CHAPTER 16

HIGH-SPEED LANS

Congratulations. I knew the record would stand until it was broken.

Yogi Berra

KEY POINTS

- The IEEE 802.3 standard, known as Ethernet, now encompasses data rates of 10 Mbps, 100 Mbps, 1 Gbps, and 10 Gbps. For the lower data rates, the CSMA/CD MAC protocol is used. For the 1-Gbps and 10-Gbps options, a switched technique is used.
- Fibre Channel is a switched network of nodes designed to provide high-speed linkages for such applications as storage area networks.
- A variety of signal encoding techniques are used in the various LAN standards to achieve efficiency and to make the high data rates practical.

Recent years have seen rapid changes in the technology, design, and commercial applications for local area networks (LANs). A major feature of this evolution is the introduction of a variety of new schemes for high-speed local networking. To keep pace with the changing local networking needs of business, a number of approaches to high speed LAN design have become commercial products. The most important of these are

- **Fast Ethernet and Gigabit Ethernet:** The extension of 10-Mbps CSMA/CD (carrier sense multiple access with collision detection) to higher speeds is a logical strategy because it tends to preserve the investment in existing systems.
- **Fibre Channel:** This standard provides a low-cost, easily scalable approach to achieving very high data rates in local areas.
- **High-speed wireless LANs:** Wireless LAN technology and standards have at last come of age, and high-speed standards and products are being introduced.

Table 16.1 lists some of the characteristics of these approaches. The remainder of this chapter fills in some of the details on Ethernet and Fibre Channel. Chapter 17 covers wireless LANs.

16.1 THE EMERGENCE OF HIGH-SPEED LANs

Personal computers and microcomputer workstations began to achieve widespread acceptance in business computing in the early 1980s and have now achieved the status of the telephone: an essential tool for office workers. Until relatively recently, office LANs provided basic connectivity services—connecting personal computers

Table 16.1 Characteristics of Some High-Speed LANs

	Fast Ethernet	Gigabit Ethernet	Fibre Channel	Wireless LAN
Data Rate	100 Mbps	1 Gbps, 10 Gbps	100 Mbps–3.2 Gbps	1 Mbps–54 Mbps
Transmission Media	UTP, STP, optical Fiber	UTP, shielded cable, optical fiber	Optical fiber, coaxial cable, STP	2.4-GHz, 5-GHz microwave
Access Method	CSMA/CD	Switched	Switched	CSMA/Polling
Supporting Standard	IEEE 802.3	IEEE 802.3	Fibre Channel Association	IEEE 802.11

and terminals to mainframes and midrange systems that ran corporate applications, and providing workgroup connectivity at the departmental or divisional level. In both cases, traffic patterns were relatively light, with an emphasis on file transfer and electronic mail. The LANs that were available for this type of workload, primarily Ethernet and token ring, are well suited to this environment.

In recent years, two significant trends have altered the role of the personal computer and therefore the requirements on the LAN:

- The speed and computing power of personal computers has continued to enjoy explosive growth. Today's more powerful platforms support graphics-intensive applications and ever more elaborate graphical user interfaces to the operating system.
- MIS organizations have recognized the LAN as a viable and indeed essential computing platform, resulting in the focus on network computing. This trend began with client/server computing, which has become a dominant architecture in the business environment and the more recent intranetwork trend. Both of these approaches involve the frequent transfer of potentially large volumes of data in a transaction-oriented environment.

The effect of these trends has been to increase the volume of data to be handled over LANs and, because applications are more interactive, to reduce the acceptable delay on data transfers. The earlier generation of 10-Mbps Ethernets and 16-Mbps token rings are simply not up to the job of supporting these requirements.

The following are examples of requirements that call for higher-speed LANs:

- **Centralized server farms:** In many applications, there is a need for user, or client, systems to be able to draw huge amounts of data from multiple centralized servers, called server farms. An example is a color publishing operation, in which servers typically contain hundreds of gigabytes of image data that must be downloaded to imaging workstations. As the performance of the servers themselves has increased, the bottleneck has shifted to the network.
- **Power workgroups:** These groups typically consist of a small number of cooperating users who need to draw massive data files across the network. Examples are a software development group that runs tests on a new software version, or a computer-aided design (CAD) company that regularly runs simulations of new designs. In such cases, large amounts of data are distributed to several workstations, processed, and updated at very high speed for multiple iterations.

- **High-speed local backbone:** As processing demand grows, LANs proliferate at a site, and high-speed interconnection is necessary.

16.2 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 considered 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.

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.

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.

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 16.1. 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.

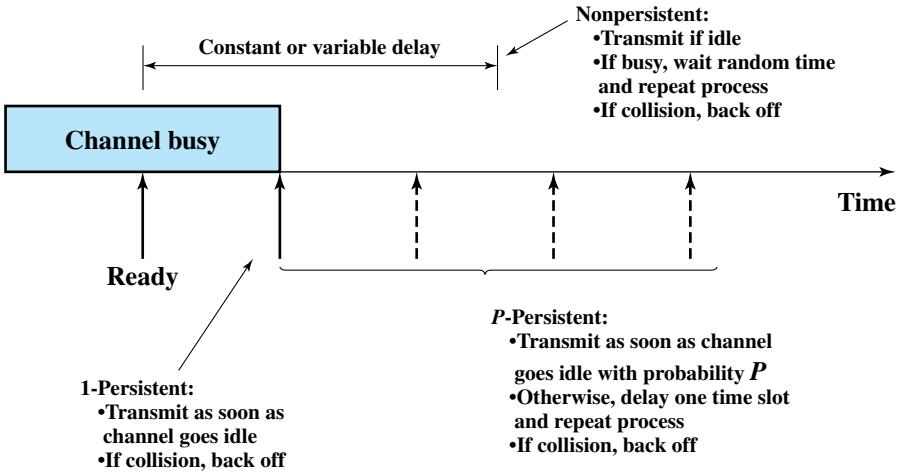


Figure 16.1 CSMA Persistence and Backoff

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.

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 as follows:

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.

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

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 16.2). 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 16.2 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 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 16.2, 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,

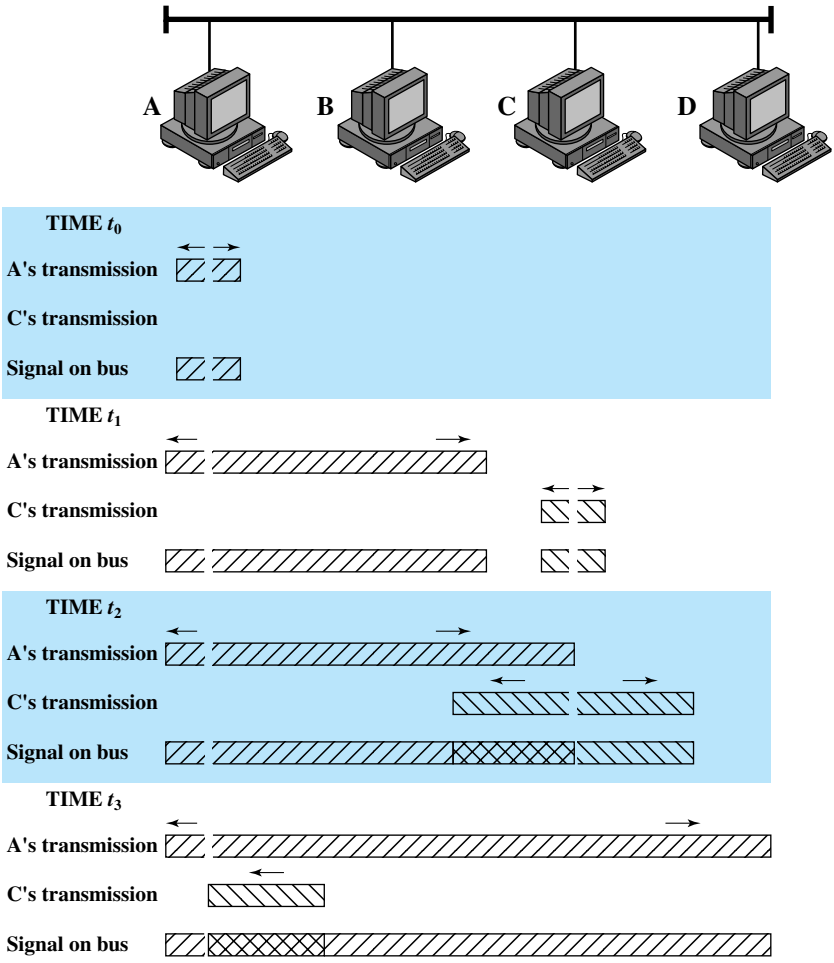


Figure 16.2 CSMA/CD Operation

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 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.

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.

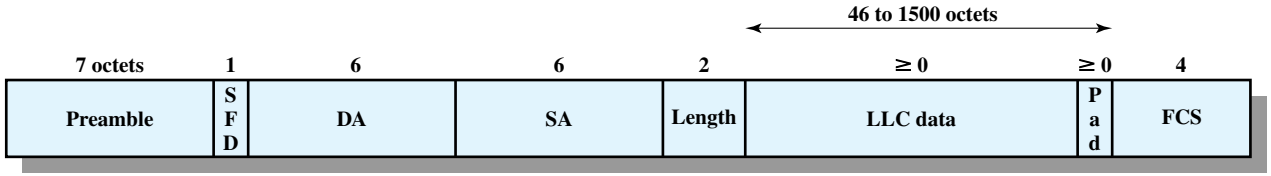
The beauty of the 1-persistent algorithm with binary exponential backoff is that it is efficient over a wide range of loads. At low loads, 1-persistence guarantees that a station can seize the channel as soon as it goes idle, in contrast to the non- and p-persistent schemes. At high loads, it is at least 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.

A much simpler collision detection scheme is possible with the twisted-pair star-topology approach (Figure 15.12). 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.

MAC Frame Figure 16.3 depicts the frame format for the 802.3 protocol. It consists of the following fields:

- **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 indicates the actual start of the frame and enables the receiver to locate the first bit of the rest of the frame.
- **Destination Address (DA):** Specifies the station(s) for which the frame is intended. It may be a unique physical address, a group address, or a global address.



SFD = Start of frame delimiter
DA = Destination address
SA = Source address
FCS = Frame check sequence

Figure 16.3 IEEE 802.3 Frame Format

- **Source Address (SA):** Specifies the station that sent the frame.
- **Length/Type:** Length of LLC data field in octets, or Ethernet Type field, depending on whether the frame conforms to the IEEE 802.3 standard or the earlier Ethernet specification. In either case, the maximum frame size, excluding the Preamble and SFD, is 1518 octets.
- **LLC Data:** Data unit supplied by LLC.
- **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.

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 as follows:[1]

- **10BASE5:** Specifies the use of 50-ohm coaxial cable and Manchester digital signaling.[2] The maximum length of a cable segment is set at 500 meters. 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 kilometers.
- **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.
- **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 meters. As an alternative, an optical fiber link may be used. In this case, the maximum length is 500 m.

---

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

Table 16.2 IEEE 802.3 10-Mbps Physical Layer Medium Alternatives

	10BASE5	10BASE2	10BASE-T	10BASE-FP
Transmission medium	Coaxial cable (50 ohm)	Coaxial cable (50 ohm)	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 to 0.6	62.5/125 μm

- **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 or repeaters at up to 2 km; a point-to-point link that can be used to connect repeaters at up to 2 km.

Note that 10BASE-T and 10-BASE-F do not quite follow the notation: "T" stands for twisted pair and "F" stands for optical fiber. Table 16.2 summarizes the remaining options. All of the alternatives listed in the table specify a data rate of 10 Mbps.

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 16.3 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] 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.

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

Table 16.3 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

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 16A 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 16A).

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.

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.

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 16A).

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.

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.

Gigabit Ethernet

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 move to 100BASE-T, putting huge traffic loads on backbone networks, demand for Gigabit Ethernet has intensified.

Figure 16.4 shows a typical application of Gigabit Ethernet. A 1-Gbps switching hub provides backbone connectivity for central servers and high-speed workgroup hubs. Each workgroup LAN switch supports both 1-Gbps links, to connect to the backbone LAN switch and to support high-performance workgroup servers, and 100-Mbps links, to support high-performance workstations, servers, and 100-Mbps LAN switches.

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 15.13b), there are two enhancements to the basic CSMA/CD scheme:

- **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.

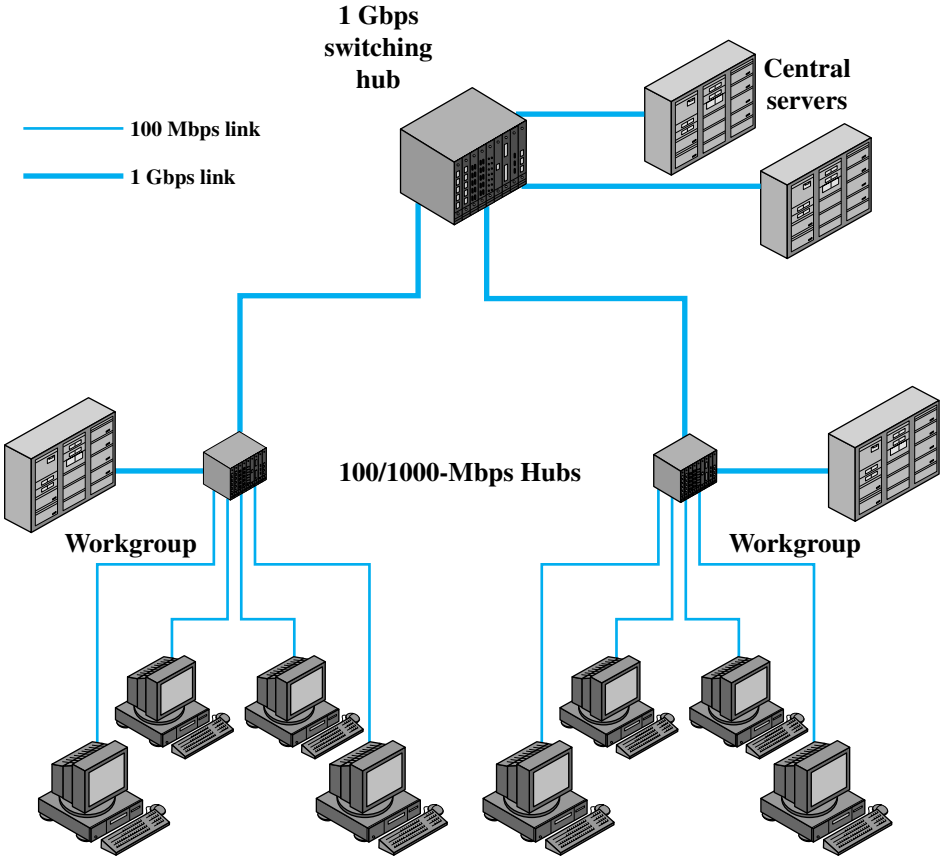


Figure 16.4 Example Gigabit Ethernet Configuration

- **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.

With a switching hub (Figure 15.13c), 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 current 1-Gbps specification for IEEE 802.3 includes the following physical layer alternatives (Figure 16.5):

- **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.

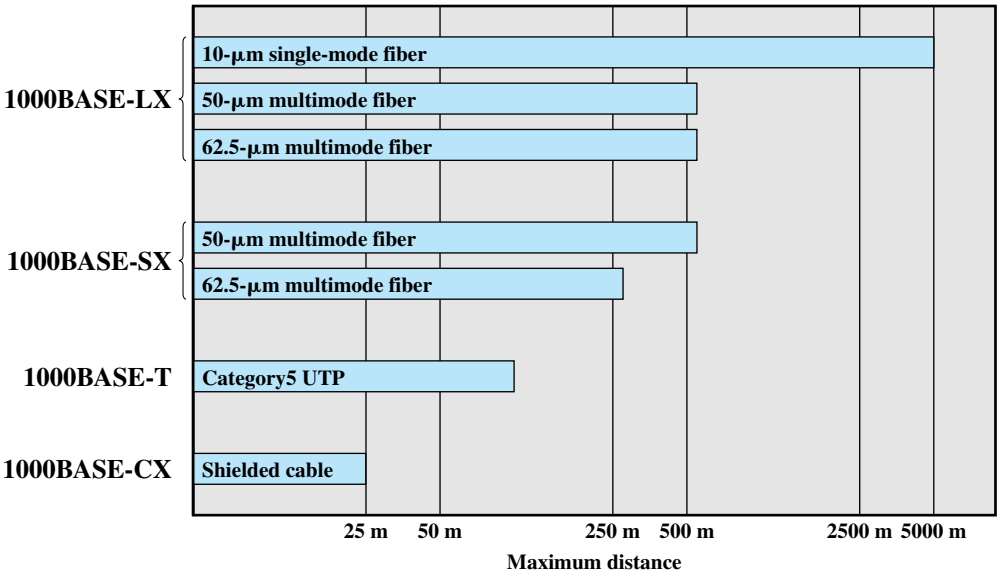


Figure 16.5 Gigabit Ethernet Medium Options (log scale)

- **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.

The signal encoding scheme used for the first three Gigabit Ethernet options just listed is 8B/10B, which is described in Appendix 16A. The signal-encoding scheme used for 1000BASE-T is 4D-PAM5, a complex scheme whose description is beyond our scope.

10-Gbps Ethernet

With gigabit products still fairly new, attention has turned in the past several years to a 10-Gbps Ethernet capability. The principle driving requirement for 10 Gigabit Ethernet is the increase in Internet and intranet traffic. A number of factors contribute to the explosive growth in both Internet and intranet traffic:

- 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
- An increase in Web hosting and application hosting traffic

Initially network managers will use 10-Gbps Ethernet to provide high-speed, local backbone interconnection between large-capacity switches. As the demand for bandwidth increases, 10-Gbps Ethernet will be deployed throughout the entire network and will 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.

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:

- 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 16.6 illustrates potential uses of 10-Gbps Ethernet. Higher-capacity backbone pipes will 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 already in widespread use.

The goal for maximum link distances cover 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 16.7). 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. 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 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

- **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.

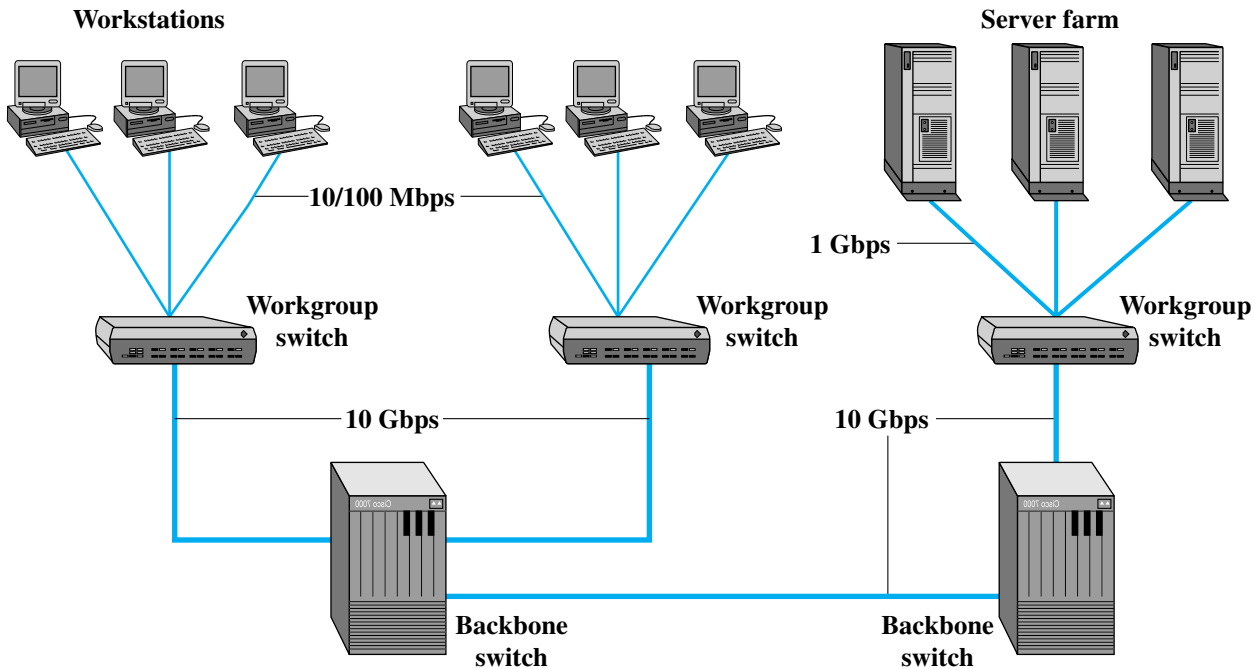


Figure 16.6 Example 10 Gigabit Ethernet Configuration

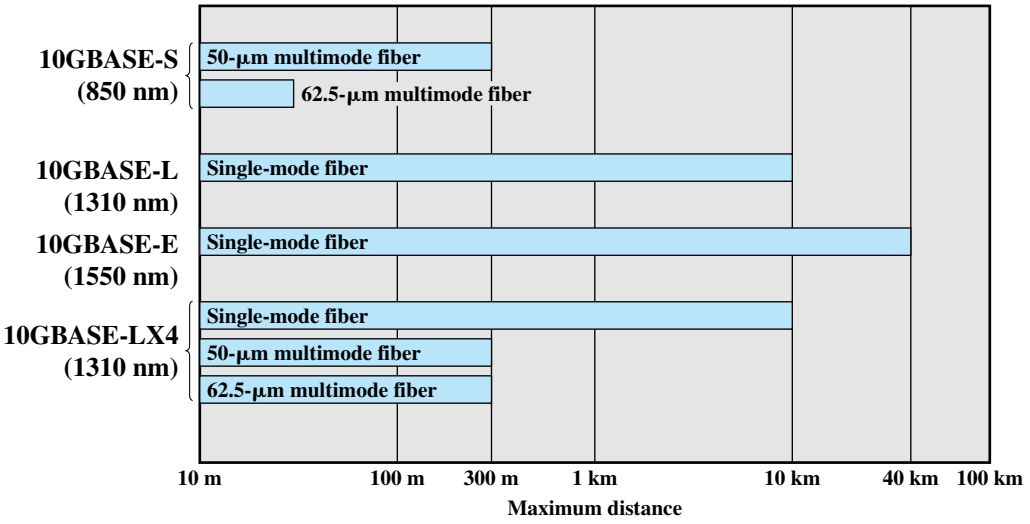


Figure 16.7 10-Gbps Ethernet Distance Options (log scale)

- **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.

The success of Fast Ethernet, Gigabit Ethernet, and 10-Gbps Ethernet highlights the importance of network management concerns in choosing a network technology. Both ATM and Fiber Channel, explored later, may be technically superior choices for a high-speed backbone, because of their flexibility and scalability. However, the Ethernet alternatives offer compatibility with existing installed LANs, network management software, and applications. This compatibility has accounted for the survival of a nearly 30-year-old technology (CSMA/CD) in today's fast-evolving network environment.

16.3 FIBRE CHANNEL

As the speed and memory capacity of personal computers, workstations, and servers have grown, and as applications have become ever more complex with greater reliance on graphics and video, the requirement for greater speed in delivering data to the processor has grown. This requirement affects two methods of data communications with the processor: I/O channel and network communications.

An I/O channel is a direct point-to-point or multipoint communications link, predominantly hardware based and designed for high speed over very short distances. The I/O channel transfers data between a buffer at the source device and a buffer at the destination device, moving only the user contents from one device to another, without regard to the format or meaning of the data. The logic associated

with the channel typically provides the minimum control necessary to manage the transfer plus hardware error detection. I/O channels typically manage transfers between processors and peripheral devices, such as disks, graphics equipment, CD-ROMs, and video I/O devices.

A network is a collection of interconnected access points with a software protocol structure that enables communication. The network typically allows many different types of data transfer, using software to implement the networking protocols and to provide flow control, error detection, and error recovery. As we have discussed in this book, networks typically manage transfers between end systems over local, metropolitan, or wide area distances.

Fibre Channel is designed to combine the best features of both technologies—the simplicity and speed of channel communications with the flexibility and interconnectivity that characterize protocol-based network communications. This fusion of approaches allows system designers to combine traditional peripheral connection, host-to-host internetworking, loosely coupled processor clustering, and multimedia applications in a single multiprotocol interface. The types of channel-oriented facilities incorporated into the Fibre Channel protocol architecture include

- Data-type qualifiers for routing frame payload into particular interface buffers
- Link-level constructs associated with individual I/O operations
- Protocol interface specifications to allow support of existing I/O channel architectures, such as the Small Computer System Interface (SCSI)

The types of network-oriented facilities incorporated into the Fibre Channel protocol architecture include

- Full multiplexing of traffic between multiple destinations
- Peer-to-peer connectivity between any pair of ports on a Fibre Channel network
- Capabilities for internetworking to other connection technologies

Depending on the needs of the application, either channel or networking approaches can be used for any data transfer. The Fibre Channel Industry Association, which is the industry consortium promoting Fibre Channel, lists the following ambitious requirements that Fibre Channel is intended to satisfy [FCIA01]:

- Full-duplex links with two fibers per link
- Performance from 100 Mbps to 800 Mbps on a single line (full-duplex 200 Mbps to 1600 Mbps per link)
- Support for distances up to 10 km
- Small connectors
- High-capacity utilization with distance insensitivity
- Greater connectivity than existing multidrop channels
- Broad availability (i.e., standard components)
- Support for multiple cost/performance levels, from small systems to supercomputers
- Ability to carry multiple existing interface command sets for existing channel and network protocols

The solution was to develop a simple generic transport mechanism based on point-to-point links and a switching network. This underlying infrastructure supports a simple encoding and framing scheme that in turn supports a variety of channel and network protocols.

Fibre Channel Elements

The key elements of a Fibre Channel network are the end systems, called **nodes**, and the network itself, which consists of one or more switching elements. The collection of switching elements is referred to as a **fabric**. These elements are interconnected by point-to-point links between ports on the individual nodes and switches. Communication consists of the transmission of frames across the point-to-point links.

Each node includes one or more ports, called N_ports, for interconnection. Similarly, each fabric-switching element includes multiple ports, called F_ports. Interconnection is by means of bidirectional links between ports. Any node can communicate with any other node connected to the same fabric using the services of the fabric. All routing of frames between N_ports is done by the fabric. Frames may be buffered within the fabric, making it possible for different nodes to connect to the fabric at different data rates.

A fabric can be implemented as a single fabric element with attached nodes (a simple star arrangement) or as a more general network of fabric elements, as shown in Figure 16.8. In either case, the fabric is responsible for buffering and for routing frames between source and destination nodes.

The Fibre Channel network is quite different from the IEEE 802 LANs. Fibre Channel is more like a traditional circuit-switching or packet-switching network, in contrast to the typical shared-medium LAN. Thus, Fibre Channel need not be concerned with medium access control issues. Because it is based on a switching network, the Fibre Channel scales easily in terms of N_ports, data rate, and distance covered.

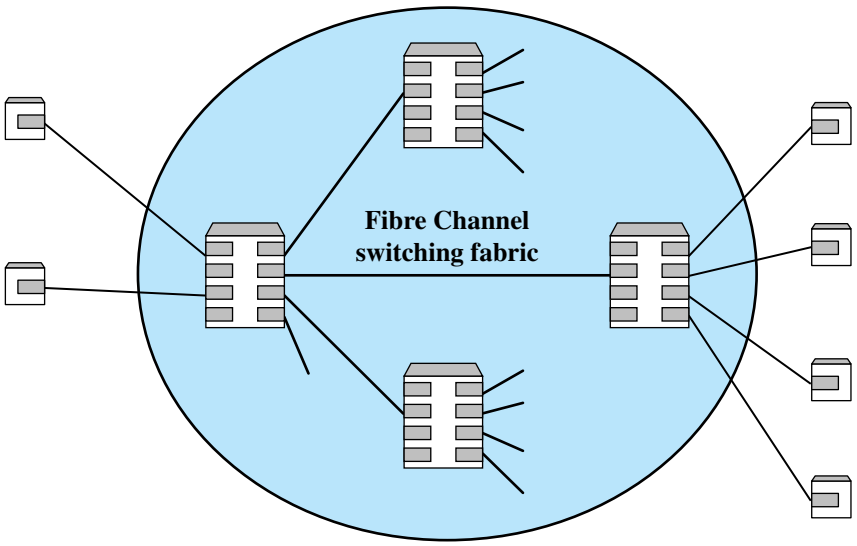


Figure 16.8 Fibre Channel Network

This approach provides great flexibility. Fibre Channel can readily accommodate new transmission media and data rates by adding new switches and F_ports to an existing fabric. Thus, an existing investment is not lost with an upgrade to new technologies and equipment. Further, the layered protocol architecture accommodates existing I/O interface and networking protocols, preserving the preexisting investment.

Fibre Channel Protocol Architecture

The Fibre Channel standard is organized into five levels. Each level defines a function or set of related functions. The standard does not dictate a correspondence between levels and actual implementations, with a specific interface between adjacent levels. Rather, the standard refers to the level as a "document artifice" used to group related functions. The layers are as follows:

- **FC-0 Physical Media:** Includes optical fiber for long-distance applications, coaxial cable for high speeds over short distances, and shielded twisted pair for lower speeds over short distances
- **FC-1 Transmission Protocol:** Defines the signal encoding scheme
- **FC-2 Framing Protocol:** Deals with defining topologies, frame format, flow and error control, and grouping of frames into logical entities called sequences and exchanges
- **FC-3 Common Services:** Includes multicasting
- **FC-4 Mapping:** Defines the mapping of various channel and network protocols to Fibre Channel, including IEEE 802, ATM, IP, and the Small Computer System Interface (SCSI)

Fibre Channel Physical Media and Topologies

One of the major strengths of the Fibre Channel standard is that it provides a range of options for the physical medium, the data rate on that medium, and the topology of the network (Table 16.4).

Transmission Media The transmission media options that are available under Fibre Channel include shielded twisted pair, video coaxial cable, and optical fiber. Standardized data rates range from 100 Mbps to 3.2 Gbps. Point-to-point link distances range from 33 m to 10 km.

Table 16.4 Maximum Distance for Fibre Channel Media Types

	800 Mbps	400 Mbps	200 Mbps	100 Mbps
Single mode fiber	10 km	10 km	10 km	—
50-μm multimode fiber	0.5 km	1 km	2 km	—
62.5-μm multimode fiber	175 m	1 km	1 km	—
Video coaxial cable	50 m	71 m	100 m	100 m
Miniature coaxial cable	14 m	19 m	28 m	42 m
Shielded twisted pair	28 m	46 m	57 m	80 m

Topologies The most general topology supported by Fibre Channel is referred to as a fabric or switched topology. This is an arbitrary topology that includes at least one switch to interconnect a number of end systems. The fabric topology may also consist of a number of switches forming a switched network, with some or all of these switches also supporting end nodes.

Routing in the fabric topology is transparent to the nodes. Each port in the configuration has a unique address. When data from a node are transmitted into the fabric, the edge switch to which the node is attached uses the destination port address in the incoming data frame to determine the destination port location. The switch then either delivers the frame to another node attached to the same switch or transfers the frame to an adjacent switch to begin routing the frame to a remote destination.

The fabric topology provides scalability of capacity: As additional ports are added, the aggregate capacity of the network increases, thus minimizing congestion and contention and increasing throughput. The fabric is protocol independent and largely distance insensitive. The technology of the switch itself and of the transmission links connecting the switch to nodes may be changed without affecting the overall configuration. Another advantage of the fabric topology is that the burden on nodes is minimized. An individual Fibre Channel node (end system) is only responsible for managing a simple point-to-point connection between itself and the fabric; the fabric is responsible for routing between ports and error detection.

In addition to the fabric topology, the Fibre Channel standard defines two other topologies. With the point-to-point topology there are only two ports, and these are directly connected, with no intervening fabric switches. In this case there is no routing. The arbitrated loop topology is a simple, low-cost topology for connecting up to 126 nodes in a loop. The arbitrated loop operates in a manner roughly equivalent to the token ring protocols that we have seen.

Topologies, transmission media, and data rates may be combined to provide an optimized configuration for a given site. Figure 16.9 is an example that illustrates the principal applications of Fiber Channel.

Prospects for Fibre Channel

Fibre Channel is backed by an industry interest group known as the Fibre Channel Association and a variety of interface cards for different applications are available. Fibre Channel has been most widely accepted as an improved peripheral device interconnect, providing services that can eventually replace such schemes as SCSI. It is a technically attractive solution to general high-speed LAN requirements but must compete with Ethernet and ATM LANs. Cost and performance issues should dominate the manager's consideration of these competing technologies.

16.4 RECOMMENDED READING AND WEB SITES

[STAL00] covers in greater detail the LAN systems discussed in this chapter.

[SPUR00] provides a concise but thorough overview of all of the 10-Mbps through 1-Gbps 802.3 systems, including configuration guidelines for a single segment of each media type, as well

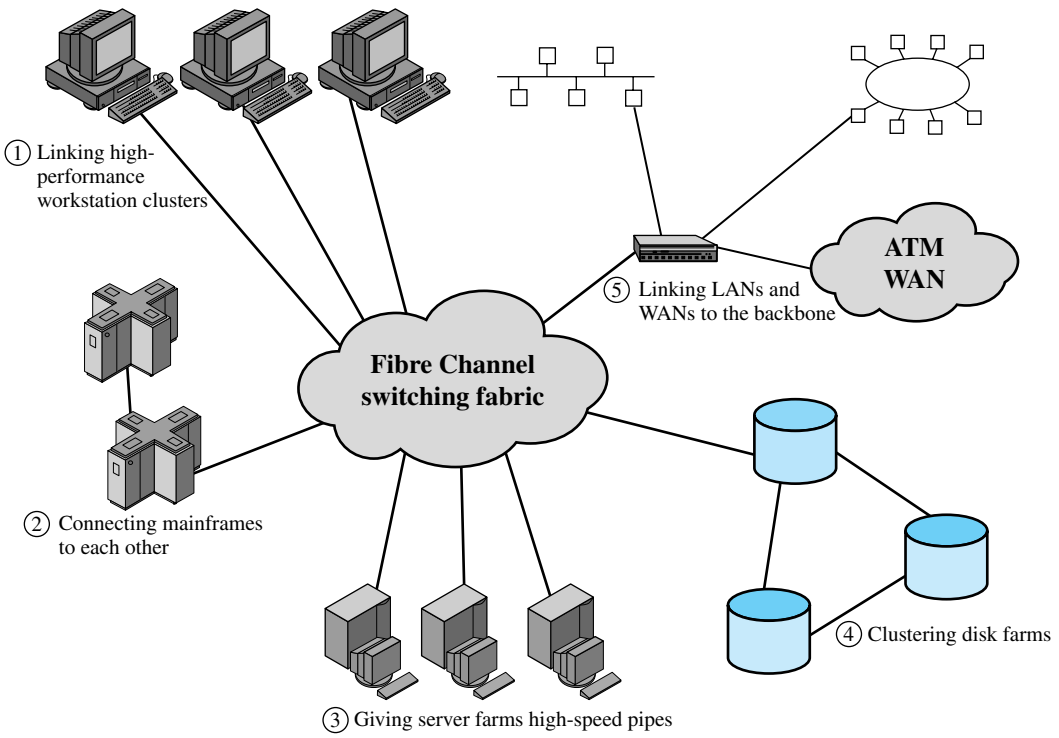


Figure 16.9 Five Applications of Fibre Channel

as guidelines for building multisegment Ethernets using a variety of media types. Two excellent treatments of both 100-Mbps and Gigabit Ethernet are [SEIF98] and [KADA98]. A good survey article on Gigabit Ethernet is [FRAZ99].

[SACH96] is a good survey of Fibre Channel. A short but worthwhile treatment is [FCIA01].

FCIA01 Fibre Channel Industry Association. *Fibre Channel Storage Area Networks.* San Francisco: Fibre Channel Industry Association, 2001.

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

KADA98 Kadambi, J.; Crayford, I.; and Kalkunte, M. *Gigabit Ethernet.* Upper Saddle River, NJ: Prentice Hall, 1998.

SACH96 Sachs. M., and Varma, A. "Fibre Channel and Related Standards." *IEEE Communications Magazine*, August 1996.

SEIF98 Seifert, R. *Gigabit Ethernet.* Reading, MA: Addison-Wesley, 1998.

SPUR00 Spurgeon, C. *Ethernet: The Definitive Guide.* Cambridge, MA: O'Reilly and Associates, 2000.

STAL00 Stallings, W. *Local and Metropolitan Area Networks, Sixth Edition.* Upper Saddle River, NJ: Prentice Hall, 2000.

Recommended Web sites:

- **Interoperability Lab:** University of New Hampshire site for equipment testing for high-speed LANs
- **Charles Spurgeon's Ethernet Web Site:** Provides extensive information about Ethernet, including links and documents
- **IEEE 802.3 10-Gbps Ethernet Task Force:** Latest documents
- **Fibre Channel Industry Association:** Includes tutorials, white papers, links to vendors, and descriptions of Fibre Channel applications
- **CERN Fibre Channel Site:** Includes tutorials, white papers, links to vendors, and descriptions of Fibre Channel applications
- **Storage Network Industry Association:** An industry forum of developers, integrators, and IT professionals who evolve and promote storage networking technology and solutions

16.5 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

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

Review Questions

16.1. What is a server farm?
16.2. Explain the three persistence protocols that can be used with CSMA.
16.3. What is CSMA/CD?
16.4. Explain binary exponential backoff.
16.5. What are the transmission medium options for Fast Ethernet?
16.6. How does Fast Ethernet differ from 10BASE-T, other than the data rate?
16.7. In the context of Ethernet, what is full-duplex operation?
16.8. List the levels of Fibre Channel and the functions of each level.
16.9. What are the topology options for Fibre Channel?

Problems

16.1 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?

16.2 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 $$\sum_{i=1}^{\infty} iX^{i-1} = \frac{1}{(1 - X)^2}.$$

16.3 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 $0 \leq r < 2^K$, 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?

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

16.5 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.

16.6 The most widely used MAC approach for a ring topology is token ring, defined in IEEE 802.5. The token ring technique is based on the use of a small frame, called a token, that circulates when all stations are idle. A station wishing to transmit must wait until it detects a token passing by. It then seizes the token by changing one bit in the token, which transforms it from a token to a start-of-frame sequence for a data frame. The station then appends and transmits the remainder of the fields needed to construct a data frame. When a station seizes a token and begins to transmit a data frame, there is no token on the ring, so other stations wishing to transmit must wait. The frame on the ring will make a round trip and be absorbed by the transmitting station. The transmitting station will insert a new token on the ring when both of the following conditions have been met: (1) The station has completed transmission of its frame. (2) The leading edge of the transmitted frame has returned (after a complete circulation of the ring) to the station.

a. An option in IEEE 802.5, known as early token release, eliminates the second condition just listed. Under what conditions will early token release result in improved utilization?

b. Are there any potential disadvantages to early token release? Explain.

16.7 For a token ring LAN, suppose that the destination station removes the data frame and immediately sends a short acknowledgment frame to the sender rather than letting the original frame return to sender. How will this affect performance?

16.8 Another medium access control technique for rings is the slotted ring. A number of fixed-length slots circulate continuously on the ring. Each slot contains a leading bit to designate the slot as empty or full. A station wishing to transmit waits until an empty slot arrives, marks the slot full, and inserts a frame of data as the slot goes by. The full slot makes a complete round trip, to be marked empty again by the station that marked it full. In what sense are the slotted ring and token ring protocols the complement (dual) of each other?

16.9 Consider a slotted ring of length 10 km with a data rate of 10 Mbps and 500 repeaters, each of which introduces a 1-bit delay. Each slot contains room for one source address byte, one destination address byte, two data bytes, and five control bits for a total length of 37 bits. How many slots are on the ring?

16.10 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?

16.11 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 − +.

16.12 Draw the MLT decoder state diagram that corresponds to the encoder state diagram of Figure 16.10.

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

16.14 Consider a token ring system with N stations in which a station that has just transmitted a frame releases a new token only after the station has completed transmission of its frame and the leading edge of the transmitted frame has returned (after a complete circulation of the ring) to the station.

a. Show that utilization can be approximated by $1/(1 + a/N)$ for $a < 1$ and by $1/(a + a/N)$ for $a > 1$,

b. What is the asymptotic value of utilization as N increases?

16.15 **a.** Verify that the division illustrated in Figure 16.18a corresponds to the implementation of Figure 16.17a by calculating the result step by step using Equation (16.7).

b. Verify that the multiplication illustrated in Figure 16.18b corresponds to the implementation of Figure 16.17b by calculating the result step by step using Equation (16.8).

16.16 Draw a figure similar to Figure 16.17 for the MLT-3 scrambler and descrambler.

APPENDIX 16A 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 one and one signal state represents binary zero. 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.

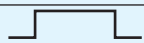
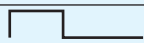
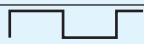
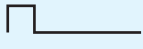
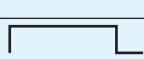
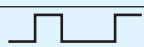
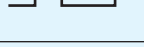
Greater efficiency can be achieved using the 4B/5B code. In this scheme, encoding is done 4 bits at a time; each 4 bits of data are encoded into a symbol with five *code bits*, such that each code bit contains a single signal element; the block of five code bits is called a *code group.* In effect, each 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 16.5 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

Table 16.5 4B/5B Code Groups (page 1 of 2)

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

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.
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 16.5. 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 ones, 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 one 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 16.10:

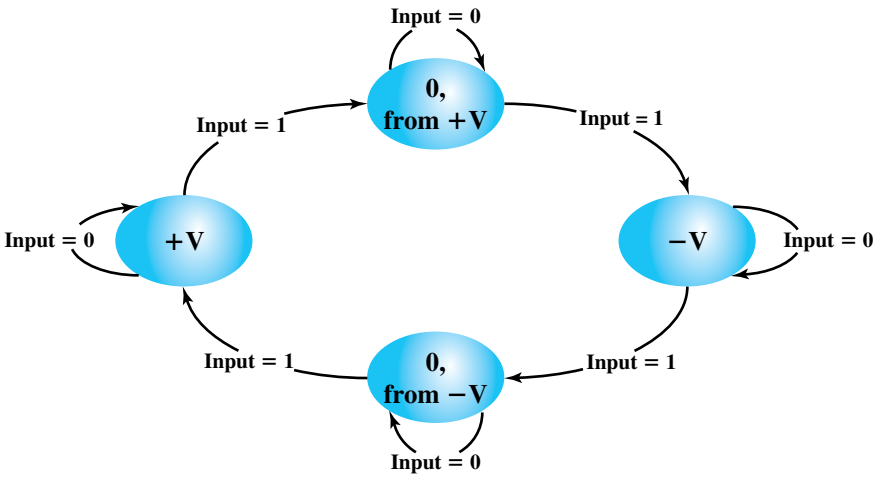


Figure 16.10 MLT-3 Encoder State Diagram

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 16.11 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 16.12). Thus the ternary transmission rate on each output channel is

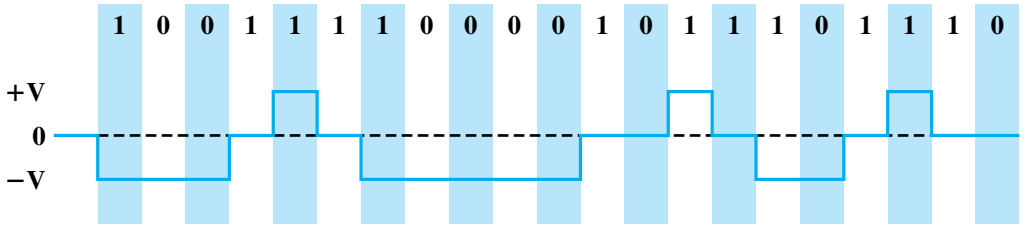


Figure 16.11 Example of MLT-3 Encoding

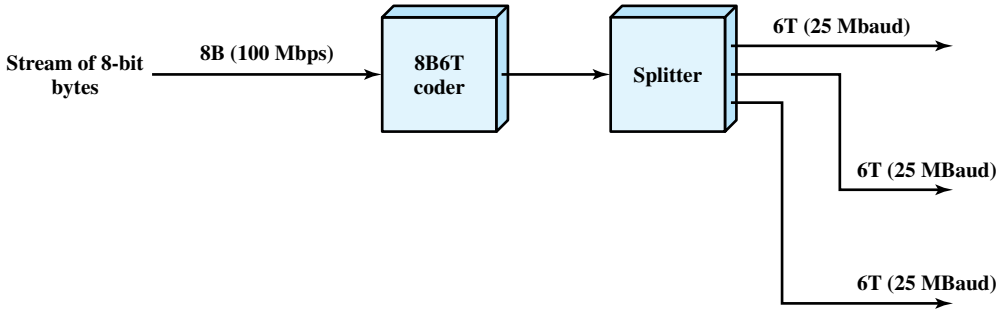


Figure 16.12 8B6T Transmission Scheme

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

Table 16.6 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 requirements in mind: synchronization and DC balance. For synchronization, the codes were chosen so 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 the of all code groups transmitted on a single pair. Each code group has a weight of 0 or 1. To maintain balance, the

Table 16.6 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 −

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.

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 *m*B*n*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 Ethernet. Figure 16.13 illustrates the process. The entire

(a) Data octets only

(b) Mixed data/control block

Figure 16.13 Encoding Using 64B/66B

Ethernet frame, including control fields, is considered "data" for this pzrocess. 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. For a 64-bit block consisting only of data octets, the entire block is scrambled. Two synchronization bits, with values 01, are prepended to the scrambled block. For a block consisting a mixture of control and data octets, a 56-bit block is used, which is scrambled; a 66-bit block is formed by prepending two synchronization bits, with values 10, and an 8-bit control type field, which defines the control functions included with this block. In both cases, scrambling is performed using the polynomial $1 + X^{39} + X^{58}$. See Appendix 16C for a discussion of scrambling. The two-bit synchronization field provides block alignment and a means of synchronizing when long streams of bits are sent.

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

APPENDIX 16B PERFORMANCE ISSUES

The choice of a LAN or MAN architecture is based on many factors, but one of the most important is performance. Of particular concern is the behavior (throughput, response time) of the network under heavy load. In this appendix, we provide an introduction to this topic. A more detailed discussion can be found in [STAL00].

The Effect of Propagation Delay and Transmission Rate

In Chapter 7, we introduced the parameter a, defined as

$$a = \frac{\text{Propagation time}}{\text{Transmission time}}$$

In that context, we were concerned with a point-to-point link, with a given propagation time between the two endpoints and a transmission time for either a fixed or average frame size. It was shown that a could be expressed as

$$a = \frac{\text{Length of data link in bits}}{\text{Length of frame in bits}}$$

This parameter is also important in the context of LANs and MANs, and in fact determines an upper bound on utilization. Consider a perfectly efficient access mechanism that allows only one transmission at a time. As soon as one transmission is over, another station begins transmitting. Furthermore, the transmission is pure data; no overhead bits. What is the maximum possible utilization of the network? It can be expressed as the ratio of total throughput of the network to its data rate:

$$U = \frac{\text{Throughput}}{\text{Data rate}} \tag{16.1}$$

Now define, as in Chapter 7:

R = data rate of the channel
d = maximum distance between any two stations
V = velocity of signal propagation
L = average or fixed frame length

The throughput is just the number of bits transmitted per unit time. A frame contains L bits, and the amount of time devoted to that frame is the actual transmission time (L/R) plus the propagation delay (d/V). Thus

$$\text{Throughput} = \frac{L}{d/V + L/R} \tag{16.2}$$

But by our preceding definition of a,

$$a = \frac{d/V}{L/R} = \frac{Rd}{LV} \tag{16.3}$$

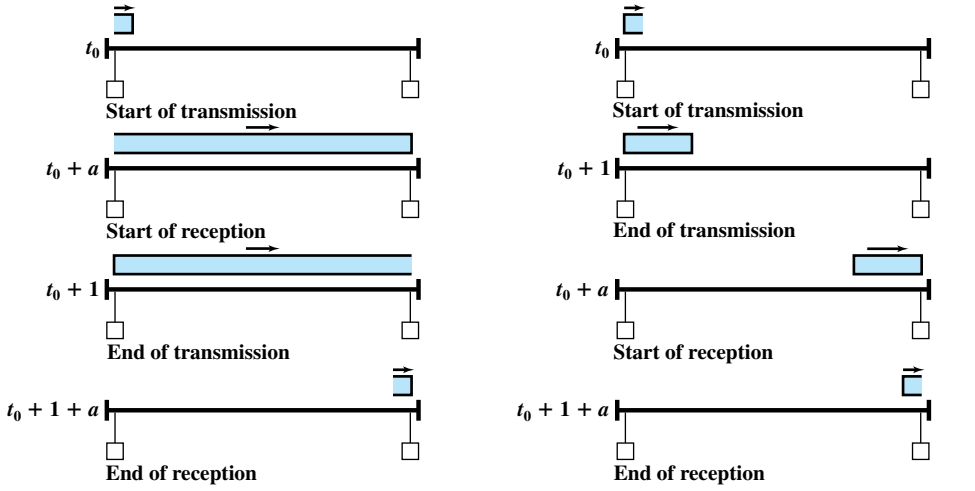
Substituting (16.2) and (16.3) into (16.1),

$$U = \frac{1}{1 + a} \tag{16.4}$$

Note that this differs from Equation (7.4) in Appendix 7A. This is because the latter assumed a half-duplex protocol (no piggybacked acknowledgments).

So utilization varies with a. This can be grasped intuitively by studying Figure 16.14, which shows a baseband bus with two stations as far apart as possible (worst case) that take turns sending frames. If we normalize time such that frame transmission time = 1, then the propagation time = a. For $a < 1$, the sequence of events is as follows:

1. A station begins transmitting at t_0.
2. Reception begins at $t_0 + a$.
3. Transmission is completed at $t_0 + 1$.
4. Reception ends at $t_0 + 1 + a$.
5. The other station begins transmitting.



(a) Transmission time = 1; propagation time = $a < 1$ **(b) Transmission time = 1; propagation time = $a > 1$**

Figure 16.14 The Effect of a on Utilization for Baseband Bus

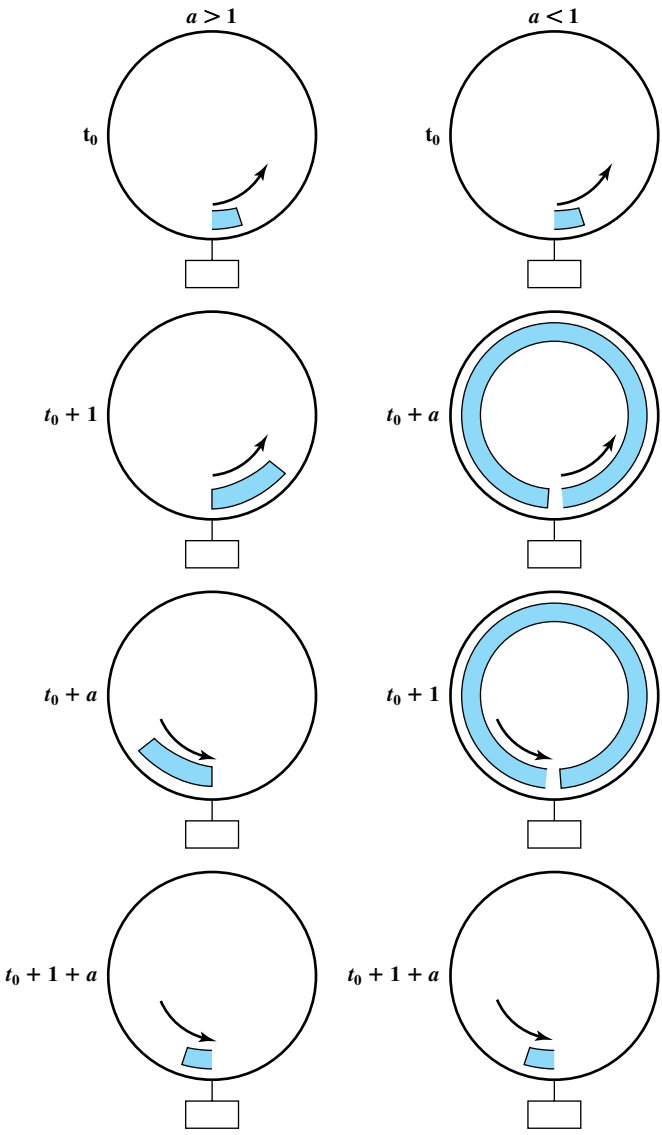


Figure 16.15 The Effect of a on Utilization for Ring

For $a > 1$, events 2 and 3 are interchanged. In both cases, the total time for one "turn" is $1 + a$, but the transmission time is only 1, for a utilization of $1/(1 + a)$.

The same effect can be seen to apply to a ring network in Figure 16.15. Here we assume that one station transmits and then waits to receive its own transmission before any other station transmits. The identical sequence of events just outlined applies.

Typical values of a range from about 0.01 to 0.1 for LANs and 0.1 to well over 1.0 for MANs. Table 16.7 gives some representative values for a bus topology. As can be seen, for larger and/or higher-speed networks, utilization suffers. For this reason, the restriction of only one frame at a time is lifted for high-speed LANs.

Table 16.7 Representative Values of *a*

Data Rate (Mbps)	Frame Size (bits)	Network Length (km)	*a*	1/(1 + *a*)
1	100	1	0.05	0.95
1	1,000	10	0.05	0.95
1	100	10	0.5	0.67
10	100	1	0.5	0.67
10	1,000	1	0.05	0.95
10	1,000	10	0.5	0.67
10	10,000	10	0.05	0.95
100	35,000	200	2.8	0.26
100	1,000	50	25	0.04

Finally, the preceding analysis assumes a "perfect" protocol, for which a new frame can be transmitted as soon as an old frame is received. In practice, the MAC protocol adds overhead that reduces utilization. This is demonstrated in the next subsection.

Simple Performance Model of CSMA/CD

The purpose of this section is to give the reader some insight into the performance of CSMA/CD by developing a simple performance models. It is hoped that this exercise will aid in understanding the results of more rigorous analyses.

For these models we assume a local network with N active stations and a maximum normalized propagation delay of a. To simplify the analysis, we assume that each station is always prepared to transmit a frame. This allows us to develop an expression for maximum achievable utilization (U). Although this should not be construed to be the sole figure of merit for a local network, it is the single most analyzed figure of merit and does permit useful performance comparisons.

Consider time on a bus medium to be organized into slots whose length is twice the end-to-end propagation delay. This is a convenient way to view the activity on the medium; the slot time is the maximum time, from the start of transmission, required to detect a collision. Assume that there are N active stations. Clearly, if each station always has a frame to transmit and does so, there will be nothing but collisions on the line. So we assume that each station restrains itself to transmitting during an available slot with probability P.

Time on the medium consists of two types of intervals. First is a transmission interval, which lasts $1/(2a)$ slots. Second is a contention interval, which is a sequence of slots with either a collision or no transmission in each slot. The throughput, normalized to system capacity, is the proportion of time spent in transmission intervals.

To determine the average length of a contention interval, we begin by computing A, the probability that exactly one station attempts a transmission in a slot and therefore acquires the medium. This is the binomial probability that any one station attempts to transmit and the others do not:

$$A = \binom{N}{1} P^1 (1 - P)^{N-1}$$
$$= NP(1 - P)^{N-1}$$

This function takes on a maximum over P when $P = 1/N$:

$$A = (1 - 1/N)^{N-1}$$

We are interested in the maximum because we want to calculate the maximum throughput of the medium. It should be clear that the maximum throughput will be achieved if we maximize the probability of successful seizure of the medium. Therefore, the following rule should be enforced: During periods of heavy usage, a station should restrain its offered load to $1/N$. (This assumes that each station knows the value of N. To derive an expression for maximum possible throughput, we live with this assumption.) On the other hand, during periods of light usage, maximum utilization cannot be achieved because the load is too low; this region is not of interest here.

Now we can estimate the mean length of a contention interval, w, in slots:

$$\begin{aligned} \mathrm{E}[w] &= \sum_{i=1}^{\infty} i \times \Pr\left(\begin{array}{l} i \text{ slots in a row with a collision or no} \\ \text{transmission followed by a slot with one} \\ \text{transmission} \end{array}\right) \\ &= \sum_{i=1}^{\infty} i(1 - A)^i A \end{aligned}$$

The summation converges to

$$\mathrm{E}[w] = \frac{1 - A}{A}$$

We can now determine the maximum utilization, which is the length of a transmission interval as a proportion of a cycle consisting of a transmission and a contention interval:

$$U = \frac{1/2a}{1/2a + (1 - A)/A} = \frac{1}{1 + 2a(1 - A)/A} \tag{16.5}$$

Figure 16.16 shows normalized throughput as a function of a for two values of N. Throughput declines as a increases. This is to be expected. Figure 16.16 also shows throughput as a function of N. The performance of CSMA/CD decreases because of the increased likelihood of collision or no transmission.

It is interesting to note the asymptotic value of U as N increases. We need to know that $\lim_{N\to\infty}\left(1 - \frac{1}{N}\right)^{N-1} = \frac{1}{\mathrm{e}}$. Then we have

$$\lim_{N\to\infty} U = \frac{1}{1 + 3.44a} \tag{16.6}$$

APPENDIX 16C SCRAMBLING

For some digital data encoding techniques, a long string of binary zeros or ones 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.

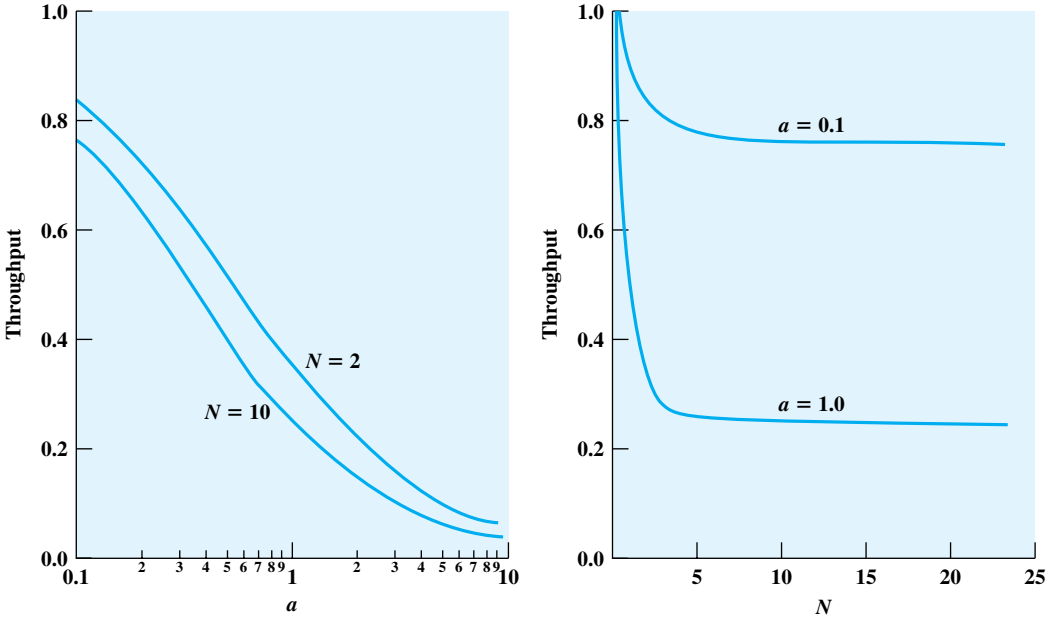


Figure 16.16 CSMA/CD Throughput as a Function of a and N

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 16.17. In this example, the scrambled data sequence may be expressed as follows:

$$B_m = A_m \oplus B_{m-3} \oplus B_{m-5} \tag{16.7}$$

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

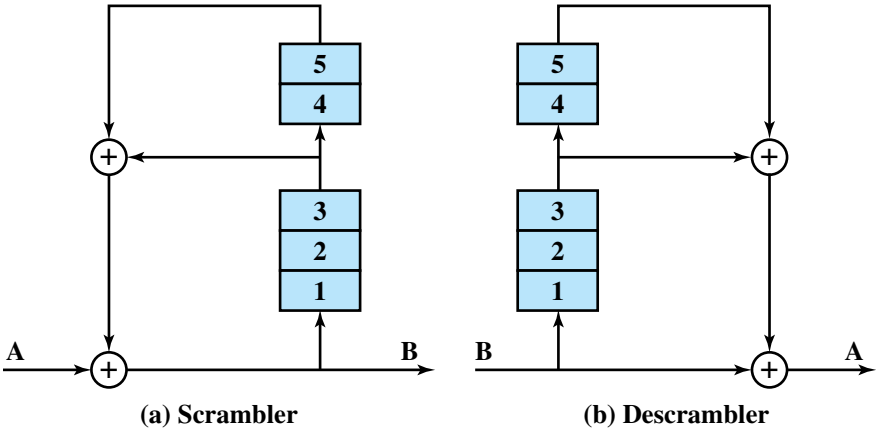
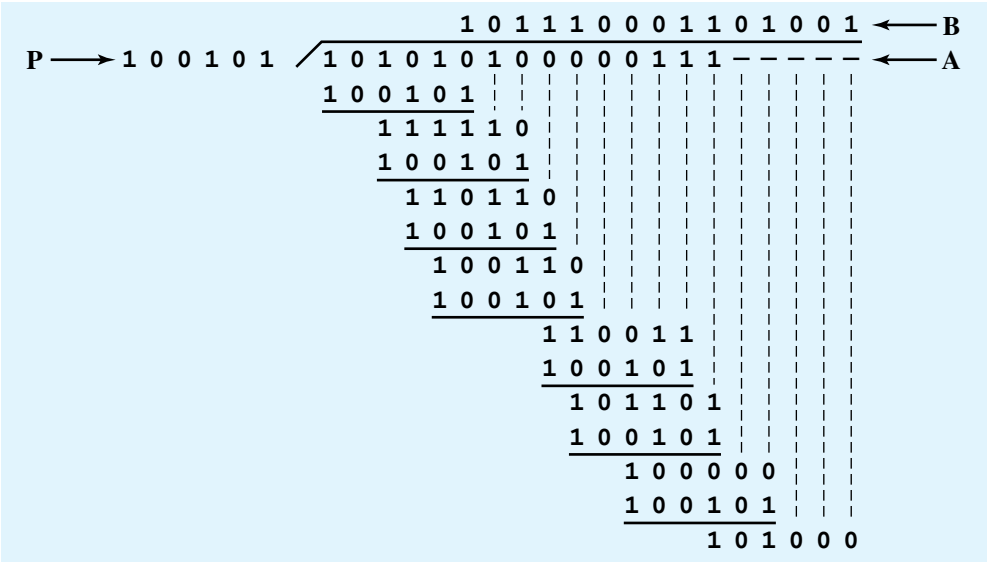
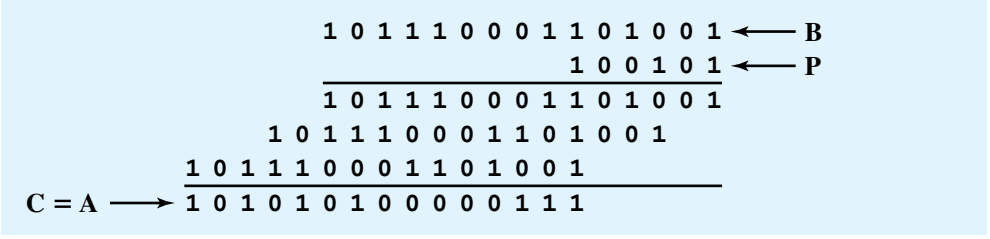


Figure 16.17 Scrambler and Descrambler



(a) Scrambling



(b) Descrambling

Figure 16.18 Example of Scrambling with $P(X) = 1 + X^{-3} + X^{-5}$

$$\begin{aligned} C_m &= B_m \oplus B_{m-3} \oplus B_{m-5} \\ &= (A_m \oplus B_{m-3} \oplus B_{m-5}) \oplus B_{m-3} \oplus B_{m-5} \\ &= A_m(\oplus B_{m-3} \oplus B_{m-3}) \oplus (B_{m-5} \oplus B_{m-5}) \\ &= A_m \end{aligned} \tag{16.8}$$

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 16.18 is an example using the polynomial $P(X)$ and an input of 101010100000111.[4] The scrambled transmission, produced by dividing by $P(X)$ (100101), is 101110001101001. When this number is multiplied by $P(X)$, we get the

[4]We use the convention that the leftmost bit is the first bit presented to the scrambler; thus the bits can be labeled $A_0A_1A_2\ldots$. 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 $B_0B_1B_2\ldots$.

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 16.17. However, in the case of MLT-3, the shift register is not fed by the output B_m. 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:

$$X_i(t) = X_{i-1}(t - 1); \qquad 2 \leq i \leq 9$$
$$X_1(t) = X_9(t - 1) \oplus X_{11}(t - 1)$$

If the shift register contains all zeros, no scrambling occurs (we just have $B_m = A_m$) the above equations produce no change in the shift register. Accordingly, the standard calls for initializing the shift register with all ones and re-initializing 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:

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