

7. Ethernet Essentials

Ethernet is the dominant local area network (LAN) technology and has enjoyed a huge success in all segments of the LAN marketplace. There are numerous reasons for this success:

- Ethernet is inexpensive
- Ethernet uses simple hardware
- Ethernet uses low-cost electrical cables or optical cables for higher speeds and longer distances
- Ethernet has a simple frame structure and link-level protocols
- Ethernet can transport information associated with a wide variety of protocols. TCP/IP is one of the most common protocols.

In the Open Systems Interconnect (OSI) reference model, Ethernet is a Layer-2 network that provides a data link for transporting higher-level information (such as TCP/IP). While not usually described as such, Fiber Channel is also essentially a Layer-2 network. The OSI reference model is shown in Figure 7-1 on page 89.

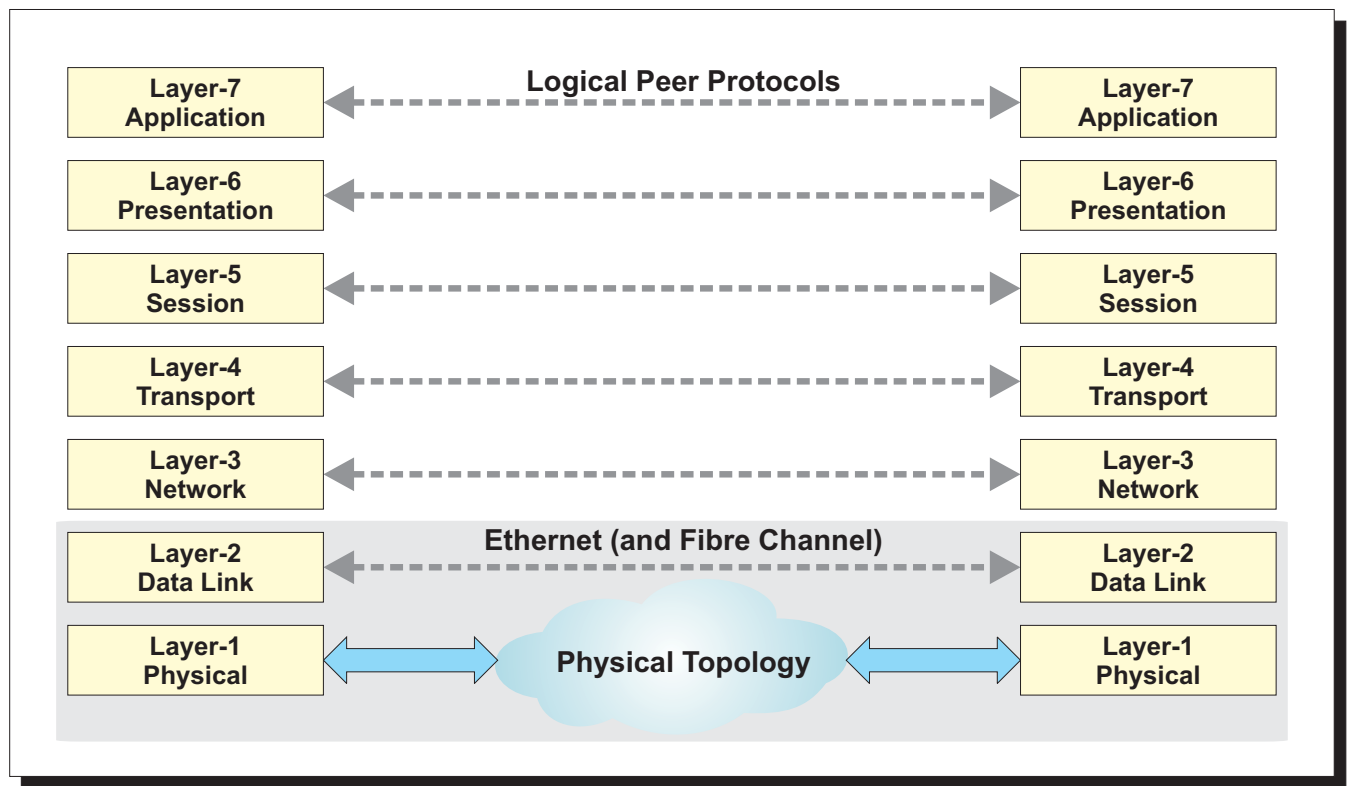


Figure 7-1. Ethernet and the OSI Reference Model

Fiber Channel over Ethernet is based on the availability of certain optional Ethernet characteristics that are not part of the behavior required by Ethernet standards. Because of this, this book uses the term “Enhanced Ethernet” to distinguish an Ethernet environment having those characteristics from a standard Ethernet environment. In some publications, you may see the terms “Data Center Ethernet (DCE)” or “Converged Enhanced Ethernet (CEE)” also used. Details of enhancements being discussed for Data Center Ethernet are described in *Data Center Ethernet* on page 241.

The Enhanced Ethernet attributes required by FCoE are:

- Lossless frame delivery (see *Making Ethernet “Lossless”* on page 104)
- In-order frame delivery (provided by the *Spanning Tree Protocol (STP)* on page 97)
- Full-duplex operation, and
- In order to encapsulate full-sized Fibre Channel frames, Ethernet support for baby-jumbo frames of at least 2.5 KB is required (see *Ethernet Jumbo Frames* on page 93)

None of these requires new functions beyond what already exists within the Ethernet standards or is commonly implemented in products. FCoE does require functions that are optional in the Ethernet standards and FCoE may benefit from functions beyond those that are currently specified in the Ethernet standards (such as an enhanced flow control method or congestion management). For performance reasons, it is also desirable to have high-speed adapters and switches with low-latency characteristics.

7.1 Ethernet Frame Format

To minimize hardware complexity and cost while providing the utmost in flexibility, Ethernet uses a very simple frame format as shown in Figure 7-2.

Preamble. An Ethernet frame begins with a Preamble followed by the Start-of-Frame delimiter and ends with an End-of-Frame delimiter. The preamble and delimiters are not considered to be part of the frame and the nature of these delimiters depends on the physical link that is being used.

Destination Address and Source Address. The Destination Address (DA) field specifies the destination of the frame and the Source Address (SA) field the source of the frame. The format of the addresses is described in *MAC Address Format* on page 91.

EtherType. The EtherType field has two different interpretations (largely based on historical reasons). If the value in the EtherType field is less than 1500 (0x5DC), it specifies the length of the frame. If the value in the EtherType field is more than 1536, it identifies the protocol carried within the frame. Using the EtherType field to identify the protocol is the more common usage of the field.

A current listing of assigned EtherType values can be found at:

<http://standards.ieee.org/regauth/ethertype/eth.txt>

The EtherType value is 8906h for FCoE and 8914h for the FCoE Initialization Protocol.

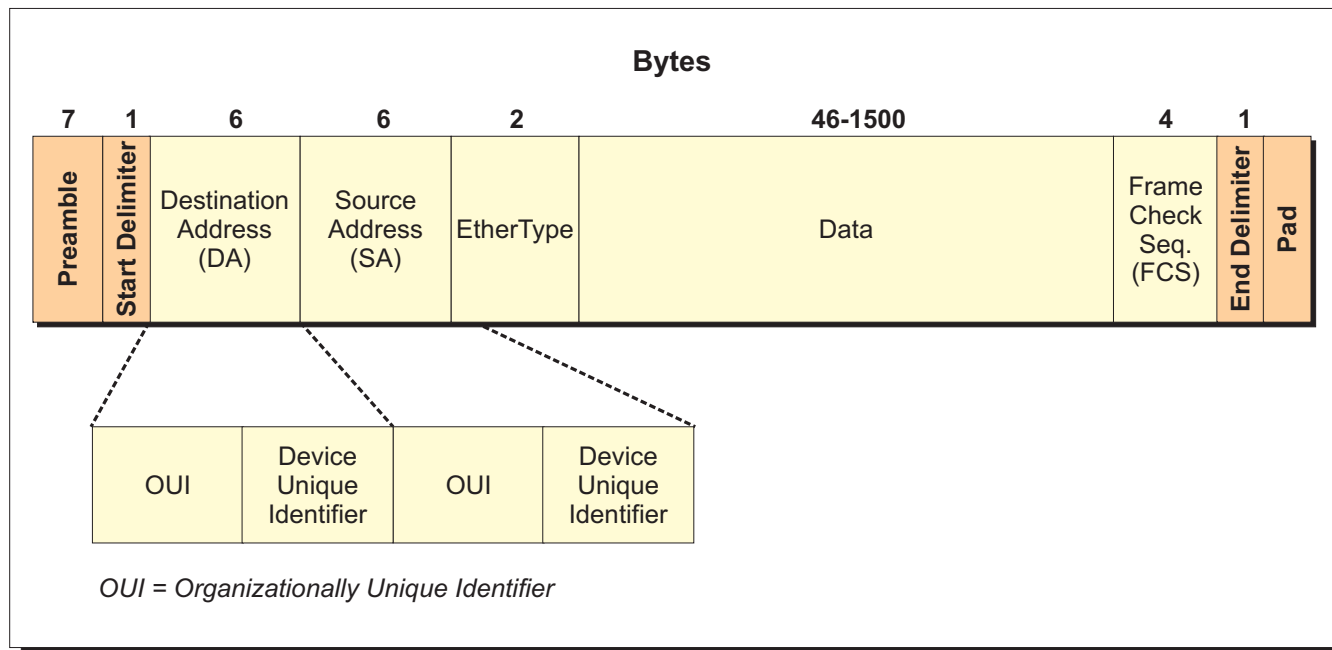


Figure 7-2. Ethernet Frame Format

Data. The Data portion of the frame contains the information being transported from the Source address to the Destination address. The size of the data portion of a standard Ethernet frame is limited to a maximum of 1500 bytes.

Frame Check Sequence (FCS). The Frame Check Sequence (FCS) is a 32-bit cyclic redundancy check (CRC) computed on the frame content beginning with the Destination Address. The algorithm is based on the same polynomial as used by Fiber Channel and is computed using the following 32-bit polynomial:

$$X^{32} + X^{26} + X^{23} + X^{22} + X^{16} + X^{12} + X^{11} + X^{10} + X^8 + X^7 + X^5 + X^4 + X^2 + X + 1$$

7.1.1 MAC Address Format

Each Ethernet adapter has an Ethernet address that is commonly referred to as the Media Access Control, or MAC address. The MAC address is usually personalized at the time of manufacture and often called the “burned-in” MAC address. An Ethernet MAC address is 48 bits long and has the format shown in Figure 7-3.

The first 24 bits are the Organizationally Unique Identifier (OUI). Normally, the OUI is a value assigned to an organization by IEEE to ensure uniqueness among different organizations. This is referred to as a Universally Administered OUI and is indicated by setting bit 41 to a zero. A list of assigned OUI values is available at:

<http://standards.ieee.org/regauth/oui/oui.txt>

An OUI may also be locally administered. This is indicated by setting bit 41 to a one. A locally administered OUI must be unique within a given Ethernet network, but may not be globally unique.

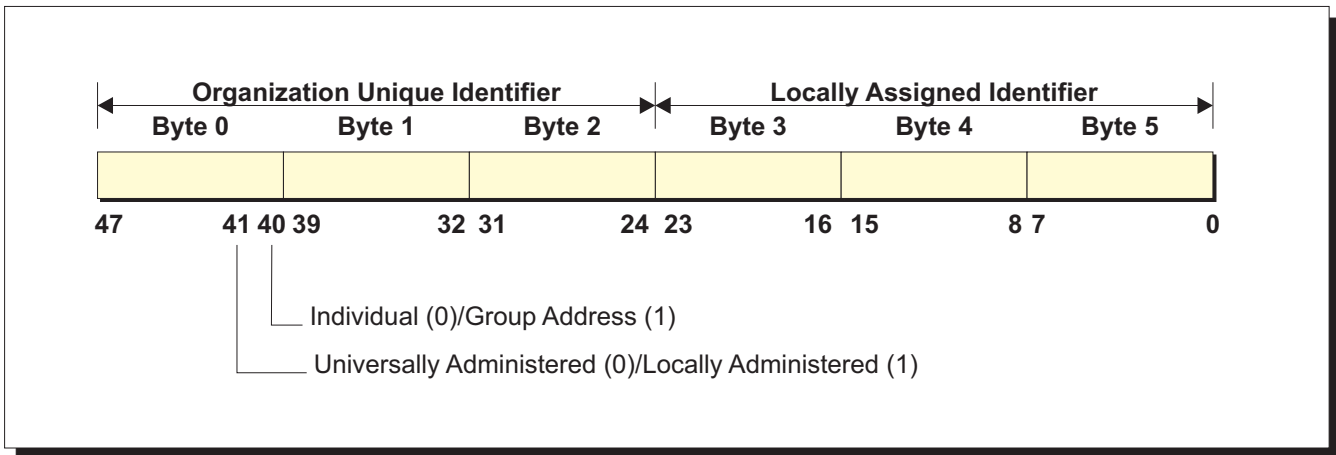


Figure 7-3. Ethernet MAC Address Format

The remaining 24 bits of the 48-bit MAC address are the device unique identifier. This is a unique value assigned within an organization.

The combination of the Universally Administered 24-bit Organizationally Unique Identifier and the 24-bit device unique identifier result in a 48-bit globally unique identifier.

7.1.2 Reserved Ethernet Group MAC Addresses

The Ethernet standards reserve a set of group addresses that are used for various link-level protocols and are not forwarded by an Ethernet switch. A listing of these Group MAC Addresses is provided in Table 7-1.

Assignment	MAC Address
Bridge Group Address	01-80-C2-00-00-00
IEEE Std 802.3x Full Duplex PAUSE operation	01-80-C2-00-00-01
IEEE Std 802.3ad Slow_Protocols_Multicast address (See <i>Link Aggregation (NIC Teaming)</i> on page 107 for an example of the usage of this address)	01-80-C2-00-00-02
IEEE P802.1X PAE address	01-80-C2-00-00-03
IEEE MAC-specific control protocols	01-80-C2-00-00-04
Reserved for future standardization	01-80-C2-00-00-05
Reserved for future standardization	01-80-C2-00-00-06
Reserved for future standardization	01-80-C2-00-00-07
Provider Bridge group address	01-80-C2-00-00-08
Reserved for future standardization	01-80-C2-00-00-09
Reserved for future standardization	01-80-C2-00-00-0A

Table 7-1. Ethernet Group MAC Addresses (Part 1 of 2)

Assignment	MAC Address
Reserved for future standardization	01-80-C2-00-00-0B
Reserved for future standardization	01-80-C2-00-00-0C
Provider Bridge MVRP address	01-80-C2-00-00-0D
IEEE Std 802.1ab Link Layer Discovery Protocol (LLDP)	01-80-C2-00-00-0E
Reserved for future standardization	01-80-C2-00-00-0F

Table 7-1. Ethernet Group MAC Addresses (Part 2 of 2)

7.1.3 FCoE Ethernet Group (Multicast) MAC Addresses

FCoE has reserved three Ethernet group addresses for multicast operations. These addresses are listed in Table 7-2 on page 93.

Assignment	MAC Address
ALL_FCoE_MACS	01-10-18-01-00-00
ALL_ENODE_MACS	01-10-18-01-00-01
ALL_FCF_MACS	01-10-18-01-00-02

Table 7-2. FCoE Group (Multicast) MAC Addresses

7.1.4 Ethernet Jumbo Frames

In an Ethernet environment where each frame interrupts the software, minimizing the number of interrupts, and associated software processing, can improve the overall efficiency. To provide better performance, some Ethernet devices support (non-standard) larger frame sizes referred to as “jumbo” frames that may be up to 9 KB in size. Because a 9 KB jumbo frame carries as much data as six standard-size Ethernet frames, the number of interrupts is reduced by a factor of six with the resulting improvement in performance.

While jumbo frames are not part of the Ethernet standard, they are widely supported by higher performance Ethernet implementations.

Jumbo frames also provide an answer to the transport of encapsulated Fiber Channel frames by FCoE. While a standard Ethernet frame cannot contain a full-sized Fiber Channel frame, an Ethernet baby jumbo frame of approximately 2.5 KB can and will be required by FCoE.

NOTE – While FCoE could use the normal Fiber Channel methods to establish a smaller Receive Data Field size during FLOGI and PLOGI so that encapsulated FC frames would fit within a standard Ethernet frame, the direction of the FCoE standard is to require support for Ethernet jumbo frames.

7.2 Ethernet Topologies

Ethernet supports multiple topology configurations and medium types. Supported topology configurations include, multi-tap cable, hubs, and switched fabrics.

7.2.1 Shared-Medium Topology

Ethernet devices can connect to a shared “bus” consisting of a single coaxial cable as shown in Figure 7-4. By using a hub, this topology can also use unshielded twisted pair cabling.

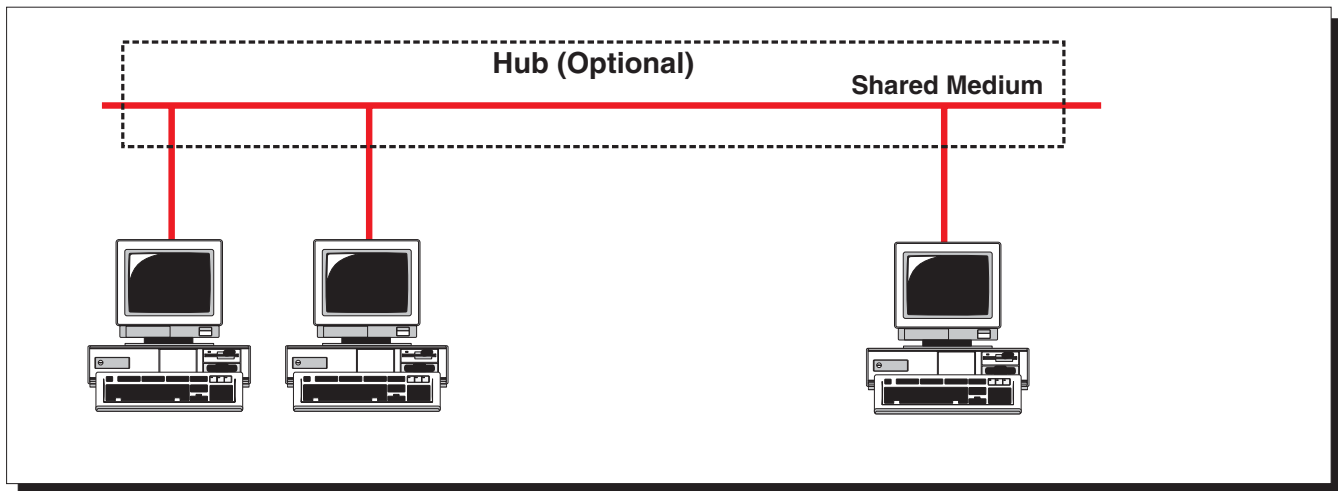


Figure 7-4. Ethernet Shared Medium Topology

Devices in this topology use the Carrier Sense Multiple Access/Collision Detect (CSMA/CD) access protocol. When a device need to transmit a frame, it:

1. Listens for traffic (Carrier Sense)
2. If none, it transmits the frame and listens to the bus at the same time
3. If another device also transmits at the same time (Multiple Access), there is a “collision”
4. The collision is detected because the transmitted frame is corrupted (Collision Detect)
5. If a collision occurs, the device backs off and tries again later

The shared medium topology using CSMA/CD is a common configuration for 10 and 100 megabit Ethernet (although most 100 megabit Ethernet is now based on the switched topology configuration).

While initially inexpensive, the shared-medium topology has been largely replaced by the switched topology. This is due to a number of limitations inherent in this type of topology:

- Because transmission and reception take place on the same coaxial cable or unshielded twisted pair, this topology only supports half-duplex behavior. That is, a device can transmit or receive a frame at any point in time but cannot do both at the same time (other than receiving the frame it is currently sending).
- When a collision occurs, no useful information is transferred. This represents wasted bandwidth on the link. As the level of activity increases, the number of collisions increase and the throughput is severely impacted.
- Only one device can be transmitting at a time. This limits the overall throughput that a shared medium topology can provide.

7.2.2 Ethernet Switched Fabric Topology

The switched topology is based on the use of one or more Ethernet switches. An illustration of a switched topology is shown in Figure 7-5. In this topology, each device connects to a port on an Ethernet switch (note that a shared-medium topology could also connect to a switch port).

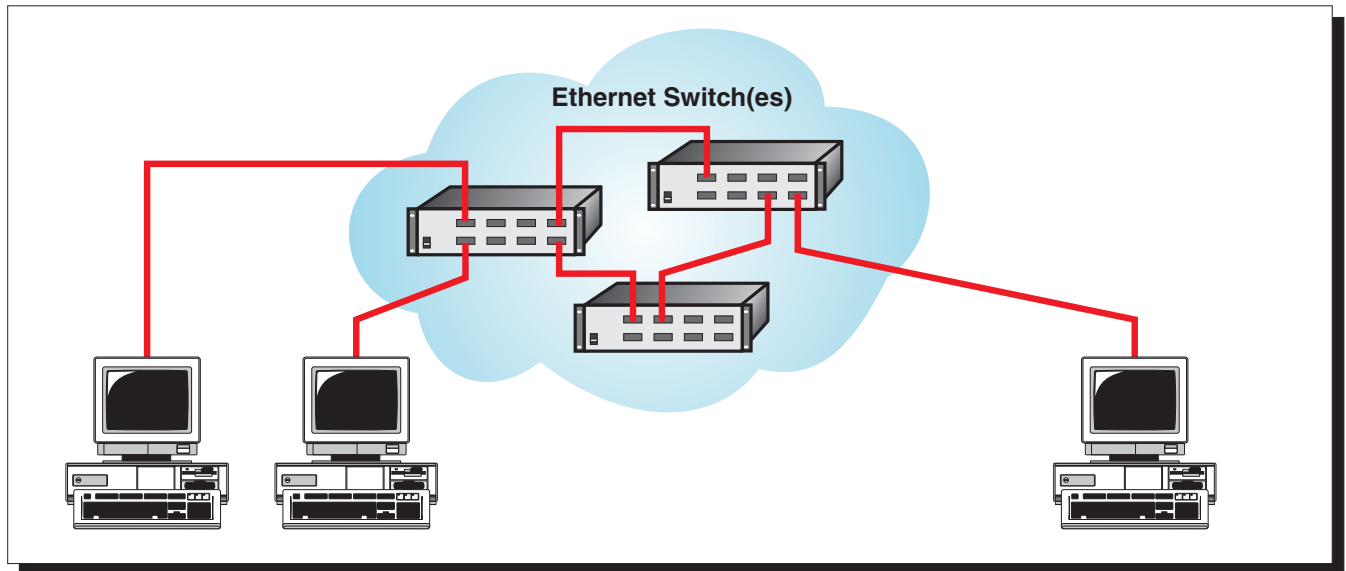


Figure 7-5. Ethernet Switched Topology

Devices that are connected to a switch may still implement the CSMA/CD access protocol to determine when a frame can be sent, but each link is in effect a separate collision domain. As a result, collisions caused by other devices are eliminated along with the wasted bandwidth and the need to retransmit frames as the result of a collision.

Finally, the links between devices and switch ports are usually full-duplex links that enable simultaneous frame transmission and reception. Full-duplex operation potentially doubles the available bandwidth when compared to half-duplex operation.

All one-gigabit and ten-gigabit Ethernet use the switched topology, effectively removing collisions and the need for the CSMA/CD protocol.

Ethernet networks may consist of multiple switches interconnected to create an Ethernet switched fabric. Using multiple switches enables larger configurations to be created (more ports and physically distributed) than would be possible by using a single switch.

7.2.3 Ethernet Switch Learning

Ethernet switches have no control over the MAC addresses of attached devices. MAC addresses are normally assigned to an Ethernet NIC and the time of manufacture and an Ethernet switch has no control over which NIC is connected to which switch port. Instead of controlling addressing as is done in Fiber Channel, Ethernet switches learn the addresses of attached devices. Each Ethernet switch has a filtering database that associates MAC addresses with switch ports.

When a switch receives a unicast frame, it looks at the Source Address (SA) in the received frame. If the Source Address it is not in the filtering database, the switch associates that switch port with the MAC address and enters it into the filtering database.

The switch also looks in its filtering database to see if it already has an entry matching the Destination Address.

- If the Destination Address (DA) is in the filtering database, the switch forwards the frame out the associated switch port. Because it had previously received a frame from that address in on that switch port, it know the destination is reachable via that port.
- If the Destination Address is not in the filtering database, the switch has no knowledge of the location of the destination and forwards the frame out all of its other ports. This ensures that the frame will reach the destination, if it exists.

Using this learning approach, each Ethernet switch learns the MAC addresses off all devices sending frames through that switch and the associated switch port. An example of the association of MAC addresses with switch ports is shown in Figure 7-6.

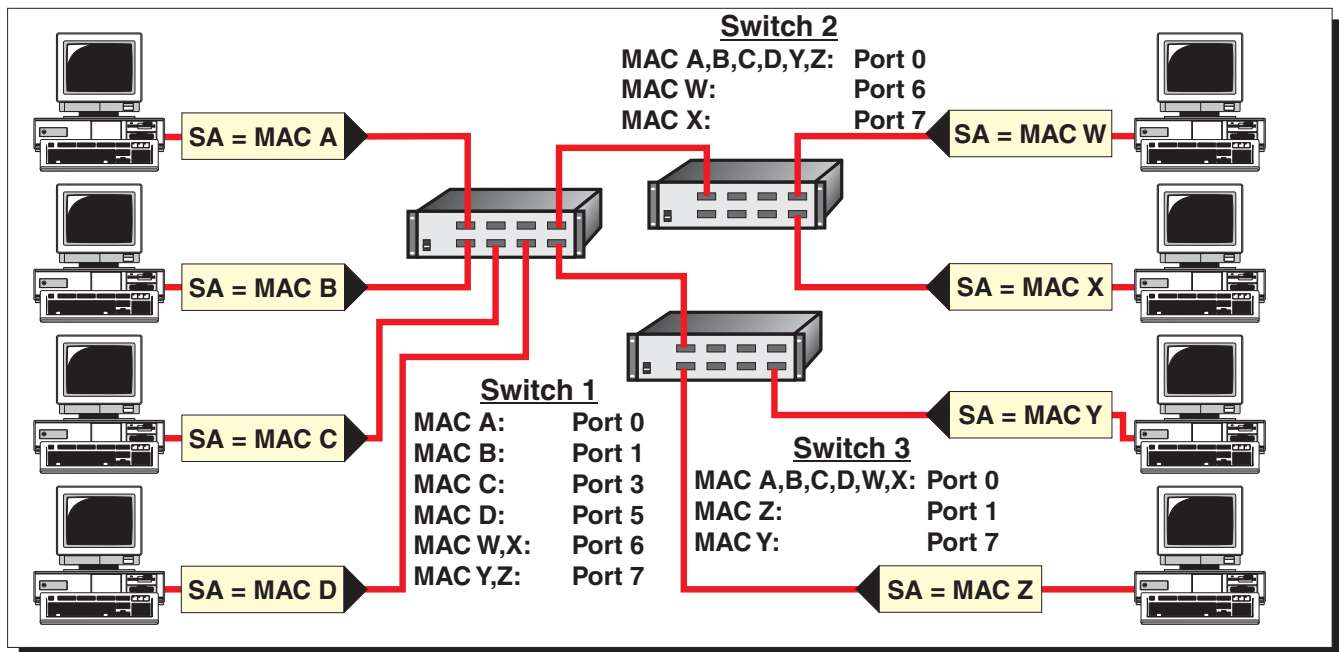


Figure 7-6. Ethernet Switch Learning Database

Because a device may be removed from the network after its MAC address has been learned by one or more switches, a method is required to remove its address. Removal is accomplished through an aging process. If there has been no activity for a given MAC address and the aging time expires (the recommended default value is 300 seconds), the entry is removed from a switch's forwarding table. An address may also be removed from a switch's forwarding table in order to make room for a newly-learned MAC address.

7.2.4 Spanning Tree Protocol (STP)

When an Ethernet network consists of multiple switches, the Ethernet switches use a Spanning Tree Protocol to identify links to other switches, prevent loops within the fabric and re-route traffic around failed inter-switch links, if possible.

The Spanning Tree Protocol creates a tree structure within the switched network by identifying a root switch and disabling redundant links that could result in loops within the network. An illustration of a network with disabled links is shown in Figure 7-7 on page 97. This example assumes that Switch B becomes the root switch (disabled links are marked with an X in the figure).

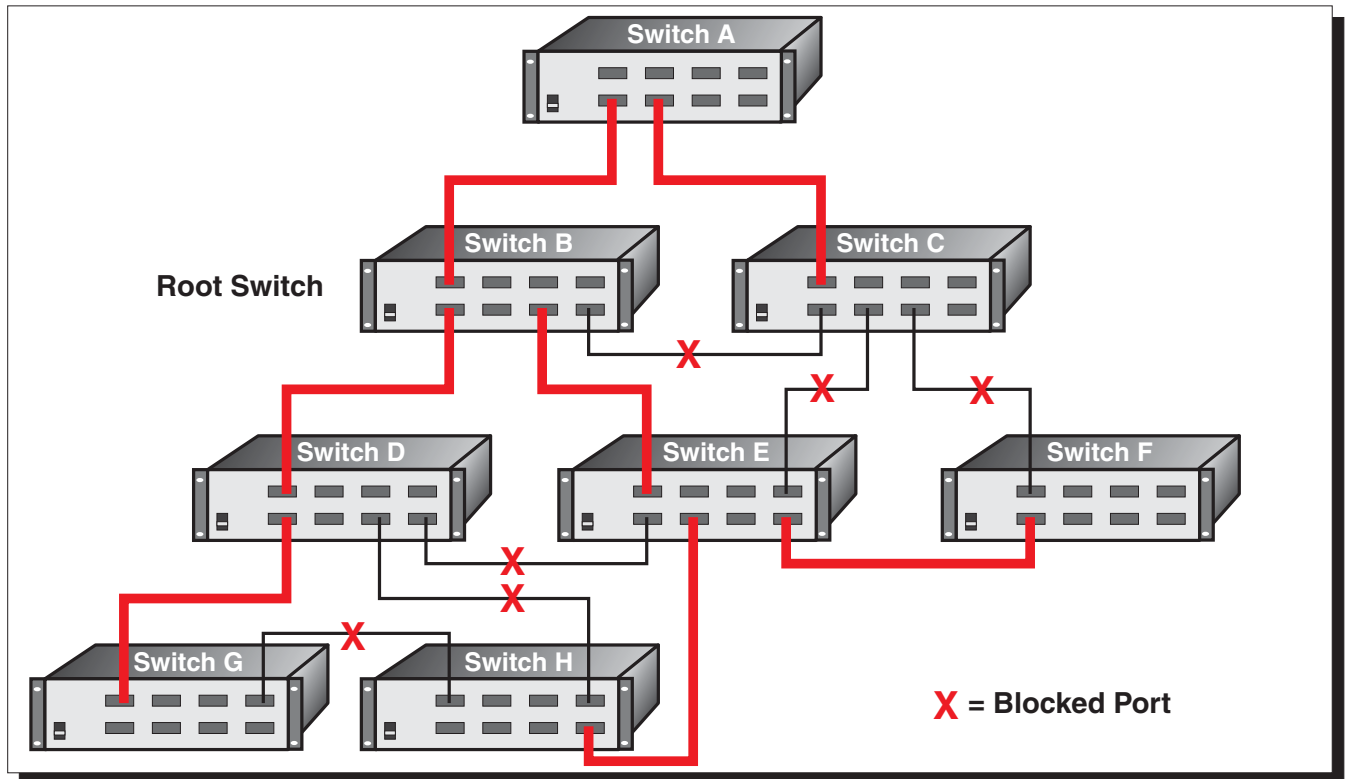


Figure 7-7. Ethernet Spanning Tree

While the tree structure created by the Spanning Tree Protocol is not immediately evident from Figure 7-7, it becomes clear when the same network is redrawn as shown in Figure 7-8. As can be seen, the result is a tree structure with each switch (and attached devices) having one, and only one, path to every other switch and device.

While a spanning tree eliminates loops within the fabric and ensures that frames are delivered in order, it does not allow redundant links or paths. Disabled links carry no traffic and the active links are the only links allowed to carry frames. This has the potential to create excessive congestion on the active links and the subsequent poor performance.

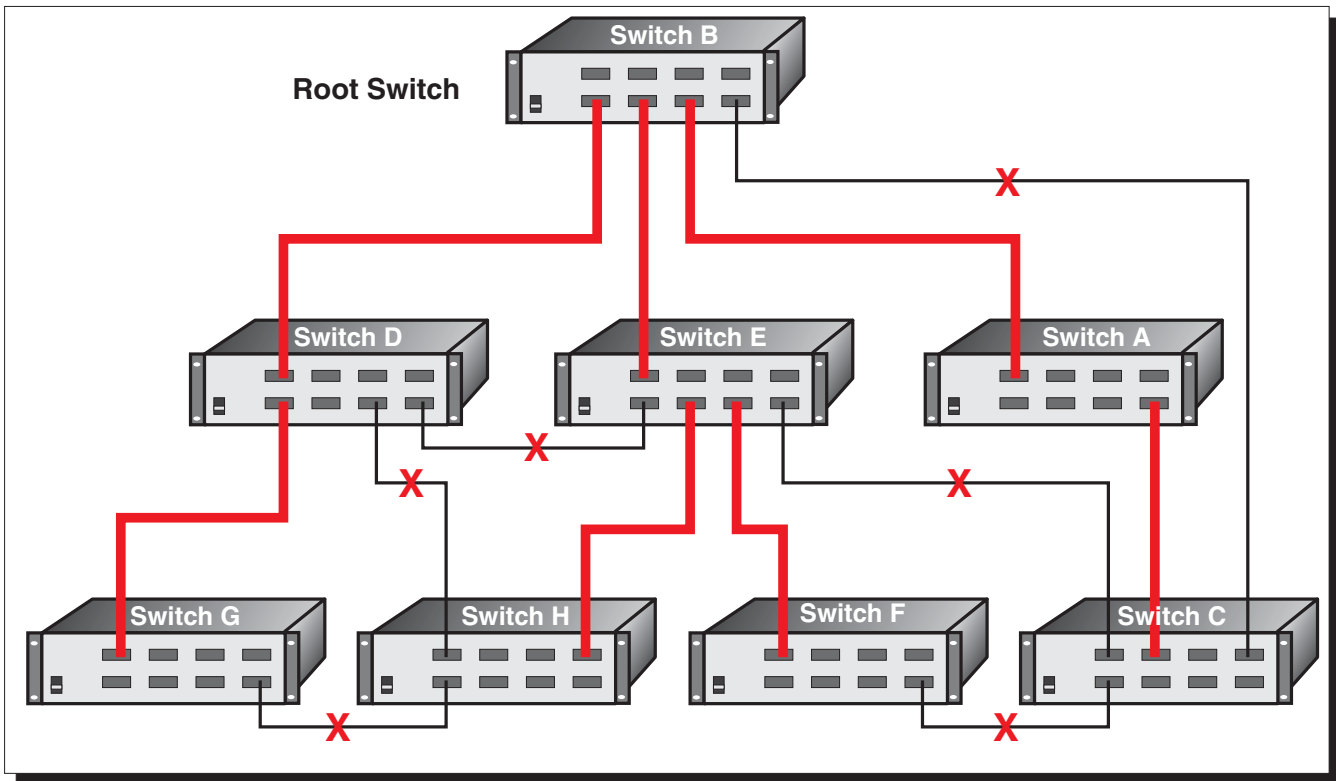


Figure 7-8. Ethernet Spanning Tree (Redrawn)

7.2.5 Per-VLAN Spanning Tree Protocol (PVST)

The basic Spanning Tree Protocol creates a single spanning tree for the entire Ethernet network. When VLANs are being used, it may be more efficient to create separate spanning trees for each VLAN. This may reduce the congestion on inter-switch links and provide better performance.

7.3 Ethernet Physical Link Variants

Like Fiber Channel, Ethernet supports multiple speeds and transmission mediums. Each variant is identified using a nomenclature consisting of the speed, signaling type and cable type. For example, 100BASE-T is 100 megabit baseband signaling over unshielded twisted pair cabling. 1000BASE-T is 1,000 megabit (1 gigabit) baseband signaling over unshielded twisted pair cabling.

Table 7-3 on page 99 summarizes some of the different physical link variants that have been defined for 100 megabit, 1 gigabit and 10 gigabit Ethernet (Note that many of the defined physical variants have not been widely used and are not listed in the table). It is also interesting to note that there are 10 gigabit variants that use multiple lanes to provide the required bandwidth (e.g., 10 GBASE-LX4 and 10 GBASE-CX4).

Name	Standard	Description
100BASE-T		A term for any of the three standards for 100 Mbit/s Ethernet over twisted pair cable up to 100 meters long. Includes 100BASE-TX, 100BASE-T4 and 100BASE-T2. All of which use a star topology.
100BASE-TX	802.3 (24)	4B5B MLT-3 coded signaling, CAT5 unshielded twisted pair (UTP) copper cabling with two twisted pairs.
100BASE-FX	802.3 (24)	4B5B NRZI coded signaling, two strands of multi-mode optical fiber. Maximum length is 400 meters for half-duplex connections (to ensure collisions are detected) or 2 kilometers for full-duplex.
100BASE-SX	TIA	100 Mbit/s Ethernet over multi-mode optical fiber. Maximum length is 300 meters. Unlike 100BASE-FX that uses a laser as the light source, 100BASE-SX uses LEDs and is less expensive.
100BASE-BX10	802.3	100 Mbit/s Ethernet bidirectionally over a single strand of single-mode optical fiber. A multiplexer is used to split transmit and receive signals into different wavelengths allowing them to share the same fiber. Supports up to 10 km.
100BASE-LX10	802.3	100 Mbit/s Ethernet up to 10 km over a pair of single mode fibers.
1 Gigabit Ethernet		
1000BASE-T	802.3 (40)	PAM-5 coded signaling using CAT5/CAT5e/CAT6 unshielded twisted pair (UTP) copper cables with four bi-directional twisted pairs.
1000BASE-SX	802.3	8B10B NRZ coded signaling, multi-mode fiber (up to 550 m).
1000BASE-LX	802.3	8B10B NRZ coded signaling, multi-mode fiber (up to 550 m) or single-mode fiber (up to 2 km; can be optimized for longer distances, up to 10 km).
1000BASE-LH	multi-vendor	A long-haul solution using 8B10B NRZ coded signaling over single-mode fiber (up to 100 km).
1000BASE-CX	802.3	8B10B NRZ coded signaling, balanced shielded twisted pair (up to 25 m) over special copper cable. Predates 1000BASE-T and rarely used.
1000BASE-BX10	802.3	Up to 10km. Bidirectional over single strand of single-mode fiber.
1000BASE-LX10	802.3	Up to 10 km over a pair of single-mode fibers.
1000BASE-PX10-D	802.3	Downstream (from head-end to tail-ends) over single-mode fiber using point-to-multipoint topology (supports at least 10 km).
1000BASE-PX10-U	802.3	Upstream (from a tail-end to the head-end) over single-mode fiber using point-to-multipoint topology (supports at least 10 km).
1000BASE-PX20-D	802.3	Downstream (from head-end to tail-ends) over single-mode fiber using point-to-multipoint topology (supports at least 20 km).
1000BASE-PX20-U	802.3	Upstream (from a tail-end to the head-end) over single-mode fiber using point-to-multipoint topology (supports at least 20 km).

Table 7-3. Ethernet Physical Link Variants (Part 1 of 2)

Name	Standard	Description
1000BASE-ZX	Unknown	Up to 100 km over single-mode fiber.[1]
10 Gigabit Ethernet		
10GBASE-SR	802.3ae	Designed to support short distances over deployed multi-mode fiber cabling, it has a range of between 26 m and 82 m depending on cable type. It also supports 300 m operation over a new 2000 MHz.km multi-mode fiber.
10GBASE-LX4	802.3ae	Uses wavelength division multiplexing to support ranges of between 240 m and 300 m over deployed multi-mode cabling. Also supports 10 km over single-mode fiber.
10GBASE-LR	802.3ae	Supports 10 km over single-mode fiber
10GBASE-ER	802.3ae	Supports 40 km over single-mode fiber
10GBASE-SW	802.3ae	A variation of 10 GBASE-SR using the WAN PHY, designed to interoperate with OC-192 / STM-64 SONET/SDH equipment
10GBASE-LW	802.3ae	A variation of 10 GBASE-LR using the WAN PHY, designed to interoperate with OC-192 / STM-64 SONET/SDH equipment
10GBASE-EW	802.3ae	A variation of 10 GBASE-ER using the WAN PHY, designed to interoperate with OC-192 / STM-64 SONET/SDH equipment
10GBASE-CX4	802.3ak	Designed to support short distances over copper cabling, it uses InfiniBand 4x connectors and CX4 cabling and allows a cable length of up to 15 m.
10GBASE-T	802.3an	Uses unshielded twisted-pair wiring.
10GBASE-LRM	draft 802.3aq	Extend to 220 meters over deployed 500 MHz.km multimode fiber
40GBASE-?	tbd	40 Gigabit Ethernet (to be defined)
100GBASE-?	tbd	100 Gigabit Ethernet (to be defined)

Table 7-3. Ethernet Physical Link Variants (Part 2 of 2)

7.3.1 Ethernet Transceivers

The majority of 100 megabit Ethernet devices use unshielded twisted pair (UTP) cabling and fixed transceivers. At the higher data rates, Ethernet devices may use pluggable transceiver modules. Pluggable transceivers are not specified by the Ethernet standards, but rather as Multi-Source Agreements (MSAs) developed by industry alliances.

XENPAK. XENPAK is a 10 Gbps Ethernet (10GbE) transceiver that incorporates the complete transmit and receive physical layer functionality from the 10.3 Gbps optical interface to the XAUI (4 lanes at 3.125 Gbps) electrical interface, including 8B/10B and 64B/66B coding.

An illustration of the XENPAK module is shown in Figure 7-9.

XPAK. XPAK is a second generation, hot pluggable, 10 Gbps optical module designed for Enterprise and SAN applications. It addresses need for smaller footprint, top side pluggable module using the industry standard, proven XAUI interface. The electrical interface is identical to 70 pin XENPAK 2.1 interface.

An illustration of the XPAK module is shown in Figure 7-10.

XPAK features a bezel opening of 1.54" by 0.506" and extends 2.685" behind the bezel. Unlike the XENPAK, which requires a cutout in the PC board, XPAK features single side mounting and allows 10 units across on a line card or can be stacked for 20 on a line card.

XPAK features 4 watts power dissipation with internal SERDES, supports uncooled laser applications up to 10 km today. It supports serial 850 nm (multi-mode) and 1310nm (single-mode) fiber with plans to include 1550nm in the future.

X2. "X2" is a new multi-source agreement (MSA) supported by leading networking component suppliers. X2 defines a smaller form-factor 10 Gbps pluggable fiber optic transceiver optimized for 802.3ae Ethernet, ANSI/ITUT OC192/STM-64 SONET/SDH interfaces, ITUT G.709, OIF OC192 VSR, INCITS/ANSI 10GFC (10 Gigabit Fibre Channel) and other 10 Gigabit applications. An illustration of the XPAK module is shown in Figure 7-11.



Figure 7-9. XENPAK Transceiver Module

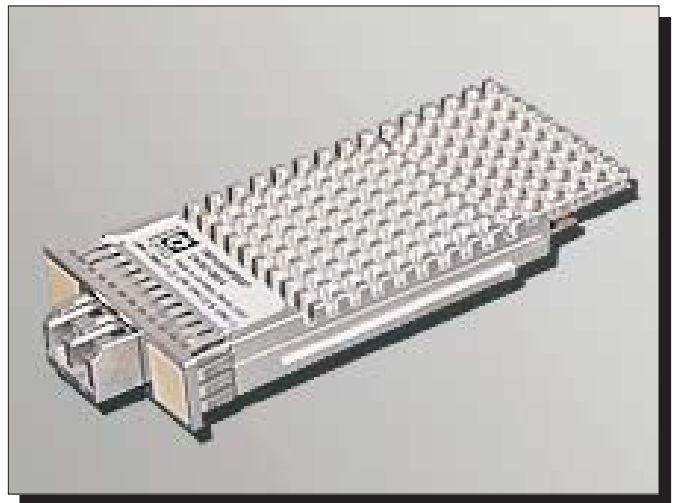


Figure 7-10. XPAK Transceiver Module

X2 is initially focused on optical links to 10 kilometers and is ideally suited for Ethernet, Fibre Channel and telecommunication switches and standard PCI (peripheral component interconnect) based server and storage connections, where a “half size” XENPAK optical transceiver is desired.

X2 is physically smaller than XENPAK but maintains the mature electrical I/O specification defined by the XENPAK MSA and continues to provide robust thermal performance and electromagnetic shielding. Electrically, X2 is compatible with the XENPAK MSA. X2 uses the same Tyco Electronics-designed, 70-pin electrical connector as XENPAK supporting four wire XAUI (10-gigabit attachment unit interface). X2 also will support the OIF SFI4_P2 interfaces and serial electrical interfaces as they emerge.

The X2 optical platform has been designed so that the heat sink and front bezel can be easily adapted to the different needs of the key 10 Gb markets. X2 can be mounted on the front panel, mid board, or in a conventional PCI card. X2's flexibility to address a wide range of high-bandwidth applications is expected to drive higher volumes on this one platform, thereby leading to lower optics costs.

XFP. The XFP (10 Gigabit Small Form Factor Pluggable) is a hot-swappable, protocol-independent optical transceiver, typically operating at 850nm, 1310nm or 1550nm, for 10 gigabit per second SONET/SDH, Fibre Channel, gigabit Ethernet, 10 gigabit Ethernet and other applications, including DWDM links. It includes digital diagnostics similar to SFF-8472, but more extensive, that provide a robust management tool. An illustration of the XFP module is shown in Figure 7-12.

The XFI electrical interface specification is a portion of the XFP Multi Source Agreement specification and uses a single lane operating at 10.3125 Gbps when using 64B/66B encoding.

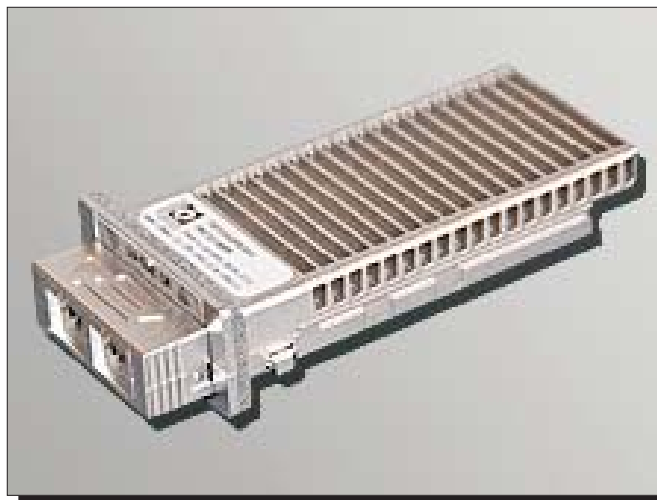


Figure 7-11. X2 Transceiver Module



Figure 7-12. XFP Transceiver Module

identify a VLAN. These two fields can be used independently of one another (e.g., you can have priority without using VLANs or vice-versa).

When present, the 802.1Q tag follows the source MAC address as shown in Figure 7-14.

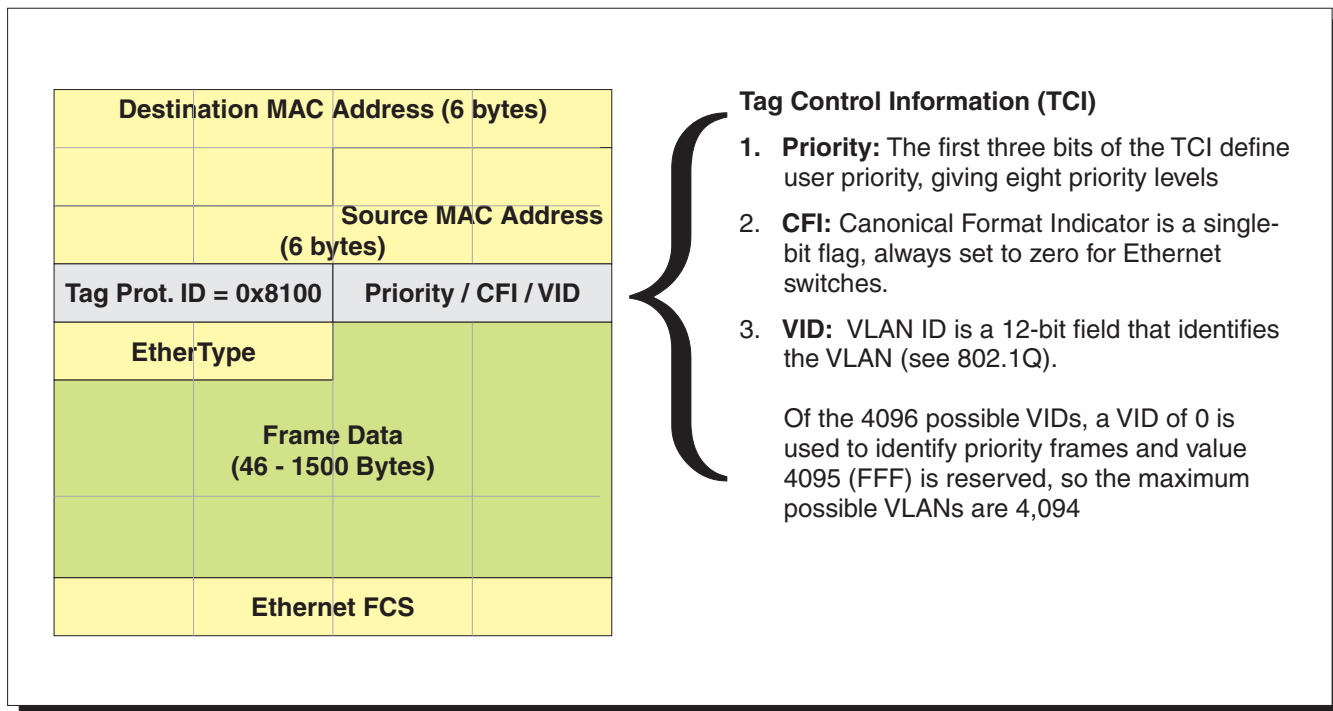


Figure 7-14. Ethernet Frame with 802.1Q VLAN Tag

7.4.2 Static and Dynamic VLANs

Static VLANs are created by assigning switch ports to a VLAN. When a device is connected to the switch port it becomes part of the associated VLAN. If the device is moved to a different switch port, it may become part of a different VLAN. The device itself probably has no awareness of the VLAN assignment and the Ethernet switch will insert or remove the VLAN Tag as appropriate.

Dynamic VLANs are created by assigning devices to a VLAN based on their MAC address or a username entered during a login. As the device enters the network, it queries a database for VLAN membership using the VLAN Query Protocol (VQP). The query goes to the VLAN Membership Policy Server (VMPS) that informs the device of its VLAN membership. If the device is moved to a different switch port, it retains its VLAN membership.

7.5 Making Ethernet “Lossless”

Storage requires “reliable” information delivery. Reliable delivery consists of two aspects, the transmission Bit Error Rate (BER) and frame loss.

7.5.1 Transmission Reliability (Bit Error Rate)

Many Ethernet physical links provide bit error rates comparable to Fiber Channel. The Bit Error Rate (BER) objective for both 1 Gb and 10 Gb Ethernet is the same objective as for Fiber Channel (10^{-12}).

Some Ethernet links may have higher bit error rates and they are not be suitable for FCoE traffic. This may occur because the Ethernet cable plant may be more variable than a Fiber Channel cable plant or Ethernet frames may be sent vial links that inherently have a higher bit error rate (such a wireless links).

The bit error rate of the links need to be taken into consideration for FCoE planning to ensure that the required level of transmission reliability is provided.

7.5.2 Fiber Channel Flow Control

Fiber Channel uses a “credit-based” flow control method. Credit is permission given by a receiver to a sender giving the sender permission to send a specified number of frames. The amount of credit given is a reflection of the buffers that are available to receive frames.

When a frame is sent, the available credit is decremented (a receiver’s buffer has been used). When the frame has been processed, and the recipient is ready for another frame, a credit reply is sent to replenish the credit. As long as a sender has credit available, it may send another frame (which of course causes the available credit to be decremented). A model of Fiber Channel’s credit-based flow control is shown in Figure 7-15.

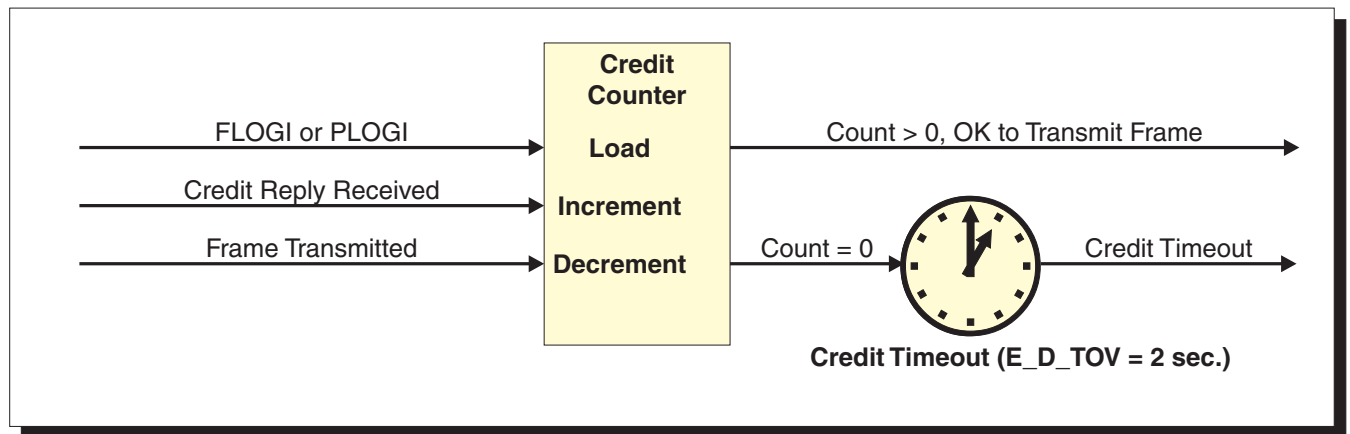


Figure 7-15. Credit-Based Flow Control

Fiber Channel provides two levels of flow control, a link-level mechanism called Buffer-to-Buffer flow control and a source to destination mechanism called End-to-End flow control. Both are based on a credit mechanism. The scope of each method is shown in Figure 7-16 on page 106.

Buffer-to-Buffer flow control controls the flow of frames on an individual link. Every Fiber Channel link is subject to link-level flow control. Buffer-to-Buffer credit is established using login parameters during Fabric Login (FLOGI) in a fabric environment and N_Port Login (PLOGI) is a

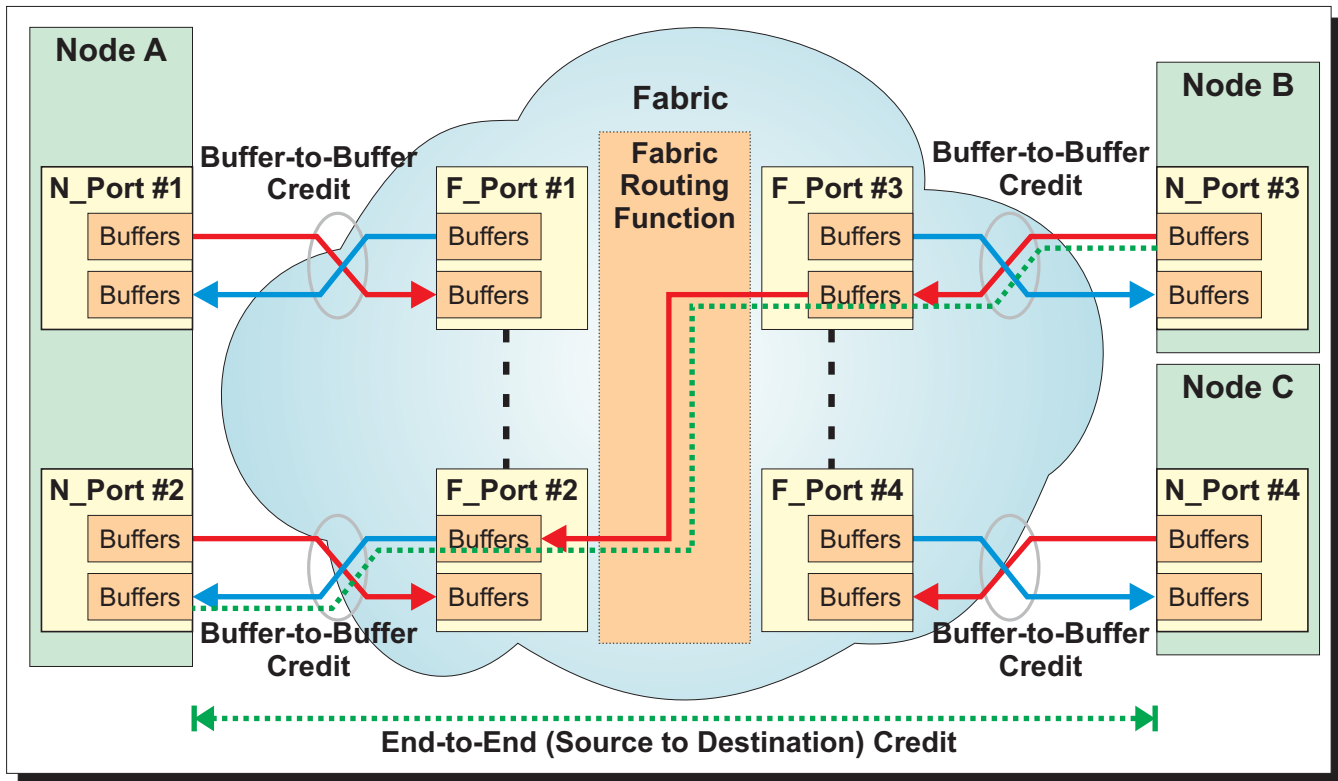


Figure 7-16. Fiber Channel Flow Control Models

point-to-point environment. The response that replenishes Buffer-to-Buffer credit is the Receiver Ready (R_RDY) Ordered Set.

End-to-End credit manages the flow of frames between a given source and destination port pair and is only used by some Fiber Channel classes of service (consequently, it may not be used in all application environments). End-to-End credit is replenished by Fiber Channel Link Control frames such as ACK and BSY.

7.5.3 Frame Loss and Ethernet Flow Control

Ethernet defines an optional “pause” based flow control described in IEEE 802.3 Annex 31B. In the pause flow control, the receiver tells the sender when to pause or resume frame transmission (done in hardware, not software). The receiver must send the pause while there is enough buffer space to accommodate frames in transit plus the time for the pause frame to be received and processed. An example of this method is shown in Figure 7-17 on page 107

While pause is part of the Ethernet standard, it is an optional feature and may not be implemented by all devices. This function, or an equivalent or enhanced flow control function is required by FCoE to prevent frame loss due to buffer overrun conditions.

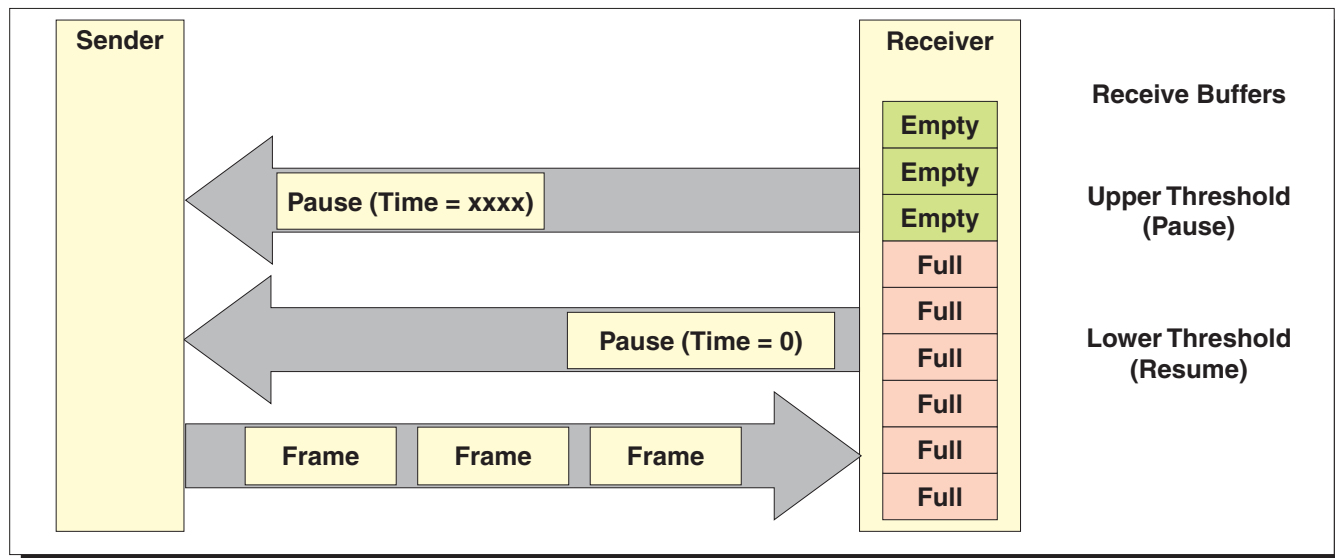


Figure 7-17. Ethernet Pause Flow Control

7.5.4 Pause Frame Format

Pause is a MAC Control frame that is created and processed by the Ethernet MAC layer, and not the software driver. MAC Control frames are identified by an EtherType value of 8808h. The format of the pause frame is shown in Figure 7-18.

The Pause frame uses a MAC Control Op-Code of 0001h to identify this as a Pause.

The Destination Address (DA) is set to a specified group address to prevent the frame from being forwarded beyond this physical link.

The Source Address is set to the NIC card's unicast address.

The Pause function has a single parameter, the `pause_time`. The `Pause_time` is specified as 512-bit increments on the associated physical link. This provides a `Pause_time` range of 0 to 33.6 msec. on a 1 gigabit link. A `Pause_time` value of zero means resume transmission.

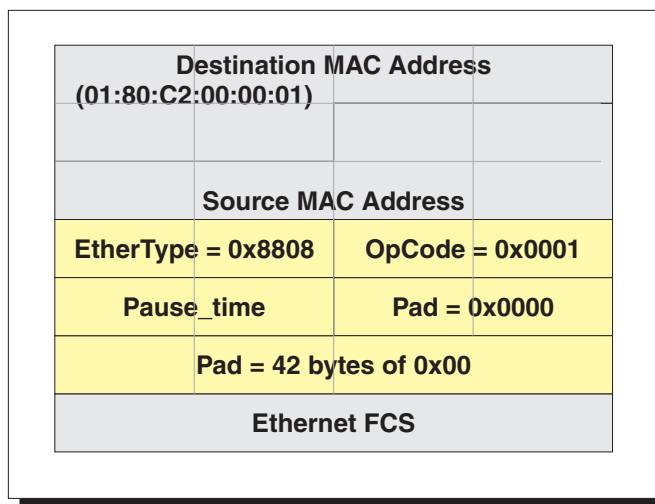


Figure 7-18. Pause Frame Format

7.6 Link Aggregation (NIC Teaming)

Link aggregation is an optional capability that enables multiple Ethernet ports (MACs) to be "aggregated" and treated as if they were a single, higher-speed port. Link aggregation was defined by the 802.3ad task force and standardized in clause 43 of IEEE 802.3 (see reference 29

in the Bibliography on page 290). There are also many proprietary implementations of link aggregation that go by a variety of names.

NOTE – Link aggregation is also known as: NIC Teaming, Ethernet trunking, port teaming, “EtherChannel”, “Multi-Link Trunking (MLT)”, “NIC bonding”, “Network Fault Tolerance (NFT)” and “link aggregate group” (LAG).

Figure 7-19 contains a block diagram of the functions associated with link aggregation. These functions may be implemented in the software driver, or by hardware or firmware associated with an adapter or switch.

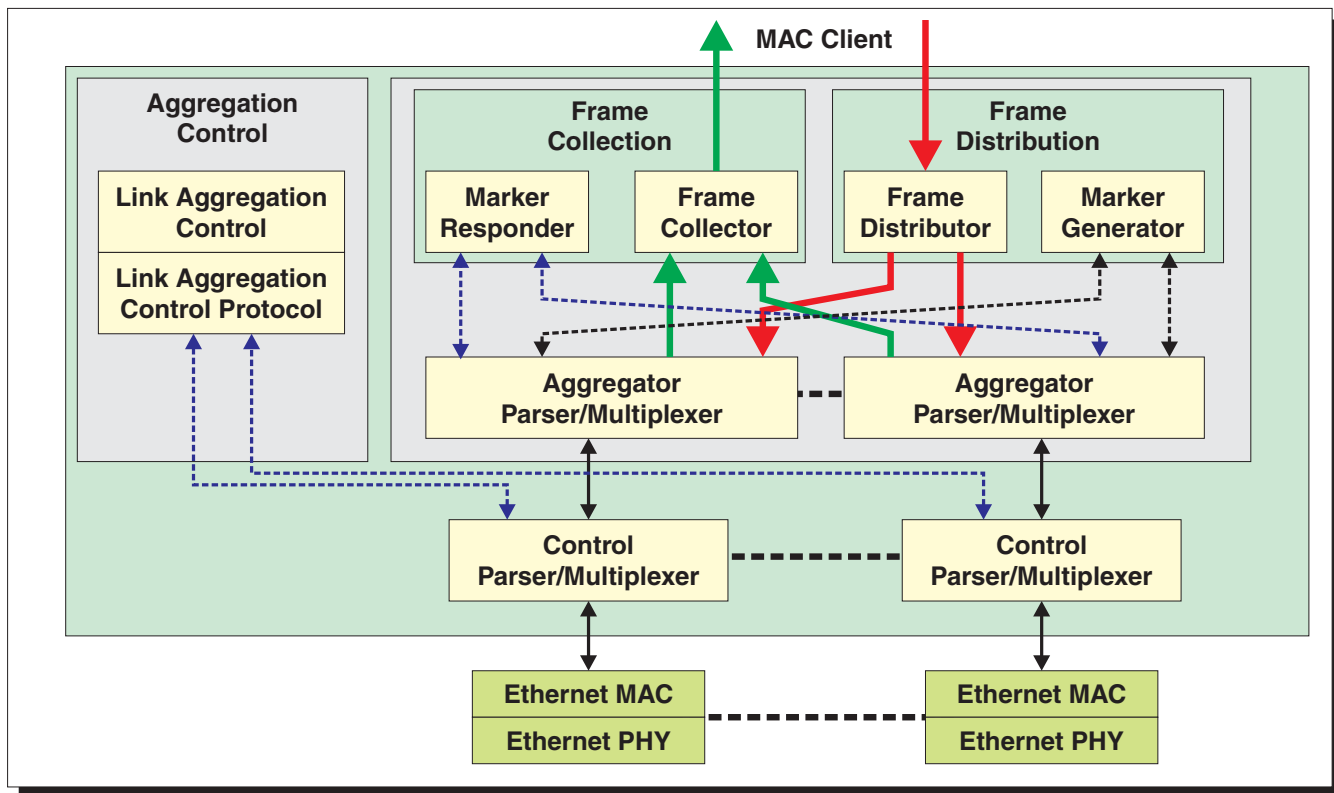


Figure 7-19. Ethernet Link Aggregation

Link aggregation is transparent to the MAC client and appears as a normal MAC function to the client. Each Ethernet MAC has a normal MAC address that is used as the source address for transmitted unicast frames and the destination address of received unicast frames. The Aggregation function is also assigned an Ethernet MAC address which is the address seen by the MAC client (this MAC address may be a unique MAC address, or the address of one of the aggregated MACs). The MAC client does not directly see the MAC addresses of the individual Ethernet MACs. MACs that are to be aggregated must operate at the same speed and need to support full-duplex operation.

Frames to be transmitted are sent by the MAC client to the frame distribution function. The frame distribution function distributes frames to the appropriate Ethernet MAC. To provide in-order frame delivery, frames associated with a given “conversation” are distributed to a specif-

ic Ethernet MAC. Frames associated with other conversations may be distributed to other MACs associated with the aggregation function. The standard does not define the algorithm to be used by the distributor, it only requires that the distributor operate in a manner to provide in-order frame delivery and prevent frame duplication. With proper attention to the in-order delivery requirements, conversations may be moved from one MAC to another within the aggregation group to provide load balancing or rerouting of traffic around failed links or MACs.

NOTE – If dissimilar NICs are aggregated (e.g., one provides TCP off load and another doesn't) performance may vary depending on which NIC is being used for a particular conversation.

NOTE – Link aggregation may apply to LAN traffic on a Converged Network Adapter, but may not apply to storage traffic using the FCoE protocol. This is due to the fact that most CNAs provide separate driver interfaces for LAN traffic and FCoE traffic.

When a frame is received by one of the MACs, it is forwarded to the frame collector for delivery to the MAC client. While frames from different conversations may be interleaved by the collector, frames within a conversation are delivered to the MAC client in order.

In 802.3ad, the Link Aggregation Protocol is used to automatically inform the switch of the ports that are to be aggregated. This is done between an end device and a switch, or between two switches using Link Aggregation Protocol Data Units sent to MAC address 01-80-C2-00-00-02, one of the addresses not forwarded by a switch (see Table 7-1 on page 92). The Link Aggregation Protocol can alleviate the need for manual configuration of the devices.