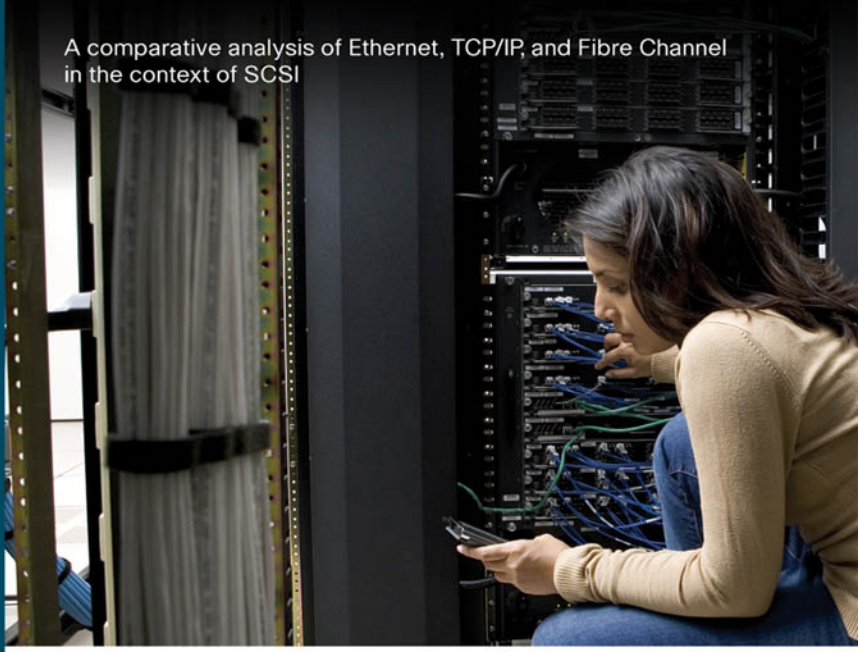
Cisco Systems

Storage Networking Protocol Fundamentals

A comparative analysis of Ethernet, TCP/IP, and Fibre Channel in the context of SCSI



ciscopress.com

James Long

Storage Networking Protocol Fundamentals

James Long

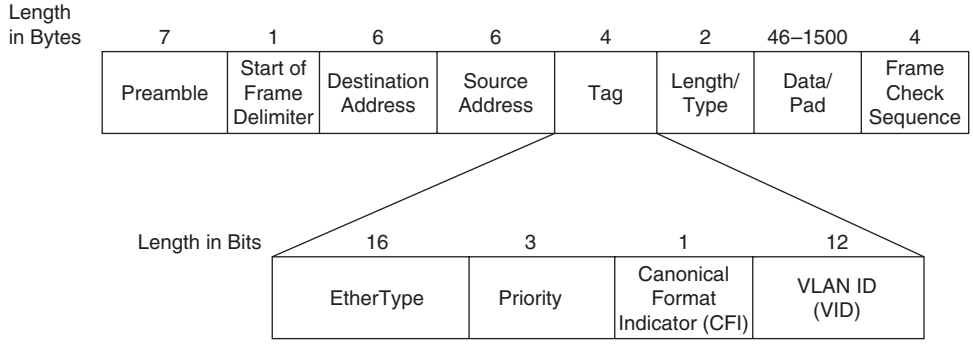
Cisco Press

800 East 96th Street
Indianapolis, Indiana 46240 USA

a switch forwards multicast and broadcast traffic, only those switch ports using the same frame format as the source node can transmit the frame without translation. All other switch ports must translate the frame format or drop the frame. Translation of every frame can impose unacceptable performance penalties on a switch, and translation is not always possible. For example, some Ethernet II frames cannot be translated to LLC format in the absence of the SNAP subheader. So, Ethernet switches do not translate frame formats. (VLAN trunking ports are a special case.) Thus, Ethernet switches drop frames when the frame format of the egress port does not match the frame format of the source node. This prevents ARP and other protocols from working properly and results in groups of devices becoming isolated. For this reason, most Ethernet networks employ a single frame format on all switch ports and attached devices.

As previously stated, VLANs require each frame sent between switches to be tagged to indicate the VLAN ID of the transmitting node. This prevents frames from being improperly delivered across VLAN boundaries. There are two frame formats for Ethernet trunking: the IEEE's 802.1Q-2003 format and Cisco Systems' proprietary ISL format. Today, most Ethernet networks use the 802.1Q-2003 frame format, which was first standardized in 1998. So, Cisco Systems' proprietary frame format is not discussed herein. Figure 5-10 illustrates the IEEE 802.1Q-2003 frame format.

Figure 5-10 *IEEE 802.1Q-2003 Frame Format*



A brief description of each Tag sub-field follows:

- EtherType—2 bytes long and must contain the value 0x8100 to indicate that the following two bytes contain priority and VLAN information. This allows Ethernet switches to recognize tagged frames so special processing can be applied.
- Priority—3 bits long. It is used to implement QoS.
- Canonical Format Indicator (CFI) bit—facilitates use of a common tag header for multiple, dissimilar network types (for example, Ethernet and Token Ring).
- VLAN ID (VID)—12 bits long. It contains a binary number between 2 and 4094, inclusive. VIDs 0, 1, and 4095 are reserved.

The brief field descriptions provided in this section do not encompass all the functionality provided by each of the fields. For more information, readers are encouraged to consult the IEEE 802.3-2002, 802.2-1998, 802-2001, and 802.1Q-2003 specifications.

Ethernet Delivery Mechanisms

Ethernet is often mistakenly considered to be a connectionless technology. In fact, Ethernet provides three types of service via the LLC sublayer. These include the following:

- Unacknowledged, connectionless service (Type 1)
- Acknowledged, connection-oriented service (Type 2)
- Acknowledged, connectionless service (Type 3)

Most Ethernet switches provide only unacknowledged, connectionless service (Type 1), which contributes to the public's misunderstanding of Ethernet's full capabilities. Because the other two service types are rarely used, the delivery mechanisms employed by the LLC sublayer to provide those types of service are outside the scope of this book. Ethernet networks that provide Type 1 service implement the following delivery mechanisms:

- Ethernet devices do not detect frames dropped in transit. When an Ethernet device drops a frame, it does not report the drop to ULPs or peer nodes. ULPs are expected to detect the drop via their own mechanisms.
- Ethernet devices do not detect duplicate frames. If a duplicate frame is received, Ethernet delivers the frame to the ULP in the normal manner. ULPs are expected to detect the duplicate via their own mechanisms.
- Ethernet devices can detect corrupt frames via the FCS field. Upon detection of a corrupt frame, the frame is dropped. Regardless of whether an intermediate switch or the destination node drops the frame, no notification is sent to any node or ULP. Some Ethernet switches employ cut-through switching techniques and are unable to detect corrupt frames. Thus, corrupt frames, are forwarded to the destination node and subsequently dropped. However, most Ethernet switches employ a store-and-forward architecture capable of detecting and dropping corrupt frames.
- Ethernet devices do not provide acknowledgement of successful frame delivery.
- Ethernet devices do not support retransmission.
- Ethernet devices support link-level flow control in a reactive manner. Ethernet devices do not support end-to-end flow control. See Chapter 9, "Flow Control and Quality of Service," for more information about flow control.
- Bandwidth is not guaranteed. Monitoring and trending of bandwidth utilization on shared links is required to ensure optimal network operation. Oversubscription on shared links must be carefully calculated to avoid bandwidth starvation during peak periods.
- Consistent latency is not guaranteed.

- The IEEE 802.3-2002 specification does not define methods for fragmentation or reassembly because the necessary header fields do not exist. An MTU mismatch results in frame drop. Thus, each physical Ethernet network must have a common MTU on all links. That means PMTU discovery is not required within an Ethernet network. MTU mismatches between physically separate Ethernet networks are handled by an ULP in the device that connects the Ethernet networks (for example, IP in a router). Likewise, an ULP is expected to provide end-to-end PMTU discovery.
- In-order delivery is not guaranteed. Ethernet devices do not support frame reordering. ULPs are expected to detect out-of-order frames and provide frame reordering.

Ethernet Link Aggregation

Clause 43 of IEEE 802.3-2002 defines a method for aggregation of multiple Ethernet links into a single logical link called a Link Aggregation Group. Link Aggregation Groups are commonly called Ethernet port channels or EtherChannels. Despite the fact that the term EtherChannel is copyrighted by Cisco Systems, the term is sometimes used generically to describe Ethernet port channels implemented on other vendors' equipment. Automation of link aggregation is supported via the IEEE's Link Aggregation Control Protocol (LACP). With LACP, links that can be aggregated will be aggregated without the need for administrative intervention. The LACP frame format contains 31 fields totaling 128 bytes. Because of the complexity of this protocol, granular description of its operation is currently outside the scope of this book. Before standardization of LACP in 2000, Cisco Systems introduced automated link aggregation via the Port Aggregation Protocol (PAgP). The details of PAgP have not been published by Cisco Systems. Thus, further disclosure of PAgP within this book is not possible. Both link aggregation protocols are in use today. The protocols are quite similar in operation, but they are not interoperable.

Automated link aggregation lowers (but does not eliminate) administrative overhead. Network administrators must be wary of several operational requirements. The following restrictions apply to Ethernet port channels:

- All links in a port channel must use the same aggregation protocol (LACP or PAgP).
- All links in a port channel must connect a single pair of devices (that is, only point-to-point configurations are permitted).
- All links in a port channel must operate in full-duplex mode.
- All links in a port channel must operate at the same transmission rate.
- If any link in a port channel is configured as non-trunking, all links in that port channel must be configured as non-trunking. Likewise, if any link in a port channel is configured as trunking, all links in that port channel must be configured as trunking.
- All links in a non-trunking port channel must belong to the same VLAN.
- All links in a trunking port channel must trunk the same set of VLANs.

- All links in a non-trunking port channel must use the same frame format.
- All links in a trunking port channel must use the same trunking frame format.

Some of these restrictions are not specified in 802.3-2002, but they are required for proper operation. Similarly, there is no de jure limit on the maximum number of links that may be grouped into a single port channel or the maximum number of port channels that may be configured on a single switch. However, product design considerations may impose practical limits that vary from vendor to vendor. The 802.3-2002 specification seeks to minimize the probability of duplicate and out-of-order frame delivery across an Ethernet port channel. However, it is possible for these outcomes to occur during reconfiguration or recovery from a link failure.

Ethernet Link Initialization

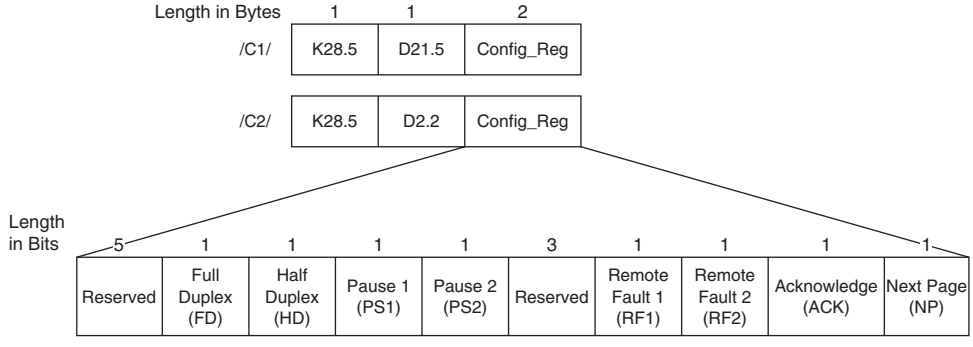
Ethernet link initialization procedures are the same for node-to-node, node-to-switch, and switch-to-switch connections. However, different procedures are observed for different types of media. FE and GE links may be configured manually or configured dynamically via auto-negotiation. 10GE does not currently support auto-negotiation. Most NICs, router interfaces, and switch ports default to auto-negotiation mode. Ethernet auto-negotiation is implemented in a peer-to-peer fashion.

Clause 37 of IEEE 802.3-2002 defines auto-negotiation for 1000BASE-X. As previously stated, auto-negotiation is accomplished via ordered sets in 1000BASE-X implementations. Therefore, 1000BASE-X implementations do not support auto-negotiation of the transmission rate because bit-level synchronization must occur before ordered sets can be recognized. So, if a 1000BASE-X device is connected to a 100BASE-FX (fiber-based FE) device, the link will not come up. When two 1000BASE-X devices are connected, operating parameters other than transmission rate are negotiated via the Configuration ordered sets /C1/ and /C2/ (collectively denoted as /C/). All capabilities are advertised to the peer device by default, but it is possible to mask some capabilities. If more than one set of operating parameters is common to a pair of connected devices, a predefined priority policy determines which parameter set will be used. The highest common capabilities are always selected. As previously stated, each /C/ ordered set carries two bytes of operating parameter information representing the transmitter's 16-bit configuration register (Config_Reg). Immediately following link power-on, alternating /C1/ and /C2/ ordered sets containing zeroes in place of the Config_Reg are transmitted by each device. This allows the other device to achieve bit-level synchronization.

Upon achieving bit-level synchronization, the receiving device begins searching the incoming bit stream for the Comma bit pattern (contained within the /K28.5/ control character) and begins transmitting alternating /C1/ and /C2/ ordered sets containing the Config_Reg. Upon recognition of the Comma bit pattern in three consecutive /C/ ordered sets without error, the receiving device achieves word alignment and begins searching the incoming bit stream for the Config_Reg. Upon recognition of three consecutive, matching Config_Regs without error, the receiving device sets the Acknowledge bit to one in its

Config_Reg, continues transmitting until the Link_Timer expires (10ms by default) and begins resolving a common parameter set. If a matching configuration is resolved, normal communication ensues upon expiration of the Link_Timer. If successful negotiation cannot be accomplished for any reason, the network administrator must intervene. Figure 5-11 illustrates the 1000BASE-X Configuration ordered sets.

Figure 5-11 *1000BASE-X Configuration Ordered Sets*



A brief description of each field follows:

- **Full duplex (FD) bit**—used to indicate whether full duplex mode is supported.
- **Half duplex (HD) bit**—used to indicate whether half duplex mode is supported.
- **Pause 1 (PS1) and Pause 2 (PS2) bits**—used together to indicate the supported flow-control modes (asymmetric, symmetric, or none).
- **Remote Fault 1 (RF1) and Remote Fault 2 (RF2) bits**—used together to indicate to the remote device whether a fault has been detected by the local device and, if so, the type of fault (offline, link error, or auto-negotiation error).
- **Acknowledge (ACK) bit**—used to indicate successful recognition of at least three consecutive matching Config_Regs.
- **Next Page (NP) bit**—indicates that one or more /C/ ordered sets follow, and each contains parameter information in one of two alternative formats: message page or unformatted page. A message page must always precede an unformatted page to indicate how to interpret the unformatted page(s). An unformatted page can be used for several purposes.

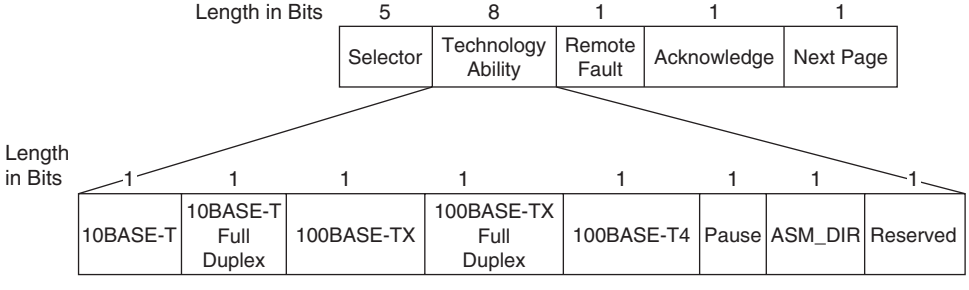
The preceding description of the 1000BASE-X link initialization procedure is simplified for the sake of clarity. For more detail about /C/ ordered set usage, Next Page formats, field interpretations, and auto-negotiation states, readers are encouraged to consult clause 37 and all associated annexes of IEEE 802.3-2002.

Clause 28 of IEEE 802.3-2002 defines auto-negotiation for all Ethernet implementations that use twisted-pair cabling. As previously stated, auto-negotiation is accomplished via the FLP in twisted-pair based GE implementations. The FLP mechanism is also used for

auto-negotiation in 100-Mbps twisted-pair based Ethernet implementations (100BASE-TX, 100BASE-T2, and 100BASE-T4). A special mechanism is defined for 10BASE-T implementations because 10BASE-T does not support the FLP. Because 10BASE-T is irrelevant to modern storage networks, only the FLP mechanism is discussed in this section. The 16 data bits in the FLP are collectively called the link code word (LCW). The LCW represents the transmitter's 16-bit advertisement register (Register 4), which is equivalent to the 1000BASE-X Config_Reg. Like 1000BASE-X, all capabilities are advertised to the peer device by default, but it is possible to mask some capabilities. If more than one set of operating parameters is common to a pair of connected devices, a predefined priority policy determines which parameter set will be used. The highest common capabilities are always selected. Unlike 1000BASE-X, the FLP is independent of the bit-level encoding scheme used during normal communication. That independence enables twisted-pair based Ethernet implementations to auto-negotiate the transmission rate. Of course, it also means that all operating parameters must be negotiated prior to bit-level synchronization. So, the FLP is well defined to allow receivers to achieve temporary bit-level synchronization on a per-FLP basis. The FLP is transmitted immediately following link power-on and is repeated at a specific time interval.

In contrast to the 1000BASE-X procedure, wherein /C/ ordered sets are initially transmitted without conveying the Config_Reg, twisted-pair based implementations convey Register 4 via the LCW in every FLP transmitted. Upon recognition of three consecutive matching LCWs without error, the receiving device sets the Acknowledge bit to one in its LCW, transmits another six to eight FLPs, and begins resolving a common parameter set. If a matching configuration is resolved, transmission of the Idle symbol begins after the final FLP is transmitted. Transmission of Idles continues until bit-level synchronization is achieved followed by symbol alignment. Normal communication then ensues. If successful negotiation cannot be accomplished for any reason, the network administrator must intervene. Figure 5-12 illustrates the Ethernet FLP LCW.

Figure 5-12 *Ethernet FLP Link Code Word*



A brief description of each field follows:

- **Selector**—5 bits long. It indicates the technology implemented by the local device. Valid choices include 802.3, 802.5, and 802.9.