The overwhelmingly dominant scheme for wired local area networks is based on the IEEE 802.3 standard, and is commonly referred to as Ethernet. Ethernet began as an experimental bus-based 3-Mbps system. The first commercially available Ethernet and the first version of IEEE 802.3 were bus-based systems operating at 10 Mbps. As technology has advanced, Ethernet has moved from bus-based to switch-based, and the data rate has periodically increased by an order of magnitude. Currently, Ethernet systems are available at speeds up to 100 Gbps. Figure 12.1, based on data from [IEEE12], shows that systems running at 1 Gbps and above dominate in data centers, and that demand is rapidly evolving toward 100-Gbps systems.

We begin this chapter with an overview of the 10-Mbps system and the basic medium access control (MAC) layer defined for 10-Mbps Ethernet. We

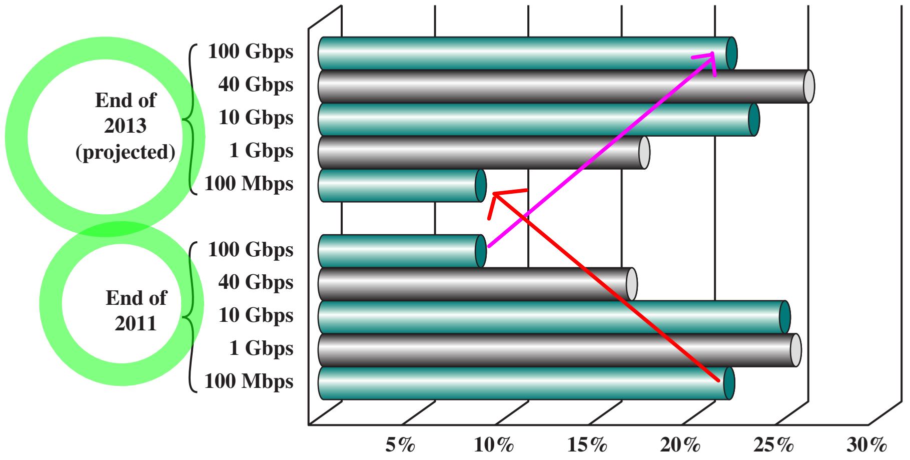


Figure 12.1 Data Center Study—Percentage of Ethernet Links by Speed

Tiempo..

ALOHAnet (o simplemente ALOHA) fue un sistema de redes de computadoras pionero desarrollado en la Universidad de Hawái. Fue desplegado por primera vez en 1970, y aunque la propia red ya no se usa, uno de los conceptos esenciales de esta red es la base para la casi universal Ethernet.

then look at subsequent generations of Ethernet, up to the 100-Gbps version, examining the physical layer definitions and the enhancements to the MAC layer. Finally, the chapter looks at the IEEE 802.1Q VLAN standard.

12.1 TRADITIONAL ETHERNET

The most widely used high-speed LANs today are based on Ethernet and were developed by the IEEE 802.3 standards committee. As with other LAN standards, there is both a medium access control layer and a physical layer, which are discussed in turn in what follows.

IEEE 802.3 Medium Access Control

It is easier to understand the operation of CSMA/CD if we look first at some earlier schemes from which CSMA/CD evolved.

Precursors CSMA/CD and its precursors can be termed *random access*, or *contention*, techniques. They are random access in the sense that there is no predictable or scheduled time for any station to transmit; station transmissions are ordered randomly. They exhibit contention in the sense that stations contend for time on the shared medium.

No hay turno. en el medio compartido.

escuchan maximo de tiempo de ida y vuelta

18.4 % Aloha

37% Aloha Ranurado

The earliest of these techniques, known as ALOHA, was developed for packet radio networks. However, it is applicable to any shared transmission medium. ALOHA, or pure ALOHA as it is sometimes called, specifies that a station may transmit a frame at any time. The station then listens for an amount of time equal to the maximum possible round-trip propagation delay on the network (twice the time it takes to send a frame between the two most widely separated stations) plus a small fixed time increment. If the station hears an acknowledgment during that time, fine; otherwise, it resends the frame. If the station fails to receive an acknowledgment after repeated transmissions, it gives up. A receiving station determines the correctness of an incoming frame by examining a frame check sequence field, as in HDLC. If the frame is valid and if the destination address in the frame header matches the receiver's address, the station immediately sends an acknowledgment. The frame may be invalid due to noise on the channel or because another station transmitted a frame at about the same time. In the latter case, the two frames may interfere with each other at the receiver so that neither gets through; this is known as a **collision**. If a received frame is determined to be invalid, the receiving station simply ignores the frame.

ALOHA is as simple as can be, and pays a penalty for it. Because the number of collisions rises rapidly with increased load, the maximum utilization of the channel is only about 18%.

To improve efficiency, a modification of ALOHA, known as **slotted ALOHA**, was developed. In this scheme, time on the channel is organized into uniform slots whose size equals the frame transmission time. Some central clock or other technique is needed to synchronize all stations. Transmission is permitted to begin only at a slot boundary. Thus, frames that do overlap will do so totally. This increases the maximum utilization of the system to about 37%.

Both ALOHA and slotted ALOHA exhibit poor utilization. Both fail to take advantage of one of the key properties of both packet radio networks and LANs, which is that propagation delay between stations may be very small compared to frame transmission time. Consider the following observations. If the station-to-station propagation time is large compared to the frame transmission time, then, after a station launches a frame, it will be a long time before other stations know about it. During that time, one of the other stations may transmit a frame; the two frames may interfere with each other and neither gets through. Indeed, if the distances are great enough, many stations may begin transmitting, one after the other, and none of their frames get through unscathed. Suppose, however, that the propagation time is small compared to frame transmission time. In that case, when a station launches a frame, all the other stations know it almost immediately. So, if they had any sense, they would not try transmitting until the first station was done. Collisions would be rare because they would occur only when two stations began to transmit almost simultaneously. Another way to look at it is that a short propagation delay provides the stations with better feedback about the state of the network; this information can be used to improve efficiency.

The foregoing observations led to the development of **carrier sense multiple access (CSMA)**. With CSMA, a station wishing to transmit first listens to the medium to determine if another transmission is in progress (carrier sense). If the medium is in use, the station must wait. If the medium is idle, the station may transmit. It may happen that two or more stations attempt to transmit at about the same time. If this happens, there will be a collision; the data from both transmissions will be garbled and not received successfully. To account for this, a station waits a reasonable amount of time after transmitting for an acknowledgment, taking into account the maximum round-trip propagation delay and the fact that the acknowledging station must also contend for the channel to respond. If there is no acknowledgment, the station assumes that a collision has occurred and retransmits.

1)

Si esta libre Tx.

Mencionar relación tiempo de propagación y tiempo de transmi ción!

One can see how this strategy would be effective for networks in which the average frame transmission time is much longer than the propagation time. Collisions can occur only when more than one user begins transmitting within a short time interval (the period of the propagation delay). If a station begins to transmit a frame, and there are no collisions during the time it takes for the leading edge of the packet to propagate to the farthest station, then there will be no collision for this frame because all other stations are now aware of the transmission.

The maximum utilization achievable using CSMA can far exceed that of ALOHA or slotted ALOHA. The maximum utilization depends on the length of the frame and on the propagation time; the longer the frames or the shorter the propagation time, the higher the utilization.

¿Que pasa si el medio está ocupado?
3 aproximaciones.

With CSMA, an algorithm is needed to specify what a station should do if the medium is found busy. Three approaches are depicted in Figure 12.2. One algorithm is **nonpersistent CSMA**. A station wishing to transmit listens to the medium and obeys the following rules:

1. If the medium is idle, transmit; otherwise, go to step 2.
2. If the medium is busy, wait an amount of time drawn from a probability distribution (the retransmission delay) and repeat step 1.

Un problema con CSMA no persistente es que la capacidad se desperdicia porque el medio generalmente permanecerá inactivo después del final de una transmisión, incluso si hay una o más estaciones esperando para transmitir.

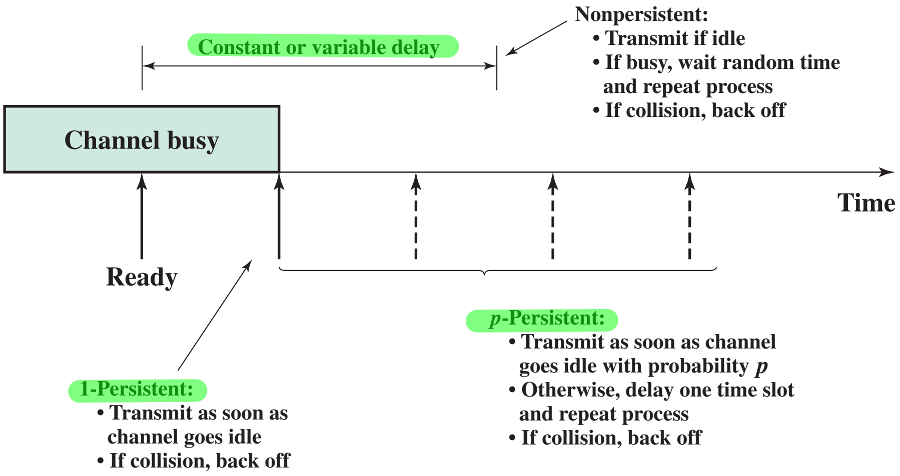

Constant or variable delay
Nonpersistent:
• Transmit if idle
• If busy, wait random time and repeat process
• If collision, back off
Channel busy
Time
Ready
p-Persistent:
• Transmit as soon as channel goes idle with probability *p*
• Otherwise, delay one time slot and repeat process
• If collision, back off
1-Persistent:
• Transmit as soon as channel goes idle
• If collision, back off

Figure 12.2 CSMA Persistence and Backoff

The use of random delays reduces the probability of collisions. To see this, consider that two stations become ready to transmit at about the same time while another transmission is in progress; if both stations delay the same amount of time before trying again, they will both attempt to transmit at about the same time. A problem with nonpersistent CSMA is that capacity is wasted because the medium will generally remain idle following the end of a transmission even if there are one or more stations waiting to transmit.

To avoid idle channel time, the **1-persistent protocol** can be used. A station wishing to transmit listens to the medium and obeys the following rules:

1. If the medium is idle, transmit; otherwise, go to step 2.
2. If the medium is busy, continue to listen until the channel is sensed idle; then transmit immediately.

as estaciones 1-persistentes son egoístas. Si dos o más estaciones están esperando para transmitir, se garantiza una colisión.
Las cosas se resuelven solo después de la colisión, no es que se evitan las coliciones.

Whereas nonpersistent stations are deferential, 1-persistent stations are selfish. If two or more stations are waiting to transmit, a collision is guaranteed. Things get sorted out only after the collision.

A compromise that attempts to reduce collisions, like nonpersistent, and reduce idle time, like 1-persistent, is ***p*-persistent**. The rules are given:

1. If the medium is idle, transmit with probability p, and delay one time unit with probability $(1 - p)$. The time unit is typically equal to the maximum propagation delay.
2. If the medium is busy, continue to listen until the channel is idle and repeat step 1.
3. If transmission is delayed one time unit, repeat step 1.

¿valor de p?
esa es la cuestión

The question arises as to what is an effective value of p. The main problem to avoid is one of instability under heavy load. Consider the case in which n stations have frames to send while a transmission is taking place. At the end of the

probabilidad de tx
= n * p

muchas estaciones=>
n grande.

n chico pocas estaciones

Solución busco
cantidad máxima de n

busco que
n*p sea menor que 1

transmission, the expected number of stations that will attempt to transmit is equal to the number of stations ready to transmit times the probability of transmitting, or np. If np is greater than 1, on average, multiple stations will attempt to transmit and there will be a collision. What is more, as soon as all these stations realize that their transmission suffered a collision, they will be back again, almost guaranteeing more collisions. Worse yet, these retries will compete with new transmissions from other stations, further increasing the probability of collision. Eventually, all stations will be trying to send, causing continuous collisions, with throughput dropping to zero. To avoid this catastrophe, np must be less than one for the expected peaks of n; therefore, if a heavy load is expected to occur with some regularity, p must be small. However, as p is made smaller, stations must wait longer to attempt transmission. At low loads, this can result in very long delays. For example, if only a single station desires to transmit, the expected number of iterations of step 1 is $1/p$ (see Problem 12.2). Thus, if $p = 0.1$, at low load, a station will wait an average of 9 time units before transmitting on an idle line.

Description of CSMA/CD CSMA, although more efficient than ALOHA or slotted ALOHA, still has one glaring inefficiency. When two frames collide, the medium remains unusable for the duration of transmission of both damaged frames. For long frames, compared to propagation time, the amount of wasted capacity can be considerable. This waste can be reduced if a station continues to listen to the medium while transmitting. This leads to the following rules for CSMA/CD:

1. If the medium is idle, transmit; otherwise, go to step 2.
2. If the medium is busy, continue to listen until the channel is idle, then transmit immediately.
3. If a collision is detected during transmission, transmit a brief jamming signal to assure that all stations know that there has been a collision and then cease transmission.
4. After transmitting the jamming signal, wait a random amount of time, referred to as the **backoff**, then attempt to transmit again (repeat from step 1).

Figure 12.3 illustrates the technique for a baseband bus. The upper part of the figure shows a bus LAN layout. At time t_0, station A begins transmitting a packet addressed to D. At t_1, both B and C are ready to transmit. B senses a transmission and so defers. C, however, is still unaware of A's transmission (because the leading edge of A's transmission has not yet arrived at C) and begins its own transmission. When A's transmission reaches C, at t_2, C detects the collision and ceases transmission. The effect of the collision propagates back to A, where it is detected by A some time later, t_3, at which time A ceases transmission.

With CSMA/CD, the amount of wasted capacity is reduced to the time it takes to detect a collision. Question: How long does that take? Let us consider the case of a baseband bus and consider two stations as far apart as possible. For example, in Figure 12.3, suppose that station A begins a transmission and that just before that transmission reaches D, D is ready to transmit. Because D is not yet aware of A's transmission, it begins to transmit. A collision occurs almost immediately and is recognized by D. However, the collision must propagate all the way back to A before

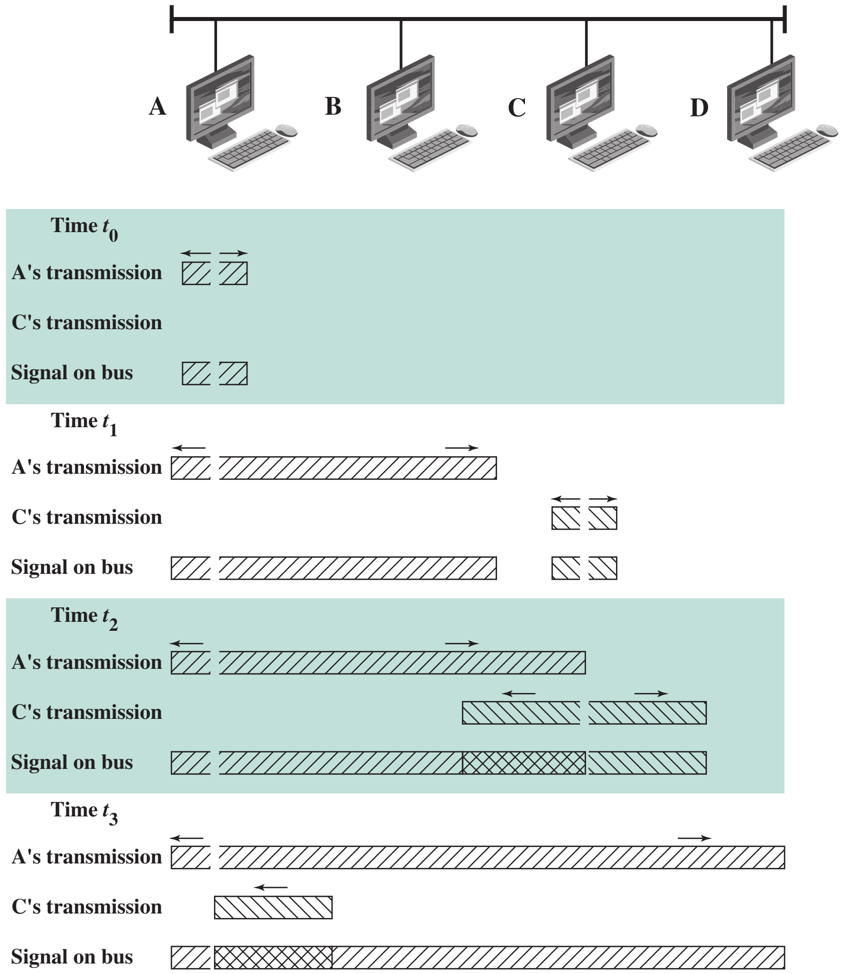


Figure 12.3 CSMA/CD Operation

2 veces T propagación

A is aware of the collision. By this line of reasoning, we conclude that the amount of time that it takes to detect a collision is no greater than twice the end-to-end propagation delay.

T trans. debe ser lo suficientemente largo=> la trama tiene un lago mínimo!!

An important rule followed in most CSMA/CD systems, including the IEEE standard, is that frames should be long enough to allow collision detection prior to the end of transmission. If shorter frames are used, then collision detection does not occur, and CSMA/CD exhibits the same performance as the less efficient CSMA protocol.

For a CSMA/CD LAN, the question arises as to which persistence algorithm to use. You may be surprised to learn that the algorithm used in the IEEE 802.3 standard is 1-persistent. Recall that both nonpersistent and *p*-persistent have performance problems. In the nonpersistent case, capacity is wasted because the medium will generally remain idle following the end of a transmission even if there are stations waiting to send. In the *p*-persistent case, *p* must be set low enough to avoid instability, with the result of sometimes atrocious delays under light load. The 1-persistent algorithm, which means, after all, that $p = 1$, would seem to be even more unstable than p-persistent due to the greed of the stations. What saves the day is that the wasted time due to collisions is mercifully short (if the frames are long relative to propagation delay), and with random backoff, the two stations involved in a collision are unlikely to collide on their next tries. To ensure that backoff maintains stability, IEEE 802.3 and Ethernet use a technique known as **binary exponential backoff**. A station will attempt to transmit repeatedly in the face of repeated collisions. For the first 10 retransmission attempts, the mean value of the random delay is doubled. This mean value then remains the same for 6 additional attempts. After 16 unsuccessful attempts, the station gives up and reports an error. Thus, as congestion increases, stations back off by larger and larger amounts to reduce the probability of collision.

Para los primeros 10 intentos de retransmisión, el valor medio del retardo aleatorio se duplica. Este valor medio permanece igual durante 6 intentos. Después de 16 intentos fallidos, la estación se da por vencida y reporta un error.

The beauty of the 1-persistent algorithm with binary exponential backoff is that it is efficient over a wide range of loads. At low loads, 1-persistence guarantees that a station can seize the channel as soon as it goes idle, in contrast to the non- and p-persistent schemes. At high loads, it is at least as stable as the other techniques. However, one unfortunate effect of the backoff algorithm is that it has a last-in first-out effect; stations with no or few collisions will have a chance to transmit before stations that have waited longer.

For baseband bus, a collision should produce substantially higher voltage swings than those produced by a single transmitter. Accordingly, the IEEE standard dictates that the transmitter will detect a collision if the signal on the cable at the transmitter tap point exceeds the maximum that could be produced by the transmitter alone. Because a transmitted signal attenuates as it propagates, there is a potential problem: If two stations far apart are transmitting, each station will receive a greatly attenuated signal from the other. The signal strength could be so small that when it is added to the transmitted signal at the transmitter tap point, the combined signal does not exceed the CD threshold. For this reason, among others, the IEEE standard restricts the maximum length of coaxial cable to 500 m for 10BASE5 and 200 m for 10BASE2.

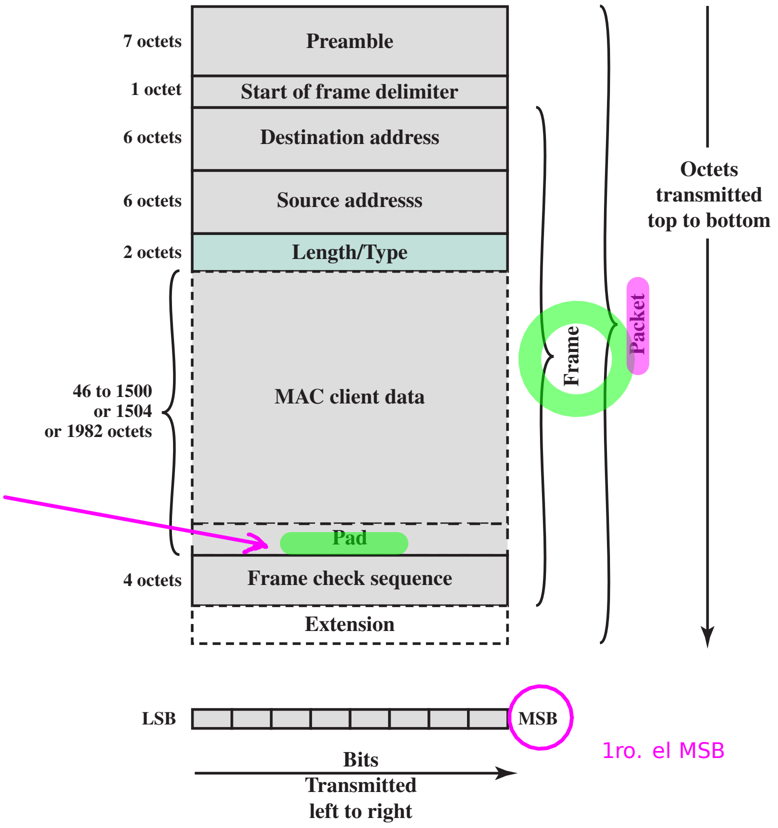
para asegurar, que lo atenuado, mas la suma no sea un valor bajo =>limito la distancia

A much simpler collision detection scheme is possible with the twisted-pair star-topology approach (Figure 11.2). In this case, collision detection is based on logic rather than sensing voltage magnitudes. For any hub, if there is activity (signal) on more than one input, a collision is assumed. A special signal called the collision presence signal is generated. This signal is generated and sent out as long as activity is sensed on any of the input lines. This signal is interpreted by every node as an occurrence of a collision.

hub: genero Señal de colisión en todas las lineas

For a discussion of LAN performance, see Appendix K.

MAC FRAME IEEE 802.3 defines three types of MAC frames. The **basic frame** is the original frame format. In addition, to support data link layer protocol encapsulation within the data portion of the frame, two additional frame types have been added. A **Q-tagged frame** supports 802.1Q VLAN capability, as described in Section 12.3. An **envelope frame** is intended to allow inclusion of additional prefixes and suffixes to the data field required by higher-layer encapsulation protocols such as those defined by the IEEE 802.1 working group (such as Provider Bridges and MAC Security), ITU-T, or IETF (such as MPLS).



Para garantizar la
Longitud míninma

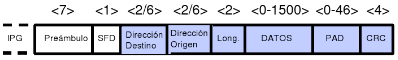
1ro. el MSB

Figure 12.4 IEEE 802.3 MAC Frame Format

Figure 12.4 depicts the frame format for all three types of frames; the differences are contained in the MAC Client Data field. Several additional fields encapsulate the frame to form an 802.3 packet. The fields are as follows:

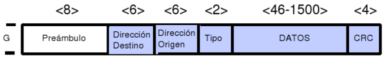
- **Preamble:** A 7-octet pattern of alternating 0s and 1s used by the receiver to establish bit synchronization.
- **Start Frame Delimiter (SFD):** The sequence 10101011, which that delimits the actual start of the frame and enables the receiver to locate the first bit of the frame.
- **Destination Address (DA):** Specifies the station(s) for which the frame is intended. It may be a unique physical address, a multicast address, or a broadcast address.
- **Source Address (SA):** Specifies the station that sent the frame.
- **Length/Type:** Takes on one of two meanings, depending on its numeric value. If the value of this field is less than or equal to 1500 decimal, then the Length/Type field indicates the number of MAC Client Data octets contained in the subsequent MAC Client Data field of the basic frame (length interpretation).

Trama 802.3



menor 1500=> long.

Trama Ethernet



mayor 1536 => tipo

Trama mas cora? 6+6+2+46+4 = 64 bytes = 512 bits (recordar este valor)

If the value of this field is greater than or equal to 1536 decimal then the Length/Type field indicates the nature of the MAC client protocol (Type interpretation). The Length and Type interpretations of this field are mutually exclusive.

- **MAC Client Data:** Data unit supplied by LLC. The maximum size of this field is 1500 octets for a basic frame, 1504 octets for a Q-tagged frame, and 1982 octets for an envelope frame.
- **Pad:** Octets added to ensure that the frame is long enough for proper CD operation.
- **Frame Check Sequence (FCS):** A 32-bit cyclic redundancy check, based on all fields except preamble, SFD, and FCS.
- **Extension:** This field is added, if required for 1-Gbps half-duplex operation. The extension field is necessary to enforce the minimum carrier event duration on the medium in half-duplex mode at an operating speed of 1 Gbps.

IEEE 802.3 10-Mbps Specifications (Ethernet)

The IEEE 802.3 committee has defined a number of alternative physical configurations. This is both good and bad. On the good side, the standard has been responsive to evolving technology. On the bad side, the customer, not to mention the potential vendor, is faced with a bewildering array of options. However, the committee has been at pains to ensure that the various options can be easily integrated into a configuration that satisfies a variety of needs. Thus, the user that has a complex set of requirements may find the flexibility and variety of the 802.3 standard to be an asset.

To distinguish the various implementations that are available, the committee has developed a concise notation:

<data rate in Mbps> <signaling method><maximum segment length in hundreds of meters>

The defined alternatives for 10-Mbps are[1]:

terminador de 50ohms

- **10BASE5:** Specifies the use of 50-Ω coaxial cable and Manchester digital signaling.[2] The maximum length of a cable segment is set at 500 m. The length of the network can be extended by the use of repeaters. A repeater is transparent to the MAC level; as it does no buffering, it does not isolate one segment from another. So, for example, if two stations on different segments attempt to transmit at the same time, their transmissions will collide. To avoid looping, only one path of segments and repeaters is allowed between any two stations. The standard allows a maximum of four repeaters in the path between any two stations, extending the effective length of the medium to 2.5 km.
- **10BASE2:** Similar to 10BASE5 but uses a thinner cable, which supports fewer taps over a shorter distance than the 10BASE5 cable. This is a lower-cost alternative to 10BASE5.

[1] There is also a 10BROAD36 option, specifying a 10-Mbps broadband bus; this option is rarely used.
[2] See Section 5.1.

Table 12.1 IEEE 802.3 10-Mbps Physical Layer Medium Alternatives

	10BASE5	10BASE2	10BASE-T	10BASE-FP
Transmission Medium	Coaxial cable (50 Ω)	Coaxial cable (50 Ω)	Unshielded twisted pair	850-nm optical fiber pair
Signaling Technique	Baseband (Manchester)	Baseband (Manchester)	Baseband (Manchester)	Manchester/ on-off
Topology	Bus	Bus	Star	Star
Maximum Segment Length (m)	500	185	100	500
Nodes per Segment	100	30	—	33
Cable Diameter (mm)	10	5	0.4–0.6	62.5/125 μm

- **10BASE-T:** Uses unshielded twisted pair in a star-shaped topology. Because of the high data rate and the poor transmission qualities of unshielded twisted pair, the length of a link is limited to 100 m. As an alternative, an optical fiber link may be used. In this case, the maximum length is 500 m.
- **10BASE-F:** Contains three specifications: a passive-star topology for interconnecting stations and repeaters with up to 1 km per segment, a point-to-point link that can be used to connect stations and repeaters at up to 2 km, and a point-to-point link that can be used to connect repeaters at up to 2 km.

Note that 10BASE-T and 10BASE-F do not quite follow the notation: "T" stands for twisted pair and "F" stands for optical fiber. Table 12.1 summarizes the remaining options. All of the alternatives listed in the table specify a data rate of 10 Mbps.

12.2 HIGH-SPEED ETHERNET

IEEE 802.3 100-Mbps Specifications (Fast Ethernet)

Fast Ethernet refers to a set of specifications developed by the IEEE 802.3 committee to provide a low-cost, Ethernet-compatible LAN operating at 100 Mbps. The blanket designation for these standards is 100BASE-T. The committee defined a number of alternatives to be used with different transmission media.

Table 12.2 summarizes key characteristics of the 100BASE-T options. All of the 100BASE-T options use the IEEE 802.3 MAC protocol and frame format. 100BASE-X refers to a set of options that use two physical links between nodes: one for transmission and one for reception. 100BASE-TX makes use of shielded twisted pair (STP) or high-quality (Category 5) unshielded twisted pair (UTP). 100BASE-FX uses optical fiber.

In many buildings, any of the 100BASE-X options requires the installation of new cable. For such cases, 100BASE-T4 defines a lower-cost alternative that can use Category 3, voice-grade UTP in addition to the higher-quality Category 5 UTP.[3]

[3] See Chapter 4 for a discussion of Category 3 and Category 5 cable.

2 Pares!

4 Pares!

Table 12.2 IEEE 802.3 100BASE-T Physical Layer Medium Alternatives

	100BASE-TX		100BASE-FX	100BASE-T4
Transmission Medium	2 pair, STP	2 pair, Category 5 UTP	2 optical fibers	4 pair, Category 3, 4, or 5 UTP
Signaling Technique	MLT-3	MLT-3	4B5B, NRZI	8B6T, NRZ
Data Rate	100 Mbps	100 Mbps	100 Mbps	100 Mbps
Maximum Segment Length	100 m	100 m	100 m	100 m
Network Span	200 m	200 m	400 m	200 m

To achieve the 100-Mbps data rate over lower-quality cable, 100BASE-T4 dictates the use of four twisted-pair lines between nodes, with the data transmission making use of three pairs in one direction at a time.

For all of the 100BASE-T options, the topology is similar to that of 10BASE-T, namely a star-wire topology.

100BASE-X For all of the transmission media specified under 100BASE-X, a unidirectional data rate of 100 Mbps is achieved transmitting over a single link (single twisted pair, single optical fiber). For all of these media, an efficient and effective signal encoding scheme is required. The one chosen is referred to as 4B/5B-NRZI. This scheme is further modified for each option. See Appendix 12A for a description.

The 100BASE-X designation includes two physical medium specifications: one for twisted pair, known as 100BASE-TX, and one for optical fiber, known as 100-BASE-FX.

100BASE-TX makes use of two pairs of twisted-pair cable, one pair used for transmission and one for reception. Both STP and Category 5 UTP are allowed. The MTL-3 signaling scheme is used (described in Appendix 12A).

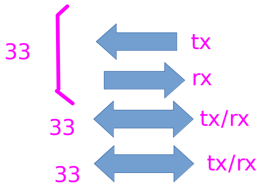
100BASE-FX makes use of two optical fiber cables: one for transmission and one for reception. With 100BASE-FX, a means is needed to convert the 4B/5B-NRZI code group stream into optical signals. The technique used is known as intensity modulation. A binary 1 is represented by a burst or pulse of light; a binary 0 is represented by either the absence of a light pulse or a light pulse at very low intensity.

100BASE-T4 100BASE-T4 is designed to produce a 100-Mbps data rate over lower-quality Category 3 cable, thus taking advantage of the large installed base of Category 3 cable in office buildings. The specification also indicates that the use of Category 5 cable is optional. 100BASE-T4 does not transmit a continuous signal between packets, which makes it useful in battery-powered applications.

trata de usar el cableado existente

For 100BASE-T4 using voice-grade Category 3 cable, it is not reasonable to expect to achieve 100 Mbps on a single twisted pair. Instead, 100BASE-T4 specifies that the data stream to be transmitted is split up into three separate data streams, each with an effective data rate of $33\frac{1}{3}$ Mbps. Four twisted pairs are used. Data are transmitted using three pairs and received using three pairs. Thus, two of the pairs must be configured for bidirectional transmission.

● 33+33+33



sería como 3 pares...

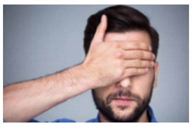
As with 100BASE-X, a simple NRZ encoding scheme is not used for 100BASE-T4. This would require a signaling rate of 33 Mbps on each twisted pair and does not provide synchronization. Instead, a ternary signaling scheme known as 8B6T is used (described in Appendix 12A).

Full-Duplex Operation A traditional Ethernet is half duplex: A station can either transmit or receive a frame, but it cannot do both simultaneously. With full-duplex operation, a station can transmit and receive simultaneously. If a 100-Mbps Ethernet ran in full-duplex mode, the theoretical transfer rate becomes 200 Mbps.

Several changes are needed to operate in full-duplex mode. The attached stations must have full-duplex rather than half-duplex adapter cards. The central point in the star wire cannot be a simple multiport repeater but rather must be a switching hub. In this case each station constitutes a separate collision domain. In fact, there are no collisions and the CSMA/CD algorithm is no longer needed. However, the same 802.3 MAC frame format is used and the attached stations can continue to execute the CSMA/CD algorithm, even though no collisions can ever be detected.

ambas estaciones
en el modo full duplex.

debe ser switch



Mixed Configuration One of the strengths of the Fast Ethernet approach is that it readily supports a mixture of existing 10-Mbps LANs and newer 100-Mbps LANs. For example, the 100-Mbps technology can be used as a backbone LAN to support a number of 10-Mbps hubs. Many of the stations attach to 10-Mbps hubs using the 10BASE-T standard. These hubs are in turn connected to switching hubs that conform to 100BASE-T and that can support both 10-Mbps and 100-Mbps links. Additional high-capacity workstations and servers attach directly to these 10/100 switches. These mixed-capacity switches are in turn connected to 100-Mbps hubs using 100-Mbps links. The 100-Mbps hubs provide a building backbone and are also connected to a router that provides connection to an outside WAN.

comunica
lo que venga!

pero el switch debe
soportar los
distintos modos
o velcocidades.

Gigabit Ethernet

hace 25 años!

In late 1995, the IEEE 802.3 committee formed a High-Speed Study Group to investigate means for conveying packets in Ethernet format at speeds in the gigabits per second range. The strategy for Gigabit Ethernet is the same as that for Fast Ethernet. While defining a new medium and transmission specification, Gigabit Ethernet retains the CSMA/CD protocol and Ethernet format of its 10-Mbps and 100-Mbps predecessors. It is compatible with 100BASE-T and 10BASE-T, preserving a smooth migration path. As more organizations moved to 100BASE-T, putting huge traffic loads on backbone networks, demand for Gigabit Ethernet intensified.

nuevo medio..

retro compatibilidad

Media Access Layer The 1000-Mbps specification calls for the same CSMA/CD frame format and MAC protocol as used in the 10-Mbps and 100-Mbps version of IEEE 802.3. For shared-medium hub operation (Figure 11.11b), there are two enhancements to the basic CSMA/CD scheme:

2 Mejoras

- **Carrier extension:** Carrier extension appends a set of special symbols to the end of short MAC frames so that the resulting block is at least 4096 bit-times in duration, up from the minimum 512 bit-times imposed at 10 and 100 Mbps. This is so that the frame length of a transmission is longer than the propagation time at 1 Gbps.

1

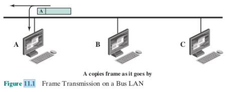
hacer una trama mas larga para extender el tiempo de trasmisión, de maneras que sea mayor que el de propagación. Recordar que el CSMA/CD funciona en base a esta relación.

Se la trama mas corta de 512 bits a 4096 bits! (8 veces mayor)

ráfaja de tramas: en forma consecutiva varias tramas cortas (sin superar un límite)
de trama máxiama y sin necesidad de dejar el control del CSMA/CD, con esto evit
la logro eficiencia extensión de portadora que introduce una sobrecarga.

- **Frame bursting:** This feature allows for multiple short frames to be transmitted consecutively, up to a limit, without relinquishing control for CSMA/CD between frames. Frame bursting avoids the overhead of carrier extension when a single station has a number of small frames ready to send.

2


A
B
C
A copies frame as it goes by
Figure 11.1 Frame Transmission on a Bus LAN

With a switching hub (Figure 11.11c), which provides dedicated access to the medium, the carrier extension and frame bursting features are not needed. This is because data transmission and reception at a station can occur simultaneously without interference and with no contention for a shared medium.

Physical Layer The 1-Gbps specification for IEEE 802.3 includes the following physical layer alternatives (Figure 12.5):

- **1000BASE-SX:** This short-wavelength option supports duplex links of up to 275 m using 62.5-μm multimode or up to 550 m using 50-μm multimode fiber. Wavelengths are in the range of 770 to 860 nm.
- **1000BASE-LX:** This long-wavelength option supports duplex links of up to 550 m of 62.5-μm or 50-μm multimode fiber or 5 km of 10-μm single-mode fiber. Wavelengths are in the range of 1270 to 1355 nm.

- **1000BASE-CX:** This option supports 1-Gbps links among devices located within a single room or equipment rack, using copper jumpers (specialized shielded twisted-pair cable that spans no more than 25 m). Each link is composed of a separate shielded twisted pair running in each direction.
- **1000BASE-T:** This option makes use of four pairs of Category 5 unshielded twisted pair to support devices over a range of up to 100 m, transmitting and receiving on all four pairs at the same time, with echo cancellation circuitry.

tx/rx en 4 pares!

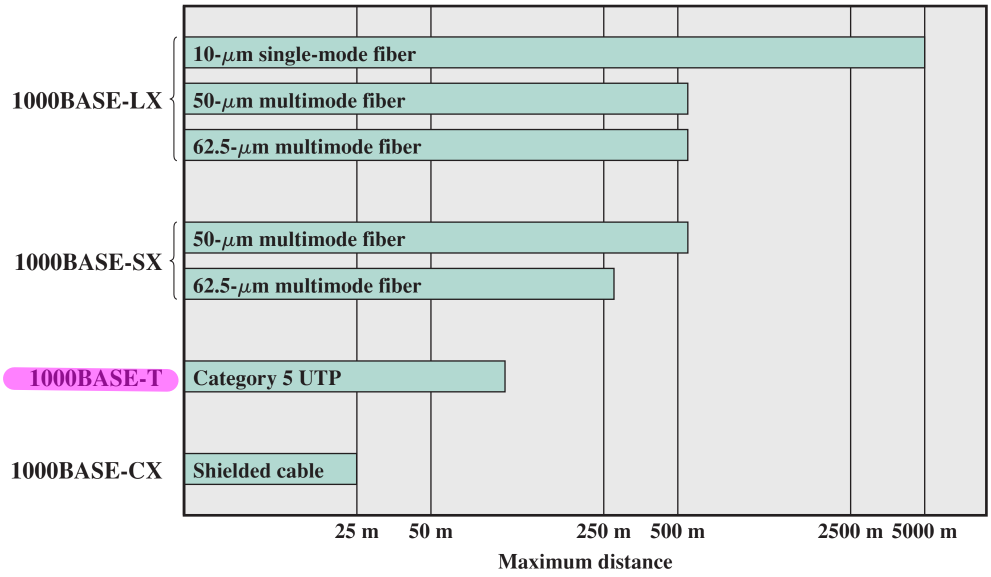
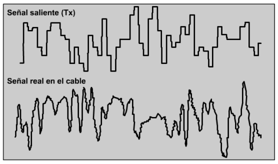


Figure 12.5 Gigabit Ethernet Medium Options (log scale)

4D-PAM5: 4 dimensiones que llevan codificación, los datos que provienen de la estación transmisora se dividen cuidadosamente en cuatro corrientes paralelas; luego se codifican, se transmiten y se detectan en paralelo y finalmente se reensemblan en una solo flujo de bits recibido. con modulación de ancho de pulso de 5 niveles.
Esta codificación tiene nueve (9) niveles de voltaje en periodos de inactividad y 17 niveles en periodos de transmisión.
Con este gran número de estados y con los efectos del ruido, la señal en el cable parece más analógica que digital.



ya mencionamos
8B/10B es para
todos menos
1000B-T
Para 1000B-T
es 4D-PAM5

The signal encoding scheme used for the first three Gigabit Ethernet options just listed is 8B/10B, which is described in Appendix 12A. The signal encoding scheme used for 1000BASE-T is 4D-PAM5, a complex scheme whose description is beyond our scope.

10-Gbps Ethernet

While gigabit products were still fairly new, attention turned to a 10-Gbps Ethernet capability. The principle driving requirement for 10-Gigabit Ethernet was the increase in Internet and intranet traffic. A number of factors contributed to the explosive growth in both Internet and intranet traffic:

Factores ----------->

- An increase in the number of network connections
- An increase in the connection speed of each end-station (e.g., 10 Mbps users moving to 100 Mbps, analog 56-kbps users moving to DSL and cable modems)
- An increase in the deployment of bandwidth-intensive applications such as high-quality video 4K
- An increase in Web hosting and application hosting traffic

backbone ->
demanda ->
toda la red!

Initially network managers used 10-Gbps Ethernet to provide high-speed, local backbone interconnection between large-capacity switches. As the demand for bandwidth increased, 10-Gbps Ethernet was deployed throughout the entire network, to include server farm, backbone, and campuswide connectivity. This technology enables Internet service providers (ISPs) and network service providers (NSPs) to create very high-speed links at a low cost, between co-located, carrier-class switches and routers.

Ethernet para
Wan y Man en P2P.

Foto vieja

The technology also allows the construction of metropolitan area networks (MANs) and WANs that connect geographically dispersed LANs between campuses or points of presence (PoPs). Thus, Ethernet begins to compete with ATM and other wide area transmission and networking technologies. In most cases where the customer requirement is data and TCP/IP transport, 10-Gbps Ethernet provides substantial value over ATM transport for both network end users and service providers:

1990 10BaseT
1995 100BaseT
ATM 1996
1999 1000BaseT
2003 10GBase-X
2009 40G y 100G
7 años

- No expensive, bandwidth-consuming conversion between Ethernet packets and ATM cells is required; the network is Ethernet, end to end.
- The combination of IP and Ethernet offers quality of service and traffic-policing capabilities that approach those provided by ATM, so that advanced traffic-engineering technologies are available to users and providers.
- A wide variety of standard optical interfaces (wavelengths and link distances) have been specified for 10-Gbps Ethernet, optimizing its operation and cost for LAN, MAN, or WAN applications.

Figure 12.6 illustrates potential uses of 10-Gbps Ethernet. Higher-capacity backbone pipes help relieve congestion for workgroup switches, where Gigabit Ethernet uplinks can easily become overloaded, and for server farms, where 1-Gbps network interface cards are in widespread use.

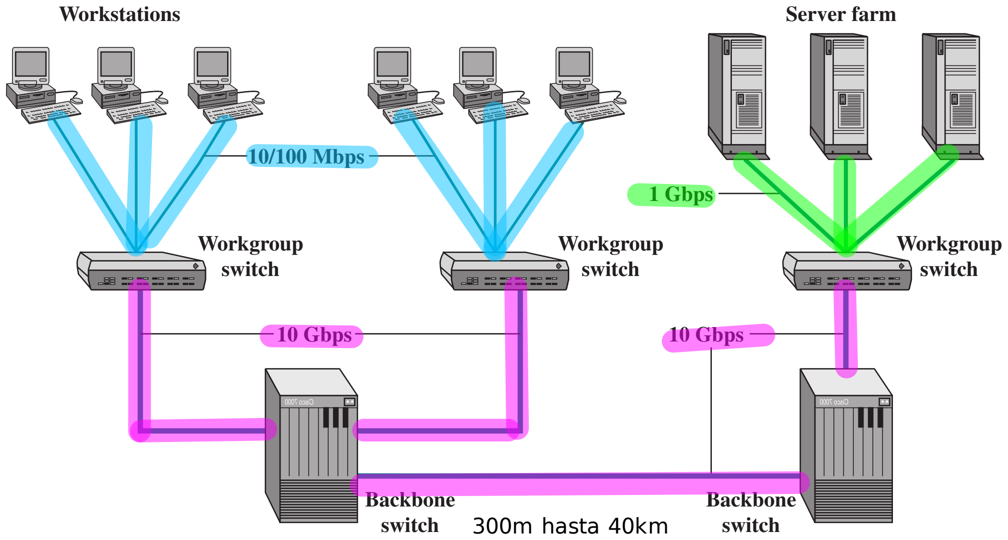
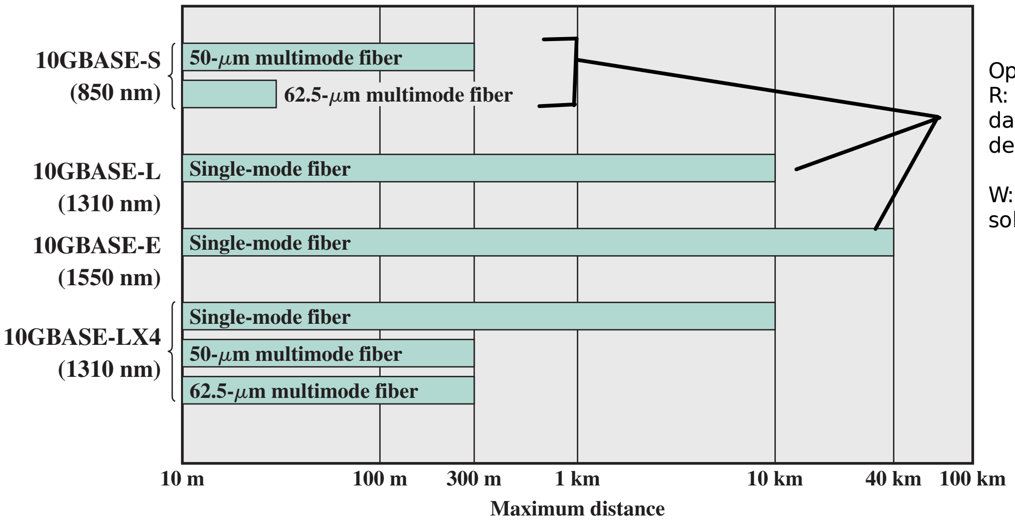


Figure 12.6 Example 10-Gigabit Ethernet Configuration

The goal for maximum link distances covers a range of applications: from 300 m to 40 km. The links operate in full-duplex mode only, using a variety of optical fiber physical media.

Four physical layer options are defined for 10-Gbps Ethernet (Figure 12.7). The first three of these have two suboptions: an "R" suboption and a "W" suboption. The R designation refers to a family of physical layer implementations that use a signal encoding technique known as 64B/66B, described in Appendix 12A. The R implementations are designed for use over *dark fiber*, meaning a fiber optic cable that is not in use and that is not connected to any other equipment. The



Opciones
R: 64B/66B para
dark fiber (uso exclusivo
de 10Gbps Ethernet

W: 64B/66B para uso
sobre SONET

Figure 12.7 10-Gbps Ethernet Distance Options (log scale)

La Red Óptica Síncrona, también llamada SONET, es un estándar creado para la transmisión digital de grandes cantidades de información en redes de fibra óptica mediante el uso de láser o diodos emisores de luz LED. Este estándar, definido por el ANSI para la red pública de telefonía empleada en EE.UU. a mediados de los años ochenta. Sistema que permite el envío de varios canales sobre un mismo medio mediante la multiplexación. En la mayoría de los casos SONET (EEUU) es un subconjunto de SDH (ITU)

W designation refers to a family of physical layer implementations that also use 64B/66B signaling but that are then encapsulated to connect to SONET equipment.

The four physical layer options are as follows:

- **10GBASE-S (short):** Designed for 850-nm transmission on multimode fiber. This medium can achieve distances up to 300 m. There are 10GBASE-SR and 10GBASE-SW versions.
- **10GBASE-L (long):** Designed for 1310-nm transmission on single-mode fiber. This medium can achieve distances up to 10 km. There are 10GBASE-LR and 10GBASE-LW versions.
- **10GBASE-E (extended):** Designed for 1550-nm transmission on single-mode fiber. This medium can achieve distances up to 40 km. There are 10GBASE-ER and 10GBASE-EW versions.
- **10GBASE-LX4:** Designed for 1310-nm transmission on single-mode or multimode fiber. This medium can achieve distances up to 10 km. This medium uses wavelength division multiplexing (WDM) to multiplex the bit stream across four light waves.

100-Gbps Ethernet

Ethernet is widely deployed and is the preferred technology for wired local area networking. Ethernet dominates enterprise LANs, broadband access, data center networking, and has also become popular for communication across metropolitan and even wide area networks. Further, it is now the preferred carrier wire line vehicle for bridging wireless technologies, such as Wi-Fi and WiMAX, into local Ethernet networks.

back bone de las Wireless

This popularity of Ethernet technology is due to the availability of cost-effective, reliable, and interoperable networking products from a variety of vendors. The development of converged and unified communications, the evolution of massive server farms, and the continuing expansion of VoIP, TVoIP, and Web 2.0 applications have driven the need for ever faster Ethernet switches. The following are market drivers for 100-Gbps Ethernet:

Web 2.0:Web deja de ser un simple contenedor o fuente de información; este caso se convierte en una plataforma de trabajo colaborativo

VoIP:Voz sobre protocolo de IP

TVoIP:Televisión sobre IP

- **Data center/Internet media providers:** To support the growth of Internet multimedia content and Web applications, content providers have been expanding data centers, pushing 10-Gbps Ethernet to its limits. Likely to be high-volume early adopters of 100-Gbps Ethernet.
- **Metro-video/service providers:** Video on demand has been driving a new generation of 10-Gbps Ethernet metropolitan/core network buildouts. Likely to be high-volume adopters in the medium term.
- **Enterprise LANs:** Continuing growth in convergence of voice/video/data and in unified communications is driving up network switch demands. However, most enterprises still rely on 1-Gbps or a mix of 1-Gbps and 10-Gbps Ethernet, and adoption of 100-Gbps Ethernet is likely to be slow.
- **Internet exchanges/ISP core routing:** With the massive amount of traffic flowing through these nodes, these installations are likely to be early adopters of 100-Gbps Ethernet.

+ consumo =>
+ capacidad

In 2007, the IEEE 802.3 working group authorized the *IEEE P802.3ba 40Gb/s and 100Gb/s Ethernet Task Force*. The 802.3ba project authorization request cited a number of examples of applications the require greater data rate capacity than 10-Gbps Ethernet offers, including internet exchanges, high performance computing and video-on-demand delivery. The authorization request justified the need for two different data rates in the new standard (40 Gbps and 100 Gbps) by recognizing that aggregate network requirements and end-station requirements are increasing at different rates.

The first products in this category appeared in 2009, and the IEEE 802.3ba standard was finalized in 2010.

An example of the application of 100-Gbps Ethernet is shown in Figure 12.8, taken from [NOWE07]. The trend at large data centers, with substantial banks of blade servers,[4] is the deployment of 10-Gbps ports on individual servers to handle the massive multimedia traffic provided by these servers. Such arrangements are stressing the on-site switches needed to interconnect large numbers of servers. A 100GbE rate was proposed to provide the bandwidth required to handle the

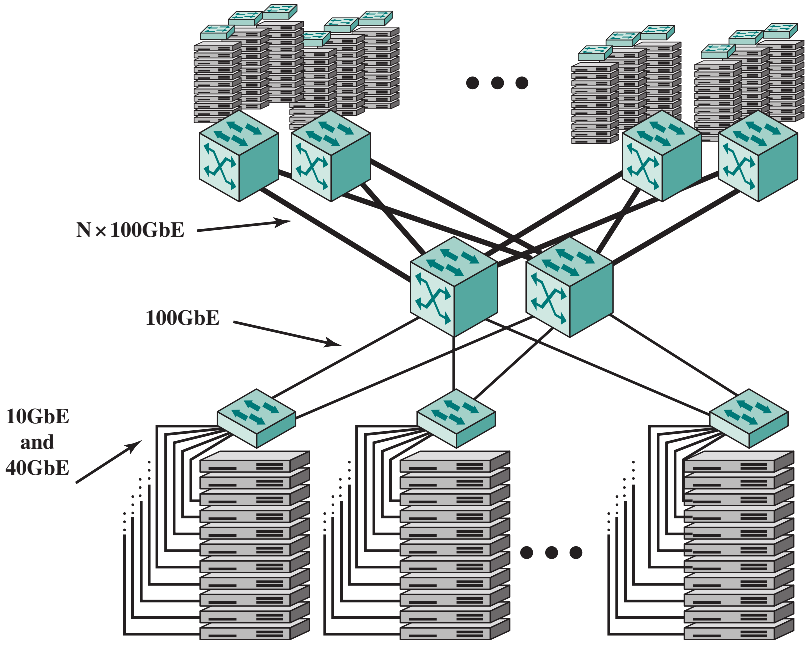


Figure 12.8 Example 100-Gbps Ethernet Configuration for Massive Blade Server Site

[4] A blade server is a server architecture that houses multiple server modules ("blades") in a single chassis. It is widely used in data centers to save space and improve system management. Either self-standing or rack mounted, the chassis provides the power supply, and each blade has its own CPU, memory, and hard disk (definition from pcmag.com encyclopedia).

Blades: Conjunto de Servidores en un Chasis o Rack increased traffic load. It is expected that 100GbE will be deployed in switch uplinks inside the data center as well as providing interbuilding, intercampus, MAN, and WAN connections for enterprise networks.

The success of Fast Ethernet, Gigabit Ethernet, and 10-Gbps Ethernet highlights the importance of network management concerns in choosing a network technology. The 40-Gbps and 100-Gbps Ethernet specifications offer compatibility with existing installed LANs, network management software, and applications. This compatibility has accounted for the survival of 30-year-old technology in today's fast-evolving network environment.

Multilane Distribution The 802.3ba standard uses a technique known as multilane distribution to achieve the required data rates. There are two separate concepts we need to address: multilane distribution and virtual lanes.

The general idea of **multilane distribution** is that, in order to accommodate the very high data rates of 40 and 100 Gbps, the physical link between an end station and an Ethernet switch or the physical link between two switches may be implemented as multiple parallel channels. These parallel channels could be separate physical wires, such as the use of four parallel twisted-pair links between nodes. Alternatively, the parallel channels could be separate frequency channels, such as provided by wavelength division multiplexing over a single optical fiber link.

multiples cables
o fibras.
Alternativa:
una fibra y multiples
longitudes de onda.

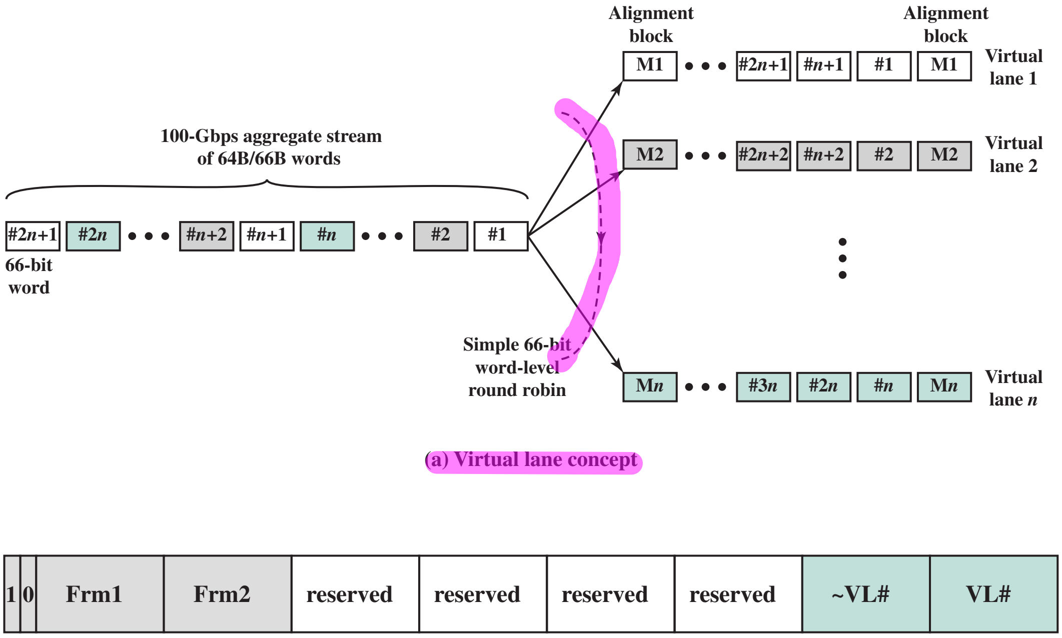
For simplicity and manufacturing ease, we would like to specify a specific multiple-lane structure in the electrical physical sublayer of the device, known as the physical medium attachment (PMA) sublayer. The lanes produced are referred to as **virtual lanes**. If a different number of lanes are actually in use in the electrical or optical link, then the virtual lanes are distributed into the appropriate number of physical lanes in the physical medium dependent (PMD) sublayer. This is a form of inverse multiplexing.

Virtual Lanes
Multiplaxación lógica
sobre vínculos.

Figure 12.9a shows the virtual lane scheme at the transmitter. The user data stream is encoded using the 64B/66B, which is also used in 10-Gbps Ethernet. Data is distributed to the virtual lanes one 66-bit word at a time using a simple round robin scheme (first word to first lane, second word to second lane, etc.). A unique 66-bit alignment block is added to each virtual lane periodically. The alignment blocks are used to identify and reorder the virtual lanes and thus reconstruct the aggregate data stream.

The virtual lanes are then transmitted over physical lanes. If the number of physical lanes is smaller than the number of virtual lanes, then bit-level multiplexing is used to transmit the virtual lane traffic. The number of virtual lanes must be an integer multiple (1 or more) of the number of physical lanes.

Figure 12.9b shows the format of the alignment block. The block consists of 8 single-byte fields preceded by the two-bit synchronization field, which has the value 10. The Frm fields contain a fixed framing pattern common to all virtual lanes and used by the receiver to locate the alignment blocks. The VL# fields contain a pattern unique to the virtual lane: one of the fields is the binary inverse of the other.



(a) Virtual lane concept

(b) Alignment block

Figure 12.9 Multilane Distribution for 100-Gbps Ethernet

3 tipos.

Media Options IEEE 802.3ba specifies three types of transmission media (Table 12.3): copper backplane, twin axial (a type of cable similar to coaxial cable), and optical fiber. For copper media, four separate physical lanes are specified. For optical fiber, either 4 or 10 wavelength lanes are specified, depending on data rate and distance.

Un backplane o placa de bus común es un grupo de conectores electrónicos en paralelo
de tal forma que cada pin de un conector está enlazado con el mismo pin del resto de conectores
formando un bus o canal de transferencia

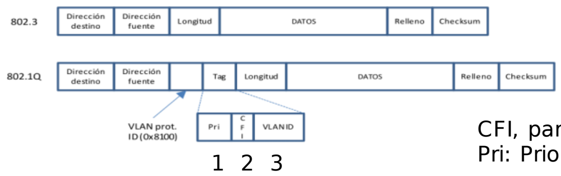
Table 12.3 Media Options for 40-Gbps and 100-Gbps Ethernet

	40 Gbps	**100 Gbps**
1m backplane	40GBASE-KR4	
10 m copper	40GBASE-CR4	1000GBASE-CR10
100 m multimode fiber	40GBASE-SR4	1000GBASE-SR10
10 km single-mode fiber	40GBASE-LR4	1000GBASE-LR4
40 km single-mode fiber		1000GBASE-ER4

Naming nomenclature:

Copper: K = backplane; C = cable assembly
Optical: S = short reach (100 m); L = long reach (10 km); E = extended long reach (40 km)
Coding scheme: R = 64B/66B block coding
Final number: number of lanes (copper wires or fiber wavelengths)

El protocolo 802.1Q propone añadir 4 bytes al encabezado Ethernet original en lugar de encapsular la trama original. El valor del campo *EtherType* se cambia a 0x8100 para señalar el cambio en el formato de la trama.



CFI, para compatibilidad.
Pri: Prioridad

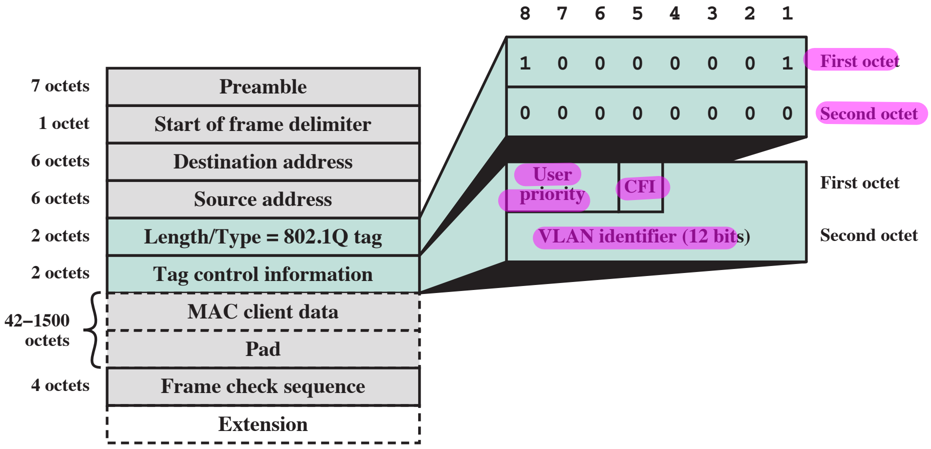
12.3 IEEE 802.1Q VLAN STANDARD

The IEEE 802.1Q standard, last updated in 2005, defines the operation of VLAN bridges and switches that permits the definition, operation, and administration of VLAN topologies within a bridged/switched LAN infrastructure. In this section, we will concentrate on the application of this standard to 802.3 LANs.

Recall from Chapter 11 that a A VLAN is an administratively configured broadcast domain, consisting of a subset of end stations attached to a LAN. A VLAN is not limited to one switch but can span multiple interconnected switches. In that case, traffic between switches must indicate VLAN membership. This is accomplished in 802.1Q by inserting a tag with a VLAN identifier (VID) with a value in the range from 1 to 4094. Each VLAN in a LAN configuration is assigned a globally unique VID. By assigning the same VID to end systems on many switches, one or more VLAN broadcast domains can be extended across a large network.

Figure 12.10 shows the position and content of the 802.1 tag, referred to as Tag Control Information (TCI). The presence of the 2-octet TCI field is indicated by setting the Length/Type field in the 802.3 MAC frame to a value of 8100 hex. The TCI consists of three subfields:

- **User priority (3 bits):** The priority level for this frame.
- **Canonical format indicator (1 bit):** Is always set to zero for Ethernet switches. CFI is used for compatibility between Ethernet-type networks and token-ring-type networks. If a frame received at an Ethernet port has a CFI set to 1, then that frame should not be forwarded as it is to an untagged port.
- **VLAN identifier (12 bits):** The identification of the VLAN. Of the 4096 possible VIDs, a VID of 0 is used to identify that the TCI contains only a priority value, and 4095 (FFF) is reserved, so the maximum possible number of VLAN configurations is 4094.



Longitud=x8100 =>
2 Bytes para ser
interpretados por el
Switch : VLAN

CFI = canonical format indicator
VLAN = virtual local area network

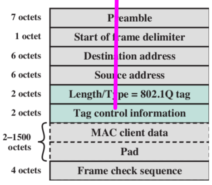
Figure 12.10 Tagged IEEE 802.3 MAC Frame Format

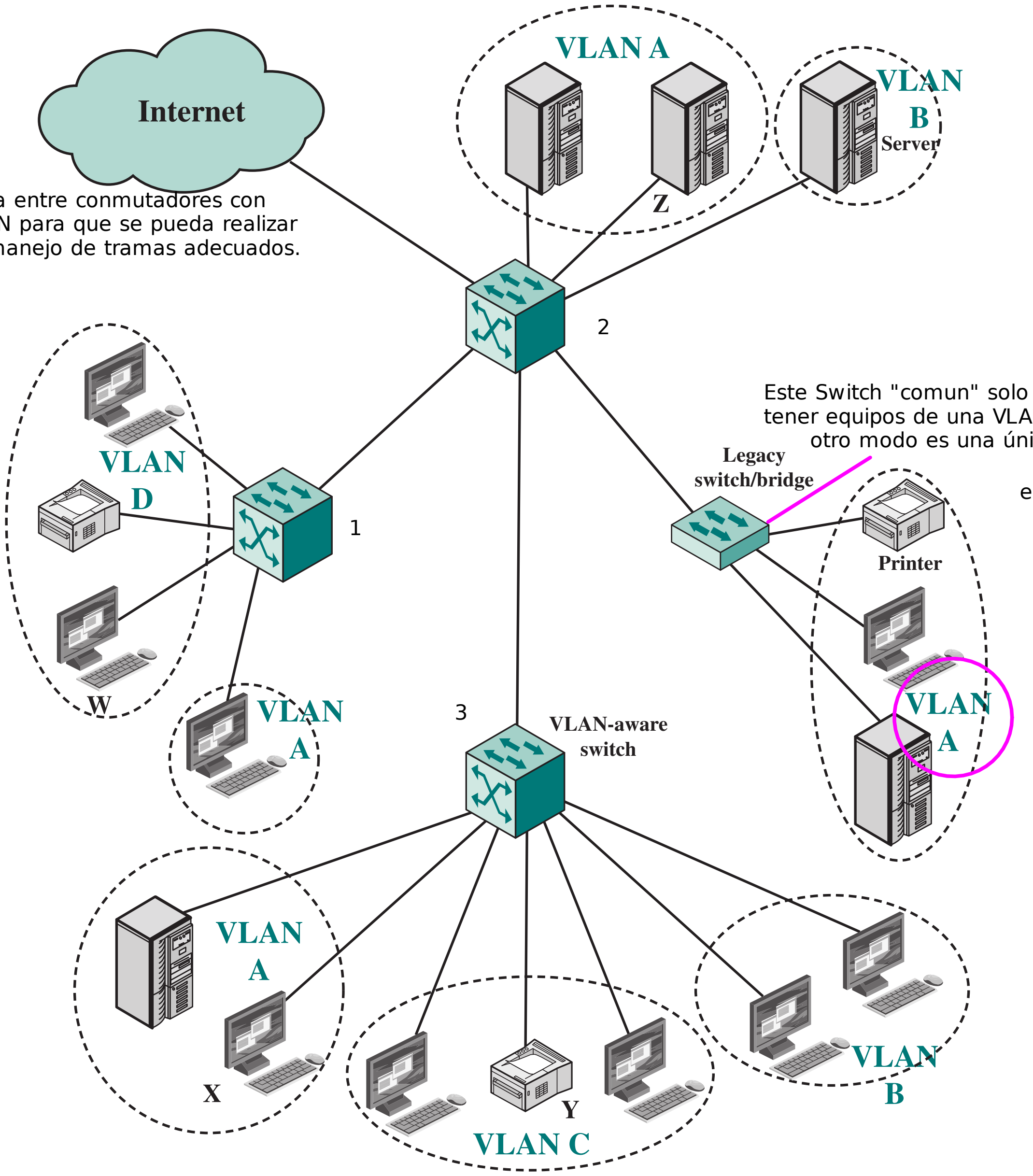
802.1q Tagged- VLAN
La VLAN basada en etiquetas identifica a su miembro por VID.
Si hay más reglas en la lista de filtrado de entrada o la lista de filtrado de salida,
el paquete se examinará con más criterios, esto si el Swithc admite 802.1q.
A cada VLAN basada en etiquetas, el ID de VLAN válido es 1-4094.

Port-Based VLAA (Basado en puerto:)
La VLAN basada en puerto se define por puerto.
No se aplica ningún criterio de filtrado en la VLAN basada en puertos.
El único criterio es el puerto físico al que se conecta.

Figure 12.11 illustrates a LAN configuration that includes three switches that implement 802.1Q and one "legacy" switch or bridge that does not. In this case, all of the end systems of the legacy device must belong to the same VLAN. The MAC frames that traverse trunks between VLAN-aware switches include the 802.1Q TCI tag. This tag is stripped off before a frame is routed to a legacy switch. For end systems connected to a VLAN-aware switch, the MAC frame may or may not

la etiqueta TCI se utiliza entre conmutadores con
reconocimiento de VLAN para que se pueda realizar
el enrutamiento y el manejo de tramas adecuados.





Este Switch "comun" solo puede
tener equipos de una VLAN o dicho de
otro modo es una única VLAN para

este switch!

Figure 12.11 A VLAN Configuration

include the TCI tag, depending on the implementation. The important point is that the TCI tag is used between VLAN-aware switches so that appropriate routing and frame handling can be performed.

12.4 RECOMMENDED READING AND ANIMATIONS

The classic paper on Ethernet is [METC76]. A good survey article on Gigabit Ethernet is [FRAZ99]. [TOYO10] provides an overview of 100-Gbps Ethernet and looks at implementation issues.

FRAZ99 Frazier, H., and Johnson, H. "Gigabit Ethernet: From 100 to 1,000 Mbps." *IEEE Internet Computing*, January/February 1999.

METC76 Metcalfe, R., and Boggs, D. "Ethernet: Distributed Packet Switching for Local Computer Networks." *Communications of the ACM,* July 1976.

TOYO10 Toyoda, H.; Ono, G.; and Nishimura, S. "100 GbE PHY and MAC Layer Implementation." *IEEE Communications Magazine,* March 2010.

Animations

Animations that illustrate CSMA/CD and VLAN concepts are available at the Premium Web site. The reader is encouraged to view these animations to reinforce concepts from this chapter.

12.5 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

1-persistent CSMA
ALOHA
binary exponential backoff
carrier sense multiple access (CSMA)
carrier sense multiple access with collision detection (CSMA/CD)
collision
Ethernet
full-duplex operation
nonpersistent CSMA
p-persistent CSMA
repeater
scrambling
slotted ALOHA

Review Questions

12.1 What is a server farm?
12.2 Explain the three persistence protocols that can be used with CSMA.
12.3 What is CSMA/CD?

12.4 Explain binary exponential backoff.

12.5 What are the transmission medium options for Fast Ethernet?

12.6 How does Fast Ethernet differ from 10BASE-T, other than the data rate?

12.7 In the context of Ethernet, what is full-duplex operation?

Problems

12.1 A disadvantage of the contention approach for LANs, such as CSMA/CD, is the capacity wasted due to multiple stations attempting to access the channel at the same time. Suppose that time is divided into discrete slots, with each of N stations attempting to transmit with probability p during each slot. What fraction of slots are wasted due to multiple simultaneous transmission attempts?

12.2 For p-persistent CSMA, consider the following situation. A station is ready to transmit and is listening to the current transmission. No other station is ready to transmit, and there will be no other transmission for an indefinite period. If the time unit used in the protocol is T, show that the average number of iterations of step 1 of the protocol is $1/p$ and that therefore the expected time that the station will have to wait after the current transmission is $T\left(\frac{1}{P} - 1\right)$. *Hint:* Use the equality $\sum_{i=1}^{\infty} iX^{i-1} = \frac{1}{(1 - X)^2}$.

12.3 The binary exponential backoff algorithm is defined by IEEE 802 as follows:

> The delay is an integral multiple of slot time. The number of slot times to delay before the nth retransmission attempt is chosen as a uniformly distributed random integer r in the range $0 \le r < 2^K$, where $K = \min(n,10)$.

Slot time is, roughly, twice the round-trip propagation delay. Assume that two stations always have a frame to send. After a collision, what is the mean number of retransmission attempts before one station successfully retransmits? What is the answer if three stations always have frames to send?

12.4 Describe the signal pattern produced on the medium by the Manchester-encoded preamble of the IEEE 802.3 MAC frame.

12.5 Analyze the advantages of having the FCS field of IEEE 802.3 frames in the trailer of the frame rather than in the header of the frame.

12.6 With 8B6T coding, the effective data rate on a single channel is 33 Mbps with a signaling rate of 25 Mbaud. If a pure ternary scheme were used, what is the effective data rate for a signaling rate of 25 Mbaud?

12.7 With 8B6T coding, the DC algorithm sometimes negates all of the ternary symbols in a code group. How does the receiver recognize this condition? How does the receiver discriminate between a negated code group and one that has not been negated? For example, the code group for data byte 00 is $+ - 0\,0 + -$ and the code group for data byte 38 is the negation of that, namely, $- + 0\,0 - +$.

12.8 Draw the MLT decoder state diagram that corresponds to the encoder state diagram of Figure 12.12.

12.9 For the bit stream 0101110, sketch the waveforms for NRZ-L, NRZI, Manchester, and differential Manchester, and MLT-3.

12.10 **a.** Verify that the division illustrated in Figure 12.18a corresponds to the implementation of Figure 12.17a by calculating the result step by step using Equation (12.1).

b. Verify that the multiplication illustrated in Figure 12.18b corresponds to the implementation of Figure 12.17b by calculating the result step by step using Equation (12.2).

12.11 Draw a figure similar to Figure 12.16 for the MLT-3 scrambler and descrambler.

APPENDIX 12A DIGITAL SIGNAL ENCODING FOR LANs

In Chapter 5, we looked at some of the common techniques for encoding digital data for transmission, including Manchester and differential Manchester, which are used in some of the LAN standards. In this appendix, we examine some additional encoding schemes referred to in this chapter.

4B/5B-NRZI

This scheme, which is actually a combination of two encoding algorithms, is used for 100BASE-X. To understand the significance of this choice, first consider the simple alternative of a NRZ (nonreturn to zero) coding scheme. With NRZ, one signal state represents binary 1 and one signal state represents binary 0. The disadvantage of this approach is its lack of synchronization. Because transitions on the medium are unpredictable, there is no way for the receiver to synchronize its clock to the transmitter. A solution to this problem is to encode the binary data to guarantee the presence of transitions. For example, the data could first be encoded using Manchester encoding. The disadvantage of this approach is that the efficiency is only 50%. That is, because there can be as many as two transitions per bit time, a signaling rate of 200 million signal elements per second (200 Mbaud) is needed to achieve a data rate of 100 Mbps. This represents an unnecessary cost and technical burden.

Greater efficiency can be achieved using the 4B/5B code. In this scheme, encoding is done 4 bits at a time; each 4 bits of data are encoded into a symbol with five *code bits*, such that each code bit contains a single signal element; the block of five code bits is called a *code group*. In effect, each set of 4 bits is encoded as 5 bits. The efficiency is thus raised to 80%: 100 Mbps is achieved with 125 Mbaud.

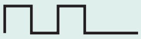
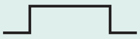
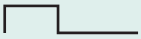
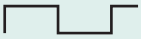
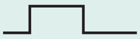
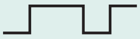
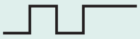
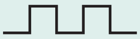
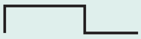
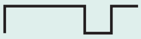
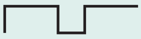
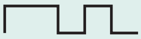
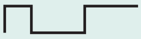
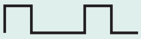
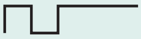
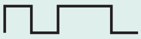
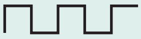
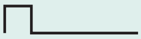
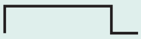
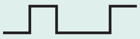
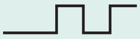
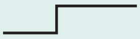
To ensure synchronization, there is a second stage of encoding: Each code bit of the 4B/5B stream is treated as a binary value and encoded using nonreturn to zero inverted (NRZI) (see Figure 5.2). In this code, a binary 1 is represented with a transition at the beginning of the bit interval and a binary 0 is represented with no transition at the beginning of the bit interval; there are no other transitions. The advantage of NRZI is that it employs differential encoding. Recall from Chapter 5 that in differential encoding, the signal is decoded by comparing the polarity of adjacent signal elements rather than the absolute value of a signal element. A benefit of this scheme is that it is generally more reliable to detect a transition in the presence of noise and distortion than to compare a value to a threshold.

Now we are in a position to describe the 4B/5B code and to understand the selections that were made. Table 12.4 shows the symbol encoding. Each 5-bit code group pattern is shown, together with its NRZI realization. Because we are encoding 4 bits with a 5-bit pattern, only 16 of the 32 possible patterns are needed for data encoding. The codes selected to represent the 16 4-bit data blocks are such that a transition is present at least twice for each 5-code group code. No more than three zeros in a row are allowed across one or more code groups.

The encoding scheme can be summarized as follows:

1. A simple NRZ encoding is rejected because it does not provide synchronization; a string of 1s or 0s will have no transitions.
2. The data to be transmitted must first be encoded to assure transitions. The 4B/5B code is chosen over Manchester because it is more efficient.

Table 12.4 4B/5B Code Groups

Data Input (4 bits)	Code Group (5 bits)	NRZI Pattern	Interpretation
0000	11110		Data 0
0001	01001		Data 1
0010	10100		Data 2
0011	10101		Data 3
0100	01010		Data 4
0101	01011		Data 5
0110	01110		Data 6
0111	01111		Data 7
1000	10010		Data 8
1001	10011		Data 9
1010	10110		Data A
1011	10111		Data B
1100	11010		Data C
1101	11011		Data D
1110	11100		Data E
1111	11101		Data F
	11111		Idle
	11000		Start of stream delimiter, part 1
	10001		Start of stream delimiter, part 2
	01101		End of stream delimiter, part 1
	00111		End of stream delimiter, part 2
	00100		Transmit error
	other		Invalid codes

3. The 4B/5B code is further encoded using NRZI so that the resulting differential signal will improve reception reliability.
4. The specific 5-bit patterns for the encoding of the 16 4-bit data patterns are chosen to guarantee no more than three zeros in a row to provide for adequate synchronization.

Those code groups not used to represent data are either declared invalid or assigned special meaning as control symbols. These assignments are listed in Table 12.4. The nondata symbols fall into the following categories:

- **Idle:** The idle code group is transmitted between data transmission sequences. It consists of a constant flow of binary 1s, which in NRZI comes out as a continuous alternation between the two signal levels. This continuous fill pattern establishes and maintains synchronization and is used in the CSMA/CD protocol to indicate that the shared medium is idle.
- **Start of stream delimiter:** Used to delineate the starting boundary of a data transmission sequence; consists of two different code groups.
- **End of stream delimiter:** Used to terminate normal data transmission sequences; consists of two different code groups.
- **Transmit error:** This code group is interpreted as a signaling error. The normal use of this indicator is for repeaters to propagate received errors.

MLT-3

Although 4B/5B-NRZI is effective over optical fiber, it is not suitable as is for use over twisted pair. The reason is that the signal energy is concentrated in such a way as to produce undesirable radiated emissions from the wire. MLT-3, which is used on 100BASE-TX, is designed to overcome this problem.

The following steps are involved:

1. **NRZI to NRZ conversion**. The 4B/5B NRZI signal of the basic 100BASE-X is converted back to NRZ.
2. **Scrambling.** The bit stream is scrambled to produce a more uniform spectrum distribution for the next stage.
3. **Encoder**. The scrambled bit stream is encoded using a scheme known as MLT-3.
4. **Driver**. The resulting encoding is transmitted.

The effect of the MLT-3 scheme is to concentrate most of the energy in the transmitted signal below 30 MHz, which reduces radiated emissions. This in turn reduces problems due to interference.

The MLT-3 encoding produces an output that has a transition for every binary 1 and that uses three levels: a positive voltage (+V), a negative voltage (−V), and no voltage (0). The encoding rules are best explained with reference to the encoder state diagram shown in Figure 12.12:

1. If the next input bit is zero, then the next output value is the same as the preceding value.
2. If the next input bit is one, then the next output value involves a transition:
   a. If the preceding output value was either +V or −V, then the next output value is 0.
   b. If the preceding output value was 0, then the next output value is nonzero, and that output is of the opposite sign to the last nonzero output.

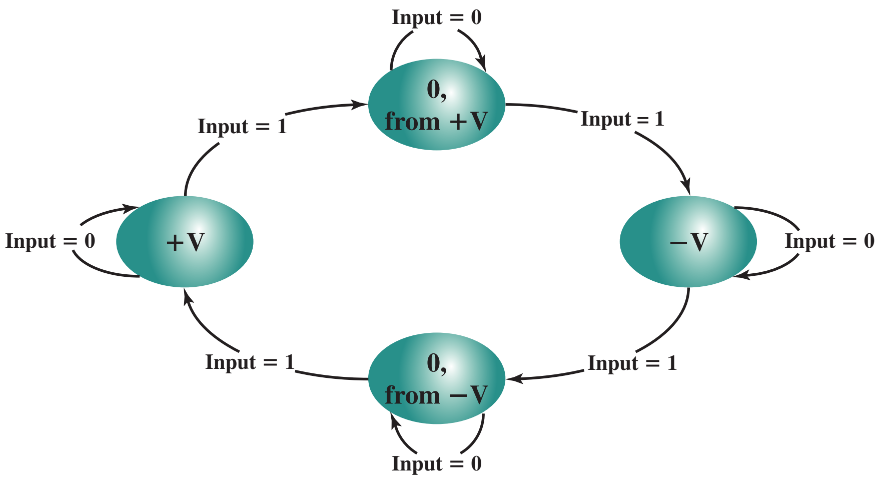


Figure 12.12 MLT-3 Encoder State Diagram

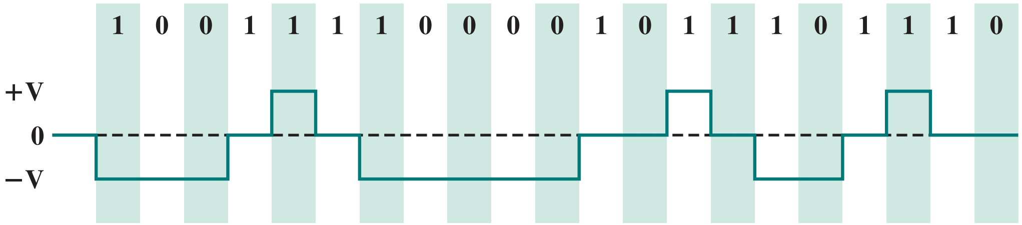


Figure 12.13 Example of MLT-3 Encoding

Figure 12.13 provides an example. Every time there is an input of 1, there is a transition. The occurrences of +V and −V alternate.

8B6T

The 8B6T encoding algorithm uses ternary signaling. With ternary signaling, each signal element can take on one of three values (positive voltage, negative voltage, zero voltage). A pure ternary code is one in which the full information-carrying capacity of the ternary signal is exploited. However, pure ternary is not attractive for the same reasons that a pure binary (NRZ) code is rejected: the lack of synchronization. However, there are schemes referred to as *block-coding methods* that approach the efficiency of ternary and overcome this disadvantage. A new block-coding scheme known as 8B6T is used for 100BASE-T4.

With 8B6T the data to be transmitted are handled in 8-bit blocks. Each block of 8 bits is mapped into a code group of 6 ternary symbols. The stream of code groups is then transmitted in round-robin fashion across the three output channels (Figure 12.14). Thus the ternary transmission rate on each output channel is

$$\frac{6}{8} \times 33\frac{1}{3} = 25 \text{ Mbaud}$$

Table 12.5 shows a portion of the 8B6T code table; the full table maps all possible 8-bit patterns into a unique code group of 6 ternary symbols. The mapping was chosen with two

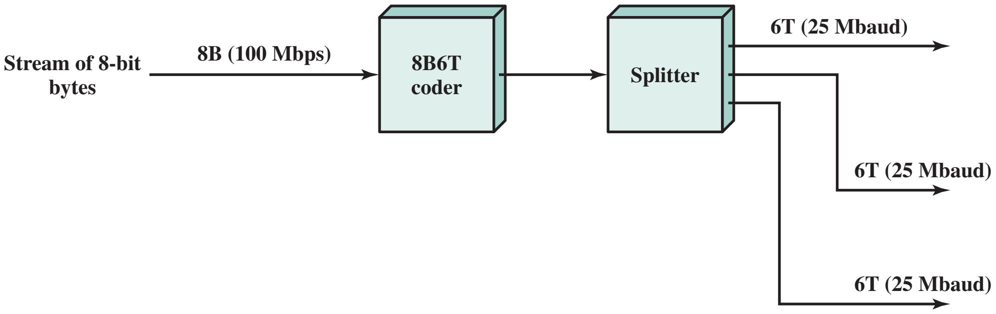


Figure 12.14 8B6T Transmission Scheme

requirements in mind: synchronization and DC balance. For synchronization, the codes were chosen to maximize the average number of transitions per code group. The second requirement is to maintain DC balance, so that the average voltage on the line is zero. For this purpose all of the selected code groups either have an equal number of positive and negative symbols or an excess of one positive symbol. To maintain balance, a DC balancing algorithm is used. In essence, this algorithm monitors the cumulative weight of all code groups transmitted on a single twisted pair. Each code group has a weight of 0 or 1. To maintain balance, the algorithm may negate a transmitted code group (change all + symbols to − symbols and all − symbols to + symbols), so that the cumulative weight at the conclusion of each code group is always either 0 or 1.

Table 12.5 Portion of 8B6T Code Table

Data Octet	6T Code Group	Data Octet	6T Code Group	Data Octet	6T Code Group	Data Octet	6T Code Group
00	+ − 0 0 + −	10	+ 0 + − − 0	20	0 0 − + + −	30	+ − 0 0 − +
01	0 + − + − 0	11	+ + 0 − 0 −	21	− − + 0 0 +	31	0 + − − + 0
02	+ − 0 + − 0	12	+ 0 + − 0 −	22	+ + − 0 + −	32	+ − 0 − + 0
03	− 0 + + − 0	13	0 + + − 0 −	23	+ + − 0 − +	33	− 0 + − + 0
04	− 0 + 0 + −	14	0 + + − − 0	24	0 0 + 0 − +	34	− 0 + 0 − +
05	0 + − − 0 +	15	+ + 0 0 − −	25	0 0 + 0 + −	35	0 + − + 0 −
06	+ − 0 − 0 +	16	+ 0 + 0 − −	26	0 0 − 0 0 +	36	+ − 0 + 0 −
07	− 0 + − 0 +	17	0 + + 0 − −	27	− − + + + −	37	− 0 + + 0 −
08	− + 0 0 + −	18	0 + − 0 + −	28	− 0 − + + 0	38	− + 0 0 − +
09	0 − + + − 0	19	0 + − 0 − +	29	− − 0 + 0 +	39	0 − + − + 0
0A	− + 0 + − 0	1A	0 + − + + −	2A	− 0 − + 0 +	3A	− + 0 − + 0
0B	+ 0 − + − 0	1B	0 + − 0 0 +	2B	0 − − + 0 +	3B	+ 0 − − + 0
0C	+ 0 − 0 + −	1C	0 − + 0 0 +	2C	0 − − + + 0	3C	+ 0 − 0 − +
0D	0 − + − 0 +	1D	0 − + + + −	2D	− − 0 0 + +	3D	0 − + + 0 −
0E	− + 0 − 0 +	1E	0 − + 0 − +	2E	− 0 − 0 + +	3E	− + 0 + 0 −
0F	+ 0 − − 0 +	1F	0 − + 0 + −	2F	0 − − 0 + +	3F	+ 0 − + 0 −

8B/10B

The encoding scheme used for Fibre Channel and Gigabit Ethernet is 8B/10B, in which each 8 bits of data is converted into 10 bits for transmission. This scheme has a similar philosophy to the 4B/5B scheme discussed earlier. The 8B/10B scheme, developed and patented by IBM for use in its 200-megabaud ESCON interconnect system [WIDM83], is more powerful than 4B/5B in terms of transmission characteristics and error-detection capability.

The developers of this code list the following advantages:

- It can be implemented with relatively simple and reliable transceivers at low cost.
- It is well balanced, with minimal deviation from the occurrence of an equal number of 1 and 0 bits across any sequence.
- It provides good transition density for easier clock recovery.
- It provides useful error-detection capability.

The 8B/10B code is an example of the more general mBn B code, in which m binary source bits are mapped into n binary bits for transmission. Redundancy is built into the code to provide the desired transmission features by making $n > m$.

The 8B/10B code actually combines two other codes, a 5B/6B code and a 3B/4B code. The use of these two codes is simply an artifact that simplifies the definition of the mapping and the implementation; the mapping could have been defined directly as an 8B/10B code. In any case, a mapping is defined that maps each of the possible 8-bit source blocks into a 10-bit code block. There is also a function called *disparity control*. In essence, this function keeps track of the excess of zeros over ones or ones over zeros. An excess in either direction is referred to as a disparity. If there is a disparity, and if the current code block would add to that disparity, then the disparity control block complements the 10-bit code block. This has the effect of either eliminating the disparity or at least moving it in the opposite direction of the current disparity.

64B/66B

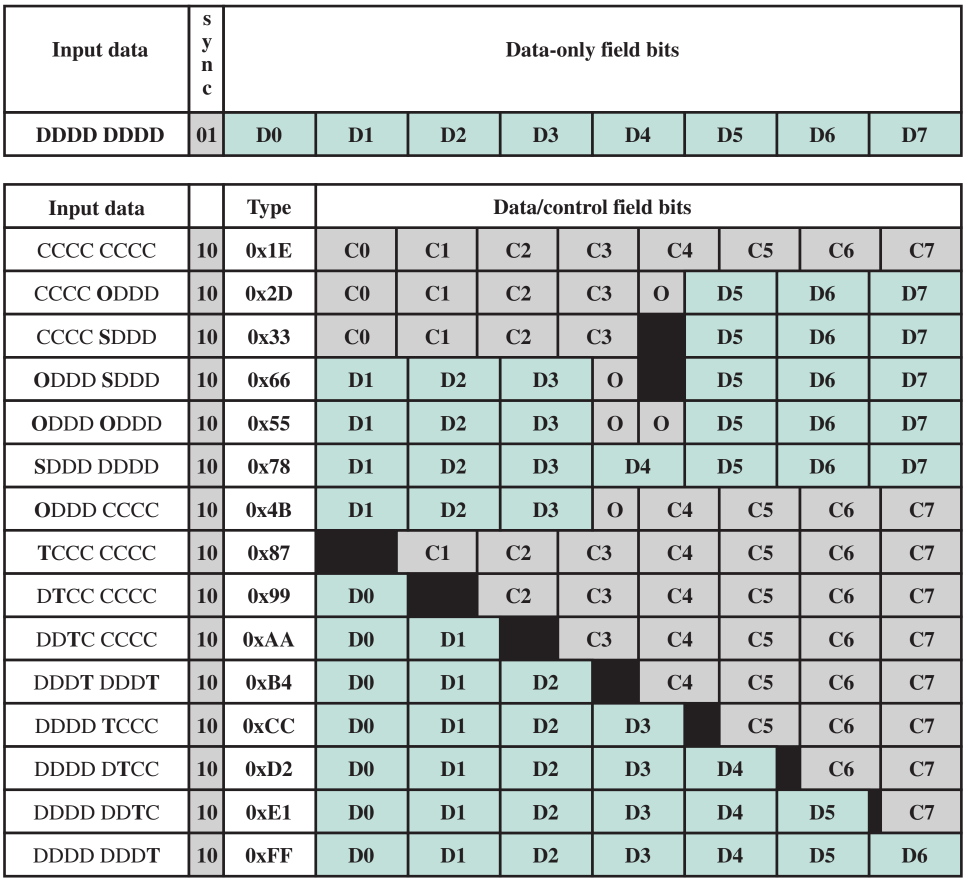
The 8B/10B code results in an overhead of 25%. To achieve greater efficiency at a higher data rate, the 64B/66B code maps a block of 64 bits into an output block of 66 bits, for an overhead of just 3%. This code is used in 10-Gbps and 100-Gbps Ethernet. The entire Ethernet frame, including control fields, is considered “data” for this process. In addition, there are nondata symbols, called “control,” and which include those defined for the 4B/5B code discussed previously plus a few other symbols.

The first step in the process is to encode an input block into a 64-bit block, to which is preopended a 2-bit synchronization field, as show in Figure 12.15. If the input block consists entirely of data octets, then the encoded block consists of the sync field value 10 followed by the 8 data octets unchanged. Otherwise, the input block consists of 8 control octets or a mixture of control octets and data octets. In this case the sync value is 01. This is followed by an 8-bit control type field, which defines the format of the remaining 56 bits of the block. To understand how the 56-bit block is formed, we need to indicate the types of control octets that might be included in the input block, which include the following:

- **Start of packet (S):** Indicates the start of a stream that includes an entire 802.3 MAC packet plus some 64B/66B control characters. This octet is encoded as either 4 bits or 0 bits.
- **End of packet (T):** Marks the termination of the packet. It is encoded using from 0 through 7 bits.

- **Ordered set (0):** Used to adapt clock rates. It is encoded in 4 bits.
- **Other control octets:** Includes idle and error control characters. These octets are encoded in 7 bits.

It is necessary to reduce the number of bits in the input control characters so that the 64-bit input block can be accommodated in 56 bits. Figure 12.15 indicates how this is done. In the input block, the start of packet character is always aligned to be the first or fifth octet. If it occurs as the first octet in the input block, then the remaining seven octets are always data octets. To accommodate all seven data octets, the S field is implied by the block type field but takes up no bits in the encoded block. If the S character is the fifth input octet, then it occupies 4 bits of the encoded block. Similarly, the position and size of the T field is specified



Input data	sync	Data-only field bits							
DDDD DDDD	01	D0	D1	D2	D3	D4	D5	D6	D7

Input data		Type	Data/control field bits							
CCCC CCCC	10	0x1E	C0	C1	C2	C3	C4	C5	C6	C7
CCCC ODDD	10	0x2D	C0	C1	C2	C3	O	D5	D6	D7
CCCC SDDD	10	0x33	C0	C1	C2	C3	■	D5	D6	D7
ODDD SDDD	10	0x66	D1	D2	D3	O	■	D5	D6	D7
ODDD ODDD	10	0x55	D1	D2	D3	O	O	D5	D6	D7
SDDD DDDD	10	0x78	D1	D2	D3	D4	D5	D6	D7	
ODDD CCCC	10	0x4B	D1	D2	D3	O	C4	C5	C6	C7
TCCC CCCC	10	0x87	■	C1	C2	C3	C4	C5	C6	C7
DTCC CCCC	10	0x99	D0	■	C2	C3	C4	C5	C6	C7
DDTC CCCC	10	0xAA	D0	D1	■	C3	C4	C5	C6	C7
DDDT DDDT	10	0xB4	D0	D1	D2	■	C4	C5	C6	C7
DDDD TCCC	10	0xCC	D0	D1	D2	D3	■	C5	C6	C7
DDDD DTCC	10	0xD2	D0	D1	D2	D3	D4	■	C6	C7
DDDD DDTC	10	0xE1	D0	D1	D2	D3	D4	D5	■	C7
DDDD DDDT	10	0xFF	D0	D1	D2	D3	D4	D5	D6	

D = data octet
C = input control octet
Ci = 7-bit output control field
S = start of packet field
T = terminate = end of packet field
O = ordered set control character

Figure 12.15 64B/66B Block Formats

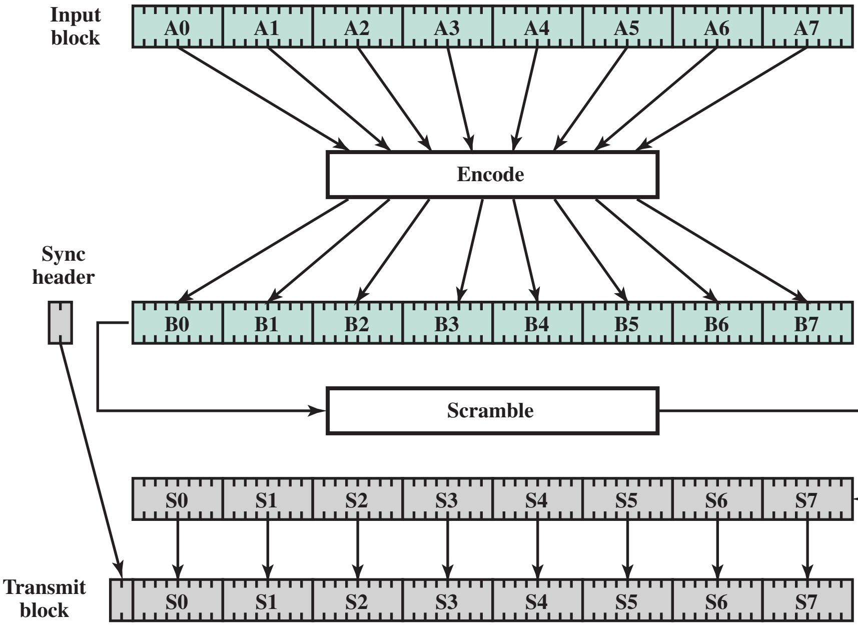


Figure 12.16 64B/66B Transmission Scheme

by the block type field and varies from 0 bits to 7 bits depending on the mixture of control and data octets in the input block.

Figure 12.16 shows the overall scheme for 64B/66B transmission. First, the input block is encoded and the 2-bit sync field is added. Then, scrambling is performed on the encoded 64-bit block using the polynomial $1 + X^{39} + X^{58}$. See Appendix 12B for a discussion of scrambling. The unscrambled 2-bit synchronization field is then prepended to the scrambled block. The sync field provides block alignment and a means of synchronizing when long streams of bits are sent.

Note that for this scheme, no specific coding technique is used to achieve the desired synchronization and frequency of transitions. Rather the scrambling algorithm provides the required characteristics.

APPENDIX 12B SCRAMBLING

For some digital data encoding techniques, a long string of binary 0s or 1s in a transmission can degrade system performance. Also, other transmission properties, such as spectral properties, are enhanced if the data are more nearly of a random nature rather than constant or repetitive. A technique commonly used to improve signal quality is scrambling and descrambling. The scrambling process tends to make the data appear more random.

The scrambling process consists of a feedback shift register, and the matching descrambler consists of a feedforward shift register. An example is shown in Figure 12.17. In this example, the scrambled data sequence may be expressed as follows:

$$B_m = A_m \oplus B_{m-3} \oplus B_{m-5} \qquad \textbf{(12.1)}$$

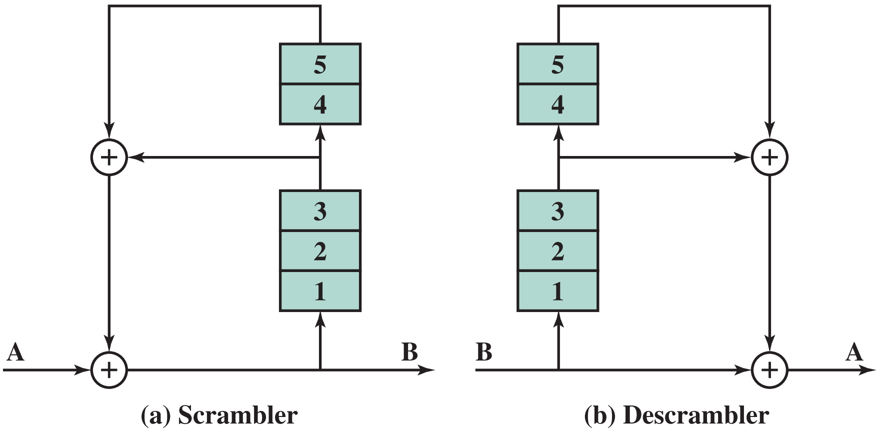


Figure 12.17 Scrambler and Descrambler

where $\oplus$ indicates the exclusive-or operation. The shift register is initialized to contain all zeros. The descrambled sequence is

$$\begin{aligned} C_m &= B_m \oplus B_{m-3} \oplus B_{m-5} \\ &= (A_m \oplus B_{m-3} \oplus B_{m-5}) \oplus B_{m-3} \oplus B_{m-5} \\ &= A_m (\oplus B_{m-3} \oplus B_{m-3} \oplus) B_{m-5} \oplus B_{m-5} \\ &= A_m \end{aligned} \tag{12.2}$$

As can be seen, the descrambled output is the original sequence.

We can represent this process with the use of polynomials. Thus, for this example, the polynomial is $P(X) = 1 + X^3 + X^5$. The input is divided by this polynomial to produce the scrambled sequence. At the receiver the received scrambled signal is multiplied by the same polynomial to reproduce the original input. Figure 12.18 is an example using the polynomial $P(X)$ and an input of 101010100000111.[5] The scrambled transmission, produced by dividing by $P(X)$ (100101), is 101110001101001. When this number is multiplied by $P(X)$, we get the original input. Note that the input sequence contains the periodic sequence 10101010 as well as a long string of zeros. The scrambler effectively removes both patterns.

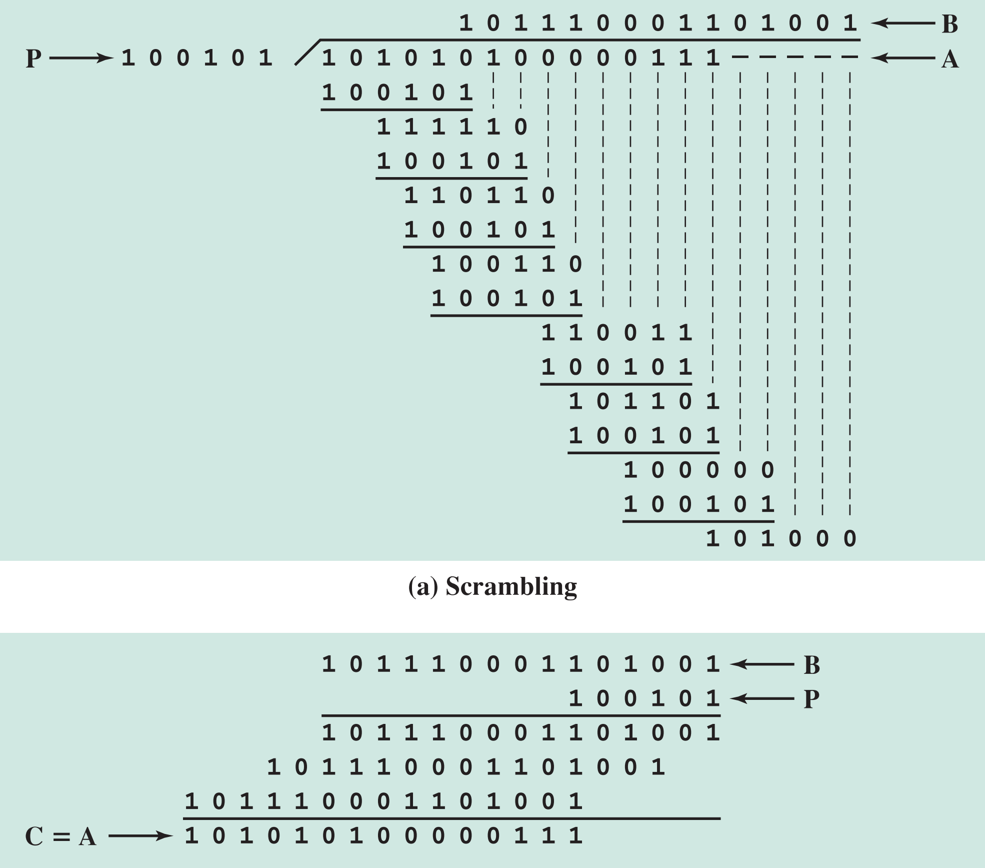
For the MLT-3 scheme, which is used for 100BASE-TX, the scrambling equation is:

$$B_m = A_m \oplus X_9 \oplus X_{11}$$

In this case the shift register consists of nine elements, used in the same manner as the 5-element register in Figure 12.17. However, in the case of MLT-3, the shift register is not fed by the output . Instead, after each bit transmission, the register is shifted one unit up, and the result of the previous XOR is fed into the first unit. This can be expressed as:

$$\begin{aligned} X_i(t) &= X_{i-1}(t-1); \quad 2 \le i \le 9 \\ X_i(t) &= X_9(t-1) \oplus X_{11}(t-1) \end{aligned}$$

[5] We use the convention that the leftmost bit is the first bit presented to the scrambler; thus the bits can be labeled $A_0A_1A_2$.... Similarly, the polynomial is converted to a bit string from left to right. The polynomial $B_0 + B_1X + B_2X^2 + \ldots$ is represented as $B_0B_1B_2 \ldots$



(a) Scrambling

(b) Descrambling

Figure 12.18 Example of Scrambling with $P(X) = 1 + X^{-3} + X^{-5}$

If the shift register contains all zeros, no scrambling occurs (we just have $B_m = A_m$) and the above equations produce no change in the shift register. Accordingly, the standard calls for initializing the shift register with all ones and reinitializing the register to all ones when it takes on a value of all zeros.

For the 4D-PAM5 scheme, two scrambling equations are used, one in each direction:

$$B_m = A_m \oplus B_{m-13} \oplus B_{m-33}$$
$$B_m = A_m \oplus B_{m-20} \oplus B_{m-33}$$