The term Ethernet refers to the family of local-area network (LAN) products covered by the IEEE 802.3 standard that defines what is commonly known as the CSMA/CD protocol. Three data rates are currently defined for operation over optical fiber and twisted-pair cables:

- 10 Mbps-10Base-T Ethernet
- 100 Mbps-Fast Ethernet
- 1000 Mbps-Gigabit Ethernet

10-Gigabit Ethernet is under development and will likely be published as the IEEE 802.3ae supplement to the IEEE 802.3 base standard in late 2001 or early 2002.

Other technologies and protocols have been touted as likely replacements, but the market has spoken. Ethernet has survived as the major LAN technology (it is currently used for approximately 85 percent of the world's LAN-connected PCs and workstations) because its protocol has the following characteristics:

- Is easy to understand, implement, manage, and maintain
- Allows low-cost network implementations
- Provides extensive topological flexibility for network installation
- Guarantees successful interconnection and operation of standards-compliant products, regardless of manufacturer

Guide Contents
Internetworking Basics
LAN Technologies
WAN Technologies
Internet Protocols
Bridging and Switching
Routing
Network Management
Voice/Data Integration Technologies
Wireless Technologies
Cable Access Technologies
Dial-up Technology
Security Technologies
Quality of Service Networking
Network Caching Technologies
IBM Network Management
Multiservice Access Technologies

Contents

Ethernet-A Brief History

The original Ethernet was developed as an experimental coaxial cable network in the 1970s by Xerox Corporation to operate with a data rate of 3 Mbps using a carrier sense multiple access collision detect (CSMA/CD) protocol for LANs with sporadic but occasionally heavy traffic requirements. Success with that project attracted early attention and led to the 1980 joint development of the 10-Mbps Ethernet Version 1.0 specification by the three-company consortium: Digital Equipment Corporation, Intel Corporation, and Xerox Corporation.

The original IEEE 802.3 standard was based on, and was very similar to, the Ethernet Version 1.0 specification. The draft standard was approved by the 802.3 working group in 1983 and was subsequently published as an official standard in 1985 (ANSI/IEEE Std. 802.3-1985). Since then, a number of supplements to the standard have been defined to take advantage of improvements in the technologies and to support additional network media and higher data rate capabilities, plus several new optional network access control features.

Throughout the rest of this article, the terms Ethernet and 802.3 will refer exclusively to network implementations compatible with the IEEE 802.3 standard.

Ethernet Network Elements

Ethernet LANs consist of network nodes and interconnecting media. The network nodes fall into two major classes:

- **Data terminal equipment (DTE)** - Devices that are either the source or the destination of data frames. DTEs are typically devices such as PCs, workstations, file servers, or print servers that, as a group, are all often referred to as end stations.
- **Data communication equipment (DCE)** - intermediate network devices that receive and forward frames across the network. DCEs may be either standalone devices such as repeaters, network switches, and routers, or communications interface units such as interface cards and modems.

Throughout this article, standalone intermediate network devices will be referred to as either *intermediate nodes or DCEs*. Network interface cards will be referred to as *NICs*.

The current Ethernet media options include two general types of copper cable: unshielded twisted-pair (UTP) and shielded twisted-pair (STP), plus several types of optical fiber cable.

Ethernet Network Topologies and Structures

LANs take on many topological configurations, but regardless of their size or complexity, all will be a combination of only three basic interconnection structures or network building blocks.

The simplest structure is the point-to-point interconnection, shown in <u>Figure: Example Point-to-Point Interconnection</u>. Only two network units are involved, and the connection may be DTE-to-DTE, DTE-to-DCE, or DCE-to-DCE. The cable in point-to-point interconnections is known as a network link. The maximum allowable length of the link depends on the type of cable and the transmission method that is used.

Figure: Example Point-to-Point Interconnection

The original Ethernet networks were implemented with a coaxial bus structure, as shown in <u>Figure: Example Coaxial Bus Topology</u>. Segment lengths were limited to 500 meters, and up to 100 stations could be connected to a single segment. Individual segments could be interconnected with repeaters, as long as multiple paths did not exist between any two stations on the network and the number of DTEs did not exceed 1024. The total path distance between the most-distant pair of stations was also not allowed to exceed a maximum prescribed value.

Figure: Example Coaxial Bus Topology
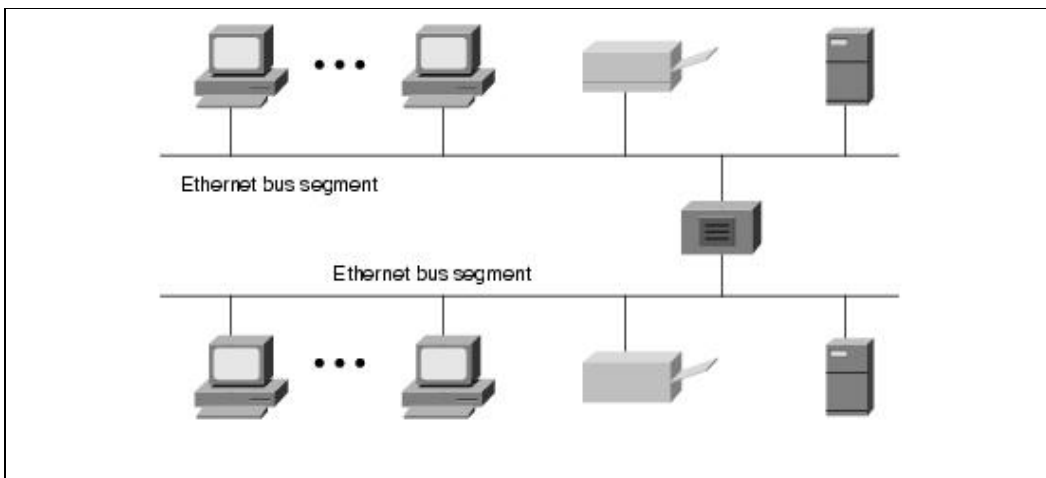
Although new networks are no longer connected in a bus configuration, some older bus-connected networks do still exist and are still useful.

Since the early 1990s, the network configuration of choice has been the star-connected topology, shown in <u>Figure: Example Star-Connected Topology</u>. The central network unit is either a multiport repeater (also known as a hub) or a network switch. All connections in a star network are point-to-point links implemented with either twisted-pair or optical fiber cable.

Figure: Example Star-Connected Topology
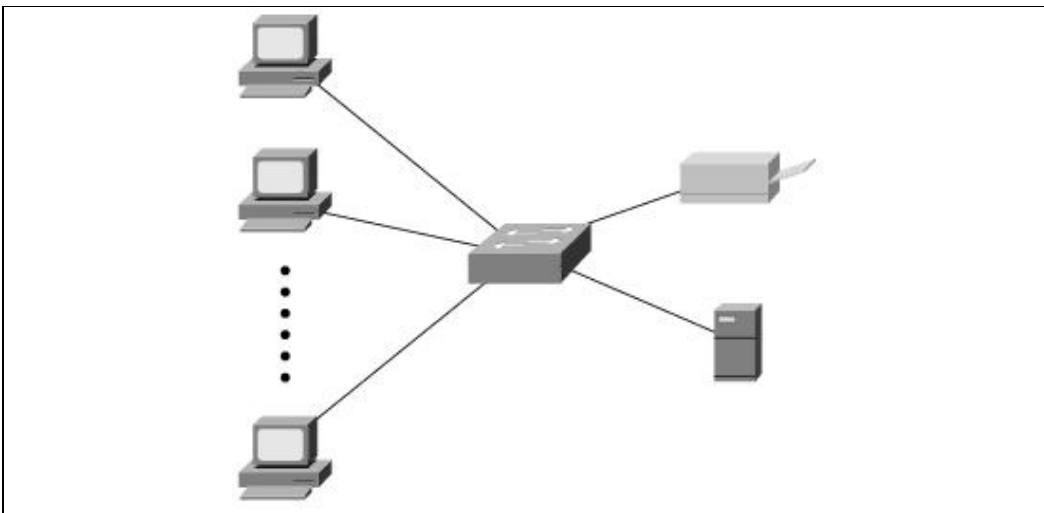
The IEEE 802.3 Logical Relationship to the ISO Reference Model

Figure: Ethernet?s Logical Relationship to the ISO Reference Model shows the IEEE 802.3 logical layers and their relationship to the OSI reference model. As with all IEEE 802 protocols, the ISO data link layer is divided into two IEEE 802 sublayers, the Media Access Control (MAC) sublayer and the MAC-client sublayer. The IEEE 802.3 physical layer corresponds to the ISO physical layer.

Figure: Ethernet?s Logical Relationship to the ISO Reference Model
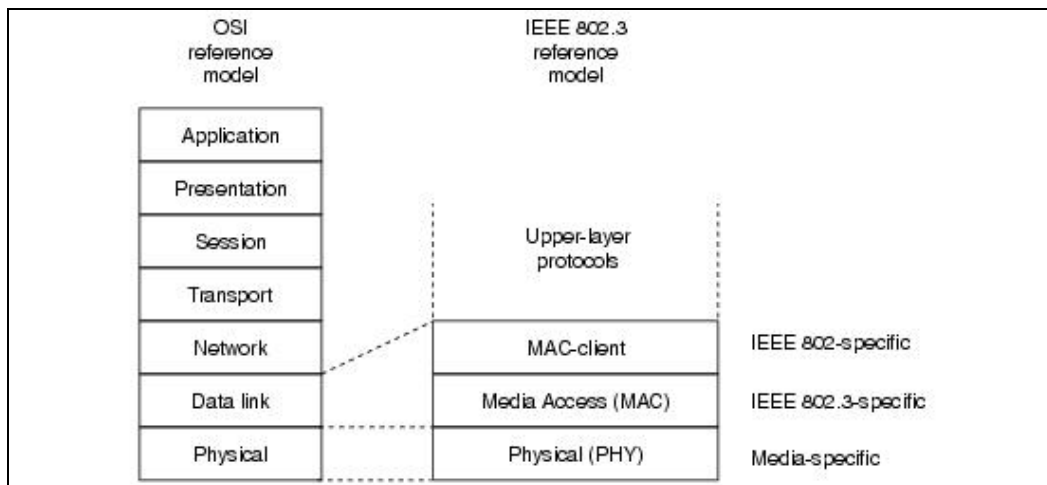
The MAC-client sublayer may be one of the following:

- Logical Link Control (LLC), if the unit is a DTE. This sublayer provides the interface between the Ethernet MAC and the upper layers in the protocol stack of the end station. The LLC sublayer is defined by IEEE 802.2 standards.
- Bridge entity, if the unit is a DCE. Bridge entities provide LAN-to-LAN interfaces between LANs that use the same protocol (for example, Ethernet to Ethernet) and also between different protocols (for example, Ethernet to Token Ring). Bridge entities are defined by IEEE 802.1 standards.

Because specifications for LLC and bridge entities are common for all IEEE 802 LAN protocols, network compatibility becomes the primary responsibility of the particular network protocol.

Figure: MAC and Physical Layer Compatibility Requirements for Basic Data Communication shows different compatibility requirements imposed by the MAC and physical levels for basic data communication

Figure: Example Star-Connected Topology   5

over an Ethernet link.

Figure: MAC and Physical Layer Compatibility Requirements for Basic Data Communication
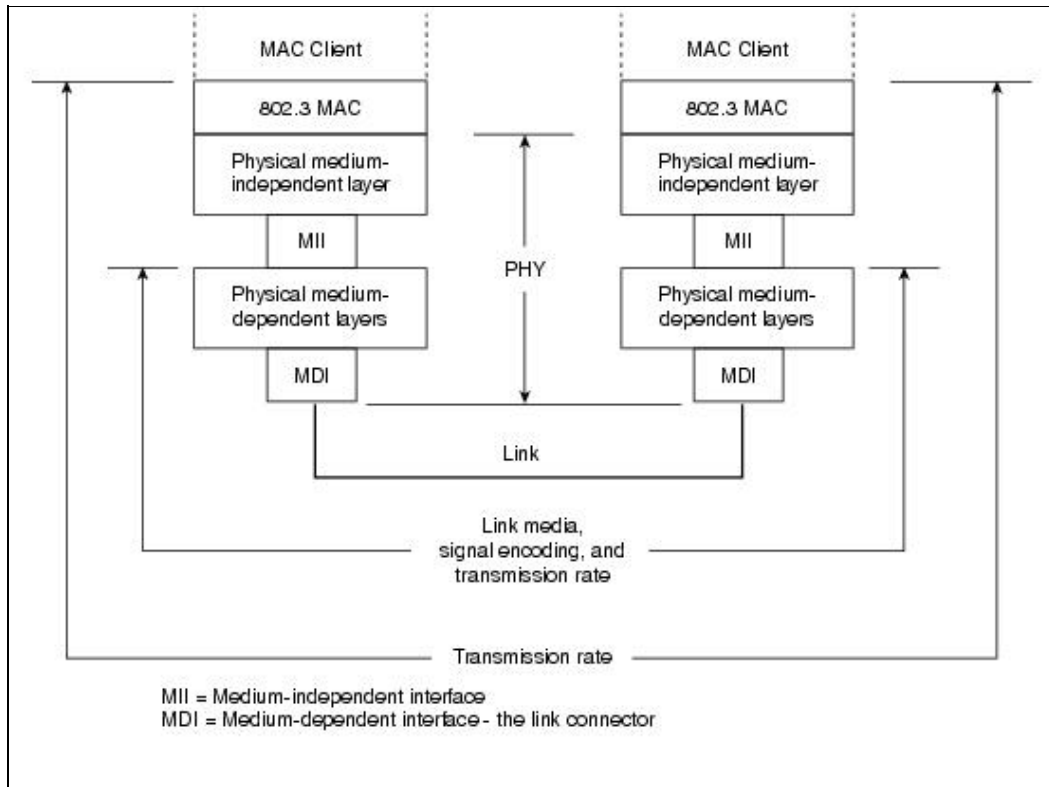
The MAC layer controls the node's access to the network media and is specific to the individual protocol. All IEEE 802.3 MACs must meet the same basic set of logical requirements, regardless of whether they include one or more of the defined optional protocol extensions. The only requirement for basic communication (communication that does not require optional protocol extensions) between two network nodes is that both MACs must support the same transmission rate.

The 802.3 physical layer is specific to the transmission data rate, the signal encoding, and the type of media interconnecting the two nodes. Gigabit Ethernet, for example, is defined to operate over either twisted-pair or optical fiber cable, but each specific type of cable or signal-encoding procedure requires a different physical layer implementation.

The Ethernet MAC Sublayer

The MAC sublayer has two primary responsibilities:

- Data encapsulation, including frame assembly before transmission, and frame parsing/error detection during and after reception
- Media access control, including initiation of frame transmission and recovery from transmission failure

The Basic Ethernet Frame Format

The IEEE 802.3 standard defines a basic data frame format that is required for all MAC implementations, plus several additional optional formats that are used to extend the protocol's basic capability. The basic data

Figure: Ethernet?s Logical Relationship to the ISO Reference Model   6

frame format contains the seven fields shown in Figure: The Basic IEEE 802.3 MAC Data Frame Format.

- **Preamble (PRE)** - Consists of 7 bytes. The PRE is an alternating pattern of ones and zeros that tells receiving stations that a frame is coming, and that provides a means to synchronize the frame-reception portions of receiving physical layers with the incoming bit stream.
- **Start-of-frame delimiter (SOF)** - Consists of 1 byte. The SOF is an alternating pattern of ones and zeros, ending with two consecutive 1-bits indicating that the next bit is the left-most bit in the left-most byte of the destination address.
- **Destination address (DA)** - Consists of 6 bytes. The DA field identifies which station(s) should receive the frame. The left-most bit in the DA field indicates whether the address is an individual address (indicated by a 0) or a group address (indicated by a 1). The second bit from the left indicates whether the DA is globally administered (indicated by a 0) or locally administered (indicated by a 1). The remaining 46 bits are a uniquely assigned value that identifies a single station, a defined group of stations, or all stations on the network.
- **Source addresses (SA)** - Consists of 6 bytes. The SA field identifies the sending station. The SA is always an individual address and the left-most bit in the SA field is always 0.
- **Length/Type** - Consists of 2 bytes. This field indicates either the number of MAC-client data bytes that are contained in the data field of the frame, or the frame type ID if the frame is assembled using an optional format. If the Length/Type field value is less than or equal to 1500, the number of LLC bytes in the Data field is equal to the Length/Type field value. If the Length/Type field value is greater than 1536, the frame is an optional type frame, and the Length/Type field value identifies the particular type of frame being sent or received.
- **Data** - Is a sequence of n bytes of any value, where n is less than or equal to 1500. If the length of the Data field is less than 46, the Data field must be extended by adding a filler (a pad) sufficient to bring the Data field length to 46 bytes.
- **Frame check sequence (FCS)** - Consists of 4 bytes. This sequence contains a 32-bit cyclic redundancy check (CRC) value, which is created by the sending MAC and is recalculated by the receiving MAC to check for damaged frames. The FCS is generated over the DA, SA, Length/Type, and Data fields.

Figure: The Basic IEEE 802.3 MAC Data Frame Format
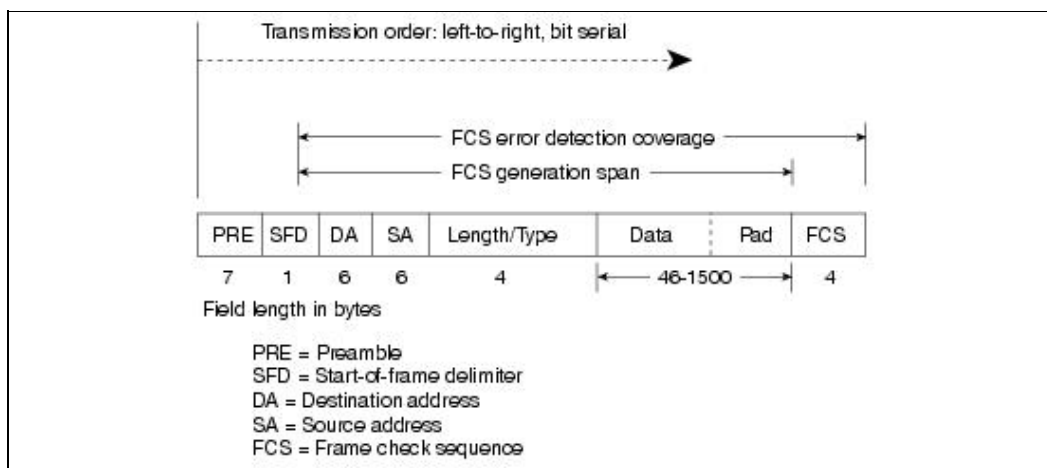
{{note:Individual addresses are also known as unicast addresses because they refer to a single MAC and are assigned by the NIC manufacturer from a block of addresses allocated by the IEEE. Group addresses (a.k.a. multicast addresses) identify the end stations in a workgroup and are assigned by the network manager. A special group address (all 1s-the broadcast address) indicates all stations on the network.}}

Frame Transmission

Whenever an end station MAC receives a transmit-frame request with the accompanying address and data information from the LLC sublayer, the MAC begins the transmission sequence by transferring the LLC information into the MAC frame buffer.

- The preamble and start-of-frame delimiter are inserted in the PRE and SOF fields.
- The destination and source addresses are inserted into the address fields.
- The LLC data bytes are counted, and the number of bytes is inserted into the Length/Type field.
- The LLC data bytes are inserted into the Data field. If the number of LLC data bytes is less than 46, a pad is added to bring the Data field length up to 46.
- An FCS value is generated over the DA, SA, Length/Type, and Data fields and is appended to the end of the Data field.

After the frame is assembled, actual frame transmission will depend on whether the MAC is operating in half-duplex or full-duplex mode.

The IEEE 802.3 standard currently requires that all Ethernet MACs support half-duplex operation, in which the MAC can be either transmitting or receiving a frame, but it cannot be doing both simultaneously. Full-duplex operation is an optional MAC capability that allows the MAC to transmit and receive frames simultaneously.

Half-Duplex Transmission-The CSMA/CD Access Method

The CSMA/CD protocol was originally developed as a means by which two or more stations could share a common media in a switch-less environment when the protocol does not require central arbitration, access tokens, or assigned time slots to indicate when a station will be allowed to transmit. Each Ethernet MAC determines for itself when it will be allowed to send a frame.

The CSMA/CD access rules are summarized by the protocol's acronym:

- **Carrier sense** - Each station continuously listens for traffic on the medium to determine when gaps between frame transmissions occur.
- **Multiple access** - Stations may begin transmitting any time they detect that the network is quiet (there is no traffic).
- **Collision detect** - If two or more stations in the same CSMA/CD network (collision domain) begin transmitting at approximately the same time, the bit streams from the transmitting stations will interfere (collide) with each other, and both transmissions will be unreadable. If that happens, each transmitting station must be capable of detecting that a collision has occurred before it has finished sending its frame. Each must stop transmitting as soon as it has detected the collision and then must wait a quasirandom length of time (determined by a back-off algorithm) before attempting to retransmit the frame.

The worst-case situation occurs when the two most-distant stations on the network both need to send a frame and when the second station does not begin transmitting until just before the frame from the first station arrives. The collision will be detected almost immediately by the second station, but it will not be detected by the first station until the corrupted signal has propagated all the way back to that station. The maximum time that is required to detect a collision (the collision window, or "slot time") is approximately equal to twice the signal propagation time between the two most-distant stations on the network.

This means that both the minimum frame length and the maximum collision diameter are directly related to the slot time. Longer minimum frame lengths translate to longer slot times and larger collision diameters; shorter minimum frame lengths correspond to shorter slot times and smaller collision diameters.

The trade-off was between the need to reduce the impact of collision recovery and the need for network diameters to be large enough to accommodate reasonable network sizes. The compromise was to choose a maximum network diameter (about 2500 meters) and then to set the minimum frame length long enough to ensure detection of all worst-case collisions.

The compromise worked well for 10 Mbps, but it was a problem for higher data-rate Ethernet developers. Fast Ethernet was required to provide backward compatibility with earlier Ethernet networks, including the existing IEEE 802.3 frame format and error-detection procedures, plus all applications and networking software running on the 10-Mbps networks.

Although signal propagation velocity is essentially constant for all transmission rates, the time required to transmit a frame is inversely related to the transmission rate. At 100 Mbps, a minimum-length frame can be transmitted in approximately one-tenth of the defined slot time, and any collision that occurred during the transmission would not likely be detected by the transmitting stations. This, in turn, meant that the maximum network diameters specified for 10-Mbps networks could not be used for 100-Mbps networks. The solution for Fast Ethernet was to reduce the maximum network diameter by approximately a factor of 10 (to a little more than 200 meters).

The same problem also arose during specification development for Gigabit Ethernet, but decreasing network diameters by another factor of 10 (to approximately 20 meters) for 1000-Mbps operation was simply not practical. This time, the developers elected to maintain approximately the same maximum collision domain diameters as 100-Mbps networks and to increase the apparent minimum frame size by adding a variable-length nondata extension field to frames that are shorter than the minimum length (the extension field is removed during frame reception).

Figure: MAC Frame with Gigabit Carrier Extension shows the MAC frame format with the gigabit extension field, and the following table shows the effect of the trade-off between the transmission data rate and the minimum frame size for 10-Mbps, 100-Mbps, and 1000-Mbps Ethernet.

Figure: MAC Frame with Gigabit Carrier Extension
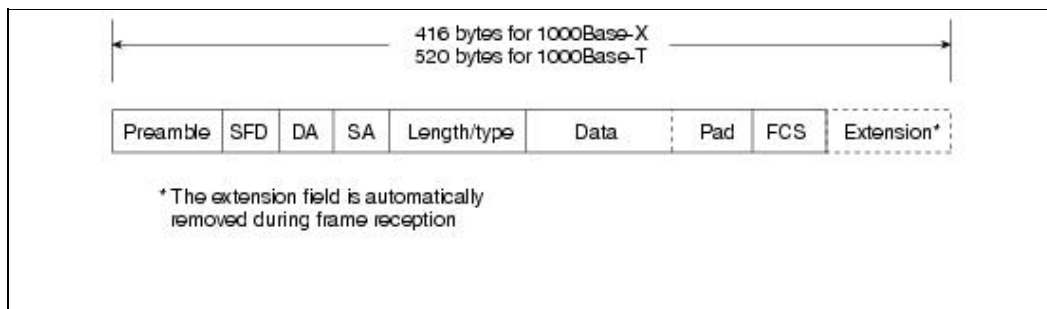
Table: Limits for Half-Duplex Operation

Parameter	10 Mbps	100 Mbps	1000 Mbps
Minimum frame size	64 bytes	64 bytes	520 bytes (with extension field added)
Maximum collision diameter, DTE to DTE	100 meters UTP	100 meters UTP 412 meters	100 meters UTP 316 meters fiber

		fiber	
Maximum collision diameter with repeaters	2500 meters	205 meters	200 meters
Maximum number of repeaters in network path	5	2	1

Note: 520 bytes applies to 1000Base-T implementations. The minimum frame size with extension field for 1000Base-X is reduced to 416 bytes because 1000Base-X encodes and transmits 10 bits for each byte.

Another change to the Ethernet CSMA/CD transmit specification was the addition of frame bursting for gigabit operation. Burst mode is a feature that allows a MAC to send a short sequence (a burst) of frames equal to approximately 5.4 maximum-length frames without having to relinquish control of the medium. The transmitting MAC fills each interframe interval with extension bits, as shown in Figure: A Gigabit Frame-Burst Sequence, so that other stations on the network will see that the network is busy and will not attempt transmission until after the burst is complete.

Figure: A Gigabit Frame-Burst Sequence
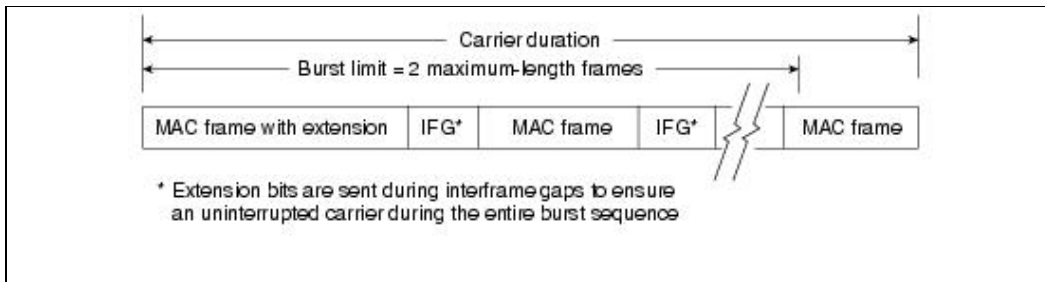
If the length of the first frame is less than the minimum frame length, an extension field is added to extend the frame length to the value indicated in Table: Limits for Half-Duplex Operation. Subsequent frames in a frame-burst sequence do not need extension fields, and a frame burst may continue as long as the burst limit has not been reached. If the burst limit is reached after a frame transmission has begun, transmission is allowed to continue until that entire frame has been sent.

Frame extension fields are not defined, and burst mode is not allowed for 10 Mbps and 100 Mbps transmission rates.

Full-Duplex Transmission-An Optional Approach to Higher Network Efficiency

Full-duplex operation is an optional MAC capability that allows simultaneous two-way transmission over point-to-point links. Full duplex transmission is functionally much simpler than half-duplex transmission because it involves no media contention, no collisions, no need to schedule retransmissions, and no need for extension bits on the end of short frames. The result is not only more time available for transmission, but also an effective doubling of the link bandwidth because each link can now support full-rate, simultaneous, two-way transmission.

Transmission can usually begin as soon as frames are ready to send. The only restriction is that there must be a minimum-length interframe gap between successive frames, as shown in Figure: Full Duplex Operation Allows Simultaneous Two-Way Transmission on the Same Link, and each frame must conform to Ethernet frame format standards.

Figure: Full Duplex Operation Allows Simultaneous Two-Way Transmission on the Same Link
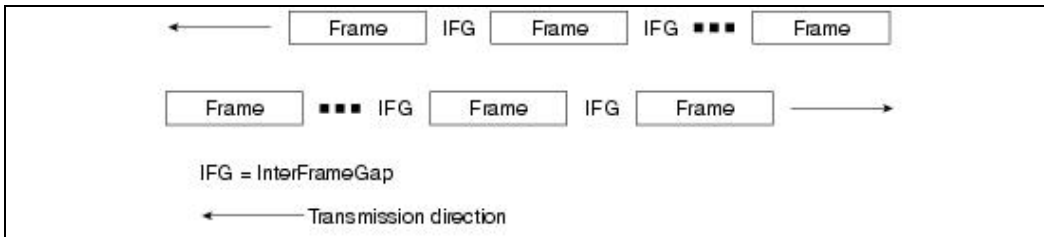
Flow Control

Full-duplex operation requires concurrent implementation of the optional flow-control capability that allows a receiving node (such as a network switch port) that is becoming congested to request the sending node (such as a file server) to stop sending frames for a selected short period of time. Control is MAC-to-MAC through the use of a pause frame that is automatically generated by the receiving MAC. If the congestion is relieved before the requested wait has expired, a second pause frame with a zero time-to-wait value can be sent to request resumption of transmission. An overview of the flow control operation is shown in Figure: An Overview of the IEEE 802.3 Flow Control Sequence.

Figure: An Overview of the IEEE 802.3 Flow Control Sequence
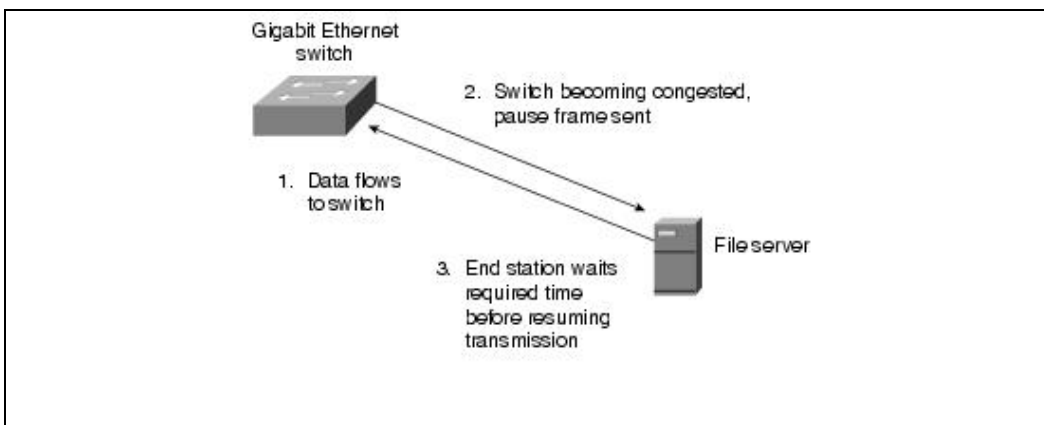
The full-duplex operation and its companion flow control capability are both options for all Ethernet MACs and all transmission rates. Both options are enabled on a link-by-link basis, assuming that the associated physical layers are also capable of supporting full-duplex operation.

Pause frames are identified as MAC control frames by an exclusive assigned (reserved) length/type value. They are also assigned a reserved destination address value to ensure that an incoming pause frame is never forwarded to upper protocol layers or to other ports in a switch.

Frame Reception

Frame reception is essentially the same for both half-duplex and full-duplex operations, except that full-duplex MACs must have separate frame buffers and data paths to allow for simultaneous frame transmission and reception.

Frame reception is the reverse of frame transmission. The destination address of the received frame is checked and matched against the station's address list (its MAC address, its group addresses, and the broadcast address) to determine whether the frame is destined for that station. If an address match is found, the frame length is checked and the received FCS is compared to the FCS that was generated during frame

reception. If the frame length is okay and there is an FCS match, the frame type is determined by the contents of the Length/Type field. The frame is then parsed and forwarded to the appropriate upper layer.

The VLAN Tagging Option

VLAN tagging is a MAC option that provides three important capabilities not previously available to Ethernet network users and network managers:

- Provides a means to expedite time-critical network traffic by setting transmission priorities for outgoing frames.
- Allows stations to be assigned to logical groups, to communicate across multiple LANs as though they were on a single LAN. Bridges and switches filter destination addresses and forward VLAN frames only to ports that serve the VLAN to which the traffic belongs.
- Simplifies network management and makes adds, moves, and changes easier to administer.

A VLAN-tagged frame is simply a basic MAC data frame that has had a 4-byte VLAN header inserted between the SA and Length/Type fields, as shown in Figure: VLAN-Tagged Frames Are Identified When the MAC Finds the LAN Type Value in the Normal Length/Type Field Location.

Figure: VLAN-Tagged Frames Are Identified When the MAC Finds the LAN Type Value in the Normal Length/Type Field Location
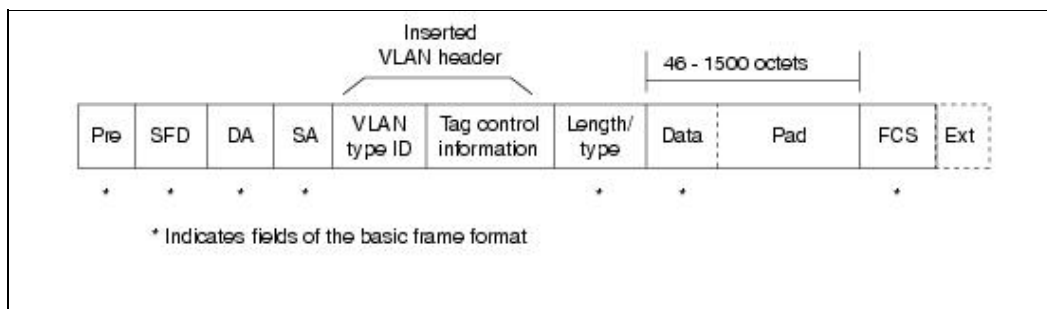
The VLAN header consists of two fields:

- A reserved 2-byte type value, indicating that the frame is a VLAN frame
- A two-byte Tag-Control field that contains both the transmission priority (0 to 7, where 7 is the highest) and a VLAN ID that identifies the particular VLAN over which the frame is to be sent

The receiving MAC reads the reserved type value, which is located in the normal Length/Type field position, and interprets the received frame as a VLAN frame. Then the following occurs:

- If the MAC is installed in a switch port, the frame is forwarded according to its priority level to all ports that are associated with the indicated VLAN identifier.
- If the MAC is installed in an end station, it removes the 4-byte VLAN header and processes the frame in the same manner as a basic data frame.

VLAN tagging requires that all network nodes involved with a VLAN group be equipped with the VLAN option.

The Ethernet Physical Layers

Because Ethernet devices implement only the bottom two layers of the OSI protocol stack, they are typically implemented as network interface cards (NICs) that plug into the host device's motherboard. The different

NICs are identified by a three-part product name that is based on the physical layer attributes.

The naming convention is a concatenation of three terms indicating the transmission rate, the transmission method, and the media type/signal encoding. For example, consider this:

- 10Base-T = 10 Mbps, baseband, over two twisted-pair cables
- 100Base-T2 = 100 Mbps, baseband, over two twisted-pair cables
- 100Base-T4 = 100 Mbps, baseband, over four-twisted pair cables
- 1000Base-LX = 100 Mbps, baseband, long wavelength over optical fiber cable

A question sometimes arises as to why the middle term always seems to be "Base." Early versions of the protocol also allowed for broadband transmission (for example, 10Broad), but broadband implementations were not successful in the marketplace. All current Ethernet implementations use baseband transmission.

Encoding for Signal Transmission

In baseband transmission, the frame information is directly impressed upon the link as a sequence of pulses or data symbols that are typically attenuated (reduced in size) and distorted (changed in shape) before they reach the other end of the link. The receiver's task is to detect each pulse as it arrives and then to extract its correct value before transferring the reconstructed information to the receiving MAC.

Filters and pulse-shaping circuits can help restore the size and shape of the received waveforms, but additional measures must be taken to ensure that the received signals are sampled at the correct time in the pulse period and at same rate as the transmit clock:

- The receive clock must be recovered from the incoming data stream to allow the receiving physical layer to synchronize with the incoming pulses.
- Compensating measures must be taken for a transmission effect known as baseline wander.

Clock recovery requires level transitions in the incoming signal to identify and synchronize on pulse boundaries. The alternating 1s and 0s of the frame preamble were designed both to indicate that a frame was arriving and to aid in clock recovery. However, recovered clocks can drift and possibly lose synchronization if pulse levels remain constant and there are no transitions to detect (for example, during long strings of 0s).

Baseline wander results because Ethernet links are AC-coupled to the transceivers and because AC coupling is incapable of maintaining voltage levels for more than a short time. As a result, transmitted pulses are distorted by a droop effect similar to the exaggerated example shown in Figure: A Concept Example of Baseline Wander. In long strings of either 1s or 0s, the droop can become so severe that the voltage level passes through the decision threshold, resulting in erroneous sampled values for the affected pulses.

Figure: A Concept Example of Baseline Wander
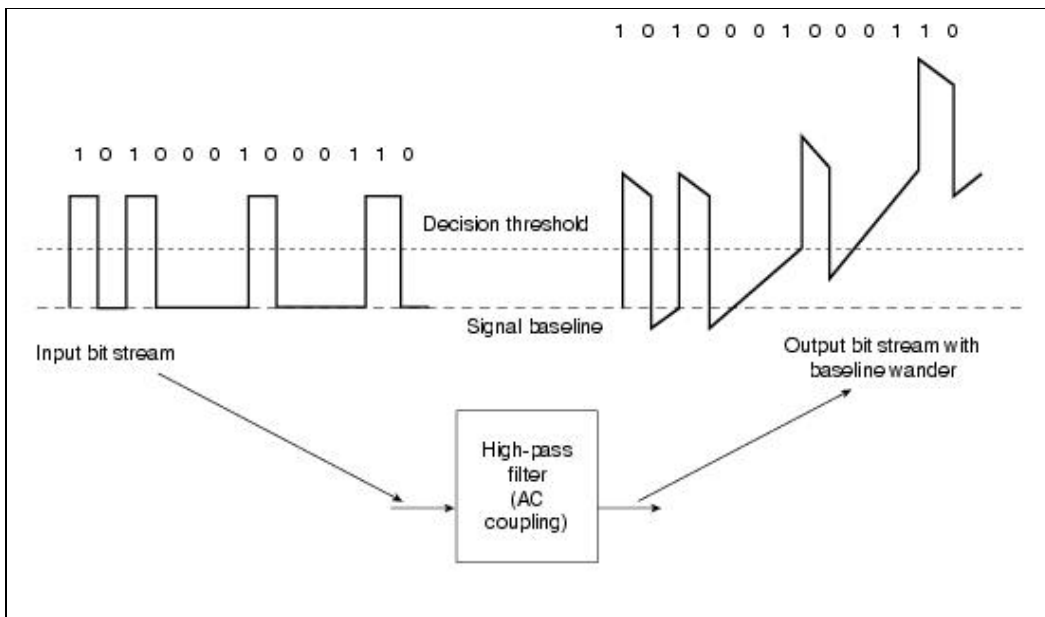
Fortunately, encoding the outgoing signal before transmission can significantly reduce the effect of both these problems, as well as reduce the possibility of transmission errors. Early Ethernet implementations, up to and including 10Base-T, all used the Manchester encoding method, shown in <u>Figure: Transition-Based Manchester Binary Encoding</u>. Each pulse is clearly identified by the direction of the midpulse transition rather than by its sampled level value.

Figure: Transition-Based Manchester Binary Encoding
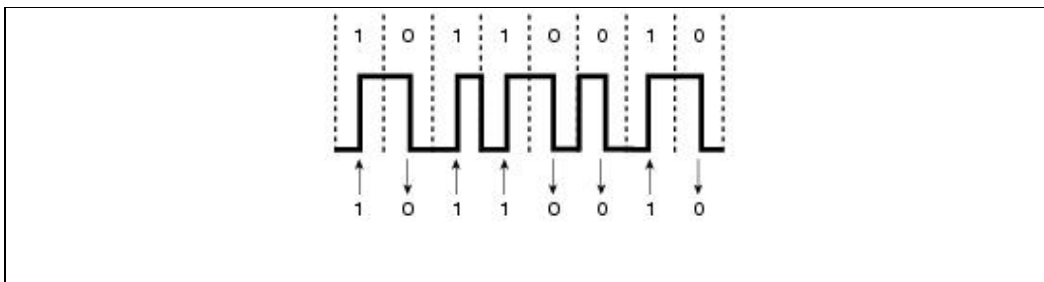
Unfortunately, Manchester encoding introduces some difficult frequency-related problems that make it unsuitable for use at higher data rates. Ethernet versions subsequent to 10Base-T all use different encoding procedures that include some or all of the following techniques:

- **Using data scrambling** - A procedure that scrambles the bits in each byte in an orderly (and recoverable) manner. Some 0s are changed to 1s, some 1s are changed to 0s, and some bits are left the same. The result is reduced run-length of same-value bits, increased transition density, and easier clock recovery.
- **Expanding the code space** - A technique that allows assignment of separate codes for data and control symbols (such as start-of-stream and end-of-stream delimiters, extension bits, and so on) and that assists in transmission error detection.
- **Using forward error-correcting codes** - An encoding in which redundant information is added to the transmitted data stream so that some types of transmission errors can be corrected during frame reception.

Figure: A Concept Example of Baseline Wander

✎**Note**: Forward error-correcting codes are used in 1000Base-T to achieve an effective reduction in the bit error rate. Ethernet protocol limits error handling to detection of bit errors in the received frame. Recovery of frames received with uncorrectable errors or missing frames is the responsibility of higher layers in the protocol stack.

The 802.3 Physical Layer Relationship to the ISO Reference Model

Although the specific logical model of the physical layer may vary from version to version, all Ethernet NICs generally conform to the generic model shown in <u>Figure: The Generic Ethernet Physical Layer Reference Model</u>:

Figure: The Generic Ethernet Physical Layer Reference Model
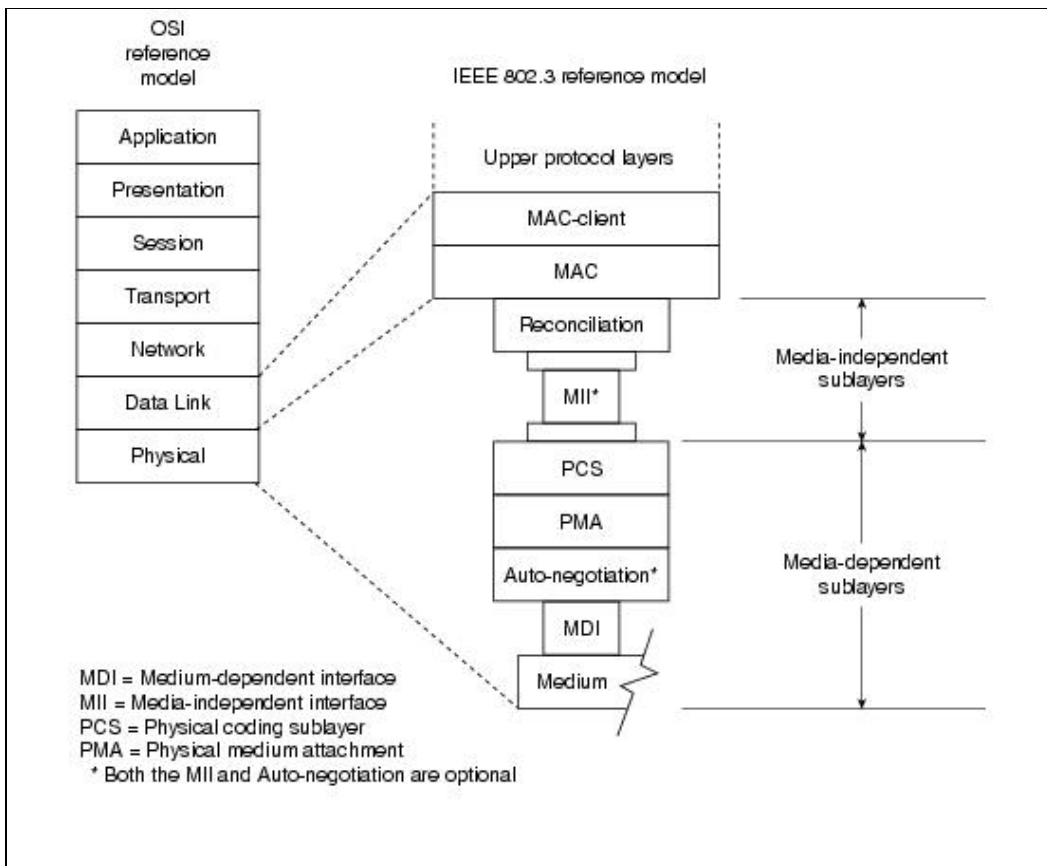
The physical layer for each transmission rate is divided into sublayers that are independent of the particular media type and sublayers that are specific to the media type or signal encoding.

- The reconciliation sublayer and the optional media-independent interface (MII in 10-Mbps and 100-Mbps Ethernet, GMII in Gigabit Ethernet) provide the logical connection between the MAC and the different sets of media-dependent layers. The MII and GMII are defined with separate transmit and receive data paths that are bit-serial for 10-Mbps implementations, nibble-serial (4 bits wide) for 100-Mbps implementations, and byte-serial (8 bits wide) for 1000-Mbps implementations. The media-independent interfaces and the reconciliation sublayer are common for their respective transmission rates and are configured for full-duplex operation in 10Base-T and all subsequent Ethernet versions.
- The media-dependent physical coding sublayer (PCS) provides the logic for encoding, multiplexing,

Figure: Transition-Based Manchester Binary Encoding   15

and synchronization of the outgoing symbol streams as well symbol code alignment, demultiplexing, and decoding of the incoming data.
- The physical medium attachment (PMA) sublayer contains the signal transmitters and receivers (transceivers), as well as the clock recovery logic for the received data streams.
- The medium-dependent interface (MDI) is the cable connector between the signal transceivers and the link.
- The Auto-negotiation sublayer allows the NICs at each end of the link to exchange information about their individual capabilities, and then to negotiate and select the most favorable operational mode that they both are capable of supporting. Auto-negotiation is optional in early Ethernet implementations and is mandatory in later versions.

Depending on which type of signal encoding is used and how the links are configured, the PCS and PMA may or may not be capable of supporting full-duplex operation.

10-Mbps Ethernet-10Base-T

10Base-T provides Manchester-encoded 10-Mbps bit-serial communication over two unshielded twisted-pair cables. Although the standard was designed to support transmission over common telephone cable, the more typical link configuration is to use two pair of a four-pair Category 3 or 5 cable, terminated at each NIC with an 8-pin RJ-45 connector (the MDI), as shown in <u>Figure: The Typical 10Base-T Link Is a Four-Pair UTP Cable in Which Two Pairs Are Not Used</u> pair is configured as a simplex link where transmission is in one direction only, the 10Base-T physical layers can support either half-duplex or full-duplex operation.

Figure: The Typical 10Base-T Link Is a Four-Pair UTP Cable in Which Two Pairs Are Not Used
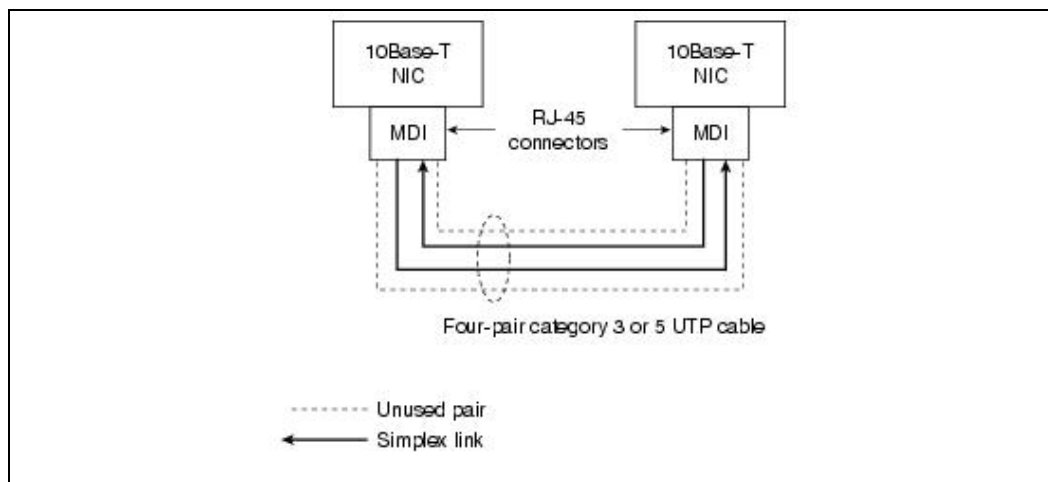
Although 10Base-T may be considered essentially obsolete in some circles, it is included here because there are still many 10Base-T Ethernet networks, and because full-duplex operation has given 10BaseT an extended life.

10Base-T was also the first Ethernet version to include a link integrity test to determine the health of the link. Immediately after powerup, the PMA transmits a normal link pulse (NLP) to tell the NIC at the other end of the link that this NIC wants to establish an active link connection:

- If the NIC at the other end of the link is also powered up, it responds with its own NLP.
- If the NIC at the other end of the link is not powered up, this NIC continues sending an NLP about once every 16 ms until it receives a response.

The link is activated only after both NICs are capable of exchanging valid NLPs.

100 Mbps-Fast Ethernet

Increasing the Ethernet transmission rate by a factor of ten over 10Base-T was not a simple task, and the effort resulted in the development of three separate physical layer standards for 100 Mbps over UTP cable: 100Base-TX and 100Base-T4 in 1995, and 100Base-T2 in 1997. Each was defined with different encoding requirements and a different set of media-dependent sublayers, even though there is some overlap in the link cabling.

Table: Summary of 100Base-T Physical Layer Characteristics compares the physical layer characteristics of 10Base-T to the various 100Base versions.

Table: Summary of 100Base-T Physical Layer Characteristics

Ethernet Version	Transmit Symbol Rate	Encoding	Cabling	Full-Duplex Operation
10Base-T	10 MBd	Manchester	Two pairs of UTP Category -3 or better	Supported
100Base-TX	125 MBd	4B/5B	Two pairs of UTP Category -5 or Type 1 STP	Supported
100Base-T4	33 MBd	8B/6T	Four pairs of UTP Category -3 or better	Not supported
100Base-T2	25 MBd	PAM5x5	Two pairs of UTP Category -3 or better	Supported

Note: One baud is equal to one transmitted symbol per second, where the transmitted symbol may contain the equivalent value of 1 or more binary bits.

Although not all three 100-Mbps versions were successful in the marketplace, all three have been discussed in the literature, and all three did impact future designs. As such, all three are important to consider here.

100Base-X

100Base-X was designed to support transmission over either two pairs of Category 5 UTP copper wire or two strands of optical fiber. Although the encoding, decoding, and clock recovery procedures are the same for both media, the signal transmission is different-electrical pulses in copper and light pulses in optical fiber. The signal transceivers that were included as part of the PMA function in the generic logical model of the following figure were redefined as the separate physical media-dependent (PMD) sublayers shown in Figure: The 100Base-X Logical Model.

Figure: The 100Base-X Logical Model
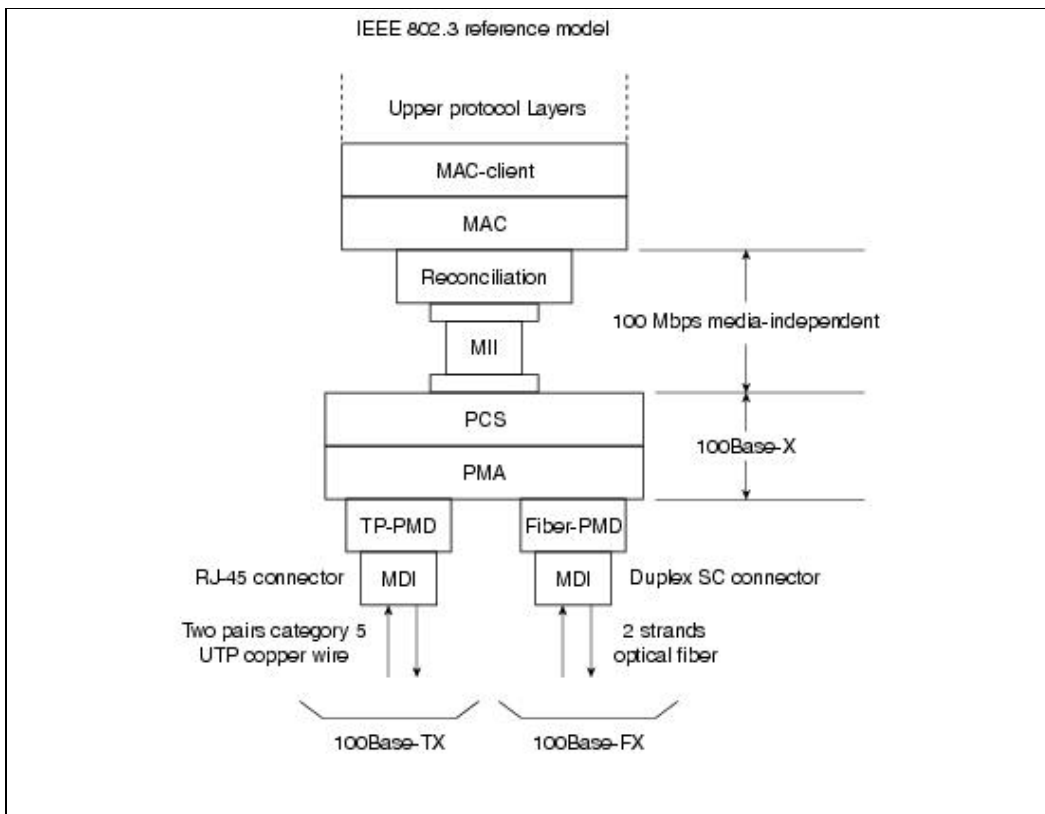
The 100Base-X encoding procedure is based on the earlier FDDI optical fiber physical media-dependent and FDDI/CDDI copper twisted-pair physical media-dependent signaling standards developed by ISO and ANSI. The 100Base-TX physical media-dependent sublayer (TP-PMD) was implemented with CDDI semiconductor transceivers and RJ-45 connectors; the fiber PMD was implemented with FDDI optical transceivers and the Low Cost Fibre Interface Connector (commonly called the duplex SC connector).

The 4B/5B encoding procedure is the same as the encoding procedure used by FDDI, with only minor adaptations to accommodate Ethernet frame control. Each 4-bit data nibble (representing half of a data byte) is mapped into a 5-bit binary code-group that is transmitted bit-serial over the link. The expanded code space provided by the 32 5-bit code-groups allow separate assignment for the following:

- The 16 possible values in a 4-bit data nibble (16 code-groups).
- Four control code-groups that are transmitted as code-group pairs to indicate the start-of-stream delimiter (SSD) and the end-of-stream delimiter (ESD). Each MAC frame is "encapsulated" to mark both the beginning and end of the frame. The first byte of preamble is replaced with SSD code-group pair that precisely identifies the frame's code-group boundaries. The ESD code-group pair is appended after the frame's FCS field.
- A special IDLE code-group that is continuously sent during interframe gaps to maintain continuous synchronization between the NICs at each end of the link. The receipt of IDLE is interpreted to mean that the link is quiet.
- Eleven invalid code-groups that are not intentionally transmitted by a NIC (although one is used by a repeater to propagate receive errors). Receipt of any invalid code-group will cause the incoming frame to be treated as an invalid frame.

Figure: The 100Base-X Code-Group Stream with Frame Encapsulation shows how a MAC frame is encapsulated before being transmitted as a 100Base-X code-group stream.

Figure: The 100Base-X Logical Model   18

Figure: The 100Base-X Code-Group Stream with Frame Encapsulation
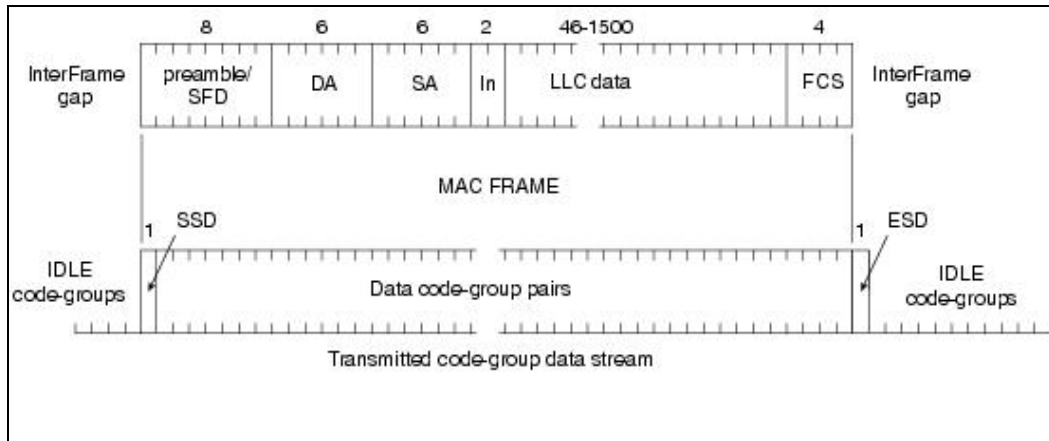
100Base-TX transmits and receives on the same link pairs and uses the same pin assignments on the MDI as 10Base-T. 100Base-TX and 100Base-FX both support half-duplex and full-duplex transmission.

100Base-T4

100Base-T4 was developed to allow 10BaseT networks to be upgraded to 100-Mbps operation without requiring existing four-pair Category 3 UTP cables to be replaced with the newer Category 5 cables. Two of the four pairs are configured for half-duplex operation and can support transmission in either direction, but only in one direction at a time. The other two pairs are configured as simplex pairs dedicated to transmission in one direction only. Frame transmission uses both half-duplex pairs, plus the simplex pair that is appropriate for the transmission direction, as shown in Figure: The 100Base-T4 Wire-Pair Usage During Frame Transmission. The simplex pair for the opposite direction provides carrier sense and collision detection. Full-duplex operation cannot be supported on 100Base-T4.

Figure: The 100Base-T4 Wire-Pair Usage During Frame Transmission
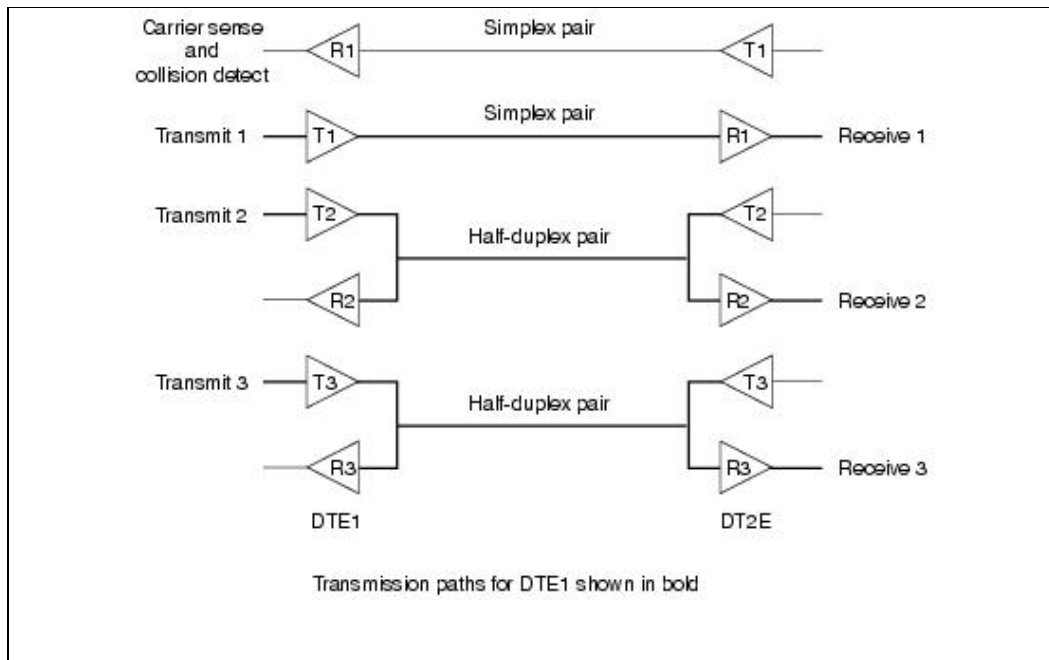
100Base-T4 uses an 8B6T encoding scheme in which each 8-bit binary byte is mapped into a pattern of six

ternary (three-level: +1, 0, -1) symbols known as 6T code-groups. Separate 6T code-groups are used for IDLE and for the control code-groups that are necessary for frame transmission. IDLE received on the dedicated receive pair indicates that the link is quiet.

During frame transmission, 6T data code-groups are transmitted in a delayed round-robin sequence over the three transmit wire-pairs, as shown in Figure: The 100Base-T4 Frame Transmission Sequence. Each frame is encapsulated with start-of-stream and end-of-packet 6T code-groups that mark both the beginning and end of the frame, and the beginning and end of the 6T code-group stream on each wire pair. Receipt of a n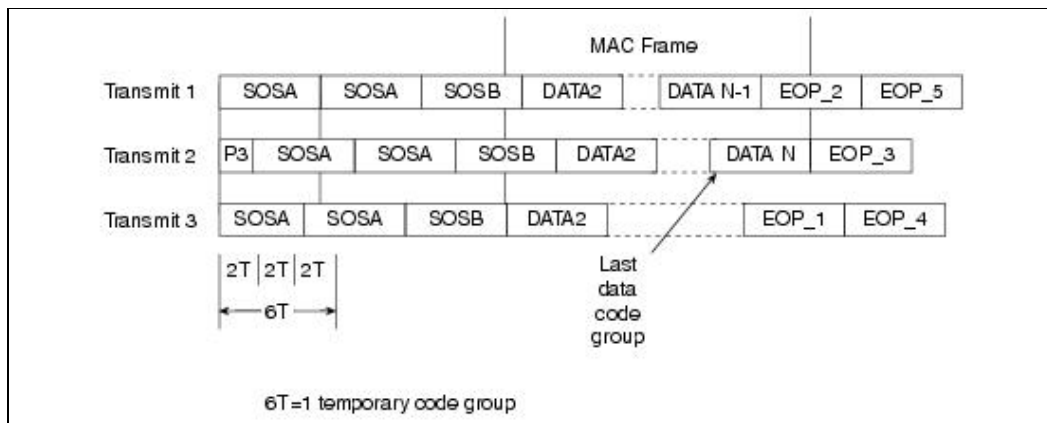on-IDLE code-group over the dedicated receive-pair any time before the collision window expires indicates that a collision has occurred.

Figure: The 100Base-T4 Frame Transmission Sequence

100Base-T2

The 100Base-T2 specification was developed as a better alternative for upgrading networks with installed Category 3 cabling than was being provided by 100Base-T4. Two important new goals were defined:

- To provide communication over two pairs of Category 3 or better cable
- To support both half-duplex and full-duplex operation

100Base-T2 uses a different signal transmission procedure than any previous twisted-pair Ethernet implementations. Instead of using two simplex links to form one full-duplex link, the 100Base-T2 dual-duplex baseband transmission method sends encoded symbols simultaneously in both directions on both wire pairs, as shown in Figure: The 100Base-T2 Link Topology. The term "TDX<3:2>" indicates the 2 most significant bits in the nibble before encoding and transmission. "RDX<3:2>" indicates the same 2 bits after receipt and decoding.

Figure: The 100Base-T2 Link Topology
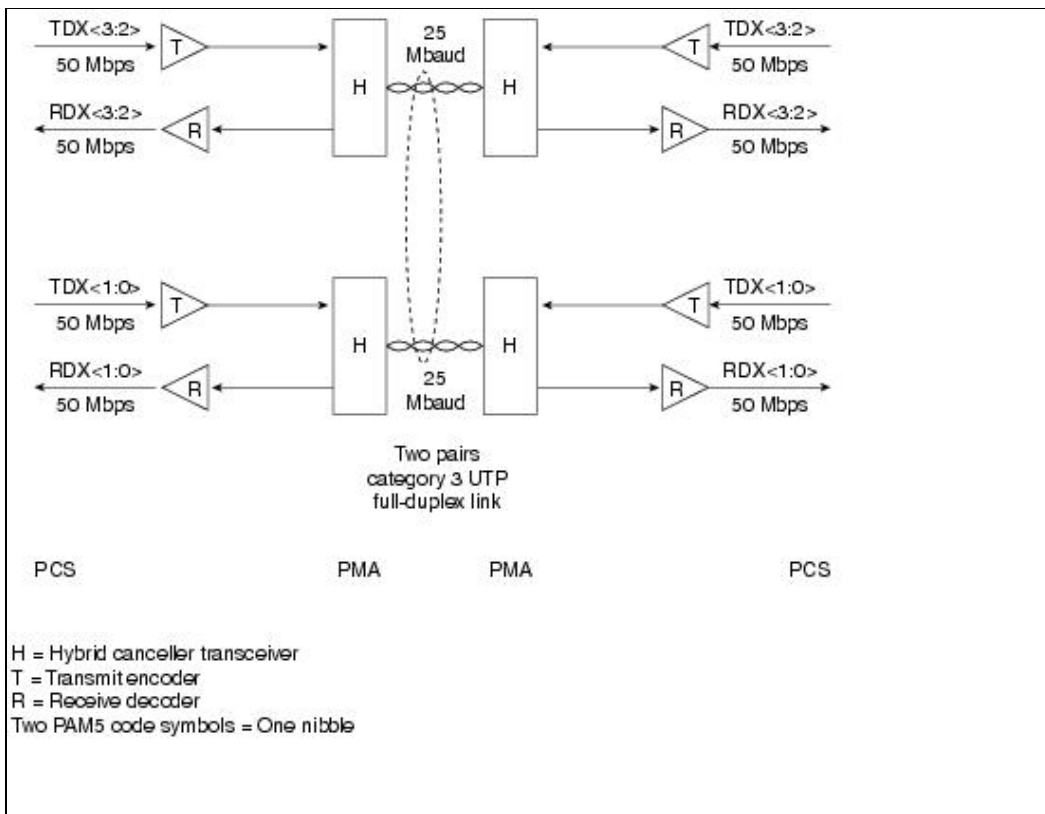
Dual-duplex baseband transmission requires the NICs at each end of the link to be operated in a master/slave loop-timing mode. Which NIC will be master and which will be slave is determined by autonegotiation during link initiation. When the link is operational, synchronization is based on the master NIC's internal transmit clock. The slave NIC uses the recovered clock for both transmit and receive operations, as shown in Figure: The 100Base-T2 Loop Timing Configuration. Each transmitted frame is encapsulated, and link synchronization is maintained with a continuous stream of IDLE symbols during interframe gaps.

Figure: The 100Base-T2 Loop Timing Configuration
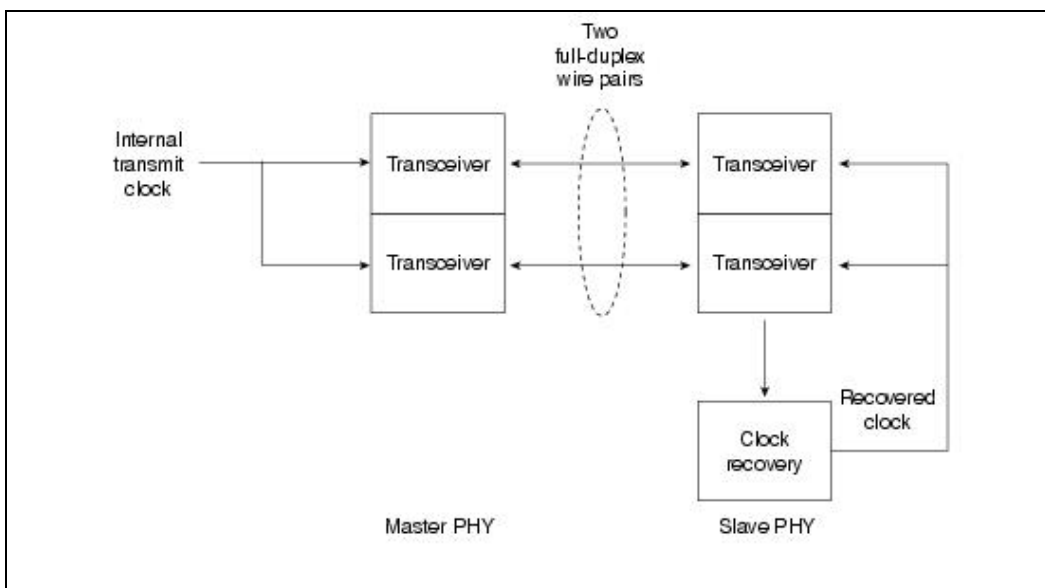
Figure: The 100Base-T2 Link Topology   21

The 100Base-T2 encoding process first scrambles the data frame nibbles to randomize the bit sequence. It then maps the two upper bits and the two lower bits of each nibble into two five-level (+2, +1, 0, -1, -2) pulse amplitude-modulated (PAM5) symbols that are simultaneously transmitted over the two wire pairs (PAM5x5). Different scrambling procedures for master and slave transmissions ensure that the data streams traveling in opposite directions on the same wire pair are uncoordinated.

Signal reception is essentially the reverse of signal transmission. Because the signal on each wire pair at the MDI is the sum of the transmitted signal and the received signal, each receiver subtracts the transmitted symbols from the signal received at the MDI to recover the symbols in the incoming data stream. The incoming symbol pair is then decoded, unscrambled, and reconstituted as a data nibble for transfer to the MAC.

1000 Mbps-Gigabit Ethernet

The Gigabit Ethernet standards development resulted in two primary specifications: 1000Base-T for UTP copper cable and 1000Base-X STP copper cable, as well as single and multimode optical fiber (see Figure: Gigabit Ethernet Variations).

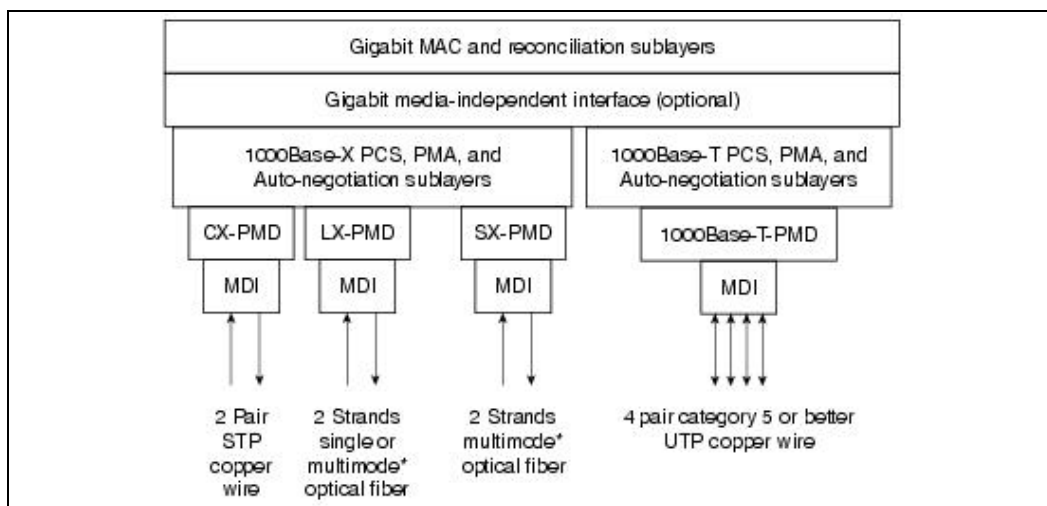**Figure 7-22 Gigabit Ethernet Variations**

1000Base-T

1000Base-T Ethernet provides full-duplex transmission over four-pair Category 5 or better UTP cable. 1000Base-T is based largely on the findings and design approaches that led to the development of the Fast Ethernet physical layer implementations:

- 100Base-TX proved that binary symbol streams could be successfully transmitted over Category 5 UTP cable at 125 MBd.
- 100Base-T4 provided a basic understanding of the problems related to sending multilevel signals over four wire pairs.
- 100Base-T2 proved that PAM5 encoding, coupled with digital signal processing, could handle both simultaneous two-way data streams and potential crosstalk problems resulting from alien signals on adjacent wire pairs.

1000Base-T scrambles each byte in the MAC frame to randomize the bit sequence before it is encoded using a 4-D, 8-State Trellis Forward Error Correction (FEC) coding in which four PAM5 symbols are sent at the

same time over four wire pairs. Four of the five levels in each PAM5 symbol represent 2 bits in the data byte. The fifth level is used for FEC coding, which enhances symbol recovery in the presence of noise and crosstalk. Separate scramblers for the master and slave PHYs create essentially uncorrelated data streams between the two opposite-travelling symbol streams on each wire pair.

The1000Base-T link topology is shown in Figure: The 1000Base-T Link Topology. The term "TDX<7:6>" indicates the 2 most significant bits in the data byte before encoding and transmission. "RDX<7:6>" indicates the same 2 bits after receipt and decoding.

Figure: The 1000Base-T Link Topology
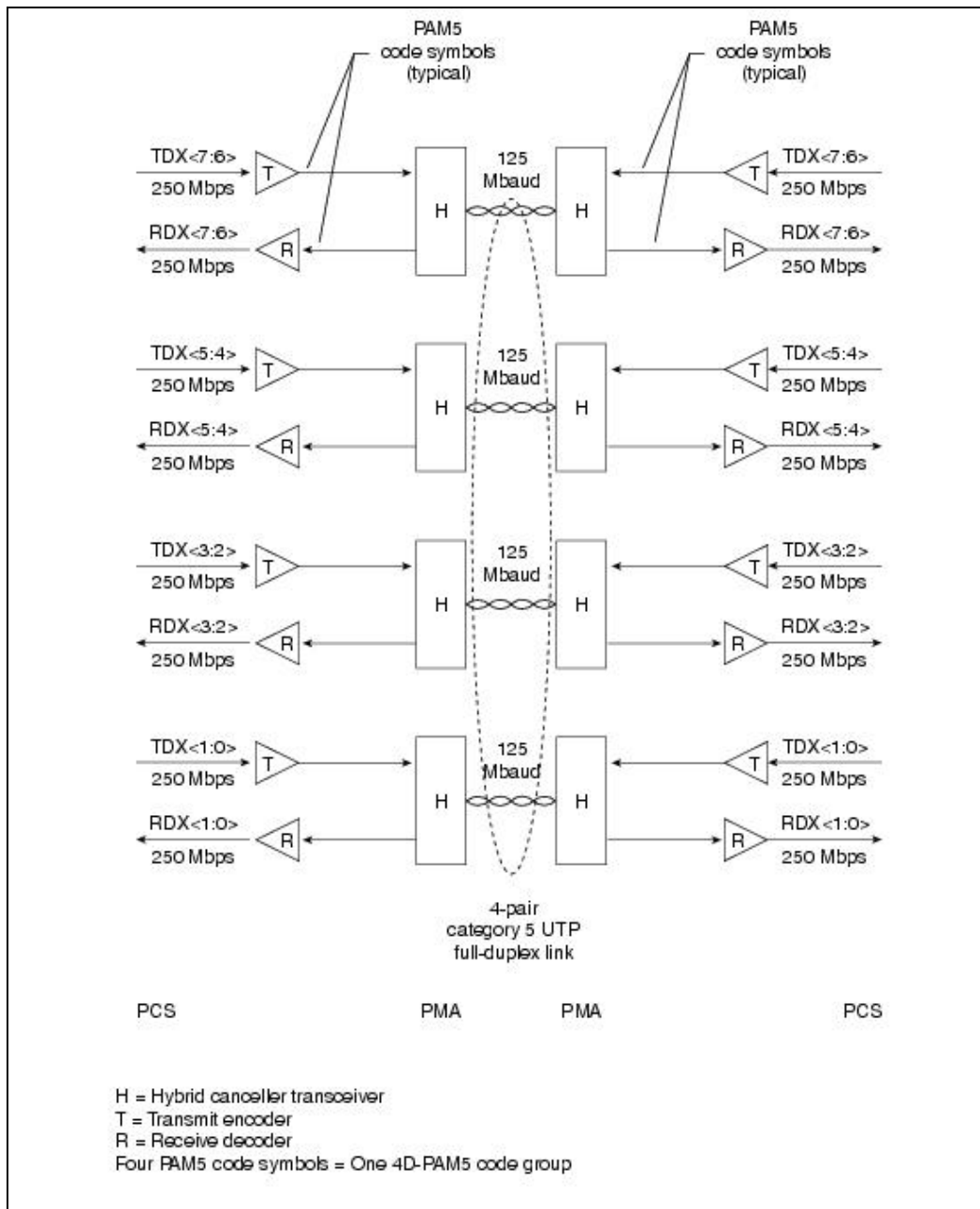
The clock recovery and master/slave loop timing procedures are essentially the same as those used in 100Base-T2 (see Figure: 1000Base-T Master/Slave Loop Timing Configuration). Which NIC will be master (typically the NIC in a multiport intermediate network node) and which will be slave is determined during

autonegotiation.

Figure: 1000Base-T Master/Slave Loop Timing Configuration
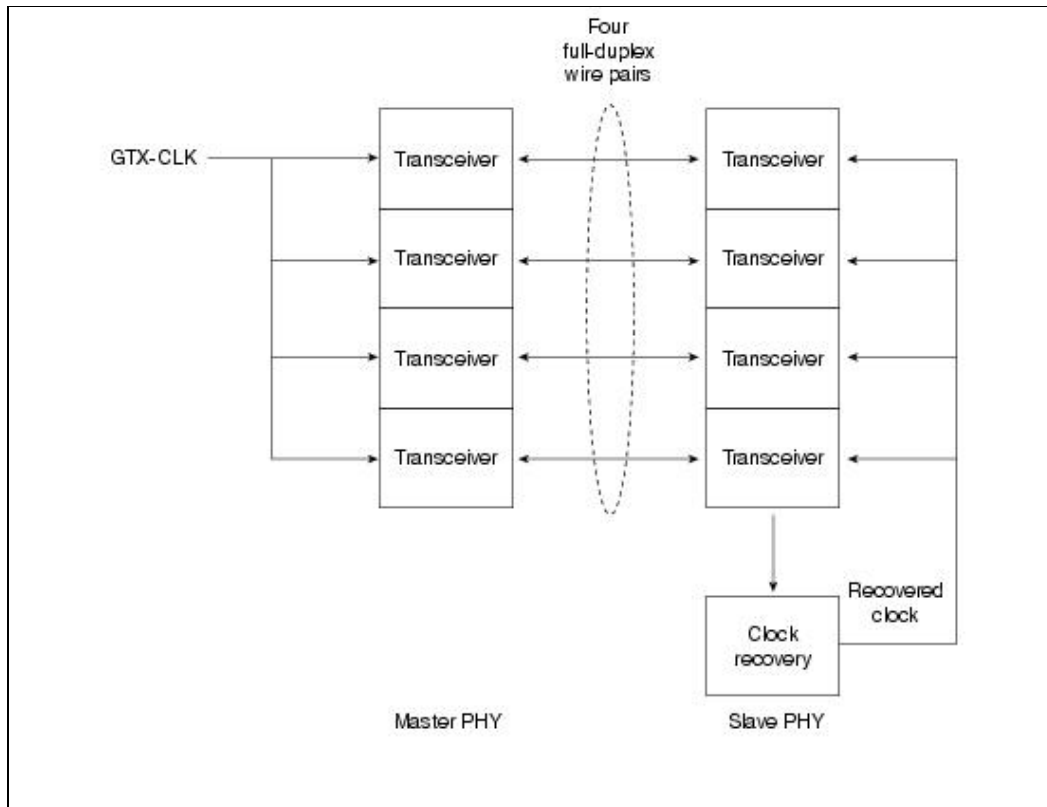
Each transmitted frame is encapsulated with start-of-stream and end-of-stream delimiters, and loop timing is maintained by continuous streams of IDLE symbols sent on each wire pair during interframe gaps. 1000Base-T supports both half-duplex and full-duplex operation.

1000Base-X

All three 1000Base-X versions support full-duplex binary transmission at 1250 Mbps over two strands of optical fiber or two STP copper wire-pairs, as shown in Figure: 1000Base-X Link Configuration. Transmission coding is based on the ANSI Fibre Channel 8B/10B encoding scheme. Each 8-bit data byte is mapped into a 10-bit code-group for bit-serial transmission. Like earlier Ethernet versions, each data frame is encapsulated at the physical layer before transmission, and link synchronization is maintained by sending a continuous stream of IDLE code-groups during interframe gaps. All 1000Base-X physical layers support both half-duplex and full-duplex operation.

Figure: 1000Base-X Link Configuration

Simplex link
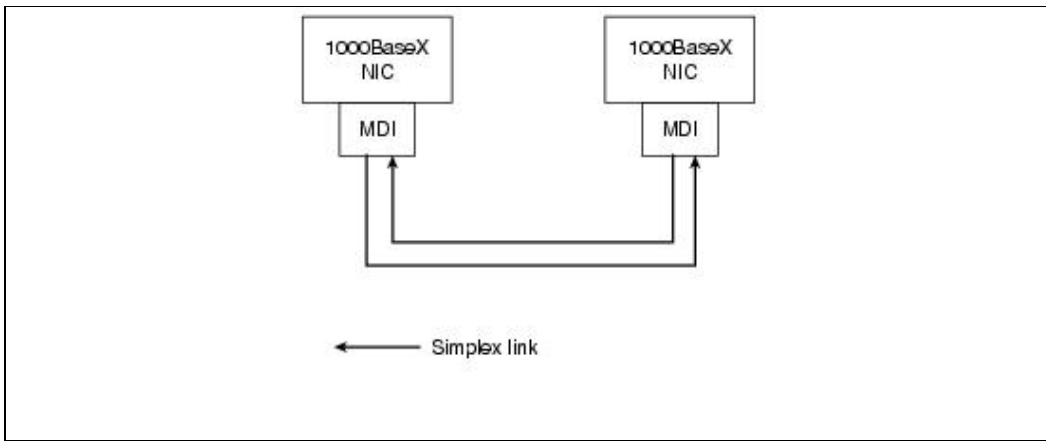
The principal differences among the 1000Base-X versions are the link media and connectors that the particular versions will support and, in the case of optical media, the wavelength of the optical signal (see Table: 1000Base-X Link Configuration Support).

Table: 1000Base-X Link Configuration Support

Link Configuration	1000Base-CX	1000Base-SX (850 nm Wavelength)	1000Base-LX (1300 nm Wavelength)
150 STP copper	Supported	Not supported	Not supported
125/62.5 µm multimode optical fiber	Not supported	Supported	Supported
125/50 µm multimode optical fiber	Not supported	Supported	Supported
125/10 µm single mode optical fiber	Not supported	Not supported	Supported
Allowed connectors	IEC style 1 or Fibre Channel style 2	SFF MT-RJ or Duplex SC	SFF MT-RJ or Duplex SC

The 125/62.5 µm specification refers to the cladding and core diameters of the optical fiber.

Network Cabling-Link Crossover Requirements

Link compatibility requires that the transmitters at each end of the link be connected to the receivers at the other end of the link. However, because cable connectors at both ends of the link are keyed the same, the conductors must cross over at some point to ensure that transmitter outputs are always connected to receiver inputs.

Unfortunately, when this requirement first came up in the development of 10Base-T, IEEE 802.3 chose not to make a hard rule as to whether the crossover should be implemented in the cable as shown in Figure: Alternative Ways for Implementing the Link Crossover Requirement (a) or whether it should be implemented internally as shown in Figure: Alternative Ways for Implementing the Link Crossover Requirement (b).

Figure: Alternative Ways for Implementing the Link Crossover Requirement

Figure: 1000Base-X Link Configuration   25
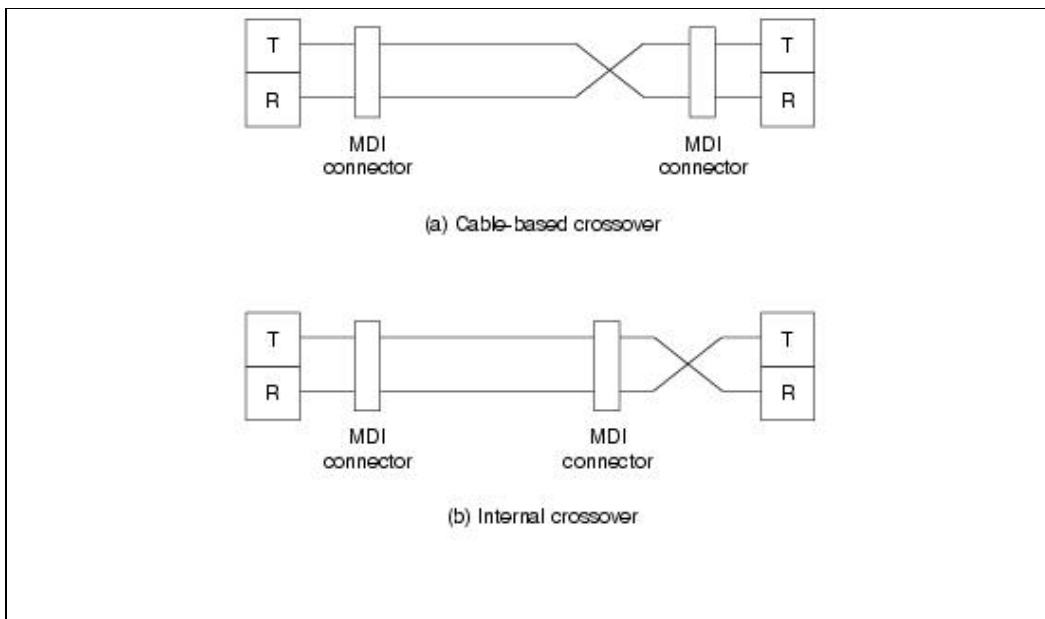
Instead, IEEE 802.3 defined two rules and made two recommendations:

- There must be an odd number of crossovers in all multiconductor links.
- If a PMD is equipped with an internal crossover, its MDI must be clearly labeled with the graphical X symbol.
- Implementation of an internal crossover function is optional.
- When a DTE is connected to a repeater or switch (DCE) port, it is recommended that the crossover be implemented within the DCE port.

The eventual result was that ports in most DCEs were equipped with PMDs that contained internal crossover circuitry and that DTEs had PMDs without internal crossovers. This led to the following oft-quoted de facto installation rule:

- Use a straight-through cable when connecting DTE to DCE. Use a crossover cable when connecting DTE to DTE or DCE to DCE.

Unfortunately, the de facto rule does not apply to all Ethernet versions that have been developed subsequent to 10Base-T. As things now stand, the following is true:

- All fiber-based systems use cables that have the crossover implemented within the cable.
- All 100Base systems using twisted-pair links use the same rules and recommendations as 10Base-T.
- 1000Base-T NICs may implement a selectable internal crossover option that can be negotiated and enabled during autonegotiation. When the selectable crossover option is not implemented, 10Base-T rules and recommendations apply.

System Considerations

Given all the choices discussed previously, it might seem that it would be no problem to upgrade an existing network or to plan a new network. The problem is twofold. Not all the choices are reasonable for all networks, and not all Ethernet versions and options are available in the market, even though they may have been specified in the standard.

Figure: Alternative Ways for Implementing the Link Crossover Requirement   26

Choosing UTP-Based Components and Media Category

By now, it should be obvious that UTP-based NICs are available for 10-Mbps, 100-Mbps, and 1000-Mbps implementations. The choice is relatively simple for both 10-Mbps and 1000-Mbps operation: 10Base-T and 1000Base-T. From the previous discussions, however, it would not seem to be that simple for 100-Mbps implementations.

Although three UTP-based NICs are defined for 100 Mbps, the market has effectively narrowed the choice to just 100Base-TX, which became widely available during the first half of 1995:

- By the time 100Base-T4 products first appeared on the market, 100Base-TX was well entrenched, and development of the full-duplex option, which 100Base-T4 could not support, was well underway.
- The 100Base-T2 standard was not approved until spring 1997, too late to interest the marketplace. As a result, 100Base-T2 products were not even manufactured.

Several choices have also been specified for UTP media: Category 3, 4, 5, or 5E. The differences are cable cost and transmission rate capability, both of which increase with the category numbers. However, current transmission rate requirements and cable cost should not be the deciding factors in choosing which cable category to install. To allow for future transmission rate needs, cables lower than Category 5 should not even be considered, and if gigabit rates are a possible future need, Category 5E should be seriously considered:

- Installation labor costs are essentially constant for all types of UTP four-pair cable.
- Labor costs for upgrading installed cable (removing the existing and installing new) are typically greater than the cost of the original installation.
- UTP cable is backward-compatible. Higher-category cable will support lower-category NICs, but not vice versa.
- The physical life of UTP cable (decades) is much longer than the useable life of the connected equipment.

Auto-negotiation-An Optional Method for Automatically Configuring Link Operational Modes

The purpose of autonegotiation is to find a way for two NICs that share a UTP link to communicate with each other, regardless of whether they both implemented the same Ethernet version or option set.

Autonegotiation is performed totally within the physical layers during link initiation, without any additional overhead either to the MAC or to higher protocol layers. Autonegotiation allows UTP-based NICs to do the following:

- Advertise their Ethernet version and any optional capabilities to the NIC at the other end of the link
- Acknowledge receipt and understanding of the operational modes that both NICs share
- Reject any operational modes that are not shared
- Configure each NIC for highest-level operational mode that both NICs can support

Autonegotiation is specified as an option for 10Base-T, 100Base-TX, and 100Base-T4, but it is required for 100Base-T2 and 1000Base-T implementations.

Table: The Defined Autonegotiation Selection Levels for UTP NICs lists the defined selection priority levels (highest level = top priority) for UTP-based Ethernet NICs.

Table: The Defined Autonegotiation Selection Levels for UTP NICs

Selection Level	Operational Mode	Maximum Total Data Transfer Rate (Mbps)
9	1000Base-T full-duplex	2000
8	1000Base-T half-duplex	1000
7	100Base-T2 full-duplex	200
6	100Base-TX full-duplex	200
5	100Base-T2 half-duplex	100
4	100Base-T4 half-duplex	100
3	100Base-TX half-duplex	100
2	10Base-T full-duplex	20
1	10Base-T half-duplex	10

Note: Because full-duplex operation allows simultaneous two-way transmission, the maximum total transfer rate for full-duplex operation is double the half-duplex transmission rate.

The autonegotiation function in UTP-based NICs uses a modified 10Base-T link integrity pulse sequence in which the NLPs are replaced by bursts of fast link pulses (FLPs), as shown in Figure: Autonegotiation FLP Bursts Replace NLPs During Link Initiation. Each FLP burst is an alternating clock/data sequence in which the data bits in the burst identify the operational modes supported by the transmitting NIC and also provide information used by the autonegotiation handshake mechanism. If the NIC at the other end of the link is a compatible NIC but does not have autonegotiation capability, a parallel detection function still allows it to be recognized. A NIC that fails to respond to FLP bursts and returns only NLPs is treated as a 10Base-T half-duplex NIC.

Figure: Autonegotiation FLP Bursts Replace NLPs During Link Initiation
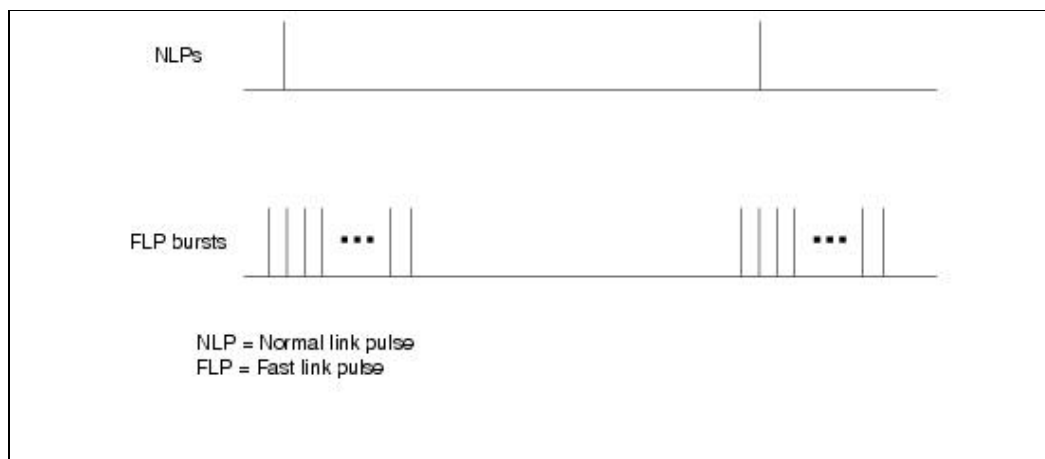
At first glance, it may appear that the autonegotiation process would always select the mode supported by the NIC with the lessor capability, which would be the case if both NICs use the same encoding procedures and link configuration. For example, if both NICs are 100Base-TX but only one supports full-duplex operation, the negotiated operational mode will be half-duplex 100Base-TX. Unfortunately, the different 100Base versions are not compatible with each other at 100 Mbps, and a 100Base-TX full-duplex NIC would autonegotiate with a 100Base-T4 NIC to operate in 10Base-T half-duplex mode.

Autonegotiation in 1000Base-X NICs is similar to autonegotiation in UTP-based systems, except that it currently applies only to compatible 1000Base-X devices and is currently constrained to negotiate only half-duplex or full-duplex operation and flow control direction.

Network Switches Provide a Second, and Often Better, Alternative to Higher Link Speeds in CSMA/CD Network Upgrades

Competitively priced network switches became available on the market shortly after the mid-1990s and essentially made network repeaters obsolete for large networks. Although repeaters can accept only one frame at a time and then send it to all active ports (except the port on which it is being received), switches are equipped with the following:

- MAC-based ports with I/O frame buffers that effectively isolate the port from traffic being sent at the same time to or from other ports on the switch
- Multiple internal data paths that allow several frames to be transferred between different ports at the same time

These may seem like small differences, but they produce a major effect in network operation. Because each port provides access to a high-speed network bridge (the switch), the collision domain in the network is reduced to a series of small domains in which the number of participants is reduced to two-the switch port and the connected NIC (see Figure: Replacing the Network Repeaters with Switches Reduces the Collision Domains to Two NICs Each). Furthermore, because each participant is now in a private collision domain, his or her available bandwidth has not only been markedly increased, it was also done without having to change the link speed.

Consider, for example, a 48-station workgroup with a couple of large file servers and several network printers on a 100-Mbps CSMA/CD network. The average available bandwidth, not counting interframe gaps and collision recovery, would be $100/50 = 2$ Mbps (network print servers do not generate network traffic). On the other hand, if the same workgroup were still on a 10Base-T network in which the repeaters had been replaced with network switches, the bandwidth available to each user would be 10 Mbps.

Clearly, network configuration is as important as raw link speed.

Note: To ensure that each end station will be capable of communicating at full rate, the network switches should be nonsaturating (be capable of accepting and transferring data at the full rate from each port simultaneously).

Multispeed NICs

Auto-negotiation opened the door to the development of low-cost, multispeed NICs that, for example, support both half- and full-duplex operation under either 100Base-TX or 10Base-T signaling procedures. Multispeed NICs allow staged network upgrades in which the 10Base-T half-duplex end stations can be connected to 100Base-TX full-duplex switch ports without requiring the NIC in the PC to be changed. Then, as more bandwidth is needed for individual PCs, the NICs in those PCs can be upgraded to 100Base-TX full-duplex mode.

Figure: Autonegotiation FLP Bursts Replace NLPs During Link Initiation   29

Figure: Replacing the Network Repeaters with Switches Reduces the Collision Domains to Two NICs Each
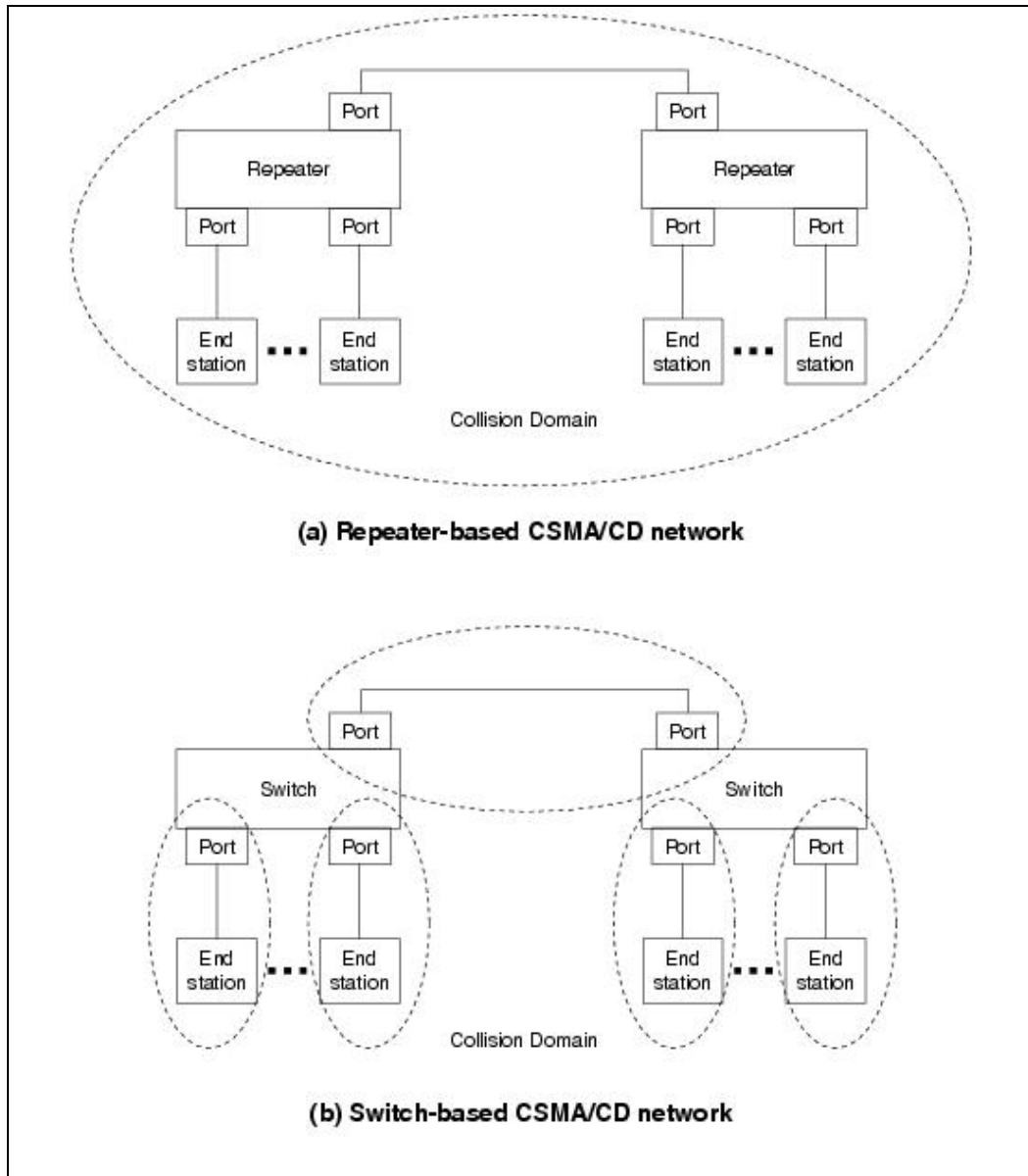
Choosing 1000Base-X Components and Media

Although the above table shows that there is considerable flexibility of choice in the 1000Base-X link media, there is not total flexibility. Some choices are preferred over others:

- NICs at both ends of the link must be the same 1000Base-X version (CX, LX, or SX), and the link connectors must match the NIC connectors.
- The 1000Base-CX specification allows either style 1 or style 2 connectors, but style 2 is preferred because some style 1 connectors are not suitable for operation at 1250 Mbps. 1000Base-CX links are intended for patch-cord use within a communications closet and are limited to 25 meters.
- The 1000Base-LX and 1000Base-SX specifications allow either the small form factor SFF MT-RJ or the larger duplex SC connectors. Because SFF MT-RJ connectors are only about half as large as duplex SC connectors, and because space is a premium, it follows that SFF MT-RJ connectors may become the predominant connector.

- 1000Base-LX transceivers generally cost more than 1000Base-SX transceivers.
- The maximum operating range for optical fibers depends on both the transmission wavelength and the modal bandwidth (MHz.km) rating of the fiber. See <u>Table: Maximum Operating Ranges for Common Optical Fibers</u>.

Table: Maximum Operating Ranges for Common Optical Fibers

Fiber Core Diameter/Modal Bandwidth	1000Base-SX (850 nm Wavelength)	1000Base-LX (1300 nm Wavelength)
62.5 μ m multimode fiber (200/500) MHz.km	275 meters	550 meters
50 μ m multimode fiber (400/400) MHz.km	500 meters	550 meters
50 μ m multimode fiber (500/500) MHz.km	550 meters	550 meters
10 μ m single-mode fiber	Not supported	5000 meters

Note: 1000Base-LX transceivers may also require use of an offset-launch, mode-conditioning patch cord when coupling to some existing multimode fibers.

The operating ranges shown in the above table are those specified in the IEEE 802.3 standard. In practice, however, the maximum operating range for LX transceivers over 62.5 μ m multimode fiber is approximately 700 meters, and some LX transceivers have been qualified to support a 10,000-meter operating range over single-mode fiber.

Multiple-Rate Ethernet Networks

Given the opportunities shown by the example in the previous sections, it is not surprising that most large Ethernet networks are now implemented with a mix of transmission rates and link media, as shown in the cable model in <u>Figure: An Example Multirate Network Topology?the ISO/IEC 11801 Cable Model</u>.

Figure: An Example Multirate Network Topology?the ISO/IEC 11801 Cable Model
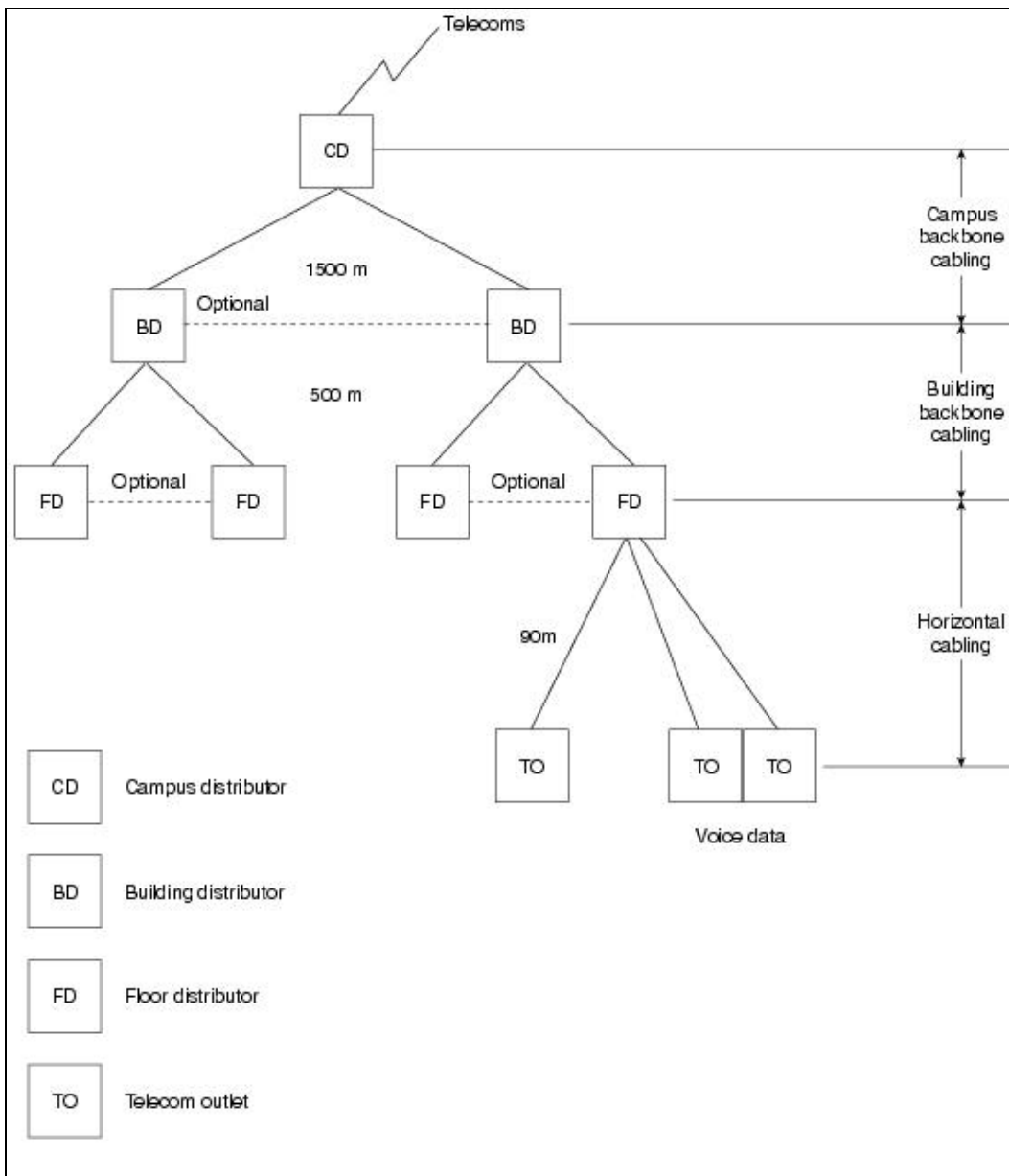
The ISO/IEC 11801 cable model is the network model on which the IEEE 802.3 standards are based:

- **Campus distributor** - The term campus refers to a facility with two or more buildings in a relatively small area. This is the central point of the campus backbone and the telecom connection point with the outside world. In Ethernet LANs, the campus distributor would typically be a gigabit switch with telecom interface capability.
- **Building distributor** - This is the building's connection point to the campus backbone. An Ethernet building distributor would typically be a 1000/100- or 1000/100/10-Mbps switch.
- **Floor distributor** - This is the floor's connection point to the building distributor. ISO/IEC 11801 recommends at least one floor distributor for every 1000 m2 of floor space in office environments, and, if possible, a separate distributor for each floor in the building. An Ethernet floor distributor would typically be a 1000/100/10- or 100/10-Mbps switch.
- **Telecom outlet** - This is the network connection point for PCs, workstations, and print servers. File servers are typically colocated with and directly connected to the campus, building, or floor distributors, as appropriate for their intended use.

Figure: An Example Multirate Network Topology?the ISO/IEC 11801 Cable Model   32

- **Campus backbone cabling** - This is typically single- or multimode cable that interconnects the central campus distributor with each of the building distributors.
- **Building backbone cabling** - This is typically Category 5 or better UTP or multimode fiber cable that interconnects the building distributor with each of the floor distributors in the building.
- **Horizontal cabling** - This is predominantly Category 5 or better UTP cable, although a few installations are using multimode fiber.

As with UTP cable selection, the choice of link media and intermediate network nodes should always be made with an eye to future transmission rate needs and the life expectancy of the network elements, unpredictable though they may be. In the 1990s, LAN transmission rates increased 100 times and, by 2002, will increase yet another 10 times.

This does not mean that all-or even some-end stations and their interconnecting links will require gigabit capability. It does mean, however, that more central network nodes (such as most campus distributors and many building distributors) should be equipped with gigabit capability, and that all floor distributors should have at least 100 Mbps capability. It also means that all network switches should be nonblocking and that all ports should have full-duplex capability, and that any new campus backbone links should be installed with single-mode fiber.

Link Aggregation-Establishing Higher-Speed Network Trunks

Link aggregation is a recent optional MAC capability that allows several physical links to be combined into one logical higher-speed trunk. It provides the means to increase the effective data rate between two network nodes in unit multiples of the individual link transmission rate rather than in an order-of-magnitude step.

Link aggregation can be a cost-effective way to provide higher-speed connections in Ethernet LANs that are reaching saturation with 100 Mbps transmission rates but that won't require gigabit capability, at least in the short term. For example, the maximum length for 62.5 μ m multimode fiber links is 2000 meters at 100 Mbps, and multimode fiber has been often used for campus backbone links. The logical upgrade would seem to be to reuse these links for 1000 Mbps operation, but the maximum supportable length for multimode fiber is only 700 meters and only with 1000Base-LX. If the existing links are longer than 700 meters, aggregating n existing links will support an effective transmission rate of (100 n) Mbps.

Link aggregation should be viewed as a network configuration option that is primarily used in the few interconnections that require higher data rates than can be provided by single links, such as switch-to-switch and in switch-to-file server. It can also be used to increase the reliability of critical links. Aggregated links can be rapidly reconfigured (typically in about 1 second or less) in case of link failure, with low risk of duplicated or reordered frames.

Link aggregation does not affect either the IEEE 802.3 data frame format(s) or any higher layers in the protocol stack. It is backward-compatible with "aggregation-unaware" devices and can be used with any Ethernet data rate (although it does not make sense for 10 Mbps because it would likely cost less to procure a pair of 100-Mbps NICs). Link aggregation can be enabled only on parallel point-to-point links and those that support full-duplex same-speed operation.

Network Management

All higher-speed Ethernet specifications include definitions for managed objects and control agents that are compatible with Simple Network Management Protocol (SNMP) and that can be used to gather statistics about the operation of the network nodes and to assist in network management. Because user information is anecdotal at best and usually comes long after the fact, all larger networks should at least be configured with managed switches and network servers to ensure that potential problems and bottlenecks can be identified before they cause serious network deterioration.

Migrating to Higher-Speed Networks

By now, it should be apparent that upgrading existing networks typically does not require wholesale equipment or media changes, but it does require knowledge of the current network configuration and the network location of potential problems. This means that a network management system should be in place and that a cable plant database should be both available and accurate. It is time-consuming and often difficult to determine link type and availability after the cables have been pulled through conduit, buried in walls, and layered in cable trays.

Links are often the limiting factors in network upgrades. Existing Category 5 links should support all current Ethernet rates from 10 Mbps to 1000 Mbps, although they should be tested to ensure their capability to support gigabit rates. If the network is equipped with only Category 3 cable, some links will have to be replaced before upgrading to 1000 Mbps. A similar situation exists with single- and multimode fiber. Multimode fiber cannot be used for all backbone installations. Single-mode fiber, on the other hand, not only can support all backbone lengths up to 10,000 meters at 1000 Mbps, but it also will be capable of supporting backbone use at 10-gigabit data rates in the future.

Switch replacement can begin as soon as the necessary links are available. Existing switches at the campus and building distributor levels can often be reused at the building or floor distributor level. NICs can generally be replaced to extend the useful life of end stations. And so on.

Summary

The article began with an overview of the Ethernet technology, the network building blocks, and Ethernet's relationship to the ISO seven-layer reference model. The requirements for MAC and PHY compatibility also were introduced.

The basic MAC responsibilities were defined:

- **Data encapsulation** - Assembling the frame into the defined format before transmission begins, and disassembling the frame after it has been received and checked for transmission errors.
- **Media access control** - In the required CSMA/CD half-duplex mode, and in the optional full-duplex mode.

Two optional MAC capability extensions and their associated frame formats were discussed. The VLAN tagging option allows network nodes to be defined with logical as well as physical addresses, and provides a means to assign transmission priorities on a frame-by-frame basis. A specific format for the pause frame, which is used for short-term link flow control, is defined in the standard but was not covered here because it is automatic MAC capability that is invoked as needed to prevent input buffer overrun.

The PHY layer discussions included descriptions of the signaling procedures and media requirements/limitations for the following:

- 10Base-T
- 100Base-TX, 100Base-T4, and 100Base-T2
- 1000Base-T, 1000Base-CX, 1000Base-LX, and 1000Base-SX

Although 100Base-FX was not specifically discussed, it uses the same signaling procedure as 100Base-TX, but over optical fiber media rather than UTP copper.

The remaining sections of the article were devoted to systems considerations for both twisted-pair and optical fiber LAN implementations:

- Link crossover requirements in UTP networks
- Matching of PMDs and network media to ensure desired data rates
- Use of link aggregation to create higher-speed logical trunks
- Implementation of multispeed networks

After essentially finishing the article, you should have a reasonable working knowledge of the Ethernet protocol and network technology. The next section should help determine whether you need to go back and re-read the article.

Review Questions

Q - *Shouldn't all 10Base-T networks just be upgraded to 100 Mbps? Why or why not?*

A - Not necessarily-if the current 10Base-T network is repeater-based, replacing the repeaters with 10/100 nonsaturating switches would result in an automatic n times increase in the average available bandwidth for each end station.

Q - *Which 100Base version(s) are recommended? Why?*

A - 100Base-TX is recommended if the horizontal wiring is Category5 or better UTP. If the horizontal cabling is Category 3, 100BaseT4 can be used, but it may be difficult to acquire (some reports indicate that because 100Base-TX was available more than a year before T4, it captured as much as 95 percent of the market). 100Base-T2 is not available.

Q - *Which 1000Base version(s) are recommended? Where would they be used?*

A - 1000Base-T, is recommended if the horizontal cabling is Category 5 or better UTP. 1000Base-SX can be used if the horizontal cabling is multimode optical fiber, as well as for some multimode backbones. 1000Base-LX can be used for either single-mode or multimode optical fiber (see Table 7-5). 1000Base-CX can be used for short-haul equipment-room jumpers up to 25 meters.

Q - *What cable types should be used for new networks? For upgrading existing networks? Why?*

A - New or replacement UTP links may be Category 5E or better to allow for data rate growth to 1000 Mbps. Multimode fiber may be used as indicated in Table 7-5 for 1000Base-SX, or as noted in the paragraph following Table 7-5 for 1000Base-LX. (These fibers will also provide future support for shorter distances [between 100 and 300 meters, depending on the wavelength] at 10,000 Mbps.) To be truly future-proof and to ensure that you will be able to operate longer-distance backbones, choose single-mode fiber.

Q - *How do you know when a network needs to be upgraded? Where do you start?*

A - There are several ways:

- Your users will tell you (but often only after they have crossed the frustration threshold).
- Your network management system should be capable of indicating the load characteristics for each DCE port.
- Your organization is considering adding new applications (such as multimedia) that will require more communication bandwidth.
- Your organization is growing, and there are not sufficient DCE ports in the right locations to accommodate the additional users.

After you have determined the need, you can consider the options. Remember that the network elements with the longest useful life (the link media, followed by the network servers and network switches) can also be the

most expensive to replace. Choose with an eye to future growth, and consider reusing these elements wherever possible.