE Network Magazine*, July/August 2000, pp. 46-57.

[14] Xiao X., Ni L. M., “Internet QoS: A Big Picture,” *IEEE Network Magazine*, March/April 1999, pp. 8-18.

[15] White, P. P., “RSVP and Integrated Services in the Internet: A Tutorial,” *IEEE Communications Magazine*, May 1997, pp. 100-106.

[16] 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.

[17] Mase K., “Toward Scalable Admission Control for VoIP Networks,” *IEEE Communications Magazine*, July 2004, pp. 42-47.

[18] Welzl M., Franzens L., Mühlhäuser M., “Scalability and Quality of Service: A Trade-off?,” *IEEE Communications Magazine*, June 2003, pp. 32-36.

[19] 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.

[20] Braden R., Clark D., Shenker S., “Integrated Services in the Internet Architecture: an Overview,” IETF Request For Comments RFC 1633, June 1994.

[21] 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.

[22] Heinanen J., Baker F., Weiss W., Wrockawski J., “Assured Forwarding PHB Group”, IETF Request for Comments RFC 2597, June 1999.

[23] Davie B. et al., “An Expedited Forwarding PHB (Per-Hop Behavior)”, IETF Request For Comments RFC 3246, March 2002.

[24] 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.

[25] Wroclawsky J., “The use of RSVP with IETF Integrated Services”, IETF Request For Comments RFC 2210, September 1997.

[26] Shenker S., Wroclawski J., “General Characterization Parameters for Integrated Service Network Elements”, IETF Request For Comments RFC 2215, September 1997.

Chapter 7

Gigabit Ethernet Testing

New fast-growing services, such as e-commerce, *Voice over IP* (VoIP) applications, Internet radio and telephony, and *Video on Demand* (VoD) are together pushing so strongly that core networks are also migrating to *Next Generation SDH* (NG-SDH) to provide efficient and seamless Ethernet transport over longer distances.

7.1 GIGABIT ETHERNET MIGRATION

Today GbE and IP are very attractive to network providers, and they are ready to compete with many legacy networks based on PDH and xDSL. The adoption of *Multiprotocol Label Switching* (MPLS) technology can add connection-oriented features. For those services and fault-tolerant architectures designed to deliver mission-critical services, *Resilient Packet Ring* (RPR) can improve resilience to a level where it is close to SDH. In other words, adoption of GbE is under way. Integration of much of the infrastructure is in progress, and many datacom and telecom services are being moved from circuit-oriented and connection-oriented networks to packet-oriented networks.

Lower capital expenditure (capex) and easy integration into the enterprise are well known aspects of GbE that will speed up its widespread adoption. While this is true, it is not enough. We need to consider the operational side of designing, installing, and maintaining the new networks. This can account for around 50% of the service cost [6].

Some studies carried out by the *Metro Ethernet Forum* (MEF) suggest that operational cost models (opex) of GbE can be 22% cheaper than their equivalent TDM, ISDN, FRL and ATM opex in western markets. Despite this positive perspective, operators need to know which methodology should be used for bringing into service and maintenance, including the parameters to be tested, *Quality of Service* (QoS) levels to be provided, and the SLAs to be offered with parameters such as availability, bandwidth profile, jitter, frame loss, and delay.

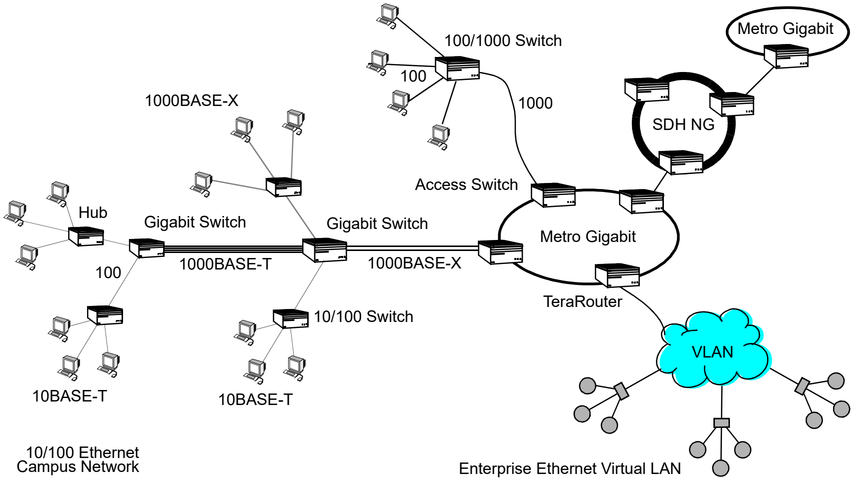


Figure 7.1 Ethernet topology and Gigabit Ethernet testing points.

Carrier-Class Services

The MEF has set a number of standards to provide carrier-class Ethernet services with quality levels equivalent to the existing data networks such as FRL and ATM. However, in spite of the expectations, they have not been widely accepted or implemented by carriers. A wide range of implementations exist, from complex layer-2 services to simple point-to-point connections [1]. This diversity makes it difficult to unify services and service extensions, and crossing more than one carrier becomes very complicated. For example, jumbo frame support, bandwidth profiles, link aggregations, and performance and remote testing facilities are limited.

Most of the Ethernet services lack a class-of-service system to prioritize traffic, making Ethernet just a best-effort service. Often, it is necessary to use a high-bandwidth backbone to guarantee a *Service Level Agreement* (SLA) that specifies that there are no delays in traffic throughput. The problem is that this policy leads to wasted resources, and it does not fully guarantee that the service would be achieved all the time.

Another limitation is the lack of standardized resilience and fast re-routing capabilities, often due to a combination of *Resilient Packet Ring* (RPR), classical SDH protection architectures and LCAS implemented in Next Generation SDH. This situation makes Ethernet not very well positioned, and often unsuitable for critical applications or services, such as real-time video, or critical applications that have strong surveillance requirements.

One more difficulty that holds up quicker GbE acceptance is the lack of demarcation points, which makes it difficult to separate the network belonging to the carrier from the part of the network that belongs to the customer. This is why *Operation and Maintenance* (OAM), monitoring and far-end testing are not easy in Ethernet. However, certain proprietary solutions have been developed that allow loop-back frames and pattern generation/detection, but these are just temporary solutions.

7.2 The Testing Challenge

The use of Ethernet beyond the LAN is yet another challenge and it is necessary to ensure that the provided service is deployed correctly and in a reliable way. This is especially true as these type of services are typically provided as an expensive lease contract with a service provider, and they can only rarely be totally controlled by the user. A range of test scenarios and test methods must be developed to allow the operator and end user to prove that a service is providing the required QoS, and to assist the investigation and resolution of faults or disputes.

There are certain keywords that seem to crop up again and again when talking about test and measurement applications: Research and Development (R&D), manufacturing, approval, acceptance, installation, bringing-into-service, maintenance, and monitoring.

In the area of R&D and the manufacture of *Network Elements* (NE), testers can be used as laboratory generators for carrying out measurements on prototypes of the NEs or parts of elements.

7.2.1 Approval and Acceptance

Many times Ethernet operators or users wish to approve a specific node or station, or compare competing pieces of equipment from different manufacturers. An equipment that has been purchased must also be tested to see that it is working correctly before it can be accepted by the buyer.

These tests are known as *approval and acceptance tests*, and they are described more thoroughly in see Chapter 8 "Approval and Acceptance Tests".

7.2.2 Installation

Many times, *installation tests* are defined together with acceptance tests, and for this reason they are often called installation/acceptance tests.

Certain NEs, with the transmission media to which they are connected, go together to make up clearly defined links, which must be tested when they are brought into operation for the first time. For this reason a series of *bringing-into-service tests* is defined.

Later, it will be necessary to set up and verify every single link that makes up the network; but now is the time to connect the node to the physical media. Once nodes are connected, this is the setting up or turning up stage, and it often requires configuration of IP, address, routing algorithms and VLAN definitions. A *continuity test* must be performed on the three different layers - physical, MAC and IP, to make sure that traffic can flow through the network.

7.2.3 Maintenance

Once the GbE network has been installed and brought into service, it must be maintained to guarantee its operation for the users. This involves making repairs when faults occur, as well as performing routine checks on the network when everything is working correctly. For this reason, *maintenance* tasks are performed that also require testers to locate faults and make sure that parts of the network are working properly.

Once all the links have been brought into service, the applications must be downloaded and executed. End-to-end verification is also needed here.

7.2.4 Monitoring

One final application is *monitoring*, which is normally done using testers connected to elements. The tester monitors the network over a long period of time to check that the network is working properly. These testers can usually be operated remotely. Finally, once the whole installation is ready, monitoring is needed to troubleshoot any problem that may occur, and finally to analyze the traffic statistic to get the best possible performance from the network.

7.3 Hands-On in Testing: Connection Modes

One consequence of the difficulties mentioned before is that more accurate testing is required now that Ethernet installations go further than the LAN.

A connection mode is basically how the tester is connected to a GbE interface. However, it is more than this, because it often determines the type of tests than can be performed.

7.3.1 Termination Mode

In *termination mode* the tester emulates an active station to stress part of the network with traffic, for example a physical medium, a device, a link; or to stress the whole network. The tester activates the transmitter and receiver of one interface (see Figure 7.2). Both ports of the tester, Rx and Tx, are activated for full emulation.

Most tests are available in this mode, for example BERT, G.821, G.826, traffic statistics, alarms, performance, and continuity tests such as trace route or IP ping.

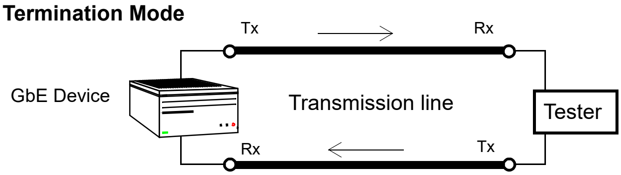


Figure 7.2 Termination connection mode for signal analysis on gigabit optical and electrical interfaces.

7.3.2 Loop-Back Mode

Loop-back mode is used to send the frames or packets back to the source station (see Figure 7.3). Looping back the information can be done at different layers:

- *Physical layer* – frames are looped unchanged including invalid frames
- *Frame layer* – frames are looped unchanged excluding invalid frames
- *MAC layers* – source and destination MAC addresses are swapped
- *IP layer* – source and destination IP and MAC addresses are swapped

Loop-back mode enables the tester to perform a full analysis of the traffic, because the same traffic that the tester generates is returned for further analysis.

This may be a necessary configuration with RFC 2544 or BERT measurements, because the procedure typically demands that the same tester is used for both test pattern generation and analysis.

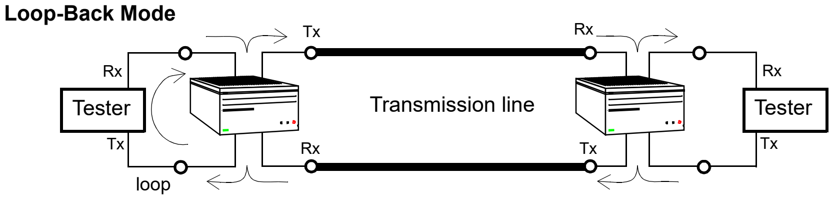


Figure 7.3 Loop-back mode for optical and electrical interfaces.

The far-end unit provides an intelligent loopback at Layer 3 (IP address layer). The tester must provide Address Resolution Protocol so that the network route could be discovered. Filters are useful so that broadcast traffic (or any IP address that is not relevant to the test procedure) does not need to undergo the loopback process. This will control the amount of extra traffic on the network and prevents possible overload.

7.3.3 Monitor Mode

Monitor mode is useful for obtaining traffic statistics and performance analysis. It is a basic mode for troubleshooting and maintenance of existing networks (see Figure 7.4). In this mode only the receive ports are active; the transmitter ports are disabled.

When an optical physical medium is being used, the connection can be made using optical splitters. If the media is UTP cabling, it is necessary to use a node for re-transmission.

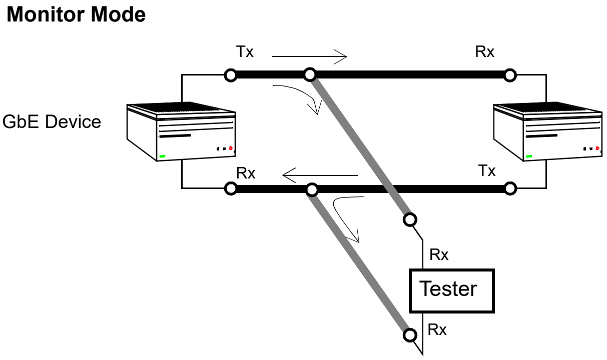


Figure 7.4 Monitor mode suitable only for optical interfaces.

7.3.4 Pass-Through Mode

The *pass-through mode* enables you to simultaneously monitor and insert traffic into an existing link. Two ports (transmit and receive) are needed so that traffic can be passed through the tester for monitoring and generating purposes (see Figure 7.5).

The most common tests to carry out in pass-through mode are traffic statistics, alarms, packet export, trace route, or IP ping.

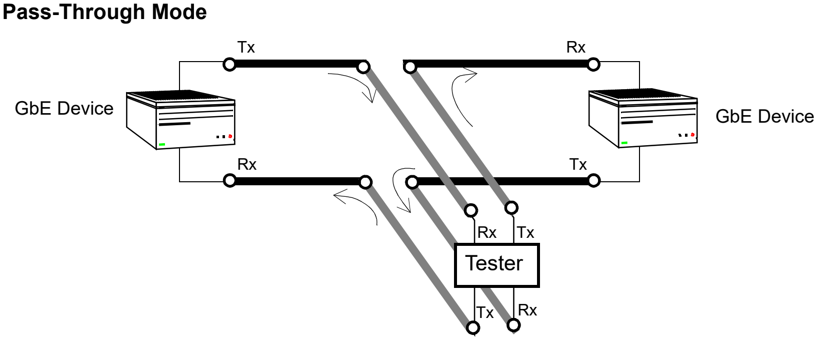


Figure 7.5 Pass-through mode, suitable for optical and electrical interfaces.

7.4 Frame-Oriented vs. Bit-Oriented

Ethernet is a *frame-oriented* technology, which means that information is put into frames and then sent to the destination.

In contrast, *bit-oriented* technologies, which are commonly used in networks like POTS, ISDN, GSM, PDH or SDH, constantly transmit a stream of bits in a predictable *Time-Division Multiplexing* (TDM) sequence. As a result of this, BERT end-to-end tests do not have the same meaning in Ethernet. In Gigabit Ethernet, stations drop entire frames as soon as a single bit causes a *Frame Check Sequence* (FCS) error (see Figure 1.16). This means that if an errored frame of 1 500 bytes is dropped because it had one bit error, then 12 000 bits are lost.

This does not mean that BER tests should not be used, as there are still many valuable tests that can be performed. In particular for transparent link-to-link Ethernet connections, or even in Metro Ethernet, a BER test is useful whenever the traffic does not cross nodes that can drop frames. IEEE 802.3 Annex 36A contains descriptions of the test patterns that are used for stressing transmit and receive devices to ensure best performance in relation to jitter, and according to each definition these may be used in a framed or unframed mode. In addition, operators familiar with

BERT may specify use of *Pseudo-Random Bit Sequences* (PRBS), though the applicability of these is debatable due to the frame format of Ethernet compared with traditional constant bit stream of TDM circuits.

Since Ethernet is a store-and-forward technology, link-to-link testing procedures are more valid than end-to-end procedures. It does not necessarily mean that tests such as from customer premises to service providers cannot be executed, but it does mean that Ethernet layer link-to-link tests are better. When it is necessary to verify the service, it is also better to use the higher layers, such as IP and Applications, in addition to the Ethernet layers.

7.5 ACCEPTANCE TEST

This is a benchmark test to verify that a product performs the required functions and meets specified operational parameters. Tests include traffic generation, and analysis of the data received from the *Device Under Test*, DUT. This approach tests the physical interface, and then hardware and software functionality.

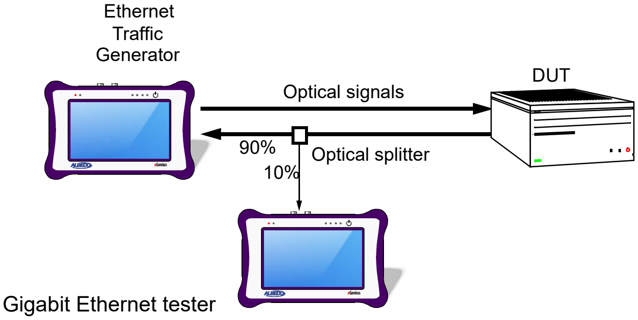


Figure 7.6 Setup for measuring optical signals.

The tester should be connected directly to the device to minimize the effects of the transmission medium. It includes PCS, PCA, Auto-negotiation, Flow control, Performance and Synchronization tests (see Chapter 8).

7.6 Installation Cabling Test

The demand for regular performance checks in mission-critical networks is a business opportunity for the cabling installer, with the chance to set a contract in place for regular re-testing. A proactive approach involving regular testing of the cabling system can prevent incidences of costly downtime in Gigabit Ethernet, where sensitivity to cable and connector related problems is higher [7].

Mission-Critical Networks

A proactive approach involving regular testing and monitoring of the cabling system can prevent costly downtime, which in today's technology-focused business world is more critical than ever. Today's networks, for so many organizations, are literally essential to the business, meaning the same level of reliability expected from telephone voice services and electrical utilities is required from the network.

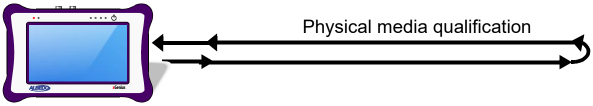


Figure 7.7 Cabling test has to verify that the fiber of the UTP cable is suitable to transport Gigabit Ethernet traffic. Analog, digital and TDR measurements are carried out.

Physical-layer problems are among the most troublesome and expensive causes of network downtime and slow operation. It is widely accepted that around 70% of network downtime can be attributed to cabling. This is not a result of faulty systems, instead network failures are effectively human failings in the form of poor installation and control. Companies making the transition from 10 Mbit/s to 100 Mbit/s to 1000 Mbit/s Ethernet are discovering that higher-speed transmission is more susceptible to cable and connector related problems.

The growth in mission-critical networks and the development of sensitive high-bandwidth systems has put regular testing high on the agenda for many end-users as a means to ensure that the cabling infrastructure they have paid for is consistently performing at optimum efficiency. Through keeping and understanding test records, end-users are able to identify deficiencies and remedy them quickly. In addition, an end-user will be in a much stronger position when dealing with the manufacturer, should system warranty issues arise. Not keeping accurate test records can create difficulties for effective troubleshooting.

The Role of the Installer

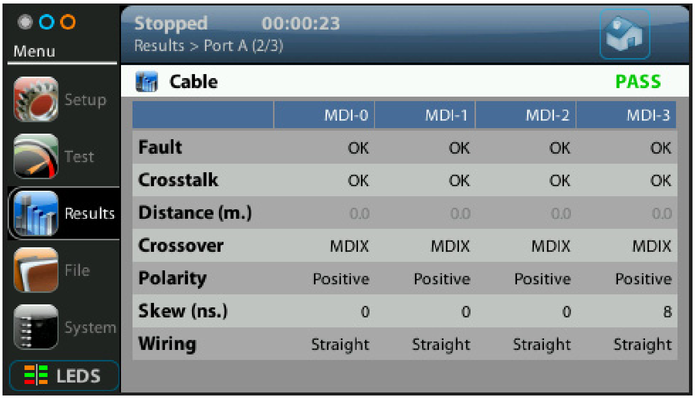
There is a growing trend among IT managers working with mission-critical networks to carry their own cable testers for regular performance checks following moves, adds or changes. The important test data for copper and fiber testing is explained in the following sections.

7.6.1 UTP Cable Test

Installation guidelines for Category 6 systems are the same as those for Category 5e and Category 5, but the margins in the pass/fail limits are significantly reduced. This means issues such as cable pulling tension, bend radius and cable ties must be given close attention to ensure that the performance of the system is not degraded.

Category 5 specified only four test parameters; wiremap, length, attenuation and crosstalk. Category 6 specifies twelve parameters to be measured over a wider frequency range and to tighter tolerances.

- *Near-End Crosstalk (NEXT)* and *Equal-Level Far-End Crosstalk (ELFEXT)* are measures of signal coupling from one cable pair to an adjacent pair at the near and far ends of the cable in one automatic test. ELFEXT is measured at the far end, by injecting a signal on to a cable pair at one end of the link, and measuring the induced signal on another pair at the other end. High levels of crosstalk can cause excessive retransmissions, data corruption, and other problems that slow down the network system.
- *Return Loss (RL)* measures the ratio of reflected to transmitted signal strength. Good-quality cable runs have little reflected signal, indicating good impedance matches in the links. Like attenuation, excessive return loss reduces signal strength at the receive end, and indicates a mismatched impedance at some point along the cable run. A value of 20 dB or greater generally indicates a good twisted pair cable. A value of 10 dB or less is bad, indicating large reflections.
- *Powersum NEXT and Powersum ELFEXT* are calculated values which measure the crosstalk effects of three transmitting pairs on the fourth, in the same cable sheath. This parameter is particularly important for Gigabit Ethernet and other applications that transmit and then receive on all four pairs.



Stopped 00:00:23
Results > Port A (2/3)

Cable PASS

	MDI-0	MDI-1	MDI-2	MDI-3
Fault	OK	OK	OK	OK
Crosstalk	OK	OK	OK	OK
Distance (m.)	0.0	0.0	0.0	0.0
Crossover	MDIX	MDIX	MDIX	MDIX
Polarity	Positive	Positive	Positive	Positive
Skew (ns.)	0	0	0	8
Wiring	Straight	Straight	Straight	Straight

Figure 7.8 Wire cable test with ALBEDO Ether.Giga

- *Attenuation to Crosstalk Ratio (ACR)* is a difference calculation between the attenuation and NEXT for each pair. This gives an indication of how ‘problem-free’ the cable pair will be for transmissions. A large difference reading is desirable, since it indicates a strong signal and little noise interference.
- *Powersum ACR* measurements are calculated by summing the NEXT between a selected pair and the other three pairs in the same cable sheath.
- *Delay* measures the period of time for a test signal applied to one end of the cable run to reach the other end.
- *Skew* indicates the difference between the measured time delay for that pair and the pair with the lowest value. This timing is critical for Gigabit Ethernet applications that transmit on all four pairs simultaneously, switch and then receive on all four pairs.

7.6.2 Fiber Optic Cabling Test

Traditionally, testing the fiber backbone of a structured cabling system just meant checking signal loss. Now, with many systems incorporating a much greater proportion of fiber and much higher demands on performance, more test functionality is required.

The latest testers allow presentation of measurements for length, delay and signal loss. Some go even further by allowing selection of network-specific certification tests, including 1000BASE-LX and 1000BASE-SX for Gigabit Ethernet. This allows an installer to produce data from which he can certify fiber for the common network types.

For troubleshooting, a full trace analysis, showing all the events (splices, breaks etc.) on the fiber link is ideal. The *Optical Time-Domain Reflectometer* (OTDR) uses backscattered light to determine the loss in the cable. As a result, this indirect test method can show significant deviations from the insertion loss tests compared with the results obtained with a tester emulating Gigabit Ethernet stations.

At present, OTDR instrumentation of the sort used routinely in long-haul public service networks is well beyond the budget of many installers. Modern high-performance cable testers offer an OTDR-like capability, enabling full trace data acquisition and analysis over LAN links. This allows documentation to be presented on every event on the fiber link at installation, so that trace comparisons can be made in the future, if problems arise.

7.7 INSTALLATION

Once the network nodes have been tested and are working, and the physical medium has been qualified for transporting Ethernet, it is time for installation. The operations involved are configuration of nodes, setting up links and interconnections. Once the segment, the subnetwork, or the whole network is interconnected it must be checked to verify the continuity, link-to-link or end-to-end, at different layers including physical, MAC and IP.

7.7.1 Configuration

It is always necessary to configure the network and nodes including protocols to use, IP addresses, networks, subnetworks, masks, routing tables, mappings, gateways, etc. The last step checks that the application is reachable.

7.7.2 Continuity Tests

Continuity tests must be performed at different layers. First at the physical layer to make sure that the signal arrives at the nodes and stations. Secondly at the MAC layer to check that the stations are properly connected, and that repeaters and switches are forwarding the frames. Finally at IP layers to check the configuration and the mask being used.

7.7.2.1 Physical Continuity

Physical continuity tests are run by generating traffic with a tester, or simply sending an electrical or optical signal that must be read at the far end of the link. The simplest way to do this is by looping the cable at the far end, but if it is convenient, you may want to connect a second tester to measure physical parameters as well.

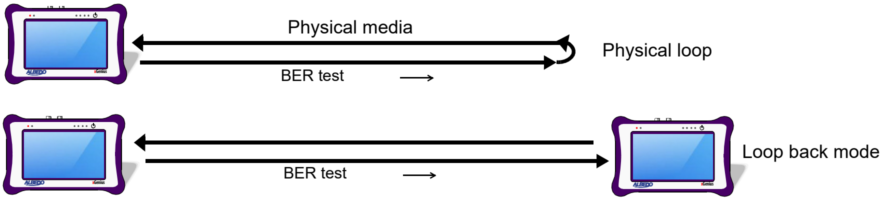


Figure 7.9 Physical layer continuity and BER test using G.821.

Performing a proper BER test at the physical layer is a good first test to measure the quality of the link. For example, the results of a G.821 test (see Paragraph 7.3.1) will give a good evaluation of the medium being used. A poor result may suggest that you should perform deeper testing, often using specific tests for each media type. This will help you to find the causes of poor performance (see Paragraph Figure 7.6).

7.7.2.2 MAC Continuity

After interconnecting the Ethernet nodes, station and cables, it is time to check the continuity of Layer 2. The set of tests to be performed must guarantee that MAC frames can be exchanged in the LAN, VLAN, or Metro network you are rolling out.

The topology of the network to be tested may be very different; from simple point-to-point topologies up to complex network and subnetworks, with different types of cables, hubs, routers, and switches. On Metro networks we may find links where Ethernet frames are transported by a core network like SDH or DWDM. In this case Ethernet mapping to SDH and media converters for DWDM transport must also be considered as a possible cause of discontinuity (see Figure 7.10).

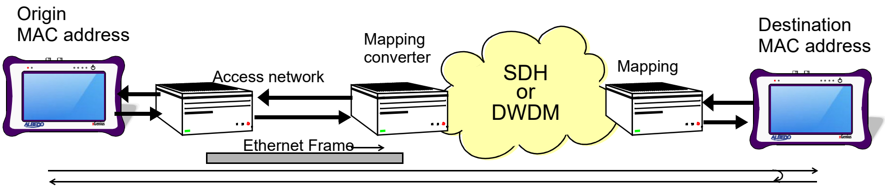


Figure 7.10 Continuity test at MAC layer. Media converters to long-haul optical signal care usec in DWDM. Classical mappings in Virtual Containers, GFP and LCAS are used to transport Ethernet over SDH.

7.7.2.3 IP Ping and Trace Route

IP ping tests are typically end-to-end continuity tests, but they can also be very useful for troubleshooting segments and subnetworks. An IP ping test checks if the address, masks, routing tables, and routing algorithms have been properly configured in stations, routers and servers. Ping is a basic Internet program that lets you verify that a particular IP address exists and can accept requests. Ping is used diagnostically to ensure that a host computer you are trying to reach is actually operating. If, for example, users cannot ping a host, they will be unable to use HTTP to browse or FTP to send files to that host. Ping can also be used with a host that is operating to see how long it takes to get a response back. Unfortunately ping is very limited; it can only verify that a particular IP resource is addressable from one other point on the network.

Trace route can be very useful to find paths and delays. It is a utility that records the route (the specific gateway computers at each hop) through the Internet between the point from where the test is being executed, and a specified IP destination. It also calculates and displays the time each hop took. Trace route is a useful tool both for understanding where problems are in the Internet network, and for getting a detailed sense of the Internet itself.

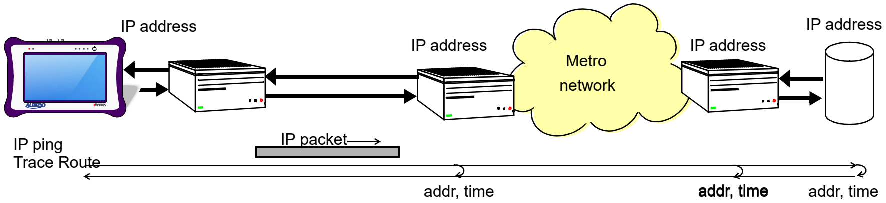


Figure 7.11 Continuity test at IP layer verified by means of IP ping and trace route.

7.8 RFC 2544 or Performance Test

Many Gigabit Ethernet test solutions have adopted the guidelines of the Internet Engineering Task Force recommendation RFC 2544, which was originally intended for verifying the performance of LAN devices at the MAC layer. The document describes a set of procedures for measuring the performance of Ethernet equipment, but it is also used for the overall networksee Chapter 9 "Performance Testing".

The procedures to measure performance are:

- *Throughput*: Defined as the number of bits transmitted per second through the DUT or the network without losing data or dropping frames.

- *Latency*: Measures the average time that elapses between sending traffic and receiving it. This measurement can be end-to-end or round trip delay.
- *Frame loss*: Measures the offered load as a percentage of the maximum line rate at which no frames are lost.
- *Burstability or back-to-back*: Defined as the maximum number of frames that can be sent in a fixed period of time without frames being dropped.
- *Recovery*: Characterizes how quickly the network recovers from an overload condition.
- *Reset*, or the time at which a network or station recovers from a reset.

7.9 Performance Measurements

The aim of *performance measurements* is to provide operators and users with information on the bit errors that occur on their digital links. For this, there is a series of ITU-T recommendations on the performance of digital links, including international connections.

Generally, transport networks can be monitored in an end-to-end or segmented way. In the case of Ethernet we have a difficulty mentioned earlier in this chapter (see Paragraph 7.4). Measurements on paths give information to the user about the overall *Quality of Service* (QoS). Measurements on lines and sections are performed during repair, installation, and maintenance, to make sure that previously established performance objectives are met. The main ITU-T recommendations concerning quality and performance measurements are G.821, G.826.

7.9.1 Physical Layer Test Using G.821

ITU-T Rec. G.821 was originally defined to measure errors on an international digital connection at *n* x 64 kbit/s. The scope of this recommendation is currently limited to simple digital links, however it can be used on Ethernet networks to measure the quality of the physical medium.

G.821 is always an *Out-of-Service Measurement* (OSM), and the test equipment generates a signal containing a known pattern, PRBS or user-defined, that can be analyzed at the far end by another tester (see Figure 7.1). The far end can be looped if attenuation or other physical parameters are not a limiting factor.

7.9.2 Performance Measurements with G.826

ITU-T Recommendation G.826 was defined to complement, improve, and substitute for G.821 (ISM and OSM measurements). It is based on the concept of *errored blocks*. The recommendation originally specified the events and parameters that define the performance of the carrier network paths, such as PDH, SDH/SONET, Frame Relay, or ATM, over metallic cables, fiber optics and wireless.

More recently, G.826 has been used by ALBEDO to create a G.826-like test to measure the performance of Gigabit Ethernet, matching blocks to Ethernet frames. This G.826-like measurement uses the Ethernet frames and the FCS (see Chapter 1 *"Ethernet and Gigabit Ethernet"*) to monitor real traffic in an ISM. It can also generate frames with FCS for the most common cases of OSM.

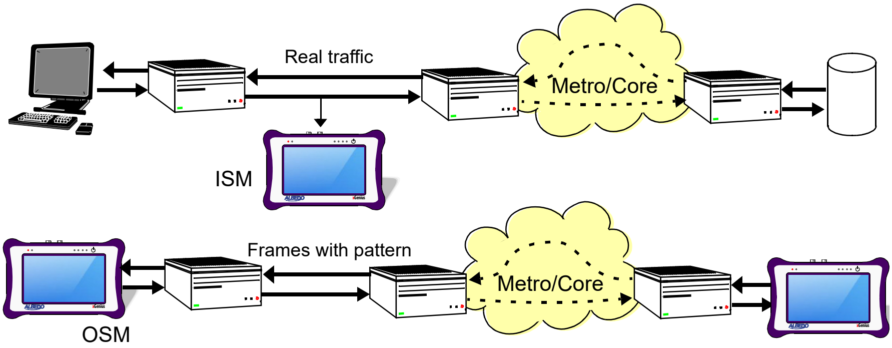


Figure 7.12 Performance measurement using xGenius.

7.9.2.1 Error Performance Events for Paths

The main differences between Ethernet and TDM networks were explained above (see Paragraph 7.4). The key point is that Ethernet nodes drop errored frames that do not arrive at the far end. This means that G.826 results in Ethernet are equivalent to PDH/SDH ones, but not in end-to-end test.

The frame-based events are:

- *Errored Frame* (EF) – A frame in which one or more bits are errored. Note that the frame will be dropped at the next Ethernet node, becoming a lost frame.
- *Frame Loss* (FL) – A frame lost or dropped.
- *Errored Second* (ES) – A period of a second with one or more errored or lost frames.

- *Severely Errored Second* (SES) – One-second period that contains $\geq$ 30% errored or lost frames.
- *Unavailable Seconds* (US) – Starts when 10 consecutive severely errored seconds have been detected. These 10 seconds are part of the US. Availability starts again after 10 consecutive non-SES have passed. If a bidirectional path is considered, a US occurs if one or both directions are unavailable.
- *Background Frame Error* (BFE) – An errored frame not occurring as a part of an SES or an US period.

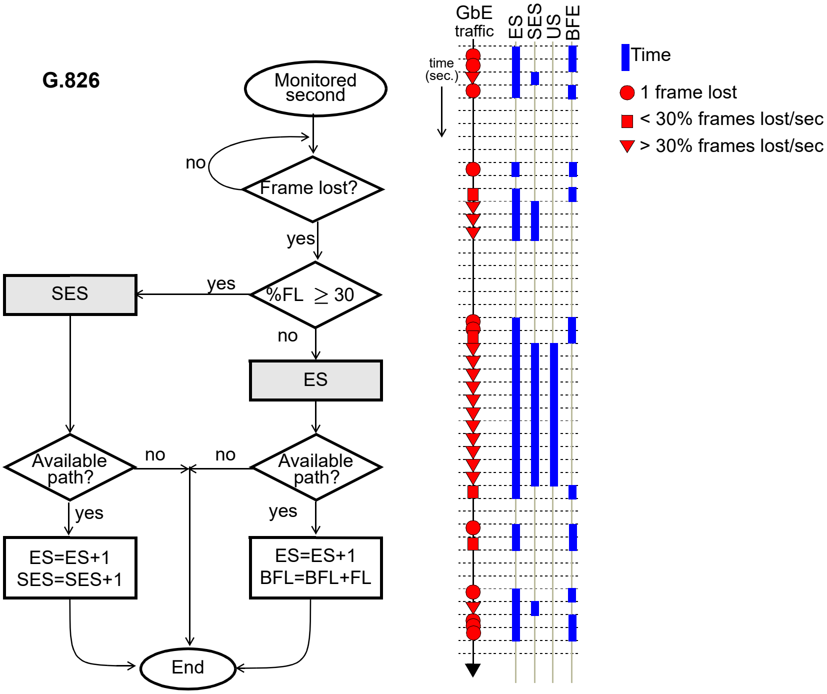


Figure 7.13 Service Level Agreement test, G.826-like, for defects, errored blocks, ES, SES, US BFE recognition.

The following is a list of error performance parameters, which should only be evaluated while the path is available:

- *Errored Second Ratio* (ESR) – The ratio of ES to total seconds in the available time during a fixed measurement interval.
- *Severely Errored Second Ratio* (SESR) – The ratio of SES to total seconds in available time during a fixed measurement interval.

- *Unavailable Seconds Ratio* (USR) – The ratio of US to *Available Seconds* (AS).

7.9.3 The Effect of Frame Loss

Frame loss is a very common event in Ethernet, because nodes drop frames if there is a mismatch in the FCS, or if any other serious defect is detected (see Paragraph 7.4). Let's now have a look at the consequences of this on performance.

If the network is transmitting isochronic information which is time-dependent, for example audio or video, then higher layers, such as UDP, do not guarantee delivery of all the packets. Furthermore, if real-time information is being streamed, it does not make any sense to retransmit lost packets. The slot time for that specific packet from a conversation or a movie would definitely be gone. So the main consequence is the *loss of quality* of the streamed information (see Figure 7.14).

When packet delivery must be assured, the higher-level protocols must guarantee that every single packet arrives at the receiver. Connection-oriented protocols, like TCP, have packet number IDs which allow the transmission of *Acknowledge Packets* (ACK) when the packet has arrived correctly, or for requesting retransmission (*Negative Acknowledgement*, NACK) if it does not. Transmitters do not assume that a packet has been received correctly until they receive the ACK, but if the packet does not arrive, a time-out indicates that the packet must be retransmitted. This strategy can have a devastating effect on the throughput if the algorithm is not properly adjusted.

The performance of TCP over WAN (the Internet) has been extensively modelled. A well-known paper by Matt Mathis [8] explains how TCP throughput has an upper limit which can be expressed independently of the bit rate:

$$Throughput \leq 0{,}7MSS/RTD\sqrt{FLR}$$

The expression says that TCP throughput is proportional to the *Maximum Segment Size* (MSS), which is the *Maximum Transmission Unit* (MTU) minus TCP/IP headers, and inversely proportional to *Round Trip Delay* (RTD) and *Packet Loss Ratio* (PLR).

Example: We have an Ethernet connection from London to Barcelona with RTD = 30 ms, and FLR = 0.09% (0.0009). If an MTU = 1500 bytes then MSS = 1460), TCP throughput will have an upper limit of about 9 Mbit/s, independently of the Ethernet bit rate! (see Figure 7.6) This is based on TCP's ability to detect and recover lost packets. Note that using the same network, it would be difficult to change any parameters except MSS or packet size.

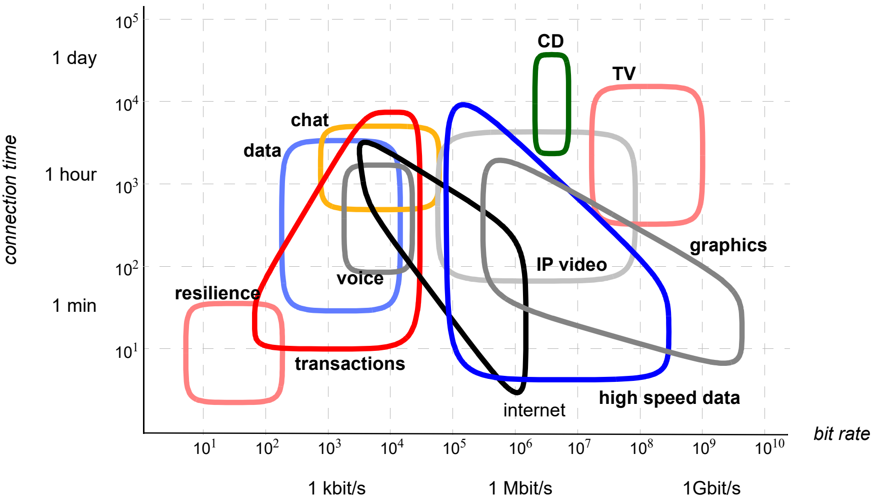


Figure 7.14 Bandwidth requirements for several applications.

In LAN and Campus networks, RTD and PL are both usually small enough, so that factors other than the above equation set the performance limit, for example raw available link bandwidths, packet forwarding speeds, host CPU limitations, etc.
In the WAN, however, RTD and FLR are often quite large, and something that the end systems cannot control.

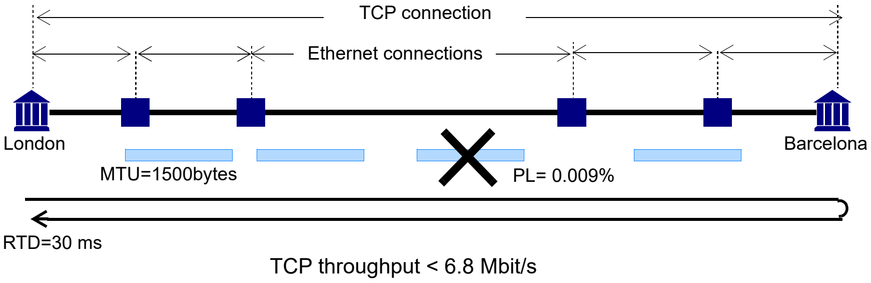


Figure 7.15 Sample of throughput limits at higher layers. TCP does not scales very well in Metro.

7.9.3.1 Performance and Frame Size

Let's say we want to achieve a throughput of 500 Mbit/s on the same Barcelona – London connection. To achieve this we have several options: reduce the latency, reduce the frame loss ratio, or increase the frame size.

a. *Reducing latency* should not be considered in a well designed network, except if there is a problem. In this case it would be unrealistic, because it would require a RTD = 0,5ms
b. *Reducing the FLR* is the second option, we would need a FLR $< 3 \cdot 10^{-7}$
c. *Increasing the frame size* to 9000 bytes.

Note that options a. and b. can be difficult and expensive to achieve, as they may require important changes in the infrastructure that are often impossible to implement.

Increasing frame sizes could be an alternative, but jumbo frames are not standard. Furthermore, this solution does not work in practice, because the cause of bit errors increases the FLR size as well. One of the conclusions is that TCP does not scale very well for high-speed network outside the LAN context.

7.10 MAINTENANCE

Once the network has been properly installed and applications are running, maintenance is needed, mainly to troubleshoot problems, but also to modify the topologies, or just update software or hardware. The same tools and similar procedures that were used for bringing into service can now be used.
Additionally, it will be necessary to analyze real traffic, packet statistics, and special maintenance information to find the causes of low performance, or a drop in service.

7.10.1 Out-of-Service Measurements

Once the Ethernet network is operational, this is a good time to measure the service level. Typically, Ethernet installers will stress the network in different ways, generating a number of traffic patterns and different statistical distributions. Measurements should be taken at significant points to get a complete picture of the network.

To set up the traffic generator, a set of parameters is used to define the characteristics of the traffic:

1. *Bit Rate* (BR), the line rate where the node or the station is connected, and where engineers feed the network with traffic

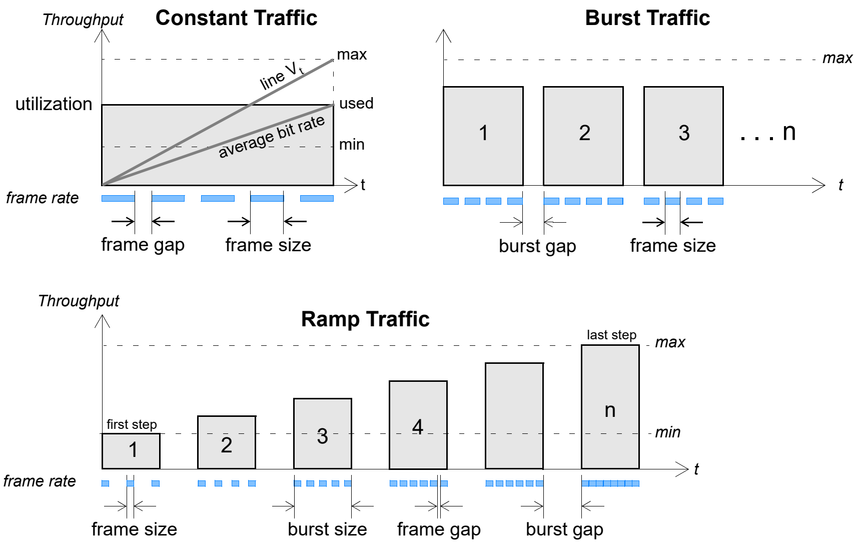


Figure 7.16 Traffic profiles for testing user bandwidth profiles.

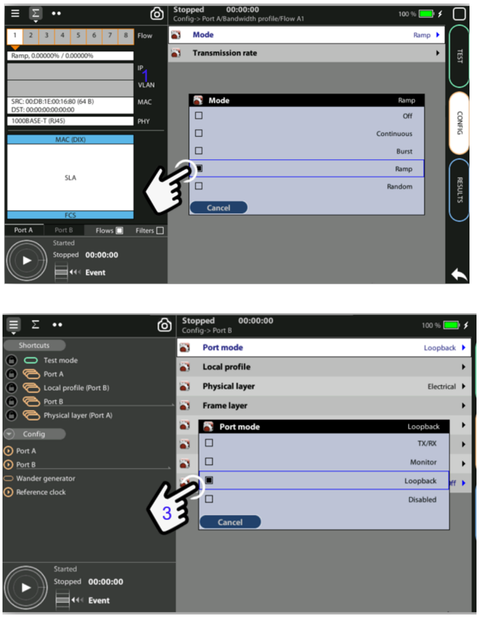
2. *Bandwidth Utilization* (BU), the percentage of line rate to be used. BU is useful to check the throughput offered by the carrier, which usually is not the bit rate of the line.
3. *Frame Size* (FS), this parameter is specified to make sure that the throughput is independent of the packet size. Small packets are more demanding than larger, because more are necessary to provide the same throughput.
4. *Traffic Profile*, the generation of frames can have a number of shapes including constant, bursty and ramp (see Figure 7.1).

7.10.1.1 Traffic Profiles

Constant rate is useful for testing. Bursty traffic is a sophisticated version of constant rate, and it can be used to discover problems with queue management. It is useful to measure network performance under constant load, and it can be an acceptable test for simple topologies with no traffic profile defined.

However, this is an unrealistic way to gauge performance, since real network traffic normally consists of bursts of frames. Frames within a burst are transmitted with the minimum inter-frame gap. The objective of the test is to determine the minimum interval between bursts which the node or the network can process with no frame loss. During each test, the number of frames in each burst is held constant and the burst gap varied.

For those networks with a bandwidth profile definition, or any kind of SLA, it is practical to use *ramp generation*. Ramp is a profile with an increasing number of frames in each burst, while the burst gap can be held constant or varied. A ramp profile is very good for testing network performance, and to discover at what level and under which circumstances the node or the network delivers error-free frames.



Stopped 00:00:00
Config-> Port A/Bandwidth profile/Flow A1
Mode: Ramp
Ramp, 0.00000% / 0.00000%
SRC: 00:DB:1E:00:16:80 (64 B)
DST: 00:00:00:00:00:00
1000BASE-T (RJ45)

Mode: Ramp
□ Off
□ Continuous
□ Burst
■ Ramp
□ Random

Stopped 00:00:00
Config-> Port B
Port mode: Loopback
Physical layer: Electrical

Port mode: Loopback
□ TX/RX
□ Monitor
■ Loopback
□ Disabled

Figure 7.17 Configuring traffic generation and loo back with ALBEDO xGenius.

7.10.2 In-Service Measurements

In-Service Measurements (ISM) are performed with real traffic, without interfering with normal services.

7.10.2.1 Traffic Control

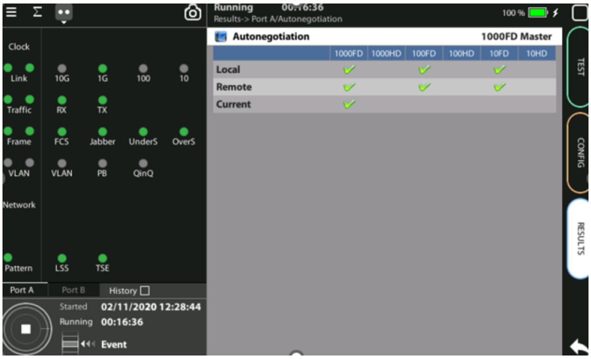
Some switches and routers can limit the bandwidth capacity. This means that it is feasible to have a GbE network offering just 150 Mbit/s. This is important for managing the whole network, and to execute UPC functions (ATM-like) to create a more predictable GbE. It is essential to test these limitations at installation, but also when the customer wants to upgrade the pipe size.

7.10.2.2 Traffic Statistics

Traffic statistics are an important source of information that help to plan and develop networks.

- *Common address* – This type of test is run to find the most popular talkers and listeners. The results can be useful so that you know how to adapt topologies for the biggest providers and consumers of information.
- *Packet sizes* – The efficiency of the network can be evaluated using different packet sizes. If there are many small packets, the proportion of overhead compared with payload data will be higher. The network can also be re-engineered if necessary, diverting traffic to other routes to avoid congested points.
- *Patterns* – It is also possible to identify traffic patterns (daily or seasonal) which could be modified to balance the load better and get more revenue.

- *Counts* provide generic information about traffic conditions and quality. Counts can include frames (total, broadcast, multicast, unicast), bytes (unidirectional or bidirectional), FCS errors, PAUSE messages, and VLAN.
- *Sizes* are separated into several ranges from 64 bytes up to >1518 bytes. For an equal number of bytes transferred, it is not the same to transmit small frames as large frames, because the payload-to-overhead ratio is not the same for the different frame sizes. You can also see if non-standard frames are being used; that is, jumbo frames that extend the frame size up to 9000 bytes.



Running 00:16:36
Results-> Port A/Autonegotiation

Autonegotiation 1000FD Master

	1000FD	1000HD	100FD	100HD	10FD	10HD
Local	✓		✓		✓	
Remote	✓		✓		✓	
Current	✓					

Started 02/11/2020 12:28:44
Running 00:16:36

Figure 7.18 Autonegotiation.

- *Errors* are always the first symptom that something is going wrong. It is useful to have a comprehensive classification of errors on the network because they show different causes: alignment, FCS, undersized frames, oversize frames, fragments, and jabberssee Chapter 9 "Performance Testing".

Selected Bibliography

[1] Paul Savill, *Metro Ethernet: What's taking so long*, Telecommunications Americas, Nov 2004

[2] Ralph Santitoro, *Metro Ethernet Services,* Metro Ethernet Forum, April 04

[3] Ralph Santitoro, *Bandwidth profiles,* Metro Ethernet Forum, March 03

[4] Paul Savill, *Metro Ethernet What's taking so long?*, Telecommunications Americas Nov 04

[5] JM Caballero, *Installation and Maintenance of SDH/SONET, ATM, xDSL and Synchronization Networks*, Artech House Aug. 2003, Addison Wesley

[6] Gary Southwell, *Metro Ethernet Offers Opex Gains*, Business Communications Review, July 2004

[7] Tony Kumeta, *All in the Presentation,* Network Cabling News Magazine, Dec 2003

[8] Matthew Mathis, Jeffrey Semke, Jamshid Mahdavi, and Teunis Ott, *The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm,* Computer Communications Review, vol. 27, #3, 1997

Chapter 8

Approval and Acceptance Tests

Ethernet operators and users may wish to approve specific nodes or stations, such as network adapters, hubs, servers, switches, or routers, which will be used as standard equipment on their network. Testing will help operators to decide which brand and model to use. Often a specific physical medium must also be selected, from among several optical and metallic alternatives, and this medium has to be qualified as well.

Approval and acceptance tests help operators to compare devices from different vendors, with a view to choosing one. Some years ago it was also common that once the equipment was purchased, a series of tests were carried out to every single device to confirm that they were working properly before they were accepted by the buyer. Acceptance and approval tests are defined by laboratories, and carried out by engineers working for the purchasers.

These tests, performed by the purchaser, usually go hand in hand with installation tests performed by the installer. Both types of tests are often defined together, and for this reason the documents where they are described usually refer to them as approval/acceptance/installation tests.

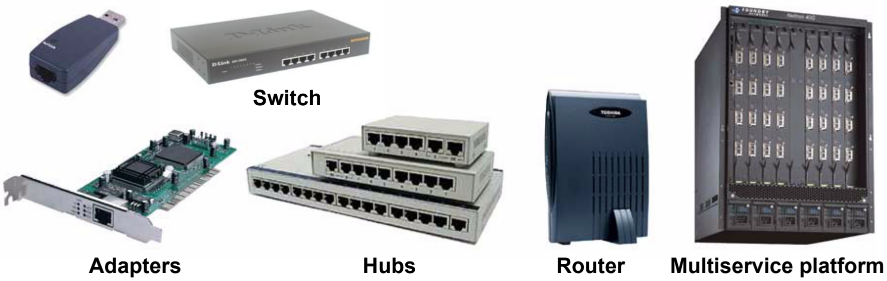


Figure 8.1 Ethernet devices

8.1 ETHERNET TESTING SUITE

Figure 8.2 Field testing GbE.

This *Ethernet testing suite* is an example of a set of test procedures that can be used to prove that an Ethernet device complies with certain specifications.

In this book we have basically referred to the IEEE 802.3ab and IEEE 802.3z standards. Nevertheless there are proprietary specifications as well, intended to cover non-standard features that could be eventually included. These procedures can be developed by installers or carriers to facilitate product development and certification by Ethernet equipment suppliers.

Approval and acceptance testing suites are not necessarily methods for demonstrating compliance, because they may be designed to verify certain features only. It may be possible to demonstrate compliance using other procedures. Laboratories of vendors, installers or service providers can use testing suites when auditing or certificating equipment on behalf of the certification authority.

They do not require opening the *Device Under Test* (DUT) to access special test points or to invoke test modes of operation. There are requirements that cannot be verified by black-box techniques, and supplier-proprietary procedures are required to test such requirements. These supplier-proprietary test procedures would be beyond the scope of an acceptance suite of a service provider.

Any testing suite has to be structured and adapted to the device features and communication layers. A hub is different to a router, because hubs do not have the IP functionalities that routers have, and this is why routers require an additional series of testing procedures. In the following test we have addressed only common features of a generic gigabit device, intentionally excluding half-duplex mode and STP cabling, because it is unlikely to find them on the market.

8.2 Physical Interfaces Test

This is probably the first test to be performed, the verification of operating ranges of transmitters and receivers, and data pulse shapes of both electrical and optical interfaces.

8.2.1 Optical Interfaces

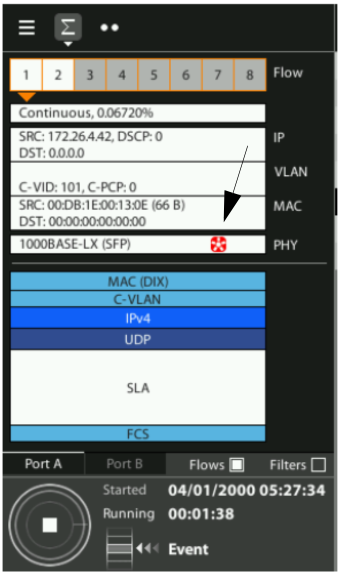

Continuous, 0.06720%
SRC: 172.26.4.42, DSCP: 0
DST: 0.0.0.0
C-VID: 101, C-PCP: 0
SRC: 00:DB:1E:00:13:0E (66 B)
DST: 00:00:00:00:00:00
1000BASE-LX (SFP)
Started 04/01/2000 05:27:34
Running 00:01:38

Figure 8.3 Summary pannel with laser indicator

Testing optical interfaces is more important than analyzing their electrical equivalents, because optical components are more prone to damage and contamination. The following tests can detect most failures:

- *Transmission power* – An optical power meter is needed for testing transmission power. This test is used to measure the output power of the device's transmit laser or LED.
- *Receiver sensitivity* – When optical power falls below a certain limit, the level of noise above the signal starts to become significant, causing errors. Optical sensitivity testing is key in evaluating the quality of optical receivers. The values for sensitivity are specified by the IEEE.
- *Spectral density* – Spectral density tests are quite sophisticated, and they require an optical spectrum analyzer. The result is compared with the acceptance mask for the signal.
- *Eye mask* – An eye mask test uses an optical communications analyzer to check transmit port compliance with industry standards. The eye mask can verify characteristics such as jitter, data rate and optical overload (see Figure 8.4).
- *Extinction ratio* – Extinction ratio test determines the ratio of optical power for a logical 1 versus a logical 0. The ratio must be large enough to ensure that the detector circuits can differentiate a high data bit from a low data bit.

- *Detecting optical overload* – The optical receiver has a limit for the maximum power it can receive. The purpose of optical overload testing is to check that this limit is adhered to. If it is not, optical saturation or *overload* occurs, which makes it impossible to detect the pulses.
- *Jitter tolerance* – Nodes must tolerate a certain amount of jitter in their inputs without losing synchronization or introducing errors. Jitter tolerance tests are used to measure how much jitter nodes can tolerate. The maximum jitter amplitude that nodes must be able to tolerate is specified by the IEEE (see Figure 8.5).

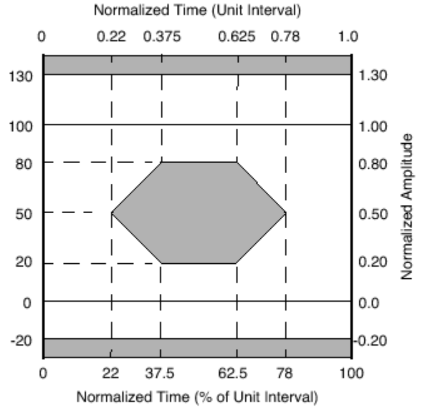


Figure 8.4 Transmitter eye mask definition.

8.2.2 Electrical Interfaces

Tests based on the Gigabit Ethernet specifications may include:

- *Transmitter electrical specifications* – In transmitter tests various attributes of the electrical transmitter are verified including differential amplitude, skew, and rise and fall times.
- *Receiver electrical specifications* – Receiver tests measure the maximum input and sensitivity, input impedance and jitter.
- *Peak differential output voltage and Level accuracy* – Voltage measurements are made at precise points in the output data stream, using predefined test setups.
- *Maximum output droop* – In receiver maximum output droop testing pairs of points on particular pulses are measured and compared, to ensure that excessive droop is not present, and to verify the receiver's ability to exceed certain levels of voltage droop.
- *Differential output templates* – These measurements look at various positions within a data stream to check that a standard waveform falls within a standard template defined by the IEEE.
- *Return Loss (RL)* – These tests measure the return loss for both the RX and TX ports.

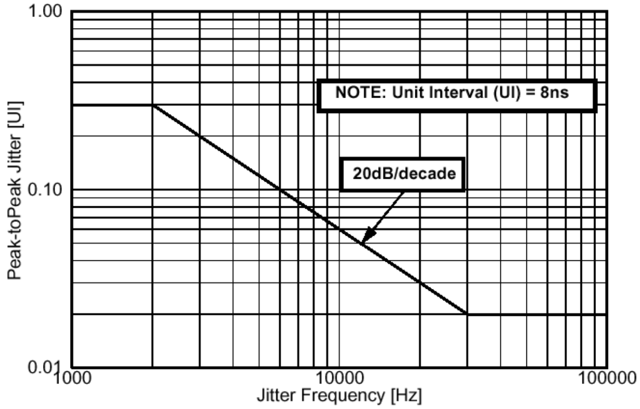


Figure 8.5 Jitter tolerance for GbE.

8.2.3 Measuring Frequency

Low-quality synchronization sources that deviate from the nominal value of the signal they supply, or badly synchronized clock recovery circuits can give rise to problems in the operation of NEs. For this reason, it is necessary to measure the frequency of the line signal at all hierarchical levels, and check its deviations.

The GbE tester needs to be capable of measuring the frequency of the signals received, and, in some cases, the associated line codes as well. The frequency of the signal is usually given in Hz, and its deviation relative to the nominal hierarchical value in ppm. This shows whether the frequency measured is inside or outside the range defined.

8.3 Jitter Test

Line jitter occurs when bits on a link are received either earlier or later than expected, causing sampling at a non-optimum instant and increasing the probability of bit errors and slips.

Line jitter should not be confused with *frame jitter*. Frame jitter is a quantity-related with the delay variation of frames delivered between the same transmitter and receiver along the network (see Paragraph 6.5.2). It depends on queuing times in the intermediate nodes as well as on the route followed by the frames.

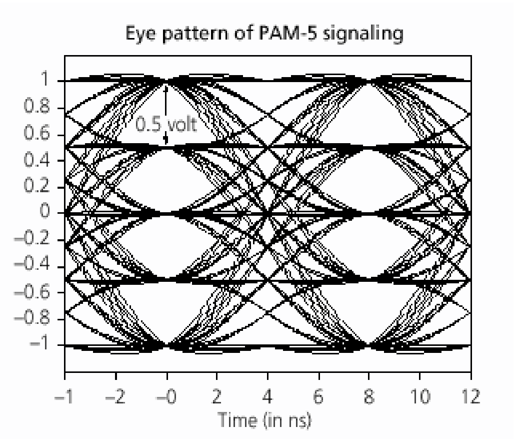


Figure 8.6 Eye pattern of PAM-5 signaling.

8.3.1 Phase Fluctuation and Jitter

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 relative to its origin in time. A significant instant is defined arbitrarily; for instance, it may be a trailing edge, or a leading edge if the signal is a square wave.

In digital communications, phase is related to clock signals. Every data signal has an associated clock signal that makes it possible to determine, on reception, when to read the value of the bits this signal is made up of. Clocks can be generated by the receiver, obtained from a high-quality timing source, or derived from the data signal. Phase fluctuation is an impairment of the data signal due to a (time-dependent) offset of this signal regarding its clock (see Figure 8.7).

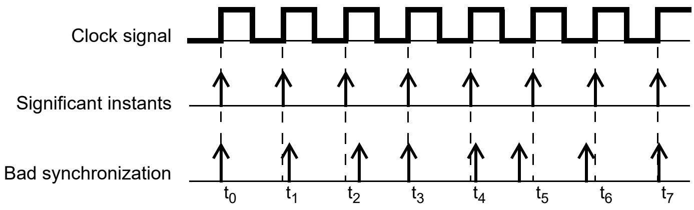


Figure 8.7 Phase error of a signal in relation to its ideal frequency

If the clock signal is generated by the receiver or obtained from a high-quality timing source, phase fluctuation is always a potential cause of problems, but if the clock is derived by the data signal, as is the case of Ethernet, then only phase fluctuation above a certain frequency needs to be considered.

In Ethernet, the clock signal is recovered from the data signal with the help of a *Phase-Locked Loop* (PLL) that is frequency-sensitive. A PPL can track low-frequency phase fluctuations on the data signal, and only high-frequency fluctuation can cause sampling errors. This high-frequency phase fluctuation is known as *line jitter* or just jitter.

8.3.2 Jitter Measurement

The main causes of line jitter in Ethernet networks are the following:

- *Random phase noise* in the clock circuits caused by random movement of the current carriers. This impairment is independent of the data signal.
- *Pattern-dependent jitter* due to limitations of the receiver PLL. Ideally, the clock can be embedded in the data signal and recovered transparently by the receiver PLL, but in real situations the clock signal is different for every possible data signal. This type of jitter is accumulative, which means that it increases together with the increase in the number of repeaters looked at.

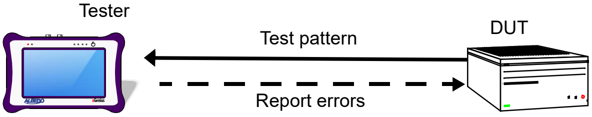


Figure 8.8 Jitter measurement of an Ethernet node. The DUT generates a test pattern. The tester measures the amplitude of the jitter generated by the DUT or just the transmission errors due to the presence of jitter.

Pattern-dependent jitter makes the test signal important when testing jitter. The IEEE defines test patterns to measure jitter. Each pattern has a different purpose, but the measurement is always performed in the same way. A predictable signal is injected by the transmitter that is being tested, and pattern errors are detected by the tester.

The patterns are given to test GbE-level signals only. The reason is that it is easier for jitter to damage a high bit rate signal, because synchronization constraints are more exigent for them.

Table 8.1
Deterministic patterns to test jitter

Test pattern	*Bit sequence*	*Code-group*	*Purpose*
High-frequency	1010101010101010101010...	D21.5	Test random jitter at a BER of 10-12 and test the asymmetry of transition times
Low-frequency	1111100000111110000011...	K28.7	Test low frequency random jitter and PLL tracking errors
Mixedfrequency	1111101011000001010011...	K28.5	Test combined random jitter and deterministic jitter

The IEEE defines deterministic and random test patterns. In fact, all the test patterns are deterministic, but the patterns classified as random have a more complex structure than the rest. *Deterministic test patterns* are obtained by looped transmission of a specific 8B/10B code group (see Table 8.1), whereas *random test patterns* are streams of identical packets separated by a minimum *Inter-Packet Gap* (IPG) and encapsulated in the usual way (see Figure 8.9).

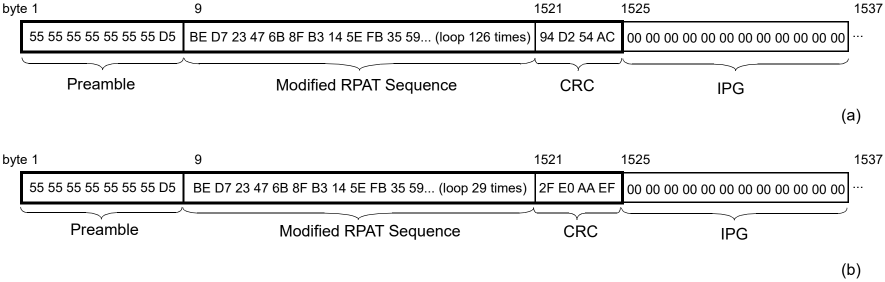


Figure 8.9 Random patterns to test jitter, (a) long continuous random test pattern, (b) short continuous random test pattern.

8.4 1000BASE-T PMA Testing

These tests should determine if a product conforms to IEEE 802.3 (see Figure 8.10). For testing *Physical Medium Attachment* (PMA) electrical characteristics, the standard defines a number of tests:

- *Peak differential output voltage and level accuracy* – to verify that the value of the wave form falls within the specified range.
- *Maximum output droop* – to verify that the transmitted signal decays according to the specification, but not faster.
- *Differential output templates* – to verify that output falls within the time domain template.

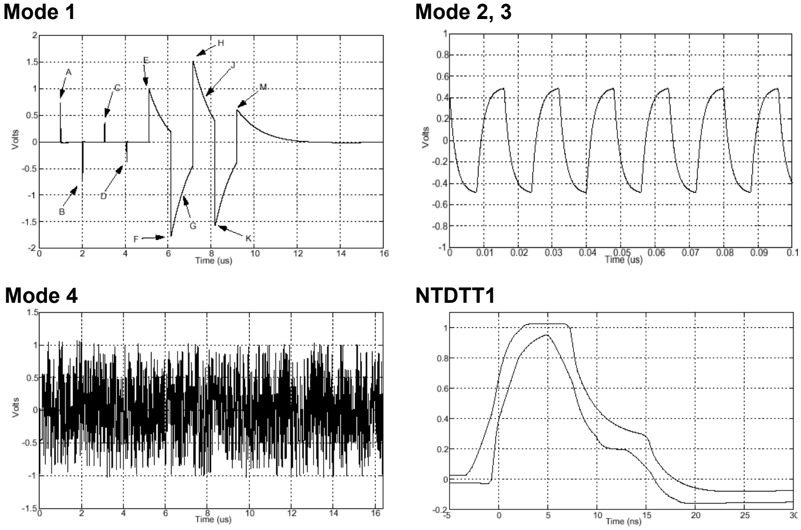


Figure 8.10 1000BASE-T PMA test. Example of transmitter test modes of 1, 2, 3 and 4 waveforms. Bottom right, the *Normalized Time-Domain Transmit Template* (NTDTT).

- *MDI return loss* – to measure the RL at the MDI for all four channels.

8.5 1000BASE-X PCS Testing

The PCS functions include the PCS transmit and receive, Carrier sense, Synchronization and Auto-Negotiation.

8.5.1 Synchronization Check

Synchronization ensures the alignment of multicode-group ordered sets to *n*-numbered code group boundaries. Synchronization is acquired by the detection of three ordered sets containing commas in their leftmost bit positions without intervening in valid code group errors (see Figure 8.11).

The DUT should be able to maintain synchronization while sending frames, and for a specific set of invalid code group sequences.

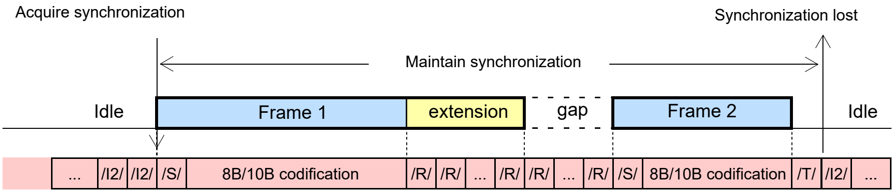


Figure 8.11 In the line coding for 1000BASE-X, code-words are used to encapsulate the frame to unambiguously distinguish data from control information.

8.5.2 PCS Transmission

The *Physical Coding Sublayer* (PCS) uses a transmission code to improve the transmission characteristics of the information to be transferred across the link (see Figure 8.11). Some tests can be performed to verify that the DUT transmits information correctly:

- *8B/10B conversion* – each 8-bit data octet is mapped into 10-bit symbols (see Paragraph 1.3.2)
- *Frame encapsulation* – The Ethernet frame is encoded using the 8B/10B code rules. However, control groups are also required for the line code: */S/* indicates the start of a packet, */T/* and */R/* frame termination, */I1/* and */I2/* inter-frame gaps, etc.

8.5.3 PCS Reception

At the receiver side, the inverse functions should be implemented including 8B/10B decoding, detecting end of packet, or the reception of */C/* during IDLE.

8.6 Auto-Negotiation

Many early implementations of the *auto-negotiation feature* are not compliant with the final standard. Some may be fixed with driver updates, while others require new hardware. For example, many Ethernet products older than 1997 do not support auto-negotiation. This has created a situation where the new standard-compliant products appear to be causing problems, when in fact it is the older, non-compliant hardware that cannot take advantage of this new valuable feature.

Auto-negotiation varies depending on the standard. 10/100/1000BASE-T are potentially able to establish links between them, after a successful auto-negotiation process has finished and common parameters have been agreed. 1000BASE-X also has auto-negotiation capabilities, but reduced to the flow control mechanism only.

The following tests are intended to verify the capacity of the DUT to manage the auto-negotiation protocol according to the IEEE 802.3.

8.6.1 10/100/1000BASE-T Auto-Negotiation

8.6.1.1 Enable/Disable

If the auto-negotiation function of the DUT is enabled, the device should always send auto-negotiation messages (see Figure 8.12) as soon it is connected to a link and before it reaches its final stage, which could be 'link established' or 'impossible to establish link'. After this stage these messages should cease. If the link is broken and reconnected, the DUT must restart the process.

It is possible to disable the auto-negotiation feature in some devices. If this is done, the DUT should not send any auto-negotiation messages at all.

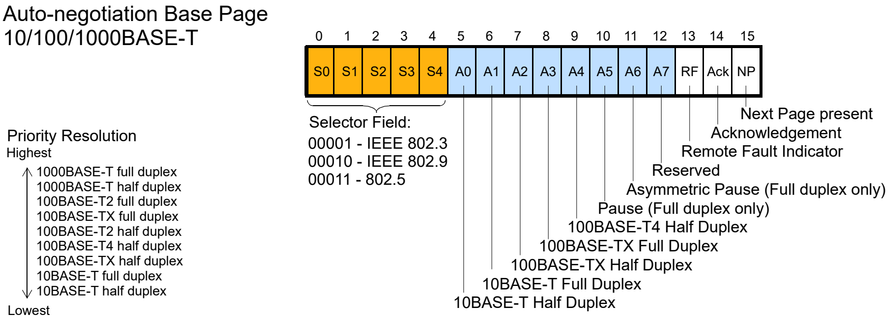


Figure 8.12 Auto-negotiation base page for 10/100/1000BASE-T using UTP and RJ-45. The base page includes information about the abilities of the device (bit rate, duplex mode, and flow control mode).

8.6.1.2 Base Page

The first action the auto-negotiation protocol must take is to declare itself to advertise the abilities of the device by means of a especial message known as *base page* (see Figure 8.12). The message must reflect the exact speed, duplex mode and flow control features of the DUT. Naturally, a device should not advertise abilities it does not have.

8.6.1.3 Priority Resolution

When both devices attached to a link have sent the base page (if auto-negotiation is enabled), they must proceed to the resolution of speed, duplex, pause mode, and master/slave:

- *Speed resolution* – The DUT must select the highest bit rate that both devices can manage. The top–down priority list is, obviously: 1 Gbit/s, 100 Mbit/s, and 10 Mbit/s.
- *Duplex mode* – The DUT must select full duplex (FDX) if both devices support it, and if not, half-duplex mode is used.
- *Pause mode* – If FDX has been resolved, the devices must also agree the flow control scheme or Pause protocol support. Transmit (Tx) and Receive (Rx) capabilities can be independent, and this way we can get up to four combinations per Tx/Rx pair. The possibilities are: a) Enabled Tx and Rx; b) Disabled Tx and Rx, c) Enabled Tx, disabled Rx; d) Disabled Tx, enabled Rx.
- *Master–Slave* – In 1000BASE-T, one device must be declared master; the other one will be the slave. To decide which one is the master, the top–down priorities are: manual configuration, multiport and single port. If both are multiport or single port, it is necessary to send unformatted page 2, and the device with the highest seed value will be declared master (see Figure 8.13). Faulty situations occur when both devices are manually configured to be either master or slave.

8.6.2 1000BASE-X Auto-Negotiation

The 1000BASE-X auto-negotiation mechanism is gigabit-specific. This means that it is not possible to negotiate bit rate, but duplex operation, flow control and similar capabilities. This mechanism cannot resolve any improper configurations of wavelength of fiber optics. If that should happen, the link just wouldn't work.

1000BASE-X has a set of 8B/10B codes to make sure that auto-negotiation messages are not interpreted as data frames.

8.6.2.1 Enable/Disable

If the DUT's auto-negotiation feature is enabled, it should always send auto-negotiation messages (see Figure 8.12) as soon as it is connected to a link and before it reaches its final stage, which could be 'link established' or 'impossible to establish link'. After this stage, the messages should cease. If the link is broken and reconnected, the DUT must restart the process.

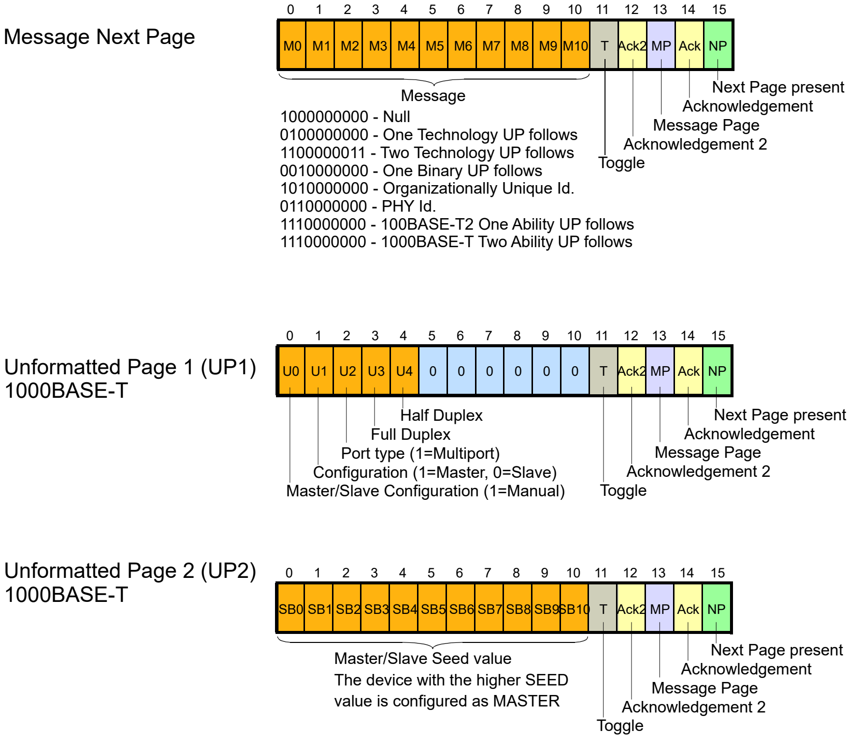

Message Next Page
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
M0 M1 M2 M3 M4 M5 M6 M7 M8 M9 M10 T Ack2 MP Ack NP
Message
1000000000 - Null
0100000000 - One Technology UP follows
1100000011 - Two Technology UP follows
0010000000 - One Binary UP follows
1010000000 - Organizationally Unique Id.
0110000000 - PHY Id.
1110000000 - 100BASE-T2 One Ability UP follows
1110000000 - 1000BASE-T Two Ability UP follows
Next Page present
Acknowledgement
Message Page
Acknowledgement 2
Toggle
Unformatted Page 1 (UP1)
1000BASE-T
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
U0 U1 U2 U3 U4 0 0 0 0 0 0 T Ack2 MP Ack NP
Half Duplex
Full Duplex
Port type (1=Multiport)
Configuration (1=Master, 0=Slave)
Master/Slave Configuration (1=Manual)
Next Page present
Acknowledgement
Message Page
Acknowledgement 2
Toggle
Unformatted Page 2 (UP2)
1000BASE-T
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
SB0 SB1 SB2 SB3 SB4 SB5 SB6 SB7 SB8 SB9 SB10 T Ack2 MP Ack NP
Master/Slave Seed value
The device with the higher SEED
value is configured as MASTER
Next Page present
Acknowledgement
Message Page
Acknowledgement 2
Toggle

Figure 8.13 Auto-negotiation. Unformatted pages 1 and 2 hold the information needed to resolve the Master-Slave mode.

It is possible to disable the auto-negotiation feature in some devices. In that case, logically, the DUT should not send any auto-negotiation messages.

8.6.2.2 Base Page

The first action the auto-negotiation protocol must take is to declare itself to advertise the abilities of the device by means of a base page (see Figure 8.14). The message must reflect the exact speed, duplex mode and flow control features of the DUT.

8.6.2.3 Priority Resolution

When both devices attached to the link have sent the base page (if auto-negotiation is enabled) they must proceed to the resolution of speed, duplex, pause mode, and master/slave:

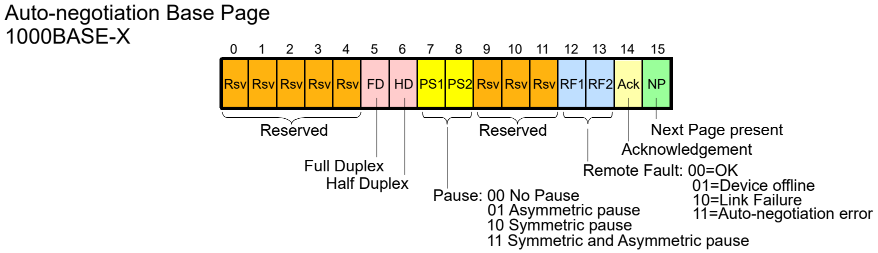
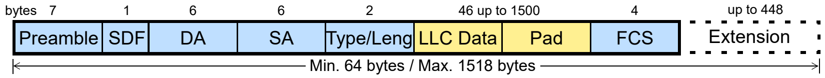


Figure 8.14 1000BASE-X auto-negotiation base page message includes information about the device's abilities, such as duplex mode and flow control mode.

- *Duplex mode* – see Paragraph 8.6.1.3.
- *Pause mode* – see Paragraph 8.6.1.3.
- *Remote Fault (RF)* – If the RF feature is supported, it is used to indicate error conditions to the link partner. Any indication other than 'No error' can make the link non-operational. If a problem is detected, a device can use an RF to inform its partner about the problem.



Preamble: Synchronization pattern
SDF: Start Frame Delimiter (10101011)
DA: Destination Address
SA: Source Address
Type: Indicates the nature of the client protocol
Length: Number of bytes of the LLC data
LLC data: Information supplied by LLC layer
Pad: Ensures a minimum frame size of 64 bytes
Extension: Ensures a minimum frame size (only GigE)
FCS: Frame Check Sequence or CRC

Figure 8.15 The basic IEEE 802.3 MAC Frame format (1997).

8.7 MAC Layer Testing

The MAC is the data link sublayer in charge of transferring data to and from the physical layer. It can be considered as the core of all Ethernet versions, and it is essential for normal operation.

A number of benchmarking tests can be performed to accept a specific Ethernet device. However, many of the new Ethernet installations are full duplex only, to avoid collisions and to increase throughput. In this case, the MAC layer is much simpler, because it has to manage the CSMA/CD protocol needed when the physical media is shared by two or more stations.

The majority of gigabit installations use full duplex only, which is also why we have not considered half duplex verification in the following test.

8.7.1 Frame Error Management

When frames are detected by an Ethernet node, the complete frame is immediately dropped. This is one of the peculiarities of Ethernet. We can carry on generating special test traffic, and then analyzing how the DUT manages frames.

8.7.1.1 Receiving Frames

- *Preamble errors* – This field is used to allow synchronization with the received frame's timing. An invalid preamble should not interfere with the reception of a correct MAC frame, and the DUT should be able to get the packet.

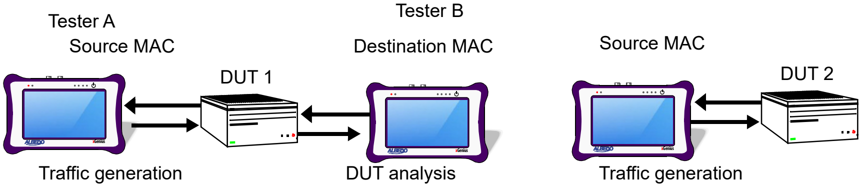


Figure 8.16 MAC frames analysis. At left, tester A generates a special frame to test how the DUT works. Tester B should check if DUT1 forwards or drops every single frame. While DUT1 could be a hub or a router, the right side testing fixture is for end stations, such as a server or a PCs, and it would require DUT2 to offer a view on the Ethernet traffic.

- *SDF errors* – This field is the sequence *10101011*. It immediately follows the preamble pattern and indicates the start of a frame. Any frame with an invalid SDF should be discarded.
- *Source and Destination address* – This field is 48 bits in length; 16-bit addresses are rejected. End stations should discard any unicast frames with address different to the own.
- *Type/Length errors* – If the value of this field is greater than or equal to 1536, that indicates a Type client protocol. If the value is less than or equal to 1500, it indicates the number of MAC client data octets contained in the LLC-Pad

field. If the length is less than the data field, the rest is considered a pad and removed. If the length is greater, there is an error, and the frame must be dropped. An additional difficulty is presented by the DUT supporting proprietary jumbo frames, which can have a size up to 9000 bytes.

- *Fragments of frames* – Any frame which is less than 512 in full-duplex Gigabit Ethernet is illegal and must be dropped by the DUT.
- *Runts* are fragments with a valid CRC. The DUT should drop them.
- *Jabber frames* are described most often as a frame greater than the maximum of 1518 bytes, with a bad CRC. The DUT should discard all of them.
- *FCS error* – This field is a CRC, generated from an algorithm, and it is derived from the data in the frame. If the frame is altered between the source and the destination, the receiving station will recognize that the CRC does not match the contents of the packet. The frame must be dropped by the DUT.

8.7.1.2 Transmitting Frames

Acceptance tests should also verify that the DUT can properly encapsulate client information in MAC frames. This information includes headers, such as preamble and SDH, source and destination addresses.

It would also be convenient to verify that the DUT calculates the frame length, padding and FCS fields correctly. The inter-frame gap, which must be at least 96 bytes, would require additional verification.

8.7.2 Full Duplex Verification

Full-duplex operation ignores the CSMA-CD protocol and implements a point-to-point Ethernet connection between two devices. In a full-duplex network there are no collisions, and end stations at opposite ends of a full-duplex Ethernet link may transmit simultaneously.

- *No frame collisions* – Collisions cannot occur, and the DUT should be able to transmit and receive simultaneously, without observing any collisions or jam signals.
- *Carrier Sense (CS) holdover* – The CS is a physical signal used to take up the media during transmission. In half duplex it has to be held over CS after sending the frame to allow transmission by other stations. However, in full duplex the MAC layer does not have to defer the CS because the channel is devoted to it.

- *No frame extensions* – Frame extensions ensuring a minimum size in Gigabit Ethernet are not needed anymore. Minimum size was a requirement to guarantee that collisions wouldn't happen while transmitting.
- *No frame bursting* – In half duplex a station is allowed to send multiple frames to avoid collisions and improve efficiency (see Paragraph 1.3.2.3). Without collisions the DUT should not use the frame bursting technique.

8.8 Physical-Layer Interoperability Test

Physical-layer interoperability tests ensure that the DUT is able to establish a link and exchange packets with a device of similar characteristics. At the physical layer, the GBIC transceiver must be of good quality, the type of cable must be appropriate, and the configuration must be correct. The integrity of the physical layer is key to getting good performance.

8.8.1 Link Establishment

A *link establishment test* checks if the DUT is able to establish a link at the optimal bit rate. Many Ethernet products have auto-negotiation capability, and others have manual configuration or nothing at all. In any case, they should be able to detect the link speed of the partner device. Once the link has been successfully established, the DUT should be able to recover after it has been disconnected and connected again.

Another interesting test is to verify that the manually configured devices can establish a link with remote stations with the same manual configuration and with a station with compliant auto-negotiation features. Furthermore, it should be verified that a DUT with an optical layer should not establish any links if the partner station has a similar, but not the same, wavelength.

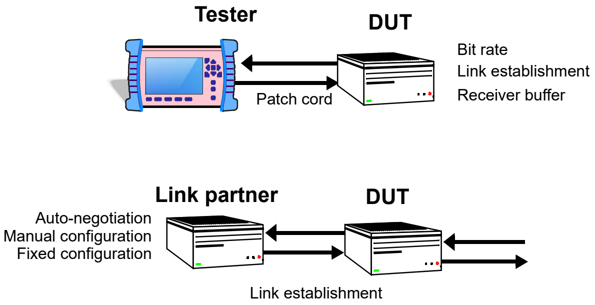


Figure 8.17 Acceptance test should begin by verifying physical layer interoperability.

8.8.2 Frame Error Ratio

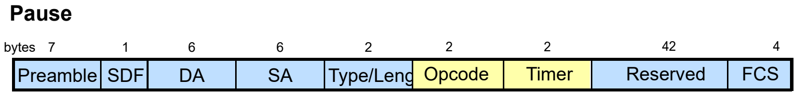
The *frame error ratio test* determines the percentage of error-free frames that the DUT can generate. The ratio must be below a threshold which will depend on parameters such as bit rate, type of traffic to transport or SLA to achieve. The test can be performed according to the RFC 2544 (see Paragraph 9.3.3).

8.8.3 Receiver Buffer Test

A *receiver buffer test* determines the buffer management ability of the DUT under heavy traffic conditions. When frames start to be dropped, the overload level has been achieved. The test can be performed according to the back-to-back test of the RFC 2544 (see Paragraph 9.3.4)

8.9 Flow Control Test

The f*low control protocol*, which can be implemented in full-duplex systems, is based on a short packet known as the *Pause frame* (see Figure 8.18). This frame provides a mechanism whereby a congested receiver can ask the transmitter to stop the transmission (see Paragraph 2.3).



Opcode: indicates Pause frame, hexa value = 0001
Pause Time: time is requested to prevent transmission

Figure 8.18 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.

The DUT may be able to work in several flow control modes, such as *Symmetric Flow Control* (SFC), *Asymmetric Flow Control* (AFC), and no flow control at all. The SFC means that Pause frames may flow in either direction. The AFC means that Pause frames may only flow in one direction, whether that direction is; towards or away from the local device. No flow control means that Pause frames are not allowed.

We can verify how the DUT manages the flow control by analyzing the Pause frame:

- Send a Pause frame with a valid pause time value. The DUT will stop sending frames until the countdown value reaches zero.

- Send a Pause frame with a pause time equal to zero. The DUT should resume sending frames, if it were previously stopped, otherwise it shall continue normal operation.
- MAC control frame reception and handling, to verify that the DUT rejects invalid MAC control Pause frames.
- Receive MAC Control Pause frames of incorrect size. This is to determine how the DUT handles Pause frames that are of incorrect size.
- Pause frame transmission. To verify that the DUT transmits properly encapsulated Pause frames.

Selected Bibliography

[1] IEEE Std 802.3™-2002, *Revision of IEEE Std 802.3*, 2000 Edition

[2] Research Computing Center, *Ethernet Interoperability*, The University of New Hampshire Sep 2003

[3] Kevin L. Paton, *Gigabit Ethernet Test Challenges*, Oct 2001 Test and Measurement World Magazine

[4] RFC 2544, *Benchmarking Methodology for Network Interconnect Devices*, S. Bradner and J. McQuaid, March 1999

[5] RFC 1242 - *Benchmarking terminology for network interconnection devices,* S. Bradner July 1991

[6] RFC 2285, *Benchmarking Terminology for LAN Switching Devices*, R. Mandeville Network Laboratories February 1998

[7] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.

Chapter 9

Performance Testing

The RFC 2544 (*Request For Comments*) is a document that describes benchmarking tests for network devices. Vendors can use these tests to measure and outline the performance characteristics of their 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 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.

9.1 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.

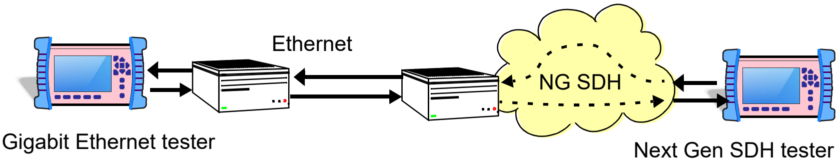


Figure 9.1 WAN performance may be tested by setting up two identical devices connected by the appropriate short-haul versions of the WAN modems. Performance is then measured between a LAN interface on one DUT, and a LAN interface on the other DUT.

9.1.1 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 shape.

9.1.2 Traffic Used in the Test

- *Traffic pattern* – The traffic on a real network is not constant, but occurs in bursts. The RFC 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 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 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 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.

9.1.3 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.

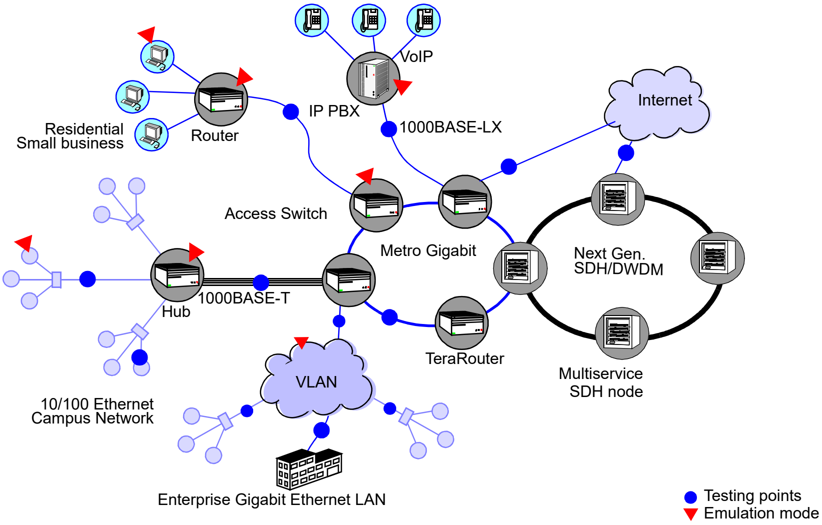


Figure 9.2 Ethernet topology and tester Gigabit Ethernet testing points

9.1.4 Test Setup

The aim of this set of tests is to evaluate the performance of equipment in real-world situations. The RFC 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

9.2 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:

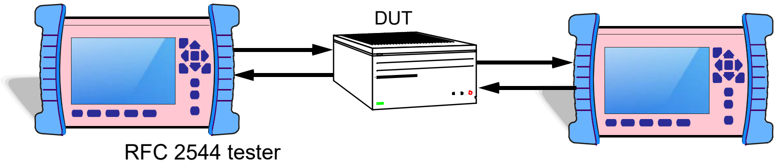
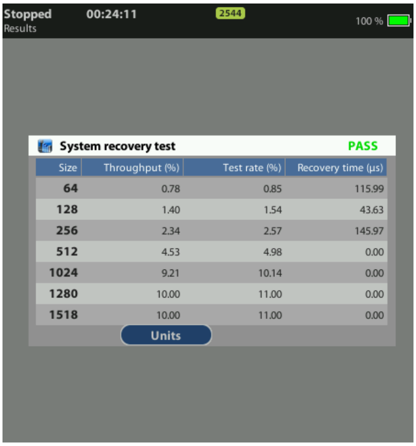


Figure 9.3 Performance evaluation using two testers

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.

9.3 The Tests



System recovery test PASS

Size	Throughput (%)	Test rate (%)	Recovery time (µs)
64	0.78	0.85	115.99
128	1.40	1.54	43.63
256	2.34	2.57	145.97
512	4.53	4.98	0.00
1024	9.21	10.14	0.00
1280	10.00	11.00	0.00
1518	10.00	11.00	0.00

Figure 9.4 RFC 2544 with xGenius.

The RFC 2544 discusses and defines a number of tests that may be used to describe the performance characteristics of a network interconnecting device. Besides defining the tests, this document describes specific formats for reporting the results.

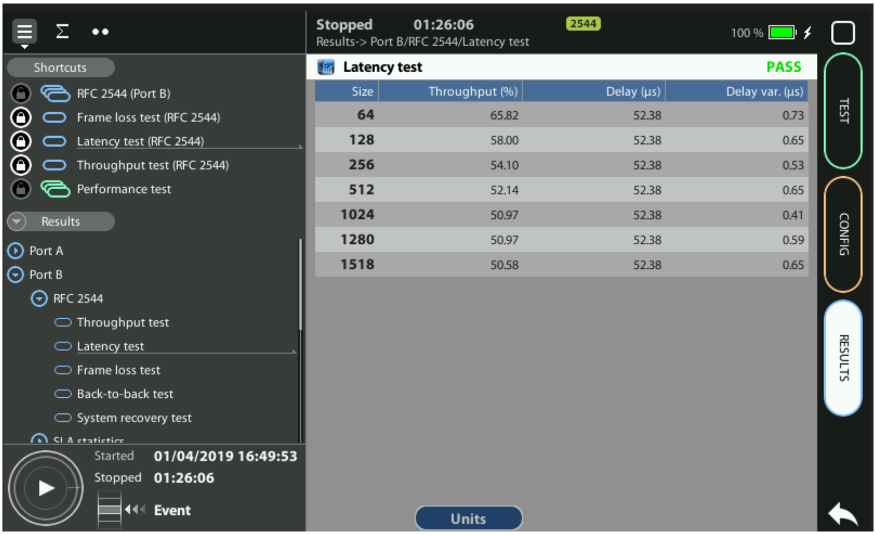
9.3.1 Throughput

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.

If a single value (minimum frame size) is desired, it must be expressed in frames per second, or alternatively in bits or bytes per second.



Stopped 01:26:06
Results-> Port B/RFC 2544/Latency test

Latency test PASS

Size	Throughput (%)	Delay (µs)	Delay var. (µs)
64	65.82	52.38	0.73
128	58.00	52.38	0.65
256	54.10	52.38	0.53
512	52.14	52.38	0.65
1024	50.97	52.38	0.41
1280	50.97	52.38	0.59
1518	50.58	52.38	0.65

Started 01/04/2019 16:49:53
Stopped 01:26:06

Figure 9.5 RFC 2544 throughput results on tester xGenius.

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.

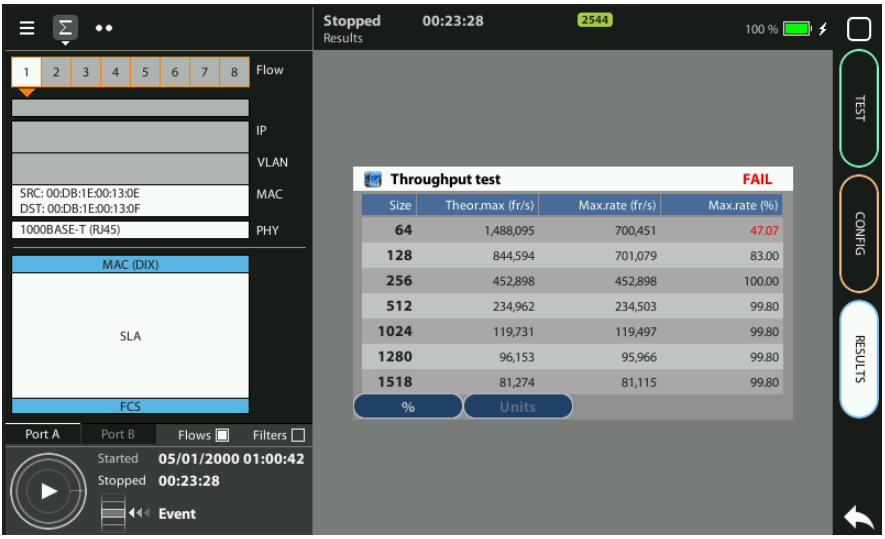
9.3.2 Latency

This test determines the *latency* inherent in the DUT. The initial data rate is based on the results of the throughput test. Typically, a packet is time stamped and inserted in the middle of a burst, and the time the stamped packet takes 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.



Stopped 00:23:28
Results

SRC: 00:DB:1E:00:13:0E
DST: 00:DB:1E:00:13:0F
1000BASE-T (RJ45)

Throughput test FAIL

Size	Theor.max (fr/s)	Max.rate (fr/s)	Max.rate (%)
64	1,488,095	700,451	47.07
128	844,594	701,079	83.00
256	452,898	452,898	100.00
512	234,962	234,503	99.80
1024	119,731	119,497	99.80
1280	96,153	95,966	99.80
1518	81,274	81,115	99.80

Started 05/01/2000 01:00:42
Stopped 00:23:28

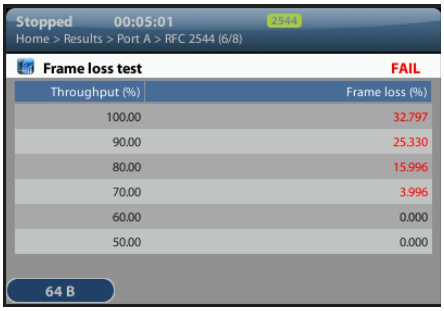
Figure 9.6 Tester Gigabit Ethernet

9.3.3 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 must be 10% of the maximum rate, although you may want to define a finer granularity.

In this test the test frame is addressed to the same destination as the rest of the data stream, and each test frame addressed to a new destination network.



Stopped 00:05:01 2544
Home > Results > Port A > RFC 2544 (6/8)

Frame loss test FAIL

Throughput (%)	Frame loss (%)
100.00	32.797
90.00	25.330
80.00	15.996
70.00	3.996
60.00	0.000
50.00	0.000

64 B

Figure 9.7 Frame loss results on tester GbE.

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.

9.3.4 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.

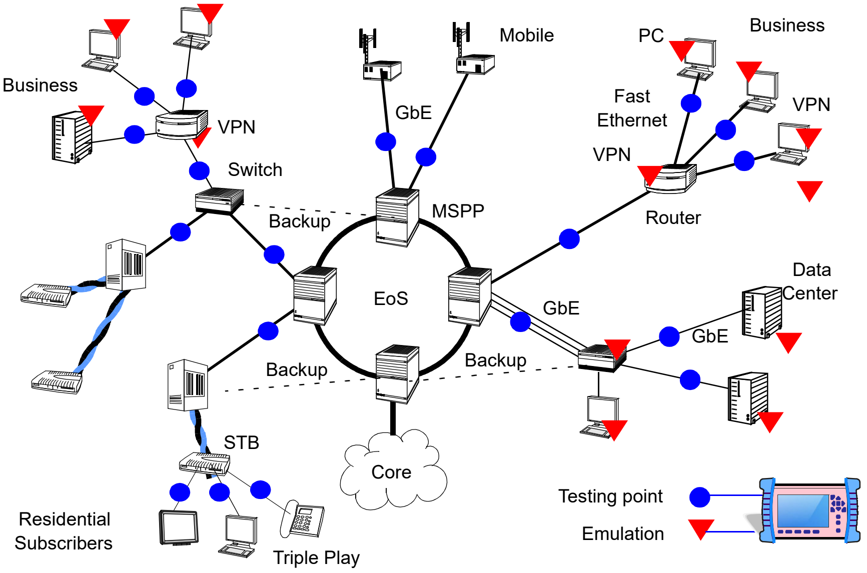


Figure 9.8 Carrier Ethernet network and tester Gigabit Ethernet testing points.

9.3.5 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.

9.3.6 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.

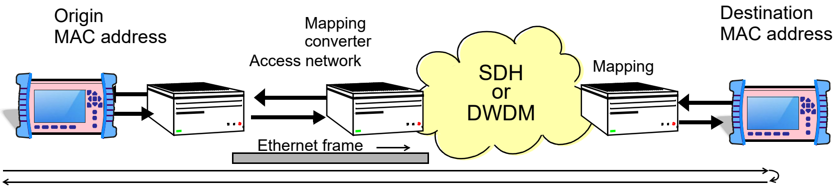


Figure 9.9 Continuity test at MAC layer. Media converters to long-haul optical signal are usec in DWDM. Classical mappings in Virtual Containers and GFP and LCAS are used to transport Ethernet over SDH.

9.4 Traffic Generation

A major factor when determining hardware and software needs for the Gigabit Ethernet DUT test approach is the *internal data handling capability* of the DUT. When testing the product, it is essential to pass data through the product. Some products can be put into a mode where an unframed PRBS stream can be passed without any issues. Other products must receive and transmit actual IP data, which may be looped back internally or externally.

Still other DUTs that require actual IP data also contain smart routing algorithms built into the on-board firmware. This adds additional constraints, since if the IP data is looped back to the same port, the router knows this is not correct, and the data will be destroyed. In some cases a DUT may even have the built-in ability to internally generate its own data stream and to verify that data when it returns to the Rx port.

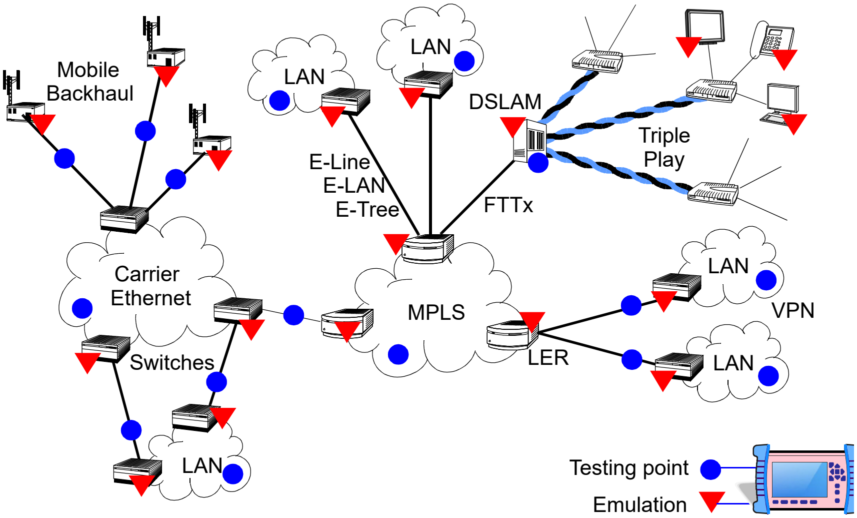


Figure 9.10 Carrier Ethernet network delivering Triple Play services and tester Gigabit Ethernet testing points.

Because of these varying internal operating characteristics, it is imperative to know the internal hardware and software design of the DUT. In the first case above, a simple 1-giga PRBS BER test instrument can be used. The second case calls for a tester that supports IP traffic. In the third case, not only will IP traffic support be needed, but the test set-up will also need either a golden module or chassis that can accept the test traffic and route it legally without losing any data. This requirement makes a stand-alone fixture approach impractical. In the final case where no external traffic test equipment is required, traffic testing can be accomplished using only loop-back cabling. To support traffic tests, there must be some means of loop-back to provide a complete path through the product under test.

These loop-backs are most effective when they are achieved externally to the product under test, either via an external jumper or other modules capable of handling the test traffic. This approach ensures that the module's backplane connections are tested. To provide a more granular level of fault isolation, some products are designed with internal loop-back capabilities. If a test fails, these localized loop-backs can be employed to isolate the fault location. If this capability is desired, the test engineer will need to contact the product design group during the design phase, as this capability must be designed into the product.

There is not much difference between testing done by using traffic streams of mixed protocols and testing with individual protocols. That is, if protocol A testing and protocol B testing give two different performance results, mixed protocol testing appears to give a result which is the average of the two [4].

The maximum frame rate to be used for LAN-WAN-LAN configurations is a judgment that can be based on known characteristics of the overall system, including compression effects, fragmentation, and gross link speeds. Practice suggests that the rate should be at least 110% of the slowest link speed.

Selected Bibliography

[1] RFC 2544, *Benchmarking Methodology for Network Interconnect Devices*, S. Bradner and J. McQuaid, March 1999

[2] RFC 1242 - *Benchmarking terminology for network interconnection devices,* S. Bradner July 1991

[3] RFC 2285, *Benchmarking Terminology for LAN Switching Devices*, R. Mandeville Network Laboratories February 1998

[4] Kevin L. Paton, *Gigabit Ethernet Test Challenges on the Manufacturing Floor*, Test and Measurement World Magazine Oct 2001

[5] Robert Breyer, *Sean Riley, Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.

[6] *Gigabit Ethernet: Accelerating the Standard for Speed*, Gigabit Ethernet, May 1999

Chapter 10

Cable Testing for 1000BASE-T

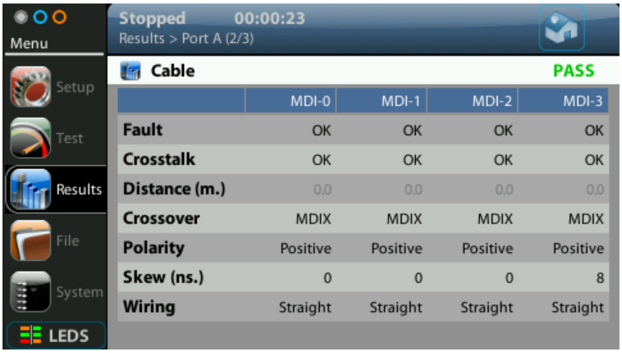
10.1 Cable Categories

The good news is that Gigabit Ethernet offers a cost-effective migration to 1000BASE-T from those installations running 10/100BASE-T over Category 5 cabling systems up to 100 meters. The bad news is that you cannot expect a Cat. 5 installation to meet all the requirements, and new performance tests must be run to ensure that the cabling supports the new requirements. New cable categories, such as Cat. 6 or Cat 7, could eventually support Gigabit Ethernet, but would require the installation of a new cable system.

In the case of new installations, the category does not matter; it will be necessary to verify the electrical connections in a process often called *cable testing*. This should be performed before analyzing performance.

10.2 Cable Testing

The verification of the electrical connections and cabling errors is called *wiremap testing*. Tests on cabling can go further, using testers with more sophisticated features, such as the ability to measure crosstalk, capacitance, resistance or TDR.



Stopped 00:00:23
Results > Port A (2/3)

Cable PASS

	MDI-0	MDI-1	MDI-2	MDI-3
Fault	OK	OK	OK	OK
Crosstalk	OK	OK	OK	OK
Distance (m.)	0.0	0.0	0.0	0.0
Crossover	MDIX	MDIX	MDIX	MDIX
Polarity	Positive	Positive	Positive	Positive
Skew (ns.)	0	0	0	8
Wiring	Straight	Straight	Straight	Straight

Figure 10.1 Ether.Giga, a wiremap tester suitable for Cat 5 verification of all four twisted pairs.

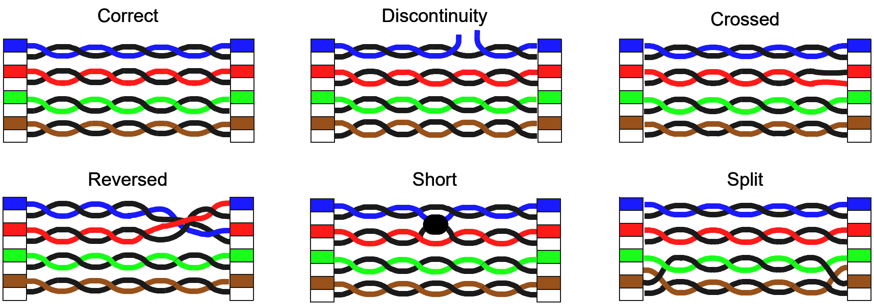


Figure 10.2 Correctly installed, Reversed, Discontinuity, Short, Crossed and Split pairs of four pairs in UTP cabling.

10.2.1 Wiremap

Wiremap is used to identify installation wiring errors. In the case of UTP cabling systems, for each one of the 8 conductors in the link, wiremap should indicate:

- *proper pin termination* at each end;
- *continuity* to the remote end;
- *shorts* between any two or more conductors;
- *crossed pairs* or polarity swap, split pairs, reversed pairs or pair swap;
- *shorted pairs* and any other mis-wiring.

10.2.2 Advanced Cable Testing

By using simple analyzers based on LEDs, it is possible to detect most wiring errors, however these simple analyzers have certain limitations. For example, to detect split pairs, it is necessary to connect the tester at the near end, and terminators at the far end. More sophisticated wiremap testers with capacitance, NEXT or impedance measuring capabilities can perform single-sided tests to detect any wiring errors without using far-end terminators. For example, split pairs can be identified by measuring crosstalk, because split pairs cause a high NEXT (over 20 dB) that limits the available bandwidth on the installed cabling.

The most advanced cabling testers may include a *Time Domain Reflectometer* (TDR) to determine the quality of the cables, connectors, and terminations. Some of the possible problems that can be diagnosed include opens, shorts, cable impedance mismatch, bad connectors, and termination mismatch.

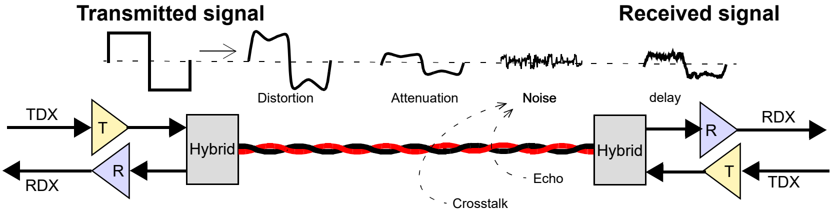


Figure 10.4 Effects of attenuation, distortion, and noise on transmission. Note that impairments are typically reflected as noise.

10.3 Performance Parameters

Wiremap is a basic test, and it does not guarantee that the cabling can support Gigabit Ethernet. It is needed to verify bandwidth performance which relies on frequency-dependent parameters.

Appendix 1 to 5 of ANSI/TIA/EIA-568B and the technical system bulletin (TSB) is the reference taken over from ANSI/TIA/EIA-568A (1995) to ensure that a Category 5 cabling provides a reliable medium for 1000BASE-T. These documents define a transmission model with a characteristic impedance, and a number of parameters. They include the limits of power loss, crosstalk, and delay (see Figure 10.4):

- *Characteristic Impedance* (Z_0) of a line is the resistance it would exhibit if it were infinite in length. Because it is a pure function of parasitic capacitance and inductance distributed along the line, then Z_0 reduces with increasing frequency.

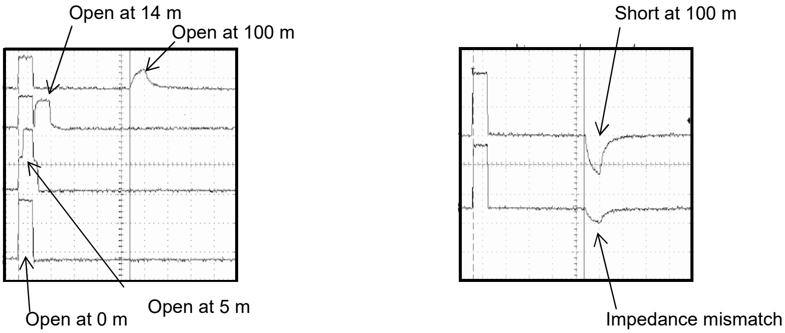


Figure 10.3 TDR shapes for UTP cable testing.

- *Power Loss*. Not all the power transmitted by the source arrives at the destination; a portion is reflected back to the source because of impedance mismatch, another portion is lost as cable losses. The first effect is called *Return Loss* (RL), and the second *Insertion Loss* (IL) or *attenuation*.
- *Crosstalk* (XT) is the noise induced by a disturbing transmitter pair in a victim receiver pair due electromagnetic coupling. Several parameters characterize XT, such as *Near-End Crosstalk* (NEXT), *Far-End Crosstalk* (FEXT) and *Equal-Level Far-End Crosstalk* (ELFEXT).
- *Delay* is the time a signal needs to arrive at the far end. *Delay skew* is the difference in delay between the four pairs transmitting signals simultaneously.

Return Loss and ELFEXT measurements were not required when qualifying links for 10/100BASE-T, but they are now, because 1000BASE-T transmission is bidirectional on a single pair (see Figure 10.5).

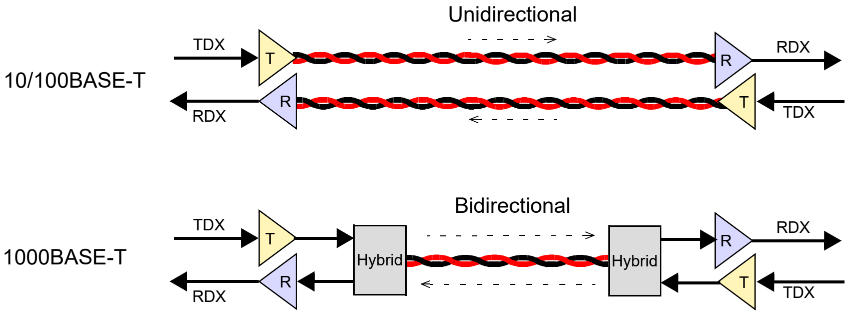


Figure 10.5 Bidirectional and unidirectional transmission.

The bidirectional transmission results in disturbing echo signals that must be removed by the hybrids, to prevent them from being mixed with the local received signals. As result of this, performance meters now include Insertion Loss, Return Loss, NEXT, ELFEXT, Delay and Delay Skew. All of them are a function of frequency and proportional to the cable length.

When qualifying UTP cabling system, a field tester should compare successive readings across the frequency range against a typical pass/fail line. If the measurement line crosses the pass/fail curve (often called a mask) at any point, then the link does not meet the stated requirement see Figure 10.6 and Figure 10.9.

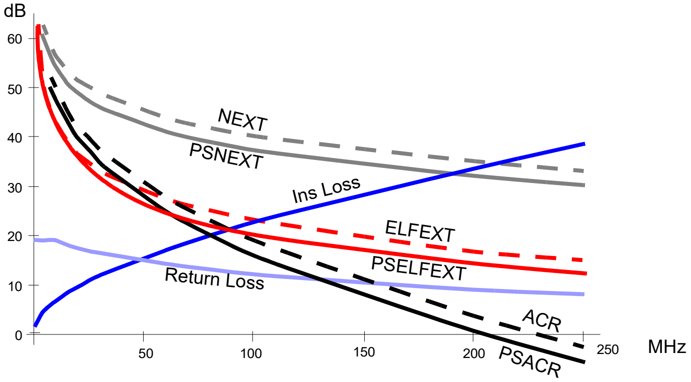


Figure 10.6 Cat 6 cabling mask up to 250 MHz. Includes Parameters Insertion Loss (IL), Return Loss (RL), Near-End Cross Talk (NEXT), Power Sum Next (PSNEX), Attenuation to Crosstalk Ratio (ACR), Power Sum ACR (PSACR), Equal Level Far-End Crosstalk (ELFEXT) and Power Sum ELFEXT (PSELFEXT).

Table 10.1 Cabling Systems

	Category 5	*Category 5e*	*Category 6*	*Category 7*
Bandwidth	100 MHz	100 MHz	250 MHz	600 MHz
Standard	ANSI/TIA/EIA 568A: 1995	TIA/EIA-568-A Addendum 5	TIA TR 42.7.1 and ISO/IEC/SC25/WG3	In development by ISO/IEC/SC25/WG3
Testing	TIA/EIA TSB 67, TIA/EIA TSB 95	Includes return loss and FEXT		

10.3.1 Characteristic Impedance

Characteristic impedance (Z_0) corresponds to the input impedance of a uniform transmission line of infinite length. It also corresponds to the input impedance of a transmission line of finite length that is terminated with its own characteristic impedance.

Z_0 is a function of the frequency of the applied signal, and it is unrelated to the cable length. At very high frequencies, the characteristic impedance asymptotes to a fixed value which is resistive. For example, coaxial cables have an impedance of 50 or 75 Ohms at high frequencies. Typically, twisted-pair lines used in local loops have an impedance above 600 Ohms for telephony, and below 150 Ohms when they are used for xDSL (see Figure 10.8).

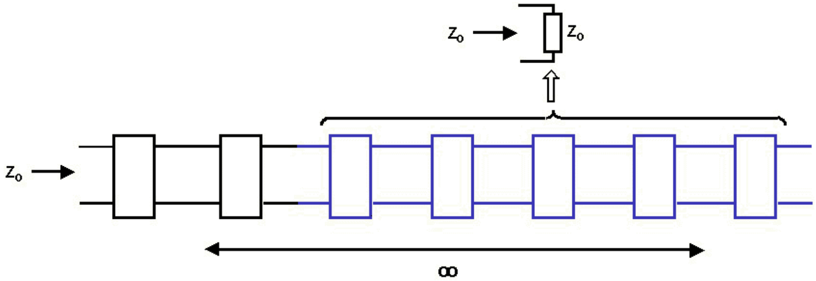


Figure 10.7 Characteristic impedance of a transmission line.

For 1000BASE-T the characteristic impedance of each link, which includes cable cords and connecting hardware, is 100 Ohms for all frequencies between 1 MHz and 100 MHz [1]. This means that tests should be conducted using source and load impedances of 100 Ohms.

The major influence on characteristic impedance is the *capacitance*, which is largely determined by the type of dielectric used. For high frequencies, it can be stated in terms of the physical dimensions of the cable:

$$Z_0 = 276 \log D/r \qquad (Ohm)$$

D is the spacing between the centers of two conductors pair
r is the radius of each conductor

Characteristic impedance is of prime importance for good transmission. Maximum power transfer occurs when the source has the same impedance as the load. Thus for sending signals over a line, the transmitting equipment must have the same characteristic impedance as the line, in order to put the maximum signal into the line. At the other end of the line, the receiving equipment must also have the same impedance as the line, to be able to get the maximum signal out of the line. Cat 5e segment testing must be conducted using source and load impedances of 100 Ohm.

10.3.2 Insertion Loss or Attenuation

When an electrical signal is inserted into a transmission line, only one part of the power arrives at the receiver. One part of the power is reflected to the transmitter; another part turns into ohmic losses; and another part is lost due to outward radiation, especially for very high frequencies.

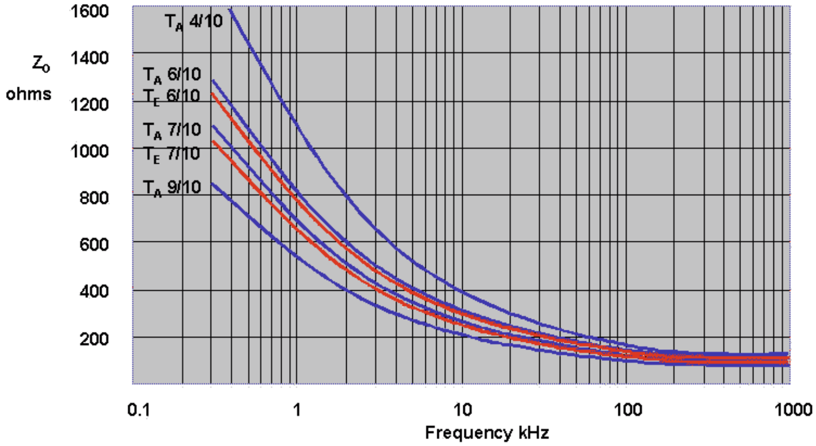


Figure 10.8 Characteristic impedance of twisted pairs of difference gauge.

Insertion loss (L) of a line is a measure of the reduction in signal power due to cabling losses. The simplest way to characterize attenuation is by means of a parameter called attenuation, L, that is obtained from the relationship between the power of the transmitted, P_T, and the received, P_R, signal:

$$L = \frac{P_R}{P_T}$$

However, 1000BASE-T transmission uses sinusoidal signals, and insertion loss is a function of frequency. Insertion loss is therefore to be measured over the frequency range of 1 MHz - 100 MHz.

Excessive length is the most common reason for problems with attenuation, but it can also be caused by poorly terminated connectors/plugs or impurities in the copper cable. If only one or two pairs have high attenuation, this suggests an installation issue. If all pairs have high attenuation, check for excess length. Connection problems and impurities typically occur on one pair only.

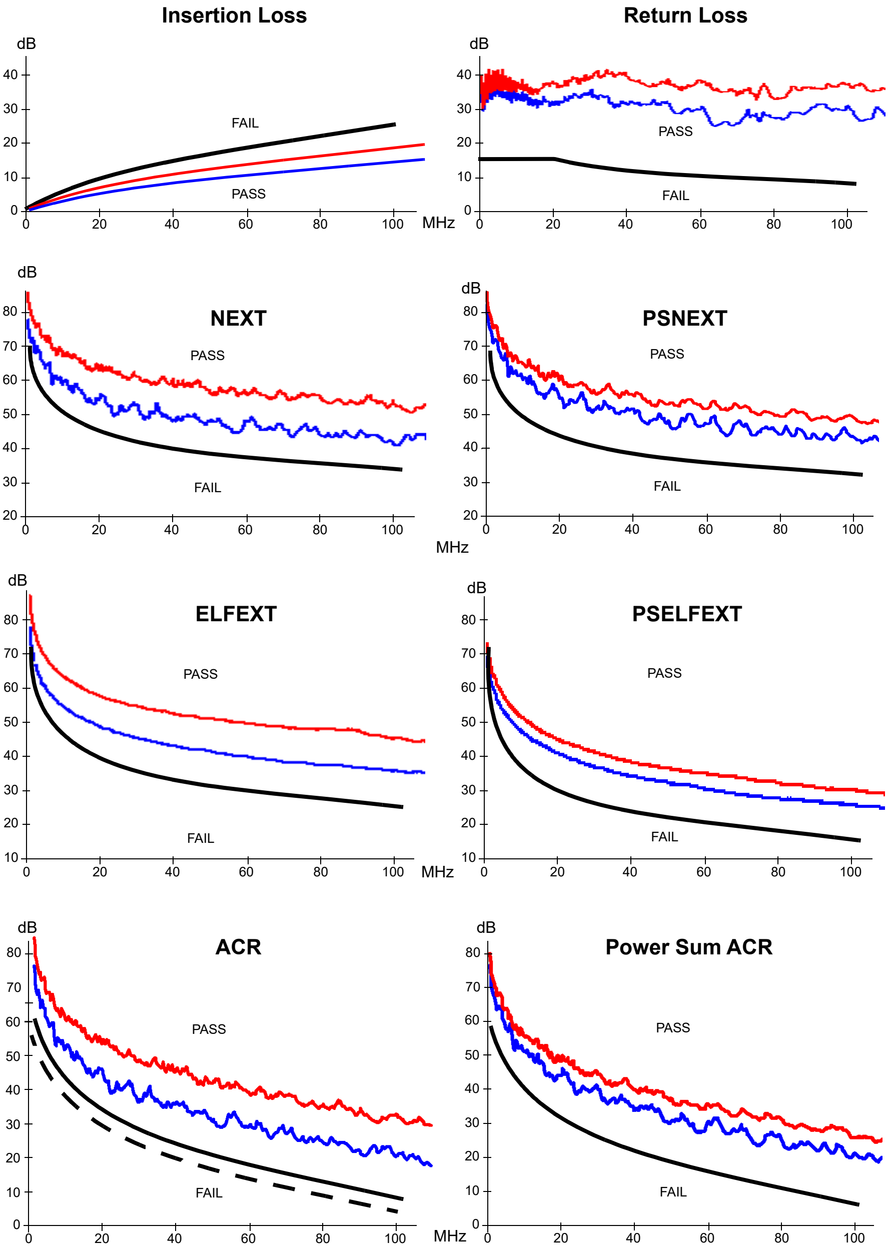


Figure 10.9 Cat 5e mask (black color) and measurement plots (blue worst case, red average). Parameters Insertion Loss (IL) and Return Loss (RL), Near End Cross Talk (NEXT), Power Sum Next (PSNEX), Equal Level Far End Cross Talk (ELFEXT) and Power Sum ELFEXT (PSELFEXT), Attenuation to Crosstalk Ratio (ACR), Power Sum ACR (PSACR).

The standard for 1000BASE-T [1] states link insertion loss shall be less than:

$$InsertionLoss(f) < 2{,}1f^{0{,}529} + 0{,}4/f \qquad (dB)$$

10.3.3 Return Loss

Where line and transmitter impedances do not match, some of the signal is reflected back towards the source as echo that disturbs the receiver-reflected signal; this causes problems and is therefore undesirable.

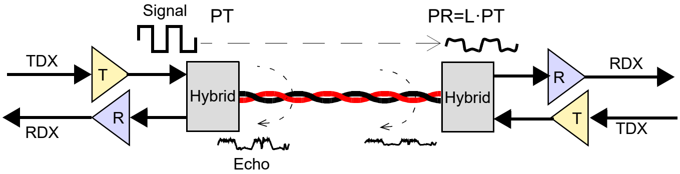


Figure 10.10 Echo and attenuation in Bidirectional Transmission Systems.

Return Loss (RL) of a line is a measure of all reflections caused by impedance mismatches. It is calculated as the ratio of the power reflected back from the line to the power transmitted into the line.

The standard for 1000BASE-T [1] states link requirements for return loss limits

$$ReturnLoss(f) > \begin{Bmatrix} 15 & [1,\ 20MHz] \\ 15 - 10log(f/20) & [20,\ 100MHz] \end{Bmatrix} \qquad (dB)$$

10.3.4 Crosstalk

Crosstalk is analyzed by means of many parameters in which the signal power received at one end is compared to the disturbing power. A disturbing signal generates crosstalk at both ends of the victim line. The end where the power is inserted is called *near end*, and the opposite is the *far end* (see Figure 10.13).

- *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. NEXT is independent of line length.

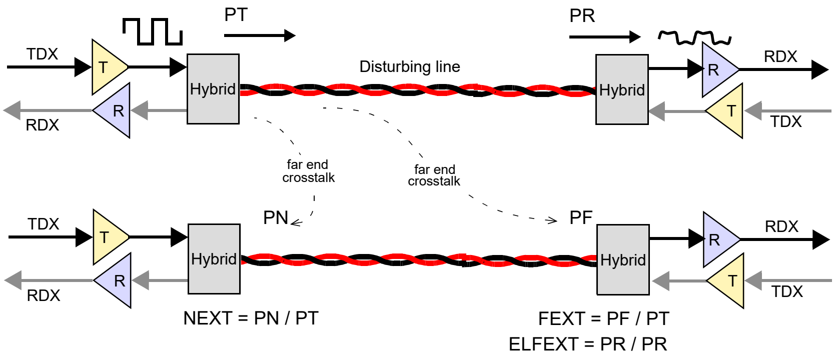


Figure 10.11 Near-End Crosstalk (NEXT), Far-End Crosstalk (FEXT) and Equal-Level Far-End Crosstalk (ELFEXT).

- *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. FEXT depends on the line length.
- *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. Unlike NEXT, FEXT depends on the length of the pairs in question: the longer the pairs, the less FEXT there is. However, this may be misleading, as the useful signal is also affected by the same attenuation factor.

$$NEXT = \frac{P_N}{P_T} \qquad FEXT = \frac{P_F}{P_T} \qquad ELFEXT = \frac{P_F}{P_R} = \frac{P_F}{LP_T} = \frac{FEXT}{L}$$

P_T: power of the transmitted signal
P_R: power of the signal received at the far-end
L: attenuation
P_N: disturbing power received by the victim pair at the near-end
P_F: disturbing power received by the victim pair at the far-end of the transmitter

Note that FEXT is not a required parameter for Cat 5e, but it makes it possible to calculate ELFEXT.

10.3.4.1 *Near-End Crosstalk* (NEXT)

NEXT varies significantly with frequency, and it is therefore important to measure it across a range of frequencies, typically 1 – 100 MHz.

Often, excessive crosstalk is due to poorly twisted terminations at connection points. Since NEXT characteristics are unique to each end of the link, six NEXT results should be obtained per line, and 6 x 4 = 4 per UTP link! Nevertheless, testers can simplify the qualification task by reporting the worst case of NEXT, the average and the Power Sum (PSNEXT) (see Figure 10.11).

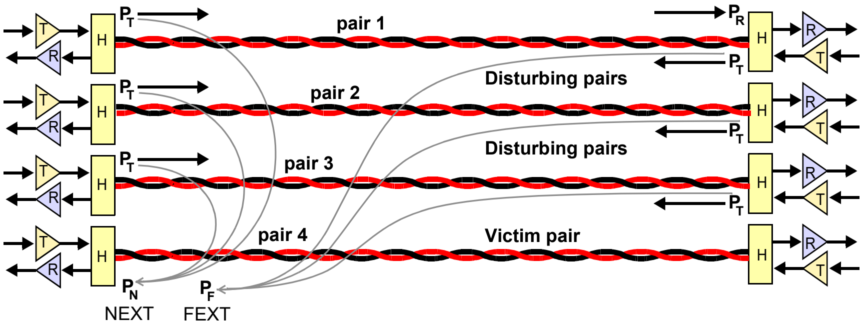


Figure 10.13 Due to electromagnetic coupling, not only the line receives the signal power at the far end, but the victim lines as well, both at the same end where the disturbing power is inserted (NEXT) and at the opposite end (FEXT).

The first thing to do in the event of a NEXT failure is to use the field tester to determine at which end the NEXT failure occurred. Once this is known, check the connections at that end and replace or re-terminate as appropriate. If this does not appear to be the problem, check for the presence of lower Category patch cords. Another possible cause of NEXT failures are split pairs (see Paragraph 10.2). These can be identified automatically with the wiremap function of your field tester.

A tester with TDR capabilities gives the ability to show the fault by distance, pinpointing the problem. This diagnostic function clearly identifies the cause of the NEXT failure, whether it is the patch cord, connection, or horizontal cable.

Figure 10.12 LANTEK 7 cable tester compliant with Categories 3/5e/6 and 7/ISO F standards, with a frequency range up to 750 MHz.

The standard for 1000BASE-T [1] states that link requirements for NEXT loss between a pair and the other three should be at least:

$$NEXT(f) > 27{,}1 - 16{,}8\log_{10}(f/100) \qquad (dB)$$

10.3.4.2 *Equal Level Far-End Crosstalk* (ELFEXT)

ELFEXT loss is critical when two or more wire-pairs carry bidirectional signals in the same direction, because it is a kind of indication of *Signal-to-Noise Ratio* (SNR).

The standard for 1000BASE-T states link requirements for ELFEXT to limit the crosstalk at the far end and meet the BER objectives. The worst pair ELFEXT loss should be greater than:

$$ELFEXT(f) > 17 - 20\log(f/100) \qquad (dB)$$

ELFEXT is a calculated result rather than a measurement. It is derived by subtracting the attenuation of the disturbing pair from the FEXT that this pair induces in an adjacent pair. This normalizes the results for length.

To ensure that the total FEXT coupled into a pair is limited, multiple disturber ELFEXT loss is specified as the power sum (PSELFEXT) of the three adjacent disturbers, which shall be:

$$PSELFEXT(f) > 14{,}4 - 20\log(f/100) \qquad (dB)$$

Consider if the FEXT is equal to 45 dB and attenuation equal to 11 dB, then:

$$ELFEXT = 45 - 11 = 34 \qquad (dB)$$

10.3.4.3 *Power-Sum Crosstalk* (PSACR)

Every pair receives, simultaneously with its data signal, NEXT and FEXT from the other three pairs (see Figure 10.13). *Power-Sum Crosstalk* (PSACR) is the combination of all the interferences received in a pair and measured at both ends. To meet Cat 5e requirements, the worst PSACR shall be limited (see Figure 10.15) according to the TIA/EIA 568-A.

This also guarantees *Collision Detection* (CD) when half-duplex transmission is used. This is because round trip delay is important to guarantee that collision detection functions properly.

10.3.5 Delay Parameters

10.3.5.1 Delay

It is necessary to guarantee that the data, which has been divided and sent separately across four channels, can be properly reassembled at the receiver side. The propagation delay of the single channels, as well as variation of the delays between the four channels must be limited. The delay limit for frequencies between 2 and 100 MHz must not exceed [1]:

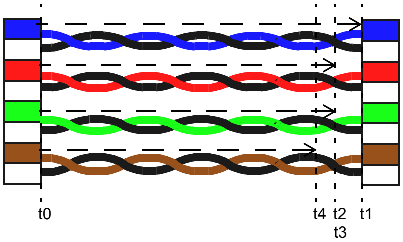


Figure 10.14 Skew is important because Gigabit Ethernet, use all four pairs in the cable. If the delay on one or more pairs is very different to recover the original signal.

$$Delay(f) < 570ns$$

10.3.5.2 Propagation Delay Skew

Different types of insulation materials on each pair produces variation in delays. *Propagation Delay Skew* is the difference between the delay in the fastest and slowest pairs in a UTP cable. Reception buffers re-synchronize the four signals, however there is a limit that cannot be broken over 100 m links [1]:

$$DelaySkew < 50ns$$

10.4 Countering Transmission Impairments

The transmit signals in 1000BASE-T are subject to impairments introduced by the cabling and external noise sources (see Figure 10.4). To operate reliably, the impairments to the transmit signal need to be controlled. The ratio between the impairments, which are generally manifested as noise, and the transmit signal, shall be maintained to achieve an acceptable Bit Error Rate (BER).

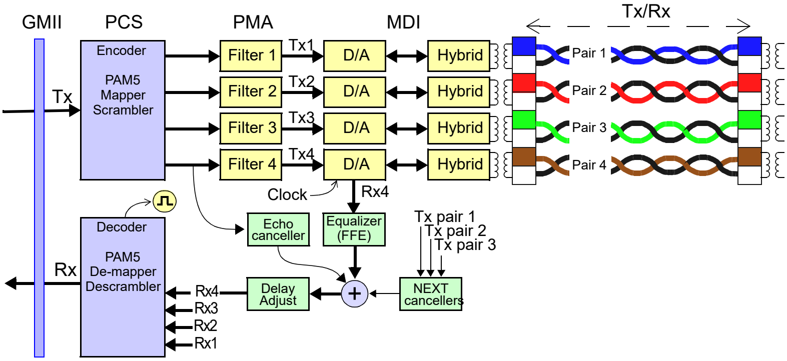


Figure 10.15 1000BASE-T circuits to counter transmission disturbances.

1000BASE-T uses several technologies to reduce disturbances. *Digital Signal Processing* (DSP) is used to cancel crosstalk and echo (return loss). To overcome attenuation and distortion, each frequency band of the signal must be equalized or amplified. Attenuation is compensated at the receiver by the equalization of the signal. Delay skew can be cancelled using buffers (see Figure 10.15).

Selected Bibliography

[1] IEEE 802.3-2002, *Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method and physical layer specifications*, March 2002

[2] Transmission Performance Specifications for 4 pair 100 Ohms Category 5e cabling

[3] Alan Keene, *Qualifying Copper*, Trend Communications, 2001.

[4] Francisco Hens and José Caballero, *SDH/SONET, ATM, xDSL, and Synchronization. Installation and Maintenance*, Aug 2003, Artech-House in the US. ISBN: 1-58053-525-9.

[5] TIA/EIA Telecommunications Systems Bulletin, TSB67, “Transmission Performance Specifications for Field Testing of Unshielded Twisted-Pair Cabling Systems”, October 1995

[6] ANSI/TIA/EIA-568-A, "Commercial Building Telecommunications Cabling Standard", October 6, 1995

[7] ISO/IEC 11801, "Generic Cabling for Customer Premises", 1995

[8] TIA TSB XX (draft) “Additional Transmission Performance Specifications for 100 W 4-Pair Category 5 Cabling”, 5/98

[9] TIA-568-A Addendum 5 (presently under ballot) "Additional Transmission Performance Specifications for 4-Pair 100 W Enhanced Category 5 Cabling", February 26, 1998

[10] Contribution to TIA TR41.8.1 UTP Systems Task Group, “Far End Crosstalk of Cat 5 Connecting Hardware”, 6/97, by Sterling Vaden of Superior Modular Products

[11] Contribution to TIA TR41.8.1-97-08-46, “Relationship [Between] Near End and Far End Crosstalk in Modular RJ-45 Connectors”, 8/97, by Henriecus Koeman & Andrew Bennett of Fluke

[12] ASTM D 4566-94, “Standard Test Methods for Electrical Performance Properties of Insulations and Jackets for Telecommunications Wire and Cable”, 8/94, Section 24

[13] G. Ungerboeck, "Trellis-Coded Modulation with Redundant Signal Sets, Part I and II", IEEE Communications Magazine, vol.25, no.2, 2/87

Chapter 11

Higher-Layer Testing

11.1 Testing Tools

There are many hardware and software tools that can carry out all types of tests on the higher layers based on TCP/IP. These tests vary from simple connectivity tests such as IP ping, up to detailed traffic statistics and tracing.

In this chapter we will discuss some standard ways to analyze packets. We will start by having a look at the ICMP protocol and see how to take advantage of it by using ping and trace route. We will also have a look at two programs; *Ethereal*, an open-source solution for packet capture; and *Observer*, a network monitoring and LAN troubleshooting tool from Network Instruments.

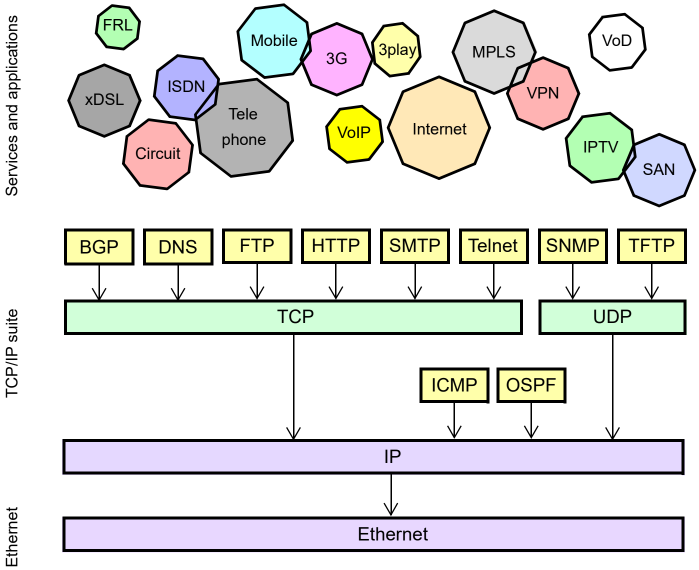


Figure 11.1 TCP/IP protocol suite and applications

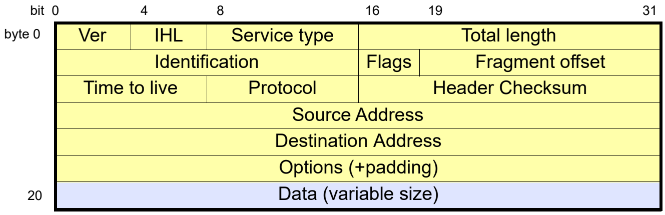


Figure 11.2 IP packet format

11.2 ICMP Analysis

The *Internet Control Message Protocol* (ICMP) is an ‘auxiliary’ protocol that works closely together with the TCP/IP protocol. It is used for error reporting and analysis, transferring messages (not data!) from routers and stations, and for reporting network configuration and performance problems.

11.2.1 What Does ICMP Do?

The ICM protocol is basically used to find out if the part of the network you are connected to is working well or not. You will, naturally, need this information so that you could transfer data (by using TCP, for instance).

The ICMP is used to:

- *Report network errors* – If a part of the network cannot be reached due to a failure.
- *Inform about network congestion* – When a router starts to buffer too many packets, not being able to send them as fast as they are being received, it will generate ICMP Source Quench messages and transmit them to the sender.
- *Help in troubleshooting* – ICMP supports an Echo function, where a packet is sent on a round trip between two hosts. This feature is the basis of the ping function, for example. (Ping is discussed in more detail in Paragraph 11.2.2.1 on page 187.)
- *Announce time-outs* – If an IP packet's TTL field is set to zero, the router discarding the packet will send an ICMP message announcing this. Trace route is a tool used to map network routes, and it is described more thoroughly in Paragraph 11.2.2.2 on page 189.

11.2.2 ICMP Formats and Protocols

ICMP messages are used by routers, hosts and intermediary devices. The messages are sent within an IP packet by encapsulating the ICMP information into the IP packet. More concretely, an ICMP header is added to a normal IP header, with the Protocol field of the latter set to 1, to inform that what follows is an ICMP datagram.

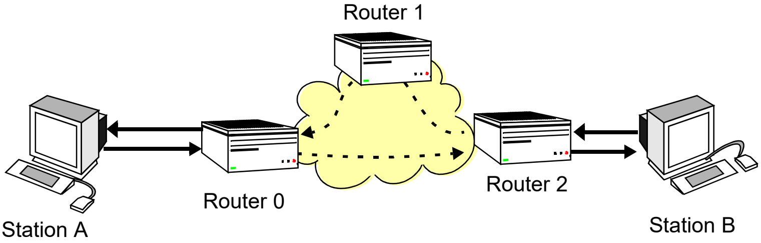


Figure 11.3 ICMP messages generated by Router 1, in response to a message sent by Station A to Station B via Router 0. The ICMP message is returned to Station A, as it is the source address specified in the problematic IP packet.

All ICMP messages contain the following fields:

- *Type* – to describe the type of the ICMP message in question
- *Code* – for additional information on the message type, to help identify the problem that is occurring
- *Checksum* – calculated from the ICMP packet

The most common ICMP messages are described in the following list. Each message has its own number that is used to fill the Type field of the header (see Table 11.1). ICMP messages include the following:

- *Echo request/reply* – An echo request, also known as *ping,* is the most common ICMP message for testing IP connectivity (see Paragraph 11.2.2.1 on page 187). This request must be answered with an echo reply.
- *Destination unreachable* – If the packet cannot reach its destination, the router informs the source that a problem has occurred during delivery. You will need to analyze traffic to know the cause of these messages, and, of course, you will need to figure out a solution, to make these messages disappear from your network.
  There are several types of 'destination unreachable' messages, and the Code field is used to identify which type of message we are dealing with. The following types tend to be the most common:
  Network Unreachable: This message is sent if the router does not know how to

get to the requested network.
Host Unreachable: This message is sent if the router can 'see' the requested network, but not the destination node.
Protocol Unreachable: This message is generated if the destination host is not running UDP or TCP.
Port Unreachable: This message occurs if within the destination host a particular service is not running.
Fragmentation needed: This message is sent if the router needs to fragment a packet, but the *Do not fragment* (DF) bit is found in the IP header.

Table 11.1
Type fields for ICMP messages according to RFC 792

Type	*Description*
1	Echo reply
3	Destination unreachable
4	Source quench
5	Redirect message
8	Echo request
11	Time exceeded
12	Parameter problem
13	Timestamp request
14	Timestamp reply
15	Information request
16	Information reply
17	Address mask request
18	Address mask reply

- *Time exceeded* – A host will send this message to indicate that the packet has been discarded, because it has taken too long for it to be delivered.
- *Source quench* – This message is a request for the source to slow down its transmission rate. It is used to indicate that the source is sending data too fast for the receiver to process it.
- *Redirect* – A redirect message is generated by an intermediary device to tell the host that there is a more direct route for the packet being transmitted.
  As a consequence, the host will update its routing tables.
- *Parameter problem* – This message is sent as a reply to an error that is not covered by any other ICMP message.
- *Timestamp request/reply* – This method measures packet delay between routers, server and stations. Basically, one system is just asking the other system to express the current time for synchronization purposes. (The current time is expressed in milliseconds since midnight in co-ordinated universal time, UTC.)

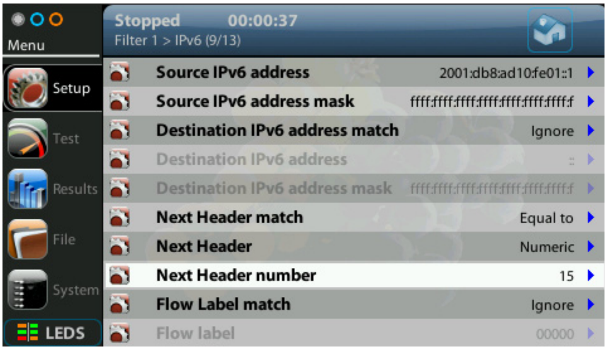
Note that this mechanism is not considered a very reliable synchronization method.

- *Information request/reply* – These messages are a way for a host to discover the IP address of the network it is on. The module that receives the request must reply to it with its full IP address. Note that this method is becoming obsolete, and other methods, such as *Bootstrap protocol* (BOOTP) or *Dynamic Host Configuration Protocol* (DHCP) are used instead.
- *Address mask request/reply* – Request by and reply to a host, to determine the correct subnet mask to be used. (A subnet mask will give us information on how the network is divided.)

These ICMP messages are presented graphically in Figure 11.4.

11.2.2.1 Ping

The *ping* functionality can be used to check if an end-to-end Internet path is operational, i.e., if two devices can communicate with each other or not. It is also used to collect performance statistics; the measured round trip time and the number of times the remote server fails to reply.



Stopped 00:00:37
Filter 1 > IPv6 (9/13)

Source IPv6 address	2001:db8:ad10:fe01::1
Source IPv6 address mask	ffff:ffff:ffff:ffff:ffff:ffff:ffff:f
Destination IPv6 address match	Ignore
Destination IPv6 address	::
Destination IPv6 address mask	ffff:ffff:ffff:ffff:ffff:ffff:ffff:f
Next Header match	Equal to
Next Header	Numeric
Next Header number	15
Flow Label match	Ignore
Flow label	00000

Figure 11.5 IP configuration with Ether.Genius

Each time an ICMP Echo reply message is received, the ping program displays a line of text. This text shows the received sequence number and the measured round trip time (in milliseconds). Each ICMP Echo message contains a sequence number (starting at 0) that is incremented after each transmission, and a timestamp value indicating the transmission time.

Typically, ping is used for short-term connectivity testing, which is why you wouldn't normally see many pings on a network. (A high number of pings may mean that somebody is doing a *ping scan* on the network, and this is not a good sign at all, because it probably means that a hacker is trying to map your network.)

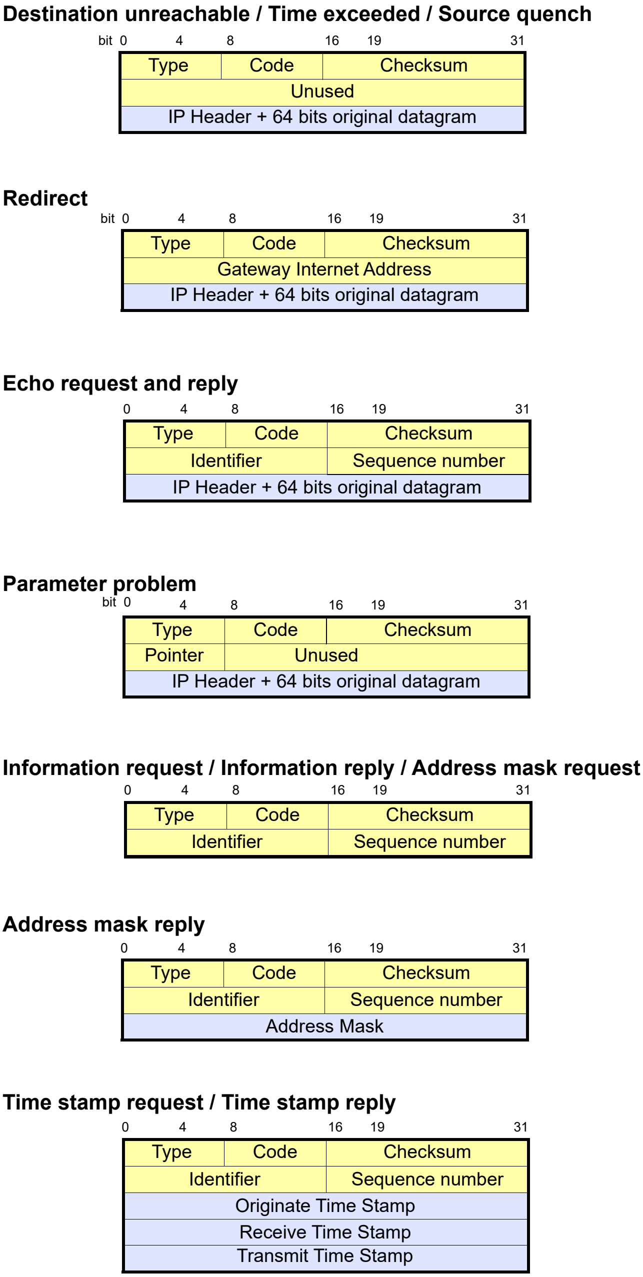


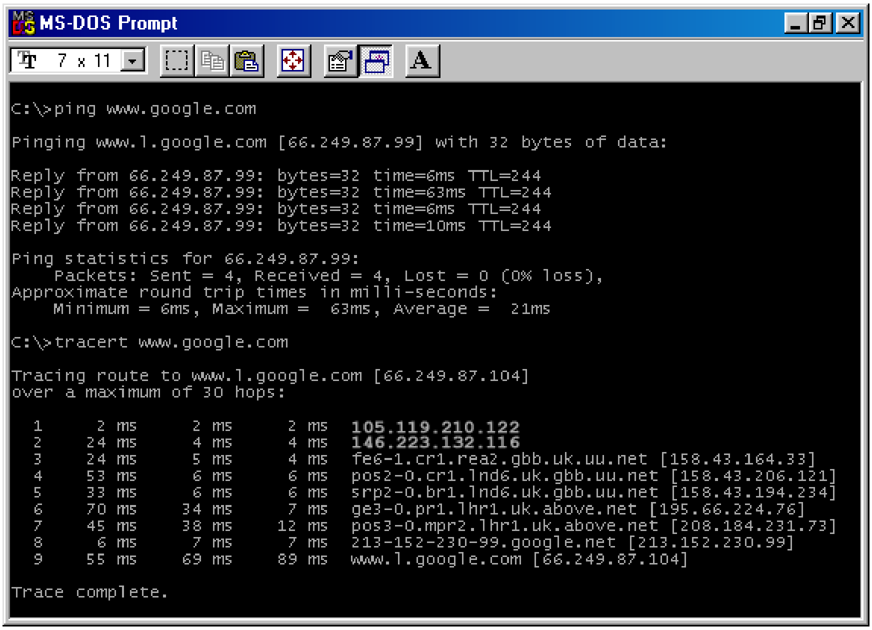
Figure 11.4 ICMP message formats

11.2.2.2 Trace Route

A trace route program is used in TCP/IP to check that two devices can communicate, and through which exact IP routers this communication is being carried out.

The trace route command (`tracert`) displays the routes that IP datagrams take to arrive to their destinations. It also provides information about the time required to send a packet and get a response from every hop. This time is called *Round Trip Delay* (RTD), and it gives an approximate idea of the delay of the link.

Trace route sends out a sequence of *User Datagram Protocol* (UDP) packets from the router to an invalid port address on the destination. For the first sequence of packets, the *Time to Live* (TTL) field value is set to 1. This causes the datagram to time out at the first router in the path. The first hop in the path then responds with ICMP TEMs (*Time Exceed Message*), indicating that the packets have expired.



```
C:\>ping www.google.com

Pinging www.l.google.com [66.249.87.99] with 32 bytes of data:

Reply from 66.249.87.99: bytes=32 time=6ms TTL=244
Reply from 66.249.87.99: bytes=32 time=63ms TTL=244
Reply from 66.249.87.99: bytes=32 time=6ms TTL=244
Reply from 66.249.87.99: bytes=32 time=10ms TTL=244

Ping statistics for 66.249.87.99:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 6ms, Maximum =  63ms, Average =  21ms

C:\>tracert www.google.com

Tracing route to www.l.google.com [66.249.87.104]
over a maximum of 30 hops:

  1     2 ms     2 ms     2 ms  105.119.210.122
  2    24 ms     4 ms     4 ms  146.223.132.116
  3    24 ms     5 ms     4 ms  fe6-1.cr1.rea2.gbb.uk.uu.net [158.43.164.33]
  4    53 ms     6 ms     6 ms  pos2-0.cr1.lnd6.uk.gbb.uu.net [158.43.206.121]
  5    33 ms     6 ms     6 ms  srp2-0.br1.lnd6.uk.gbb.uu.net [158.43.194.234]
  6    70 ms    34 ms     7 ms  ge3-0.pr1.lhr1.uk.above.net [195.66.224.76]
  7    45 ms    38 ms    12 ms  pos3-0.mpr2.lhr1.uk.above.net [208.184.231.73]
  8     6 ms     7 ms     7 ms  213-152-230-99.google.net [213.152.230.99]
  9    55 ms    69 ms    89 ms  www.l.google.com [66.249.87.104]

Trace complete.
```

Figure 11.6 Ping and Trace Route to www.google.com

For testing the remaining route hops, trace route generates a sequence of UDP packets with increasing TTL to the destination, and gets ICMP TEMs from every intermediate hop of the path. This process continues until the packets reach the destination or packets with the maximum TTL are transmitted.

Since these datagrams are trying to access an invalid port at the destination host, ICMP Port Unreachable messages are returned instead of ICMP TEMs when the packets arrive to their destination. As a consequence of this, the trace route program exits.

The TTL value plays a very important role in this process. When the first Echo message is sent, the TTL value in the IP header is set to 1. The first system receiving the message decrements this value, and if the result is 0, as it obviously is in this case, the message is discarded and the above-mentioned ICMP error message is sent.

Each time the Echo message is repeated and sent through the same path, the TTL value is incremented by one, and each time a system that is not the intended destination finds itself setting this value to 0, it sends an ICMP error message to the sender. This process is repeated until the sender gets a response from the intended destination, or when the maximum TTL value is reached.

There may be the case of a router that cannot respond with an ICMP message, because some routers are not configured to do so. In this case, the trace route program indicates that the router is not responding.

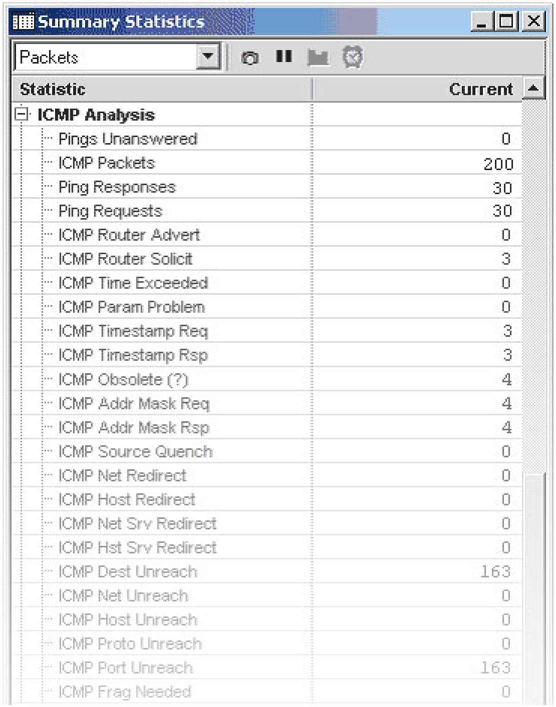
11.2.3 Address Resolution Protocol

The *Address Resolution Protocol* (ARP) is a method used to find out a host's Ethernet (hardware) address from its Internet (protocol) address. ARP is a request-response method, where the sender transmits a packet that contains the Internet address of the destination device, and the destination device replies, sending back its Ethernet or hardware address.

If an ARP is not responded, this naturally means that the destination device is not responding - perhaps because it is not working, or because the network mask has not been correctly configured.

If a station suddenly sends a large number of ARP requests, this probably means that an ARP scan is being carried out. This may be due to an application that is not set up correctly, but it may also mean that a hacker is scanning your network, or that a network virus is trying to attack your system.

Traffic capture and analysis are needed so that we could find out what is really going on.



Statistic	Current
ICMP Analysis	
Pings Unanswered	0
ICMP Packets	200
Ping Responses	30
Ping Requests	30
ICMP Router Advert	0
ICMP Router Solicit	3
ICMP Time Exceeded	0
ICMP Param Problem	0
ICMP Timestamp Req	3
ICMP Timestamp Rsp	3
ICMP Obsolete (?)	4
ICMP Addr Mask Req	4
ICMP Addr Mask Rsp	4
ICMP Source Quench	0
ICMP Net Redirect	0
ICMP Host Redirect	0
ICMP Net Srv Redirect	0
ICMP Hst Srv Redirect	0
ICMP Dest Unreach	163
ICMP Net Unreach	0
ICMP Host Unreach	0
ICMP Proto Unreach	0
ICMP Port Unreach	163
ICMP Frag Needed	0

Figure 11.7 The ICMP analysis section can help identify configuration problems.

11.3 Network Analysis

Network analysis is defined as “the process of capturing network traffic and inspecting it closely to determine what is happening on the network.” [1] Network analysis may also be called traffic analysis, protocol analysis, or packet sniffing.

Traffic can be analyzed for both legal and illegal purposes. Legal traffic analysis may be carried out for instance to monitor the performance of the network, verify its security, log network traffic, or see if ARP or ping scans are being carried out by unwanted visitors (hackers). The aim of an illegal analysis may be to capture network information or passwords for hacking purposes, or get access to confidential information.

11.3.1 Before Starting

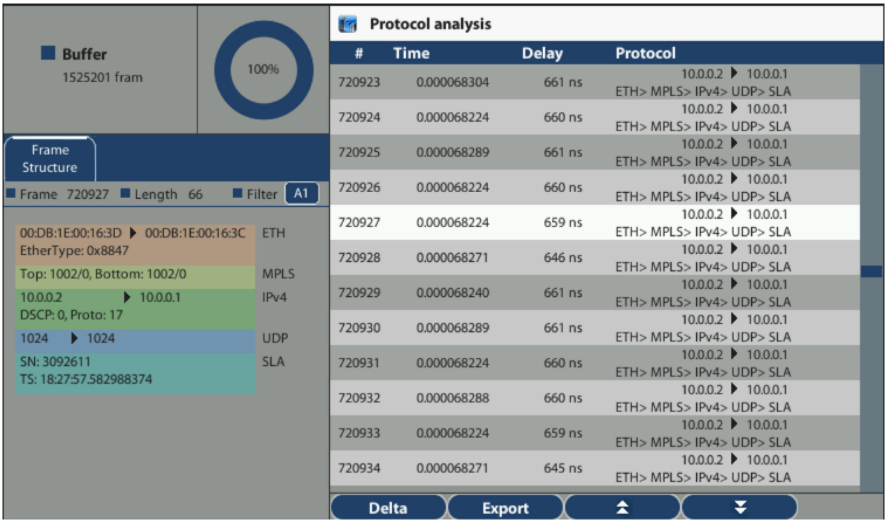
Today's network analysis tools tend to be so easy to use that you do not need to be an expert in packet sniffing to use them.

However, to know how to interpret the results given by your network analysis software, you do need to understand some basic concepts about network communications in general and the protocols used in particular. You will also need to know the server, IP address and application information specific to your network, and, of course, know how your network is structured.

11.3.2 What is Packet Capture?

Packet capture means collecting packets without modifying them in any way. So, network traffic is captured, monitored and logged, but it remains unaltered. Packets are captured as binary data, and a *network analyzer/packet sniffer* is needed, first to capture the data and then to convert this information into a readable form.

Packets are captured in *promiscuous mode*, which means that the computer you are using for packet sniffing will see *all* the traffic that is being sent across the network you are on, not just the traffic that is being sent to you. (You would normally be running your computer in non-promiscuous mode, only receiving the information that is being sent to you.)



Frame 720927 Length 66 Filter A1

00:DB:1E:00:16:3D ▶ 00:DB:1E:00:16:3C ETH
EtherType: 0x8847
Top: 1002/0, Bottom: 1002/0 MPLS
10.0.0.2 ▶ 10.0.0.1 IPv4
DSCP: 0, Proto: 17
1024 ▶ 1024 UDP
SN: 3092611 SLA
TS: 18:27:57.582988374

#	Time	Delay	Protocol
720923	0.000068304	661 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720924	0.000068224	660 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720925	0.000068289	661 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720926	0.000068224	660 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720927	0.000068224	659 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720928	0.000068271	646 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720929	0.000068240	661 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720930	0.000068289	661 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720931	0.000068224	660 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720932	0.000068288	660 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720933	0.000068224	659 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA
720934	0.000068271	645 ns	10.0.0.2 ▶ 10.0.0.1 ETH> MPLS> IPv4> UDP> SLA

Figure 11.8 MPLS test with xGenius.

11.3.3 Who Captures What, and Why?

Network analyzers are used by security engineers, system administrators, operators–and, unfortunately, also by hackers.

- When a network analyzer is used for legal purposes, the most typical aim is to use it for troubleshooting problems on the network, or for security purposes, for example to detect unauthorized intrusions, spyware, ping scans, and so on.
- When you are capturing packets for legal purposes, you are interested in information such as 'top talkers', traffic to a certain Internet address, traffic coming from a specific IP address, or any other security/troubleshooting-specific data.

Troubleshooting and security are the most common reasons for packet capture, and they are *very* important reasons for any organization, but if you work for law enforcement officials, you might also use a network analyzer in crime investigation[1]. As there are many freeware tools available for packet sniffing, home users may want to use these tools to monitor their Internet connection, or 'just out of curiosity'. These tools are an excellent way to learn what Internet traffic is all about, which is why they are also used for educational purposes.

Whatever your needs and motives for packet capture are, you will need a network analyzer to accomplish this task. There are many commercial solutions available, but if they do not support your operating system, or if you cannot find a solution that adapts to your budget, you may also want to consider using a freeware sniffer.

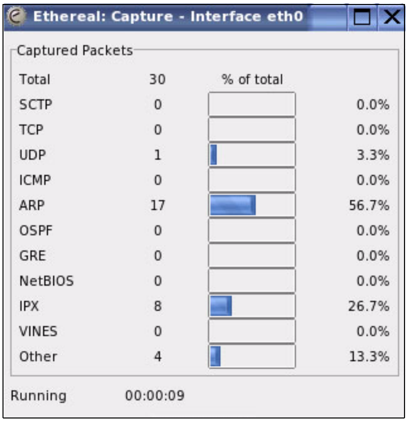
One of the most common freeware solutions is Ethereal, which will be briefly discussed in the following.

11.4 WIRE.SHARK – A FREEWARE SOLUTION

WireShark is an open-source packet sniffer that has originally been developed for UNIX and Linux, but it can also be used for Windows platforms by installing a special 'driver' called WinPcap.

[1] Network analyzers are increasingly used in crime investigation, and some law enforcement entities have even developed software of their own to meet their specific needs.

You can use Ethereal for both capture and analysis. Network administrators can use this tool to troubleshoot their network, developers might use it to debug protocol implementations, and if you work in network security, you can use this tool to solve security problems.



Ethereal: Capture - Interface eth0

Captured Packets

Total	30	% of total
SCTP	0	0.0%
TCP	0	0.0%
UDP	1	3.3%
ICMP	0	0.0%
ARP	17	56.7%
OSPF	0	0.0%
GRE	0	0.0%
NetBIOS	0	0.0%
IPX	8	26.7%
VINES	0	0.0%
Other	4	13.3%

Running 00:00:09

Figure 11.9 WireShark Packet capture window

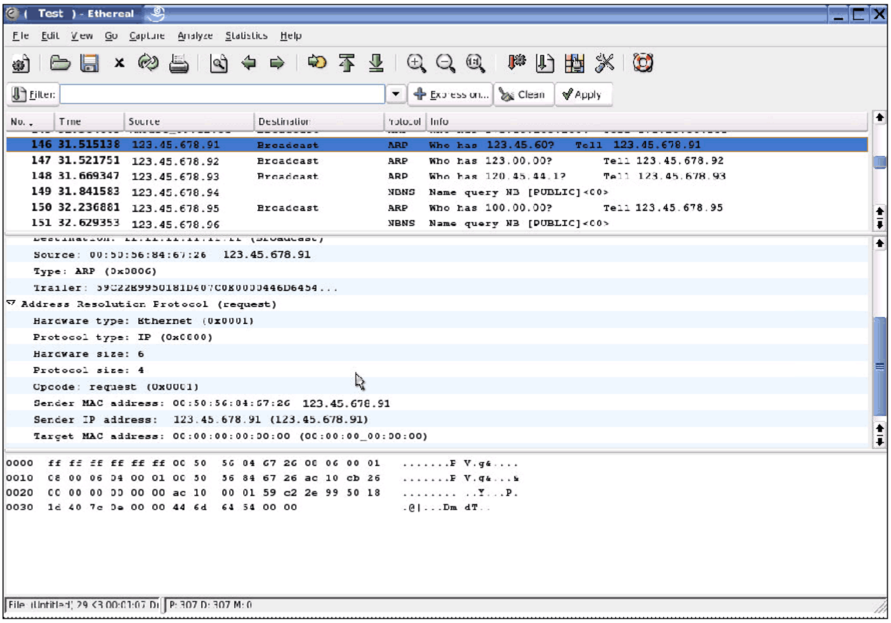
You can capture data directly from a live network, or alternatively read it from capture files. These capture files can originate from *tcpdump* or from a large number of commercial solutions. Ethereal can also decompress *gzip* files. Depending on the platform used, you can read live data from Ethernet, FDDI, PPP, Token Ring, IEEE 802.11, IP over ATM and loop-back interfaces.

You can browse the captured network data via the graphical user interface, or in TTY (*teletype*) mode[2]. You can use a display filter to filter the data you wish to display, highlighting or coloring packet summary information.

Output can be saved or printed as plain text or PostScript®, and each captured network trace can be saved to disk.

Ethereal currently supports more than 600 protocols, and you can download the software (or just the source code, if this is what you prefer) at: www.ethereal.com.

[2] The TTY version of Ethereal is called Tethereal.



```
No. .  Time       Source           Destination   Protocol Info
  146 31.515138  123.45.678.91    Broadcast     ARP      Who has 123.45.60?    Tell 123.45.678.91
  147 31.521751  123.45.678.92    Broadcast     ARP      Who has 123.00.00?         Tell 123.45.678.92
  148 31.669347  123.45.678.93    Broadcast     ARP      Who has 120.45.44.1?       Tell 123.45.678.93
  149 31.841583  123.45.678.94                  NBNS     Name query NB [PUBLIC]<00>
  150 32.236881  123.45.678.95    Broadcast     ARP      Who has 100.00.00?         Tell 123.45.678.95
  151 32.629353  123.45.678.96                  NBNS     Name query NB [PUBLIC]<00>

    Source: 00:50:56:84:67:26  123.45.678.91
    Type: ARP (0x0806)
    Trailer: 59C22B9950181D407C0B0000446D6454...
▽ Address Resolution Protocol (request)
    Hardware type: Ethernet (0x0001)
    Protocol type: IP (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (0x0001)
    Sender MAC address: 00:50:56:04:67:26  123.45.678.91
    Sender IP address:  123.45.678.91 (123.45.678.91)
    Target MAC address: 00:00:00:00:00:00 (00:00:00_00:00:00)

0000  ff ff ff ff ff ff 00 50  56 04 67 26 08 06 00 01   .......P V.g&....
0010  08 00 06 04 00 01 00 50  56 84 67 26 ac 10 cb 26   .......P V.g&...&
0020  00 00 00 00 00 00 ac 10  00 01 59 c2 2e 99 50 18   ........ ..Y...P.
0030  1d 40 7c 0a 00 00 44 6d  64 54 00 00               .@|...Dm dT..
```

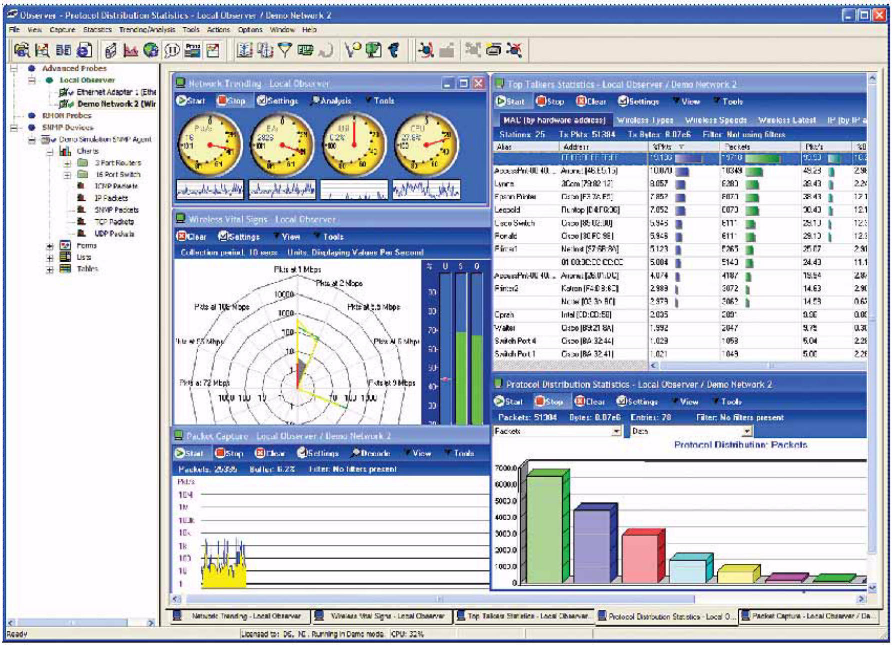
Figure 11.10 WireShark window

11.5 Observer – Network Monitoring and LAN Troubleshooting

Observer is a Windows-based family of network monitoring and LAN troubleshooting tools designed by Network Instruments. This software family includes three different versions: the basic Observer, Expert Observer and Observer Suite, of which the latter is the most complete, as it includes all the features of the basic and expert versions.[3]

Observer can monitor Ethernet (10/100/1000), Wireless 802.11a/b/g, Token Ring, Full Duplex Gigabit and WAN networks, and it provides metrics, capture and trending for both shared and switched environments. It can also be used as an add-on for xGenius Gigabit Ethernet (see Chapter 7 *"Gigabit Ethernet Testing"*), and it is compliant with RMON1 and RMON2, which are the industry standards for traffic management and packet-level data collection for multi-segment LANs.[4]

[3] For simplicity, we use the name Observer in this chapter to refer to all the three versions of the product. If you are interested in more detailed descriptions of each product and the differences between them, check out the manufacturer's website at: www.networkinstruments.com.


Observer - Protocol Distribution Statistics - Local Observer / Demo Network 2
File View Capture Statistics Trending/Analysis Tools Actions Options Window Help
Advanced Probes
Local Observer
Ethernet Adapter 1 (Ethe
Demo Network 2 (Wir
RMON Probes
SNMP Devices
Demo Simulation SNMP Agent
Charts
3 Port Routers
16 Port Switch
ICMP Packets
IP Packets
SNMP Packets
TCP Packets
UDP Packets
Forms
Lists
Tables
Network Trending - Local Observer
Start Stop Settings Analysis Tools
Pkt/s
B/s
Util
CPU
Wireless Vital Signs - Local Observer
Clear Settings View Tools
Collection period: 10 secs Units: Displaying Values Per Second
Pkts at 1 Mbps
Pkts at 2 Mbps
Pkts at 5.5 Mbps
Pkts at 6 Mbps
Pkts at 9 Mbps
Pkts at 72 Mbps
Pkts at 55 Mbps
Pkts at 11 Mbps
10000
1000
100
10
Packet Capture - Local Observer / Demo Network 2
Start Stop Clear Settings Decode View Tools
Packets: 25395 Buffer: 6.2% Filter: No filters present
Pkt/s
Top Talkers Statistics - Local Observer / Demo Network 2
Start Stop Clear Settings View Tools
MAC (by hardware address) Wireless Types Wireless Speeds Wireless Latest IP (by IP a
Stations: 25 Tx Pkts: 51384 Tx Bytes: 8.07e6 Filter: Not using filters
Protocol Distribution Statistics - Local Observer / Demo Network 2
Start Stop Clear Settings View Tools
Packets: 51384 Bytes: 8.07e6 Entries: 78 Filter: No filters present
Packets
Data
Protocol Distribution: Packets
7000.0
6000.0
5000.0
4000.0
3000.0
2000.0
1000.0
0
Network Trending - Local Observer
Wireless Vital Signs - Local Observer
Top Talkers Statistics - Local Observer
Protocol Distribution Statistics - Local O...
Packet Capture - Local Observer / Da...
Ready
Licensed to: DS, NC. Running in Demo mode. CPU: 32%

Alias	Address	%Pkts	Packets	Pkt/s	%B
	FF:FF:FF:FF:FF:FF	19.100	19710	90.00	10.2
AccessPnt-00 40...	Aironet [46:E5:15]	10.070	10349	49.23	2.98
Lynne	3Com [79:82:12]	8.057	8280	38.43	2.24
Epson Printer	Cisco [F3:7A:F5]	7.852	8070	38.43	12.1
Leopold	Dunlop [D4:FG:30]	7.052	C070	30.40	12.1
Cisco Switch	Cisco [88:02:00]	5.945	6111	29.10	12.3
Ronald	Cisco [80:FC:9E]	5.946	6111	29.10	12.3
Printer1	Netinst [57:6B:8A]	5.123	5265	25.07	2.91
	01:00:0C:CC:CC:CC	5.004	5143	24.40	11.1
AccessPnt-00 40...	Aironet [36:01:DC]	4.074	4187	19.94	2.8
Printer2	Katron [F4:D3:6C]	2.989	3072	14.63	2.90
	Nortel [03:36:8C]	2.979	3062	14.58	0.63
Oprah	Intel [CD:CD:58]	2.035	2091	9.96	0.00
Walter	Cisco [B9:21:8A]	1.992	2047	9.75	0.30
Switch Port 4	Cisco [BA:32:44]	1.029	1058	5.04	2.28
Switch Port 1	Cisco [BA:32:41]	1.021	1049	5.00	2.28

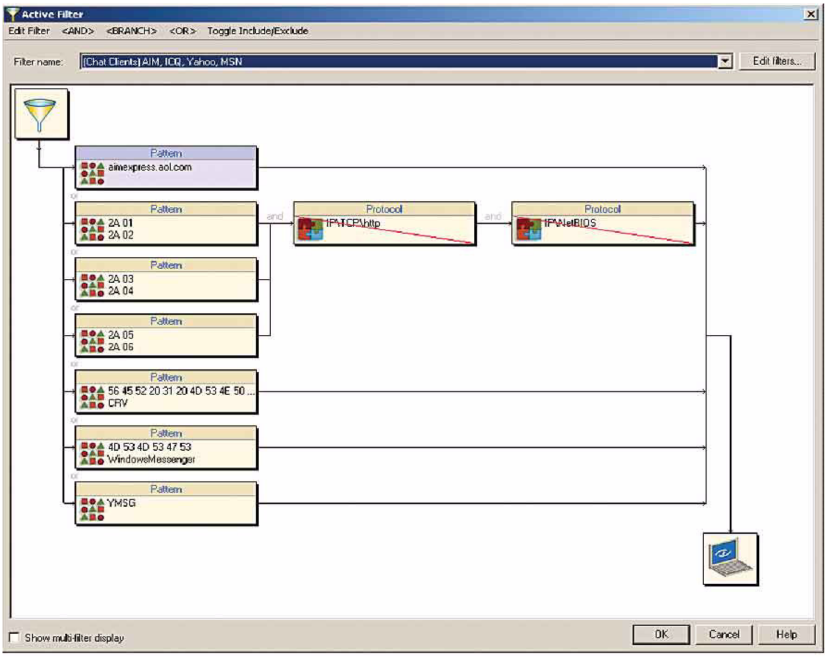
Figure 11.11 Observer's main window

Observer includes decode of more than 500 primary protocols, and if sub-protocols are included, it supports more than 1000 protocols.

Normally, if you wish to sniff traffic, your sniffer has to be on the same wire as the data you wish to analyze. However, some tools, including Observer, can be attached to a remote probe, which makes it possible for you to manage remote networks as if they were local. In the case of Observer, multiple sessions and multiple users are also supported. Observer filters traffic on- and off-line, by MAC/IP address, IP address range, name, protocol type, or by any value at byte offset. Although not open source, this software has some functions that you can customize; for example, you can use Boolean logic to create complex filters, and the decoding functions are customizable in C++.

The software includes a web publishing service for publishing your data and network health reports, and you can control the level of access of each user. For example, you might not want your organization's internal and external employees view the same information, so, you can choose the data each user can view.

[4] According to the manufacturer, all RMON and HCRMON groups are fully supported.



Filter name: [Chat Clients] AIM, ICQ, Yahoo, MSN

☐ Show multi-filter display

Figure 11.12 Observer's Active filter window.

You can download an evaluation version of the software from the manufacturer's website: www.networkinstruments.com.[5]

Selected Bibliography

[1] Angela Orebaugh, *Ethereal Packet Sniffing*. Rockland, MA: Syngress Publishing, Inc., 2004.

[2] *RFC 792,* Internet Control Message Protocol (ICMP)

[3] Rich Seifert, *Gigabit Ethernet Technology and Applications for High/Speed LANs*, Addison Wesley Oct 1999

[4] William Stallings, *Data and Computer Communications*, Prentice Hall, 1997.

[5] Kevin L. Paton, *Gigabit Ethernet Test Challenges*, Oct 2001 Test and Measurement World Magazine

[6] RFC 2544, *Benchmarking Methodology for Network Interconnect Devices*, S. Bradner and J. McQuaid, March 1999

[5] Note that your operating system should be Windows 2000 or XP.

[7] RFC 1242 - *Benchmarking terminology for network interconnection devices,* S. Bradner July 1991

[8] RFC 2285, *Benchmarking Terminology for LAN Switching Devices*, R. Mandeville Network Laboratories February 1998

[9] RFC 792, *Internet Control Message Protocol*, at: http://www.ietf.org/rfc/rfc792.txt

[10] Robert Breyer, Sean Riley, *Switched, Fast and Gigabit Ethernet*, 3rd edition 1999.

[11] www.networkinstruments.com

[12] www.ethereal.com

[13] www.trendcomms.com

Appendix 1

Testing Applications

A1.1 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.

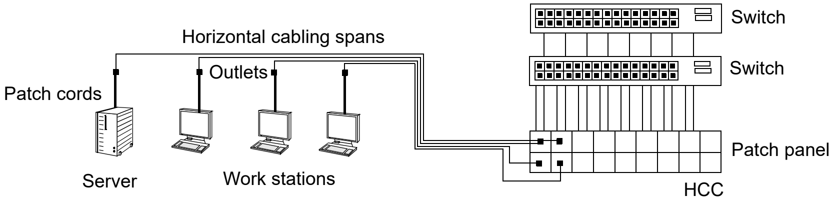


Figure 1.1 Simple, structured cabling layout. All stations are connected to a central location in a star topology.

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.

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 or range and bandwidth, but also to avoid potential earthing/grounding problems.

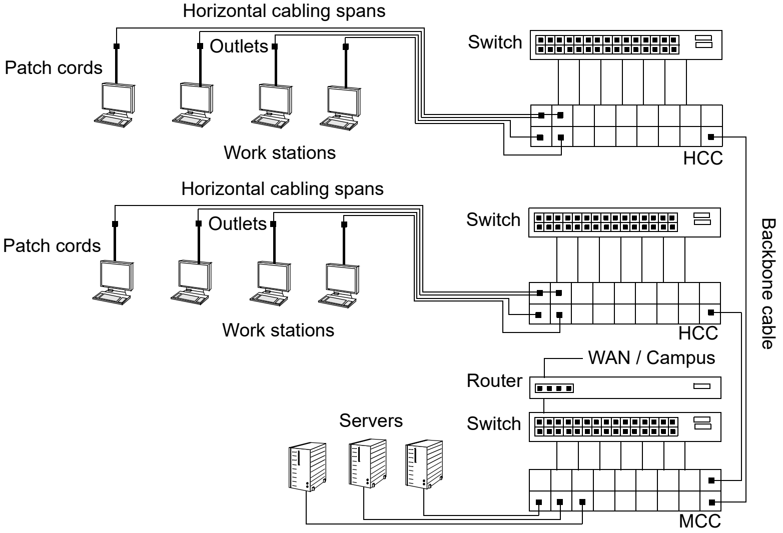


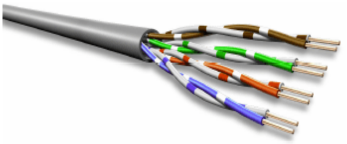
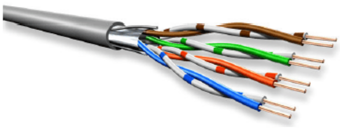
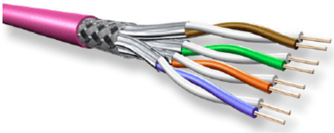
Figure 1.2 Typical cabling layout, including both horizontal and vertical cable spans. The network is structured with an extended star topology centered in the MCC.

If twisted pair is chosen, one must still decide which type of cable should be used (see Table 1.1). 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 1.2). 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 Gbit/s. Ethernet at 1 Gbit/s should also work in existing networks using Cat 5 cables, but not in those that use Cat 3. All common cable categories can be implemented with UTP. In fact, balanced transmission is good enough to protect from external interference, so shielding is not needed.

Cables of different categories are different in many ways. Performancewise, what really makes them different is crosstalk (see Table 1.3). 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 pairs 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.

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

Table 1.1
Common twisted pair types for data networks.

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.

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.

Table 1.2
Twisted Pair Categories.

EIA/TIA Category	*Bandwidth*	*Common Application*
1	-	Telephony, ISDN BRI
2	4 MHz	4 Mbit/s Token Ring
3	16 MHz	Telephony, 10BASE-T, 100BASE-T4 (four wires)
4	20 MHz	16 Mbit/s Token Ring
5	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mbit/s ATM
5e	100 MHz	100BASE-T, 1000BASE-T (four wires), short haul 155 Mbit/s ATM
6	250 MHz	1000BASE-T (four wires)

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 Gbit/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 (see Figure 1.3). All four pairs are always connected, but for 10 and 100 Mbit/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 1.4)..

Table 1.3
Typical specifications of commercially available UTP cables.

Category	*Cat 6*	*Cat 5e*	*Cat 5*
DC resistance (20ºC, 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
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

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.
- *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.

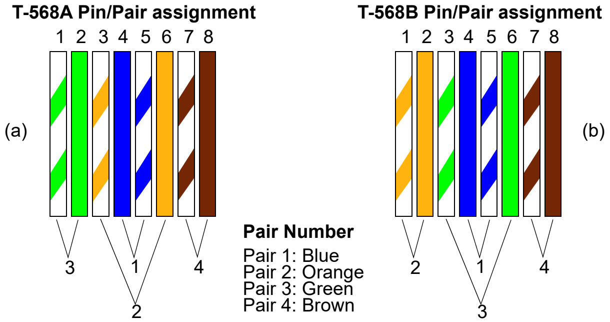


Figure 1.3 T-568A and T-568B wiring standards.

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.

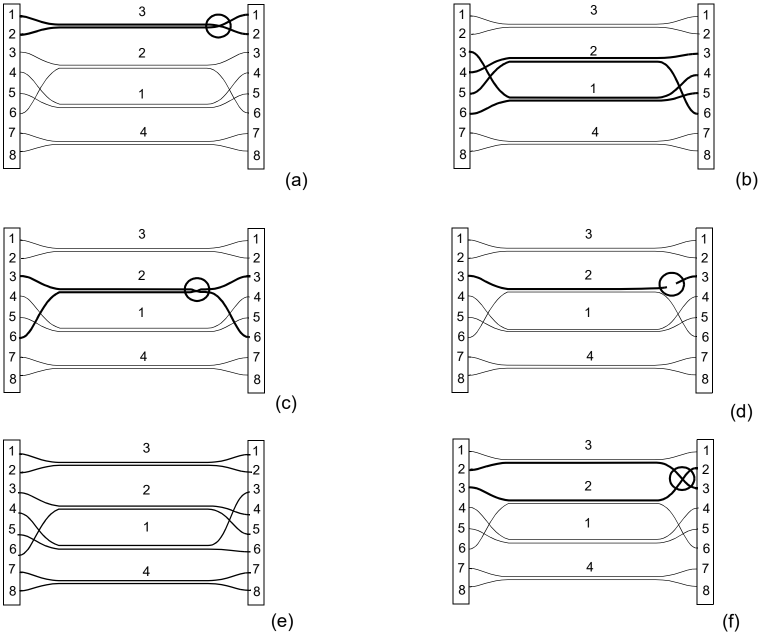


Figure 1.4 Common cabling faults. (a) Inverted cable, polarity is inverted in both ends of the same pair. (b) Pairs 1 and 2 are split. (c) There is a short circuit in pair 2. (d) There is an open circuit in pair 2. (e) Swapped pair, connections are OK but wires of the same connection are twisted in different pairs. (f) Miswired cable.

Another issue addressed by cabling standards is how and where to connect the tester(s) (see Figure 1.5). 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 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 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.

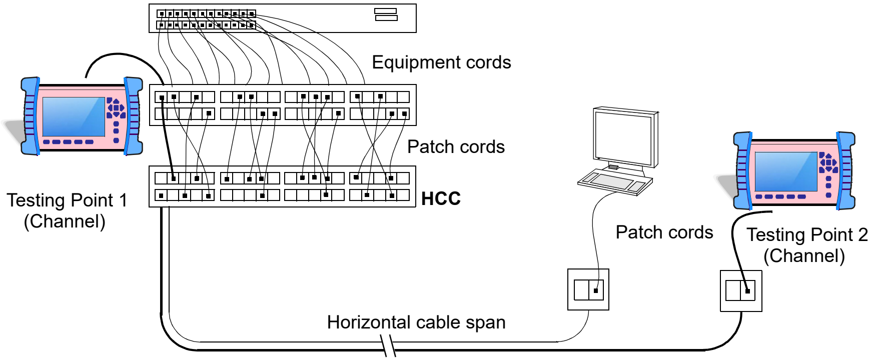


Figure 1.5 Connecting test equipment to the network to check compliance with cabling standards.

A1.2 INSTALLATION AND QUALIFICATION OF OPTICAL FIBER CABLING

Optical fiber is a good replacement for copper in long-haul and high-bandwidth applications. Also, optical fiber has low maintenance needs and it is immune to electromagnetic noise caused by radio sources, motors or nearby cables.

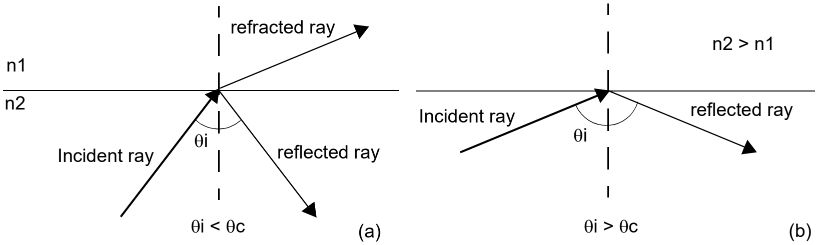


Figure 1.6 The total internal reflection principle. (a) If the incident angle of a ray of light in the boundary of the material is smaller than the critical angle, the power is distributed between a reflected ray and a refracted ray, and no effective propagation is achieved. (b) If the incident ray is bigger than the critical angle, all incident power is reflected back and the light is propagated through the fiber.

Light transmission in optical fibers is based on the total internal reflection principle (see Figure 1.6). Total reflection is possible only if light is reflected from a material with higher refractive coefficient to a material with smaller refractive coefficient. Optical fibers therefore consist of a core and cladding with different refractive indices. The core, of high refractive index, is surrounded by a cladding layer of lower refractive index. The index difference forms a boundary which constrains most of the light within the core.

Earlier, optical fibers were limited by their high attenuation caused by intrinsic absorption, scattering, impurities, physical bending or other factors. Today, improvements in the manufacturing process have brought high-quality, low-attenuation fiber to the market. As a result, the optical loss spectrum of today's fiber is almost free of resonance peaks that may be caused by impurities (see Figure 1.7).

When optical fiber networks are designed, it is important to keep the following in mind: There are many different types of fiber, light sources and receivers available in the market. Some fibers support light transmission over dozens of kilometers, whereas some others can only reach a few meters. The same fiber may work for all possible wavelengths and light sources, but it may be optimized only for some of them. There are two important families of optical fiber:

- *Multimode Fiber* (MMF) has a core with a diameter larger than 10 µm (usually 50 µm or 62.5 µm). This fiber can transport many simultaneous transverse modes. It is usually fed by a *Light Emitting Diode* (LED), but there are some MMFs that work with laser as well.
- *Single Mode Fiber* (SMF) has a core with a diameter less than 10 µm (typically 9 µm). As a result, it only propagates one transverse mode from the optical source. LEDs are not suitable for injecting optical power to SMF, and this is why it uses lasers as optical sources.

Generally speaking, SMF offers better performance than MMF. SMF is used for long-haul applications. It is perfectly suitable for WAN applications, but it is used for LANs and campus networks as well. MMF is often used in LAN backbone cabling. MMF is more expensive than SMF, however, lasers are more expensive than LEDs. This is why, even though SMF is cheaper, SMF-based systems are more expensive.

Optical transmission systems usually work at 850 nm (first transmission window), 1310 nm (second window) and 1550 nm (third window). The operating transmission window depends on the light source, not on the optical fiber. Optical fiber should work at any transmission window, but it can only be optimized for one of them. Common LEDs usually work in the first transmission window, while lasers are built to work in the second and third transmission window. Each transmission window has its own particular properties. For example, minimum attenuation is obtained in the third transmission window. Currently, optical fibers and light sources enable operation in new windows, offering almost unlimited bandwidth.

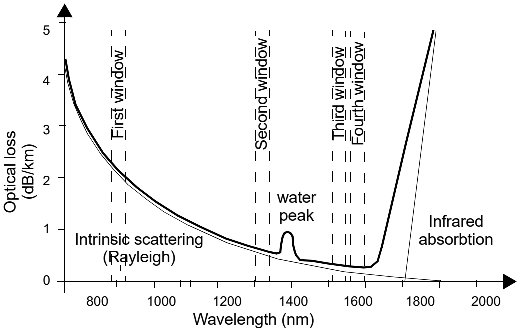


Figure 1.7 Light attenuation in a silica crystal optical fiber and transmission windows.

Each MMF mode has its own propagation properties, including propagation delay. As a result, the fiber filters the incoming signal and transmission pulses are wider when they are received, which causes *Inter-Symbol Interference* (ISI). This effect is known as modal dispersion, and it may be severe for high transmission rates (narrow pulses), thus making signal decoding impossible. Graded-index fibers offer a partial solution to this problem. Refractive index of graded-index fibers decreases gradually away from its center (see Figure 1.15). The result is delay equalization of transmission modes, which allows higher bit rates without ISI.

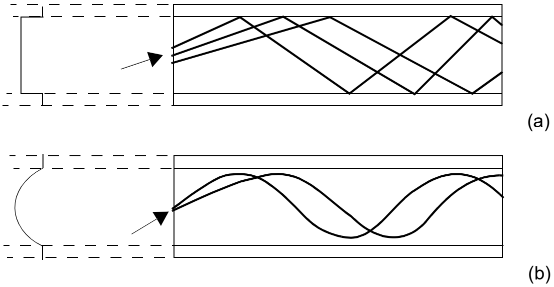


Figure 1.8 Refractive index of MMFs. (a) Step-index MMF. (b) Graded-index MMF.

The ability of a MMF to transport high bit rate signals is specified by means of a distance/bandwidth product known as modal bandwidth. Modal bandwidth is always specified for a specific transmission window. For example, an MMF may have a 1500 MHz*km modal bandwidth operating at 850 nm, and only 500 MHz*km at 1310 nm.

Modal bandwidth is a critical parameter that must be checked when installing MMF. Special care must be taken with 10 GbE applications, because the same fiber used for 1 GbE may not be valid for 10 Gbit/s (see Table 1.4). The preferred choice for 10 Gbit/s applications over MMF is laser-optimized fiber rated with OM3 by ISO/IEC 11801 (see Table 1.5). These MMFs are designed to transport a 10 Gbit/s signal using a *Vertical Cavity Surface Emitting Laser* (VCSEL) that has a better performance than traditional LED sources. VCSELs are a low-cost alternative to common laser sources such as *Fabry Perot* (FP) or *Distributed-FeedBack* (DFB) lasers. Like most LEDs, VCSELs operate in the 850 nm band.

Table 1.4
Performance of MMF at 1 Gbit/s and 10 Gbit/s, with MMF operating at 850 nm with an LED source

Fiber type	*Modal bandwidth*	*1000BASE-SX*	*10GBASE-S*
62.5 μs MMF	160 MHz*km	2 ~220 m	2 ~ 26 m
62.5 μs MMF	200 MHz*km	2 ~275 m	2 ~ 33 m
50 μs MMF	400 MHz*km	2 ~500 m	2 ~ 66 m
50 μs MMF	500 MHz*km	2 ~550 m	2 ~ 82 m

When they work with MMFs, lasers (and VCSELs in particular) are different from LEDs, because they excite just a few transverse modes in the fiber, limiting the light propagation to some mode groups. Modal bandwidth is measured with the assumption that the so-called Over-Filled Launch (OFL) conditions are met. However, VCSELs under-fill fiber modes and the OFL conditions are not met. When OFL conditions fail, the impulse response of optical fiber becomes strongly dependent on how the power is injected. When laser sources are used, the modal bandwidth of fiber can be lower or higher than the modal bandwidth of OFL. LEDs inject power into the entire core of the MMF, while a laser injects power only into a small area of the fiber. Modal bandwidth with laser sources is measured by a parameter known as Effective Modal Bandwidth (EMB). When power is correctly injected in the MMF, the EMB is better than the OFL modal bandwidth. Specifically, the minimum OFL bandwidth for an OM3 MMF operating in the first transmission window is 1500 MHz*km, but the minimum EMB is 2000 MHz*km.

Table 1.5
ISO/IEC 11801 and EN50173 standard fiber categories.

	OM1	*OM2*	*OM3*	*OS1*
Core diameter	50 μs, 62.5 μs	50 μs, 62.5 μs	50 μs	9 μs
Max. attenuation (850 nm)	3.5 dB/km	3.5 dB/km	3.5 dB/km	-
Max. attenuation (1300 nm)	1.5 dB/km	1.5 dB/km	1.5 dB/km	-
Max. attenuation (1310 nm)	-	-	-	1.0 dB/km
Max. attenuation (1550 nm)	-	-	-	1.0 dB/km
Min. OFL (850 nm)	200 MHz*km	500 MHz*km	1500 MHz*km	-
Min. OFL (1300 nm)	500 MHz*km	500 MHz*km	500 MHz*km	-

SMF propagates only one transverse mode, which is why modal bandwidth is not an issue for SMF. As a result, SMF transmits higher bit rates over larger distances than MMF. These fibers are, however, limited by second-order dispersive effects:

- Different colors or wavelengths travel at different speeds in the fiber. This effect is called *chromatic dispersion*, and it makes optical pulses spread as they propagate along the fiber. Chromatic dispersion is quantified by the chromatic dispersion coefficient that gives the amount of broadening in an optical fiber per unit of distance and per unit of spectral width of the power source. The chromatic dispersion coefficient uses to be expressed in ps/nm*km units.
- The electromagnetic wave that propagates along the fiber can be decomposed into two ortogonal polarizations. Due to imperfections in the fiber, these polarizations have different propagation speeds. This effect is called *Polarization Mode Dispersion* (PMD). As a result of PMD, optical pulses are spread in the fiber. The amount of spreading caused by optical fiber is given by a PMD coefficient proportional to the square root of the length of the fiber. The PMD coefficient is therefore measured in ps/km1/2. PMD can be ignored for low and medium bit rate applications and it only is an issue for 10 Gbit/s and higher bit rate applications.

The fiber commonly used in telecommunications is made of silica glass. ITU-T Recommendation G.652 specifies a silica glass optical fiber known as *Non-Dispersion-Shifted-Fiber* (NDSF). This fiber has a zero dispersion point near 1300 nm, but minimum attenuation is close to 1550 nm. An important NDSF subclass is the G.652.C fiber, with a low water peak of around 1400 nm. G.652.C fibers are optimized for *Dense Wavelength Division Multiplexing* (DWDM). NDSF is the most commonly deployed SMF, but there are other possibilities as well (see Figure 1.9):

- ITU-T G.653 defines a *Dispersion-Shifted-Fiber* (DSF) with zero dispersion point shifted near to 1550 nm. This fiber is therefore optimized for third-window operation. DSF, however, is not suitable for DWDM applications, due to its increased non-linearity when compared with standard DSF.
- ITU-T G.655 defines a *Non-Zero Dispersion Shifted Fiber* (NZDSF) that has a small, but non-zero chromatic dispersion near 1550 nm. This way, non-linear effects such as “four-wave mixing” are minimized, and at the same time, good performance in DWDM is achieved.

Ethernet standards assume that G.652 fiber is used in SMF installations. This does not, however, prevent the use of other types of SMF, since they may potentially enhance the performance of the optical link. Attenuation is the limiting factor in almost all installations. However, in the third window, chromatic dispersion could be important enough to limit operation over long distances for 10 GbE applications.

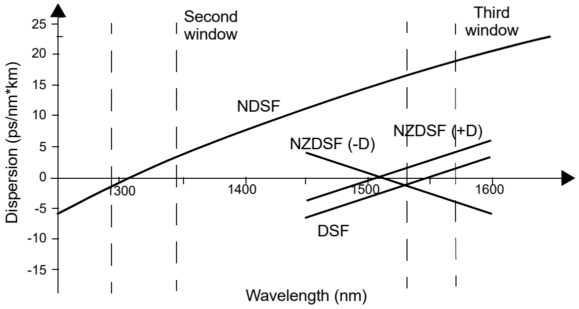


Figure 1.9 Chromatic dispersion of different standard SMF types.

Fibers are packed in many different types of cables. Some of them only contain one fiber, but others may contain hundreds. Optical fiber cables can be roughly divided into two groups:

- *Indoor cables*: The main requirement for this type of cables is fire safety. This is particularly true for fiber that is installed in public spaces. These fibers are manufactured with flame-resistant materials.
- *Outdoor cables* must deal with moisture, high temperature variations and ultraviolet rays. Aerial fiber cables must also be weatherproof.

Most fiber cables use one of the following designs: 900 μm tight buffered fiber and 250 μm coated fiber or bare fiber. The former simply adds a soft plastic coating to a 125 μm silica glass fiber, whereas the latter adds two. The internal coating is made of soft plastic, while the external coating uses harder plastic. This provides extra protection and makes handling easier.

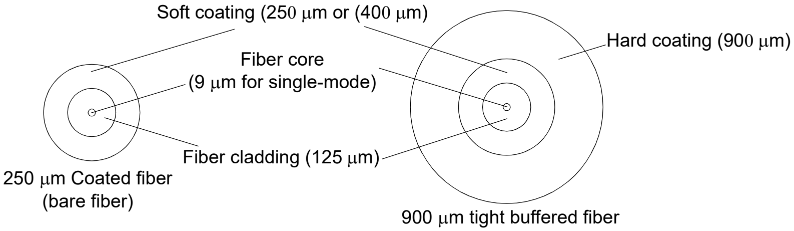


Figure 1.10 Two basic fiber cable configurations.

All available optical fiber cables are variations of two basic configurations known as tight-buffered cables and loose-tube cables (see Figure 1.10). They have the following properties:

- *Tight-buffered* cables contain buffering material for the fiber. This configuration is very suitable for indoor applications, but there are others that are specially designed for outdoor applications. Some tight-buffered cables contain just one fiber, but others may contain many of them. Some of them are specified for riser or plenum cables.
- *Loose-tube* cables house and protect fibers with plastic buffer tubes. Sometimes, a gel filling is used to make the cable waterproof. If fibers are longer than the tube, they are better protected from stress and environmental effects. Loose-tube cables usually contain various (even hundreds of) fibers, and they are typically used for outside-plant applications in aerial, duct or direct-buried installations.

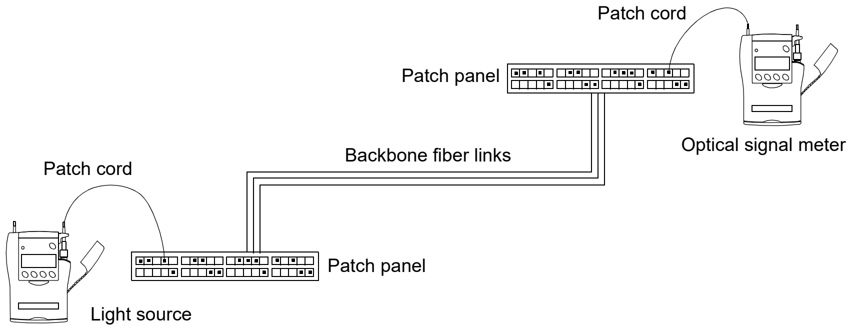


Figure 1.11 Typical optical loss test setup

When designing fiber links, the first step is the characterization of the link power budget. The power budget defines the difference between the transmitter's minimum power out and the receiver's minimum sensitivity. This figure determines the total loss due to attenuation, as well as some other factors that can be introduced between the transmitter and the receiver. Other parameters, such as modal bandwidth and dispersion may be important as well, but as they cannot be affected by installation practices, they are tested by the fiber manufacturer and not in the field. The total insertion loss in a fiber link must be smaller than the available power budget. The insertion loss can be estimated by using this formula:

Loss = Fiber Attenuation(dB/km)xlength(km) + Connector loss(dB)xConnectors + Splice loss(dB)xSplices

The number of splices and connectors in an optical link may be different, and the value for length is also different for each link. Cabling standards provide limits for the attenuation caused by fiber and insertion loss of connectors and splices. The attenuation for MMF must be smaller than 3.5 dB/km while in the first window, and smaller than 1.5 dB/km in the second window. The attenuation for SMF is usually required to be smaller than 1.0 dB/km. The insertion loss must be smaller than 0.3 dB per splice and 0.75 dB per connector. The overall insertion loss allowed for an optical link is defined by IEEE 802.3 (see Table 1.6).

Fiber links must be tested for insertion loss to make sure that they have been installed according to cabling standards. A typical optical loss test set is made up of an optical signal generator and an optical power meter (see Figure 1.11). End-to-end loss depends on the quality of the materials used for the link

Table 1.6
Maximum channel loss and distance for some Ethernet fiber interfaces
(patch cables, cross-connection cables and equipment cords are included in the results).

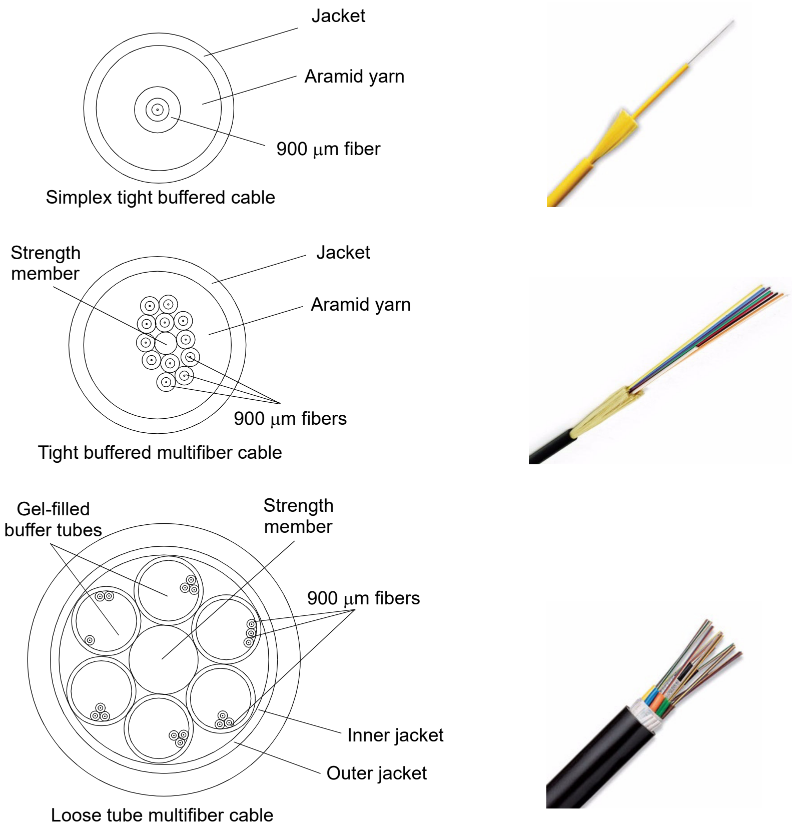
Transmission Rate	*Fiber type / Wavelength*	*Insertion Loss*	*Reach*
1 000 Mb/s	62.5 µs MMF, 160 MHz*km / 850 nm	2.33 dB	220 m
	62.5 µs MMF, 200 MHz*km / 850 nm,	2.53 dB	275 m
	62.5 µs MMF, 500 MHz*km / 1300 nm,	2.32 dB	550 m
	50 µs MMF, 400 MHz*km / 850 nm	3.25 dB	500 m
	50 µs MMF, 500 MHz*km / 850 nm	3.43 dB	550 m
	50 µs MMF, 400 MHz*km / 1300 nm	2.32 dB	550 m
	10 µs SMF / 1310 nm	4.5 dB	5000 m
10 000 Mb/s	62.5 µs MMF, 160 MHz*km / 850 nm	2.6 dB	26 m
	62.5 µs MMF, 200 MHz*km / 850 nm,	2.5 dB	33 m
	50 µs MMF, 400 MHz*km / 850 nm	2.2 dB	66 m
	50 µs MMF, 500 MHz*km / 850 nm	2.3 dB	82 m
	50 µs MMF, 2000 MHz*km / 850 nm	2.6 dB	300 m
	10 µs SMF / 1310 nm	6.0 dB	10 km
	10 µs SMF / 1550 nm	11.0 dB	40 km

and on whether they have been installed correctly or not. To control the amount of insertion loss, connectors must be spliced and added correctly. Optical fiber is also sensitive to bending. Furthermore, the effects of bending are different for each wavelength. This is why insertion loss must be checked at all operating wavelengths. MMF links must be tested at 850 nm and 1310 nm, and SMF links should be tested at 1310 nm and 1550 nm.

A1.3 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 retransmission attempts. 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.



Simplex tight buffered cable

Tight buffered multifiber cable

Loose tube multifiber cable

Figure 1.12 Commercial fiber optic cable designs.

Optical configuration is easier 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 1.14).

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 is that the half-duplex switch detects part of the incoming traffic as a collision and attempts to retransmit data. Retransmissions result in transmission

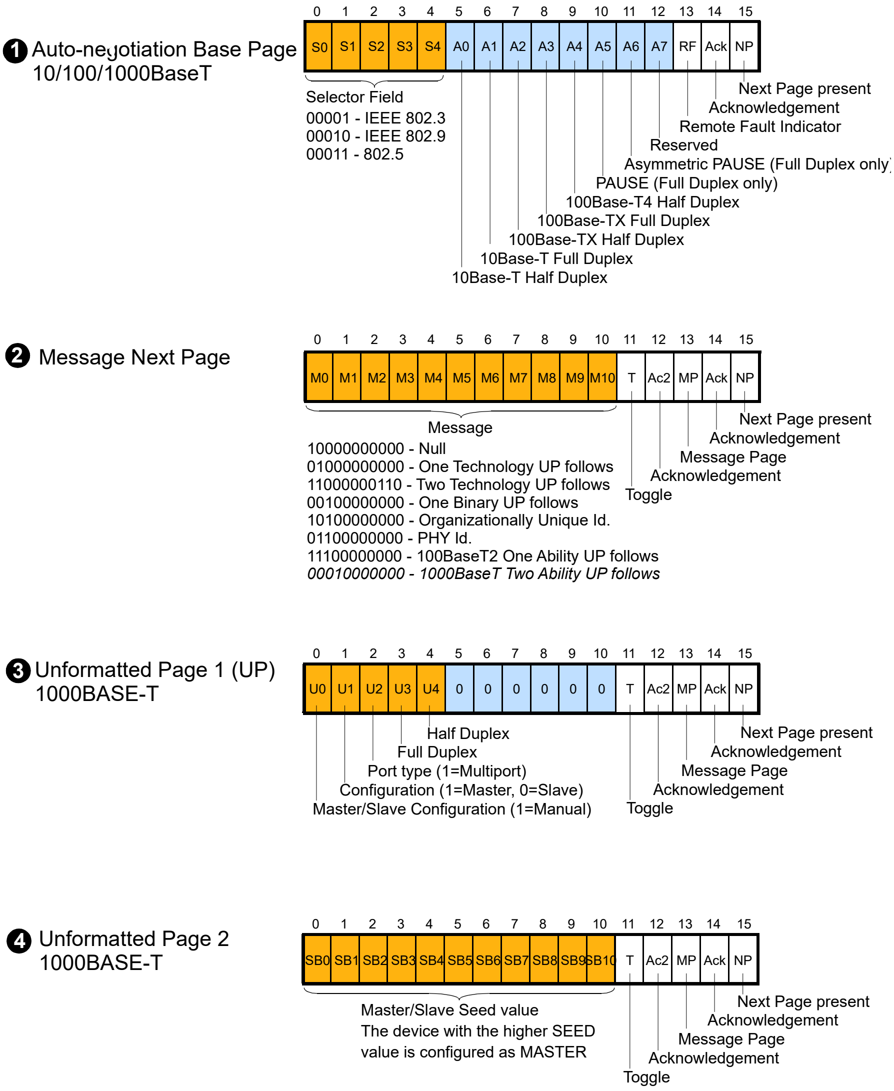


Figure 1.13 Twisted-pair Ethernet auto-negotiation protocol. Auto-negotiation for 10/100 Mbit/s requires exchange of one singe message (a auto-negotiation base page). To communicate their ability to transmit and receive 1000BASE-T signals, a device must exchange four auto-negotiation messages (a base page, a next page and two unformatted pages) with its remote peer.

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.

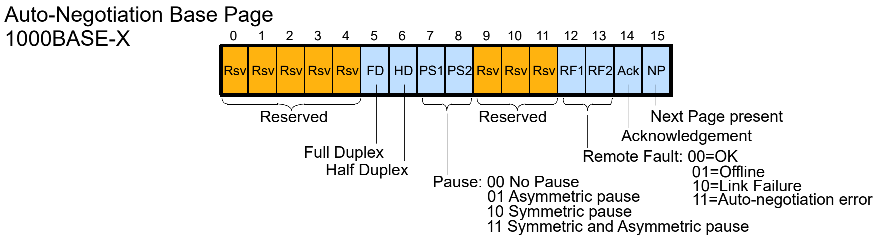
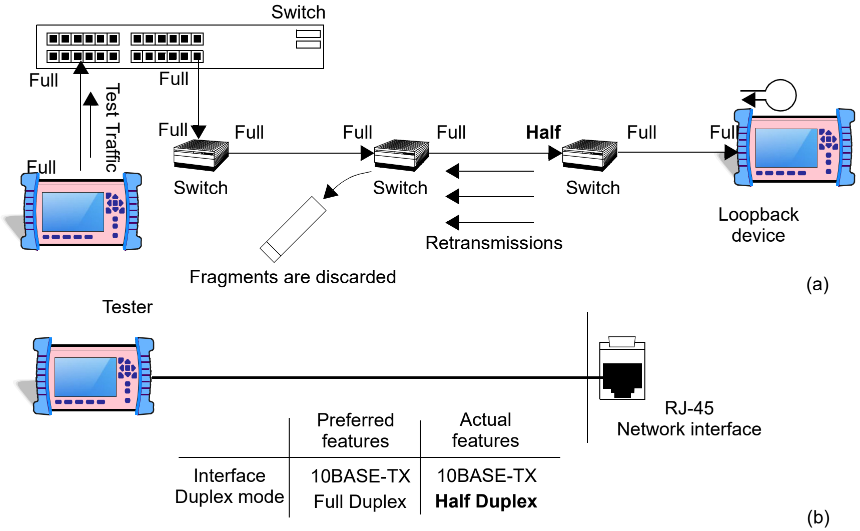


Figure 1.14 1000BASE-X auto-negotiation page description.

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.



	Preferred features	Actual features
Interface Duplex mode	10BASE-TX Full Duplex	10BASE-TX **Half Duplex**

Figure 1.15 Tests related with auto-negotiation. (a) Loopback test to check traffic loss due to a duplex mismatch. (b) Comparison of preferred and actual transmision values in a network interface with unknown features.

A1.4 Determining support of Jumbo frames in a switch chain

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 frame, an encapsulation used to tag frames in trunk links (see Figure 1.17).

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.

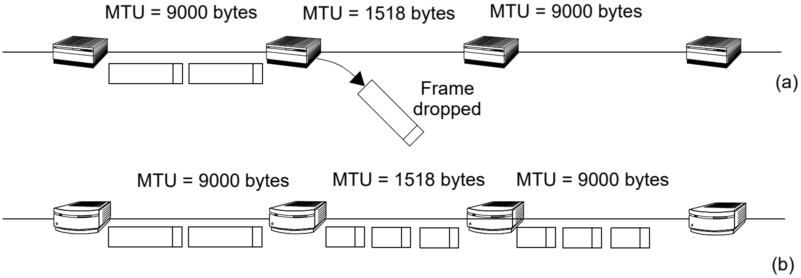


Figure 1.16 (a) Switched path, frames exceeding the S2 MTU are dropped. (b) Routed path, packets are fragmented.

Those who support 9000-byte long frames claim that while Ethernet is now 1 000 times faster than the original 10 Mbit/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 Gbit/s, but even in this case, data may have been aggregated from slower interfaces such as 100 Mbit/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 different from 1580 bytes, they are usually 9216, 9192, 9180, 9176 or 4470 bytes.

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 1.16). Sometimes an *Internet Group Message Protocol* (IGMP) message is generated to inform the transmitter about the event, but this does not always

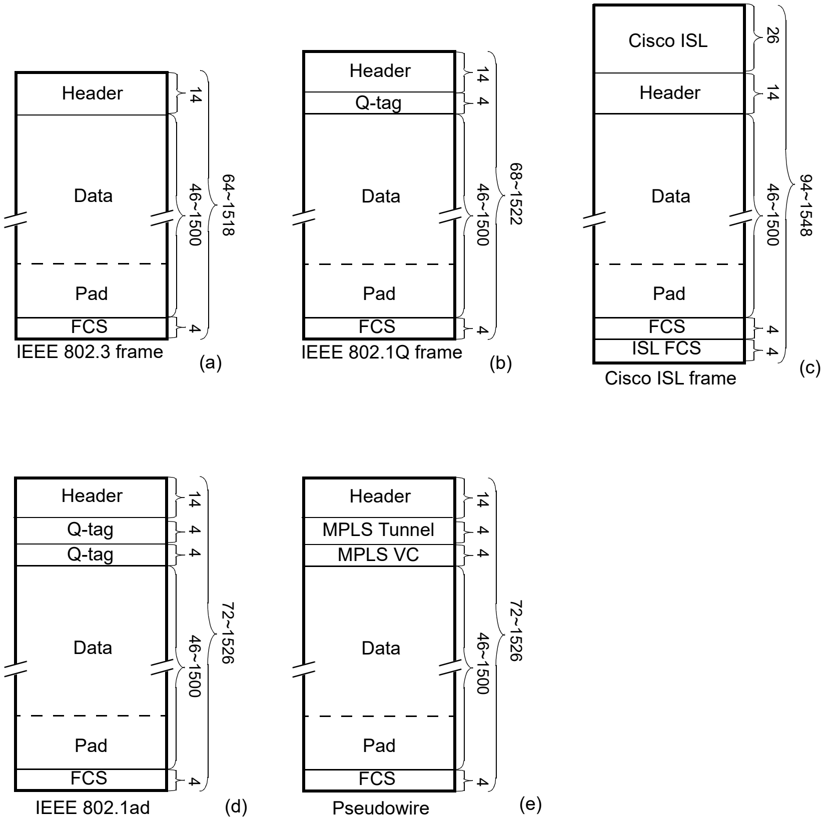


Figure 1.17 (a) The standard Ethernet frame has an MTU of 1518 bytes, (b) Q-tagged VLAN frames have an MTU of 1522 bytes, (c) Cisco Inter-Switch Link (ISL) encapsulated frames have an MTU of 1548 bytes, (d) Q-in-Q Ethernet frames have an MTP of 1526 bytes, (e) MPLS pseudowires built over Ethernet infrastructure have an MTU of 1526 bytes.

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 1.18). A tester that can check the frame size and compute a histogram can be used to analyze the traffic that passes 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 non-real 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.

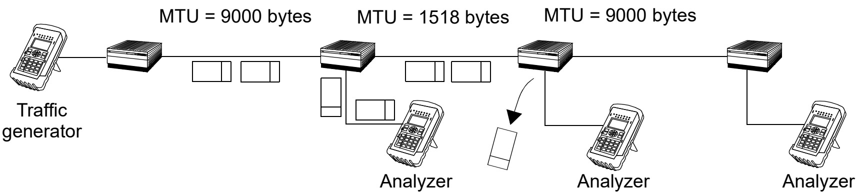


Figure 1.18 (a) Switched path; frames exceeding the S2 MTU are dropped. (b) Routed path; packets are fragmented.

A1.5 CHECKING ISOLATION AND ROUTING BETWEEN 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 (see Figure 2.10). As a result, the fields the network has at its disposal for forwarding packets are: Destination MAC address, Destination IP address and VID. 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, for example 192.168.101.0/24, 192.16.102.0/24 and 192.168.16.103.0/24.

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 user port 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 (a 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 1.20)). 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. 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.

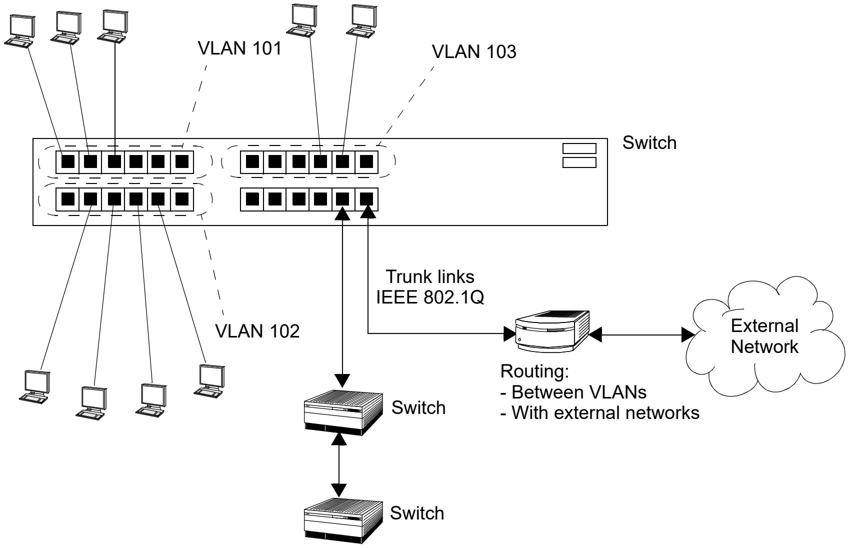


Figure 1.19 A switch with two trunks carrying VLAN-tagged frames. One of the links connects VLAN users from different switches across the network. The other one is used for connection between VLANs and with external networks. All frames that need to change their VID must pass through this router.

All networks and network segments with VLANs should be tested before bringing them into service. Especially, it is very important to test isolation and routing between VLANs. Isolation can be tested in two different ways (see Figure 1.21):

- *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 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.

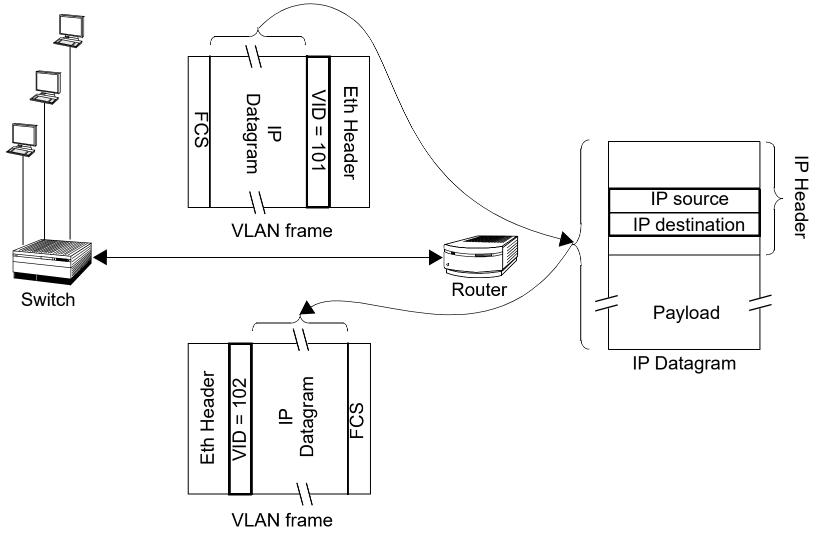
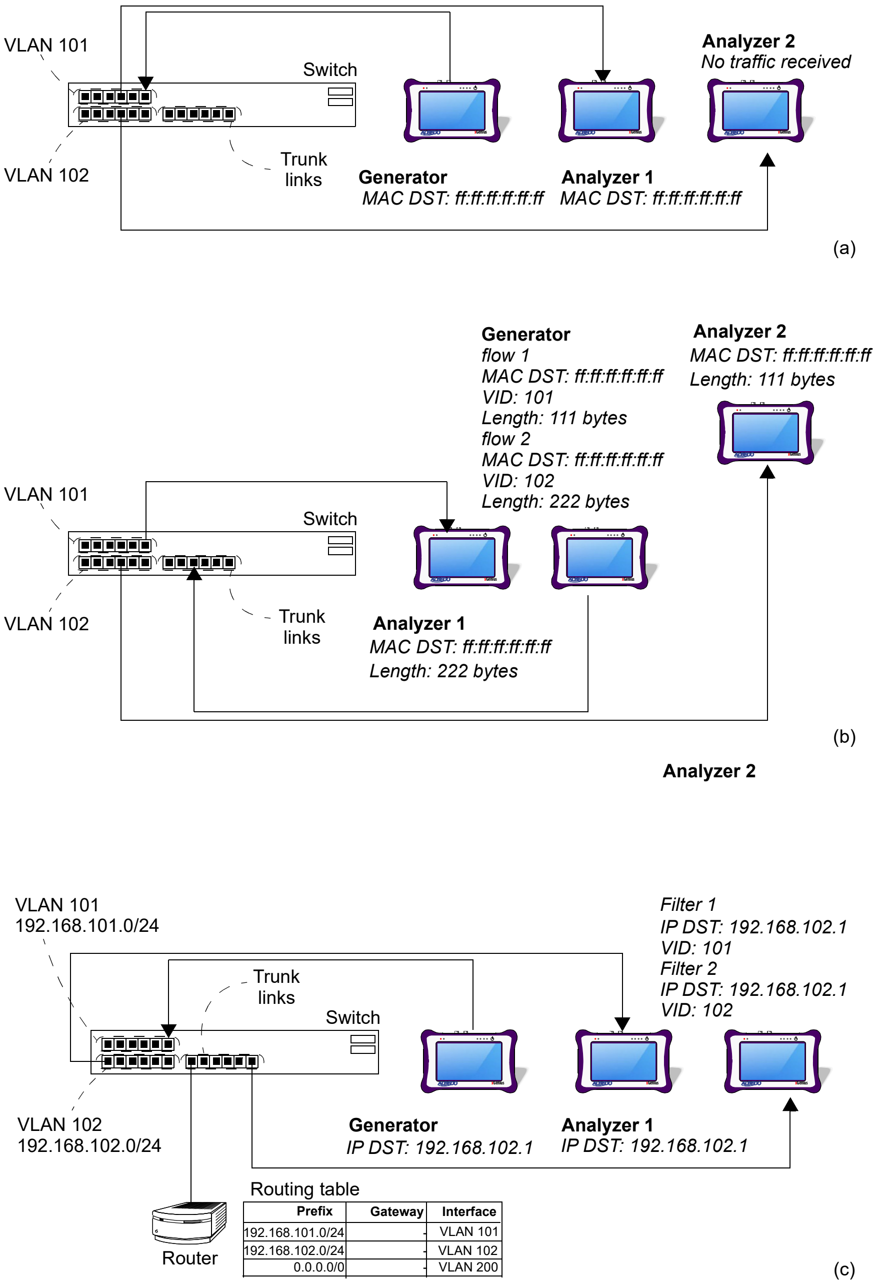


Figure 1.20 Routing between VLANs. Traffic is delivered to a router. The router checks the destination IP address and replaces the source VLAN tag with the destination VLAN tag. The network then forwards the traffic to the destination by bridging.

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.

A1.6 Checking Ethernet Admission Control Mechanisms

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.



Routing table

Prefix	Gateway	Interface
192.168.101.0/24	-	VLAN 101
192.168.102.0/24	-	VLAN 102
0.0.0.0/0	-	VLAN 200

Figure 1.21 (a) Testing VLAN isolation by injecting broadcast traffic into a user port.
(b) Testing VLAN isolation by injecting broadcast traffic into a trunk port.
(c) Testing routing between VLANs by injecting IP traffic into a user port.

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 differentiate 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. Traffic marking at the Ethernet layer is usually not available for LAN users, because Ethernet CoS marks are implemented in VLAN tags, and LAN user interfaces do not support this function. Using source or destination addresses as CoS marks is generally not a good idea, and it is better to separate routing from QoS provisioning.

Things are different in MAN applications. 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 preconfigured rates through the device under test, and then check the frames that are transmitted through the DUT without errors. The number of frames offered and forwarded is compared, and depending on the result, a new iteration starts and the test is performed again with a different frame rate. After some iterations, the test rate converges to the throughput of the device under test.
- *Back-to-back* tests measure the length of the longest maximum-rate frame burst a device can accept without dropping any frames. To perform this measurement, the RFC 2544 compliant tester sends a burst of frames with minimum interframe gaps to the DUT and counts the number of frames forwarded by this device. If the number of transmitted frames is equal to the number of frames forwarded, the length of the burst is increased and the test is performed again. If the number of forwarded frames is less than the number of frames transmitted, the length of the burst is reduced and the test is performed again. Finally, the burst length converges to the longest possible back-to-back burst.

The RFC 2544 throughput test is used to check the steady-state bandwidth of an Ethernet connection. If the average transmission rate is higher the CIR (or EIR, depending on the admission control filter), frames will be dropped sooner or later. If the transmission rate is constant, and smaller than the

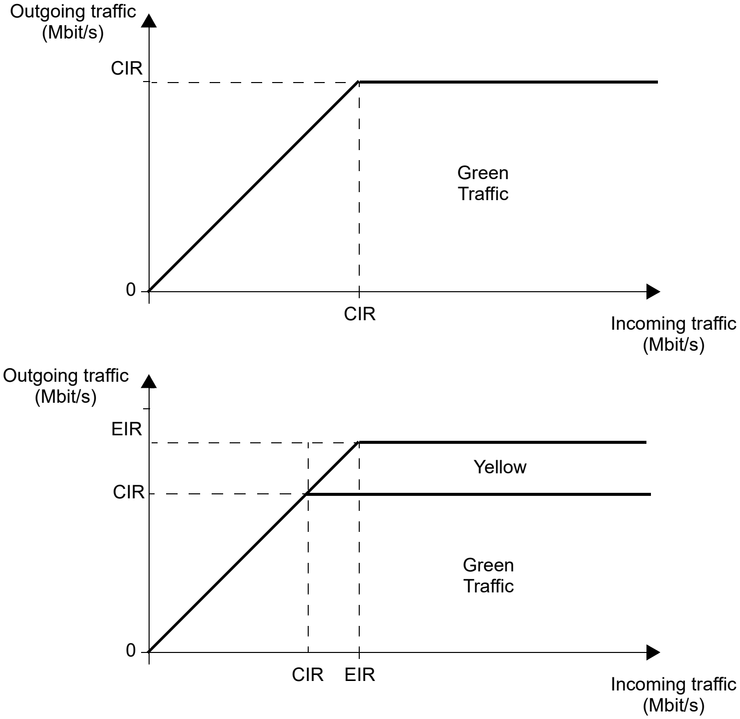


Figure 1.22 The amount of traffic that crosses an admission control filter. Graphics represent steady states, traffic is usually allowed to be greater than the CIR and EIR for short periods of time. (a) The CIR is equal to the EIR, the network guarantees traffic delivery if incoming traffic is smaller 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.

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:

$$\text{CBS} = \text{I}_{\text{CBS}} - \text{CIRxT}_{\text{CBS}} \qquad (1.1)$$

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

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

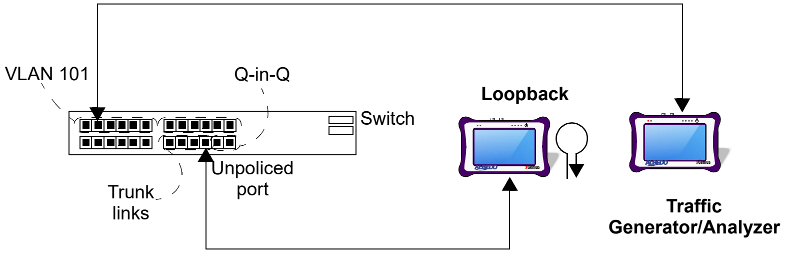


Figure 1.23 In this test, traffic is delivered through an IEEE 802.3 interface to a device connected to an IEEE 802.1Q interface.

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

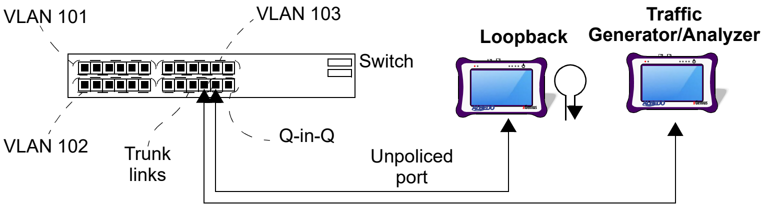


Figure 1.24 In this test, traffic is delivered through an IEEE 802.1Q interface to an IEEE 802.1ad (Q-in-Q). This is a very typical situation in a service provider network.

A1.7 Checking End-to-End Performance Parameters

Once devices are interconnected and remote applications accessible, it is time to test performance and resource availability. QoS tests check *frame loss*, *latency* and *jitter*, and in some cases some other parameters as well. Frame loss, latency and jitter are all important, but there are applications that are not sensitive to some of them (see Table 1.7). For example, VoIP is sensitive to jitter and latency. On the other hand, streamed video and business data are sensitive to frame loss ratio.

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.

Table 1.7
ITU-T Y.1541 Network Performance Objectives.

QoS Class	*Applications*	*Packet Loss*	*Delay*	*Jitter*
0	Real-time, jitter-sensitive, highly interactive traffic (VoIP, videoconference)	1x10^{-3}	100 ms	50 ms
1	Real-time, jitter-sensitive, interactive traffic (VoIP, videoconference)	1x10^{-3}	400 ms	50 ms
2	Transaction data, highly interactive traffic (signalling)	1x10^{-3}	100 ms	Unspecified
3	Transaction data, interactive traffic (signalling)	1x10^{-3}	400 ms	Unspecified

Table 1.7
ITU-T Y.1541 Network Performance Objectives.

QoS Class	*Applications*	*Packet Loss*	*Delay*	*Jitter*
4	Low-loss data traffic (short transactions, bulk data, video streaming)	$1\text{x}10^{-3}$	Unspecified	Unspecified
5	Best-effort traffic (traditional IP data)	Unspecified	Unspecified	Unspecified
6	Real-time, jitter-sensitive, highly interactive, low error-tolerant traffic	$1\text{x}10^{-5}$	100 ms	50 ms
7	Real-time, jitter sensitive, interactive, low error-tolerant traffic	$1\text{x}10^{-5}$	400 ms	50 ms

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, 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.

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.

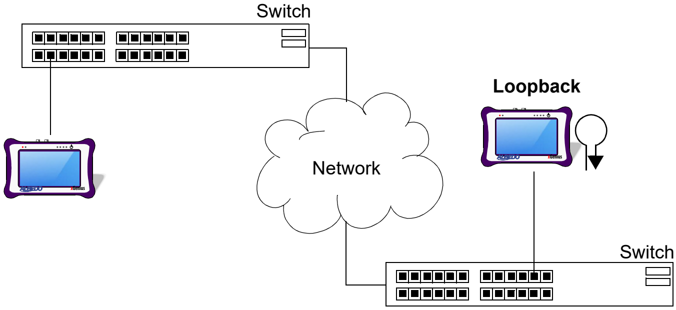


Figure 1.25 QoS test setup. Atraffic generator/analyzer and a loopback device are connected to remote devices. The traffic crosses the network in two directions. The traffic generator/analizer collects statistics on the test traffic.

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 bitrate 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 1.7). This ITU-T standard defines eight traffic classes numbered from 0 to 7. Classes 6 and 7 are provisional. Classes 1 and 2 are defined for interactive traffic, such as VoIP or videoconferencing. Classes 2 and 3 are designed to transport short transactions sensitive to delay, mainly signalling. Classes 4 and 5 are for data traffic and non-interactive multimedia, such as video streaming. The provisional traffic classes are for interactive traffic with low tolerance to errors and packet loss. High-quality IPTV is well suited to these traffic classes.

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

Table 1.8
VoIP Service Degradation under Different Transmission Conditions

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

The VoIP service benchmarking parameter chosen by the ITU-T is the R-Factor, defined in ITU-T G.107 (the so-called E-model). The R-Factor rates the conversational quality of voice communications on a scale from 0 to 100. The R-Factor should be better than 80, and it should never drop below 70. The ITU-T results (see Table 6.4) 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.

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.

Table 1.9
Digital Television Loss/Error Ratio Requirements

Application	*One performance hit per 10 days*	*One performance hit per day*	*10 Performance hits per day*
Contribution (270 Mb/s)	$4x10^{-11}$	$4x10^{-10}$	$4x10^{-9}$
Primary distribution (40 Mb/s)	$3x10^{-10}$	$3x10^{-9}$	$3x10^{-8}$
Access distribution (3 Mb/s)	$4x10^{-9}$	$4x10^{-8}$	$4x10^{-7}$

- *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 6.5). There is no Y.1541 traffic class that meets this requirement. Even the provisional low-loss ratio traffic classes (6 and 7) are unable to provide the desired packet loss ratio. This shows the importance of FEC in video transport to correct errors at the destination, at the price of increased overhead during transmission (see Table 1.10).

Table 1.10
Approximate FEC overhead for different channels, necessary to achieve acceptable overhead in video transmission.

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

Index