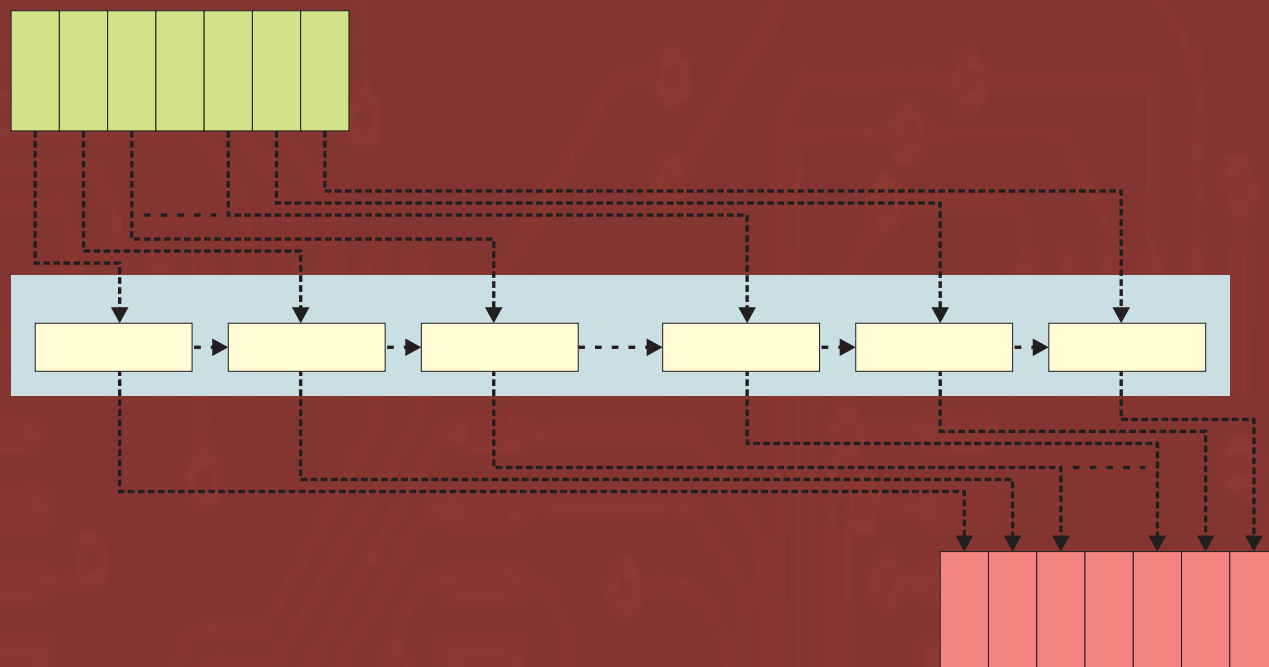


The Fibre Channel Consultant Series

Fibre Channel over Ethernet



Robert W. Kembel

Copyright © 2008 by Robert W. Kembel

All rights reserved. Except for brief passages to be published in a review or as citation of authority, no part of this book may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without prior written permission from the publisher.

If trademarks or tradenames of any companies or products have been used within this book, no such uses are intended to convey endorsement or other affiliations with the book. Any brand names or products used within this book are trademarks or registered trademarks of their respective holders.

Though the author and publisher have made every attempt to ensure the accuracy and completeness of the information contained in this book, they assume no responsibility for errors, inaccuracies, omissions, or any inconsistency therein. The reader is strongly advised to refer to the appropriate standards documents before beginning any design activities.

Disclaimer:

This is a pre-publication edition of the book and has not been spell-checked, proofread or checked for accuracy or completeness and may be incomplete, incorrect, or just downright wrong. At the time of writing, the FCoE standard had not been approved by INCITS Technical Committee T11 and information contained in this publication should be considered preliminary.

ISBN 978-0-931836-75-6

Published by:

**Northwest Learning Associates, Inc.
3061 N. Willow Creek Drive
Tucson, AZ 85712
520-881-0877, Fax: 520-881-0632
email: info@nlabooks.com
Visit our web site at www.NLAbooks.com**

Printed in the United States of America

First edition: 10 9 8 7 6 5 4 3 2 1

Contents

List of Figures	vii
List of Tables	xiii
Foreword	xv
Preface	xvii
Acknowledgments	xviii
Section I. Introduction and Concepts	
1. FCoE Introduction and Concepts	1
1.1 The “Converged” Data Center Network	2
1.2 Fibre Channel and FCoE Roadmap	3
1.3 Benefits of a Converged Network Approach	4
1.4 What is Fibre Channel over Ethernet (FCoE)?	4
1.5 What’s Different About FCoE?	9
1.6 Benefits of FCoE	13
1.7 Software FCoE vs. Hardware FCoE	14
1.8 Chapter Summary	16
2. FCoE, Fibre Channel and Ethernet Standards	19
2.1 FCoE Development Timeline	19
2.2 FCoE and Fibre Channel Standards	19
2.3 SCSI Standards	25
2.4 Ethernet Standards	26
2.5 Internet Engineering Task Force (IETF) Standards	26
2.6 Standards Summary	26
2.7 Chapter Summary	28
3. Deployment Scenarios	31
3.1 FCoE Point-to-Point Evaluation Scenario	31
3.2 FCoE Gateway Evaluation Scenario	32
3.3 Standalone FCoE Evaluation Scenario	32
3.4 FCoE Blade Server Deployment	33
3.5 FCoE Rack Server Deployment	35
3.6 Deploying FCoE via a Fibre Channel Switch Line Card	36
3.7 Deploying FCoE via an FCoE Forwarder	37
3.8 FCoE via Enhanced Ethernet “Transit” Switches	38
3.9 FCoE Servers and Virtual Machines	40
3.10 Chapter Summary	41
4. FCoE Configurations	43
4.1 Ethernet Point-to-Point	43
4.2 FCoE to Native Fibre Channel Point-to-Point	44
4.3 FCoE Devices Connected to FCoE Forwarders	45

4.4	FCoE ENode to FCF via Intervening Ethernet Network	46
4.5	Hybrid Ethernet and FCoE Fabrics	47
4.6	Hybrid Configuration with Intervening Enhanced Ethernet Switches	49
4.7	FCoE Concentrator Configuration	50
4.8	Fibre Channel and Ethernet Forwarding Protocols	52
4.9	Virtual Machine Environments	56
4.10	Switch Forwarding Strategies	63
4.11	Direct VN_Port to VN_Port Communication?	66
4.12	Chapter Summary	68
5.	Looking at the Benefits	75
5.1	Benefits of a Converged Network	75
5.2	Blade Server Example Case Study	78
5.3	Chapter Summary	81
 Section II. FCoE Technology		
6.	FCoE Technology	85
6.1	FCoE and Enhanced Ethernet	85
6.2	FCoE Frame Encapsulation	85
6.3	Architectural Models and Configurations	85
6.4	MAC Addressing	85
6.5	Discovery and Virtual Link Initialization	86
6.6	Access Control	86
6.7	Error Scenarios and Handling	86
6.8	Chapter Summary	87
7.	Ethernet Essentials	89
7.1	Ethernet Frame Format	90
7.2	Ethernet Topologies	93
7.3	Ethernet Physical Link Variants	98
7.4	Virtual LANs (VLANs)	103
7.5	Making Ethernet “Lossless”	104
7.6	Link Aggregation (NIC Teaming)	107
7.7	Chapter Summary	110
8.	Architecture Models	117
8.1	FCoE Virtual Links	117
8.2	FCoE Link Endpoint (FCoE_LEP)	118
8.3	Node and Node Port Models	119
8.4	Fabric and Fabric Port Models	122
8.5	Architecture Models Conclusion	128
8.6	Chapter Summary	129
9.	FCoE Addressing	133
9.1	Fibre Channel and Ethernet Addresses	133
9.2	MAC Addressing Methods	133
9.3	Server Provided MAC Addresses (SPMA)	134
9.4	Fabric Provided MAC Addresses (FPMA)	136

9.5	Comparison of SPMA and FPMA	137
9.6	MAC Addressing Summary	139
9.7	Chapter Summary	140
10.	FCoE Discovery and Virtual Link Initialization	143
10.1	FCoE Discovery	143
10.2	ENode Virtual Link Initialization	145
10.3	FCF Interswitch Link (ISL) Discovery and Initialization	153
10.4	Chapter Summary	156
11.	FCoE Initialization Protocol (FIP)	159
11.1	FCoE Initialization Protocol (FIP)	159
11.2	FIP Discovery Solicitation From an ENode	167
11.3	FIP Discovery Solicitation From FCF	168
11.4	FIP Discovery Advertisement	169
11.5	FIP FLOGI Request and Reply	171
11.6	FIP NPIV FDISC Request and Reply	173
11.7	FIP Fabric LOGO Request and Reply	175
11.8	FIP ELP Request and Reply	178
11.9	FIP Clear Virtual Link	180
11.10	FIP Keep Alive (FKA)	181
11.11	FIP Protocol Timers	181
11.12	FCoE Initialization Protocol Error Processing	182
11.13	Chapter Summary	184
12.	FCoE Encapsulation	189
12.1	Start-of-Frame and End-of-Frame	190
12.2	Determining the Ethernet Destination Address (DA)	190
12.3	Fibre Channel and Ethernet Frame Sizes	192
12.4	Ethernet and Fibre Channel Comparison	192
12.5	Encapsulation/Decapsulation Flow	195
12.6	Chapter Summary	203
13.	Access Control	207
13.1	Fibre Channel Zoning	207
13.2	Ethernet Access Control Lists (ACLs) and Static Forwarding Tables	212
13.3	Chapter Summary	216
14.	FCoE Error Conditions	219
14.1	Ethernet Duplicate MAC Address Errors	219
14.2	Ethernet Rogue Host Attacks	222
14.3	Ethernet Forwarding Errors	224
14.4	Ethernet and Fiber Channel's R_A_TOV	226
14.5	Ethernet Link Down Detection	232
14.6	Chapter Summary	234
15.	Summary of FCoE Changes	237
15.1	Physical Link (FC-0)	237
15.2	Fibre Channel FC-1 Differences	237
15.3	Fibre Channel FC-2 Differences	238

15.4	Fibre Channel FC-3 Differences	238
15.5	Fibre Channel FC-4 Differences	238
15.6	Upper-Level Protocol (ULP) Differences	239
15.7	Link Services (ELS and SW_ILS) Differences	239
15.8	Chapter Summary	240
16.	Data Center Ethernet	241
16.1	Per-Priority Pause Flow Control (802.1Qbb)	241
16.2	Priority and Bandwidth Management	242
16.3	Ethernet Enhanced Transmission Selection (802.1Qaz)	247
16.4	Ethernet Congestion Notification (802.1Qau)	248
16.5	Link-Level Discovery Protocol (LLDP)	256
16.6	Data Center Bridge Capability Exchange Protocol (DCBX)	259
16.7	Chapter Summary	269
 Section III. Reference Information		
III.	Glossary	275
	Bibliography	289
	Index	293

List of Figures

Figure 1-1. Server with Multiple Interfaces	1
Figure 1-2. Ethernet As A “Fat Pipe”.	2
Figure 1-3. Server with Converged Network.	3
Figure 1-4. Fibre Channel and FCoE Roadmap	4
Figure 1-5. Pure Ethernet FCoE Configuration.	5
Figure 1-6. Protocol Stack Comparison	10
Figure 1-7. Fibre Channel over IP (FCIP) Frame Format	11
Figure 1-8. FCoE Encapsulation Concept	12
Figure 2-1. FCoE Timeline	20
Figure 2-2. Fibre Channel Standards	21
Figure 2-3. SCSI Standards Structure	26
Figure 2-4. Ethernet Standards.	27
Figure 3-1. Standalone FCoE Evaluation Scenario	31
Figure 3-2. FCoE Evaluation Gateway Scenario	32
Figure 3-3. FCoE Switch Evaluation Scenario	33
Figure 3-4. FCoE Blade Server Scenario	34
Figure 3-5. FCoE Rack Server Scenario	35
Figure 3-6. Deploying FCoE via a Fibre Channel Switch Line Card	37
Figure 3-7. FCoE via FCoE Forwarder (FCF).	38
Figure 3-8. FCoE via Enhanced Ethernet Switches Scenario.	39
Figure 3-9. FCoE and Server with Virtual Machines.	40
Figure 4-1. FCoE Point-to-Point Configuration	43
Figure 4-2. FCoE to FC Point-to-Point Configuration	44
Figure 4-3. FCoE Forwarder (FCF) Topology.	45
Figure 4-4. Intervening Enhanced Ethernet Network	46
Figure 4-5. Hybrid FCoE Configuration.	47
Figure 4-6. Hybrid FCoE Configuration with Intervening Ethernet Switches.	49
Figure 4-7. FCoE Concentrator.	50
Figure 4-8. FCoE Forwarding: No Ethernet Switching	52
Figure 4-9. FCoE Forwarding: Non-Overlapping Ethernet networks	53
Figure 4-10.FCoE Forwarding: Overlapping Ethernet networks	54
Figure 4-11.FCoE Forwarding: Ethernet Link Aggregation	55

Figure 4-12. Para Virtualization (e.g., XEN)	56
Figure 4-13. Para Virtualization (e.g., XEN) with FCoE	57
Figure 4-14. Para Virtualization (e.g., XEN) with Software FCoE	57
Figure 4-15. Hardware Virtualization (e.g., VMware) Model	58
Figure 4-16. Hardware Virtualization (e.g., VMware) with FCoE	59
Figure 4-17. Hardware Virtualization (e.g., VMware) with Software FCoE	60
Figure 4-18. Moving a Virtual Machine (VMotion)	61
Figure 4-19. Moving a Virtual Machine with Fibre Channel HBAs (VMotion).	62
Figure 4-20. Moving a Virtual Machine with FCoE HBAs (VMotion)	64
Figure 4-21. Cut-Through Forwarding	64
Figure 4-22. Store and Forward Architecture.	65
Figure 4-23. Store and Forward “Pipelining”	65
Figure 4-24. Direct VN_Port to VN_Port Forwarding	66
Figure 5-1. FCoE Blade Server Configuration	78
Figure 7-1. Ethernet and the OSI Reference Model.	89
Figure 7-2. Ethernet Frame Format	91
Figure 7-3. Ethernet MAC Address Format	92
Figure 7-4. Ethernet Shared Medium Topology	94
Figure 7-5. Ethernet Switched Topology	95
Figure 7-6. Ethernet Switch Learning Database	96
Figure 7-7. Ethernet Spanning Tree.	97
Figure 7-8. Ethernet Spanning Tree (Redrawn).	98
Figure 7-9. XENPAK Transceiver Module	101
Figure 7-10. XPAK Transceiver Module.	101
Figure 7-11. X2 Transceiver Module	102
Figure 7-12. XFP Transceiver Module	102
Figure 7-13. SFP+ Transceiver Module.	103
Figure 7-14. Ethernet Frame with 802.1Q VLAN Tag	104
Figure 7-15. Credit-Based Flow Control.	105
Figure 7-16. Fiber Channel Flow Control Models	106
Figure 7-18. Pause Frame Format.	107
Figure 7-17. Ethernet Pause Flow Control.	107
Figure 7-19. Ethernet Link Aggregation	108
Figure 8-1. FCoE Virtual Links	117
Figure 8-2. FCoE Link End Point (FCoE_LEP)	118
Figure 8-3. Fibre Channel Node and Node Port Model	119

Figure 8-4. FCoE Node with Two FCoE_LEPs VN_Ports.	121
Figure 8-5. Fibre Channel VN_Ports with NPIV	123
Figure 8-6. FCoE VN_Port with NPIV	124
Figure 8-7. Fibre Channel Fabric Model	125
Figure 8-8. Fibre Channel Switch Model.	125
Figure 8-9. FCoE VF_Port and VE_Port Model	126
Figure 9-1. Multiple Protocols Sharing the Same NIC	133
Figure 9-2. Server Provided (Single) MAC Address	134
Figure 9-3. Server Provided (Multiple) MAC Addresses	135
Figure 9-4. Ethernet Address Lookup for SPMA Addressing	136
Figure 9-5. Fabric Provided MAC Addresses	137
Figure 9-6. Fabric Provided MAC Addresses	138
Figure 10-1.FCoE Operational Phases	143
Figure 10-2.FCoE Discovery	144
Figure 10-3.Enode Discovery and Virtual Link Initialization Steps (SPMA).	146
Figure 10-4.Enode Discovery and Virtual Link Initialization Steps (SPMA with NPIV)	147
Figure 10-5.Enode Discovery and Virtual Link Initialization Steps (SPMA Case 2).	148
Figure 10-6.Enode Discovery and Virtual Link Initialization Steps (SPMA with NPIV)	149
Figure 10-7.Enode Discovery and Virtual Link Initialization Steps (FPMA)	150
Figure 10-8.Enode Discovery and Virtual Link Initialization Steps (FPMA with NPIV)	151
Figure 10-9.Direct VN_Port to VN_Port Virtual Link Initialization	152
Figure 10-10.Relinquishing a VN_Port ID with LOGO.	153
Figure 10-11.FCoE FCF Interswitch Link (ISL) Discovery.	154
Figure 10-12.FCF Interswitch Link (ISL) Discovery and Virtual Link Initialization	155
Figure 11-1.FCoE Operational Phases	159
Figure 11-2.FCoE Initialization Protocol (FIP) General Frame Format	160
Figure 11-3.ENode FIP Discovery Solicitation Frame Format.	167
Figure 11-4.FCF FIP Discovery Solicitation Frame Format.	168
Figure 11-5.FIP Discovery Advertisement Frame Format	170
Figure 11-6.FIP FLOGI Request and LS_ACC Frame Format	171
Figure 11-7.FIP FLOGI LS_RJT Frame Format	172
Figure 11-8.FIP NPIV FDISC Request and LS_ACC Frame Format.	173
Figure 11-9.FIP NPIV FDISC LS_RJT Frame Format.	174
Figure 11-10.FIP Fabric LOGO Request Frame Format	175
Figure 11-11.FIP Fabric LOGO LS_ACC Reply Frame Format.	176
Figure 11-12.FIP Fabric LOGO LS_RJT Reply Frame Format	177

Figure 11-13.FIP ELP Request and SW_ACC Frame Format	178
Figure 11-14.FIP ELP SW_RJT Reply Frame Format	179
Figure 11-15.FIP Clear Virtual Link Frame Format.	180
Figure 11-16.FIP Keep Alive Frame Format.	181
Figure 11-17.FIP Error and Ethernet Switches	183
Figure 12-1.FCoE Frame Encapsulation.	189
Figure 12-2.Encapsulated Frame Format.	190
Figure 12-3.FCoE MAC “Mapped” Addressing Example	191
Figure 12-4.Ethernet Framing Efficiency.	194
Figure 12-5.Fibre Channel Framing Efficiency	194
Figure 12-7.FCoE Framing Efficiency (“Baby Jumbo” Ethernet Frame)	195
Figure 12-6.FCoE Framing Efficiency (Standard Ethernet Frame).	195
Figure 12-8.FCoE Link End Point (FCoE_LEP)	196
Figure 12-9.ENode Frame Encapsulation Flow	197
Figure 12-10.ENode Frame Decapsulation Flow	201
Figure 13-1.Fibre Channel Zoning	207
Figure 13-2.Basic Zoning Data Model and Structure	208
Figure 13-3.Fibre Channel Soft Zoning.	210
Figure 13-4.Fibre Channel Hard Zoning	211
Figure 13-5.FCoE Default Access Control List (ACL) Example	213
Figure 13-6Automatic Installation of ACL Entries at FLOGI.	214
Figure 14-1.Cross-Connect Error Due to Duplicate MAC Addresses	220
Figure 14-2.Forwarding Loop Due to Duplicate MAC Addresses	221
Figure 14-3.Learning Attack by Rogue Host on VN_Port MAC	222
Figure 14-4.Learning Attack by Rogue Host on FCF-MAC.	224
Figure 14-5.Duplicate Ethernet Frame During Reconfiguration	225
Figure 14-6.Duplicate Ethernet Frame During Reconfiguration	226
Figure 14-7.R_A_TOV and Maximum Fabric Transit Time.	227
Figure 14-8.Read Data Corruption Due to Late Frames.	228
Figure 14-9.Write Error Due to Late Frames.	229
Figure 14-10.R_A_TOV Enforcement.	230
Figure 14-12.FCoE Forwarding Delay Accumulation	231
Figure 14-11.FCoE R_A_TOV Enforcement Problem	231
Figure 14-13.Ethernet Link Down Condition	232
Figure 14-14.Ethernet Link Down Signaling	233
Figure 16-1.Ethernet Per-Priority Pause Flow Control	241

Figure 16-2.Per-Priority Pause Frame Format.	242
Figure 16-3.Priority and Bandwidth Limiting Flow Diagram (Example)	246
Figure 16-4.Enhanced Transmission Selection Mapping	249
Figure 16-5.Congestion Spreading	250
Figure 16-6.Backward Congestion Notification (BCN).	251
Figure 16-7.Congestion Point Queue Model	252
Figure 16-8.Reaction Point (RP) Rate Limiting	253
Figure 16-9.Congestion Notification Domain	255
Figure 16-10.BCN Frame Format (Proposed)	256
Figure 16-11.Rate Limited Tag (RLT) Format (Proposed).	257
Figure 16-12.LLDP Frame Format.	258
Figure 16-13.LLDP Parameter Transfer Miss at Link Up.	259
Figure 16-14.LLDP Accelerated Initial Parameter Transfer Proposal	260
Figure 16-15.DCBX Control TLV	261
Figure 16-16.DCBX Protocol Flow Example	262
Figure 16-17.DCBX Feature Sub-TLV Format.	263
Figure 16-18.DCBX Priority Group Feature TLV Format.	264
Figure 16-19.DCBX Priority Flow Control Feature TLV Format.	265
Figure 16-20.DCBX BCN Feature TLV Format	266
Figure 16-21.FCoE Application Feature TLV Format	268
Figure 16-22.Logical Link Down Feature TLV Format.	268

List of Tables

Table 5-1. 40 Server Cost Estimate (Legacy Blade Server)	79
Table 5-2. 40 Server Cost Estimate (FCoE Blade Server)	80
Table 7-1. Ethernet Group MAC Addresses	92
Table 7-2. FCoE Group (Multicast) MAC Addresses	93
Table 7-3. Ethernet Physical Link Variants	99
Table 11-1. FIP Operation Codes and SubCodes	161
Table 11-3. FIP Priority Descriptor Format	162
Table 11-2. FIP Descriptors	162
Table 11-4. FIP MAC Address Descriptor Format	163
Table 11-5. FIP FC-MAP Descriptor Format	163
Table 11-6. FIP Name_Identifier Descriptor Format	163
Table 11-7. FIP Fabric_Name Descriptor Format	163
Table 11-8. FIP Maximum Receive Size Descriptor Format	164
Table 11-9. FIP FLOGI Descriptor Format	164
Table 11-10. FIP FDISC_NPIV Descriptor Format	164
Table 11-11. FIP LOGO Descriptor Format	165
Table 11-12. FIP Exchange Link Parameters (ELP) Descriptor Format	165
Table 11-14. FIP FKA_ADV_Period Descriptor Format	166
Table 11-13. FIP VN_Port Identifier Descriptor Format	166
Table 12-1. SOF and EOF Representation in Encapsulated Frame	191
Table 12-2. Ethernet and Fibre Channel Signaling Rate Comparison	193
Table 16-1. Bandwidth Group (BWG) Allocation Table Example	243
Table 16-2. Strict Priority Example	244
Table 16-4. Round-Robin With One Priority Queue	245
Table 16-3. Bandwidth Limiting (No Priority) Example	245
Table 16-5. Priority Group ID Bandwidth Limit Example	247
Table 16-6. LLDP TLV Type Values	258

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.

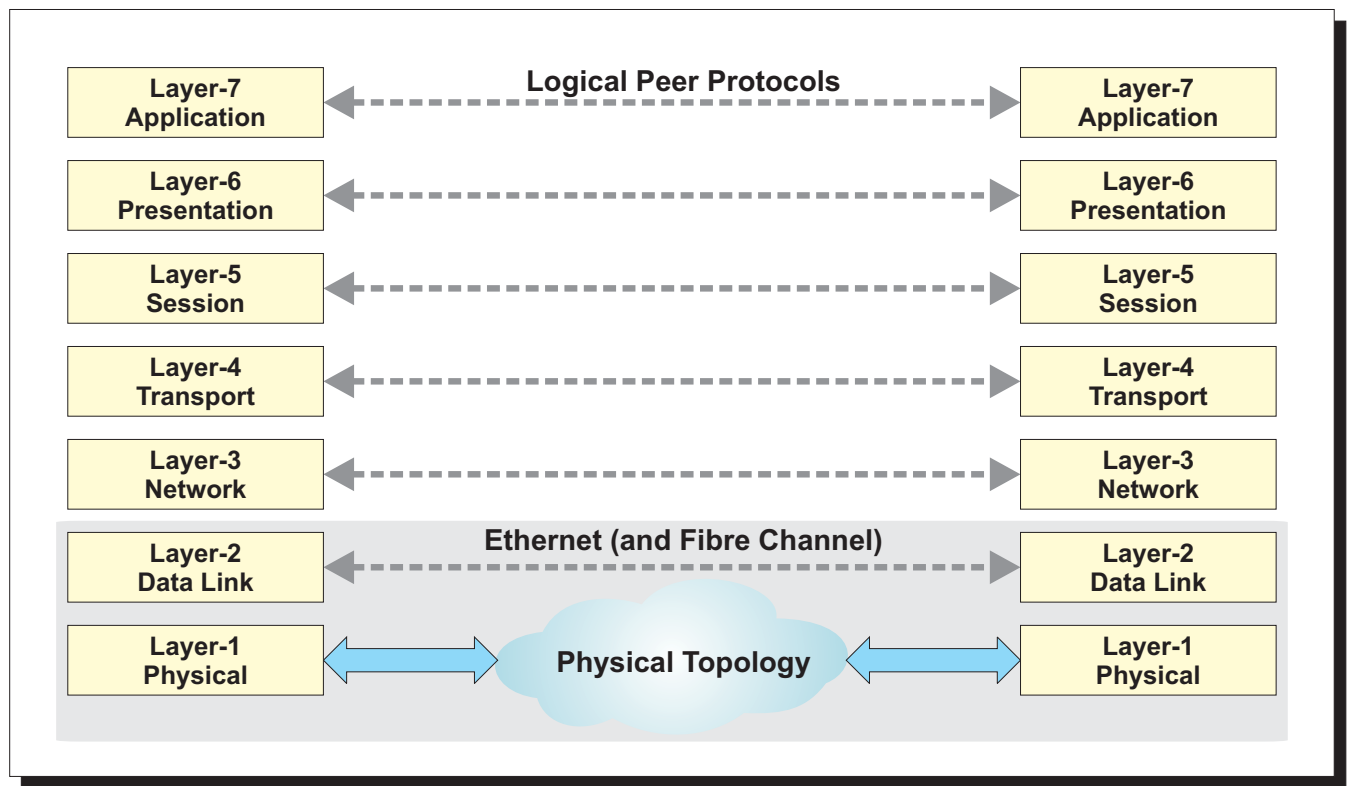


Figure 7-1. Ethernet and the OSI Reference Model

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.

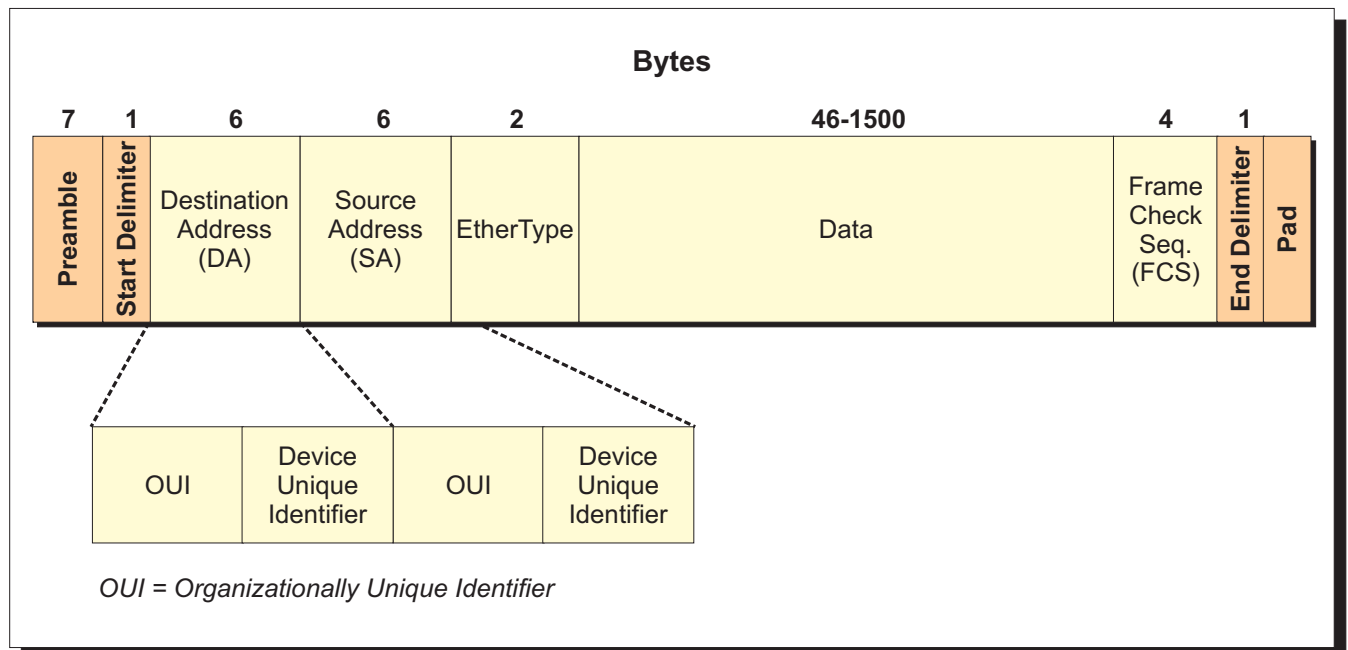


Figure 7-2. Ethernet Frame Format

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.

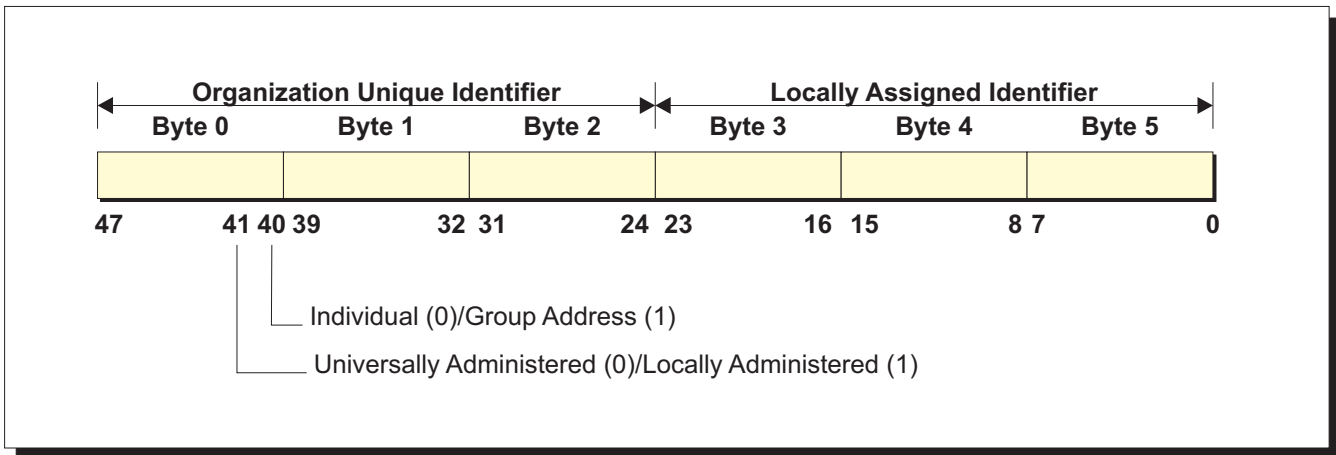


Figure 7-3. Ethernet MAC Address Format

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.

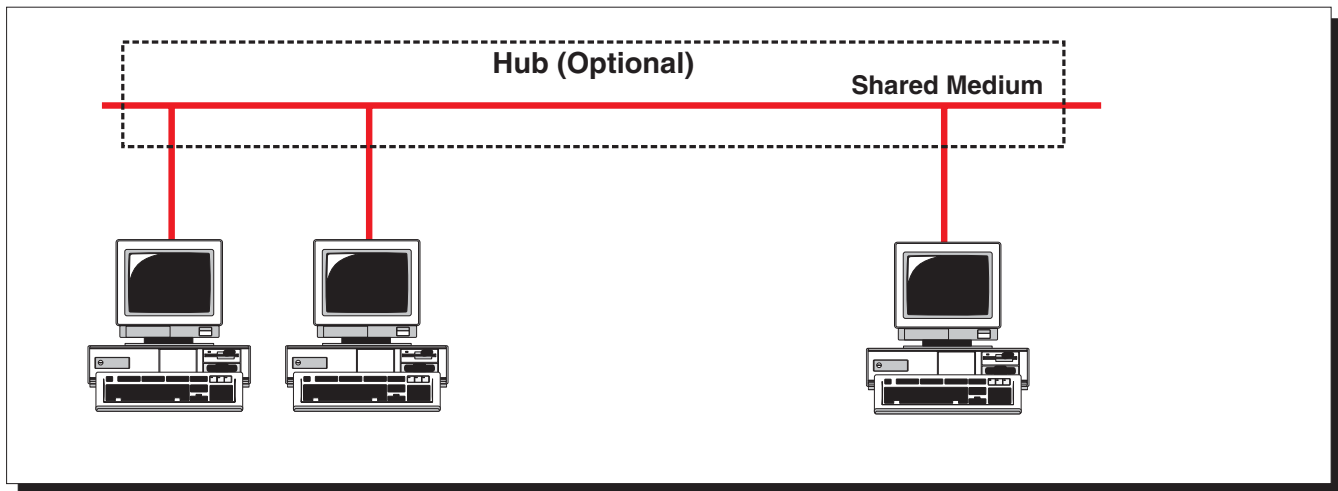


Figure 7-4. Ethernet Shared Medium Topology

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).

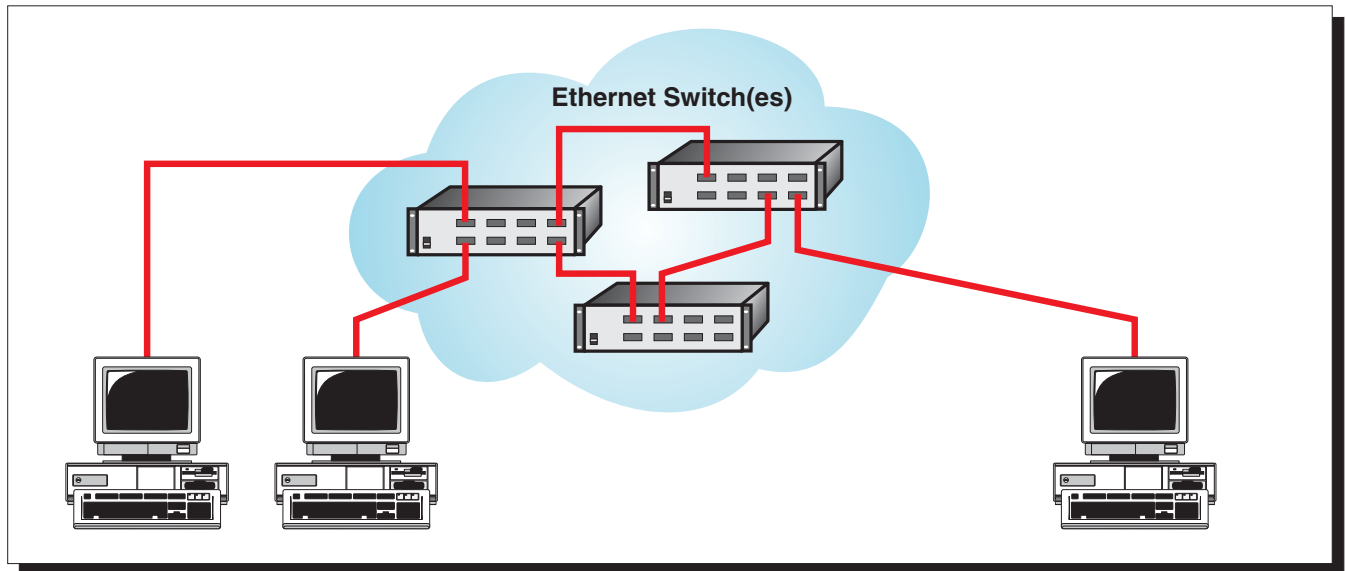


Figure 7-5. Ethernet Switched Topology

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.

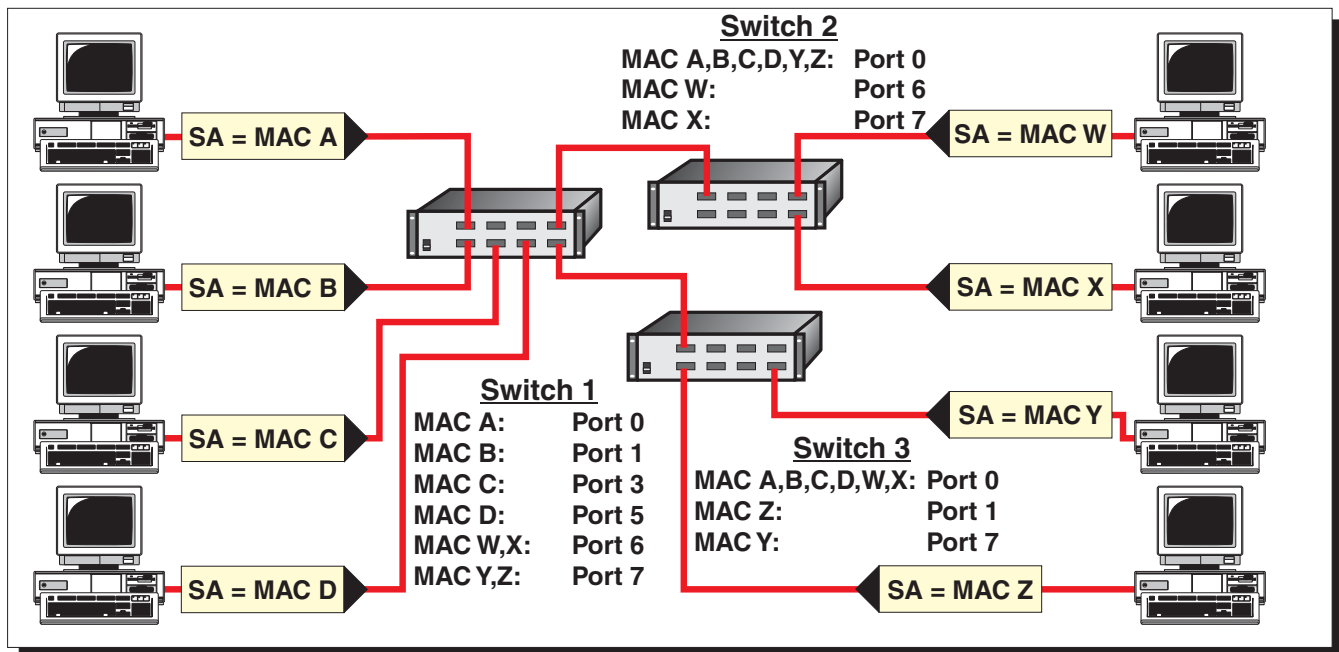


Figure 7-6. Ethernet Switch Learning Database

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).

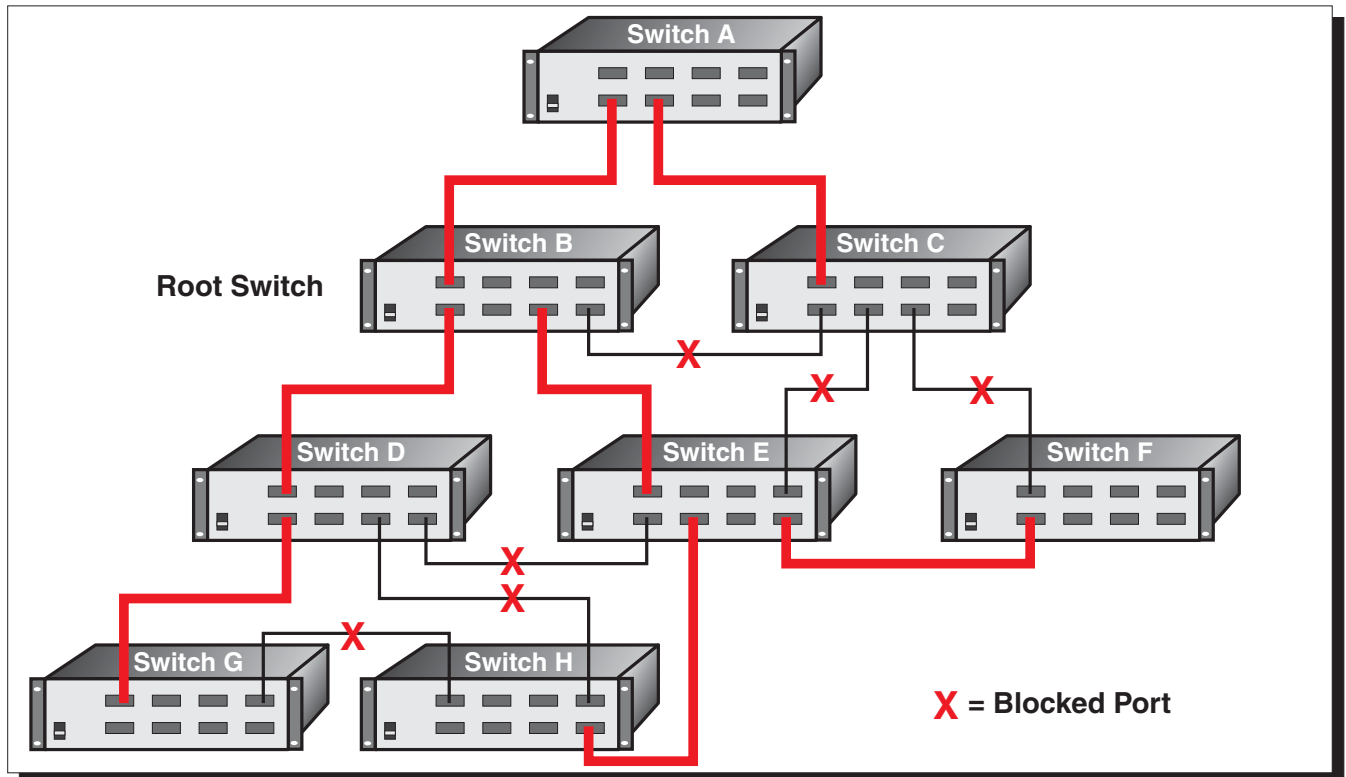


Figure 7-7. Ethernet Spanning Tree

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.

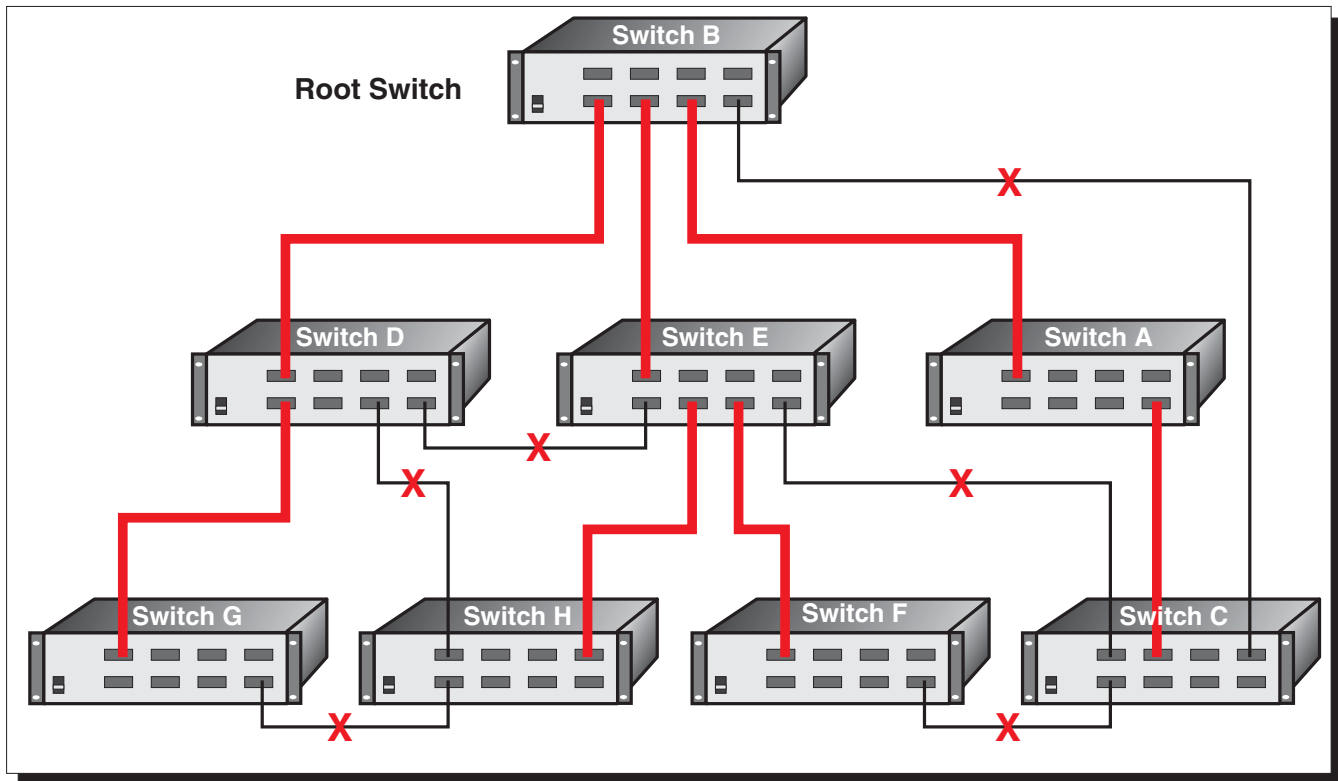


Figure 7-8. Ethernet Spanning Tree (Redrawn)

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

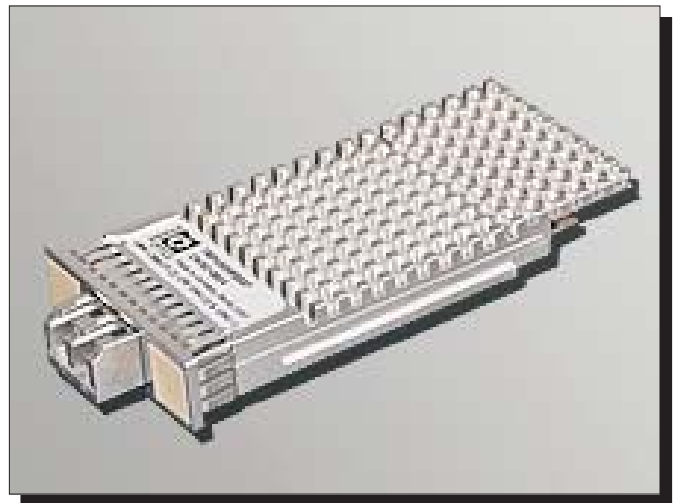


Figure 7-10. XPAK Transceiver Module

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.

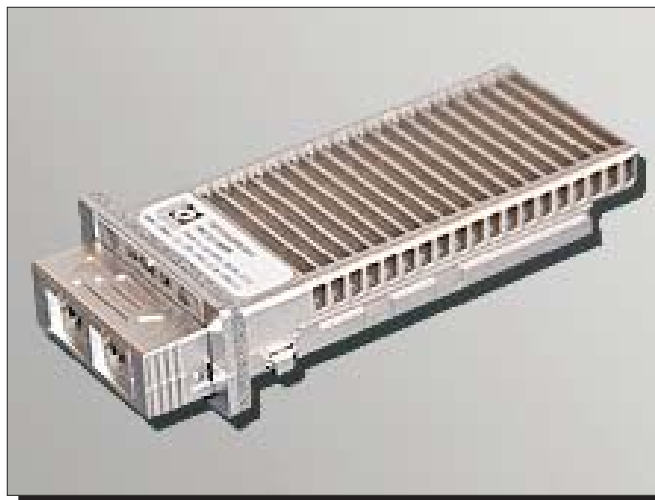


Figure 7-11. X2 Transceiver Module

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 7-12. 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.

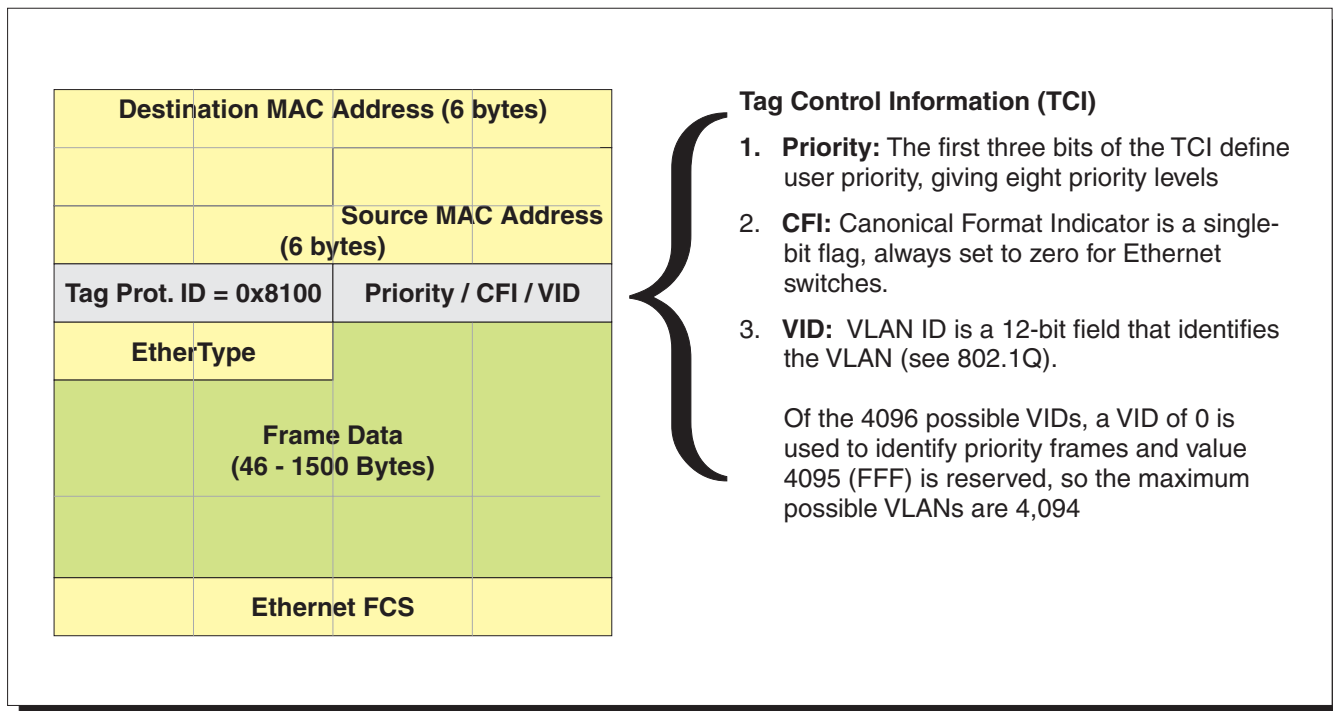


Figure 7-14. Ethernet Frame with 802.1Q VLAN Tag

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.

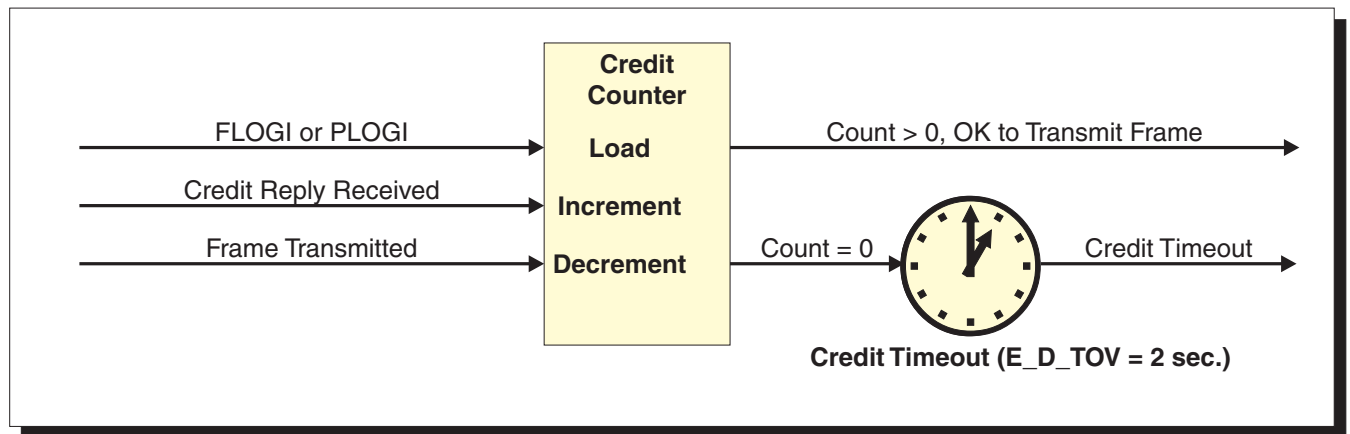


Figure 7-15. Credit-Based Flow Control

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

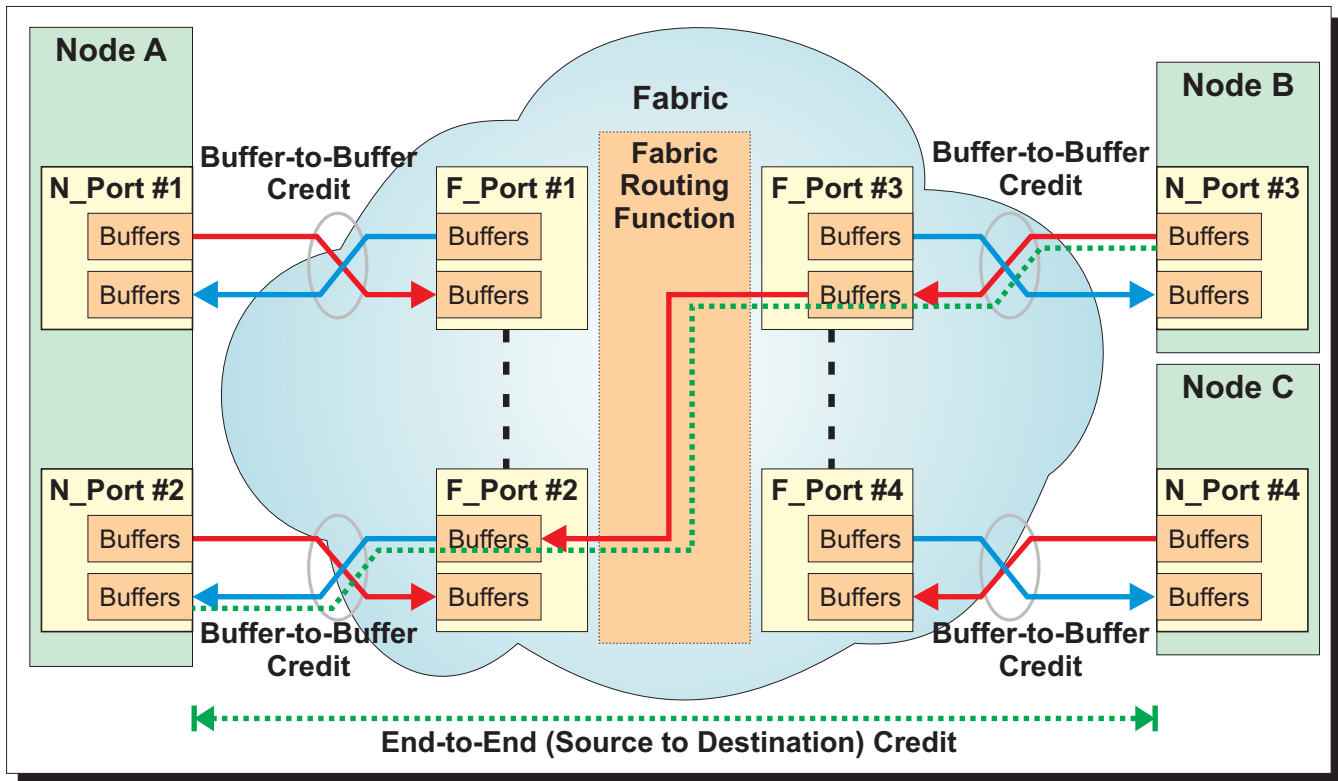


Figure 7-16. Fiber Channel Flow Control Models

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.

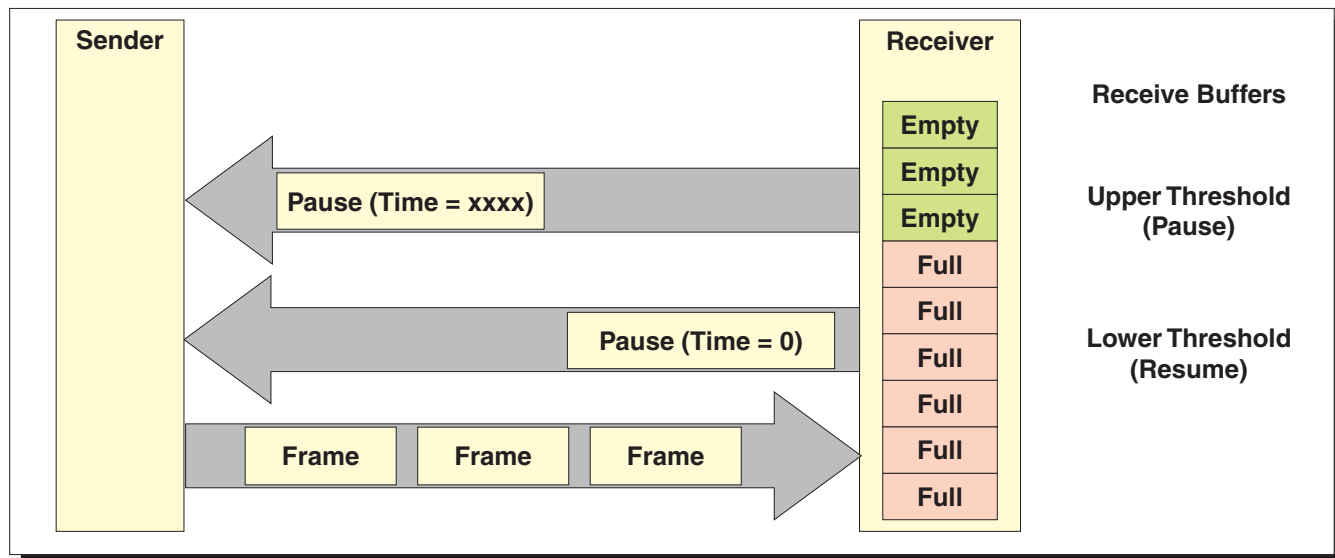


Figure 7-17. Ethernet Pause Flow Control

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.

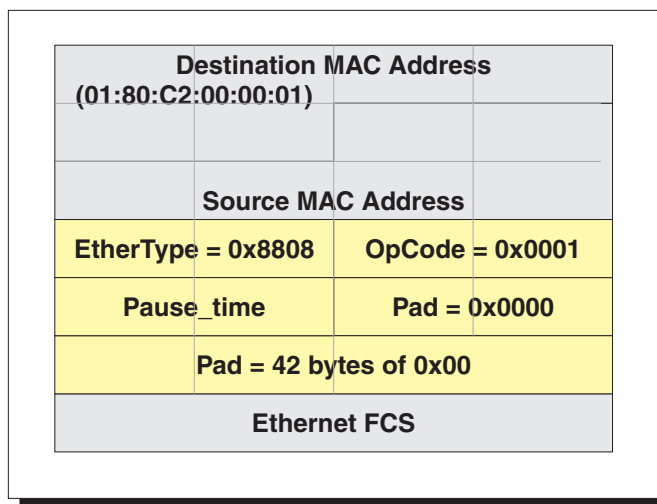


Figure 7-18. Pause Frame Format

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.

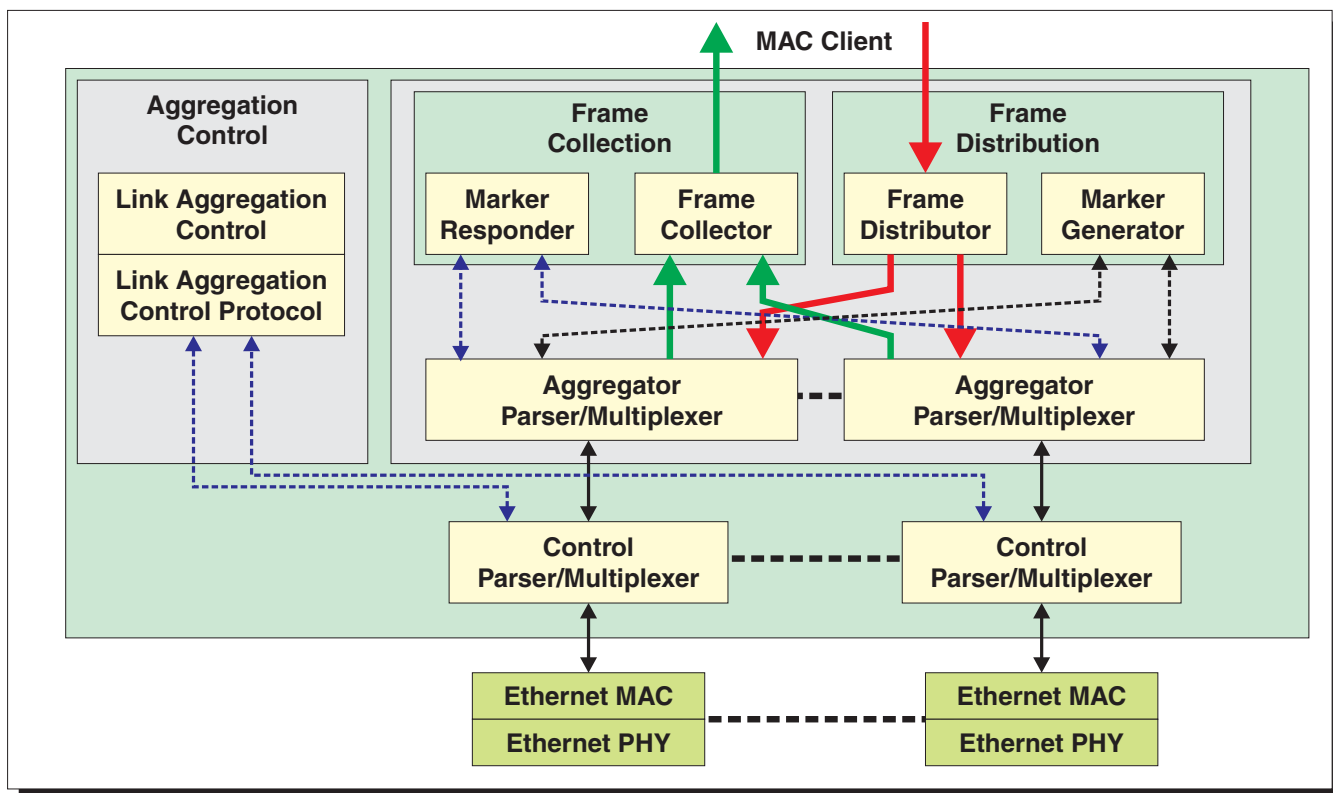


Figure 7-19. Ethernet Link Aggregation

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.

Index

Symbols

Numerics

- 10 Gigabit Fibre Channel (10GFC) 22
- 802.1AB-2005 (Station and Media Access Control Connectivity Discovery) 256

A

- A Flag 161
- AckNo Field 261
- Active zone set 210, 275
- Address
 - alias address identifier 275
 - assignment 275
 - Identifier 275
 - preferred 284
 - previously acquired 285
- Adjacent switch 275
- Alias
 - address identifier 275
 - zone 208
- Arbitrated Loop
 - circuit 282
 - defined 275
 - fabric address 282
 - failure 282
 - port 282
 - tenancy 282
- Area 275
 - Identifier 275
- Available BB_Credit 275
- Avionics Environment project 24

B

- B_Port, *see* Bridge Port (B_Port)
- Backwards Congestion Notification (BCN) 248
- Bandwidth
 - 10-gigabit Fibre Channel 22

- Bandwidth Groups (BWGs) 243
- BB_Credit 275
 - available 275
 - login 282
- BCN Frame Format 254
- BCNA Field 266
- Bits
 - CP Admin Mode 266
 - Enable 263
 - Error 264
 - Remove Tag Oper Mode 267
 - RP Admin Mode 266
 - RP Oper Mode Bit 266
 - Willing 263
- Bridge Port (B_Port) 276
- Broadcast Link 276
- Buffer-to-Buffer flow control 105
- BWG_Percentage Field 265

C

- Class Enable Vector 242
- Class-F 277
- Class-n 277
- Close 277
- Command Descriptor Block (CDB) 23
- Congestion Point (CP) 249
- Converged Enhanced Ethernet (CEE) 90
- Converged Network Adapter (CNA) 3
- Conversations 108
- Core 277
- CP Admin Mode Bit 266
- Cp Sf Field 267
- Cp Sr Field 267
- CPID 255
- CPID Field 255, 256
- Credit 277
- Current Fill Word 277

D

- D_A_TOV 169

D_S_TOV, *see Distributed Services Time-Out Value (D_S_TOV)*
Data Center Bridge Capability Exchange Protocol (DCBX) 259
Data Center Ethernet (DCE) 90
DCBX
 Application Feature 267
 Logical Link Down Feature 267
DCBX Control Sub-TLV 261
DesiredCfg Field 264
Discovery 143
Distributed Services Time-Out Value (D_S_TOV) 278
Domain 278
 address manager 278
 Identifier 278
Domain_ID
 list 278
 preferred 284
Downstream principal ISL 278
drift factor 254

E

E_D_TOV, *see Error Detect Time-Out Value (E_D_TOV)*
E_Port 279
 Identifier 278
 Name 278
Enable Bit 263
End-to-End flow control 105
Enhanced Transmission Selection 247
Entry switch 278
Error Bit 264
Error Detect Time-Out Value (E_D_TOV) 278
ESC, *see Exchange Switch Capabilities (ESC)*
ESCON 24
EtherChannel 108
Ethernet
 conversations 108
 forwarding table aging 96
 frame collector 109
 frame distribution 108
 Group MAC Addresses 92
 link aggregation 107
 Link Aggregation Protocol 109

Multi-Link Trunking (MLT) 108
NIC bonding 108
NIC teaming 107
port teaming 108
transit switches 39
trunking 108
EtherType 90
EtherType x'88CC' 257
Exchange Fabric Parameters (EFP) 278
Exchange Link Parameters (ELP) 54, 278
Exchange Switch Capabilities (ESC) 278
Exchange_ID 278
Expansion Port (E_Port) 279
Extended Link Services
 Fabric Login (FLOGI) 282
 N_Port Login (PLOGI) 282

F

F Flag 161
F_Port 21, 279, 280
F_S_TOV, *see Fabric Stability Time-Out Value (F_S_TOV)*
Fabric 279
 Controller 279
 Element 279
Fabric Discover (FDISC) 120, 122
Fabric Login (FLOGI) 120, 122
 BB_Credit 282
Fabric Port 279, 280
Fabric Provided MAC Address (FPMA) 134, 279
Fabric Provided MAC Addresses (FPMA) 136
Fabric Shortest Path First (FSPF) 24
Fabric_Name 124
FC Entity 279
FC-AL 279
FC-AL-2 279
FCF (FCoE Forwarder) 280
FC-GS 22
FC-GS-2 22
FC-GS-3 23
FC-GS-4 23
FC-GS-5 23
FC-GS-6 23
FC-MAP 279
FCoE Controller 121, 126, 280

- FCoE differences
 - Arbitrated Loop Not Supported 238
 - Class-1 Service Not Supported 238
 - Encoding/Decoding 237
 - Extended Link Services (ELS) 239
 - Fibre Channel Speed Negotiation 237
 - Flow Control 238
 - Inter-Frame Gap (IFG) 238
 - Link Initialization 237
 - Ordered Sets 237
 - Path Establishment and Removal Link Services 239
 - Read Link Status (RLS) 239
 - Start of Frame and End of Frame Delimiters 238
- FCoE End Node (Enode) 121
- FCoE Entity 280
- FCoE Forwarder (FCF) 124
- FCoE Initialization Protocol (FIP) 159, 239
- FCoE Link End Point 118
- FCoE Link Endpoint (FCoE_LEP) 280
- FCoE Node (ENode) 280
- FCoE_LEP 118
- FC-PH 279
- FC-PH-2 280
- FC-PH-3 280
- FC-RDMA 24
- FC-SB 24
- FC-SB-2 24
- FC-SB-3 24
- FC-SW 280
- Fibre Channel Base-T 22
- Fibre Channel Generic Services-2 (FC-GS-2) 22
- Fibre Channel Generic Services-3 (FC-GS-3) 23
- Fibre Channel Generic Services-4 (FC-GS-4) 23
- Fibre Channel Generic Services-5 (FC-GS-5) 23
- Fibre Channel Generic Services-6 (FC-GS-6) 23
- Fibre Channel Standards
 - FC-AE 24
 - FC-AL 25
 - FC-AV 24
 - FC-PH 20
 - FC-PH-3 21
 - FC-PI 22
 - FC-SW 24
- Fibre Channel Switch Fabric (FC-SW) 24
 - Fibre Channel Switch Fabric-2 (FC-SW-2) 24
 - Fibre Channel Switch Fabric-3 (FC-SW-3) 24
 - Fibre Channel Switch Fabric-4 (FC-SW-4) 25
- FICON 24
- Field
 - SeqNo 261
- Fields
 - AckNo 261
 - BCNA 266
 - BWG_Percentage 265
 - Cp Sf Field 267
 - Cp Sr Field 267
 - CPID 255, 256
 - DesiredCfg 264
 - FIP Descriptor List Length 160
 - FIP Operation Code 160
 - Max_Version 261, 263
 - Oper_Version 261, 263
 - Qoff 255
 - Rp Alpha Field 267
 - Rp Beta Field 267
 - Rp Gd Field 267
 - Rp Gi Field 267
 - Rp Rd Field 267
 - Rp Rmin Field 267
 - Rp Ru Field 267
 - Rp Td Field 267
 - Rp Tmax Field 267
 - Rp W Field 267
 - Strict Priority 265
 - SubCode 160
 - SubType 264
 - Timestamp 255, 256
 - User Priority Percentage 265
- Fill Word 280
- FIP Descriptor List Length 160
- FIP Descriptors 161
- FIP Fabric_Name Descriptor 163
- FIP FC-MAP Descriptor 163
- FIP FDISC_NPIV Descriptor 164
- FIP FKA_ADV_Period Descriptor 165
- FIP Flags
 - Available for Login (A) 161
 - FCF (F) 161
 - FPMA Support (FP) 161
 - Solicited (S) 161

- SPMA Support (SP) 161
- FIP FLOGI/FDISC/LOGO Descriptor 164
- FIP Link Reset 180
- FIP LOGO Descriptor 165
- FIP MAC Address Descriptor 163
- FIP Maximum Receive Size Descriptor 164
- FIP Name_Identifier Descriptor 163
- FIP Operation Code 160
- FIP Priority Descriptor 162
- FIP VN_Port Identification Descriptor 165
- FKA_ADV_Period 169, 233
- FKA_ADV_Period Descriptor 165
- FL_Port 21
- Flooding 280
 - reliable 285
- Flow control
 - Buffer-to-Buffer 105
 - End-to-End Credit 105
 - Ethernet Pause 106
- FP Flag 161
- FPMA 280
- Frame 280
- Frame Check Sequence 192
- Frame Check Sequence (FCS) 91
- Frame collector 109
- Frame distributor 108
- Framing and Signaling (FC-FS-2) standard 120
- FS-SW-2 24, 25
- FS-SW-3 24
- FS-SW-5 25
- Full-Duplex 281
- Fx_Port 281

G

- Generic Services 22

H

- Half-Duplex 281
- Hard Address 281
- HBA, *see Host Bus Adapter (HBA)*
- Hello 281
- Host Bus Adapter (HBA) 281

I

- Identifier
 - Area 275
 - Domain 278
- Information Unit 281
- Interject 281
- Intermix 281
- Internet SCSI (iSCSI) 12
- Inter-Switch Link (ISL) 24, 281
- Isolated 281

J

- JBOD 281
- jumbo Ethernet 192

K

- Key Distribution Server 23

L

- L_Port 281
- LAN 281
- Link 281
 - Point-to-Point 284
- Link Aggregation 107
- Link Aggregation Protocol 109
- Link End Point 118
- Link Failure protocol 180
- Link Initialization protocol 180
- Link Level Discovery Protocol 256
- Link State Update (LSU) 282
- Link Strict Priority 244
- Link-Level Discovery Protocol (LLDP) 256
- LIP, *see Loop Initialization Primitive Sequence (LIP)*
- LLDP Protocol Data Units (PDUs) 257
- Local Area Network (LAN) 282
- Local Switch 282
- Logout (LOGO) 122
- Loop Fabric Address 282
- Loop Port State Machine 282
- Loop_ID 282
- Lossless Ethernet Bridging Element 283
- Lossless Ethernet MAC 283
- Lossless Ethernet Network 283

LPSM 282

M

Management Information Base 260
Max_Version Field 261, 263
Metropolitan Area Network (MAN) 283
MIB 260
MIL-STD-1553 24
Multicast
 Group_ID 283
 Group_number 283
Multi-Link Trunking (MLT) 108

N

N_Port 20, 283, 284
 Identifier 283
N_Port ID Virtualization (NPIV) 122
N_Port Login (PLOGI)
 BB_Credit 282
Name 283
 Node_Name 209
Name Server 23, 122
Names
 E_Port 278
NAS 283
Neighbor 283
Network Operating System (NOS) 283
NIC bonding 108
NIC Teaming 107
NL_Port 21, 283
Node 283
 Port 283
Node_Name 209, 283
Non-L_Port 283
Non-Participating Mode 284
Nx_Port 284

O

Open 284
 Originator 284
 Recipient 284
Open Systems Interconnect (OSI) 89
Oper_Version Field 261, 263
Organizationally Unique Identifier (OUI) 91

OX_ID 284

P

Participating Mode 284
Path 284
Path Selection 284
Pause flow control 106
Pause_Time field 242
PDU 284, 285
Per VLAN Spanning Tree (PVST) 103
PLDA 285
Point-to-Point
 link 284
Point-to-Point topology 20, 24
Port 284
 Identifier 284
 Mode 284
Port teaming 108
Port_Name 120, 124, 284
Ports
 arbitrated loop 282
 F_Port 21
 FL_Port 21
 Fx_Port 281
 N_Port 20, 283
 NL_Port 21, 283
 Node 283
 Nx_Port 284
Preamble 90
Preferred Address 284
Preferred Domain_ID 284
Principal Downstream Link 278
Principal ISL 285
Principal Switch 285
Priority Group Feature Sub-TLV 264
Priority Group ID (PGID) 247
Priority Groups (PGs) 243
Private
 loop device 285
 NL_Port 285
Private Loop Direct Attach 285
Protocol 285
Protocols
 Link Failure 180
 Link Initialization 180

Public
NL_Port 285

Q

Qdelta 255
Qoff 255
Qoff Field 255

R

R_A_TOV 285
Rate Limited Tag (RLT) 251, 256
Reaction Point (RP) 250
Reliable Flooding 285
Remote switch 285
Remove Tag Oper Mode Bit 267
Request Domain Identifier (RDI) 285
Request Rate 285
Resource Allocation Time-Out Value (R_A_TOV)
285
Router 286
Routing 286
RP Admin Mode Bit 266
Rp Alpha Field 267
Rp Beta Field 267
Rp Gd Field 267
Rp Gi Field 267
RP Oper Mode Bit 266
Rp Rd Field 267
Rp Rmin Field 267
Rp Ru Field 267
Rp Td Field 267
Rp Tmax Field 267
Rp W Field 267

S

S Flag 161
SCSI FCP-3 23
SCSI FCP-4 23
SCSI, *see* *Small Computer System Interface (SCSI)*
SCSI-3
FCP protocol mapping 23
SeqNo Field 261
Server Provided MAC Address (SPMA) 134, 286

Server Provided MAC Addresses (SPMA) 134
Service Rate 286
SFP+ 103
Simple Network Management Protocol (SNMP)
260, 286
Single-Byte Command Code Sets (FC-SB-2) 24
Small Computer System Interface (SCSI) 286
SNMP 260
SNMP, *see* *Simple Network Management Protocol (SNMP)*
SONET, *see* *Synchronous Optical Network (SONET)*
SP Flag 161
SPMA 286
Storage Area Network 286
Strict Priority Field 265
SubCode 160
SubType Field 264
SW_ACC, *see* *Switch Accept (SW_ACC)*
SW_ILS, *see* *Switch Internal Link Service (SW_ILS)*
SW_RJT, *see* *Switch Reject (SW_RJT)*
Switch 286
Adjacent 275
Entry 278
Local 282
remote 285
Switch_Name 287
Switch_Priority 287
Switch Accept (SW_ACC)
Switch Internal Link Service (SW_ILS) 286
Exchange Fabric Parameters (EFP) 278
Exchange Link Parameters (ELP) 278
Exchange Switch Capabilities (ESC) 278
Link State Update (LSU) 282
Request Domain Identifier (RDI) 285
Switch Reject (SW_RJT)
Switch_Name 124, 287
Switch_Priority 287
Synchronous Optical Network (SONET) 287

T

tbd 26
Acknowledgment Section xviii
does an RP intercept frames for a SA outside

- the congestion managed region? 255
- Does SCSI need to know about FCoE? 239
- ELP as FIP? 239
- Ethernet Standards for FCoE 26
- FIP error handling 182
- FPMA FLOGI request MAC address descriptor value? 198
- how does an FCF logout a VN_Port? In FC, it can simply take the link down. 200
- jumbo frame support following Ethernet STP reconfiguration 145, 169
- LLDP TLV detail? 257
- LOGO as FIP? 239
- PLOGI as FIP? 239
- Qdelta field 255
- virtual path removal with LOGO 51
- Time To Live Field 257
- Timers
 - Arbitration Wait (AW_TOV) 275
 - Distributed Services Time-Out Value (D_S_TOV) 278
 - Error Detect Time-Out Value (E_D_TOV) 278
 - Fabric Stability Time-Out Value (F_S_TOV)
 - LIS_HOLD_TIME 282
 - Loop Master (LM_TOV) 282
 - Resource Recovery (RR_TOV) 285
 - Upper Level Protocol (ULP_TOV) 287
- Timestamp Field 255, 256
- TLVs
 - DCBX Control 261
 - Priority Group Feature 264
- Topologies
 - point-to-point 20, 24
- Topology 287
- Transfer 287
- Transit switch 39
- Transmission Character 287
- Transmission Word 287

U

- UDP 287
- Universally Administered OUI 91
- Upstream principal ISL 287
- User Priority Percentage Field 265

V

- VE_Port_Name 287
- VF_Port_Name 287
- Virtual E_Port (VE_Port) 287
- Virtual F_Port (VF_Port) 287
- Virtual LANs (VLANs) 103
- Virtual Link 288
- Virtual link 118
- Virtual N_Port (VN_Port) 288
- VKA_ADV_Period 233, 236
- VN_Port 120
- VN_Port Identification Descriptor 165
- VN_Port_Name 120, 288

W

- Wide Area Network (WAN) 288
- Willing Bit 263

Z

- Zero Domain_ID List 288
- Zone 209, 288
 - Active zone set 275
 - alias 209
 - definition 288
 - member 288
 - name 288
 - set 210
- Zone Alias 208
- Zone Set 288
 - name 288
 - state 288
- Zoning
 - at egress switch port 211
 - at ingress switch port 211
- Zoning configuration 288

