EMC²

Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Concepts and Protocols

Version 15

- Fibre Channel over Ethernet (FCoE) and Ethernet Basics
- Storage in an FCoE Environment
- EMC RecoverPoint and EMC Celerra MPFS as Solutions
- Troubleshooting Basic FCoE and CEE Problems

Mark Lippitt
Erik Smith
David Hughes

EMC²
TECHBOOKS

For the most up-to-date regulator document for your product line, go to EMC Online Support (https://support.emc.com).

Part number H6290.19

Contents

Chapter 1 Introduction to Fibre Channel over Ethernet

Chapter 2 Ethernet Basics

Chapter 5 Troubleshooting Basic FCoE and CEE Problems and Case Studies

Figures

Tables

Preface

The former EMC Engineering Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Concepts and Protocols TechBook *has been divided into two separate TechBooks. This TechBook provides an introduction to Fibre Channel over Ethernet (FCoE) along with basic information to better understand the various aspects and protocols involved with a typical Ethernet environment. FCoE connectivity and storage in an FCoE environment are discussed. Information on RecoverPoint and Celerra MPFS as solutions in an FCoE environment is included. Basic troubleshooting techniques are also provided.*

The Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook *provides supported configurations and features along with case studies to show how to incorporate EMC Connectrix B switches, Cisco Nexus switches, and Blade Servers utilizing FCoE into an existing data center. This TechBook can be found at http://elabnavigator.EMC.com,* **Documents > Topology Resource Center.**

E-Lab would like to thank all the contributors to this document, including EMC engineers, EMC field personnel, and partners. Your contributions are invaluable.

As part of an effort to improve and enhance the performance and capabilities of its product lines, EMC periodically releases revisions of its hardware and software. Therefore, some functions described in this document may not be supported by all versions of the software or hardware currently in use. For the most up-to-date information on product features, refer to your product release notes. If a product does not function properly or does not function as described in this document, please contact your EMC representative.

Note: This document was accurate at publication time. New versions of this document might be released on EMC Online Support https://support.EMC.com. Check to ensure that you are using the latest version of this document.

Audience

This TechBook is intended for EMC field personnel, including technology consultants, and for the storage architect, administrator, and operator involved in acquiring, managing, operating, or designing a networked storage environment that contains EMC and host devices.

EMC Support Matrix and E-Lab Interoperability Navigator

For the most up-to-date information, always consult the *EMC Support Matrix* (ESM), available through E-Lab Interoperability Navigator at http://elabnavigator.EMC.com.

Related documentation

Related documents include:

- The following documents, including this one, are available through the E-Lab Interoperability Navigator, **Documents > Topology Resource Center**, at http://elabnavigator.EMC.com.

  These documents are also available at the following location:

  http://www.emc.com/products/interoperability/topology-resource-center.htm

  - Backup and Recovery in a SAN TechBook
  - Building Secure SANs TechBook
  - Extended Distance Technologies TechBook
  - Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook
  - Fibre Channel SAN Topologies TechBook
  - iSCSI SAN Topologies TechBook
  - Networked Storage Concepts and Protocols TechBook
  - Networking for Storage Virtualization and RecoverPoint TechBook
  - WAN Optimization Controller Technologies TechBook
  - EMC Connectrix SAN Products Data Reference Manual
  - Legacy SAN Technologies Reference Manual
  - *Non-EMC SAN Products Data Reference Manual*
- *EMC Support Matrix*, available through E-Lab Interoperability Navigator at http://elabnavigator.EMC.com.
- RSA security solutions documentation can be found at http://RSA.com > **Content Library**.

All of the following documentation and release notes can be found at EMC Online Support at https://support.emc.com.

Hardware documents and release notes include those on:

- Connectrix B series
- Connectrix M series
- Connectrix MDS (release notes only)
- VMAX
- VNX series
- CLARiiON
- Celerra

Software documents include those on:

- EMC Ionix ControlCenter
- RecoverPoint
- Invista
- TimeFinder
- PowerPath

The following E-Lab documentation is also available:

- Host Connectivity Guides
- HBA Guides

For Cisco and Brocade documentation, refer to the vendor's website.

- http://cisco.com
- http://brocade.com

Authors of this TechBook

This TechBook was authored by Mark Lippitt, Erik Smith, and David Hughes, with contributions from other EMC engineers, EMC field personnel, and partners.

Mark Lippitt is a Technical Director in EMC E-Lab with over 30 years experience in the storage industry, including Engineering and Marketing roles at Data General, Tandem Computers, and EMC. Mark initiated and led the Stampede project in 1997, which became EMC's first Connectrix offering. Mark is an active T11 participant, a committee within the InterNational Committee for Information Technology Standards, responsible for Fibre Channel Interfaces.

Erik Smith is a Consulting Technologist for the Connectrix business unit within EMC Engineering. Over the past 17 years, Erik has held various technical roles in both EMC Engineering and Technical Support. Erik has authored and coauthored several EMC TechBooks. Erik is also a member of T11.

Dave Hughes is a Principal Systems Integration Engineer and has been with EMC for over 18 years. While at EMC, David has held various technical positions within Engineering, Customer Service Technical Support, and IT. David is currently a member of the Advanced Infrastructure and Interfaces group within E-Lab, where he evaluates new technologies and solutions.

Conventions used in this document

EMC uses the following conventions for special notices:

IMPORTANT

An important notice contains information essential to software or hardware operation.

Note: A note presents information that is important, but not hazard-related.

Typographical conventions

EMC uses the following type style conventions in this document.

Normal

Used in running (nonprocedural) text for:

- Names of interface elements, such as names of windows, dialog boxes, buttons, fields, and menus
- Names of resources, attributes, pools, Boolean expressions, buttons, DQL statements, keywords, clauses, environment variables, functions, and utilities
- URLs, pathnames, filenames, directory names, computer names, links, groups, service keys, file systems, and notifications

Bold

Used in running (nonprocedural) text for names of commands, daemons, options, programs, processes, services, applications, utilities, kernels, notifications, system calls, and man pages

Used in procedures for:

- Names of interface elements, such as names of windows, dialog boxes, buttons, fields, and menus
- What the user specifically selects, clicks, presses, or types

Italic

Used in all text (including procedures) for:

- Full titles of publications referenced in text
- Emphasis, for example, a new term
- Variables

`Courier`	Used for: • System output, such as an error message or script • URLs, complete paths, filenames, prompts, and syntax when shown outside of running text
`Courier bold`	Used for specific user input, such as commands
`Courier italic`	Used in procedures for: • Variables on the command line • User input variables
< >	Angle brackets enclose parameter or variable values supplied by the user
[ ]	Square brackets enclose optional values
|	Vertical bar indicates alternate selections — the bar means "or"
{ }	Braces enclose content that the user must specify, such as x or y or z
...	Ellipses indicate nonessential information omitted from the example

Where to get help

EMC support, product, and licensing information can be obtained as follows.

EMC support, product, and licensing information can be obtained on the EMC Online Support site as described next.

Note: To open a service request through the EMC Online Support site, you must have a valid support agreement. Contact your EMC sales representative for details about obtaining a valid support agreement or to answer any questions about your account.

Product information

For documentation, release notes, software updates, or for information about EMC products, licensing, and service, go to the EMC Online Support site (registration required) at:

https://support.EMC.com

Technical support

EMC offers a variety of support options.

Support by Product — EMC offers consolidated, product-specific information on the Web at:

https://support.EMC.com/products

The Support by Product web pages offer quick links to Documentation, White Papers, Advisories (such as frequently used Knowledgebase articles), and Downloads, as well as more dynamic content, such as presentations, discussion, relevant Customer Support Forum entries, and a link to EMC Live Chat.

EMC Live Chat — Open a Chat or instant message session with an EMC Support Engineer.

eLicensing support

To activate your entitlements and obtain your VMAX or Symmetrix license files, visit the Service Center on https://support.EMC.com, as directed on your License Authorization Code (LAC) letter e-mailed to you.

For help with missing or incorrect entitlements after activation (that is, expected functionality remains unavailable because it is not licensed), contact your EMC Account Representative or Authorized Reseller.

For help with any errors applying license files through Solutions Enabler, contact the EMC Customer Support Center.

If you are missing a LAC letter, or require further instructions on activating your licenses through the Online Support site, contact EMC's worldwide Licensing team at licensing@emc.com or call:

- North America, Latin America, APJK, Australia, New Zealand: SVC4EMC (800-782-4362) and follow the voice prompts.
- EMEA: +353 (0) 21 4879862 and follow the voice prompts.

We'd like to hear from you!

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user publications. Send your opinions of this document to:

techpubcomments@emc.com

Your feedback on our TechBooks is important to us! We want our books to be as helpful and relevant as possible. Send us your comments, opinions, and thoughts on this or any other TechBook to:

TechBooks@emc.com

1

Introduction to Fibre Channel over Ethernet

This chapter provides an introduction to Fibre Channel over Ethernet (FCoE) and includes the following information:

Note: Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Introduction

I/O consolidation has been long sought by the IT industry to unify the multiple transport protocols in the data center. This section provides a basic introduction to Fibre Channel over Ethernet (FCoE), which is an approach to I/O consolidation that was originally defined by the FC-BB-5 T11 work group and has since been extended by the FC-BB-6 T11 work group.

Much of the information provided in this introduction was derived from the following sources, which also provide more details on FCoE, including encapsulation, frame format, address mapping, lossless Ethernet, and sample topologies:

- *Fibre Channel Over Ethernet: An Introduction*, White Paper

  `http://www.fibrechannel.org`

- Silvano, Gai, *Data Center Network and Fibre Channel over Ethernet*, Nuova Systems Inc., 2008

I/O consolidation, simply defined, is the ability to carry different types of traffic, having different traffic characteristics and handling requirements, over the same physical media. I/O consolidation's most difficult challenge is to satisfy the requirements of different traffic classes within a single network. Since Fibre Channel is the dominant storage protocol in the data center, any viable I/O consolidation solution for storage must allow for the FC model to be seamlessly integrated. FCoE meets this requirement in part by encapsulating each Fibre Channel frame inside an Ethernet frame.

The goal of FCoE is to provide I/O consolidation over Ethernet, allowing Fibre Channel and Ethernet networks to share a single, integrated infrastructure, thereby reducing network complexities in the data center. An example is shown in Figure 1 on page 21.

FCoE consolidates both SANs and Ethernet traffic onto one Converged Network Adapter (CNA), eliminating the need for using separate Host Bus Adapters (HBAs) and Network Interface Cards (NICs).

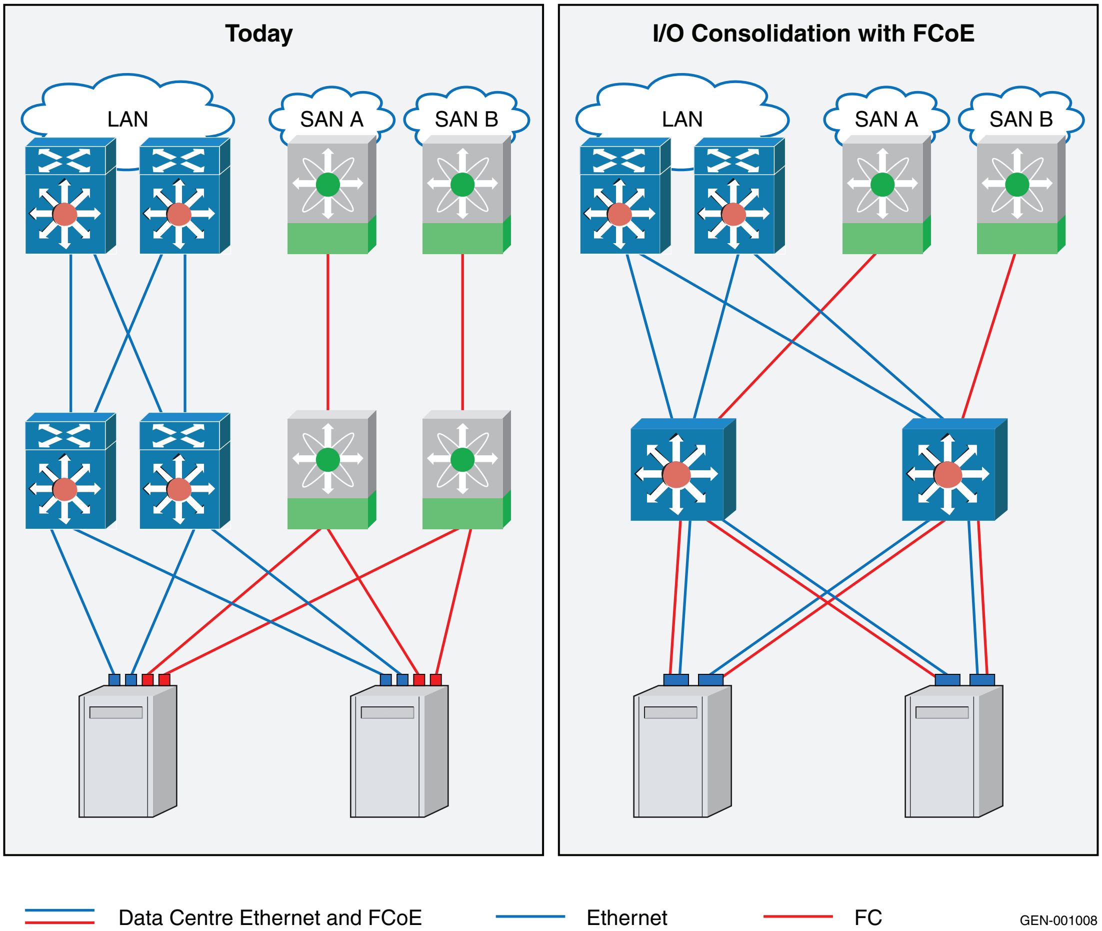


Figure 1 **Typical topology versus FCoE example using Cisco Nexus 5000**

Note: Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, , **Documents> Topology Resource Center**.

History

Customer and competitive imperatives for better, faster, cheaper IT solutions have driven convergence. For example,

- Servers have converged on the x86 architecture.
- Desktop operating systems have converged on Windows.
- Server operating systems have converged on Linux.
- LAN architectures have converged on Ethernet.

Although information technologists have made many attempts to converge the I/O of LAN and storage onto one wire, the goal of unifying the server's I/O resources has been elusive. A brief history of converging IP with storage I/O is shown in Figure 2, followed by a brief description of the emergence of FCoE technology.

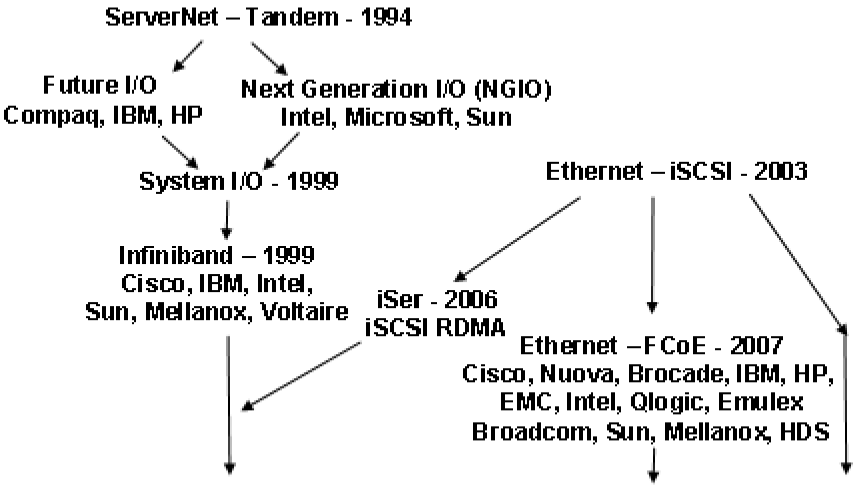


Figure 2 **Converged I/O history**

Tandem Computers in 1994 introduced the industry to ServerNet. ServerNet found limited success, but it provided a launching point for two more attempts at unified I/O. Competitors coalesced into two camps: Future I/O and NGIO. Once again the challenge was steeper than the initiative. The initiative then condensed into a single effort named System I/O, which was quickly renamed to Infiniband.

Infiniband met the challenge of providing an entire ecosystem but has found limited market acceptance. True to its ancestry, its architecture disrupted everything from physical media to operating system middleware. As a result, its success has been a narrow IT niche where reducing server-to-server latency from milliseconds to microseconds yields large financial returns.

iSCSI attacked the opportunity for Unified I/O by exploiting the ubiquity of Ethernet. iSCSI has been able to achieve significant success in the small/medium business (SMB) sector by leveraging:

- Physical media
- LAN I/O stacks
- SCSI protocol
- Low cost server interfaces
- Ethernet switching products

However, iSCSI has not penetrated the market for large-scale, data center class, block I/O. Large-scale block I/O infrastructures require a vendor-neutral, centralized management model that is able to provide data center acceptable services for the storage I/O of thousands of servers, storage, and switch ports.

Fibre Channel over Ethernet aims to meet this challenge by leveraging the proven large-scale management model of Fibre Channel switching. Like iSCSI, it leverages the ubiquity and cost effectiveness of Ethernet. Unlike Infiniband and its ancestors, it avoids making changes to many critical layers of the ecosystem.

The designers of FCoE designed compatibility into the existing infrastructure:

- Middleware products, like EMC® PowerPath®
- Storage device drivers, like those of Emulex and QLogic
- SAN hardware from Cisco and Brocade
- SAN management software
- Storage products, like EMC VMAX® and VNX®/CLARiiON®

This high degree of compatibility does the following:

- Minimizes risk and integration costs, critical values to today's data centers
- Enables the storage industry's first data center solution to Unified I/O

The industry’s innovation with Fibre Channel over Ethernet sets a new milestone in the road towards Unified I/O.

Benefits

The Fibre Channel portion of FCoE appears as normal Fibre Channel to a host or a switch, and therefore to a user. It is based completely on the FC model, which makes it easy to understand, manage, and troubleshoot. A major value is that FCoE uses Ethernet hardware to deliver an enterprise storage solution, while also using the existing FC management infrastructure.

The benefits of FCoE include:

- Becomes part of the Fibre Channel architecture, allowing for:
  - Seamless integration with existing FC SANs
  - Uses existing FC SAN admin tools and workflows
- Requires no gateway
  - Since the FC frame is untouched, the operation is completely stateless
- Provides the following current functions and services, allowing for a smooth transition:
  - Zoning
  - dNS (distributed Name Server)
  - RSCN (Registered State Change Notification)
  - FSPF (Fibre Channel Shortest Path First)
  - Management tools
  - Storage and server virtualization

Further benefits include:

- Fewer cables, simplifying cable management
- Fewer adapters and switch ports, saving in power, equipment, and cooling costs

Terminology

Table 1 provides commonly used acronyms.

Table 1 Acronyms (page 1 of 2)

Acronym	Definition
ACL	access control list
CEE	Converged Enhanced Ethernet (Deprecated, see “DCB”)
CNA	Converged Network Adapter
CRC	Cyclical Redundancy Check
DA	Destination MAC Address
DCB	Data Center Bridge
DCBX	Data Center Bridging Capability eXchange Protocol
DCFM	Data Center Fabric Manager
ETHv2	Ethernet Version 2
EOF	End of Frame
ETS	Enhanced Transmission Selection
FCF	FCoE Forwarder
FCoE	Fibre Channel over Ethernet
FPMA	Fabric Provided MAC Address
IP	Internet Protocol
LACP	Link Aggregation Control Protocol
LAG	Link Aggregation Group
LAN	Local Area Network
LLDP	Link Layer Discovery Protocol
MAC	Media Access Control
MSTP	Multiple Spanning Tree Protocol
NPV	N_Port virtualization

Table 1 Acronyms (page 2 of 2)

Acronym	Definition
NPIV	N_Port ID virtualization
NIC	Network Interface Card
PFC	Priority Flow Control
QoS	Quality of Service
SA	Source MAC Address
SAN	Storage Area Network
SPMA	Server Provided MAC Address
STP	Spanning Tree Protocol
RSTP	Rapid Spanning Tree Protocol
VE	Virtual E port
VF	Virtual Fabric port
VFC	Virtual Fibre Channel
VLAN	Virtual LAN (ID)
WAN	Wide Area Network

Management tools

The management tools used to manage FCoE and Fibre Channel environments are similar.

CNA management tools include:

- Emulex — HBAnywhere/OneCommand Manager
- QLogic — SANSurfer

Switch management tools include:

- Brocade
  - Fabric OS CLI
    - For configuration of Fibre Channel features
  - CMSH (CEE management shell) CLI
    - For configuration of Converged Enhanced Ethernet (CEE) features
  - CMCNE - Connectrix Manager Converged Network Edition
- Cisco
  - Fabric Manager
    - For configuration of Fibre Channel features
  - NX-OS CLI
    - For configuration of CEE features

Cable management recommendations

Consider the following recommendations for cable management.

The minimum bend radius for a 50 micron cable is 2 inches under full tensile load and 1.2 inches with no tensile load.

Cables can be organized and managed in a variety of ways, for example, using cable channels on the sides of the cabinet or patch panels to minimize cable management. Following is a list of recommendations:

Note: You should not use tie wraps with optical cables because they are easily overtightened and can damage the optic fibers.

- Plan for rack space required for cable management before installing the switch.
- Leave at least 1 m (3.28 ft) of slack for each port cable. This provides room to remove and replace the switch, allows for inadvertent movement of the rack, and helps prevent the cables from being bent to less than the minimum bend radius.
- If you are using Brocade ISL Trunking, consider grouping cables by trunking groups. The cables used in trunking groups must meet specific requirements, as described in the *Fabric OS Administrator's Guide*.
- For easier maintenance, label the fiber optic cables and record the devices to which they are connected.
- Keep LEDs visible by routing port cables and other cables away from the LEDs.
- Use hook and loop style straps to secure and organize fiber optic cables.

Enabling technologies

The following sections describe just a few of the technologies and protocols required to make I/O consolidation practical in large scale environments:

- "Converged Network Adapter" on page 29
- "Fibre Channel Forwarder" on page 30
- "FIP Snooping Bridge" on page 30
- "Data Center Bridging (DCB)" on page 38
- "Priority Flow Control and PAUSE" on page 39
- "Data Center Bridging eXchange" on page 40

Converged Network Adapter

A Converged Network Adapter (CNA) is similar to an HBA or a NIC, but instead of handling either FC or IP, the CNA can handle both simultaneously. The CNA presents separate networking and storage system interfaces to the operating system. The interfaces preserve compatibility with existing system software, middleware, and management tools.

The first generation CNAs used three ASICs: an Intel ASIC for networking, a FC-HBA ASIC, and a mux ASIC from Cisco. The first generation CNAs achieved time-to-market but were full-height, full-length PCIe adapters with high wattage requirements.

The second generation CNAs (QLogic QLE 81xx/ Brocade10x0/ Emulex OCe10102) feature a single ASIC implementation that helps to reduce power consumption and improve reliability.

After the second generation of CNAs were introduced, software initiators from Broadcom and Intel were released. These software initiators allowed for end users to take advantage of FCoE without needing to purchase special adapters.

Recently, EMC completed the qualification of the third-generation of CNAs. These new CNAs allow for more than two PCIe functions per physical port and as a result support more than two personalities (protocols) per physical port.

For all types of CNAs, from an end-user's perspective, the FC and Ethernet instances appear in the OS just as they would if discrete 10 GbE NICs and FC HBAs were used.

Fibre Channel Forwarder

The purpose of the Fibre Channel Forwarder is to service login requests and provide the FC services typically associated with a FC switch. FCFs may also optionally provide the means to:

- De-encapsulate FC frames that are coming from the CNA and going to the SAN.
- Encapsulate FC frames that are coming from the SAN and going to the CNA.

Examples of products that provide an FCF function are the Cisco MDS 9250i, Nexus 7000, Nexus 5596, Nexus 5548, Nexus 5020, Nexus 5010, Brocade 6740, and EMC Connectrix® DCX and MP-8000B.

FIP Snooping Bridge

A FIP Snooping Bridge is an Ethernet Bridge that supports:

- Priority Flow Control (PFC - 802.1Qbb)
- Enhanced Transmission Selection (ETS - 802.1Qaz)
- Data Center Bridging Capabilities Exchange Protocol (DCBX - 802.1Qaz)
- Dynamic ACLs as described in FC-BB-5 Annex C, discussed further on page 37.

The FCoE Initialization Protocol (FIP), described further on page 44, bridges the gap between the expectations of the Fibre Channel protocol and the reality of an Ethernet network. This section describes why FCoE requires ENodes to be either directly connected to a Fibre Channel Forwarder (FCF) or connected to a FIP Snooping Bridge (FSB) and then to an FCF in order to function properly.

IMPORTANT

FCoE cannot be guaranteed to function properly when a non-FIP-aware Ethernet Bridge is used anywhere in the data path.

Note: The terms "switch" and "bridge" are interchangeably not only in this section but in the Ethernet standards as well.

Before FCoE could be considered ready for production, two well-known networking exploits, the *denial-of-service* and *learning* attacks, had to be addressed. In order to prevent these two attack vectors, a new protocol (FIP) and a new Layer 2 feature (Dynamic ACLs) had to be created. The purpose of this section is three-fold:

- To describe the impact that each of these exploits has on the FC protocol
- To show how FIP snooping and Dynamic ACLs resolve these problems
- To prove why a non-FIP-aware Ethernet bridge should not be used when using FCoE

Case studies are used to better explain these problems:

- "Case study 1: The problem with network joins" on page 32 involves a user inadvertently causing a denial-of-service attack by joining two separate Layer 2 networks.
- "Case study 2: The Rogue Host" on page 36 describes a malicious user attaching a Rogue Host to the network and performing a learning attack.

Case study 1: The problem with network joins

Consider the topology shown in Figure 3. It consists of two physically separate, yet identical, network topologies.

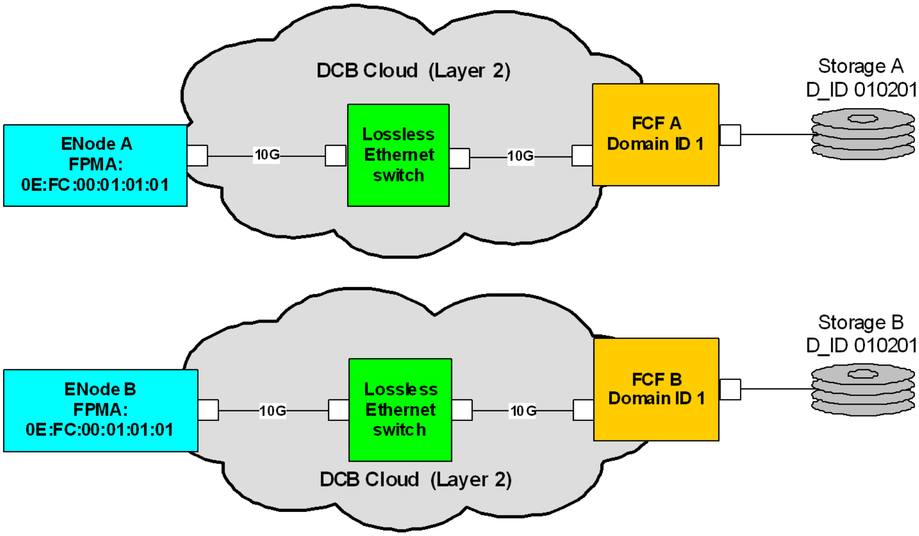


Figure 3 **Network joins topology example**

Note that the FCFs have the same Domain ID of 1. Because of this, it is *possible* that each ENode will be assigned the same Fabric Provided MAC Address (FPMA). Ordinarily, this would not a problem since the ENodes are located on two different L2 networks. However, what happens if someone accidentally connects the two lossless Ethernet switches together, as shown in Figure 4 on page 33?

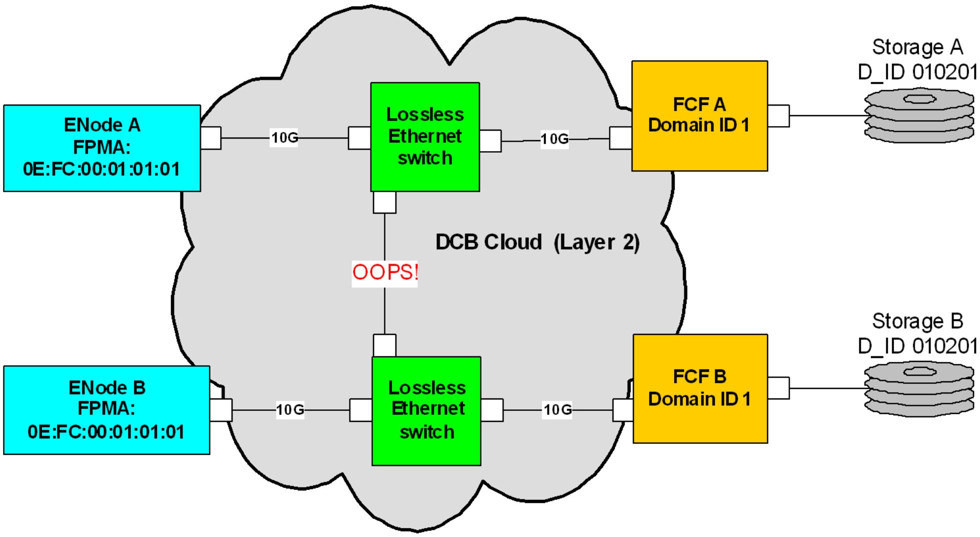


Figure 4 **Lossless Ethernet topology example**

Were this to occur, the two ENodes would have the same MAC Address, which could cause a denial-of-service condition. In fact, whether or not a denial-of-service could occur in the configuration shown in Figure 4 is not the question, but rather, how long it would take before it happened. The reason has to do with how MAC Addresses are learned, which is explained in a little more detail in "Ethernet Frame delivery."

Ethernet Frame delivery

The *Nexus 5000 Architecture* white paper, available at http://www.cisco.com, describes how frames are processed by a Layer 2 Ethernet switch. For the purpose of examining the denial-of-service condition, you only need to consider the destination lookup and Ethernet learning portions of this white paper. As a result, this section is limited to aspects of frame-forwarding.

When an Ethernet Frame is received by an Ethernet switch, one of the many things it will do is to take note of the Ethernet Frame's Source Address (SA) and Destination Address (DA).

- If the SA is unknown to the switch, it will take note of the interface the frame was received on and "learn" the MAC Address by adding it to the station table. By doing so, any frames that are received with that MAC Address in the future will be forwarded back down the interface that it was received on.
- If the DA is unknown to the switch, it will perform a unicast flood and forward the frame on all interfaces that are in the same network segment (VLAN). This process of learning and unicast flooding will be repeated throughout the broadcast domain until either the frame is received by a switch that recognizes the DA of the flooded frame or until there are no interfaces left to forward the frame on. In this way, the path back to the frame originator will be built as the frame is flooded across the broadcast domain. If an end station with that DA exists and it transmits a response back to the originator of the flooded frame, the learning process will be repeated as the frame transits the network. However, since the path back to the originator of the unicast frame was built as the original frame was flooded, the response will not be flooded. Eventually, the response will be received by the end station that transmitted the original frame. At this point, the path between the end stations has been created and no more unicast flooding will need to be performed as long as the MAC Addresses are in the station table.

In Figure 4 on page 33 , the cable was accidentally connected between the two lossless Ethernet switches. In this case, frames would continue to be delivered to the proper end stations until at least one of the end station MAC Addresses is flushed from the station table. This could happen for a number of reasons, including:

- A MAC Address is removed from the station table due to a loss of physical connectivity to the end station
- A MAC Addresses is aged out of the station table
- An administrator manually clears the station table
- A Topology Change Notification (TCN) is received by the switch.

  A TCN being received by the switch would occur after the Spanning-Tree Protocol ran and one of the switches recognized a change in the network configuration. When the TCN is received, the forwarding entries in the station table would be rapidly aged out and the process of unicast flooding would need to be performed in order to rebuild the forwarding entries.

Figure 5 shows case "A" and assumes that for some reason FCF "A" was physically disconnected from the lossless Ethernet switch.

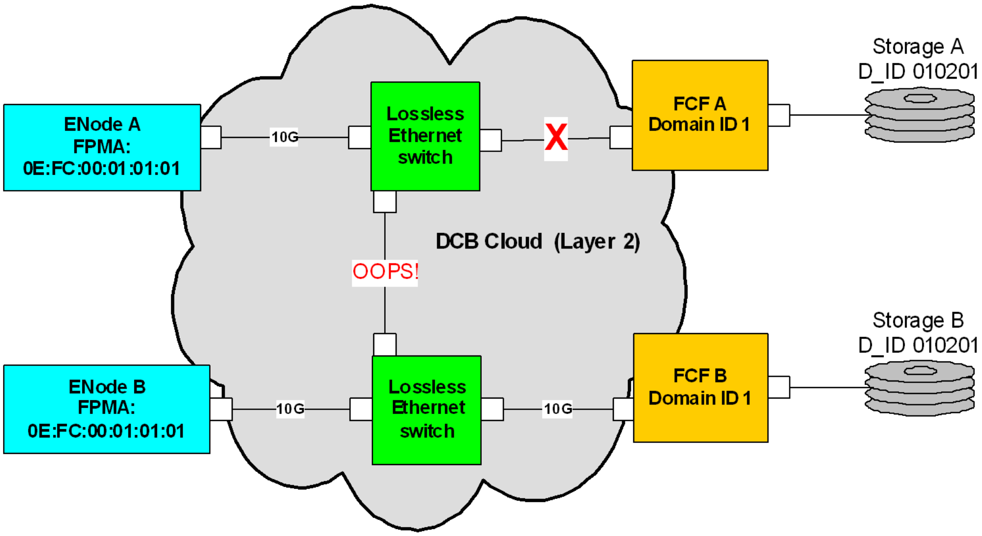


Figure 5 Flooding example

When this happens, the FPMA for FCF A will be cleared from the station table and subsequent frames from ENode "A" to array "A" will be flooded. As a part of the flooding process, this frame would be forwarded across the link between the two switches. hen the frame is received by the ingress port on the switch at the other end of the link, the ingress port will not recognize the DA since it belongs to FCF "A", and the frame will need to be flooded. It will, however, take note of the SA and update its station table to indicate that frames destined back to this FPMA should be forwarded across the link back to the other switch. This means any frames from Array "B" going back to ENode "B" will be incorrectly forwarded back across the link and on to ENode "A". As soon as the next frame is received from ENode "B", the station table would be correctly updated and frames would be forwarded to ENode "B" again. The problem is, if frames were received from ENode "A" frequently enough, it could cause many frames to be incorrectly forwarded to ENode "A" so ENode "B" would spend all of its time in error recovery and get no meaningful work done. Therefore, a denial of service condition would be created.

A FIP Snooping Bridge prevents this problem by only allowing frames either from— or to—an FCF to be forwarded out of a FIP Snooping uplink.

As defined in FC-BB-5 annex C and D, allowing the forwarding or reception of a frame to or from the FCF is something that should be explicitly enabled on a per-interface basis. As a result, a FIP Snooping Bridge would prevent the scenario described by not allowing these frames to be flooded out of non-FIP Snooping uplinks. In addition, even if one of the FIP Snooping uplinks were accidentally connected to an interface on another FIP Snooping Bridge, the receiving interface would not forward frames with an FCoE Ethertype. For more information, refer to FC-BB-5 annex C and annex D, located at For more detailed information, download a copy of FC-BB-5 at http://www.fcoe.com/09-056v5.pdf.

Case study 2: The Rogue Host

As shown in Figure 6, this problem is simple to understand and easy to write code for and perform in the lab. The topology consists of a host labeled ENode "A" (a non-FIP-aware Lossless Ethernet switch) an FCF-labeled FCF "A" (a storage port labeled Storage "A" and a "Rogue Host"). Assume that ENode "A" is logged into Storage "A".

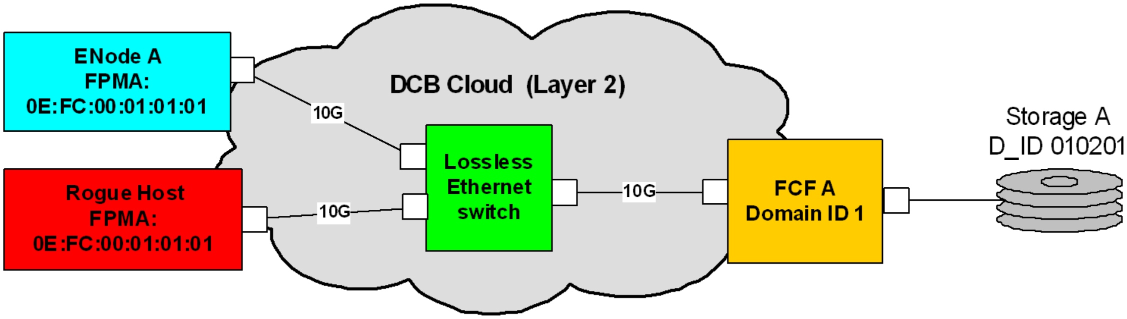


Figure 6 **Rogue Host example**

If a person were interested in using a Rogue Host to gain access to specific data on Storage "A", all they would need to know is the FPMA of an ENode that has access to that data. The FPMA of any ENode can easily be found by using "show fcoe database" on a Nexus 5k (for example). Once in possession of the FPMA, the person in control of the Rogue Host could transmit a single Ethernet Frame (e.g., FIP Solicitation), setting the Source Address (SA) to the value of the FPMA that they are interested in. This would cause a station table update and then allow them to capture whatever frames come back. These frames would contain the FCF-MAC Address and any storage port FCIDs (labeled D_ID in the diagram) that happen to be transmitting frames back to ENode "A" at the time. Once the

FCF-MAC and the D_ID of the storage port is known, the Rogue Host could transmit a SCSI-FCP READ command to the storage port and read data or a SCSI-FCP WRITE command to alter (corrupt) the data.

Preventing this problem requires the use of an ACL as defined in FC-BB-5 annex C and D. This section provides an overview of the functionality. For more detailed information, download a copy of FC-BB-5 at http://www.fcoe.com/09-056v5.pdf.

Dynamic ACLs

By default, an interface on a FIP Snooping bridge will not forward FIP or FCoE frames and will not learn any MAC Addresses contained within these frames. Because of this, the Rogue Host problem described above is prevented from happening by default. However, this default ACL does not prevent many other problems, such as someone making use of a port where the ACL was removed to allow an FCoE host to be attached, or someone hijacking an existing host that is already logged into the fabric. To prevent these types of problems, the ACL would need to specify the exact MAC Addresses that can be used on the interface and this ACL would need to be modified whenever a port was added or removed.

The downside to this solution is that it would create an enormous administrative burden on the network administrator since they would need to manually modify each ACL entry to allow FIP and FCoE frames. Because of this administrative burden and the critical nature of the problem should the ACLs not be used, a solution that would allow for the ACLs to be updated automatically at certain points in the protocol was created. This solution, *Dynamic ACLs*, works as follows:

1. By default, FIP and FCoE frames will not be forwarded from the ENode onto the network.
2. Once a Discovery Advertisement is transmitted to the ENode, the switch will take note of the Discovery Advertisement SA and modify the ACL to allow the attached ENode to transmit a FIP frames to that address.
3. For each successful FIP FLOGI, the ACL will be updated to allow FCoE frames between the ENode's MAC Address and the FCF that it completed login with.

Implementing this behavior as described would prevent the previously described learning attack from being successful.

An important point to note is that the Dynamic ACLs are not actually visible in any of the implementations. For the most part, this has been implemented by forwarding all FIP Frames to the Control Plane of the switch for processing or by discarding FCoE frames until a login has been completed.

Data Center Bridging (DCB)

A DCB Ethernet switch is an Ethernet switch implementation that has certain characteristics, the most important being that they do not drop frames under congestion, or are, in other words, considered to be *lossless*. A lossless network is very important to block I/O operations because unlike TCP/IP, the loss of a single frame typically requires the entire FC exchange to be aborted and re-driven by the upper-layer protocol (ULP), instead of just re-sending a particular missing frame.

Data Center Bridging (DCB) includes:

- Priority-based Flow Control (PFC) — IEEE 802.1Qbb provides a link level flow control mechanism that can be controlled independently for each Class of Service (CoS), as defined by 802.1p. The goal of this mechanism is to ensure zero loss under congestion in DCB networks.
- Enhanced Transmission Selection (ETS) — EEE 802.1Qaz provides a common management framework for assignment of bandwidth to 802.1p CoS-based traffic classes.
- Congestion Notification — IEEE 802.1Qau provides end-to-end congestion management for protocols that are capable of transmission rate limiting to avoid frame loss. Although not yet implemented by any products supported by EMC, it is expected to benefit protocols such as TCP that do have native congestion management, as it reacts to congestion in a more timely manner.
- Data Center Bridging Capability eXchange Protocol (DCBX) — A discovery and capability exchange protocol used for conveying capabilities and configuration of the above features between neighbors to ensure consistent configuration across the network. This protocol is expected to leverage functionality provided by IEEE 802.1AB (LLDP).

More information on Data Center Bridging can be found at http://en.wikipedia.org/wiki/Data_Center_Bridging.

Although lossless Ethernet may have wider applications in the future, such as ISCSI. At this time, due to limited test exposure, EMC does not recommend simultaneous use of both FCoE and lossless iSCSI. Traditional iSCSI is fully supported in an FCoE environment, but not lossless iSCSI. As a result, only FCoE traffic will be lossless. TCP and UDP traffic will continue to be lossy on this infrastructure.

Priority Flow Control and PAUSE

Priority Flow Control (PFC) (802.1Qbb) enables PAUSE-like (802.3x) functionality on a per-Ethernet priority basis. PFC allows for lossless Ethernet connections to be created for a given priority within an otherwise lossy Ethernet network. As shown in Figure 7, priority 3 is being paused because the receive buffer hit a threshold. This is done by the receiver transmitting a PAUSE-ON frame. The PAUSE-ON frame contains the priority to be paused, as well as the number of quantas (512-bit increments) for the pause to remain in effect.

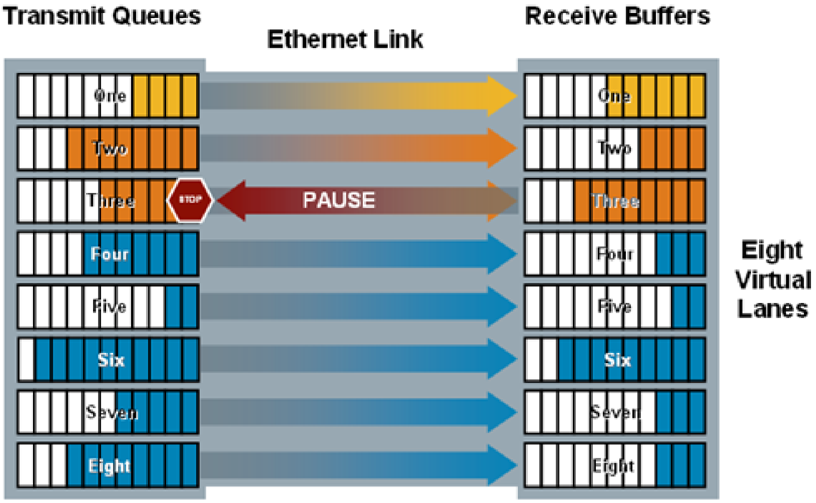


Figure 7 PFC and PAUSE example

Once the amount of data in the buffer dips below a certain threshold, either a PAUSE-OFF frame can be transmitted, or the number of quantas will expire and data will start to flow from the Transmit Queue to the Receive buffer.

As with any method of flow control, PFC does have limitations, but the most significant may be a distance limit. A distance limitation is due to the amount of buffering available on both of the CNA and FCF. In order for PFC to work properly, the receive buffer has to know the proper time to transmit a PAUSE ON. This requires the receive buffer to not only know how much data it contains, but to also predict the following:

- How much data is actually on the link
- How much additional data can be transmitted before a PAUSE ON frame from the receive buffer would actually reach the transmit queue and be processed

In order to calculate how much additional data could potentially be received, both the length of the link and the speed at which the link is operating must be known.

- Gen 1 CNAs imposed a maximum distance of 50 meters.
- Starting with Gen 2 CNAs, the maximum distance supported is determined by the physical media in use for the link.

Data Center Bridging eXchange

The Data Center Bridging Capability eXchange Protocol (DCBCXP) also known as DCBX, is a protocol that extends the Link Layer Discovery Protocol (LLDP) defined in IEEE802.1AB. For FCoE environments, DCBX allows the FCF to provide Link Layer configuration information to the CNA and allows both the CNA and FCF to exchange status.

In order for a CNA to successfully log in to the FCF, the DCBX protocol must be used. If for some reason DCBX was not being used by the CNA, or the CNA was not capable of accepting configuration information pushed to it from the FCF, the link would fail to initialize properly and the FC portion of the CNA would be unable to perform FLOGI. This typically will not be of any concern to users since DCBX is properly configured by default on CNAs and the FCFs.

DCBX frames

DCBX frames contain an LLDP PDU (Protocol Data Unit), which in turn consists of many Type Length Value (TLV) entries. Each TLV contains information for one configuration or status parameter. An example of the information contained with one of the TLVs is the Priority Flow Control Sub-TLV which allows for the exchange of Priority Flow Control (PFC) information. The exchange of this information allows for lossless ethernet for Ethernet frames with an FCoE Ether type.

The protocol starts when a physical connection has been established between the CNA and FCF. Both the CNA and FCF start to initialize DCBX by entering a state known as *fast initial LLDP retransmission*.

While in this state, each will transmit one DCBX Ethernet frame (ethertype 0x88CC) per second for five seconds. The purposes of these retransmissions are to allow the link to initialize faster than would otherwise be possible. Once the initial retransmissions have been performed, each side of the link periodically transmits status DCBX frames, either after a configurable time period or immediately after a change in the status of the link. When a DCBX frame is transmitted due to a status change, the sequence number is incremented by one.

Protocols

FCoE relies on the use of two different protocols:

- FCoE — Data plane protocol, discussed further in "FCoE encapsulation" on page 42.

  This protocol is data intensive and requires lossless Ethernet. It is typically implemented in hardware and is used to carry most of the FC frames and all the SCSI traffic.

- FIP (FCoE Initialization Protocol) — Control plane protocol, discussed further in "FCoE Initialization Protocol (FIP)" on page 44.

  This is not data intensive. It is typically implemented in software, and used to discover FCoE capable devices connected to an Ethernet network and to negotiate capabilities.

FCoE encapsulation

As the name implies and as shown in Figure 8, FCoE is literally an encapsulation of an FC frame inside an Ethernet frame. This section provides some details about that encapsulation, but the most important concept is that there is a one-to-one relationship between an FC frame and the Ethernet frame that encapsulates it. This is important because it means that FC frames are *never* segmented and transmitted as a part of multiple Ethernet frames.

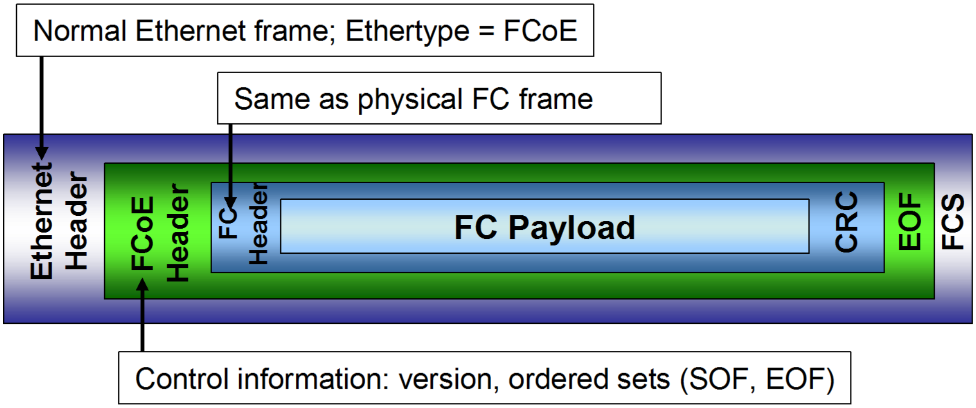

Normal Ethernet frame; Ethertype = FCoE
Same as physical FC frame
Ethernet Header
FCoE Header
FC Header
FC Payload
CRC
EOF
FCS
Control information: version, ordered sets (SOF, EOF)

Figure 8 FCoE encapsulation

The following are further discussed in this section:

- "FCoE frame size" on page 43
- "FCoE frame format" on page 43
- "FCoE frame mapping" on page 43

Refer to "FCoE Ethernet Frame example" on page 84 for an example.

FCoE frame size

The maximum field size of a FCoE frame is 2180 bytes. To support growth, FCoE requires that the Ethernet infrastructure supports frames up to 2.5 KB (baby jumbo frames).

FCoE frame format

FCoE encapsulates a Fibre Channel frame within an Ethernet frame. Figure 9 represents the frame format as agreed to by the INCITS T11.3 standards body.

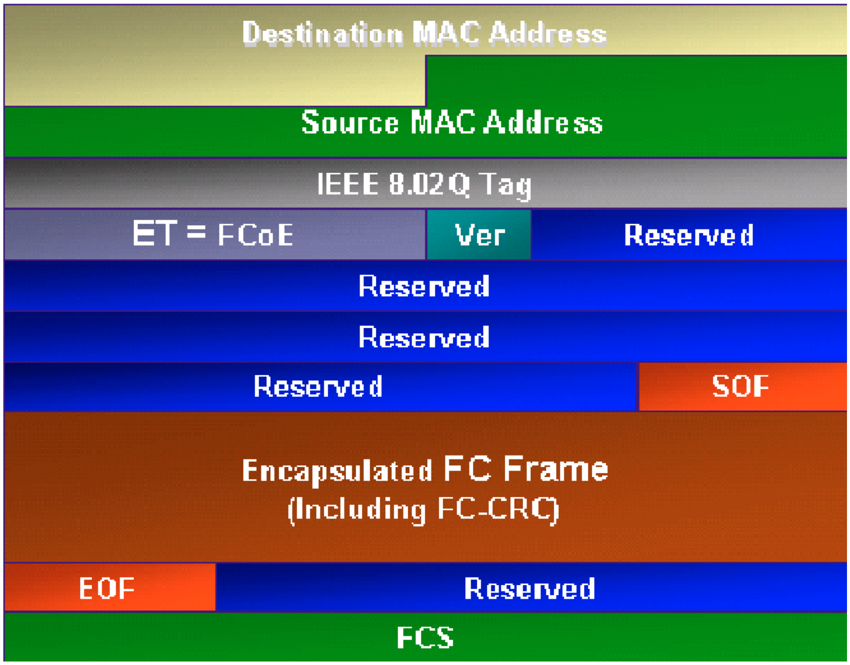


Figure 9 **FCoE frame format**

FCoE frame mapping

The encapsulation of the Fibre Channel frame occurs through the mapping of FC onto Ethernet. Fibre Channel and traditional networks have stacks of layers where each layer in the stack represents a set of functionality. The Fibre Channel stack consists of five layers, FC-0 through FC-4. Ethernet is typically considered a set of protocols in the seven-layer OSI stack that define the physical and data link layers. FCoE provides the capability to carry the FC-2 layer over the Ethernet layer, as shown in Figure 10 on page 44.

This allows Ethernet to transmit the upper Fibre Channel layers FC-3 and FC-4 over the IEEE 802.3 Ethernet layers. It is this FCoE mapping that allows FC traffic to pass over an Ethernet infrastructure.

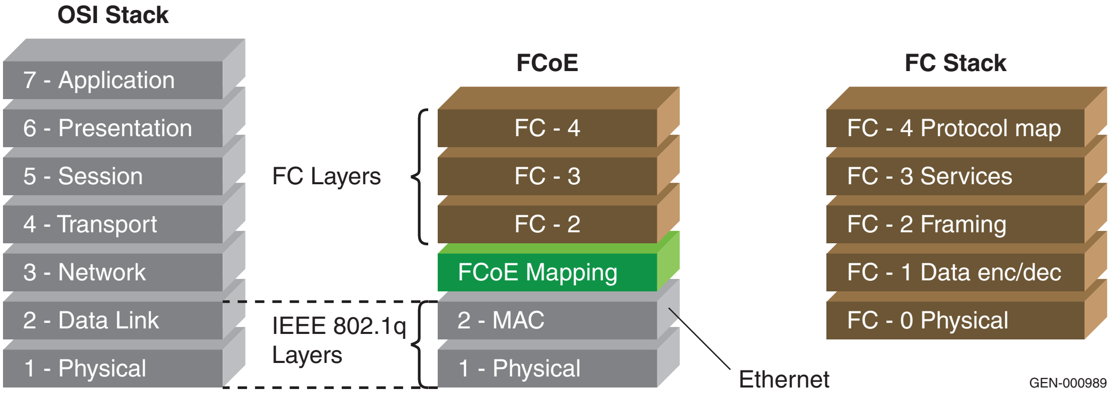


Figure 10 FCoE mapping

For additional information, refer to "Protocols" on page 42.

FCoE Initialization Protocol (FIP)

The FCoE Initialization Protocol (FIP) bridges the gap between the expectations of the Fibre Channel protocol and the reality of an Ethernet network.

The main goal of FIP is to discover and initialize FCoE capable entities connected to an Ethernet cloud. FIP uses a dedicated Ethertype, 0x8914.

This section contains the following information:

- "Overview" on page 45
- "FIP frame format" on page 47
- "FCoE VN_Port Virtual Link instantiation and FIP" on page 48
- "FCoE VE_Port Virtual Link instantiation and FIP" on page 54
- "FIP Link Keep Alive (LKA)" on page 57
- "FIP Clear Virtual Links (CVL)" on page 57

Note: More information on Fibre Channel over Ethernet is provided in Chapter 1, "Introduction to Fibre Channel over Ethernet."

Overview

The FCoE Initialization Protocol (FIP) is defined in FC-BB-5. FIP is used to not only for initialization functions such as discovering which Fibre Channel entities are available on a layer 2 Ethernet network and the creation of virtual links, but it is also used to verify the state of the virtual links and to destroy virtual links when there is a need to do so.

The role that FIP plays in both direct connect environments and CEE cloud environments (shown in Figure 11, "Direct Connect topology" and Figure 12, "CEE Cloud topology") is similar. However, while in a CEE cloud environment, FIP allows the lossless Ethernet switch(es) to perform FIP snooping.

FIP snooping is required in order to prevent man in the middle types of attacks by allowing the lossless Ethernet switches to Dynamically update ACLs and only allow the ENode that performed FIP to transmit frames with the FPMA (Fabric Provided MAC Address, see FC-BB-5) assigned to it.

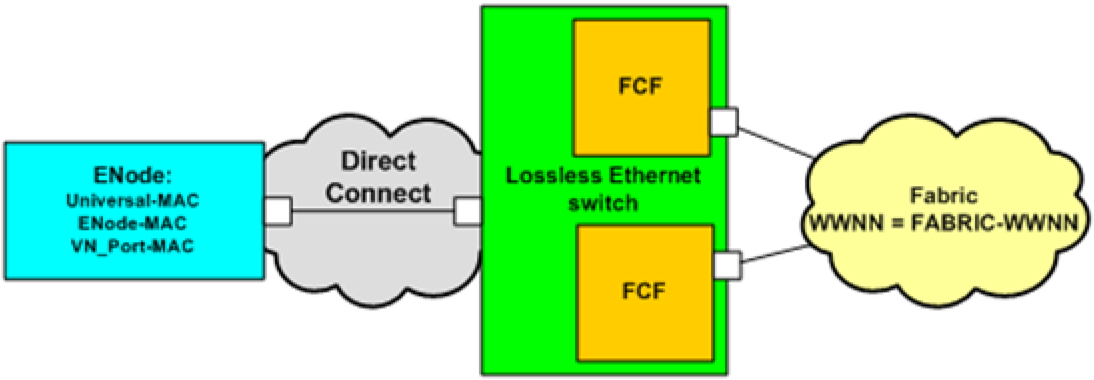


Figure 11 Direct Connect topology

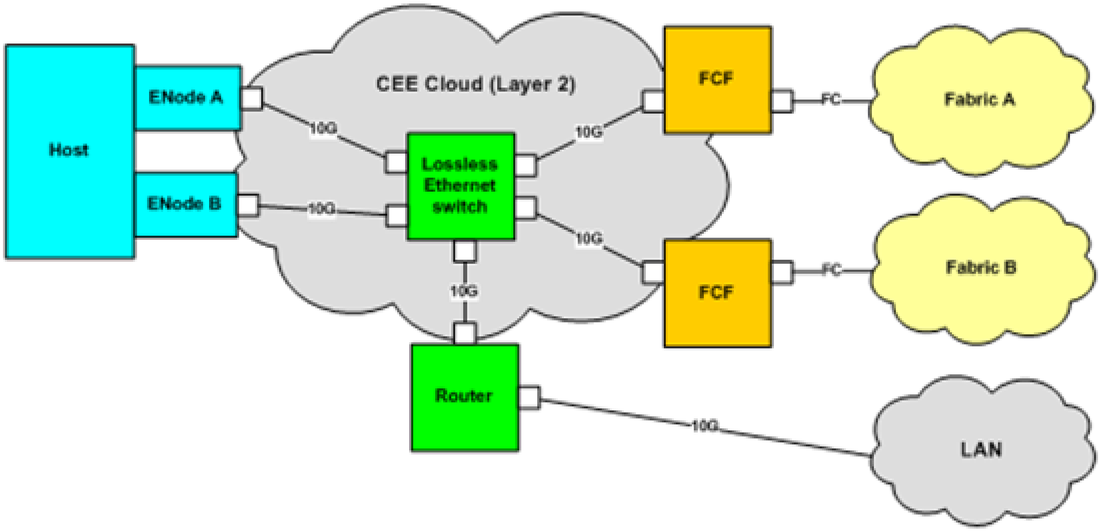


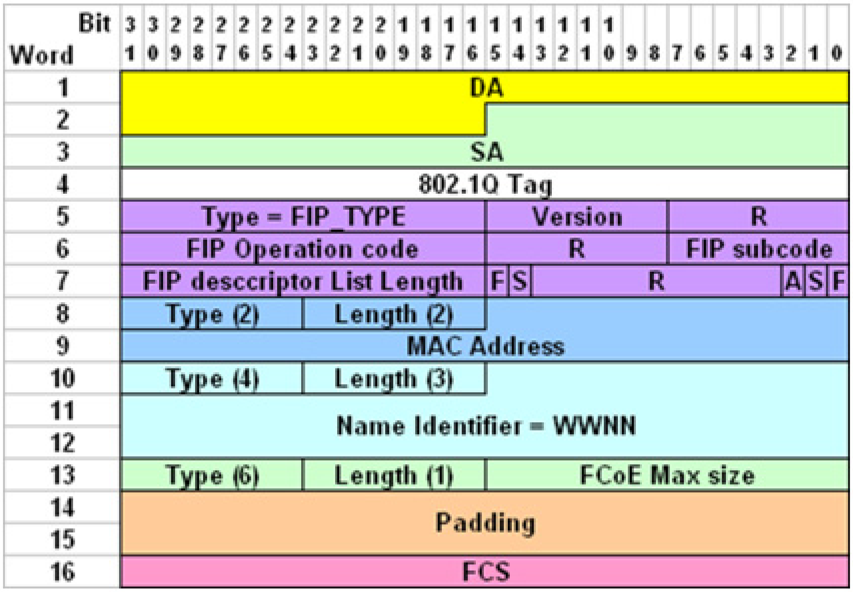
Figure 12 **CEE Cloud topology**

When an FCoE initiator or target initializes a virtual link, it is expected that it will do so in a certain order. The first thing that will need to be done is to discover which VLAN the FCoE services are being provided on. Next, the initializing port will need to discover which FCFs are available for login. Finally, FIP login shall be performed.

Once the link has been established, the FIP LKA (Link Keep Alive) protocol will be responsible maintaining that link.

FIP frame format

All of the protocols using FIP have the same basic frame format, as shown in Figure 13.



<table>
<tr><th>Bit</th><th>3</th><th>3</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>2</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th>1</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><th>Word</th><th>1</th><th>0</th><th>9</th><th>8</th><th>7</th><th>6</th><th>5</th><th>4</th><th>3</th><th>2</th><th>1</th><th>0</th><th>9</th><th>8</th><th>7</th><th>6</th><th>5</th><th>4</th><th>3</th><th>2</th><th>1</th><th>0</th><th>9</th><th>8</th><th>7</th><th>6</th><th>5</th><th>4</th><th>3</th><th>2</th><th>1</th><th>0</th></tr>
<tr><td>1</td><td colspan="32">DA</td></tr>
<tr><td>2</td><td colspan="16"></td><td colspan="16"></td></tr>
<tr><td>3</td><td colspan="32">SA</td></tr>
<tr><td>4</td><td colspan="32">802.1Q Tag</td></tr>
<tr><td>5</td><td colspan="16">Type = FIP_TYPE</td><td colspan="8">Version</td><td colspan="8">R</td></tr>
<tr><td>6</td><td colspan="16">FIP Operation code</td><td colspan="8">R</td><td colspan="8">FIP subcode</td></tr>
<tr><td>7</td><td colspan="16">FIP desccriptor List Length</td><td>F</td><td>S</td><td colspan="11">R</td><td>A</td><td>S</td><td>F</td></tr>
<tr><td>8</td><td colspan="8">Type (2)</td><td colspan="8">Length (2)</td><td colspan="16"></td></tr>
<tr><td>9</td><td colspan="32">MAC Address</td></tr>
<tr><td>10</td><td colspan="8">Type (4)</td><td colspan="8">Length (3)</td><td colspan="16"></td></tr>
<tr><td>11</td><td colspan="32" rowspan="2">Name Identifier = WWNN</td></tr>
<tr><td>12</td></tr>
<tr><td>13</td><td colspan="8">Type (6)</td><td colspan="8">Length (1)</td><td colspan="16">FCoE Max size</td></tr>
<tr><td>14</td><td colspan="32" rowspan="2">Padding</td></tr>
<tr><td>15</td></tr>
<tr><td>16</td><td colspan="32">FCS</td></tr>
</table>

Figure 13 FIP frame format

All FIP frames start with a DA, SA, an optional 802.1Q Tag and several other fields including the Ether type and ending with the FCF bit (Word 7 bit 0 / word 7 bit 31 in network order). At this point, the format of the frame changes depending upon the Operation code. However, all of the fields that follow the FCF bit are in a Type, Length, Value (TLV) format.

FCoE VN_Port Virtual Link instantiation and FIP

After the DCBX protocol has successfully completed, FCoE initiators or targets should begin the Virtual Link instantiation process by transmitting a FIP VLAN Request. See Figure 14.

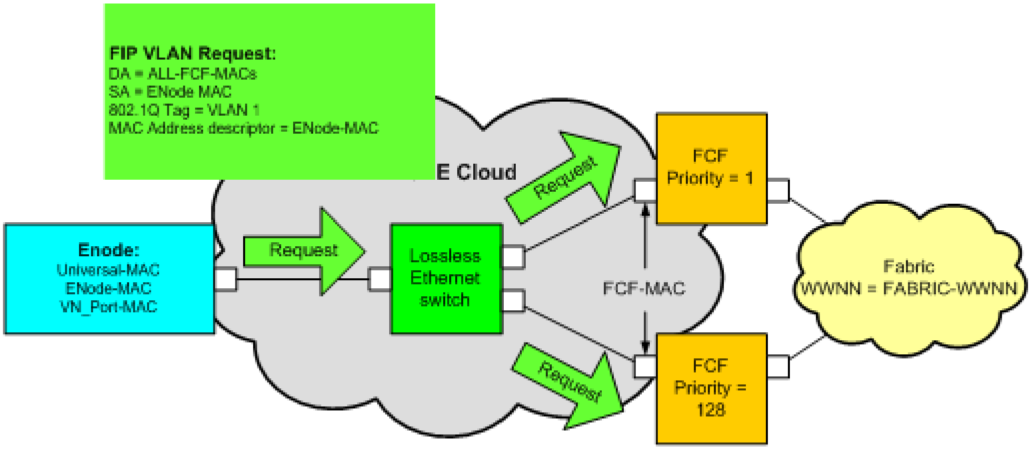


Figure 14 **FIP VLAN Request**

As shown in Figure 14, an ENode has transmitted a multicast FIP VLAN Request to the Multicast Destination Address (DA) of ALL-FCF-MACs. When this multicast frame is received by the switch, it is transmitted on all active switch interfaces other than the one that it was received on. An important point to note about this frame is that the 802.1Q Tag contains the default VLAN ID of 1. This is done because the standard specifies that all Fibre Channel Forwarders (FCFs) should be listening on VLAN 1 for FIP VLAN Requests.

When the FCF receives the FIP VLAN Request, it will respond with a unicast FIP VLAN Notification, as shown in Figure 15 on page 49.

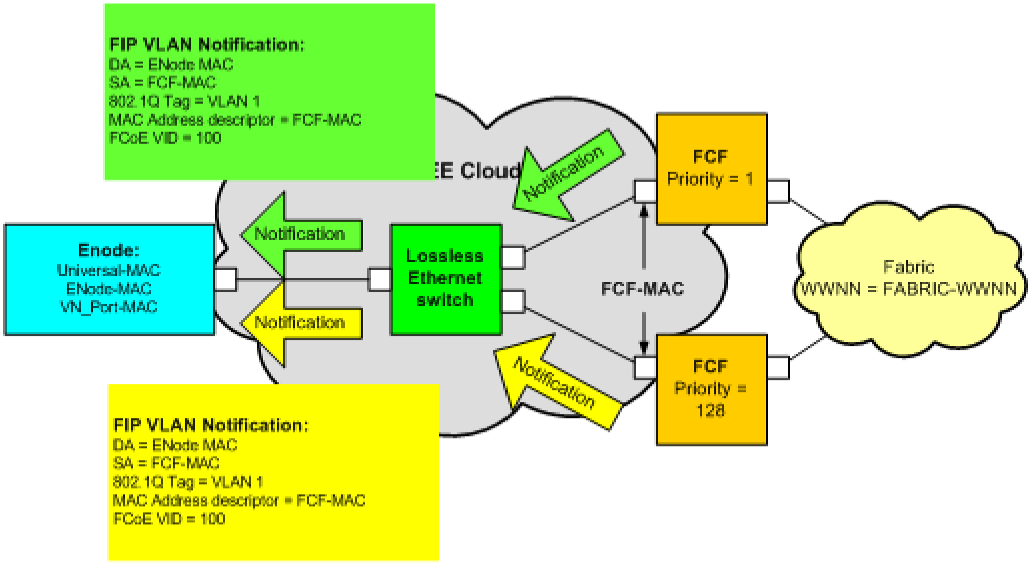


Figure 15 FIP VLAN Notification

As shown in Figure 15, each FCF responds to the FIP VLAN Request with a unicast FIP VLAN Notification. Notice that the 802.1Q Tag is set to the same value as used in the FIP VLAN Request. This is done because the standard mandates that the Notification be transmitted on the same VLAN the Request was received on. Another important point to make note of is that the FCoE VID (VLAN ID) TLV is included in the Notification. There can be several VLANs listed in the Notification and it is up to the ENode to decide how to handle this case should it arise. Ideally, the ENode would transmit a FIP Discovery on each VLAN returned in the Notification.

After the ENode receives the FIP VLAN Notification, it should transmit a multicast FIP Solicitation, as shown in Figure 16.

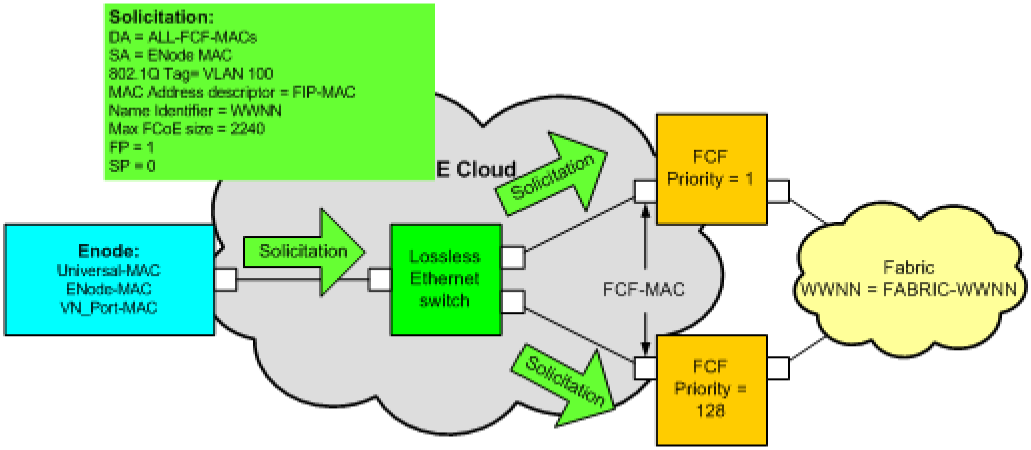
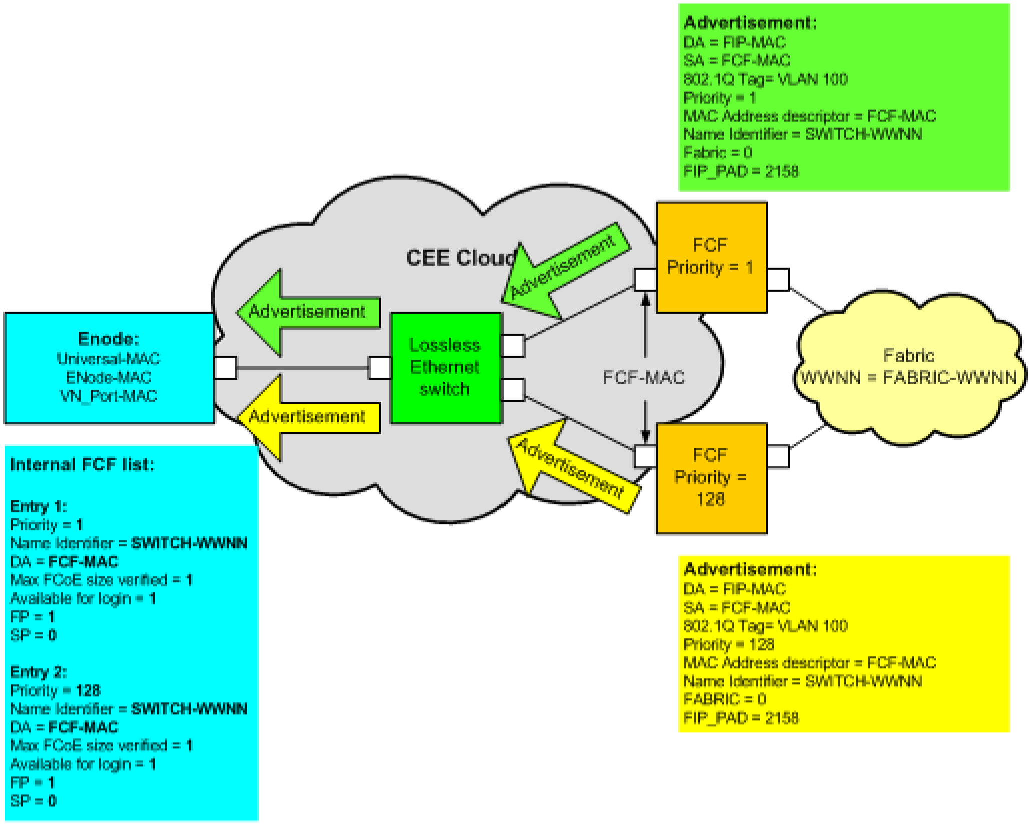


Figure 16 **FIP Solicitation**

One multicast FIP Solicitation should be transmitted on each VLAN that was returned in the FIP VLAN Notification. As shown in Figure 16, since only one VID (100) was returned in the FIP VLAN Notification, the ENode will transmit only one multicast FIP Solicitation. Note that the 802.1Q tag is set to the value of the VID returned in the FIP VLAN Notification. Other important features of the FIP Solicitation are the following fields.

- Max FCoE size — The maximum length FCoE frame that is supported by the ENode.
- FP — Fabric Provided MAC Address support (Boolean) — Setting this bit to one indicates that the ENode allows the Fabric to specify what VN_Port-MAC will be used by the ENode.
- SP — Server Provided MAC Address support (Boolean) — Setting this bit to one indicates that the ENode is capable of specifying its own VN_Port-MAC Address.

Upon receiving the FIP Solicitation, each FCF shall respond with a unicast FIP Advertisement as shown in Figure 17.



Advertisement:
DA = FIP-MAC
SA = FCF-MAC
802.1Q Tag= VLAN 100
Priority = 1
MAC Address descriptor = FCF-MAC
Name Identifier = SWITCH-WWNN
Fabric = 0
FIP_PAD = 2158

Enode:
Universal-MAC
ENode-MAC
VN_Port-MAC

Internal FCF list:

Entry 1:
Priority = 1
Name Identifier = **SWITCH-WWNN**
DA = **FCF-MAC**
Max FCoE size verified = 1
Available for login = 1
FP = 1
SP = 0

Entry 2:
Priority = 128
Name Identifier = **SWITCH-WWNN**
DA = **FCF-MAC**
Max FCoE size verified = 1
Available for login = 1
FP = 1
SP = 0

Advertisement:
DA = FIP-MAC
SA = FCF-MAC
802.1Q Tag= VLAN 100
Priority = 128
MAC Address descriptor = FCF-MAC
Name Identifier = SWITCH-WWNN
FABRIC = 0
FIP_PAD = 2158

Figure 17 FIP Advertisement

The purpose of the FIP Advertisement is to notify the ENode of the available FCFs that can support Login. It is also responsible for ensuring that the data path being used is capable of handling a full size FCoE Frame. The FIP Advertisement contains several important fields:

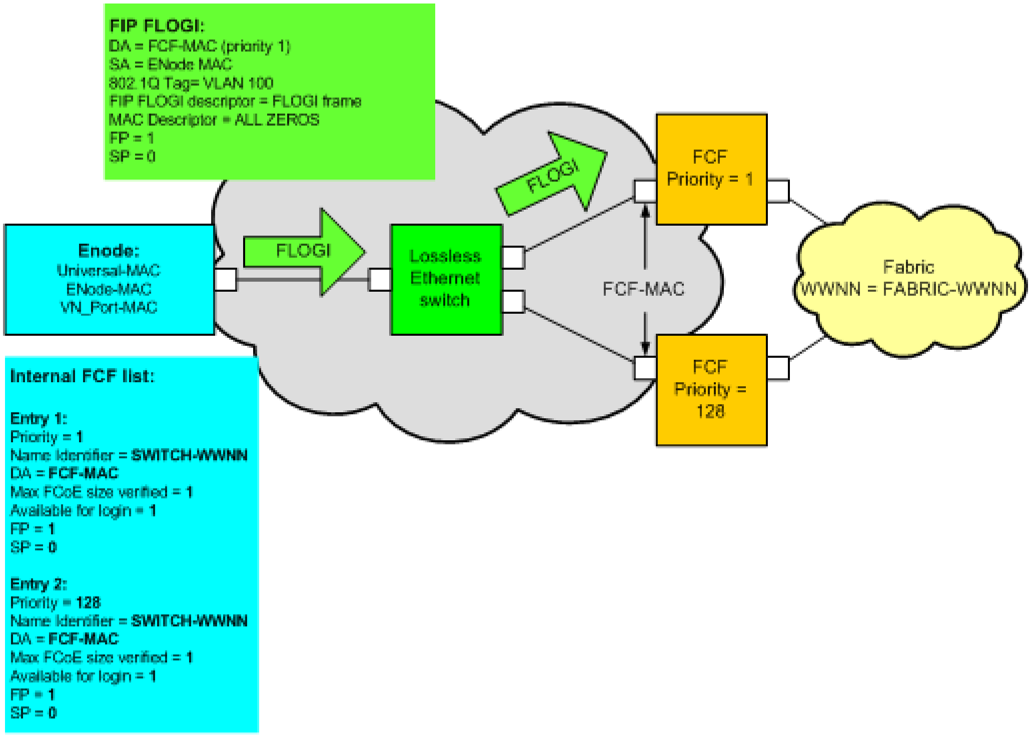
- Priority — As the name implies, the Priority field indicates the Priority that has been manually assigned to the FCF. The purpose manually configuring the priority is to allow a Network/SAN Administrator to indicate a preference for where Login should be performed. How this field is used is explained in greater detail below.

- Name Identifier — The Name Identifier is the (World Wide Node Name) WWNN of the Fabric that the FCF is either attached to or participating in. This field allows the ENode to determine which Fabric the Advertisement was received from.
- Fabric — The Fabric ID is a manually configured field and specifies the FC-MAP value for the responding FCF.
- FIP_PAD — As shown in the FIP frame format (Figure 13 on page 47), a pad field exists at the end of the FIP frame. The purpose of the pad field is to allow for a frame to always meet the minimum Ethernet Frame Length of 64 bytes. In the case of a Solicited Discovery Advertisement (i.e., when the FCF is transmitting an Advertisement in response to a Solicitation), the FIP_Pad field shall be set to the length required to create an 802.3 frame with a payload length that matches the Max_FCoE_Size field value in the Max FCoE Size descriptor in the received Discovery Solicitation. The FIP_Pad field values shall be set to reserved. For an unsolicited Discovery Advertisements, the FIP_Pad field shall be of zero length (i.e., not present).

When the ENode receives an Advertisement from a previously unknown FCF, it will add an entry to its Internal FCF list. The list contains several important pieces of information:

- Priority — The Priority returned from the FCF in the Solicited Discovery Advertisement.
- Name Identifier — The Name Identifier returned from the FCF in the Solicited Discovery Advertisement.
- DA — The DA of the FCF
- Max FCoE size verified — This is a bit that will be set to one if the length of the Solicited Discovery Advertisement is equal to the MAX FCoE size specified in the FIP Solicitation.
- Available for login — This bit is set to one if the FCF indicated that it supported login in the Solicited Discovery Advertisement.
- FP — This bit indicates if the FCF indicated support for Fabric Provided MAC Addresses in the Solicited Discovery Advertisement.
- SP — This bit indicates if the FCF indicated support for Server Provided MAC Addresses in the Solicited Discovery Advertisement.

Once a Solicited Discovery Advertisement has been received from an FCF and the MAX FCoE size has been verified, the ENODE may perform Fabric Login with the FCF. The Fabric Login request is transmitted in a FIP Frame as seen in Figure 18.



FIP FLOGI:
DA = FCF-MAC (priority 1)
SA = ENode MAC
802.1Q Tag= VLAN 100
FIP FLOGI descriptor = FLOGI frame
MAC Descriptor = ALL ZEROS
FP = 1
SP = 0

Internal FCF list:

Entry 1:
Priority = **1**
Name Identifier = **SWITCH-WWNN**
DA = **FCF-MAC**
Max FCoE size verified = **1**
Available for login = **1**
FP = **1**
SP = **0**

Entry 2:
Priority = **128**
Name Identifier = **SWITCH-WWNN**
DA = **FCF-MAC**
Max FCoE size verified = **1**
Available for login = **1**
FP = **1**
SP = **0**

Figure 18 FIP FLOGI

If two or more FCFs are present from the same Fabric (as shown in Figure 18), the ENode will transmit the FLOGIto the FCF with the higher priority (lower value). The ENode is capable of determining what FCFs are attached to the same Fabric by using the Name Identifier field. The FLOGI payload used in the FIP FLOGI is identical to the FLOGI payload from a native FC environment.

If two or more FCFs are present from different Fabrics, the ENode should transmit an FLOGI to each FCF.

When the FCF receives the FLOGI request it should transmit a FIP FLOGI Accept as shown in Figure 19.

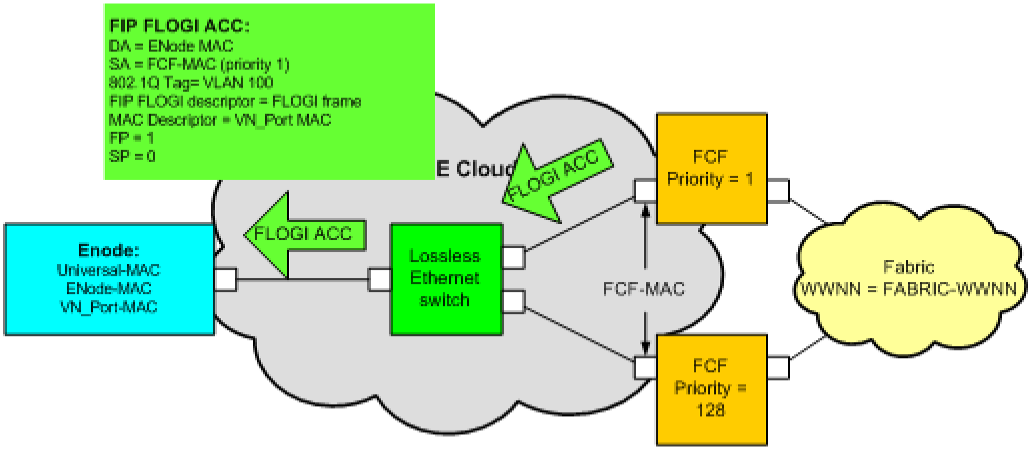


Figure 19 **FIP FLOGI ACC**

The FLOGI ACC payload used in the FIP FLOGI ACC is identical to the FLOGI ACC payload from a native FC environment.

After the FLOGI ACC has been received, the rest of the Login process (e.g., Name Server registration) may continue but the FCoE and not the FIP frame format will be used.

FCoE VE_Port Virtual Link instantiation and FIP

Note: This section describes the FCoE VE_Port Virtual Link Instantiation process currently being used between Cisco Nexus products. The reason for this is that Cisco is the only FCF vendor currently providing this functionality. When other vendors provide VE_Port functionality, this section will be updated to show the differences, if any.

The virtual links used to support the instantiation of Virtual E_Ports (VE_Ports) are instantiated in a manner that is similar to the process used to instantiate a virtual link that supports the instantiation of a Virtual F_Port (VF_Port) as described in "FCoE VN_Port Virtual Link instantiation and FIP" on page 48. One major difference is that FIP VLAN discovery is not used. For the sake of this example, the topology shown in Figure 20 on page 55 will be used.

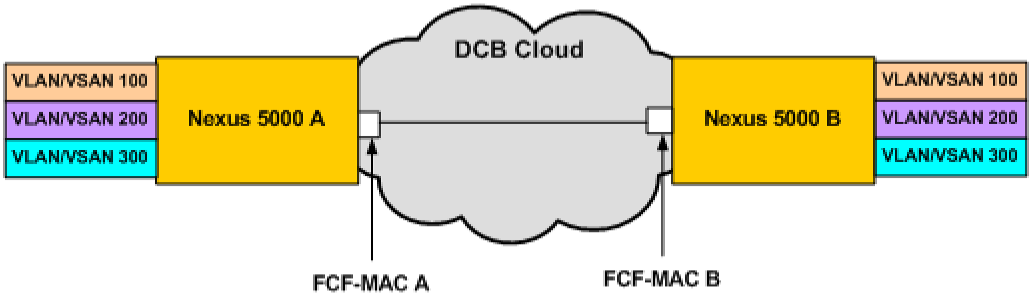
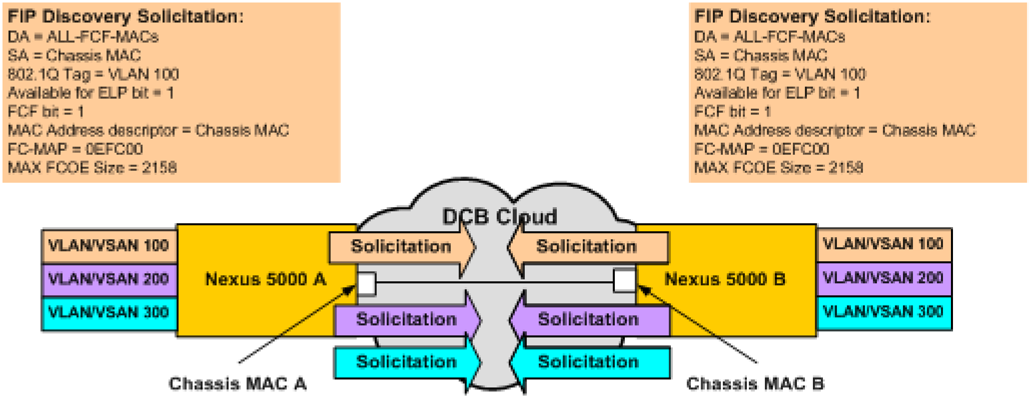


Figure 20 Topology example

The process starts with both sides of the link transmitting DCBX frames. Once the DCBX parameters have been exchanged, both sides will transmit Solicitations on every VLAN that FCoE has been allowed on. In this example, it is assumed that the Ethernet / vFC interface on the Nexus has been configured to allow all VLANs / VSANs. In Figure 21, only the details for the Solicitation frames for VLAN/VSAN 100 are shown. The Solicitations for VLAN/VSAN 200 and 300 would contain similar information.



FIP Discovery Solicitation:
DA = ALL-FCF-MACs
SA = Chassis MAC
802.1Q Tag = VLAN 100
Available for ELP bit = 1
FCF bit = 1
MAC Address descriptor = Chassis MAC
FC-MAP = 0EFC00
MAX FCOE Size = 2158

FIP Discovery Solicitation:
DA = ALL-FCF-MACs
SA = Chassis MAC
802.1Q Tag = VLAN 100
Available for ELP bit = 1
FCF bit = 1
MAC Address descriptor = Chassis MAC
FC-MAP = 0EFC00
MAX FCOE Size = 2158

Figure 21 FIP Discovery Solicitation frames

The Solicitations are transmitted to the multicast Destination Address (DA) of ALL-FCF-MACs. The SA of these frames is the Chassis MAC of the Nexus. The 802.1Q tag will be set to the VLAN that the Solicitation is being performed on.

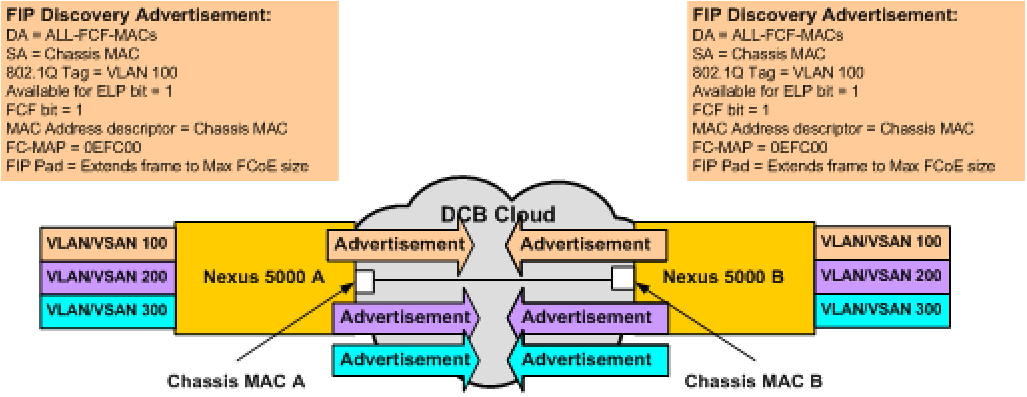
The Available for ELP bit will need to be set to **1** in order for FIP to proceed to the next phase (FIP Advertisement).

The FCF bit indicates that the Frame was transmitted by an FCF.

The FC-MAP is checked by both sides to ensure that this value matches. If it does not, FIP will not proceed to the next phase and the virtual link will not be instantiated. The FC-MAP value prevents unintentional FC fabric merges and should be administratively set to a value other than the default if you have multiple FCoE fabrics in the same data center and you do not want an accidental connection between two FCFs to result in a fabric merge.

The **Max FCoE Size** field is set to the maximum size FCoE frame supported by each side.

If the Available for ELP bit and FC-MAP values match on both sides of the link, the FIP process will continue with both sides transmitting a Discovery Advertisement, as shown in Figure 22.



FIP Discovery Advertisement:
DA = ALL-FCF-MACs
SA = Chassis MAC
802.1Q Tag = VLAN 100
Available for ELP bit = 1
FCF bit = 1
MAC Address descriptor = Chassis MAC
FC-MAP = 0EFC00
FIP Pad = Extends frame to Max FCoE size

FIP Discovery Advertisement:
DA = ALL-FCF-MACs
SA = Chassis MAC
802.1Q Tag = VLAN 100
Available for ELP bit = 1
FCF bit = 1
MAC Address descriptor = Chassis MAC
FC-MAP = 0EFC00
FIP Pad = Extends frame to Max FCoE size

Figure 22 FCoE Discovery advertisement

The information contained within the FIP Discovery Advertisement is similar to what is contained in the Solicitation. One difference between Cisco's implementation and FC-BB-5 is that the Advertisement is supposed to be padded so that the frame is equal to the Max FCoE size. In the traces captured, EMC did not see this happen.

Once both sides have received and validated the information contained within the Advertisements, the virtual link can be instantiated by both sides transmitting ELP on each VLAN/VSAN that supports FCoE. The frames exchanged to complete the VE_Port instantiation are practically identical to those frame used to initialize an FC E_Port and will not be repeated here. For more information, refer to the *Networked Storage Concepts and Protocols TechBook* available through the E-Lab Interoperability Navigator, **Documents> Topology Resource Center**, at http://elabnavigator.EMC.com.

FIP Link Keep Alive (LKA)

Historically, in a native FC SAN, the link between an N_Port and an F_Port was typically a strand of fiber. However, with the introduction of NPIV Gateways and some distance extension technologies (e.g., SONET), this one to one relationship between the physical link and the logical link between an N_Port and F_Port changed. When this change occurred, a method needed to be developed to ensure that although there could be multiple physical links and protocols that made up a logical link between an N_Port and an F_Port, the link needed to appear to have the same characteristics of a single physical link for backward compatibility reasons. The method that was designed and that has been in use for the past few years is Link Keep Alive (LKA).

FCoE introduces additional changes to the concept of a link by allowing for a CEE cloud to exist between the VN_Port and the VF_Port. Once a Virtual Link is established, after the completion of the FIP FLOGI, the link must be periodically checked to ensure that the device on the other end of the virtual link is still responding. If it is not, the logical link must be torn down and an RSCN sent.

The default for FCoE is that if two sequential LKAs are missed, the link will be torn down. The LKA interval is configurable but has a default value of 4 seconds.

FIP Clear Virtual Links (CVL)

Clear Virtual Links (CVL) allow for a Virtual Link to be torn down without the need to wait for 2 LKAs to be missed. An example of when this might be done would be when the FC functionality on an FCF was disabled for some reason. In this case, as soon as the FC functionality was disabled, the FCF would send out a CVL to any VN_Ports using the FCF.

Physical connectivity options for FCoE

There are two options available for physical connectivity when using FCoE, each discussed further in this section:

- "Optical (fiber) cable" on page 58
- "Twinax (copper) cable" on page 59

Each option has benefits and limitations to consider.

Optical (fiber) cable

You can use optical connections for any FCoE link.

If you are currently using FC or 10 GbE, you are familiar with this type of cable. Shown in Figure 23 as an LC connector, this cable is available in several different diameters and bandwidth distance product (BDP) ratings, as listed in Table 2.

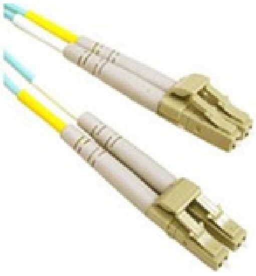


Figure 23 LC connector

Table 2 Multimode media maximum supported distances

Protocol	Transceiver type	Speed	62.5 µm/200MHz* km (OM1) [62.5 micron]	50 µm/500 MHz* km (OM2) [50 micron]	50 µm/2000 MHz* km (OM3) [50 micron]	50 µm /47000 MHz* km (OM4) [50 micron]
GbE	SW	1 Gb	300m	550m	1000m	1000m *
		10 Gb	33m	82m	300m	550m
		40 Gb	N/A	N/A	100m	125m

* Denotes at least this distance. No documented distance is available at this time.

The bandwidth distance product (BDP), measured in megahertz.km, is an indication of the overall quality of the fiber. The higher the number, the more data it can carry at a given distance and, hence, the more expensive it will be.

Due the effects of dispersion, as the speed of the transmission increases, the maximum distance supportable by each type of cable decreases until, in some cases, the distance becomes too short to provide any meaningful distance and that media type can no longer be supported at a given speed (e.g., 40 Gb and OM1 as shown in Table 2 on page 58).

Benefits include:

- Interoperable — If the other side supports optical cable, the two ports will inter-operate from a physical connectivity perspective.
- Longer distance — OM4 allows for 10 G connections up to 550m as opposed to a maximum distance of 10m with twinax.
- Physically smaller in size — The smaller size allows the cable to be easily pushed to the side of a cabinet to ensure that airflow is not restricted.
- Can use any vendor's cable.

Limitations include:

- Expensive — Up to 10 times the cost of twinax.
- Uses more power than twinax.

Twinax (copper) cable

Typically, twinax is used between the server and the Top of Rack (ToR) switch.

Unlike fiber optic cable, twinax uses two copper conductors to pass electrical signals between the two cable ends, as shown in Figure 24 on page 60.

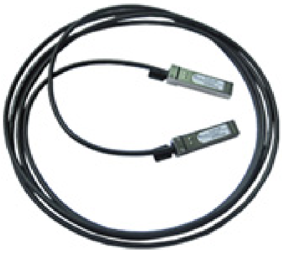


Figure 24 **Twinax cable with integrated SFP+**

Although twinax is much less expensive than optical fiber, it is susceptible to some interoperability constraints.

Benefits include:

- Less expensive than optical
- Uses less power than optical

Limitations include:

- Limited in distance
- Interoperability concerns

Logical connectivity options

Since the release of the first FCoE "Direct Connect" ToR configuration, the topologies supported by FCoE have been continuously evolving. This evolution has resulted in numerous logical connectivity options, the most recent being VE_Ports. This section provides an overview of FCoE fabrics and VE_Ports and then provides an list of rules that can be used to derive the set of topologies that are supported by FCoE today:

- "FCoE fabrics" on page 61
- "Although it is not shown inFigure 26, the FCoE ISLs must be on physically separate links rather than trunked on the Ethernet uplinks used to carry non-FCoE traffic. This is not a requirement from an FCoE protocol perspective, but it is required when using the current version of NX-OS. This requirement may eventually be removed." on page 64

FCoE fabrics

Virtual E_Ports (VE_Ports) allow for the formation of FCoE ISLs and the creation of an all FCoE fabric. Once a link between two FCFs has been established using FIP, VE_Ports and FCoE ISLs are initialized using the same protocol that is used to initialize E_Ports and FC ISLs. The FIP protocol used to establish FCF to FCF connections is described in "FCoE Initialization Protocol (FIP)" on page 44.

From a logical connectivity point of view, an all-FCoE fabric is identical to an FC fabric and supports all of the same functionality including zoning, a distributed name server, and RSCNs. As a result, the same types of scalability limits apply to both FC and FCoE fabrics, such as the maximum number of hops, VN_Ports, and domains.

From a physical connectivity point of view, the connectivity options currently supported for an all-FCoE fabric are shown in Figure 25. Each connectivity option is further explained following the figure.

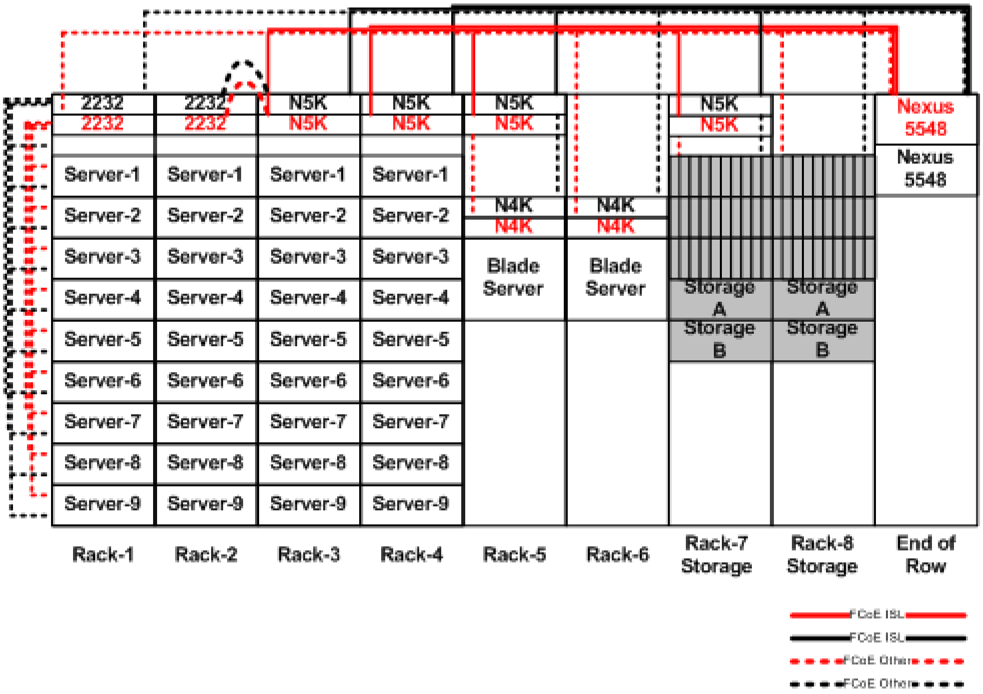


Figure 25 All-FCoE fabric example

For the sake of clarity, Figure 25 on page 62 shows the physical connectivity options that can be used within a row of equipment racks. This figure is not intended to indicate a limitation in the types of topologies that are supported; rather, it is used to help highlight all of the different possibilities. Each Rack is described in detail below.

End of Row

The End of Row (EoR) cabinet contains the aggregation layer switches. In order to maintain the highly-available characteristics of FC, two Cisco Nexus 5548s are shown and they have not been connected together via FCoE ISLs. his allows for the two fabrics to remain logically isolated. The Nexus 5548 switches may be part of the same vPC domain. For more information on vPC, refer to the"Virtual PortChannel" section in the "Nexus Series Switches Setup Examples" chapter in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging*

(DCB) Case Studies TechBook at http://elabnavigator.EMC.com, under the **Topology Resource Center** tab.

The Nexus 5548 switches in the End of Row cabinet may have:

- VE_Port (FCoE ISL) connections from other Nexus 5000 switches
- VF_Port (ENode) connections to FCoE initiators and targets
- Native FC E_Ports (ISLs)
- F_Port (FC initiator and target) connections to it
- Uplinks from an FSB
- Fabric Ports to connect to the Nexus 2232 FEX module

Rack-7 and Rack-8 Storage

Storage ports (either FC or FCoE) can be connected to a Top of Rack (ToR) switch or directly back to the Nexus 5548 in the EoR. Connecting a storage port directly to the top of rack switch in a given rack makes sense if the storage port is accessed primarily by servers residing in that rack. However, if a storage port is going to be accessed by many servers in different racks, the optimal placement of the storage ports may be on the Nexus 5548 at the EoR rack.

Rack-5 and Rack-6 FIP Snooping Bridge (FSB)

FIP Snooping Bridges, such as the Cisco or IBM 4001i, can either be connected to the Nexus 5000 at the top of the rack or connected to the Nexus 5548 at the EoR.

Rack-4 Nexus 5000 (VE_Ports)

The Nexus 5000 can be connected to the Nexus 5548 in the EoR while running in FC-SW mode via VE_Ports over FCoE, or over FC while running NPV and FC-SW modes. Currently, you cannot utilize FCoE for uplinks to the Nexus 5548 at the EoR while running in NPV mode.

Rack-2 and Rack-3 Nexus 5000 (VE_Ports) and Nexus 2232 (Fabric Ports)

The Nexus 2232 can be connected to a Nexus 5000 via fabric ports and then the Nexus 5000 can be connected to the Nexus 5548 at the EoR via E_Ports.

Rack 1 Nexus 2232 (Fabric Ports)

The Nexus 2232 at the ToR can be connected to the Nexus 5548 at the EoR via Fabric ports.

A logical representation of the physical topology is shown in Figure 26.

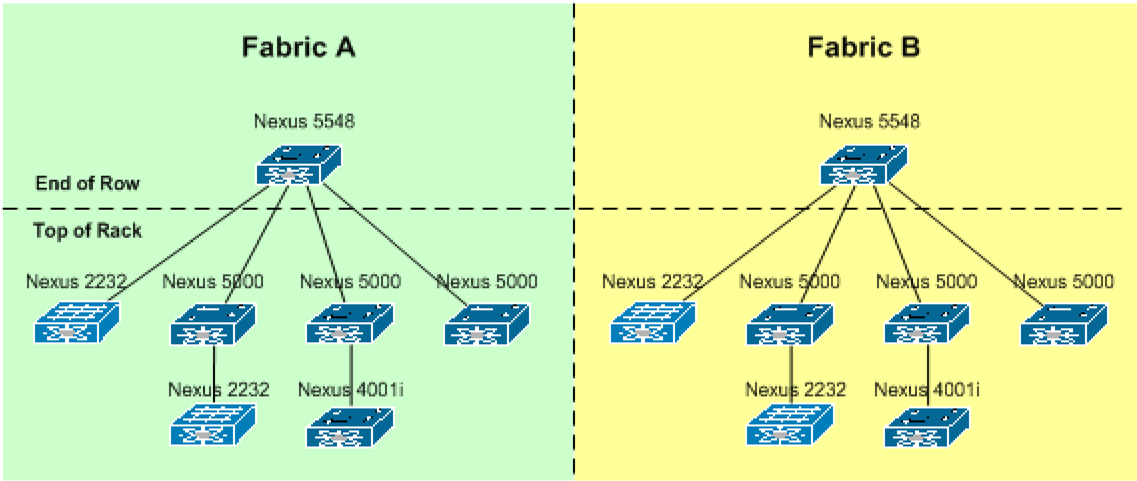


Figure 26 Physical topology example

Although it is not shown inFigure 26, the FCoE ISLs must be on physically separate links rather than trunked on the Ethernet uplinks used to carry non-FCoE traffic. This is not a requirement from an FCoE protocol perspective, but it is required when using the current version of NX-OS. This requirement may eventually be removed.

2

Ethernet Basics

This chapter provides a basic understanding of the various aspects and protocols involved with a typical Ethernet environment. Ethernet incorporates many different components and protocols to make it run successfully. The following information is included:

Ethernet history

Ethernet refers to of a family of standards defining local area networks (LANs) in terms of the Media Access Control (MAC) functions of the Data Link layer and the cabling and signaling functions of the Physical layer. Ethernet is typically used for network communications in a local area network where speed and security are paramount and distances are relatively short.

The original format for Ethernet was originally developed in 1972 at Xerox Palo Alto Research Center by Bob Metcalfe and David Boggs. Metcalfe described Ethernet in a memo as interconnecting advanced workstations, making it possible to send data from one station to one another and to high-speed printers. Ethernet first ran at 3 Mb/s and had eight-bit destination and source address fields, unlike the MAC Addresses fields used today. The design was based on an earlier experiment in networking called the *Aloha* network. This project began at the University of Hawaii in the late 1960s when Norman Abramson and his colleagues developed a radio frequency based network for communication among the Hawaiian Islands.

This section shows the development of the Ethernet through:

- "Communication modes of operation" on page 66
- "Ethernet devices" on page 69
- "Auto-negotiation" on page 70
- "Gigabit Ethernet" on page 71
- "10GBaseT" on page 71
- "40GbE technology" on page 72

Communication modes of operation

The history of the Ethernet can be briefly described in relation to the improvements in communication modes of operation: simplex, half duplex, and full duplex, as follows.

Simplex

The Aloha protocol used a *simplex* mode of communication, meaning a channel is always one way and information can only be sent in one direction. If an acknowledgment was not received within a period of time, the host assumed that another host had transmitted data simultaneously, causing a collision and corrupting the data or the response and preventing the sender from receiving an acknowledgment from the receiver. Upon detecting a collision, both transmitting hosts would choose a random back-off time and then retransmit the frames with a good probability of success. However, as volumes of traffic increased on the Aloha network, the collision rate quickly increased as well.

Half duplex

To improve the Aloha network, Metcalfe developed a new system that included a mechanism that detected when a collision occurred (collision detect). Hosts or stations listened for activity before transmitting and supported access to a shared channel by multiple stations. The mode of communication was *half duplex*, meaning data was now sent in both directions between two nodes, but only one direction could transmit on the link at a time.

Metcalfe left Xerox in 1979 and formed 3Com. He convinced Digital Equipment Corp., Intel, and Xerox to work together to promote Ethernet as a standard, (DIX). The DIX standards defined a *thick* Ethernet system (10Base5), based on a 10 Mb/s Carrier Sense Multiple Access with Collision Detect (CSMA/CD) protocol. It was known as *thick* because of the thick coaxial cable used to connect devices on the network. The first standard draft was published in 1980 within IEEE.

The CSMA/CD protocol can be better understood by defining its parts:

- *Carrier Sense* refers to the process of listening before speaking. Any host wishing to communicate listens for any active communication on the media. If communication exists it means the cable is in use and the host must wait to transmit.
- *Multiple Access* is the term for hosts in an Ethernet network attaching to the shared medium and having the opportunity to transmit. No host has any special privilege or priority over any other host in the network. Nevertheless, they do need to take turns per the access algorithm.
- *Deferral* or *back-off* counters refers to Ethernet hosts maintaining a counter of how often they need to wait before they can transmit. If the deferral counter exceeds a threshold value of 15 retries, the host attempting to transmit assumes that it will never get access to the cable to transmit the data frame. In this situation, the source host discards the frame.

  This could happen if there are too many hosts on the shared medium, meaning there is not enough bandwidth available on the network. If this situation continues, network re-design, i.e., LAN segmentation, can resolve the issue. If the link's power level exceeds a certain threshold, it implies to the system that a *collision* occurred.

An example is shown in Figure 27. When hosts detect that a collision occurs, all the hosts generate a collision enforcement signal.

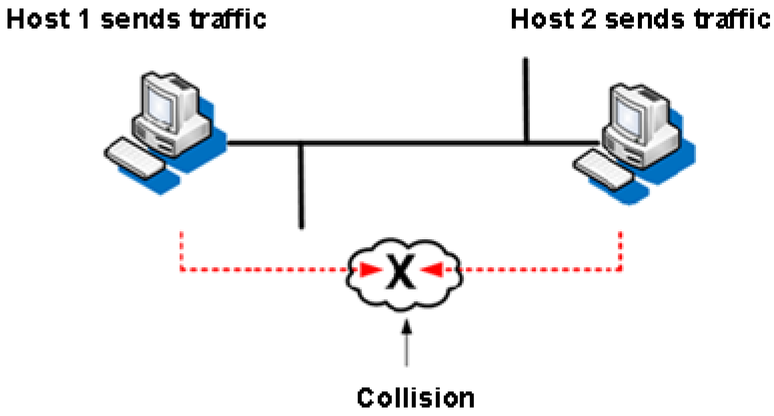


Figure 27 **Collision example**

The enforcement signal lasts as long as the smallest frame size, which in Ethernet is 64 bytes. This ensures that all hosts know about the collision and that no other host attempts to transmit during the collision event. If a host experiences too many consecutive collisions, the host stops transmitting the frame and informs the user (through an error message) that it was unable to send data.

Early Ethernet versions competed with Token Ring and Token Bus, two big proprietary technologies at that time. The International Organization for Standardization (ISO) was instrumental for Ethernet success. Proprietary systems soon found themselves buried under the ubiquity of the Ethernet. In the process, 3Com became a major company and built the first 10 Mb/s Ethernet NIC in 1983. This was quickly followed by Digital Equipment's Unibus to Ethernet adapter. Twisted-pair Ethernet systems, like 10BaseT, have been developed since the mid-1980s, replacing the early coaxial-based Ethernet implementations. Carrier Sense Multiple Access Collision Detect (CSMA/CD)

A major advance in Ethernet standards came with introduction of the IEEE 802.3i 10Base-T standard in 1990. It permitted 10 Mb/s Ethernet to operate over simple Category 3 Unshielded Twisted Pair (UTP) cable. The widespread use of UTP cabling in existing buildings created a high demand for 10Base-T technology. 10Base-T also permitted the network to be wired in a "star" topology that made it much easier to install, manage, and troubleshoot. These advantages led to a vast expansion in the use of Ethernet.

The IEEE improved the performance of Ethernet technology by a factor of 10 when it released the 100 Mb/s 802.3u 100Base T standard in 1995, commonly known as *Fast Ethernet*.

Full duplex

In 1997, the IEEE standards committee incorporated a feature called the *full duplex* mode of operation, allowing communication to occur in both directions simultaneously. In full duplex Ethernet, the links operate by using two physical pairs of wires wherein one pair is used for receiving data and one pair is used for sending data to a directly-connected device. This technology helps to maximize the bandwidth of the link and eliminate collisions.

Gigabit Ethernet was introduced in 1998 when the performance of Ethernet technology increased by a factor of 10. VLAN tagging was now supported.

In 10 Gb DCB (Data Center Bridging), full duplex is required to help support lossless communication in a Fibre Channel over Ethernet (FCoE) environment.

Ethernet devices

Repeaters and hubs were the original way to increase the distance and scalability of the Collision Domain.

A *repeater* is a layer 1 (physical layer) device that repeats a signal. The purpose of a repeater is similar to a signal amplifier. A repeater is a dual-port device and usually utilized to extend a connection between two hosts or to connect a group of hosts that exceed the distance limitation of early generations of Ethernet. Since its purpose is to regenerate the signal, it is usually placed inline to increase the reachability of the Ethernet network. It is transparent to hosts, which are unaware of the presence of the repeater across the link.

The repeater has these fundamental functions:

- Retime the signal

- Restore the symmetry of the signal
- Restore the signal amplitude

A *hub* is simply a means of connecting Ethernet cables together so that signals can be repeated to every other connected cable on the hub. For this reason, hubs are also called multi-port repeaters.

When 10BaseT Ethernet began utilizing unshielded twisted pair (UTP) cables, hubs became popular in most installations. Many companies used hubs on their LANs to also allow greater flexibility. Hubs supported UTP and BNC 10Base-2 installations, but UTP was so much easier to work with that it became the most common cable type used.

Ethernet *bridges* (*switches*) work somewhat like Ethernet hubs, passing all traffic between segments. However, as the switch discovers the addresses associated with each port, it only forwards network traffic to the necessary segments, improving overall performance. Broadcast traffic is still forwarded to all network segments. Switches also overcame the limits on total segments between two hosts and allowed the mixing of speeds, both of which became very important with the introduction of *Fast Ethernet*.

The network switch, or packet switch (or just switch), plays an integral part in most Ethernet local area networks or LANs. In the context of a standard 10/100 Ethernet switch, a switch operates at the data-link layer of the OSI model to create a different collision domain per switch port allowing data to never interfere with each others' conversations. Switches allow you to have dedicated bandwidth on point-to-point connections with every computer by running in full duplex, thereby avoiding collisions. For more information, refer to "Ethernet switching concepts" on page 86.

Auto-negotiation

Auto-negotiation is a mechanism that enables interfaces to automatically set their speed and mode in interaction with other Ethernet switches/hubs or hosts, relieving the network engineer of this configuration task. Auto-negotiation also makes the migration from 10 Mb/s to 100 Mb/s Ethernet easy to accomplish.

The basic mechanisms of auto-negotiation are:

- Operation over link segments

Auto-negotiation is designed to work across link segments only. A link segment is composed of two devices connected to each other over a single piece of media.

- Auto-negotiation occurs at link initialization

  When a cable is connected or a port comes up, the link is initialized by the Ethernet devices at each end of the link. Link initialization and auto-negotiation occurs before any data or communication is passed across the link.

- Auto-negotiation uses its own signaling system

  Each Ethernet media system has a particular way of sending signals across the link. Auto-negotiation uses its own independent signaling designed for copper based cabling. These signals are sent once during link initialization.

Gigabit Ethernet

In 1998, Gigabit Ethernet was introduced and ran over copper or fiber-optic media. At this time, auto-negotiation is still used in copper-based Ethernet to automatically adjust itself when connected to slower 10 Mb/s to 100 Mb/s nodes. Auto-negotiation was originally developed for copper-based Ethernet devices only, and therefore is not supported on all Ethernet media types.

10 GbE, or 10 Gigabit Ethernet, (fiber) was first published in 2002 as IEEE 802.3ae. Optical media types were defined by multi-source agreements (MSAs). 10 Gigabit Ethernet supports only full duplex links. Auto-negotiation and CSMA/CDare not supported with 10 GbE.

10GBaseT

The 10GBaseT standard or IEEE 802.3an was released in 2006 to provide 10 Gb/s connections over copper twisted pair cables (RJ45 media connector).

40GbE technology

In 2010, 40 Gigabit Ethernet (40GbE) became the standard that enables the transfer of Ethernet frames at speeds of up to 40 gigabits per second (Gbps). The 40GbE standard is intended for local server connectivity.

The 40GbE standard was initially intended for end node connectivity. 100GbE was designed to be used for interswitch or backbone connectivity. However, the current 40GbE implementation can be used for both server/storage and interswitch connectivity. To accommodate local server/storage, if the server/storage/switch does not support QSFP and 40GbE, a breakout cable must be used to provide 10GbE connectivity.

The following standards apply:

- IEEE Standard 802.3ba
  - Based on 4x 10GbE lanes, but supports True 40GbE flows
- Modes
  - True/Native 40GbE
  - 4x 10GbE
  - QSFP+ can operate in both modes
- Multilane distribution
  - Traffic split across multiple lanes (round robin)
  - Performed by transmitting each 66-bit word in round robin fashion across 10GbE lanes
  - Allows single flow to use aggregate bandwidth of the lanes (4)
  - Line rate bandwidth achieved by using all lanes simultaneously and without flow based constraints of 4x10GbE port channel and aggregation bonded link

Figure 28 shows an example of a multilane distribution process.

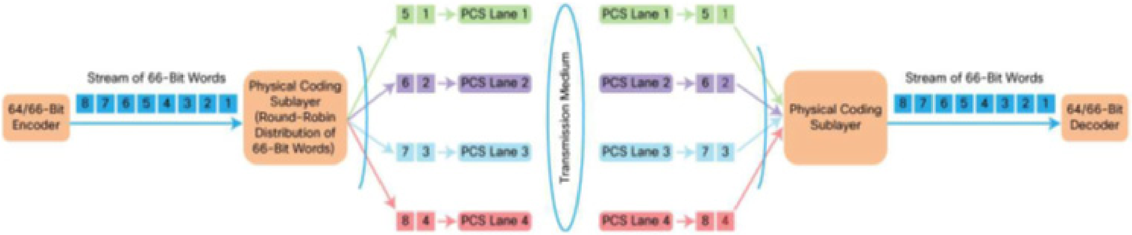


Figure 28 Multilane distribution process example

- 64/66 encoding/decoding
- Medium
  - Four Fiber Pairs
    - Each lane transmitted on single fiber pair
    - Requires four pairs of laser and optics
    - Same wavelength can be used per fiber pair

  Figure 29 shows an example of a Four Fiber Pair Transmission.

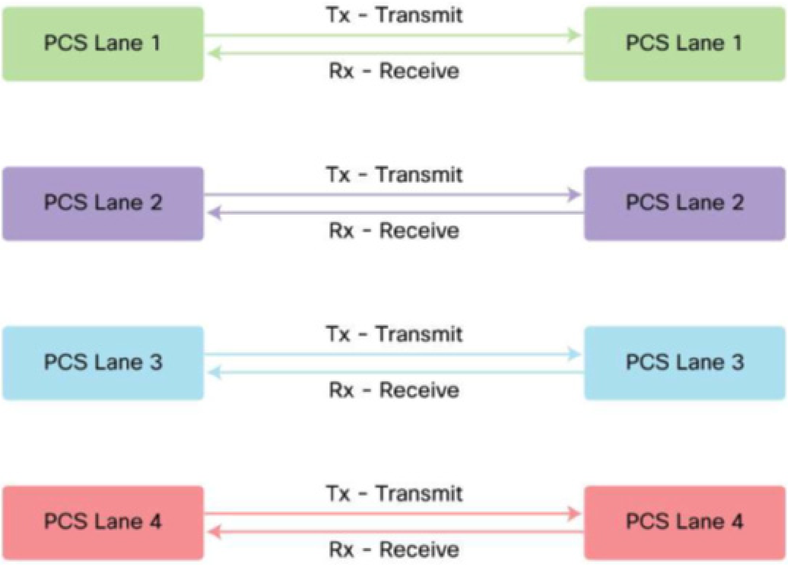


Figure 29 Four Fiber Pair Transmission

- Single Fiber Pair
  - All four lanes transmitted on single fiber pair
  - Requires single fiber pair and optic
  - 4 wavelengths/lambdas combined onto single fiber (CWDM)

Figure 30 shows an example of a Single Fiber Pair Transmission.

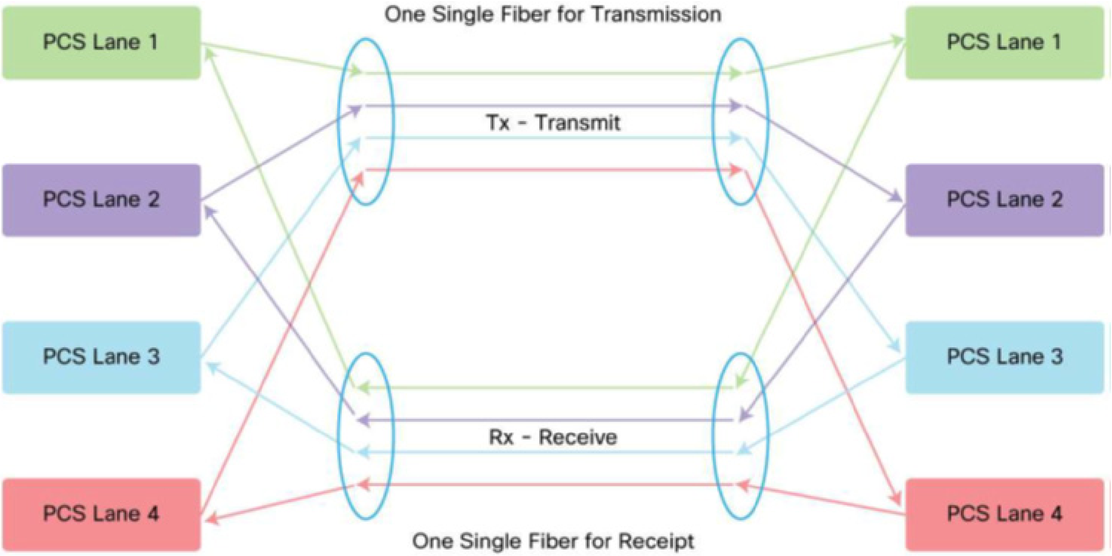


Figure 30 Single Fiber Pair Transmission

Protocols

The Open System Interconnection (OSI) protocol suite comprises of numerous standard protocols that are based on the OSI reference model. These protocols are part of an international program to develop data-networking protocols and other standards that facilitate multi-vendor equipment interoperability.

Ethernet is a frame-based protocol that is used to transport data across a Layer 2 network. A request for data from a client to a server (or host to storage) would start with the application making a request and then that request is passed down the stack. The information needed is encapsulated and mapped back to the OSI model.

This section provides further information on these two protocols:

- "OSI networking protocol" on page 75
- "Ethernet frame-based protocol" on page 80

OSI networking protocol

The OSI protocol suite is designed to facilitate communication between hardware and software systems, despite differences in underlying architectures. The OSI specifications were conceived and implemented by two international standards organizations: the International Organization for Standardization (ISO) and the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T).

Figure 31 on page 76 illustrates the entire OSI protocol suite and its relation to the layers of the OSI reference model.

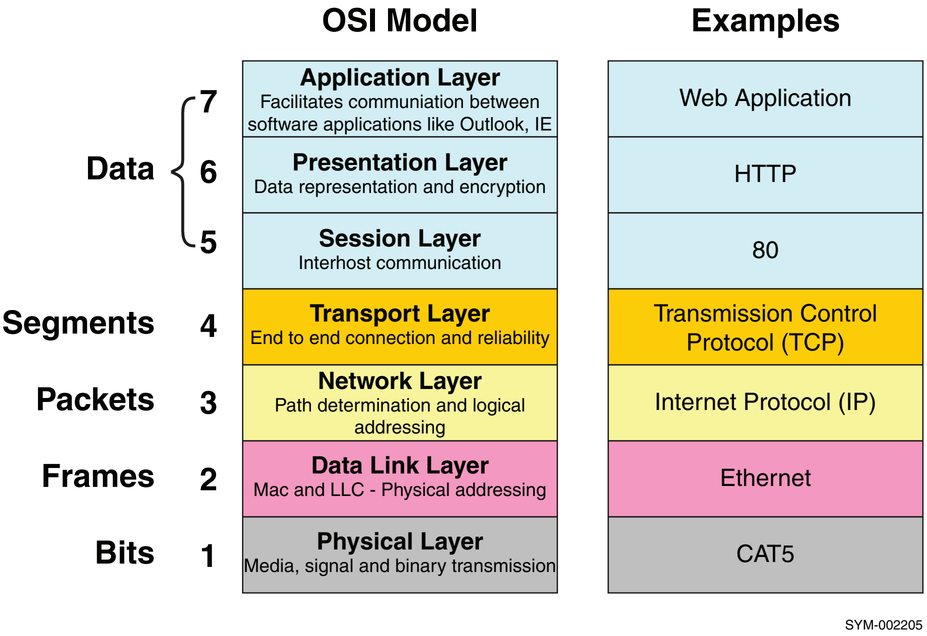

OSI Model
Examples
Data
7
6
5
Segments
4
Packets
3
Frames
2
Bits
1
Application Layer
Facilitates communiation between software applications like Outlook, IE
Presentation Layer
Data representation and encryption
Session Layer
Interhost communication
Transport Layer
End to end connection and reliability
Network Layer
Path determination and logical addressing
Data Link Layer
Mac and LLC - Physical addressing
Physical Layer
Media, signal and binary transmission
Web Application
HTTP
80
Transmission Control Protocol (TCP)
Internet Protocol (IP)
Ethernet
CAT5
SYM-002205

Figure 31 OSI protocol suite

Each layer is further explained as follows:

Layer 1 – Physical Layer

The Physical layer defines all electrical and physical specifications for devices. This includes the layout of pins, voltages, and cable specifications. The OSI protocol suite supports numerous standard media at the physical layers. Hubs and repeaters are one of the examples of physical layer devices.

Layer 2 – Data Link Layer

The Data Link layer provides the functional and procedural means to transfer frames between network entities and to detect and possibly correct errors that may occur in the Physical layer. The addressing scheme is physical which means that the addresses are hard-coded into the network cards at the time of manufacture. The addressing scheme is flat. The best known example of this layer is Ethernet. Other examples of data link protocols are Token ring, FDDI and Frame-relay networks. This is the layer at which bridges and switches operate. Connectivity is provided only among locally attached network nodes/ hosts.

Layer 3 – Network Layer

The Network layer provides the means of transferring packets from a source to a destination via one or more networks while maintaining the quality of service requested by the Transport layer. The Network layer performs network routing, flow control, error control functions, segmentation and desegmentation. The router operates at this layer which sends data throughout the extended network and making the Internet possible, although there are layer 3 switches. This uses a logical and hierarchical addressing scheme, which values are chosen by the network designer/ engineer. The best known examples of layer 4 protocols are IP, IPX and Appletalk.

Layer 4 – Transport Layer

The OSI protocol suite implements two types of services at the Transport layer: connection-oriented transport service (TCP) and connectionless transport service (UDP). Five connection-oriented transport-layer protocols exist in the OSI suite, ranging from Transport Protocol Class 0 through Transport Protocol Class 4. Connectionless transport service is supported only by Transport Protocol Class 4.

- Transport Protocol Class 0 (TP0), the simplest OSI transport protocol, performs segmentation and reassembly functions. TP0 requires connection-oriented network service.
- Transport Protocol Class 1 (TP1) performs segmentation and reassembly and offers basic error recovery. TP1 sequences protocol data units (PDUs) and will retransmit PDUs or reinitiate the connection if an excessive number of PDUs are unacknowledged. TP1 requires connection-oriented network service.
- Transport Protocol Class 2 (TP2) performs segmentation and reassembly, as well as multiplexing and demultiplexing data streams over a single virtual circuit. TP2 requires connection-oriented network service.
- Transport Protocol Class 3 (TP3) offers basic error recovery and performs segmentation and reassembly, in addition to multiplexing and demultiplexing data streams over a single virtual circuit. TP3 also sequences PDUs and retransmits them or reinitiates the connection if an excessive number are unacknowledged. TP3 requires connection-oriented network service.

- Transport Protocol Class 4 (TP4) offers basic error recovery, performs segmentation and reassembly, and supplies multiplexing and demultiplexing of data streams over a single virtual circuit. TP4 sequences PDUs and retransmits them or reinitiates the connection if an excessive number are unacknowledged. TP4 provides reliable transport service and functions with either connection-oriented or connectionless network service. It is based on the Transmission Control Protocol (TCP) in the Internet Protocols suite and is the only OSI protocol class that supports connectionless network service.

Layer 5 – Session Layer

The Session layer provides the mechanism for managing the dialogue between end-user application processes. It provides for either duplex or half duplex operation and establishes checkpointing, adjournment, termination, and restart procedures. This layer is responsible for setting up and tearing down TCP/IP sessions. This layer consists of a session protocol and a session service. The session protocol allows session-service users (SS-users) to communicate with the session service. An SS-user is an entity that requests the services of the session layer. Such requests are made at Session-Service Access Points (SSAPs), and SS-users are uniquely identified by using an SSAP address. Session service provides four basic services to SS-users. First, it establishes and terminates connections between SS-users and synchronizes the data exchange between them. Second, it performs various negotiations for the use of session-layer tokens, which must be possessed by the SS-user to begin communicating. Third, it inserts synchronization points in transmitted data that allow the session to be recovered in the event of errors or interruptions. Finally, it allows SS-users to interrupt a session and resume it later at a specific point.

Layer 6 – Presentation Layer

The Presentation layer implementation of the OSI protocol suite consists of a presentation protocol and a presentation service. The presentation protocol allows presentation-service users (PS-users)

to communicate with the presentation service. A PS-user is an entity that requests the services of the presentation layer. Such requests are made at Presentation-Service Access Points (PSAPs). PS-users are uniquely identified by using PSAP addresses. Presentation service negotiates transfer syntax and translates data to and from the transfer syntax for PS-users, which represent data using different syntaxes. The presentation service is used by two PS-users to agree upon the transfer syntax that will be used. When transfer syntax is agreed upon, presentation-service entities must translate the data from the PS-user to the correct transfer syntax. The OSI presentation-layer service is defined in the ISO 8822 standard and in the ITU-T X.216 recommendation. The OSI presentation protocol is defined in the ISO 8823 standard and in the ITU-T X.226 recommendation. A connectionless version of the presentation protocol is specified in the ISO 9576 standard.

Layer 7 – Application Layer

The Application layer interfaces directly to and performs common application services for the application processes. It consists of various application entities. An application entity is the part of an application process that is relevant to the operation of the OSI protocol suite. An application entity is composed of the user element and the application service element (ASE). The user element is part of an application entity that uses ASEs to satisfy the communication needs of the application process. The ASE is the part of an application entity that provides services to user elements and, therefore, to application processes. ASEs also provide interfaces to the lower OSI layers. Some of the standard OSI application processes include the following:

- Common Management-Information Protocol (CMIP) performs network management functions, allowing the exchange of management information between ESs and management stations. CMIP is specified in the ITU-T X.700 recommendation and is functionally similar to the Simple Network-Management Protocol (SNMP) and NetView.

- Directory Services (DS) serves as a distributed directory that is used for node identification and addressing in OSI internetworks. DS is specified in the ITU-T X.500 recommendation.

- File Transfer, Access, and Management (FTAM) provide file-transfer service and distributed file-access facilities.

- Message Handling System (MHS) provides a transport mechanism for electronic messaging applications and other applications by using store-and-forward services.

- Virtual Terminal Protocol (VTP) provides terminal emulation that allows a computer system to appear to a remote ES as if it were a directly attached terminal.

Ethernet frame-based protocol

Ethernet is a frame-based protocol that is used to transport data across a layer 2 network. As discussed in "Protocols" on page 75, a request for data from a client to a server (or host to storage) would start with the application making a request and then that request is passed down the stack.

As the request is passed down the stack, it is encapsulated with information that will allow the request to be routed to the server. The request would also need to contain information for the requested data to be sent back to the client. This encapsulation, shown in Figure 32 on page 81, is required to facilitate the transfer of information.

Fibre Channel over Ethernet (FCoE) is a new approach to I/O consolidation over Ethernet, allowing Fibre Channel and Ethernet networks to share a single, integrated infrastructure, thereby reducing network complexities in the data center. For more information, refer to Chapter 1, ”Introduction to Fibre Channel over Ethernet.”

Typical Ethernet frame example

Figure 32 on page 81 shows a typical Ethernet frame and how the encapsulation can be mapped back to the OSI model for a request from a Web application.

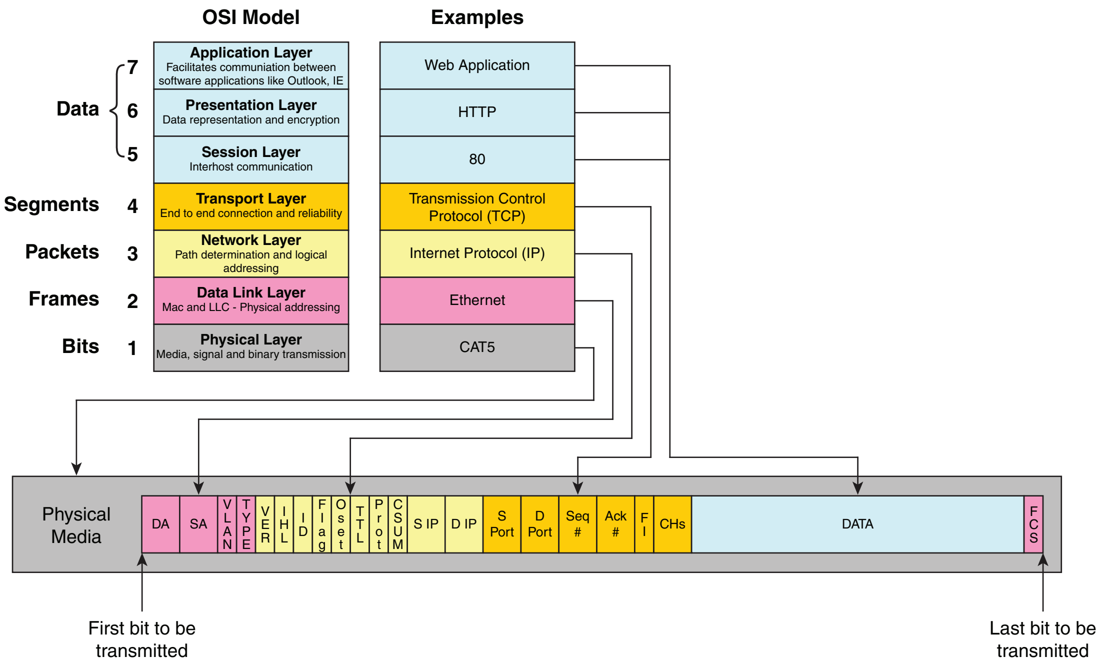

OSI Model
Examples
Data
7
6
5
Segments
4
Packets
3
Frames
2
Bits
1
Application Layer
Facilitates communiation between software applications like Outlook, IE
Presentation Layer
Data representation and encryption
Session Layer
Interhost communication
Transport Layer
End to end connection and reliability
Network Layer
Path determination and logical addressing
Data Link Layer
Mac and LLC - Physical addressing
Physical Layer
Media, signal and binary transmission
Web Application
HTTP
80
Transmission Control Protocol (TCP)
Internet Protocol (IP)
Ethernet
CAT5
Physical Media
DA
SA
VLAN
TYPE
VER
IHL
ID
Flag
Oset
TTL
Prot
CSUM
S IP
D IP
S Port
D Port
Seq #
Ack #
FI
CHs
DATA
FCS
First bit to be transmitted
Last bit to be transmitted

Figure 32 **OSI model and frame format**

A brief explanation of the fields in the Ethernet frame, shown in Figure 32, follows:

Physical media

This is the physical media (CAT5, Fiber, etc.) that carries the encoded binary data stream.

Ethernet

Note: For detailed information about Ethernet, refer to IEEE 802.3

MAC Address – A Media Access Control (MAC) Address, as shown in Figure 33 on page 82, is a 48-bit address defined in 802 – 2001 (clause 9). Each Physical Port attached to a Layer 2 Ethernet network must have at least one MAC Address associated with it in order for the port to send and receive data.

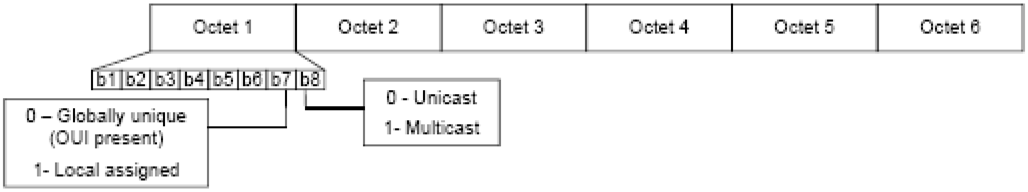


Figure 33 MAC Address

DA – Destination MAC Address - The DA is the MAC Address of the physical port where the frame will be sent to.

SA – Source MAC Address - The SA is the MAC Address of the physical port where the frame originated from.

VLAN – Virtual Local Area Network ID - The VLAN field also known as the 802.1Q tag allows multiple virtual networks to use the same physical link. The VLAN field starts with a 16-bit type field (set to 0x8100) that allows other layer 2 devices to detect the presence of the VLAN field.

FCS – Frame Check Sequence – A CRC value that covers the entire Ethernet Frame.

IP

Note: Not all of the fields in the IP header are shown. For detailed information on each field present in an IP header, refer to RFC 791 – Internet Protocol.

TYPE – Ether Type – The Ether Type is used to determine how to interpret the next bytes of data in the frame. For IP the Ether Type is 0x0800.

IHL – IP Header Length – The header length is used to specify the length of the IP header. The value is in word (32 bit) increments.

ID – Identification – An identifying value assigned by the sender to aid in assembling the fragments of a datagram.

Flag – Control flags – Indicates if fragmentation is allowed and when fragmentation is supported it also indicates if this is the last fragment or not.

Oset – Fragment Offset – Indicates where this fragment belongs in the datagram.

TTL – Time To Live – The maximum amount of time that the datagram is allowed to exist.

Prot – Protocol – Indicates the next layer (up the stack) protocol that is used in the data portion of the datagram. For example if TCP is used the value would be 0x6, if UDP is used the value would be 0x11.

CSUM – Checksum – A checksum of the header only.

S IP – Source IP Address – The IP Address of the network entity that transmitted the datagram.

D IP – Destination IP Address – The IP Address of the network entity that the datagram is destined for.

TCP

Note: Not all of the fields in the TCP header are shown. For detailed information on each field present in a TCP header, refer to RFC 793 – Transmission Control Protocol.

S port – Source Port – The port number used to identify the sending application.

D port – Destination Port – The port number used to identify the receiving application.

Seq# – Sequence number – The sequence number identifies the first byte of data in the segment.

Ack# – Acknowledgement number – The next sequence number that the sender expects to receive.

Fl – Flags – The Flags included in a TCP segment are URG, ACK, PSH, RST, SYN and FIN.

CHs – Checksum – The TCP checksum covers the entire TCP segment (both header and data).

FCoE Ethernet Frame example

Figure 34 shows an Ethernet frame and how the encapsulation can be mapped back to the OSI model in an FCoE environment.

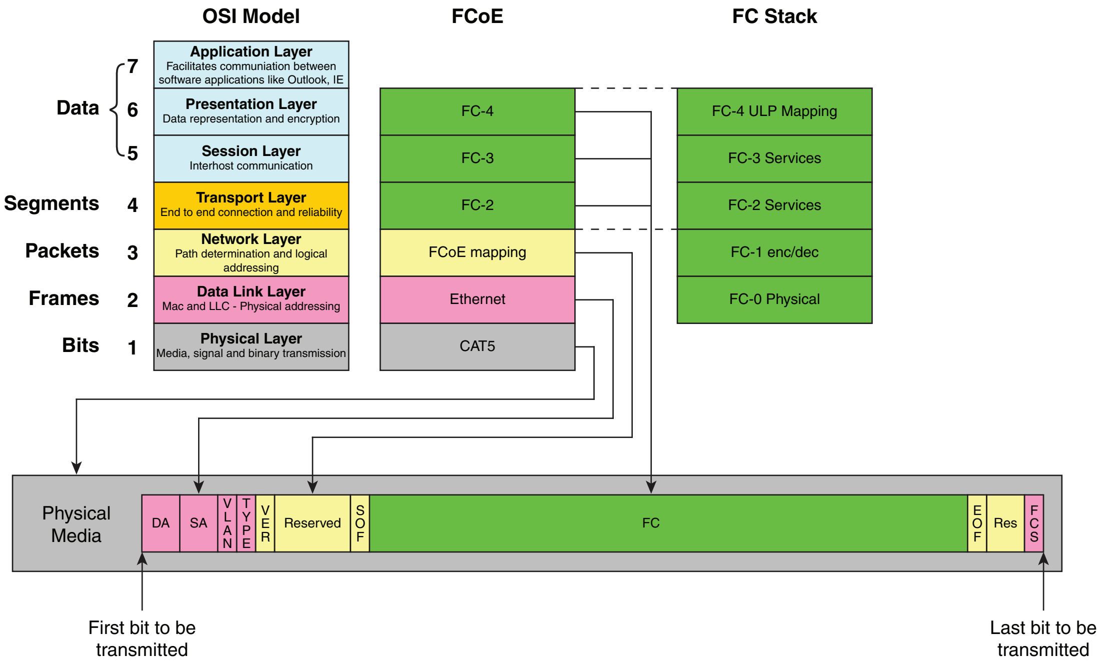


Figure 34 Encapsulated Ethernet frame in an FCoE environment

A brief explanation of the fields in the Ethernet frame, shown in Figure 34, follows:

Physical Media

This is the physical media (CAT5, Fiber, etc.) that carries the encoded binary data stream.

Ethernet

Note: For detailed information about Ethernet, refer to IEEE 802.3.

MAC Address – A Media Access Control (MAC) Address is a 48-bit address defined in 802 – 2001 (clause 9). Each Physical Port attached to a Layer 2 Ethernet network must have at least one MAC Address associated with it in order for the port to send and receive data.

DA – Destination MAC Address - The DA is the MAC Address of the physical port where the frame will be sent to.

SA – Source MAC Address - The SA is the MAC Address of the physical port where the frame originated from.

VLAN – Virtual Local Area Network ID - The VLAN field also known as the 802.1Q tag allows multiple virtual networks to use the same physical link. The VLAN field starts with a 16 bit type field (set to 0x8100) that allows other layer 2 devices to detect the presence of the VLAN field. With FCoE, all frames need to at least be priority tagged. This means that it is possible for the VLAN to be null but the COS.

FCS – Frame Check Sequence – A CRC value that covers the entire Ethernet Frame.

FCoE Mapping

VER – Version – The FCoE version being used.

Reserved – The reserved field is necessary to ensure that the minimum Ethernet Frame length of 64 bytes is always maintained.

SOF – Start of Frame delimiter – This field indicates the beginning of the FC Frame.

EOF – End of Frame delimiter – This field indicates the ending of the FC Frame.

Res – Reserved – This reserved field is necessary to ensure that the total FC Frame length is always a multiple of 4.

FC

Data – The contents of the FC frame starting with the R_CTL field and ending with the CRC.

Ethernet switching concepts

The information in this section is intended to provide a person who is familiar with Fibre Channel a basic understanding of switching concepts and the differences between FC and Ethernet switching and to explain some of the more common features used in Ethernet L2 networks.

This section includes information on the following switching concepts:

- "Fibre Channel switching versus Ethernet bridging" on page 86
- "Gratuitous ARP" on page 89
- "Unicast flood" on page 90
- "Spanning Tree Protocol (STP)" on page 92
- "Link Aggregation" on page 105
- "Access Control Lists" on page 112

Fibre Channel switching versus Ethernet bridging

Although Fibre Channel switches and Ethernet bridges (switches) both provide a similar service (such as delivering frames from one address to another), there are significant differences in how this task is accomplished. One of the biggest differences is that in Fibre Channel, frames are sent from and to addresses (N_Port_IDs) that the switch knows about ahead of time while Ethernet switches need to discover the address (MAC Address) the first time a frame with an unknown destination is received.

One of the reasons for this difference is that the addresses used in Fibre Channel are provided to a device by the switch during Fabric Login and are derived from the FC switches Domain ID. Figure 35 on page 87 shows a simple FC environment containing one host, one storage FE (front end) port and two switches connected by two ISLs (Inter Switch Links).

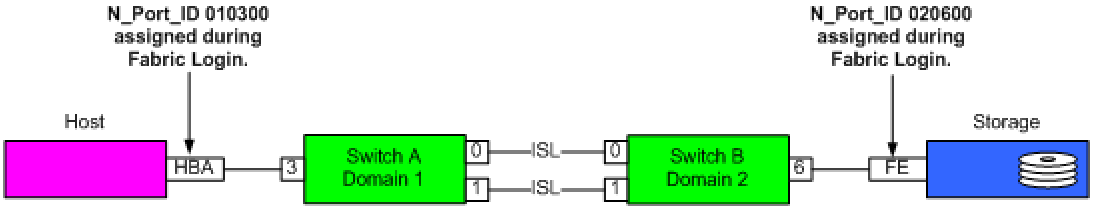


Figure 35 **FC addressing**

The Host's N_Port_ID is 010300 and this N_Port_ID can typically be broken down into three different areas, Domain, Area, and Port as shown in Figure 36.

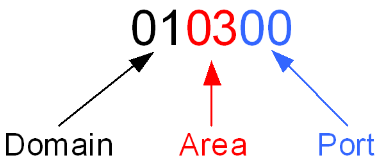


Figure 36 **N_Port_ID format**

Any host or storage port that logs in to Domain 1 (Switch A in this case) will be assigned an N_Port_ID that has the Domain portion of the N_Port_ID set to 01. The next byte (Area) is typically bound to a specific physical interface on the switch, (03 for the host above). The final byte (Port) is typically used to differentiate different N_Ports hanging off of the same physical switch port (such as a loop, more commonly found with NPIV).

One of the interesting benefits of having the Domain ID included in the N_Port_ID is that each N_Port_ID includes a form of routing information. This routing information (Domain ID) allows a switch to determine if the D_ID (Destination ID) of a frame is for the local Domain or not and if the frame is not for the local Domain, it allows the switch to determine the best path to the destination Domain. Inclusion of the Domain ID in the N_Port_ID in conjunction with the build fabric process allows for some nice features such as multiple links (ISLs) between two switches without having to worry about things like forwarding loops.

In an Ethernet environment, each end device provides its own addresses (MAC Address) and since there is no concept of a Fabric Login in Ethernet, the MAC Address of an end device cannot be determined until the device transmits a frame. For example, Figure 37 shows a host attached to Ethernet switch A and an array FE port attached to Ethernet switch port B. The Ethernet switches are physically connected by two cables but only one link is active due to Spanning Tree.

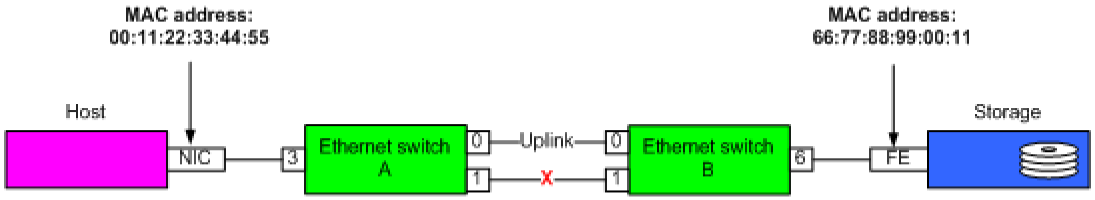


Figure 37 **Ethernet addressing**

Typically when an Ethernet end device comes up for the first time it will either transmit a gratuitous ARP (Address Resolution Protocol) frame (discussed further in "Gratuitous ARP" on page 89), a DHCP request, or another type of outbound Ethernet frame. When the switch receives the first frame, it will add the MAC Address in the SA (Source Address) field to the MAC Address Table. The following is an example of what a MAC Address Table might look like for this environment.

```
Ethernet-switch-A# show mac-address-table

VLAN       MAC Address        Type         Age        Port
------------+---------------------+-----------+---------+--------------------
1         0011.2233.4455       dynamic      0         Eth1/3
1         6677.8899.0011       dynamic      10        Eth1/0
```

```
Ethernet-switch-B# show mac-address-table

VLAN       MAC Address        Type         Age        Port
------------+---------------------+-----------+---------+--------------------
1         0011.2233.4455      dynamic      10        Eth1/0
1         6677.8899.0011      dynamic      20        Eth1/6
```

The MAC Address tables for Ethernet-switch-A and Ethernet-switch-B are shown above. The fields are explained next:

VLAN – The VLAN that the entry belong to.

MAC Address – The MAC Address for the entry in the table. These entries are listed in ascending order based on the MAC Address.

Type – The type of entry either Static or Dynamic. Static entries will not be removed from the table and have to be manually configured. Dynamic entries are learned by the switch and will be timed out after a period of time. The default timeout is 300 seconds or 5 minutes.

Age – The length of time in seconds since the last Frame with an SA equal to the MAC Address was received. By default, when the age reaches 300 seconds, the entry is removed from the MAC Address Table.

Port – The port where frames with a DA (Destination Address) equal to the MAC Address should be forwarded.

The MAC Address Table for Ethernet-switch-A shown above contains two entries, one for the host's NIC and one for the Array's FE port. Notice that on Ethernet-switch-A the outbound port for the NIC is Eth1/3 while the outbound port for the FE port is Eth1/0. This means that any frame received with a DA of 0011.2233.4455 will be transmitted on port Eth1/3 and any frame received with a DA of 6677.8899.0011 will be transmitted on port Eth1/0. For frames that are transmitted to the FE port out of port Eth1/0, they will be received at Ethernet-switch-B (on port Eth1/0). When the frame is received at Ethernet-switch-B the frame will be transmitted on port Eth1/6 due to the entry in the MAC Address Table.

Gratuitous ARP

Gratuitous ARP could mean both gratuitous ARP request or gratuitous ARP reply. Gratuitous in this case means a request/reply that is not normally needed according to the ARP specification (RFC 826) but could be used in some cases. A gratuitous ARP request is an Address Resolution Protocol request packet where the source and destination IP are both set to the IP of the machine issuing the packet and the destination MAC is the broadcast address ff:ff:ff:ff:ff:ff. Ordinarily, no reply packet will occur. A gratuitous ARP reply is a reply to which no request has been made.

Gratuitous ARPs are useful for four main reasons:

- Helps detect IP conflicts. When a machine receives an ARP request containing a source IP that matches its own, then it knows there is an IP conflict.
- Assists in the updating of other machines' ARP tables. Clustering solutions utilize this when they move an IP from one NIC to another, or from one machine to another. Other machines

maintain an ARP table that contains the MAC associated with an IP. When the cluster needs to move the IP to a different NIC, be it on the same machine or a different one, it reconfigures the NICs appropriately then broadcasts a gratuitous ARP reply to inform the neighboring machines about the change in MAC for the IP. Machines receiving the ARP packet then update their ARP tables with the new MAC.

- Informs switches of the MAC Address of the machine on a given switch port, so that the switch knows that it should transmit packets sent to that MAC Address on that switch port.
- Announces link up event. Every time an IP interface or link goes up, the driver for that interface will typically send a gratuitous ARP to preload the ARP tables of all other local hosts. Thus, a gratuitous ARP will tell us that host just has had a link up event, such as a link bounce, a machine just being rebooted or the user/sysadmin on that host just configuring the interface up. If we see multiple gratuitous ARPs from the same host frequently, it can be an indication of bad Ethernet hardware/cabling resulting in frequent link bounces.

Note: Information in this section was found at http://wiki.wireshark.org/Gratuitous_ARP, which provides more information.

Unicast flood

Although excessive Unicast flooding may lead to performance issues, unicast flooding is a normal part of the Ethernet switching process. An example of when a unicast flood would be performed would be in the case where one of the entries in the MAC Address Table expires. Figure 38 on page 91 shows an example.

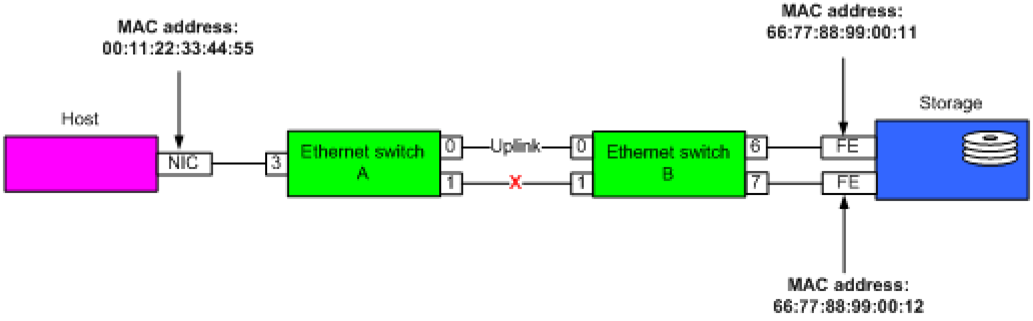


Figure 38 Unicast flood example

If every MAC Address in the L2 network shown in Figure 38 were known by both switches, the MAC Address Tables would appear similar to the following:

```
Ethernet-switch-A# show mac-address-table

VLAN      MAC Address        Type        Age         Port
------------+---------------------+-----------+---------+--------------------
1         0011.2233.4455      dynamic     0          Eth1/3
1         6677.8899.0011      dynamic    10          Eth1/0
1         6677.8899.0012      dynamic    10          Eth1/0
```

```
Ethernet-switch-B# show mac-address-table

VLAN      MAC Address        Type        Age        Port
------------+---------------------+-----------+---------+--------------------
1         0011.2233.4455     dynamic     10         Eth1/0
1         6677.8899.0011     dynamic     20         Eth1/6
1         6677.8899.0012     dynamic     20         Eth1/7
```

However, if the entry for MAC Address 0011.2233.4455 were to expire on switch B, the MAC Address Table would appear similar to the following:

```
Ethernet-switch-A# show mac-address-table

VLAN      MAC Address         Type        Age       Port
------------+---------------------+-----------+---------+--------------------
1         0011.2233.4455      dynamic     0          Eth1/3
1         6677.8899.0011      dynamic     10         Eth1/0
1         6677.8899.0012      dynamic     10         Eth1/0
```

```
Ethernet-switch-B# show mac-address-table

VLAN      MAC Address        Type        Age         Port
------------+---------------------+-----------+---------+--------------------
1         6677.8899.0011     dynamic     20           Eth1/6
1         6677.8899.0012     dynamic     20           Eth1/7
```

If this were to happen, then when either of the storage ports on switch B were to try and transmit a frame to the host, switch B would not recognize the MAC Address and would transmit it on all ports except for the one the frame was received on. The frame would be received on switch A and then transmitted only on port Eth1/3. When the host responds to the frame from the storage port, the MAC Address Table on switch B would get updated.

Spanning Tree Protocol (STP)

The Spanning Tree topics covered in this section are:

- "Overview" on page 92
- "Election of root switch" on page 94
- "BPDUs" on page 94
- "Spanning Tree port states" on page 96
- "Spanning Tree timers" on page 97
- "Spanning Tree path costs" on page 98
- "STP Topology change example" on page 98
- "Rapid Spanning Tree (802.1w)" on page 102
- "Multiple Spanning Tree (802.1s)" on page 104

Overview

The critical underlying technology to any Layer 2 network is the STP, invented by Radia Perlman. Whether this involves only two switches connected together with redundant links or many switches connected together through a mesh type topology. This protocol will ensure stability by learning the network topology and prevent forwarding loops from forming by creating a *Spanning Tree*.

A forwarding loop is a situation that can be created when a topology contains multiple paths and spanning tree is not being used. Refer to Figure 39.

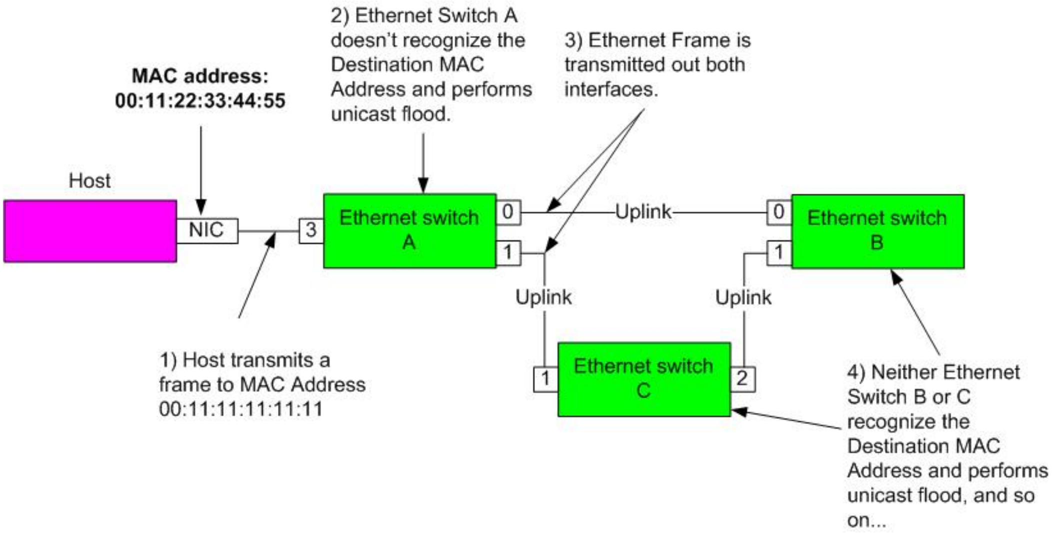

2) Ethernet Switch A doesn't recognize the Destination MAC Address and performs unicast flood.
3) Ethernet Frame is transmitted out both interfaces.
MAC address: 00:11:22:33:44:55
Host
NIC
3
Ethernet switch A
0
1
Uplink
0
Ethernet switch B
1
Uplink
Uplink
1
Ethernet switch C
2
1) Host transmits a frame to MAC Address 00:11:11:11:11:11
4) Neither Ethernet Switch B or C recognize the Destination MAC Address and performs unicast flood, and so on...

Figure 39 Forwarding loop

As shown in Figure 39, the Host transmits a frame to a MAC Address of 00:11:11:11:11:11. When Ethernet switch A receives the frame it will not recognize the DA and as a result will perform a unicast flood (see "Unicast flood" on page 90 for more information). When the frame is received by Ethernet switch B and C, they will not recognize the DA either and perform a unicast flood. This process repeats as fast as the frame can be forwarded and will eventually consume all of the bandwidth available between the three switches.

In an FCoE environment, it is important to understand how the Spanning Tree protocol functions because if it is left enabled on the interfaces where CNAs attach, it can drastically elongate the amount of time it takes for a link to initialize following a power cycle or a cable pull. For this reason, EMC recommends that users disable spanning tree on Ethernet interfaces where CNAs will be attached.

The following briefly describes how the spanning tree protocol works:

- STP is a link management protocol that provides path redundancy while preventing undesirable forwarding loops in the network. For an Ethernet network to function properly, only one active path can exist between any two switches.
- By definition, multiple active paths between switches create loops in the network. STP was created to allow for multiple paths while preventing loops by putting one of the paths into a blocking state.
- To provide path redundancy, STP defines a tree that spans all switches in an extended network. STP forces certain redundant data paths into a standby (blocked) state. If one network segment becomes unreachable, or if STP costs change, the spanning-tree algorithm reconfigures the spanning-tree topology and reestablishes the link by activating the standby path.
- STP operation is transparent to end stations, which are unaware whether they are connected to a single LAN segment or a switched LAN of multiple segments.

Election of root switch

All switches in a LAN participating in STP gather information on other switches in the network through an exchange of data messages. These messages are called *Bridge Protocol Data Units* (BPDUs). This exchange of messages results in the following:

- The election of a unique root switch for the stable spanning-tree network topology.
- The election of a designated switch for every switched LAN segment.
- The removal of loops in the switched network by placing redundant switch ports into the Blocking state.

BPDUs

Bridge Protocol Data Units, (BPDUs), are special data frames used to exchange information about Switch IDs and root path costs. The STP uses this information to elect the root switch and root port for the switched network, as well as the root port and designated port for each switched segment.

A switch sends a BPDU frame using its own unique MAC Address as the source and a destination address of the STP multicast address 01:80:C2:00:00:00.

The are three types of BPDUs:

- Configuration BPDU (CBPDU) — Used for STP computation
- Topology Change Notification (TCN) — Used to announce a change within the network topology
- Topology Change Acknowledgement (TCA) — Used to acknowledge TCNs

BPDUs are exchanged regularly (every 2 seconds by default) and enable switches to keep track of network changes and react accordingly.

The stable active topology of a switched network is determined by the following:

- The unique switch identifier (MAC Address) associated with each switch.
- The path cost to the root associated with each switch port.
- The port identifier (MAC Address) associated with each switch port.

Each configuration BPDU contains the following information:

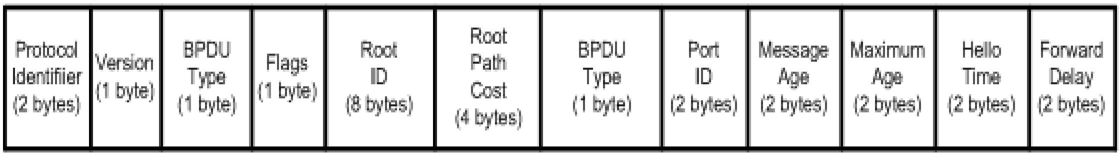
- The unique identifier (MAC Address) of the switch that the transmitting switch believes to be the root switch.
- The cost of the path to the root from the transmitting port.
- The identifier of the transmitting port.

A BPDU exchange results in the following:

- One switch is elected as the root switch. If not configured, the switch with the lowest MAC Address will win the election.
- The shortest distance to the root switch is calculated for each switch.
- A designated switch is selected. This is the switch closest to the root switch through which frames will be forwarded to the root.
- A port for each switch is selected. This is the port providing the best path from the switch to the root switch. If equal cost paths are available, the same mechanism is used as the election of the root switch. The port with the lowest MAC Address will become the designated port.
- Ports included in the STP are selected.

If all switches are enabled with default settings, the switch with the lowest MAC Address (Bridge ID) in the network becomes the root switch. However, due to traffic patterns, number of forwarding ports, or line types, the switch with the lowest MAC Address may not be best suited to be the root switch. You can force a STP recalculation to form a new, stable topology with the best functional path by increasing the priority (lowering the numerical priority number) of the ideal switch so that it then becomes the root switch.

Figure 40 shows the BPDU frame format.



Protocol Identifiier (2 bytes)	Version (1 byte)	BPDU Type (1 byte)	Flags (1 byte)	Root ID (8 bytes)	Root Path Cost (4 bytes)	BPDU Type (1 byte)	Port ID (2 bytes)	Message Age (2 bytes)	Maximum Age (2 bytes)	Hello Time (2 bytes)	Forward Delay (2 bytes)

Figure 40 BPDU frame format

Spanning Tree port states

When a switch port transitions directly from non-participation in the stable topology to the forwarding state, it creates temporary data loops. Ports must wait for new topology information to propagate through the switched LAN before starting to forward frames. They must also allow the frame lifetime to expire for frames that have been forwarded using the old topology.

Each port on a switch using STP exists in one of the following five states:

- Blocking — A port in *blocking* state does not participate in frame forwarding but does receive and respond to network management messages.
- Listening — A port in *listening* state does not learn or forward MAC Addresses, but it does send and receive BPDUs as well as receive and respond to network management messages.
- Learning – A port in *learning* state does not forward any frames, but it does send and receive BPDUs, populate its forwarding table in preparation for the forwarding state, and receive and respond to network management messages.
- Forwarding – A port in *forwarding* state does forward all frames, send and receive BPDUs, and receive and respond to network management messages.

- Disabled – (Disabled is a manual configuration step.) A port in *disabled* state is completely non-functional and does not receive or transmit any type of frame.

After a topology change or reboot, a port using STP goes through four states. A switch with redundant connections back to root and configured correctly will go into either the *blocking* or *forwarding* state. The steps a port takes before it begins to forward user traffic are as follows:

- From initializing to blocking
- From blocking to listening
- From listening to learning
- From learning to forwarding
- Forwarding to disabled (manual configuration) — Not necessary, but listed as an option.

Spanning Tree timers

Spanning Tree timers are used to make sure that the topology is stable and to ensure that duplicate frames do not make it on the wire before any frames are forwarded.

- Hello — The hello time is the time between each bridge protocol data unit (BPDU) that is sent on a port. This time is equal to 2 seconds by default, but you can tune the time to be between 1 and 10 seconds.
- Forward delay — The forward delay is the time that is spent in the listening and learning state. This time is equal to 15 seconds by default, but you can tune the time to be between 4 and 30 seconds.
- Max age — The max age timer controls the maximum length of time that passes before a bridge port saves its configuration BPDU information. This time is 20 seconds by default, but you can tune the time to be between 6 and 40 seconds.

The default values for Spanning Tree make the convergence time slow by today's ID networking standards.

The formula for convergence is:

(20 seconds for Max age + 2 x forward delay (15 seconds for listening/learning)) = 50 seconds

Spanning Tree path costs

Path cost is directly related to the bandwidth of the link. A switch will use the path with the lowest path back to root as its primary path. If there is a second path with a higher path cost, then that path will stay in the *blocking* state unless the primary path goes away. The following table shows the default cost of an interface for a given data rate.

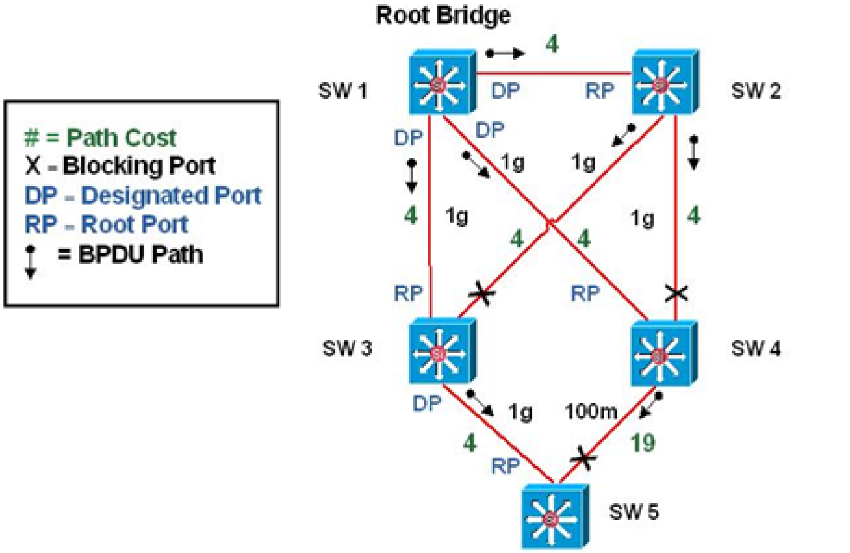
Data Rate	Path Cost (802.1D)
10 Mb/s	100
100 Mb/s	19
1 Gb/s	4
10 Gb/s	2

STP Topology change example

Figure 41 on page 99 is an example of the beginning STP topology. In this topology the STP has been completed and the environment is in a stable condition.

The things to note in Figure 41 are:

- SW 1 is the root bridge
- BPDUs are sent out from the root bridge and flow down the “Tree”
- Path costs are illustrated
- STP port states are illustrated



Path Cost for SW 5 back to Root (SW 1) via SW 3 = 8
Path Cost for SW 5 back to Root (SW 1) via SW 4 = 23
Preferred path is Through SW 3

Figure 41 Beginning STP topology example

Figure 42 is an example of an STP convergence. In this example the link between SW 3 and SW 5 has failed and there is no longer any communication between SW 3 and SW 5.

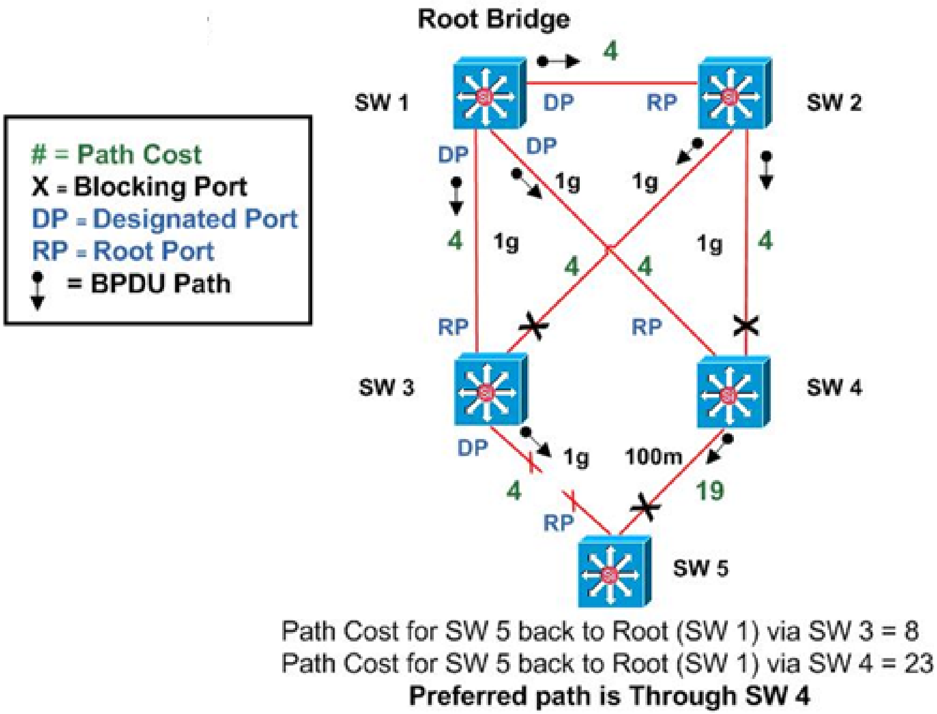


Figure 42 STP convergence example

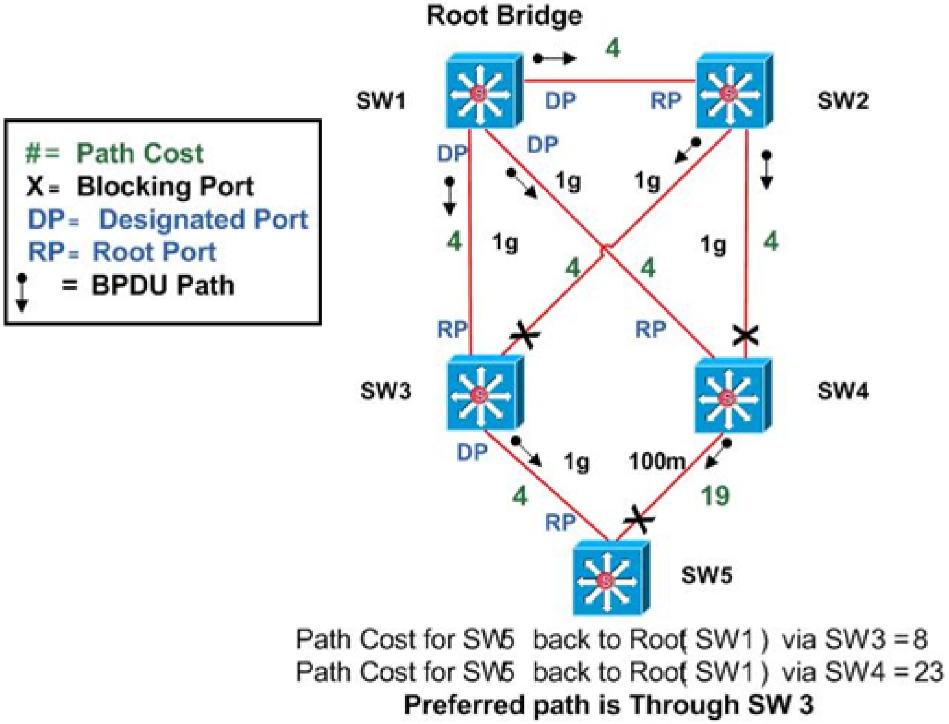
The steps included in the convergence of the STP for switch 5 are:

1. The link between SW 3 and SW 5 has just gone down and there is no longer any communication between SW 3 and SW 5.
2. SW 5 stops receiving BPDUs from SW 3.
3. SW 5 waits the Max age (20 seconds) and begins it STA.
4. SW 3 immediately sends a TCN out it RP to notify the Root Bridge (SW 1) of a topology change.
5. SW 1 acknowledges the TCN with a TCA.
6. SW 1 broadcasts out a BPDU with the TC bit set to notify all switches of a Topology Change.

After receipt of the BPDU with the TC bit set each switch will reduce its Forwarding Database aging time to 15 seconds. This is done so that any actively transmitting MAC Addresses will remain active, but any MAC Addresses that time out due to the Topology change will be flooded and immediately relearned after the STP completes and the Topology is stable again.

7. SW 5 completes its STP and begins to forward traffic to SW 4.

Figure 43 is an example of an STP re-convergence. In this example the link between SW3 and SW 5 has been restored.



Path Cost for SW5 back to Root(SW1) via SW3 =8
Path Cost for SW5 back to Root(SW1) via SW4 =23
Preferred path is Through SW 3

Figure 43 STP re-convergence example

The steps included in the restoration of links Switch 3 and 5 are:

1. SW 5 receives a Superior BPDU from SW 3.

2. SW 3 and SW 5 immediately sends out a TCN to the root bridge (SW 1).

3. SW 5 puts both its RP and newly linked port into the blocking state so it can begin its STP process.
4. SW 1 sends a TCA to acknowledge the TCN.
5. SW 1 broadcasts a BPDU with the TC bit set to notify all switches of a topology change.

   After receipt of the BPDU with the TC bit set, each switch will reduce its forwarding database aging time to 15 seconds. This is done so that any actively transmitting MAC Addresses remains active; however, any MAC Addresses that time out due to the topology change will be flooded and immediately relearned after the STP completes and the topology is stable
6. SW 5 completes its STP and realizes that the best path back to the SW 1 (root) is now through SW 3 and begins forwarding to SW 3 while putting the port to SW 4 back into a *blocking* state.

Rapid Spanning Tree (802.1w)

Rapid Spanning Tree (RSTP) was developed to provide a faster convergence after a topology change. Unlike the original STP protocol which uses time intervals to decide whether or not the topology is stable and to change through transitioning states, RSTP uses a two-way communication between active ports so that each switch will keep a table of participating ports in Spanning Tree and can transition between transitioning states quickly.

RSTP was designed around the original STP, which means the same mechanisms used to determine the topology of the network as well as the root switch are still intact and backwards-compatible. Like the original 802.1d standard, there is a Common Spanning Tree (CST) for all VLANs created.

There are proprietary methods that allow for multiple Spanning Tree instances or a separate Spanning Tree instance per VLAN.

The IEEE standards based RSTP is defined as a single instance for all VLANS within an STP domain.

There are proprietary STP protocols that allow for a per-VLAN instance of STP, but these protocols will not operate between different vendors.

How Rapid Spanning Tree works

The initial RSTP convergence time, after all switches are connected and powered up, is similar to that of STP. However, once the network becomes stable and all switches agree on the current topology, any subsequent changes (e.g., link failure) are propagated rapidly without the need for Spanning Tree timers. Depending on the complexity of the network, the time it takes to establish the new topology may vary from tens of milliseconds to seconds.

When RSTP switches experience a topology change they immediately purge their forwarding tables and forward the TCN (Topology Change Notification) to other switches so they can do the same, bypassing the slower timeouts of the original STP. The quick convergence time is accomplished by aggressively figuring out the topology in the event of a failure, rather than using the traditional timeout values. In order to accomplish the quick convergence, RSTP does the following:

- Monitors MAC operational states and retires ports that are no longer functional.
- Processes inferior (not best path – Designated Port) BPDUs to detect topology changes.
- Keeps track of ports that provide alternate paths to the root bridge. If the root port fails, RSTP quickly makes the Alternate Port the new root port and begins forwarding through the Alternate Port without delay.
- Uses point-to-point links, which use a two way handshake (sync) rather than timers to transition the Designated Port to forwarding.

Rapid Spanning Tree port states

Rapid Spanning-Tree accomplishes the quick convergence by being able to quickly transition between the STP States. To do this, it consolidated the five port states of the original Spanning Tree into four port states.

- Discarding
- Learning
- Forwarding
- Disabled – As with the original Spanning Tree, this is a manual step and the port does not forward any traffic.

The following table shows the relationship of the port states and their functions relative to 802.1D and 802.1w.

STP (802.1D) port state	RSTP (802.1w) port state	Is port included in active topology?	Is port learning MAC Addresses?
Disabled	Discarding	No	No
Blocking	Discarding	No	No
Listening	Discarding	Yes	No
Learning	Learning	Yes	Yes
Forwarding	Forwarding	Yes	Yes

Rapid Spanning Tree port roles

The role is now a variable assigned to a given port. The root port and designated port roles remain, while the blocking port role is split into the backup and alternate port roles. The STP determines the role of a port based on BPDUs.

- Root Port — The port that receives the best BPDU on a bridge is the root port. This is the port that is the closest to the root bridge in terms of path cost.
- Designated Port — The port on the root bridge that sends out the best BPDU on a segment. This is the port that is Root Port receives the best BPDU to root from.
- Alternate Port — The port on a switch that receives BPDUs but it is the secondary path back to the root bridge.
- Backup Port — This port is similar to the Alternate Port in that it receives BPDUs but is receiving them from itself. In this instance the switch would have an active Designated Port sending out BPDUs downstream and receive the BPDU on a different port.
- Edge Port (configurable) — A port directly connected to an end station that cannot create bridging loops in the network. Therefore, the edge port directly transitions to the forwarding state and skips the listening and learning stages.

Multiple Spanning Tree (802.1s)

The IEEE came out with Multiple Spanning Tree Protocol (MSTP) to allow for a standards based per-VLAN STP.

Up until MSTP there were only proprietary per-VLAN STP protocols and which were not capable of running between different vendors.

Multiple Spanning Tree Protocol is an extension of the RSTP protocol and aims to further develop the usefulness of VLANs. MSTP is used to configure a separate Spanning Tree instance per VLAN or Multiple Spanning-Tree Instances (MSTI).

Link Aggregation

This section provides the following information:

- "Port Aggregation" on page 105
- "Link Aggregation Control Protocol (LACP)" on page 107
- "NIC teaming" on page 109
- "Load-balancing method" on page 111

Port Aggregation

In a highly-scalable campus LAN or Data Center network, using different Ethernet technologies is one of the options to increase the bandwidth. Ethernet line rates had evolved from 10 Mb/s (Ethernet) to 100 Mb/s (Fast Ethernet) to 1 Gb/s (Gigabit Ethernet) up to 10 Gb/s Ethernet. However, there are times that single links, for example 10 GE, are not adequate for carrying loads of traffic from multiple VLANs (VLAN trunk) or storage initiator/target (if Ethernet is used as iSCSI/FCoE transport). The bandwidth requirements of a VLAN trunk are high and it must accommodate the total bandwidth required of all the VLANs attached to that switch. The same is true with a high demanding iSCSI host connected to a switch.

If an iSCSI host belongs to a Gigabit Ethernet network sometimes 1 Gb/s of bandwidth is not enough. Different switches have different terms for their Port Aggregation implementation. Cisco uses Etherchannel and Brocade uses the Brocade LAG for their static Port Aggregation. There is also standards-based implementation of Port Aggregation, LACP (Link Aggregation Control Protocol). This is the dynamic way of configuring Port Aggregation and designed to be vendor neutral which means most, if not all, Ethernet switches support this protocol.

All implementations perform the same function, that is to bundle two or more ports as one high bandwidth logical port to provide incremental link speeds, link redundancy, link resiliency and load-balancing. Technically, you can have multiple ports to create parallel trunk links between switches but Spanning Tree treats these as a loop and shuts down all but one link to eliminate the loop. Port Aggregation *prevents* this situation by bundling the ports into a single, logical link, which can act as either an access (for connecting to host) or a trunk link (for carrying multiple VLANs).

Switches or hosts on each end of the aggregated link must understand and use a common Port Aggregation technology for compatibility and proper operation. load-balancing is done through a hashing algorithm that allows different traffic patterns to be distributed across the individual links within the bundle. Port Aggregation also provides redundancy with several bundled physical links. If one of the links in the bundle fails, traffic sent through that link moves to an adjacent link. Failover to the adjacent links occurs in less than few milliseconds. Since the bundled link consists of more than one link, the Port Aggregation link will stay up even one of the physical links went down. This event will cause almost no performance impact to the users. As more links fail, more traffic moves to further adjacent links. Similarly, as links are restored, the load redistributes among the active links.

Figure 44 shows Port Aggregation between two switches. Each port is 10 Gigabit Ethernet links, thus providing an aggregated 20 Gb/s link.

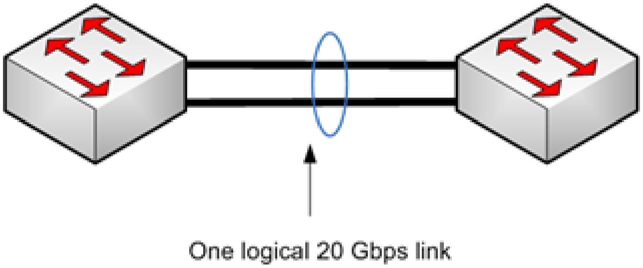


Figure 44 Port Aggregation between two switches

Figure 45 shows Port Aggregation between a switch and host. Each port is 1 Gigabit Ethernet links, thus providing an aggregated 2 Gb/s link.

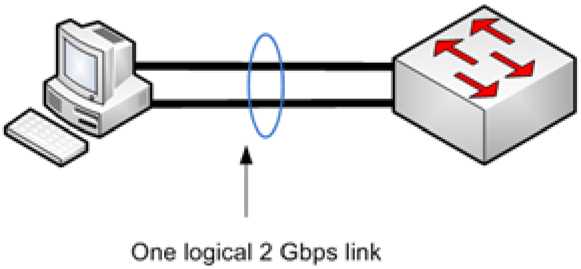


Figure 45 **Port Aggregation between a switch and host**

When selecting the interfaces to be part of the Port Aggregation or port channel, each port should be compatible. It requires checking the operational attributes of an interface before allowing it to participate in the port aggregation group, which is usually called channel group. The compatibility check includes the following operational interface attributes:

- Port mode
- Access VLAN
- Trunk native VLAN
- Allowed VLAN list
- Speed

Link Aggregation Control Protocol (LACP)

Link Aggregation Control Protocol (LACP) works by sending frames (LACPDUs) down all links that have the protocol enabled. If it finds a device on the other end of the link that also has LACP enabled, it will also independently send frames along the same links enabling the two units to detect multiple links between themselves and then combine them into a single logical link.

LACP can be configured in one of two modes: active or passive. In active mode it will always send frames along the configured links. In passive mode however, it acts as "speak when spoken to", and therefore can be used as a way of controlling accidental loops (as long as the other device is in active mode).

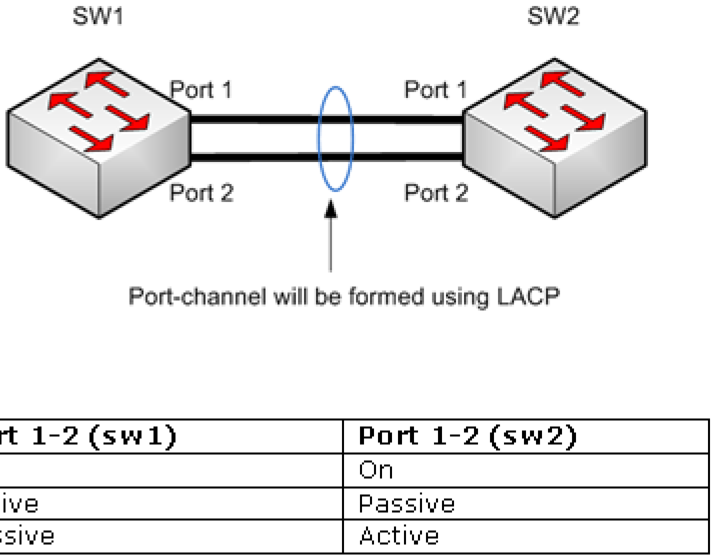
LAC, defined in IEEE 802.3ad, is a protocol that allows a switch to negotiate an automatic link bundling of ports by sending LACP frames to its peer. These frames are exchanged between switches over port channel (Port Aggregation) capable switch ports. The identification of neighbors and port group capabilities is learned and compared with local switch capabilities, and then LACP assigns roles to the port channel's end points.

The switch with the lowest system priority is allowed to make decisions about what ports are actively participating in the Port Aggregation at a given time. Ports are selected and become active according to their port priority value. For example, a set of up to eight capable links can be defined for each port channel. Through LACP, a switch selects up to four of these eight ports having the highest port priorities as active port channel links at any given time. Usually the way priority is implement is the lower the numeric value the higher the priority. The other 8 links are placed in a hot-standby state and will be enabled in the port channel if one of the active links goes down.

Port priorities are configurable, but if it is not configured certain default values are used by different vendors. If ports are using the same values it means they are competing with each other. Usually vendors implement a tie breaker like lower port numbers are used to select the active ports, i.e., port 1/1 is higher than port 1/5.

Each interface included in a single port channel bundle must be assigned to the same unique channel group number. LACP automatically configures an administrative key value equal to the channel-group number on each port configured to use LACP. This administrative key defines the ability of a port to aggregate with other ports. A port's ability to aggregate with other ports is determined by bandwidth, duplex capability, and the point-to-point or shared medium state. Channel negotiation must be set to on (unconditionally channel; no LACP negotiation), passive (passively listen and wait to be asked), or active (actively ask).

Figure 46 shows valid port configuration on both sides to be part of LACP channel group.



Port 1-2 (sw1)	Port 1-2 (sw2)
On	On
Active	Passive
Passive	Active

Figure 46 **Valid port configuration**

NIC teaming

Servers in a network environment or data center have different availability requirements. If they host business-critical applications and process-intensive applications the requirement for a highly available solution is very high. This can be implemented using load balancers, dual connections to two access switches and network adapter teaming, also referred to as *NIC teaming*.

The problem with the first and second solution is the scope of redundancy within the hierarchical layers of LAN design model. The redundancy happens at the access layer and not on the server nodes.

NIC teaming, however, provides server level high availability. NIC teaming is implemented using two or more NICs installed on a single server, which can be referred to as a dual-attached server. The deployment of dual-attached servers with NIC teaming can help to push the concept of high availability from the core of the network to each individual server in the data center server farm.

The possible modes for the deployment of dual-attached servers are fully dependent on the NIC teaming software provided by the adapter vendor. Usually it provides rudimentary high availability features like fault tolerance and load-balancing. Different algorithms are used to do the load-balancing, usually based on source or destination Mac Address, IP address, TCP/UDP port, or sometimes combination of both source and destination.

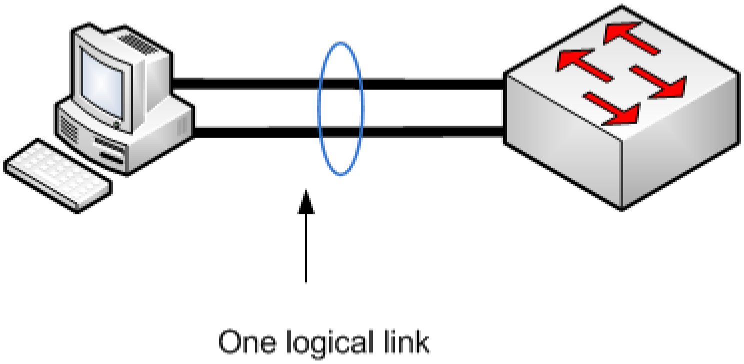


Figure 47 **One logical link**

As shown in Figure 47, the switch treats the redundant NICs as one logical NIC only. When host NICs are grouped together as a load-balancing team, a virtual adapter instance is created. You configure the virtual adapter with a unique IP address and it automatically assumes the Mac Addresses of both server adapters (for the outgoing traffic). The IP address of the virtual adapter is advertised along with the Mac Address of each adapter but it is recommended that you assign a Mac Address to the virtual adapter.

With this configuration if the main adapter fails, the remaining one will be able to receive traffic destined to the same MAC as the previous adapter. Under normal conditions, both the primary adapter and the secondary adapter remain in active state and is doing load-balancing. Under failure conditions, the failed adapter transitions from an active state to a disabled state while the other adapter remains active. The remaining link receives and transmits traffic and answers ARP requests for the IP address of the virtual adapter. Network adapter or link failures are detected by probe responses or by monitoring link status and link activity.

Load-balancing method

Traffic in a Port Aggregation is spread out across the individual bundled links in a deterministic fashion. Frames are forwarded on a specific link as a result of a hashing algorithm. In some switches the algorithm used can be configured on the switch with the following methods to load balance across the links:

- Destination MAC Address
- Source MAC Address
- Source and destination MAC Address
- Destination IP Address
- Source IP Address
- Source and destination IP Address
- Destination TCP/UDP port number
- Source TCP/UDP port number
- Source and destination TCP/UDP port number

For example, if there are two links in a Port Aggregation and you used source MAC Address, the last bit of the Address will be used as an index in the load-balancing mechanism (see Table 3). If a Port Aggregation comprises of four links you would need four index thus we need to have four possible combinations. In this case last two bits will be used to have four available indexes: 00, 01, 10, and 11. These will give use the load-balancing effect.

Table 3 shows frame distribution on a two-link Port Aggregation using a MAC Address (translated to bit).

Table 3 Frame distribution on a two-link Port Aggregation

Binary addresses	Two-link Port Aggregation and link number
Address1: ... xxxxxxx0	Use link0
Address2: ... xxxxxxx1	Use link1

Table 4 shows frame distribution on a four-link Port Aggregation using a MAC Address (the last two bits are used as an index).

Table 4 Frame distribution on a four-link Port Aggregation

Binary addresses	Four-link Port Aggregation and link number
Address1: ... xxxxxx00	Use link0
Address2: ... xxxxxx01	Use link1
Address3: ... xxxxxx10	Use link2
Address4: ... xxxxxx11	Use link3

Table 5 shows frame distribution on a two-link Port Aggregation using the source and destination MAC Address by performing XOR.

Table 5 Frame distribution on a two-link Port Aggregation by performing XOR

Binary addresses	Two-link Port Aggregation XOR and link number
Address1: ... xxxxxxx0 Address2: ... xxxxxxx0	... xxxxxxx0: Use link 0
Address1: ... xxxxxxx0 Address2: ... xxxxxxx1	... xxxxxxx1: Use link 1
Address1: ... xxxxxxx1 Address2: ... xxxxxxx0	... xxxxxxx1: Use link 1
Address1: ... xxxxxxx1 Address2: ... xxxxxxx1	... xxxxxxx0: Use link 0

Access Control Lists

Access Control Lists (ACLs) can be used for many network-related operations, including router management, controlling route access, filtering debug output on the CLI interface, and controlling exterior gateway routing protocol attributes such as BGP AS-path. ACLs can also be used for defining traffic to Network Address Translation (NAT) or filtering non-IP protocols. Depending on the IOS features installed on the network device, ACLs can also be used for encryption.

In FCoE switches, since these are deployed mainly on access layers, ACLs are primarily used for traffic filtering and controlling router management. Filtering can be done based on IP, MAC, VLAN, and even UDP/TCP numbers.

ACLs are a set of rules and actions separated with sequence numbers and read from top to bottom, or top-down. These actions are called Access Control Entries (ACEs). Each ACE performs set of conditions deciding whether a frame will be permitted or denied. A condition must be satisfied before a rule is performed. In each rule, you can specify the source and the destination of the traffic that matches the rule. You can specify both the source and destination as a specific host, any hosts, a group of hosts, or the whole network (whole subnet).

Implicit deny

ACLs have implicit rules. These are rules that do not appear in the running configuration. The switch applies them to traffic when no other rules in an ACL match. All IP ACLs (IPv4) include the following implicit rule, "deny ip any any". This implicit rule ensures that the switch denies unmatched IP traffic. When performing IP ACL, "permit ip any any" must be included at the end of the ACL in order to permit other traffic. If this line is not explicitly added, all traffic will be denied since there is always an implicit rule "deny ip any any" at the end of each ACL.

Wildcard mask

The wildcard mask is used to include many addresses in a policy statement or ACE. For example:

Example 1

```
deny ip 10.0.0.0  0.0.0.255
```

Example 2

```
deny ip 10.0.0.1  0.0.0.0
```

- The first ACE example shows that all hosts on network 10.0.0.0/24 are included in that deny statement.
- The second ACE example shows that only the host 10.0.0.1 is being denied or blocked.

The wildcard mask is interpreted as a bit mask wherein the value bit of 1 means match *anything* in the corresponding bit areas in the IP address. The 0 (zero) value bit means match the IP address *exactly* in the same bit position. Therefore, if you want to block only a specific host, you would need an all zeros mask, which is 0.0.0.0, as shown in "Example 2".

Another way of specifying a single IP address on an ACE is by using the keyword *host*. Instead of using the **deny ip 10.0.0.1 0.0.0.0** command, you can use **deny ip host 10.0.0.1**. On the other hand, if you want to block the entire subnet or the whole range of the network block, you would make the last octet (or last few bits) all **1**s to match anything on that octet.

Logical operators

ACL rules for TCP and UDP traffic can use logical operators to filter traffic based on port numbers. The most commonly used operator is the eq (*equal*) operator. These operators are used to match a specific UDP/TCP port, port ranges, or ports that are not in the statement (using the neq (*not equal)* operator). The following is a list of ACL operators used in ACEs:

Operator	Description
any	Any destination address
eq	Match only packets on a given port number
gt	Match only packets with a greater port number
lt	Match only packets with a lower port number
neq	Match only packets not on a given port number
range	Match only packets in the range of port numbers

Examples of these operators include the following:

Example 1

```
permit udp host 192.168.0.1 any eq tftp
```

Example 2

```
deny tcp host 172.16.10.1 host 10.10.10.1 gt 100
```

- The first example allows tftp traffic from host 192.168.0.1 to any host.
- The second example blocks tcp port numbers from 100 and beyond, from host 172.16.10.1 to host 10.10.10.1.

Operators help minimize the number of ACEs, decreasing the number of lines the switch NX-OS/Fabric OS needs to parse, thereby lessening the burden of the switch.

Implementing security

Security can be implemented through traffic filtering. Controlling network access, such as denying specific hosts from unnecessary or unwanted traffic, is the most common application for ACLs. Figure 48 on page 115 shows an example of IP ACL.

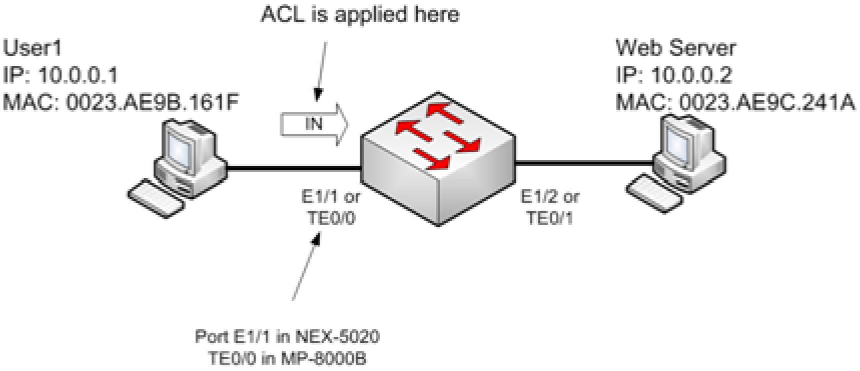


Figure 48 **IP ACL example**

In Figure 48, the network administrator blocks all the traffic coming from node *User1* to node *Web Server*. A standard IP ACL will block all traffic types from the source. If extended ACL is used, specific traffic types using UDP/TCP numbers can be blocked explicitly.

IP-based ACL

To create an IP ACL in Nexus 5020:

1. Create an ACL name and create the ACE to permit/deny a specific host. More than one entry is allowed to deny two or more hosts:

```
Nexus 5020 (config)# ip access-list ACL_name
Nexus 5020 (config-acl)# deny ip 10.0.0.1 0.0.0.0 any
Nexus 5020 (config-acl)# permit ip any any
Nexus 5020 (config-acl)# statistics
```

2. Apply the ACL as an inbound ACL on an interface:

```
Nexus 5020 (config)# interface ethernet 1/1
Nexus 5020 (config-if)# ip port access-group ACL_name in
```

IP ACLs on the MP-8000B are not currently supported. There is some indirect IPv4 ACL support in MP-8000B, but it is still using MAC as its source or destination in the ACE.

The following example is based on Figure 48 on page 115:

```
seq 1 deny host 0023.ae9b.161f any ipv4
permit any any
```

In this example, the ACE is telling all IP version 4 traffic coming from the host with the MAC address 0023.ae9b.161f that it will not be allowed into the network. ACE **permit any any** was added to allow other traffic, since there is always implicit deny at each ACL.

MAC-based ACL

MAC ACLs can be used as substitute to IP ACLs.

To create a MAC ACL on a Nexus 5020 switch:

1. Create an ACL name and create the ACE to permit/deny a specific host. More than one entry is allowed to deny two or more hosts. The **statistics** command is used to see the number of hits matching the specific ACL.

```
Nexus 5020 (config)# mac access-list ACL_name
Nexus 5020 (config-mac-acl)# deny 0023.ae9b.161f 0000.0000.0000 any
Nexus 5020 (config-mac-acl)# statistics
```

2. Apply the ACL as inbound ACL on an interface:

```
Nexus 5020 (config)# interface ethernet 1/1
Nexus 5020 (config-if)# mac port access-group ACL_name
```

To create a MAC ACL on the MP-8000B switch:

1. Create an ACL name and then create ACE to permit/deny a specific host. More than one entry is allowed to deny two or more hosts. *Count* is added to record the number of hits matching the ACL rule.

```
MP-8000B (config)# mac access-list standard ACL_name
MP-8000B (conf-macl-std)# seq 1 permit 0023.ae9b.161f count
```

2. Apply the ACL as an inbound ACL on an interface:

```
MP-8000B (config)# interface TenGigabitEthernet 0/0
MP-8000B (conf-if-te-0/0)# mac access-group ACL_name in
```

3. View the ACL statistics, using the **show statistics access-list mac** command:

```
MP-8000B # show statistics access-list mac mac_ext_acl
mac access-list extended mac_ext_acl on interface Te 0/0
 seq 1 deny host 0efc.0001.0801 any 8100
 seq 2 deny host 0efc.0001.0801 any 8914
mac access-list extended mac_ext_acl on interface Te 0/3
 seq 1 deny host 0efc.0001.0801 any 8100
 seq 2 deny host 0efc.0001.0801 any 8914
mac access-list extended mac_ext_acl on interface Te 0/0
 seq 1 deny host 0efc.0001.0801 any 8100
 seq 2 deny host 0efc.0001.0801 any 8914
mac access-list extended mac_ext_acl on interface Te 0/3
 seq 1 deny host 0efc.0001.0801 any 8100
 seq 2 deny host 0efc.0001.0801 any 8914
MP-8000B #
```

Extended MAC ACL in MP-8000B

Extended MAC ACL is an MP-8000B feature. Using the available parameters shown in the next table can give the same flexibility as available in the Nexus 5020:

Extended ACL options	Description
arp	EtherType: ARP (0x0806)
count	Packet count
fcoe	EtherType: FCOE (0x8906)
ipv4	EtherType: IPv4 (0x0800)
<1536-65535>	EtherType: Custom value between 1536 and 65535

FCoE filtering

Using the custom EtherType values in ACEs provide flexible configuration options, including FCoE filtering. FCoE filtering can be accomplished by using the 8906 Ethertype header. To block FCoE, use one of the following commands:

seq 1 deny host <source mac> any 8906 count

or

seq 1 deny host <source mac> any fcoe count

Figure 49 shows an example of FCoE traffic filtering.

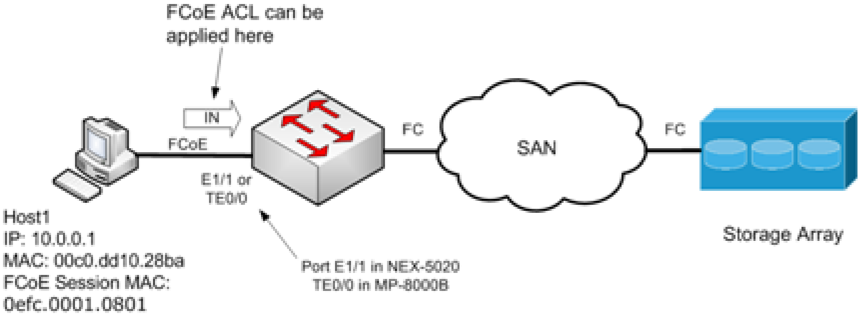


Figure 49 FCoE traffic filtering example

To filter FCoE traffic, complete the following steps:

1. Create an ACL name.
2. Create ACE to deny FCoE frames from host1.

> **Note:** More than one entry is allowed to deny two or more hosts. *Count* is added to record the number of hits matching the ACL rule.

```
MP-8000B (config)# mac access-list extended ACL_name
MP-8000B (conf-macl-ext)# seq 1 deny host 0e:fc:00:01:08:01 any 8906 count
```

Use the session MAC address of the CNA, not the physical or "actual" MAC address. This session MAC can be seen on the Fabric Login table using the **fcoe –loginshow** command:

```
MP-8000B:admin> fcoe --loginshow
==============================================================================
Port   Te port        Device WWN             Device MAC        Session MAC
==============================================================================
8      Te 0/0    21:00:00:c0:dd:10:28:bb  00:c0:dd:10:28:bb  0e:fc:00:01:08:01

MP-8000B:admin>

MP-8000B # show mac-address-table
VlanId   Mac-address         Type        State      Ports
1          00c0.dd10.28ba     Dynamic    Active     Te 0/0
1          0201.0000.0000     Dynamic    Active     Te 0/0
1002       0efc.0001.0801     Dynamic    Active     Te 0/0
```

3. Apply the ACL as an inbound ACL on an interface:

```
MP-8000B (config)# interface TenGigabitEthernet 0/0
MP-8000B (conf-if-te-0/0)# mac access-group ACL_name in
```

VLAN Access Control List (VACL)

A VLAN Access Control List (VACL) is needed since only traffic that passes between VLANs can be filtered using ACLs. A VACL is like a pointer to an ACL. Simply put, VACL is an access map that links to an IP ACL or MAC ACL. An action is performed on the VACL after mapping the IP or MAC ACL. Only the permitted traffic in the IP or MAC ACL will be accepted or denied by the configured action by the VACL. VACL is applied to the bridged VLAN segment. VACLs are not applied inbound or outbound interface.

The example in Figure 50 shows IP traffic from host1 is being blocked inside VLAN100. Notice that the VACL saves time in creating IP ACLs equal to the number of ports in that VLAN.

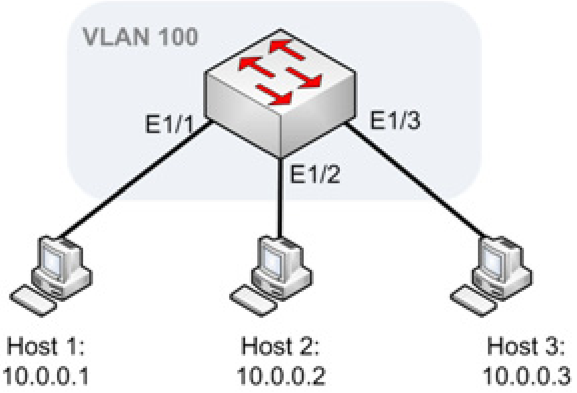


Figure 50 VACL example

To create a VACL in Nexus 5020, complete the following steps:

1. From the switch, create a standard IP ACL to specify the host 1:

```
Nexus 5020 (config)# ip access-list IP_ACL_name
Nexus 5020 (config-acl)# permit ip 10.0.0.1 0.0.0.0 any
Nexus 5020 (config-acl)# exit
```

2. Create an access-map to call the IP ACL you just created.

   a. Use the map statement to call the IP ACL.

   b. Specify the action that the switch applies to the traffic that matches the ACL:

```
Nexus 5020 (config)# vlan access-map VACL_name
Nexus 5020 (config-access-map)# match ip address IP_ACL_name
Nexus 5020 (config-access-map)# action drop
Nexus 5020 (config-access-map)# exit
```

3. Apply the VACL to the VLAN by the VLAN list you specified:

```
Nexus 5020 (config)# vlan filter access-map VACL_name vlan-list 100
```

SNMP security

Simple Network Management Protocol (SNMP) is used for managing the network devices. Its primary task is to allow a host to get statistics from any network node like hosts, switches, and routers. An MIB is used as a primary resource in SNMP. This management protocol uses UDP as transport protocol in its communications because it has lower overhead, is lightweight, and is simple.

SNMP has two types of objects:

- Read-only, used for primarily on debugging
- Write-enabled, allows for changes to be made to the network device

Both types can be found on network management suites that use SNMP to manage and configure routers, switches, and firewalls. However, many network administrators do not disable the SNMP default settings, thus creating a vulnerability that an attacker can exploit to gain important network information and the ability to change and reconfigure the network. Therefore, SNMP should be considered when implementing security policies. Restriction to SNMP access can be created using ACLs with explicit deny statements, or ACEs. A policy set can be included with the list of hosts that have the authority to have SNMP access to the device, and restrict the SNMP access for those hosts to the SNMP management stations.

The following example shows ACEs that permit SNMP access from two hosts:

1. Create an ACL name, and then create ACEs to permit Simple Network Management Protocol (161) and SNMP Traps (162) from the 10.0.0.1 host and 10.0.0.2 host.

```
ip access-list ACL_name
permit udp host 10.0.0.1 any eq snmp
permit udp host 10.0.0.1 any eq snmptrap
permit udp host 10.0.0.2 any eq snmp
permit udp host 10.0.0.2 any eq snmptrap
```

2. Apply the ACL on an interface:

```
interface ethernet 1/1
ip port access-group ACL_name in
```

Useful ACL commands

The following table lists some useful ACL commands.

Command	Description
show ip access-lists	Displays the IP ACL configuration
show mac access-lists	Displays the MAC ACL configuration
show access-lists <ACL_name> summary	Displays the summary information about the ACL, i.e. ACL name, interface where it was applied.
show access-lists <ACL_name>	Displays the detailed information about ACL, i.e. ACEs, number of matches and all information in *show access-lists <ACL_name> summary* command.
show run interface <intf>	Shows the interface configuration including what ACL is applied.
show running-config aclmgr	Shows ACL-specific running configurations.

ACL debugging

Debugging options can be enabled on the logs for more detailed and real-time events relevant to a certain feature happening on the switch. Debugging is not commonly used, unless a problem is suspected with a feature or an interaction with other switches in the network. Debug commands should be used cautiously since they can generate a huge amount of logs. The debug process itself can affect the switch's CPU and memory performance to the degree that it severely impacts traffic switching and forwarding. Therefore, always be sure to turn off the debug command when you are done using it.

The following are debugging options for ACL. Normally "all" is used to view and log all the events related to ACL.

```
NX-5010-16# debug aclmgr ?
  all         Configure all debug flags of aclmgr
  errors      Configure debugging of aclmgr errors
  events      Configure debugging of aclmgr events
  fsm         Configure debugging of aclmgr FSM events
  ha          Configure debugging of aclmgr HA
  ppf         Configure PPF debugs
  trace       Configure debugging of aclmgr trace
```

Ethernet fabric

This section contains the following information:

- "Ethernet fabric overview" on page 122
- "Transparent Interconnect of Lots of Links (TRILL)" on page 123
- "Brocade VCS Fabric technology" on page 123

Ethernet fabric overview

Compared to the traditional Cisco three layer hierarchical model, Ethernet fabric provides higher levels of network performance and utilization, availability, and simplicity.

Ethernet fabrics have the following characteristics that are improved over classical Ethernet:

- Flatter — Ethernet fabrics eliminate the need for Spanning Tree Protocol (STP), Ethernet fabrics can attach to traditional Ethernet networks, but it is huge overstatement to say they are completely interoperable.
- Greater flexibility — Ethernet fabrics can be designed in any topology to best meet different network needs and requirements.
- Better resiliency — Multiple "least cost" paths are used for high performance and high reliability.
- Improved performance — With TRILL the shortest paths through the network are all active, and traffic is automatically distributed across the equal-cost paths, unlike classic Ethernet running Spanning Tree Protocol, only 50% of the links are forwarding traffic while the rests are blocked and waiting for the primary link to fail.
- Easier to scale — Ethernet fabrics easily scale up and down on demand.

More advanced Ethernet fabrics borrow further from Fibre Channel fabric concepts. They are self-forming and function as a single logical entity, in which all switches automatically know about each other and all connected physical and logical devices. Therefore, management can be domain-based rather than device-based and defined by policy rather than by repetitive procedures.

Transparent Interconnect of Lots of Links (TRILL)

TRILL (Transparent Interconnect of Lots of Links) is an IETF standard for improved bridging loop prevention and having Layer 2 multipathing function in an Ethernet fabric. Unlike the Spanning Tree Protocol, with TRILL, all the paths through the network are active, and traffic is automatically distributed across the equal-cost paths.

Overall Ethernet fabric performance is improved while delivering bridging loop prevention on the network.

TRILL is implemented by devices called RBridges (Routing Bridges) or TRILL switches. It combines the advantages of routers and bridges. TRILL is the application of link state routing protocol to the VLAN-aware customer-bridging problem. RBridges are compatible with the legacy IEEE 802.1 Ethernet bridges. They are also compatible with IPv4 and IPv6 routers and end nodes. They are invisible to current IP routers and, like routers, RBridges terminate the bridge spanning tree protocol.

Brocade VCS Fabric technology

This section contains the following information:

- "VCS Fabric technology overview" on page 123
- "Distributed intelligence" on page 125
- "Logical chassis" on page 125
- "Examples of VCS deployments in a data center" on page 126
- "References" on page 131

VCS Fabric technology overview

The Brocade VCS Fabric is a new Layer 2 Ethernet technology. It leverages the emerging TRILL standard as well as other standards from IEEE and T11, such as Data Center Bridging (DCB) and Fibre Channel over Ethernet (FCoE). VCS eliminates many limitations of classic Ethernet networks in the data center.

Figure 51 on page 124 shows a classic Ethernet architecture and the corresponding Brocade VCS Fabric architecture. The Brocade VCS Fabric combines the Access layer and Aggregation layers. It is more scalable especially as you add and expand the network.

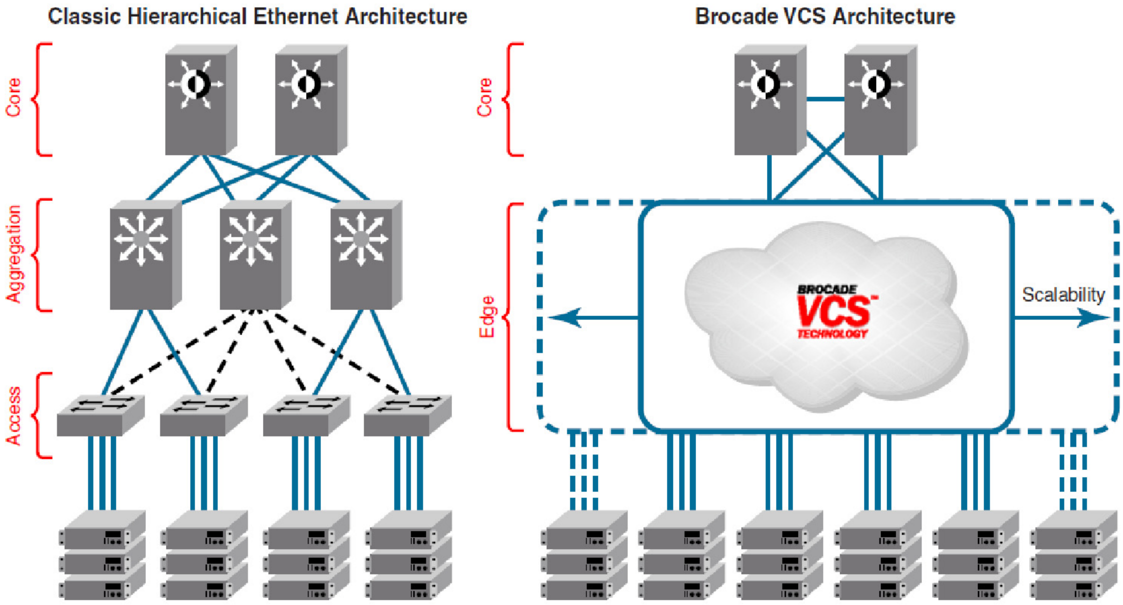


Figure 51 **Classic Ethernet and corresponding VCS Fabric architecture**

Brocade VCS Fabric technology comprises the following concepts:

- Ethernet fabric, discussed on page 122
- Distributed intelligence, discussed on page 125
- Logical chassis, discussed on page 125

When two or more Brocade VCS Fabric mode-enabled switches (such as VDX 6720 or 6730) are connected together, they form an Ethernet fabric and exchange information among each other to implement distributed intelligence. To the rest of the network, the Ethernet fabric appears as a single logical chassis.

Examples of VCS deployments in a data center can be found on page 126.

For Brocade VDX switch setup examples, refer to the *Fibre Channel over Ethernet (FCoE) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Distributed intelligence

With Brocade VCS Fabric technology, all relevant information is automatically distributed to each member switch to provide unified fabric functionality. For example, when a host connects to the fabric for the first time, all switches in the fabric learn about that server. In this way, fabric switches can be added or removed and physical or virtual servers can be relocated-without the fabric requiring manual reconfiguration.

Distributed intelligence has the following characteristics:

- The fabric is self-forming. When two Brocade VCS Fabric mode-enabled switches are connected, the fabric is automatically created and the switches discover the common fabric configuration.
- The fabric is masterless. No single switch stores configuration information or controls fabric operations. Any switch can fail or be removed without causing disruptive fabric downtime or delayed traffic.
- The fabric is aware of all members, devices, and VMs. If the VM moves from one Brocade VCS Fabric port to another Brocade VCS Fabric port in the same fabric, the port-profile is automatically moved to the new port.

Logical chassis

Regardless of the number of VCS Fabric mode-enabled switches the VCS Fabric, they are going to be managed as if they were a single logical chassis. From the visibility of the network, the fabric looks no different than any other Ethernet switch.

Figure 52 on page 126 shows an Ethernet fabric with two switches. The rest of the network is aware of only the edge ports in the fabric, and is unaware of the connections within the fabric. Each physical switch in the fabric is managed as if it were a blade in a chassis. When a Brocade VCS Fabric mode-enabled switch is connected to the fabric, it inherits the configuration of the fabric and the new ports become available immediately.

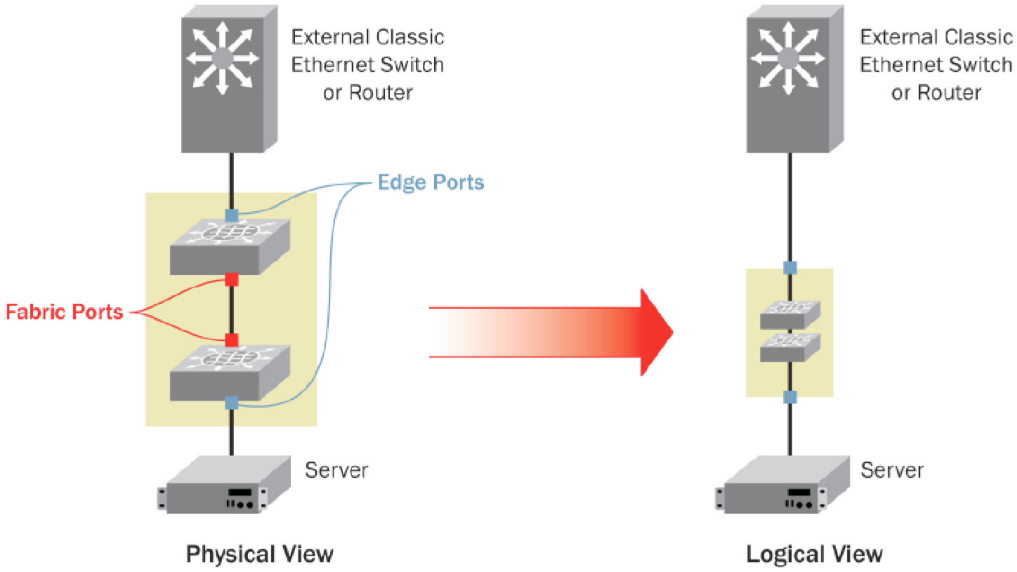


Figure 52 VCS logical chassis

Examples of VCS deployments in a data center

Brocade VCS fabric technology can be used in different locations in the network. Traditionally, data centers are built using three-tier architectures with access layers providing high port densities for server connectivity, aggregation layers for security, and aggregating the access layer devices, and the core layer linking the campus network and data center network.

This section describes different VCS deployments in various locations in the network and in data center infrastructure.

Note: For Brocade VDX switch setup examples, refer to the *Fibre Channel over Ethernet (FCoE) Case Studies TechBook* at http://elabnavigator.EMC.com, , **Documents> Topology Resource Center**.

Example 1 VCS Fabric technology with native FC SAN

Fibre Channel ports on the Brocade VDX 6730 provide support for connecting a Brocade VCS Fabric to a native Fibre Channel SAN. Fibre Channel routers provide the connectivity, which provides access to Fibre Channel devices while preserving isolation between the fabrics.

Brocade zoning allows you to determine which FCoE devices can access which storage devices on the Fibre Channel SAN. An example is shown in Figure 53.

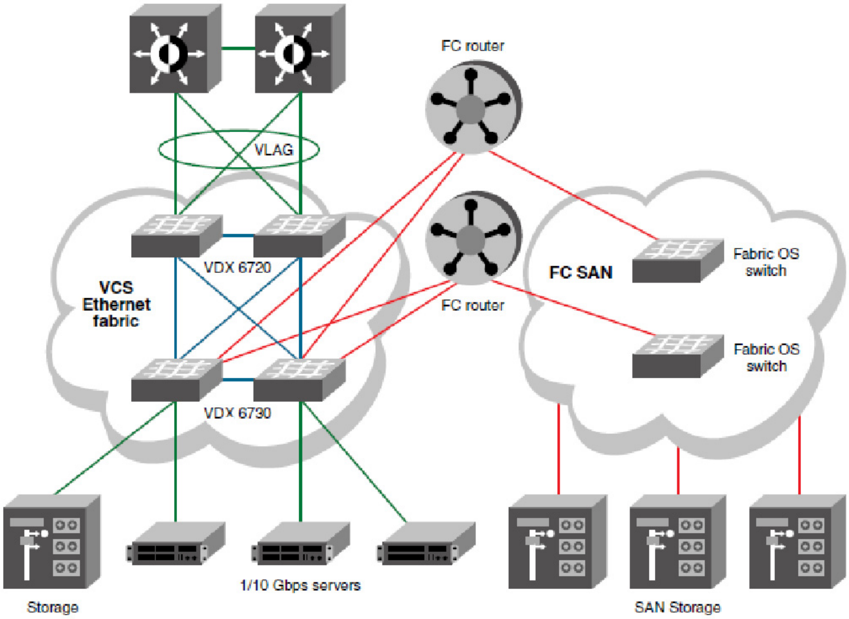


Figure 53 VCS Fabric Technology with Native FC SAN example

Example 2 VCS Fabric technology in the access layer

Figure 54 on page 128 demonstrates a typical deployment of VCS Fabric technology in the access layer. In this layer, VCS fabric technology can be inserted in existing design, as it fully interoperates with existing LAN protocols, services, and architecture. In addition, VCS Fabric technology brings greater performance by allowing active-active server connectivity to the network without additional management overhead. At the access layer, VCS Fabric technology allows Gigabit Ethernet and 10 Gigabit Ethernet server connectivity and flexibility of oversubscription ratios, and it is completely auto-forming, with zero configuration. Servers see the VCS as a single switch and can fully utilize the provisioned network capacity, thereby doubling the bandwidth of network access.

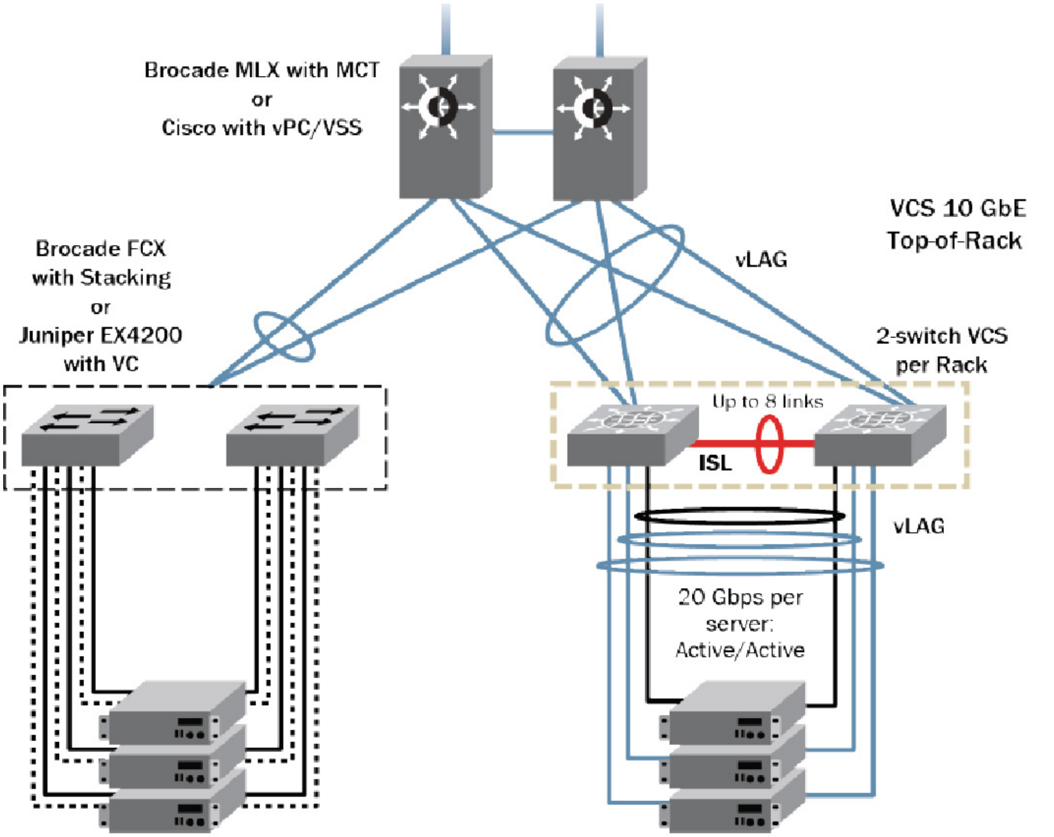


Figure 54 VCS Fabric Technology in the access layer example

Example 3 VCS Fabric technology in collapsed access/aggregation layer

Traditionally, Layer 2 networks have been broadcast traffic-heavy, which forced the data center designers to build smaller L2 domains to limit both broadcast domains. However, in order to seamlessly move VMs in the data center, it is absolutely essential that the VMs are moved within the same Layer 2 domain. In traditional network designs, therefore, VM mobility is severely limited to these small L2 domains. By using Transparent Interconnection of Lots of Links (TRILL)-based VCS Fabric technology, these issues are minimized in the data center.

Figure 55 shows how a scaled-out self-aggregating data center edge layer can be built using VCS Fabric technology. This architecture allows customers to build resilient and efficient networks by eliminating STP, as well as drastically reducing network management overhead by allowing the network administrator to manage the whole network as a single logical switch.

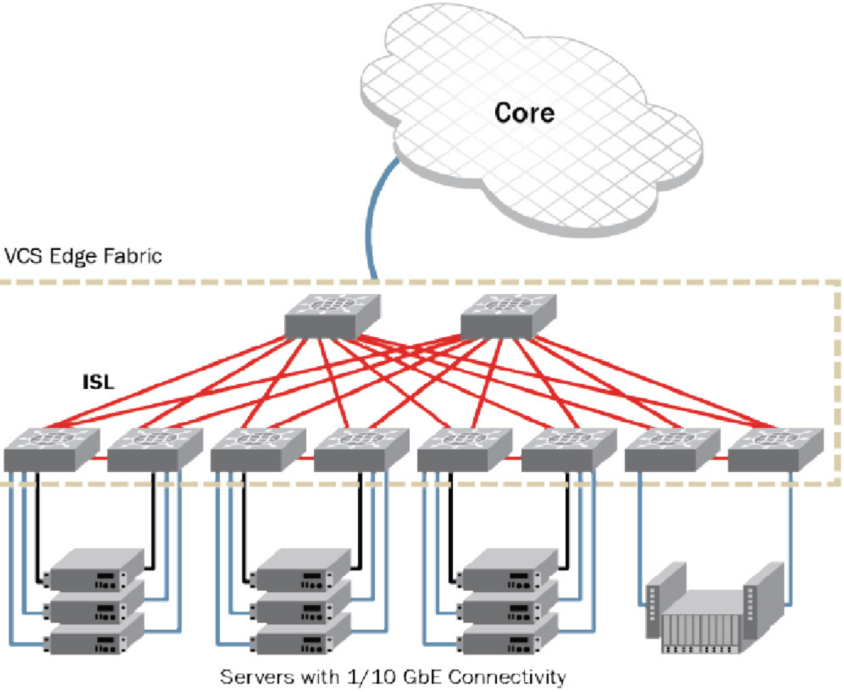


Figure 55 VCS Fabric technology in collapsed access/aggregation layer example

Example 4 VCS Fabric technology in a virtualized environment

Depending on the hypervisor in use, when a VM moves within a data center, the server administrator needs to open a service request with the network admin to provision the machine policy on the new network node where the machine is moved. This policy may include, but is not limited to, VLANs, Quality of Service (QoS), and security for the machine. VCS Fabric technology eliminates this provisioning step and allows the server admin to seamlessly move VMs within a data center by automatically distributing and binding policies in the network at a per-VM level, using the Automatic Migration of Port Profiles (AMPP) feature. AMPP enforces VM-level policies in a consistent fashion across the fabric and is completely hypervisor-agnostic. The example in Figure 56 shows the behavior of AMPP in a 10-switch VCS fabric.

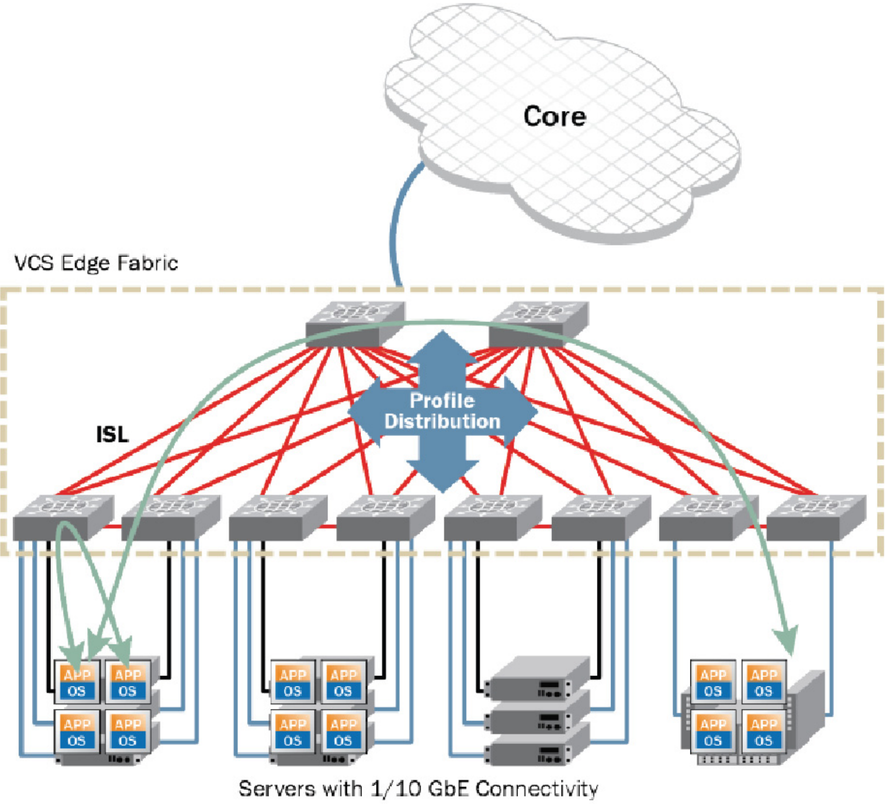


Figure 56 VCS Fabric technology in a virtualized environment example

Example 5 VCS Fabric technology in converged network environments

VCS Fabric technology allows for lossless Ethernet using DCB and TRILL, which allows VCS Fabric technology to provide multihop, multipath (for load-balancing), highly reliable and resilient FCoE and iSCSI storage connectivity. Figure 57 shows a sample configuration with FCoE and iSCSI storage connected to the fabric.

FCoE and iSCSI in a VCS Fabric with configuration examples can be found in the "VCS LAN/SAN convergence case study" in the *Fibre Channel over Ethernet (FCoE) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

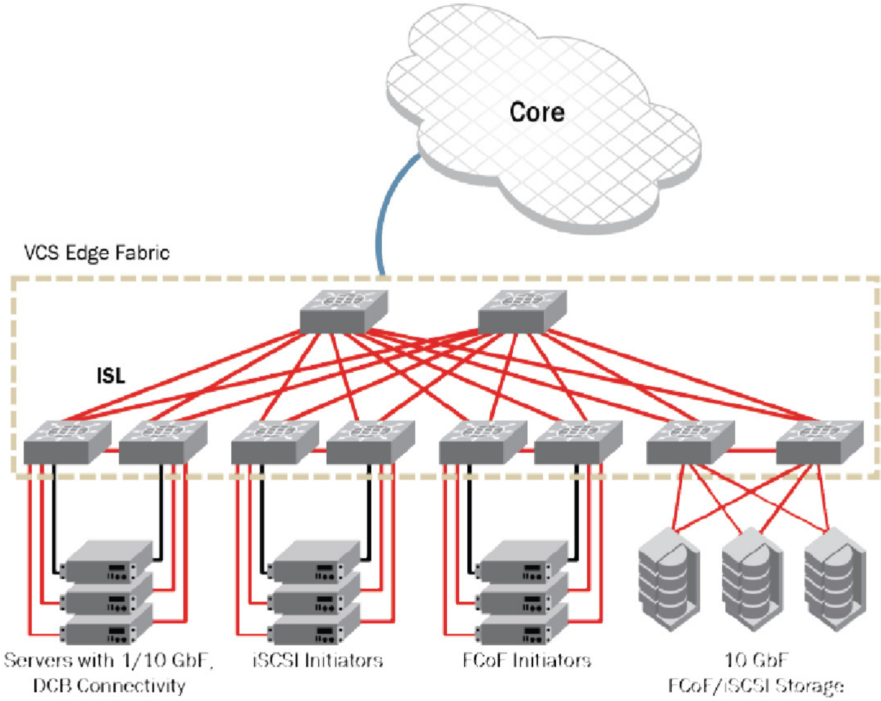

Core
VCS Edge Fabric
ISL
Servers with 1/10 GbE, DCB Connectivity
iSCSI Initiators
FCoE Initiators
10 GbE FCoE/iSCSI Storage

Figure 57 VCS Fabric technology in converged network environments example

References

For Brocade VDX switch setup examples, refer to the *Fibre Channel over Ethernet (FCoE) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Much of the information provided in this section was derived from the Brocade website, http://www.brocade.com, which provides details on VCS Fabric technology, its technical architecture, Ethernet fabrics, configuration guides, guides, case studies, and deployment scenarios.

- Refer to the *Network OS Administrator's Guide,* located at www.brocade.com, for information on the following:
  - Brocade VCS Fabric formation
  - Ethernet fabrics
  - Automatic Migration of Port Profiles
  - Configuring Classic Ethernet or IEEE 802.x standards like STP, VLANs, Link Aggregation, ACL, IGMP, etc.
  - Configuring Fibre Channel (for example, Zoning and FC ports)
  - Configuring vLAGs
- Refer to the *Network OS Command Reference*, locatedat www.brocade.com, for information on the following:
  - Configuring Classic Ethernet or IEEE 802.x standards like STP, VLANs, Link Aggregation, ACL, IGMP, etc.
  - Configuring Fibre Channel (for example, Zoning and FC ports)
  - Configuring vLAGs
- Refer to the Products section at www.brocade.com ror more product information on Brocade VDX data center switches.
- Refer to the technical documents located at www.brocade.com and internet drafts at www.ietf.org for more information on TRILL (Transparent Interconnect of Lots of Links).

VLAN

A Virtual LAN (VLAN) allows stations to communicate as if attached to the same physical medium regardless of their physical locations. Requires use of QTag in frame layout. Using a VLAN architecture provides many benefits, including increased performance, improved manageability, network tuning and simplification of software configurations, physical topology independence, and increased security options.

This section provides some basic VLAN information and details on 802.1Q (VLAN Tagging) and how the protocol works, including:

- "Description" on page 133
- "History" on page 134
- "802.1Q — VLAN tagging" on page 135

Description

A VLAN is a group of hosts with a common set of requirements that communicate as if they were attached to the same subnet (Broadcast domain) regardless of their physical location. A VLAN has the same attributes as a physical LAN, but it allows for end stations to be grouped together even if they are not located on the same network switch. Network reconfiguration can be done through software instead of physically relocating devices.

VLANs are created to provide the segmentation originally provided by routers in a traditional LAN configuration. Routers in a VLAN topology provide broadcast filtering, address summarization, and traffic flow management. By definition, switches may not bridge IP traffic between VLANs as it would violate the integrity of the VLAN broadcast domain.

VLANs operate at Layer 2 only and work with the Layer 3 router (default gateway) to provide access across VLANs or different IP networks. The router interface provides the segmentation between subnets. Normally, there is a one-to-one relationship between a VLAN operating at Layer 2 and a router interface (subnet) operating at Layer 3. Although there are exceptions to the relationship between VLANs and router interfaces, that discussion is out of the bounds of this document.

History

This section offers a brief history of VLAN.

Pre-VLAN

Before switches were being deployed, VLANs did not exist. A typical network or Local Area Network (LAN) consisted of a router that connected directly to pieces of hardware known as *bridges* or *repeaters*. In this configuration each network or sub-network (subnet) was physically separated from each other. Figure 58 illustrates this network.

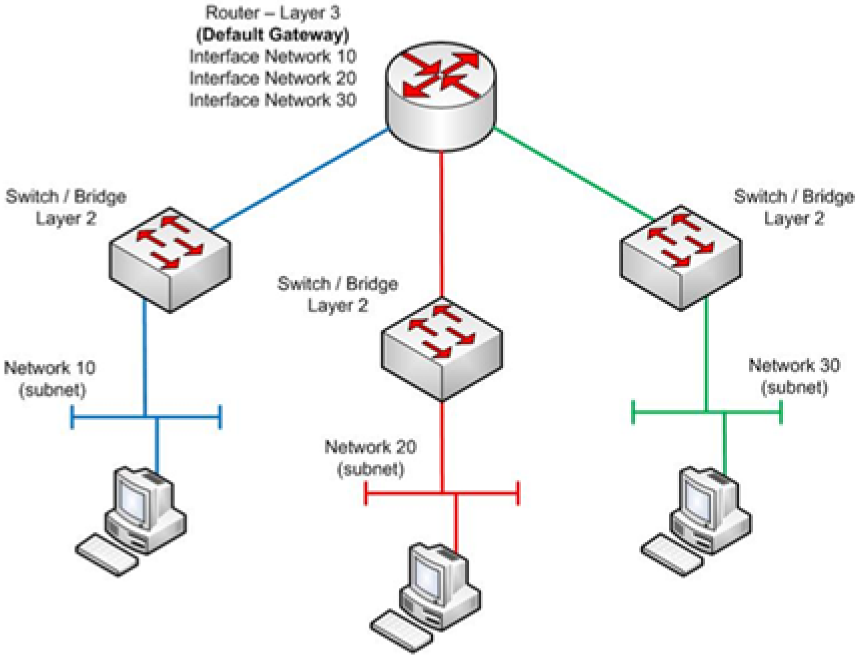


Figure 58 **Pre-VLAN network example**

Untagged

When switches started to become more prevalent, it created a need for virtual networks. At this point there was still a physical separation between the router interfaces and the different subnets being services by the router. The VLAN was created on the Layer 2 switch and each VLAN was connected back to the router independently.

The switch now had VLAN assignments and isolated the different subnets from one another, but needed a physical connection to the router for each VLAN created. Figure 59 is a representation of a VLAN creation.

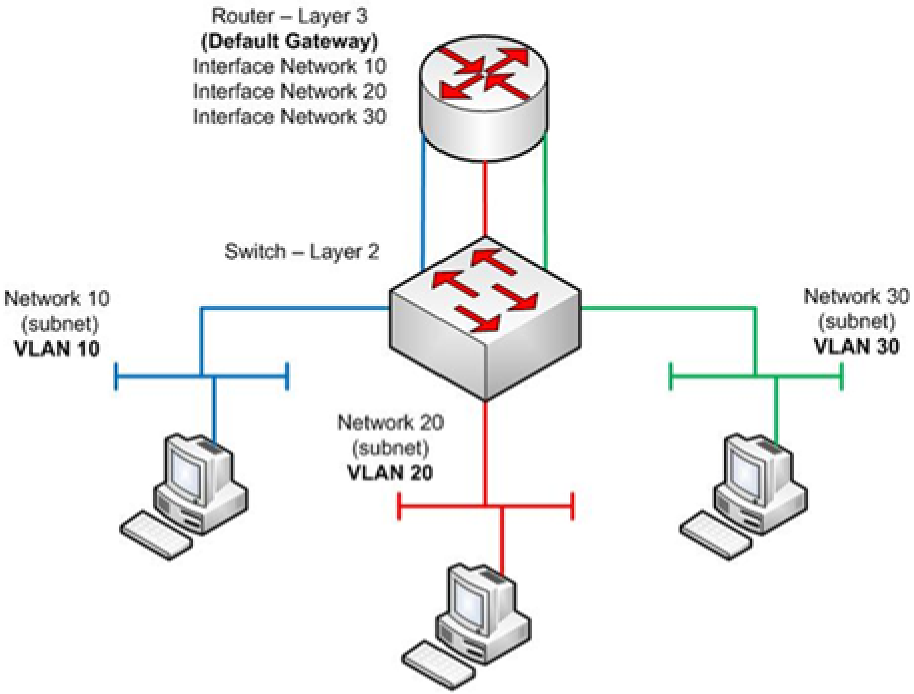


Figure 59 **Creating a VLAN**

802.1Q — VLAN tagging

This new model of a LAN uses VLAN tagged frames known as *trunking* (802.1Q) to multiplex multiple VLANs over a single physical connection. This technology allows for a single connection between any two networking devices (routers, switches, or hosts capable of trunking) with multiple VLANs traversing the same physical path. The mechanism used to achieve this is the tagging of the Ethernet frame. Tagging of frames can be between any devices capable of trunking, including a network interface card (NIC).

IEEE 802.1Q, or *VLAN Tagging*, allows multiple bridged networks to transparently share the same physical network link without leakage of information between networks. 802.1Q is the method of adding and removing a tag to the original Ethernet frame with VLAN-specific information. The IEEE committee defined this method of multiplexing VLANs in an effort to provide multi-vendor VLAN support. This standard defines all of the components used to create and transport VLAN traffic from any end station to another. 802.1Q is an addition to the original 802 standard written by IEEE and encompasses all Layer 2 connectivity.

VLAN Tagging is accomplished by configuring a switch port to send and receive tagged frames. A VLAN tag is nothing more than the VLAN ID (VID) assigned to a given switch port. If there is no VLAN ID assigned to a switch port then most vendors will default to VLAN 1, which is normally considered to be a management VLAN and not used for general purposes. The VID associated to a switch port is also known as the Port VLAN ID (PVID).

Switch ports can be configured as either an "access" port or a "trunk" port:

- Access port

  A switch port that belongs to a single VLAN. The VID assigned to the switch port is added to the incoming frames. The VID will be carried within the frame until it reaches its destination switch port, at which time it will be removed and forwarded to the original destination address.

- Trunk port

  Switch ports that are configured to carry traffic belonging to multiple VLANs between two devices over the same physical link.

802.1Q uses an internal tagging mechanism which inserts a 4-byte tag field in the original Ethernet frame between the Source Address and Type/Length fields. Because the frame is altered, the trunking device recomputes the FCS on the modified frame. This process is automatically performed by the switch right before it sends the frame down a trunk link. At the receiving end, the tag is removed and the frame is forwarded to the assigned VLAN.

802.1Q does not tag frames of the native VLAN on the trunk interfaces. It tags all other frames that are transmitted and received on the trunk. When configuring an 802.1Q trunk, you must make sure that you configure the same native VLAN on both sides of the trunk.

Figure 60 is an example of a Trunking model.

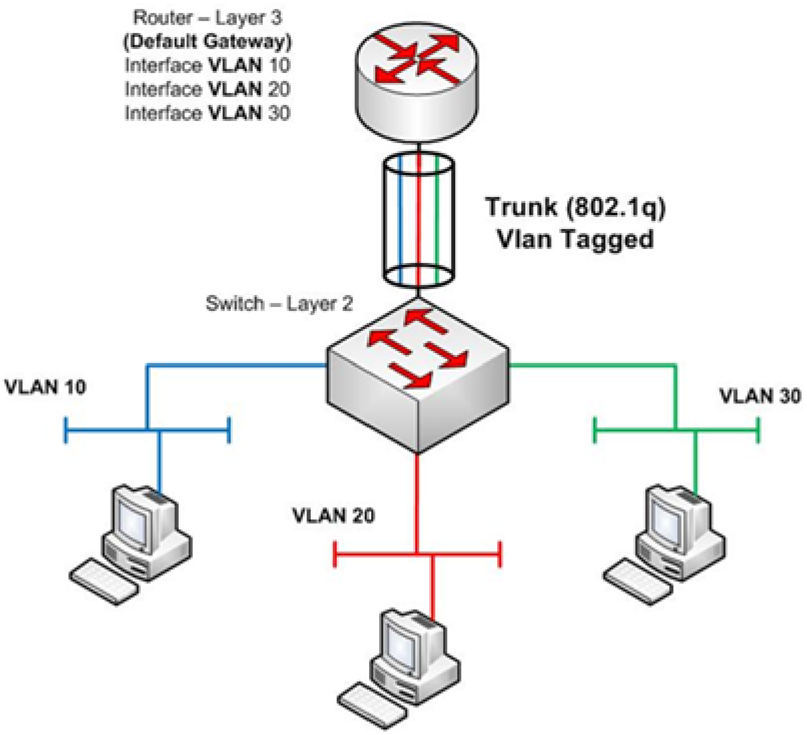


Figure 60 **Trunking (802.1q) example**

Frame format

802.1Q does not actually encapsulate the original frame. Instead, it adds a 32-bit field between the source MAC Address and the Length/Type fields of the original frame.

The VLAN tag field has the following format, as shown in Figure 61:

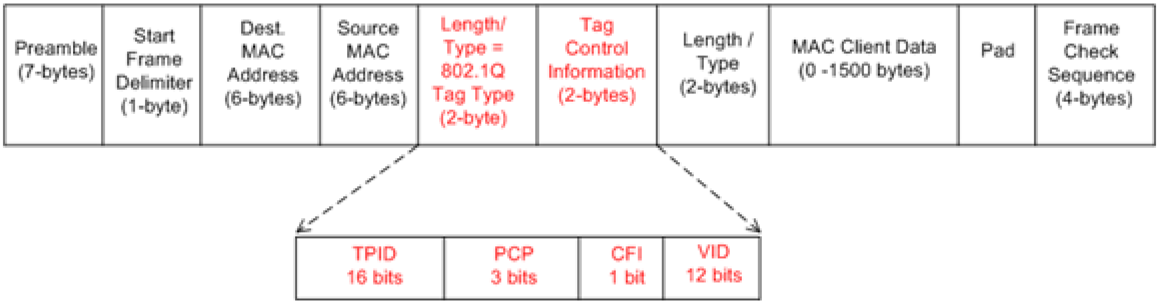

Preamble (7-bytes)
Start Frame Delimiter (1-byte)
Dest. MAC Address (6-bytes)
Source MAC Address (6-bytes)
Length/ Type = 802.1Q Tag Type (2-byte)
Tag Control Information (2-bytes)
Length / Type (2-bytes)
MAC Client Data (0 -1500 bytes)
Pad
Frame Check Sequence (4-bytes)
TPID 16 bits
PCP 3 bits
CFI 1 bit
VID 12 bits

Figure 61 **VLAN tag field**

The following fields are briefly defined:

Tag Protocol Identifier (TPID)

This is a 16-bit field set to a value of 0x8100 in order to identify the frame as an IEEE 802.1Q-tagged frame.

Priority Code Point (PCP)

This is a 3-bit field which refers to the IEEE 802.1p priority (Class of Service) to prioritize different classes of traffic, not necessary for VLAN tagging.

IMPORTANT

This field is used with FCoE and is very important as it is one of the components necessary to allow for lossless behavior.

Canonical Format Indicator (CFI)

This is always set to zero on Ethernet switches. It is used for compatibility between Ethernet and Token Ring networks.

VLAN ID (VID)

This is a 12-bit field specifying the VLAN (1 - 4094) to which the frame belongs. If the value is set to zero (null) the frame does not belong to a VLAN and the 802.1Q tag specifies only a priority, referred to as a *priority tag*. A value of hex FFF is reserved for implementation use. All other values may be used as VLAN identifiers, allowing up to 4094 VLANs. On bridges, VLAN 1 is often reserved for management.

802.1Q process

Figure 62 shows how the 802.1Q process works.

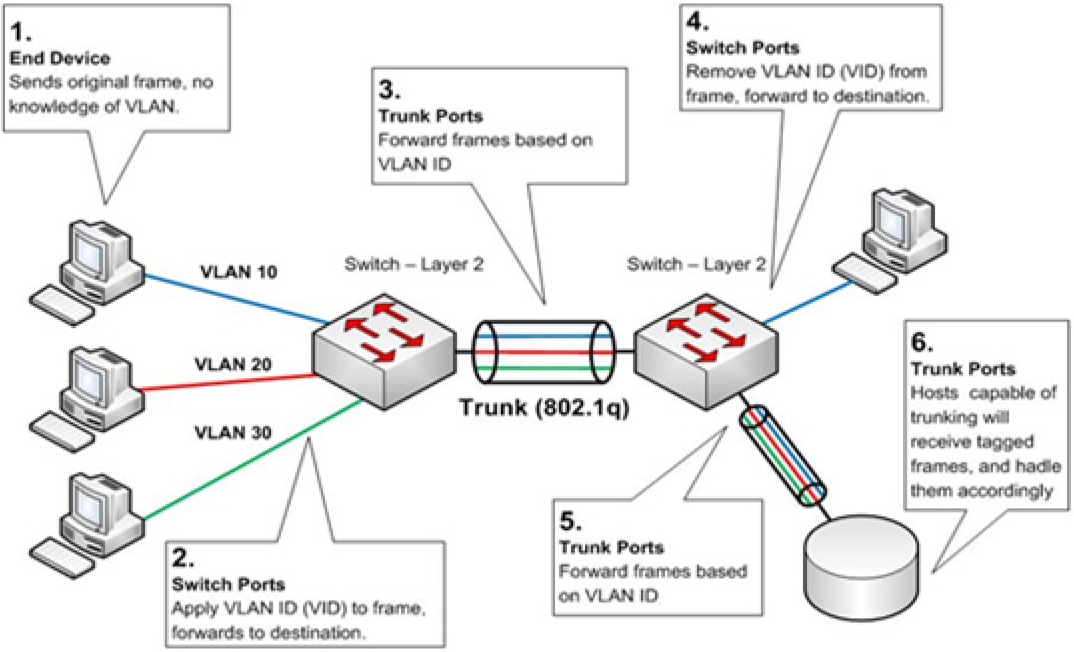

1.
End Device
Sends original frame, no knowledge of VLAN.
2.
Switch Ports
Apply VLAN ID (VID) to frame, forwards to destination.
3.
Trunk Ports
Forward frames based on VLAN ID
4.
Switch Ports
Remove VLAN ID (VID) from frame, forward to destination.
5.
Trunk Ports
Forward frames based on VLAN ID
6.
Trunk Ports
Hosts capable of trunking will receive tagged frames, and hadle them accordingly
VLAN 10
VLAN 20
VLAN 30
Switch – Layer 2
Switch – Layer 2
Trunk (802.1q)

Figure 62 802.1Q process

Step	Description
1	The end station does not have any knowledge of a VLAN. The end station will forward the frame to the switch.
2	The switch will apply the VID to the incoming frame and forward it accordingly. If the destination is within the switch it will forward it directly. If the destination is not local to the switch itself it will forward it to the trunk interface.
3	The switch will apply a 4-byte tag to the frame and recalculate the FCS before forwarding the frame to the other end of the trunk. The receiving switch will verify the FCS and remove the additional tag before forwarding the frame to the correct VLAN.
4	The switch will remove the VID from the frame and forward the frame to the destination.
5	The switch will reapply the 4-byte tag and recalculate the FCS before sending he frame to the destination.
6	The destination device will need to verify the FCS, remove the 4-byte tag and forward the frame to the correct i interface within the device itself.

3

EMC Storage in an FCoE Environment

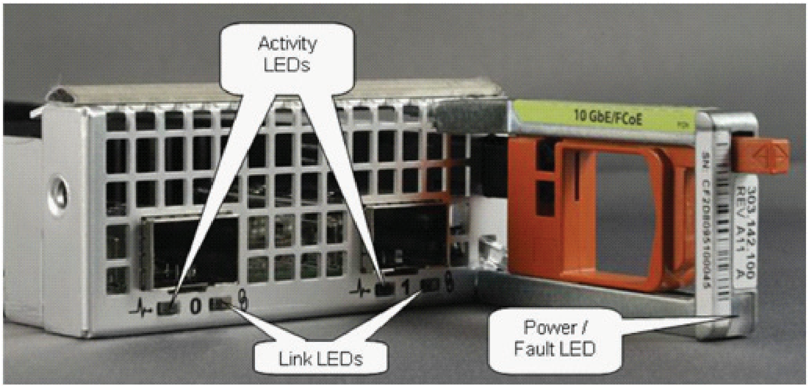
This chapter provides the following information on EMC storage in an FCoE environment:

Note: Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

FCoE connectivity

FCoE connectivity is provided through a dual port 10 Gb/s SLIC (Storage Line Card), referred to as the "UltraFlex I/O module" (Figure 63). The UltraFlex technology allows for storage systems to be easily customized and for I/O slots to be populated with the appropriate I/O modules to meet the need of each environment.

Note: For more information on SLIC, refer to "Prior to installing FCoE I/O module" on page 149.



ICO-IMG-001033

Figure 63 **FCoE UltraFlex module**

FCoE UltraFlex I/O module support two types of physical connectors:

- SFP+ (optical)
- Twinax active

Not all array types support both optical and Twinax connections. Refer to "Cabling support" on page 153 for more details.

Note: Currently, the FCoE I/O module is sold separately from the two physical connectors and requires ordering one type of connector.

EMC storage in an FCoE environment

EMC's FCoE offering adds native FCoE target support with its VMAX and VNX storage platforms. This enables the building of a fully converged, end-to-end, LAN, and SAN infrastructure with the FCoE switch products from Brocade and Cisco.

This section provides information on the following EMC products that support FCoE:

- "Symmetrix VMAX" on page 143
- "VNX series" on page 145
- "CLARiiON CX4" on page 147

VMAX

EMC VMAX systems scale from a single VMAX Engine system with one storage bay to a large eight-engine system and a maximum of ten storage bays.

Online system upgrades are achieved by adding single or multiple VMAX Engines or additional storage bays. Each VMAX Engine contains two VMAX directors with extensive CPU processing power, physical memory, front-end ports, and back-end ports. Drive capacity is increased by installing 4 Gb/s disk array enclosures (DAEs) to the storage bay.

VMAX systems are offered in following three models:

- VMAX

  Scales from one engine pair and 48 drives to a maximum of eight engines and 2,400 drives.

- VMAX SE

  Single engine array and scales from 48 drives to a maximum of 360 drives.

- VMAXe

  Scales from a single engine with 24 drives to a maximum of four engines and 960 drives.

Table 6 compares each VMAX model.

Table 6 VMAX FCoE connectivity comparison

	Maximum drives	Usable capacity	Maximum integrated directors	Maximum FCoE connectivity
VMAX	2400	2.06 PB	16	64
VMAX SE	360	303 TB	2	8
VMAXe	960	1.3 PB	8	32

IMPORTANT

The FCoE option is only available starting with EMC Enginuity version 5875.

The VMAX Engine is a system bay component that provides physical memory, front-end host connectivity (including SRDF), back-end connectivity, and connection to other VMAX Engines. Each Engine provides 32, 64, or 128 GB of physical memory, multiple host configuration options, and connection to eight disk array enclosures. An FCoE capable VMAX Engine uses the FCoE Ultra Flex IO Module as its Front End IO modules as diagramed in Figure 64:

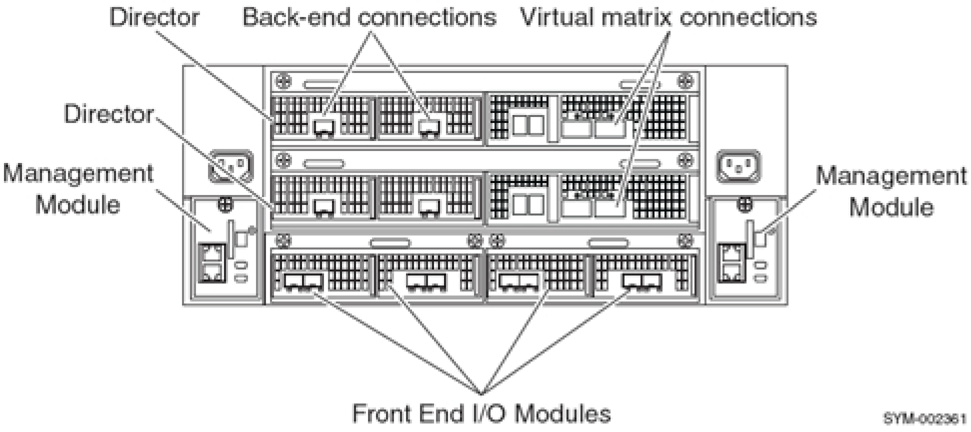


Figure 64 VMAX Engine block diagram

Scalability

Following are scalability numbers for VMAX models.

Note: Mixed I/O module configurations are allowed per engine; however, each type of I/O module must be added in pairs.

Table 7 lists port limit per engine.

Table 7 **VMAX I/O port limit per engine**

Model	Max ports per engine			Max engines per model
	FCoE	FC	10 Gb iSCSI	
VMAX	8	16	8	8
VMAX SE				1
VMAXe				4

Table 8 lists maximum number of initiators allowed per port for each VMAX model.

Table 8 **VMAX Initiator scalability per port**

Model	Max initiators per port		
	FCoE port	FC port	10 Gb iSCSI port
VMAX	32	1024	64
VMAX SE			
VMAXe			

VNX series

The EMC VNX series is designed for medium-size to enterprise storage environments.

VNX arrays have two different enclosure types:

- Onboard FC ports and two I/O module slots
  - VNX 5100, 5300, and 5500 models
- No onboard FC ports and five I/O module slots
  - VNX 5700 and 7500

VNX 5100, 5300, and 5500 models have onboard FC ports and two I/O module slots. The enclosure type for VNX 5100, 5300 and 5500 is shown in Figure 65. This type of enclosure house I/O module slots, storage processors as well as first tray of disks.

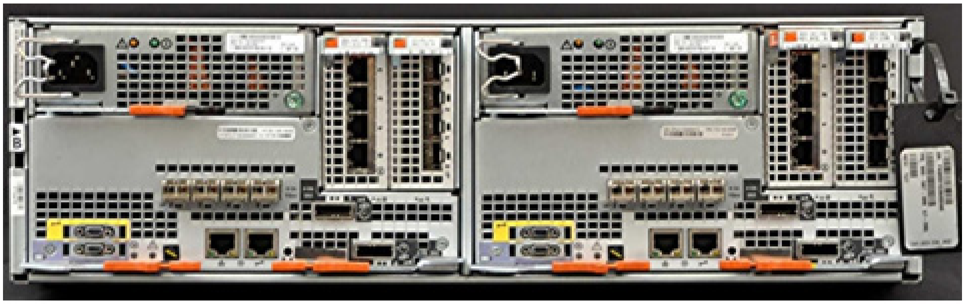
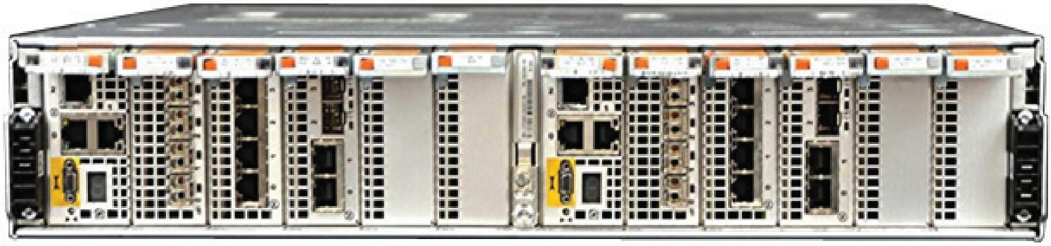


Figure 65 **Sentry model (back end of DPE)**

VNX 5700 and 7500 systems have five I/O module slots and no onboard FC ports. Enclosure types for models VNX 5700 and 7500 are as shown in Figure 66. This type of enclosure house I/O modules slots and service processors.



ICO-IMG-01032

Figure 66 **Argonauts model (back end of SPE)**

Scalability

Following are scalability limits for various VNX models.

Note: The following list highest front-end ports possible with a minimum required 1 SAS I/O module used for back-end connectivity. With more ports used for the back-end, these numbers will change.

Table 9 lists I/O module port limit for each model.

Table 9 VNX I/O port limit per SP

Model Number	Max ports per SP			Max IO modules per SP	Max FE ports per SP
	FCoE	FC	10 Gb iSCSI		
VNX5100	0	4	0	0	4
VNX5300	4	4	4	2	12
VNX5500	4	4	4	2	12
VNX5700	6	12	6	5	16
VNX7500	8	16	6	5	16

Table 10 lists maximum number of initiators allowed per port for each VNX model.

Table 10 VNX Initiator scalability per port

Model Number	Max initiators per			
	FCoE port	FC port	10 Gb iSCSI port	SP
VNX5100	0	256	0	256
VNX5300	512	512	512	1024
VNX5500	512	512	1024	2048
VNX5700	512	512	2048	2048
VNX7500	1024	1024	2048	4096

CLARiiON CX4

The minimum EMC FLARE® version required for supporting an FCoE module is: 04.30.000.5.506.

Table 11 and Table 12 on page 148 list scalability limits for various CLARiiON CX4 models.

Note: The following are the highest possible front-end ports available with a minimum required FC ports for back-end connectivity. With more ports used on the back-end, these numbers will change.

Table 11 lists I/O module port limit per model.

Table 11 **CX4 I/0 port limit per SP**

Model Number	Max ports per SP			Max IO modules per SP
	FCoE	FC	10 Gb iSCSI	
CX4-120	2	6	2	3
CX4-240	2	6	2	4
CX4-480	4	8	4	5
CX4-960	4	12	4	6

Table 12 lists the maximum number of initiators allowed per port for each CX4 model.

Table 12 **CX4 Initiator scalability per port**

Model Number	Max initiators per			
	FCoE port	FC port	10 Gb iSCSI port	SP
CX4-120	512	512	512	512
CX4-240	512	512	1024	1024
CX4-480	512	512	2048	2048
CX4-960	1024	1024	2048	8196

Prior to installing FCoE I/O module

This section provides information to note before installing an FCoE I/O module in the following storage systems:

- "VMAX" on page 149
- "VNX and CX4" on page 149

VMAX

It is important to note the following before installing modules:

- I/O modules must be added in pairs and in symmetric slot location
- FCoE modules are hot-pluggable

VNX and CX4

It is important to note the following before installing modules:

- I/O modules must be added in pairs and in symmetric slot location
- Installation process reboots one SP at a time. Therefore, make sure each host has connections through both service processors for redundancy
- FCoE modules are hot-pluggable
- To replace an I/O module of different type with an FCoE module, all the configuration data on that array must be erased before replacement (not recommended)
- An FCoE module can be added to an empty slot without erasing array configuration data
- Failed FCoE modules can be upgraded, but removing module (any type) causes the SP to reboot

Supported topologies for FCoE storage connectivity

Figure 67 illustrates various connections allowed from a FCoE capable storage array. The following are supported connection types:

- FCoE target ports can be used as targets only; hence replication technologies cannot be used on these. However FC ports on the same array can be used for replication.
- FCoE target ports can communicate concurrently with CNAs and HBAs as long as they go through proper connections
- LUNs can be shared between hosts with CNAs, HBAs and with other arrays of same type with replication technology.

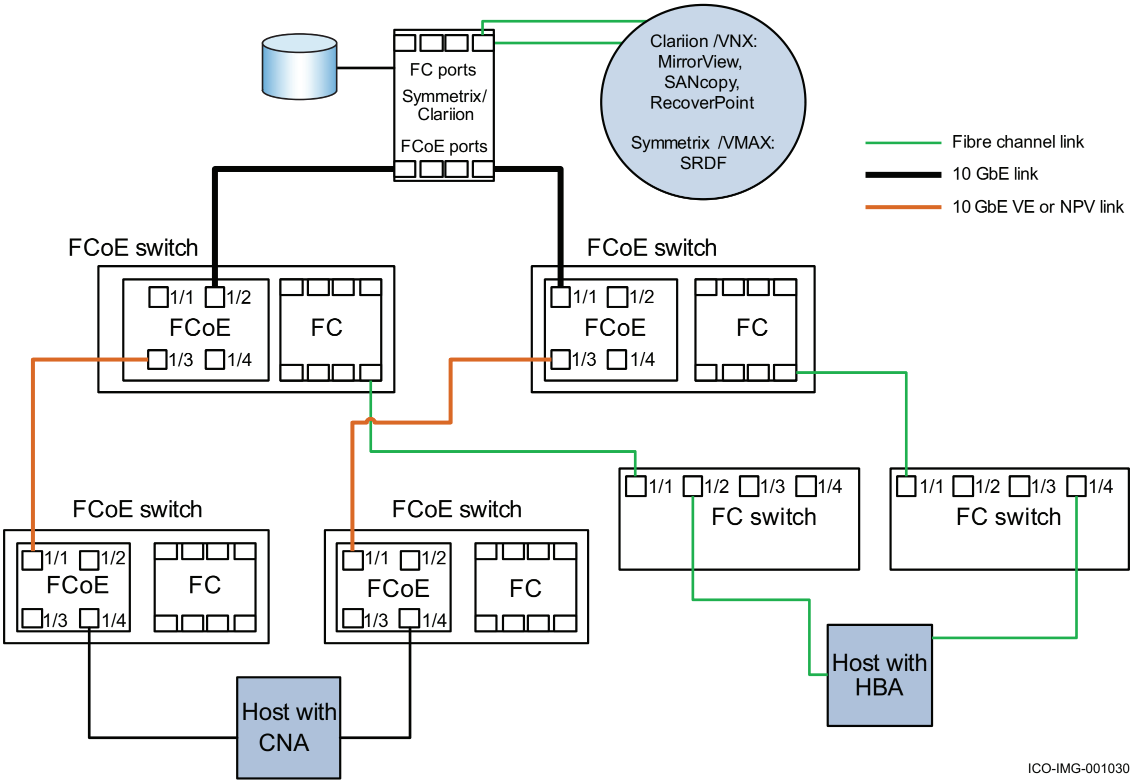


Figure 67 **Supported connection/communication types**

Note: VE port and FCoE NPV features that allow cascading FCoE switches for FCoE hop are not yet available from all switch vendors.

For more information, refer to "FCoE storage connectivity best practices and limitations" on page 152.

For Nexus and EMC Connectrix B Series switches and setup examples, refer to the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

FCoE storage connectivity best practices and limitations

This section contains best practices and limitations.

Best practices

Dual or multiple paths between the hosts and the storage system are required. This includes redundant HBAs, a robust implementation, strictly following management policies and procedures, and dual attachment to storage systems.

Path management software such as PowerPath and dynamic multipathing software on hosts (to enable failover to alternate paths and load balancing) are recommended.

Common

Common guidelines include:

- Redundant paths from host to storage
- Use of Multipathing software and use of failover modes
- Dual fabrics
- Single initiator zoning

VMAX-specific

VMAX-specific guidelines include:

- Each host should have connections to LUNs through different directors for redundancy

VNX-specific

VNX-specific guidelines include:

- Connect each SP to each fabric

Limitations

Limitations include:

- FCoE implementation is currently 'Target only', hence replication technologies (e.g., SRDF, RecoverPoint and Mirror View) are not supported, but can be used on the same array with FC ports.
- CNAs cannot be directly connected to FCoE storage ports.

FCoE storage connectivity requirements and support

This section contains requirements for switches and cabling.

Supported switches

Table 13 lists the minimum firmware version required for supported switches.

Table 13 Minimum required firmware versions

FCoE switches	Minimum firmware version required for direct attach storage
Cisco Nexus 5010, 5020, 5548 and 5548, 7000	4.2(1)N2(1)
Cisco FEX	N/A
Brocade 8000	6.3.1a
Brocade DCX	6.4.1_fcoe1
Cisco UCS 6120 and 6140	1.4(1m)
Cisco MDS 9500	NX-OS 5.2.1

Note: Directly connecting storage to Cisco UCS switches requires the 6120/6140 to be in 'FC switch mode' and connected to an uplink FC switch for zoning. Refer to the “Cisco UCS supported features and topology” and “Cisco UCS Fibre Channel Switch Mode configuration example” sections in the "Blade Server Solutions Setup Examples" chapter of the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Cabling support

UltraFlex FCoE I/O module supports two types of physical connectors:

- SFP+ (optical)
- Twinax active

Refer '“Physical connectivity options for FCoE” on page 58 for more information.

IMPORTANT

VMAX does not support Twinax cables.

4

Solutions in an FCoE Environment

This chapter provides information on the following solutions in an FCoE environment.

Note: Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

EMC RecoverPoint with Fibre Channel over Ethernet

FCoE is an open standards-based protocol that encapsulates Fibre Channel in Ethernet frames, eliminating the need for separate switches, cabling, adapters, and transceivers for each class of traffic. This decreases power consumption and reduces both capital (CAPEX) and operating expenses (OPEX) for businesses. EMC RecoverPoint may be implemented in an FCoE environment without any impact to the functionality and performance of RecoverPoint.

The EMC Connectrix Nexus 5000 or MP-8000B series FCoE switches can be integrated into a RecoverPoint environment. VNX series, CLARiiON, host, and switch-based splitters are supported. The VNX series or CLARiiON splitter requires initiator mode support. This is supported on the FC I/O module. It is *not* currently supported on the FCoE I/O module. All FCoE-connected servers can leverage the RecoverPoint services of the FC I/O module.

This section briefly discusses the following:

- "RecoverPoint replication in an FCoE environment" on page 156
- "Continuous remote replication using a VNX series or CLARiiON splitter" on page 157
- "Continuous data protection using a host-based splitter" on page 158
- "Concurrent local and remote data protection using an intelligent fabric-based splitter" on page 159

RecoverPoint replication in an FCoE environment

For information on local and remote replication, refer to the "Local and remote replication" section in the EMC RecoverPoint chapter in the *Storage Virtualization and Replication Technologies TechBook*, available through the E-Lab Interoperability Navigator, **Documents> Topology Resource Center**, at http://elabnavigator.EMC.com.

FCoE works seamlessly with the three major phases of replication: write (splitting), transfer, and distribution. Each of these phases are further described in the *EMC RecoverPoint Administrator's Guide*, located EMC Online Support website at https://support.emc.com.

Continuous remote replication using a VNX series or CLARiiON splitter

The VNX series or CLARiiON splitter, as shown in Figure 68, runs in each storage processor of a VNX series system or CLARiiON CX3 or CX4 arrays and splits ("mirrors") all writes to a volume, sending one copy to the original target and the other copy to the RecoverPoint appliance (RPA). The RPA is RecoverPoint’s intelligent data protection appliance and manages all aspects of reliable data replication at all sites. This example uses a CLARiiON.

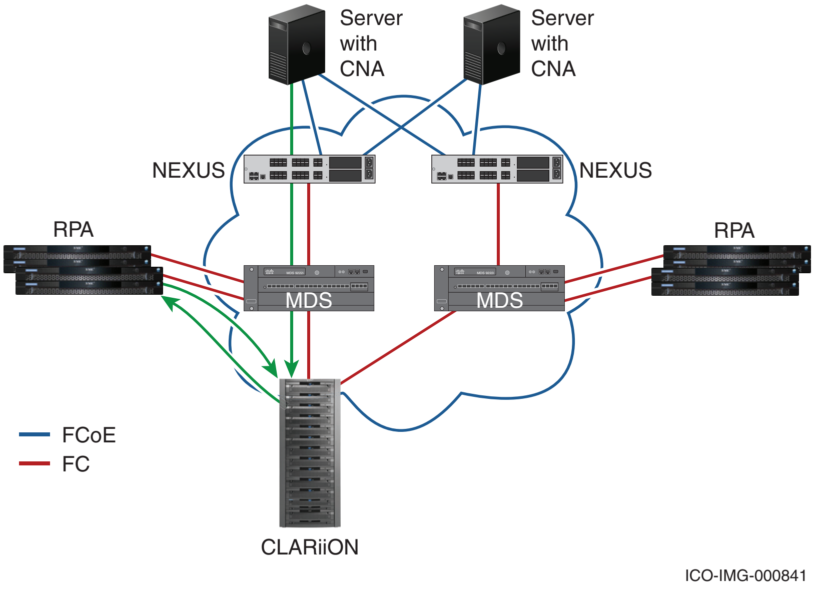


Figure 68 **FC-attached CLARiiON splitter**

During replication, the data originates as a write command from the host. This write is encapsulated as an FCoE frame and received by the FCoE switch. At the FCoE switch, the write command is de-encapsulated and then forwarded to the CLARiiON as an FC write command.

The CLARiiON splitter intercepts the write and sends it to the RPA. Upon receiving the data, it is written to the source replication volume on both the CLARiiON and the RPA.

The RPA sends an ACK to the splitter residing on the CLARiiON. The splitter then sends an ACK to the FCoE-connected host that the write has been successful.

The data is then transferred by the RPA to its peer at the remote location over WAN or Fibre Channel networks.

RecoverPoint distributes the image to the appropriate location on the remote-side storage.

Continuous data protection using a host-based splitter

The host-based splitter (kdiver) is proprietary software installed on hosts that access the volumes to be replicated. The primary function of a kdriver is to split application writes so that they are not only sent to their normally designated storage volumes, but also to the RPA. The RecoverPoint proprietary host-based utility "kutils" is automatically installed when the kdriver is installed on the host. Figure 69 shows an example of a host-based splitter on an FCoE-attached server.

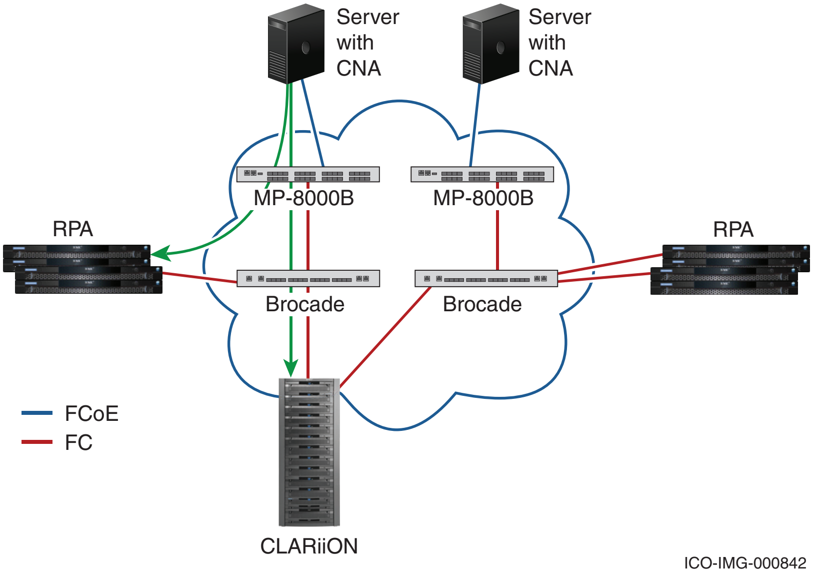


Figure 69 Host-based splitter

RecoverPoint can be used to perform replication within the same local site using continuous data protection (CDP) technology. For a CDP metavolume, the data is continuously written to the journal and to the replica image.

A write command is originated at the FCoE-connected server. The write is intercepted by the kdriver and split. The write is sent to both the source replication volume and the RPA. Both write commands are encapsulated in an FCoE frame, transmitted by the server, received by the FCoE switch, and then de-encapsulated.

The write command meant for the source replication volume is forwarded through the FC SAN to the storage device while the write meant for the RPA is forwarded to the RPA, also through the FC SAN.

Upon receiving the data, if synchronous replication is enabled, the RPA will write data to both the source and target journals before returning the ACK. In asynchronous mode, the RPA will write the data to the source journal but will not wait for the remote journal to be updated before returning the ACK.

Concurrent local and remote data protection using an intelligent fabric-based splitter

RecoverPoint can be used to perform both local and remote replication using CDP and CRR for the same set of production volumes. This type of replication is called concurrent local and remote replication (CLR), as shown in Figure 70 on page 160, in an FCoE environment with Cisco and Brocade switches.

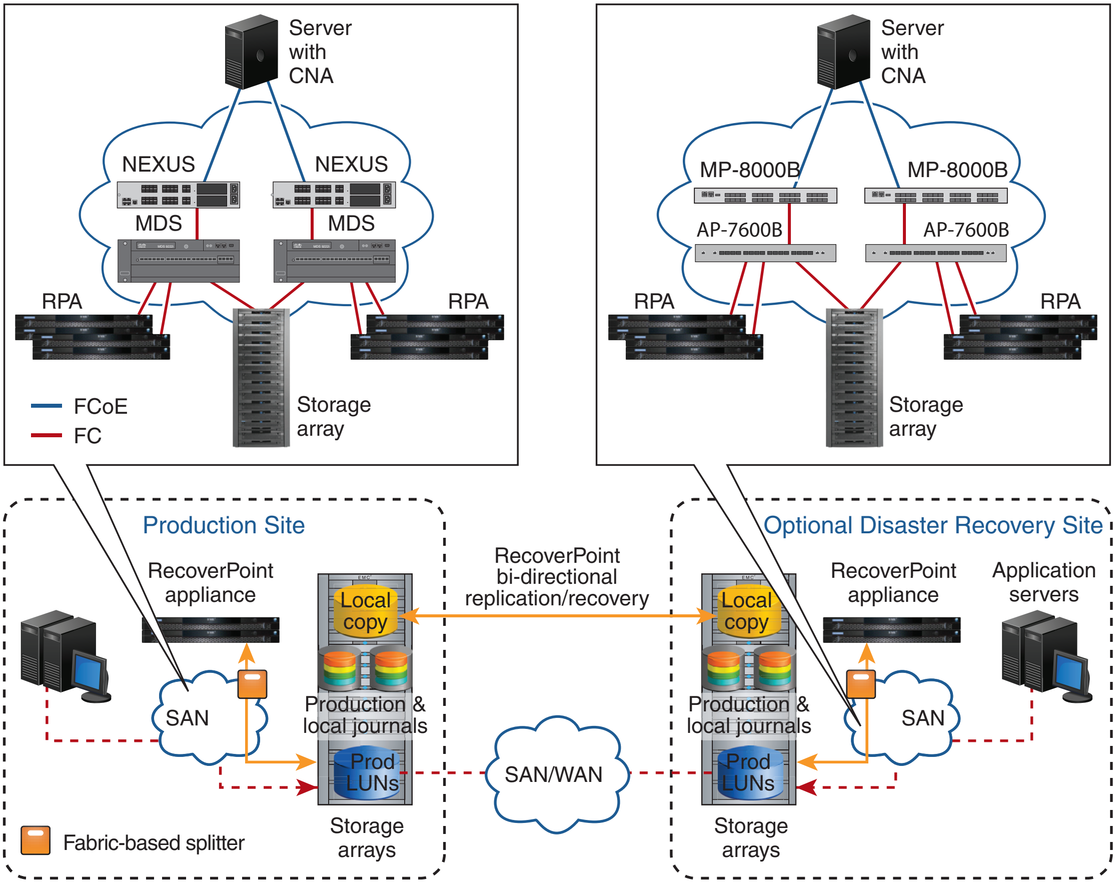


Figure 70 RecoverPoint local and remote replication example

The FCoE-connected server issues a write command encapsulated in an FCoE frame. The write is then de-encapsulated by the FCoE switch and forwarded to the intelligent FC switch running either the SANTap or the SAS service.

The write is split at the intelligent FC switch and sent to both the local replication source and the local RPA. At this point, two simultaneous data streams are created:

- CRR stream
- CDP stream

Each stream is independent of the other.

For local and remote replication, there will be three journals per consistency group: one at the remote site and two at the local site.

If the local replication is paused, it does not affect the remote replication stream, which continues, and vice versa.

FCoE connected servers can take full advantage of RecoverPoint's services. The CLARiiON FC splitter, the host-based splitter, or the intelligent switch splitters from Cisco or Brocade are supported in an FCoE infrastructure.

Related documentation

There are a number of documents for RecoverPoint-related information, all which can be found at EMC Online Support website at https://support.emc.com.

- Release notes

  Information in the release notes include:

  - Support parameters
  - Supported configurations
  - New features and functions
  - SAN compatibility
  - Bug fixes
  - Expected behaviors
  - Technical notes
  - Documentation issues
  - Upgrade information
- *EMC RecoverPoint Administrator's Guide*
- *Storage Virtualization and Replication Technologies TechBook*, available through the E-Lab Interoperability Navigator, **Documents> Topology Resource Center**, at http://elabnavigator.EMC.com
- RecoverPoint maintenance and administration documents
- RecoverPoint installation and configuration documents
- RecoverPoint-related white papers
- *EMC Support Matrix*, located at http://elabnavigator.emc.com

EMC Celerra Multi-Path File System in an FCoE environment

EMC Celerra Multi-Path File System (MPFS) delivers high performance by leveraging the throughput strengths of Fibre Channel with the management strengths of NFS. FCoE switch products integrate Fibre Channel switching with 10 Gb Ethernet technology into a single switch chassis. Combining MPFS with FCoE delivers the throughput advantages of MPFS while leveraging the equipment efficiencies of FCoE.

This section contains the following information:

- "Introduction" on page 162
- "EMC Celerra Multi-Path File System (MPFS)" on page 164
- "Setting up MPFS in an FCoE environment on a Linux host" on page 169
- "MPFS in an FCoE environment using Cisco Nexus switches with redundant path" on page 172
- "Setting up MPFS in an FCoE environment on a Windows 2003 SP2 host" on page 173

Introduction

As the demand for new types of applications increases the number of servers dynamically, the challenge to deliver high-performance file sharing, while at the same time guaranteeing data integrity, makes scalability of storage sharing even more critical. MPFS provides a solution for this type of environment. Figure 71 on page 163 shows an example of EMC Celerra Multi-Path File System (MPFS).

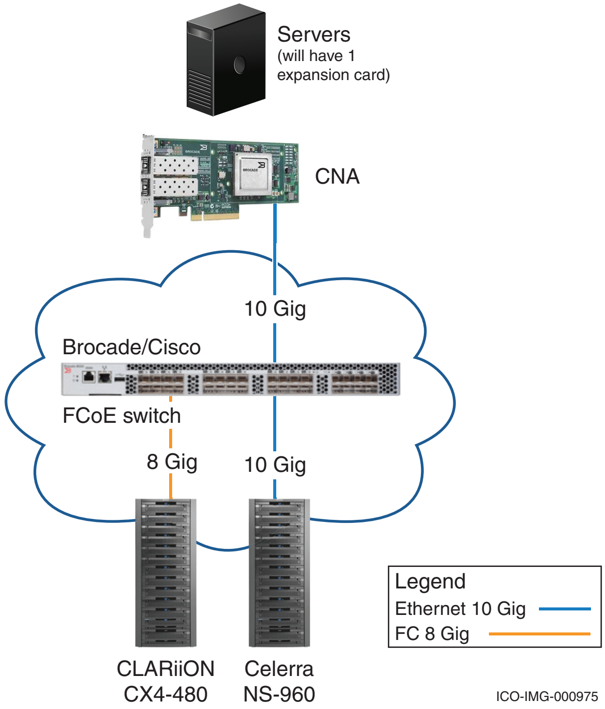


Figure 71 **EMC Celerra Multi-Path File System (MPFS)**

An MPFS solution decreases the complexity of network design by making data uniformly accessible, preventing the need to split/divide the entire dataset, change the applications, or to replicate and distribute the data.

IP connectivity is inexpensive and ubiquitous so many companies choose to leverage the same IP network for both servers and storage traffic by using network-attached storage (NAS).

Network-attached storage (NAS) addresses the IP network and sharing requirements. This approach works well for smaller deployments and some application types. For larger or more I/O-intensive workloads, more NAS servers must be used, making management and load balancing both difficult and costly.

With the upcoming 10 G FCoE network, data centers can be augmented with FCoE-capable infrastructure to support FCoE transportation.

Imagine just using one medium to transport both FC and IP traffic. This will translate to easier maintenance and better control over the network.

EMC Celerra Multi-Path File System (MPFS)

This section provides the following information on EMC Celerra Multi-Path File System (MPFS).

- "Overview" on page 164
- "MPFS advantages over NFS in an FCoE environment" on page 165
- "FCoE MPFS architecture" on page 167

For more details, refer to the Celerra MPFS documentation available at https://support.emc.com.

Overview

EMC Celerra MPFS through a Fibre Channel over Ethernet (FCoE) connection allows the MPFS client to access both metadata and shared data concurrently.

Whereas a traditional MPFS client using EMC Celerra MPFS over FC and IP requires separate paths, namely Fibre Channel to transfer the shared data path and TCP/IP via Ethernet to access the metadata. FCoE uses a common IP LAN topology with FCoE over a 10 gigabit Ethernet Data Center Bridging switch to transport both data and metadata.

Without the MPFS file system, NFS clients can access shared data using standard Network File System (NFS). The MPFS file system accelerates data access by providing separate transports for file data (file content) and metadata (control data).

For a server with an FCoE CNA, data is transferred directly between the Windows server and storage array using Fibre Channel over Ethernet. Metadata passes through the Celerra Network Server and the FCoE network, which includes the NAS portion of the configuration.

In conclusion, MPFS with FCoE offers:

- High-speed transfer of data, compared to NFS
- Converged I/O on server for port count and cabling reduction
- Ease-of-management and less resources

- Converged I/O on switch for port count and cabling reduction.
- Consolidation of IP and FC switches into one FCoE switch

MPFS advantages over NFS in an FCoE environment

FCoE is an open standards-based protocol that encapsulates Fibre Channel over Ethernet, eliminating the need for separate switches, cabling, adapters, and transceivers for each class of traffic. This decreases power consumption and reduces both capital and operating expenses for businesses. EMC MPFS may be implemented in an FCoE environment without any impact to the functionality and performance.

FCoE has enabled the ability to consolidate I/O, translating to lower costs in terms of equipment and power requirements. Implementing MPFS over FCoE utilizes some of the benefits FCoE brings to the new data center.

EMC E-Lab performed tests to compare MPFS performance with NFS and discovered that MPFS with FCoE strongly outperforms NFS. The testing was done with a topology similar to Figure 122. The NFS values were measured in the same topology but FCoE was not used. Table 14 lists the performance differences in an FCoE environment.

Table 14 MPFS versus NFS performance

	Write 4G	Read 4G	Write 5G	Read 5G	Write 7G	Read 7G	Write 9G	Read 9G
MPFS / 1024k	880247	357257	563323	365392	411704	362888	373902	378639
NFS / 1024k	167567	138152	70094	144931	34467	117531	42144	122994
MPFS Advantage	525.31%	258.60%	803.67%	252.11%	1194.49%	308.76%	887.20%	307.85%

Figure 72 on page 166 graphically compares MPFS and NFS performance.

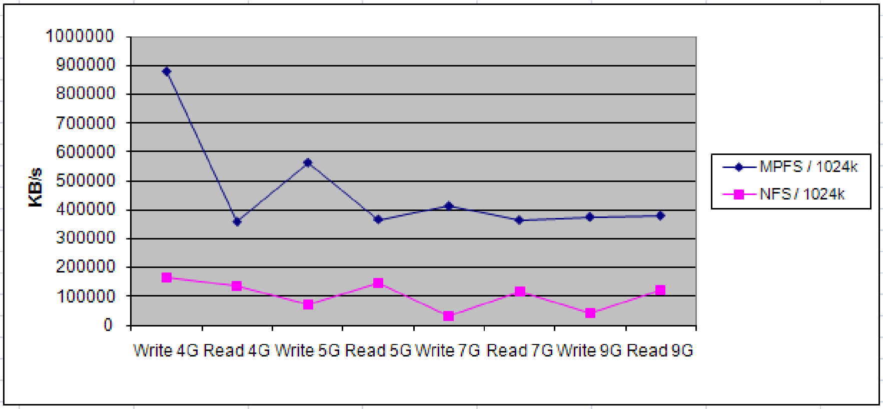


Figure 72 **MPFS versus NFS performance**

Equipment

Tables 15 through 17 show the equipment used in E-Lab's testing of MPFS in an FCoE environment.

Table 15 lists servers used in E-Lab's testing environment:

Table 15 **Servers needed**

Model	Operating system version	Comments / Use
PowerEdge 1950	RHEL 5.4	Source server for MPFS operations 15.9 GB of RAM
Brocade CNA 1020		CNA card

Table 16 lists switches used in E-Lab's testing environment:

Table 16 **Switches needed**

Model (Either one)	Comments / Use
Nexus 5020	FCoE Network
MP-8000B	FCoE Network

Table 17 lists storage used in E-Lab's testing environment:

Table 17 **Storage needed**

Model	Comments / Use
VNX, CX, VMAX, or DMX	Control LUNs for Celerra Gateways and LUN for MPFS

Figure 73 and Figure 74 on page 168 show examples of MPFS environments tested.

FCoE MPFS architecture

Figure 73 shows a typical MPFS architecture using Ethernet and Fibre Channel networks.

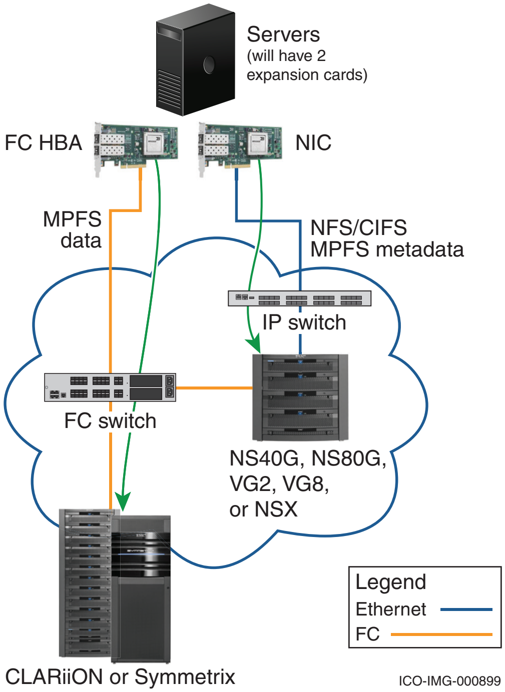


Figure 73 **Typical MPFS architecture example**

Figure 74 shows MPFS data flow in an FCoE environment.

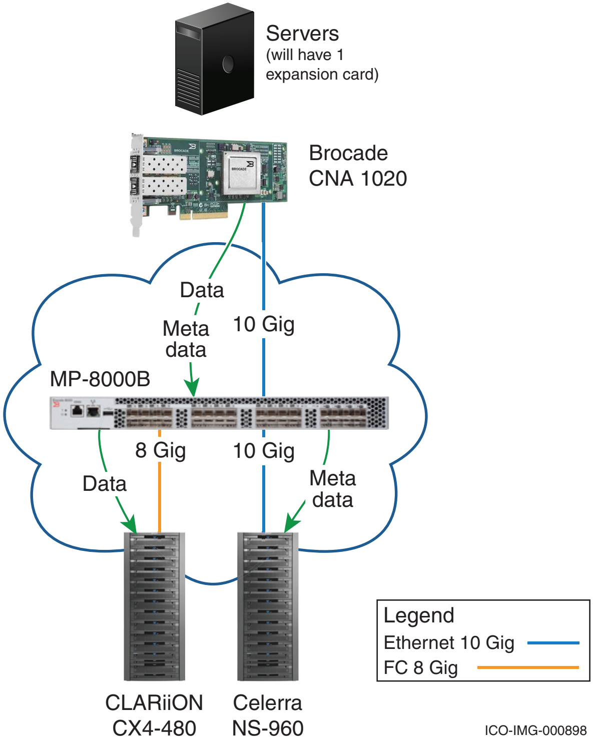


Figure 74 MPFS in an FCoE network example

Comparing Figure 73 on page 167 with Figure 74, observe that less equipment is needed for an FCoE MPFS architecture. A CNA and FCoE switch can support both IP and FC traffic, translating to a simpler management of equipment and lower power requirements.

As shown in Figure 74, the CNA card is generating two types of traffic at the same time, IP and FC. The IP protocol is used for the communication and transfer of metadata and control data, while the FC protocol is used to transport the real large volume data. All these operations are used with only one medium, saving equipment costs and improving performance and ease of management.

The EMC Connectrix Nexus 5020 or MP-8000B series FCoE switches can be used in an MPFS implementation integrated in an FCoE environment.

Setting up MPFS in an FCoE environment on a Linux host

This section contains the following information on installation of MPFS on a Linux host. The RPM to be used is available on Powerlink and the FC version of the MPFS shall be used for the configuration.

- "Enabling MPFS on a Celerra and Linux host" on page 169
- "Mounting MPFS in a FCoE environment" on page 170
- "Configuring zoning and enabling Ethernet interface" on page 170

Enabling MPFS on a Celerra and Linux host

To install MPFS on a Celerra and Linux host, complete the following steps. The RPM is available on Powerlink at **Support** > **Software Downloads and Licensing** > **Downloads C** > **Celerra MPFS Client for Linux or Windows**. The Fibre Channel version of the MPFS is used for this configuration.

1. On the Celerra, enter the following command to enable the MPFS service to run:

   $ server_setup server_2 -Protocol mpfs -option start

2. Install the MPFS Client software either using the RPM package or a CD.

```
#./install-mpfs
Installing ./EMCmpfs-5.0.32.x-i686.rpm on localhost
[ Step 1 ] Checking installed MPFSpackage ...
[ Step 2 ] Installing MPFS package ...
Preparing... ############################################ [100%]
1:EMCmpfs ############################################ [100%]
Loading EMC MPFS Disk Protection [ OK ]
Protecting EMC Celerra disks [ OK ]
Loading EMC MPFS [ OK ]
Starting MPFS daemon [ OK ]
Discover MPFS devices [ OK ]
Starting MPFS perf daemon [ OK ]
[ Done ]
```

For more information, refer to the *EMC Host Connectivity with Brocade Fibre Channel Host Bus Adapters (HBAs) and Fibre Channel over Ethernet Converged Network Adapters (CNAs) in the Linux Environment* document, available on the EMC Online Support website at https://support.emc.com.

Mounting MPFS in a FCoE environment

The command syntax for mounting a MPFS file system is similar to normal Fibre Channel and iSCSI environments:

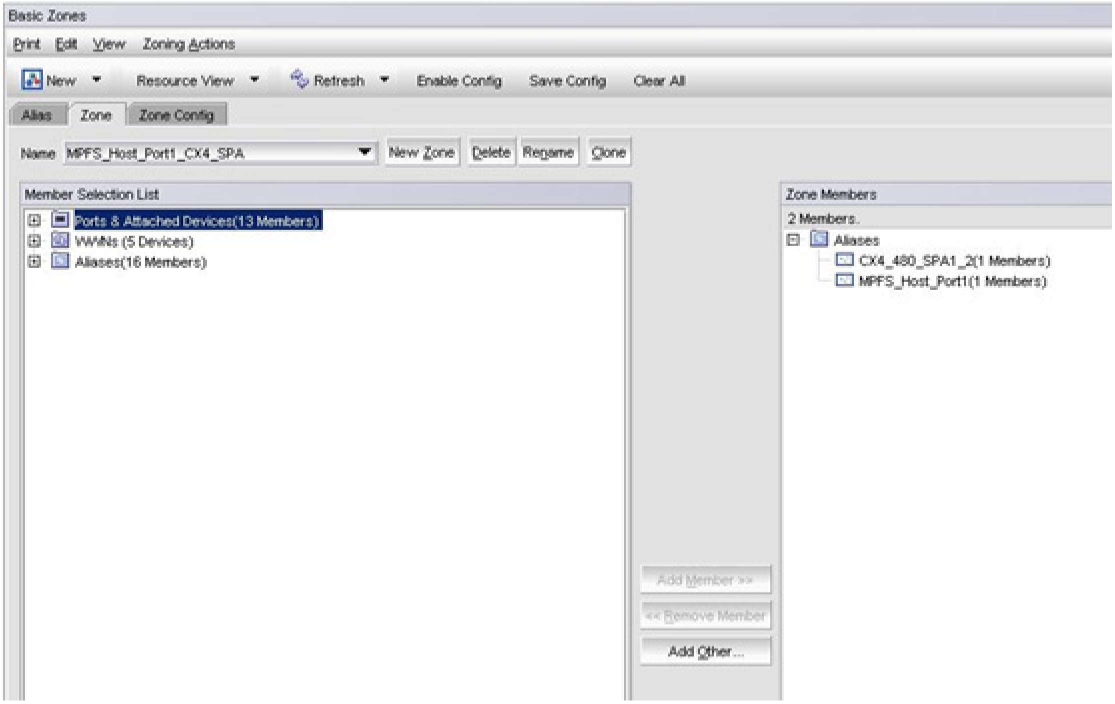
[root@Localhost ~]# mount -t mpfs lp_addr:/mount /mount_point

Working with MPFS in an FCoE environment is no different than working with a normal Fibre Channel network. Implementation of MPFS in an FCoE environment should be an easy and straight-forward task.

Configuring zoning and enabling Ethernet interface

To configure zoning and enable the Ethernet interface, complete the following steps.

1. Create a zone from the server FCoE port that has MPFS installed to the array port, as shown in the following screenshot.



Name MPFS_Host_Port1_CX4_SPA

Member Selection List

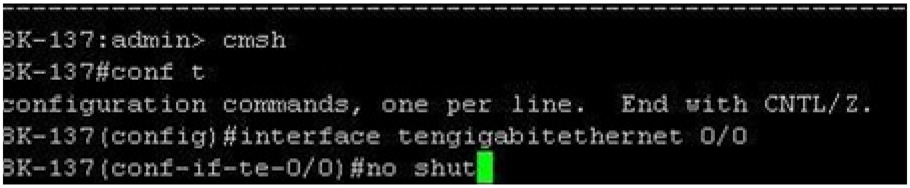
- Ports & Attached Devices(13 Members)
- WWNs (5 Devices)
- Aliases(16 Members)

Zone Members

2 Members.

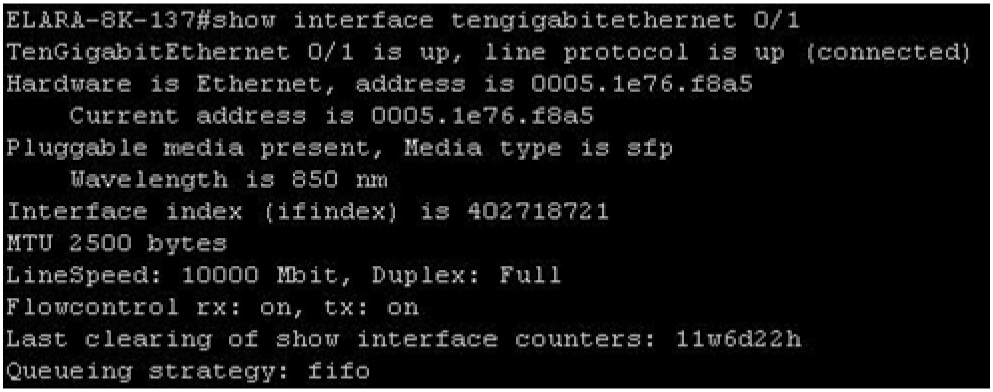
- Aliases
  - CX4_480_SPA1_2(1 Members)
  - MPFS_Host_Port1(1 Members)

2. Using the CLI or GUI, enable the Ethernet interface, connecting the host by issuing the **no shut** command, as shown next.



```
------------------------------------------------------------------
8K-137:admin> cmsh
8K-137#conf t
configuration commands, one per line.  End with CNTL/Z.
8K-137(config)#interface tengigabitethernet 0/0
8K-137(conf-if-te-0/0)#no shut
```

3. Issue the **show interface tengigabitethernet** *x/x* command to check the operations of the FCoE port, as shown next.



```
ELARA-8K-137#show interface tengigabitethernet 0/1
TenGigabitEthernet 0/1 is up, line protocol is up (connected)
Hardware is Ethernet, address is 0005.1e76.f8a5
    Current address is 0005.1e76.f8a5
Pluggable media present, Media type is sfp
    Wavelength is 850 nm
Interface index (ifindex) is 402718721
MTU 2500 bytes
LineSpeed: 10000 Mbit, Duplex: Full
Flowcontrol rx: on, tx: on
Last clearing of show interface counters: 11w6d22h
Queueing strategy: fifo
```

Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**. Refer to the "Nexus Series Switches Setup Examples" and "EMC Connectrix B Setup Examples" chapters in the *Fibre Channel over Ethernet Case Studies TechBook* for detailed instructions on setting up configurations and switch interfaces in an FCoE environment.

MPFS in an FCoE environment using Cisco Nexus switches with redundant path

Figure 74 on page 168 shows MPFS data flow in an FCoE environment.

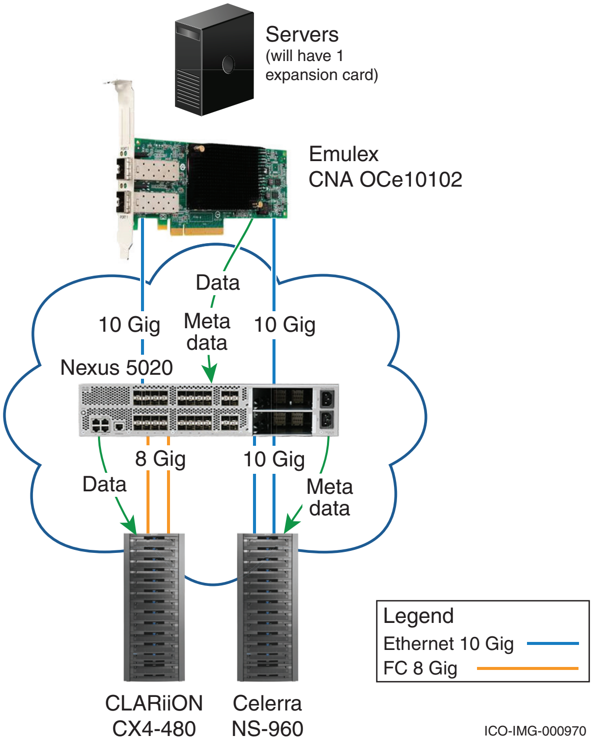


Figure 75 MPFS over FCoE network example

There is less equipment is needed for an FCoE MPFS architecture. A CNA and FCoE switch can support both IP and FC traffic, translating to a simpler management of equipment and lower power requirements.

As shown in Figure 75, the CNA card is generating two types of traffic at the same time, IP and FC. The IP protocol is used for the communication and transfer of metadata and control data, while the FC protocol is used for the transport of large volume of data. All these operations are used with only one medium.

The EMC Connectrix Nexus 5020 or MP-8000B series FCoE switches can be used in an MPFS implementation integrated in an FCoE environment.

Equipment

Tables 18 through 20 show the equipment used in E-Lab's testing of MPFS in an FCoE environment.

Table 18 lists servers used in E-Lab's testing environment:

Table 18 **Servers needed**

Model	Operating system version	Comments / Use
PowerEdge 1950	Windows 2003 SP2	Source server for MPFS operations
Emulex OneConnect OCe10102-F		CNA card

Table 19 lists switches used in E-Lab's testing environment:

Table 19 **Switches needed**

Model (Either one)	Comments / Use
Nexus 5020	FCoE Network

Table 20 lists storage used in E-Lab's testing environment:

Table 20 **Storage needed**

Model	Comments / Use
VNX, C, VMAX, or DMX	Control LUNs for Celerra Gateway and LUN for MPFS

Setting up MPFS in an FCoE environment on a Windows 2003 SP2 host

This section contains the following information on installation of MPFS on a Windows 2003 SP2 host. The executable file to be used is available on Powerlink and the FC version of the MPFS shall be used for the configuration.

- "Enabling MPFS on a Celerra and Windows 2003 SP2 host" on page 174

- "Enabling NIC teaming" on page 174
- "Mapping an MPFS share to a network drive" on page 176

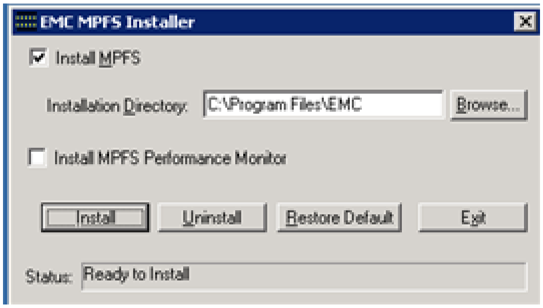
Enabling MPFS on a Celerra and Windows 2003 SP2 host

To install MPFS on a Celerra and Windows 2003 SP2 host, complete the following steps.

1. On the Celerra, enter the following command to enable the MPFS service to run:

   $ server_setup server_2 -Protocol mpfs -option start

   The **EMC MPFS Installer** dialog box displays.

   EMC MPFS Installer

   ☑ Install MPFS

   Installation Directory: C:\Program Files\EMC

   ☐ Install MPFS Performance Monitor

   Status: Ready to Install

2. In the **Installation Directory** field, select the folder you want the installation saved to.
3. Click **Install**.

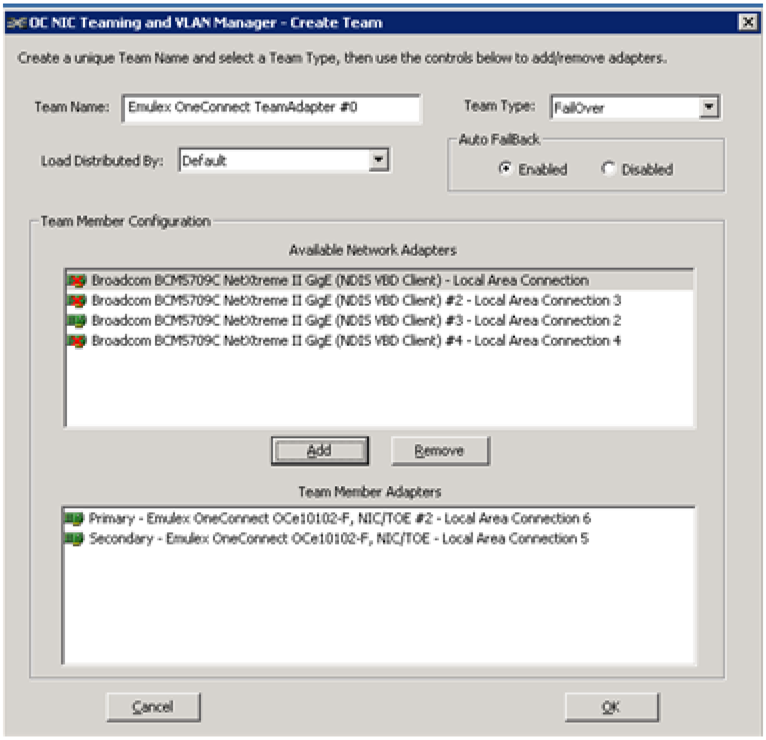
Enabling NIC teaming

Creating NIC teaming allows you to have high availability on the Ethernet network.

To enable NIC teaming for Emulex OneConnect, complete the following steps.

1. In the **EMC NIC Teaming and VLAN Manager - Create Team** window, complete the following fields.
   a. In the **Team Name** field, enter a team name.
   b. In the **Team Type** field, enter the team type.
   c. In the **Load Distributed By** field, select **Default**.
   d. In the **Auto Failback** field, select **Enabled**.
   e. Create a team by selecting the two ports of Emulex OneConnect and click **Add**.

f. Click **OK**.



OC NIC Teaming and VLAN Manager - Create Team

Create a unique Team Name and select a Team Type, then use the controls below to add/remove adapters.

Team Name: Emulex OneConnect TeamAdapter #0

Team Type: FailOver

Load Distributed By: Default

Auto FailBack

● Enabled ○ Disabled

Team Member Configuration

Available Network Adapters

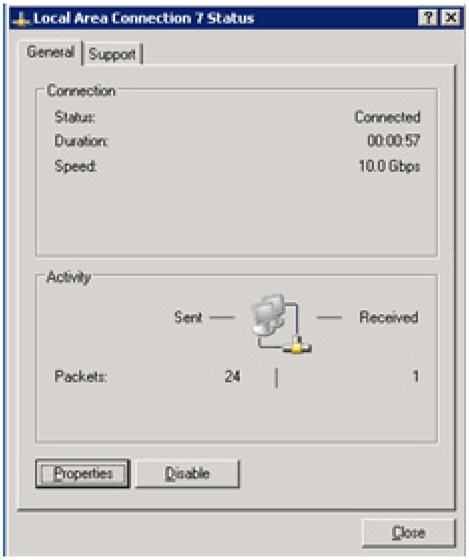
- Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) - Local Area Connection
- Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) #2 - Local Area Connection 3
- Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) #3 - Local Area Connection 2
- Broadcom BCM5709C NetXtreme II GigE (NDIS VBD Client) #4 - Local Area Connection 4

Team Member Adapters

- Primary - Emulex OneConnect OCe10102-F, NIC/TOE #2 - Local Area Connection 6
- Secondary - Emulex OneConnect OCe10102-F, NIC/TOE - Local Area Connection 5

Forming a NIC team will create a new connection under **Network Connections**.

Assign the IP address to the new connection, as shown in the next screen.

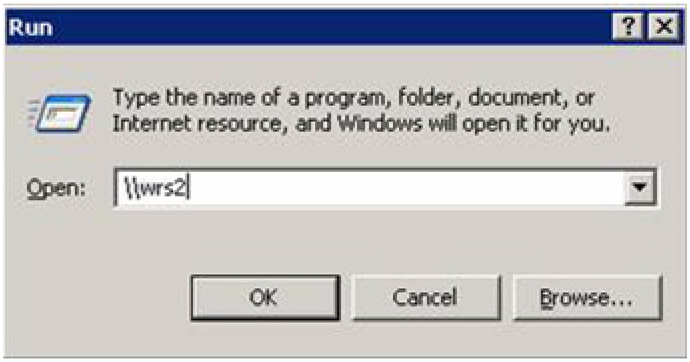


Status: Connected
Duration: 00:00:57
Speed: 10.0 Gbps
Packets: 24 | 1

Mapping an MPFS share to a network drive

To map an MPFS share to a network drive, complete the following steps:

1. Select **Start > Run** to open the **Run** window.

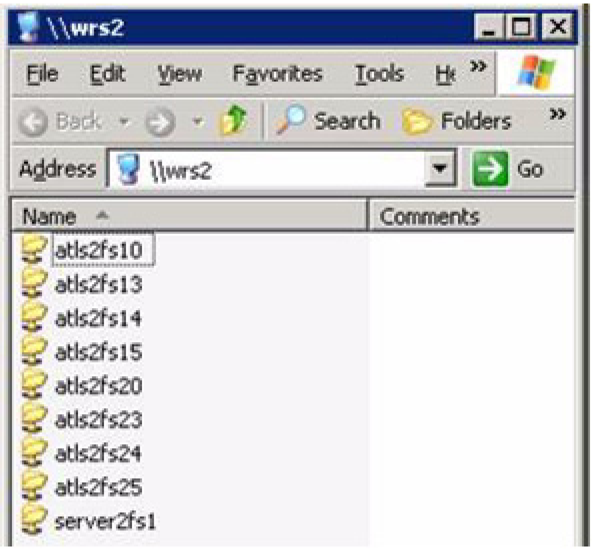


Type the name of a program, folder, document, or Internet resource, and Windows will open it for you.

Open: \\wrs2

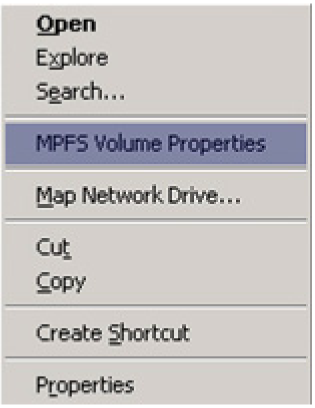
2. Type the Celerra Network Server (Data Mover interface) containing the CIFS shares to be mapped. The above example shows a Celerra Network Server named wrs2.

3. Click **OK** and a **List of Available Shares** window displays, as shown next.

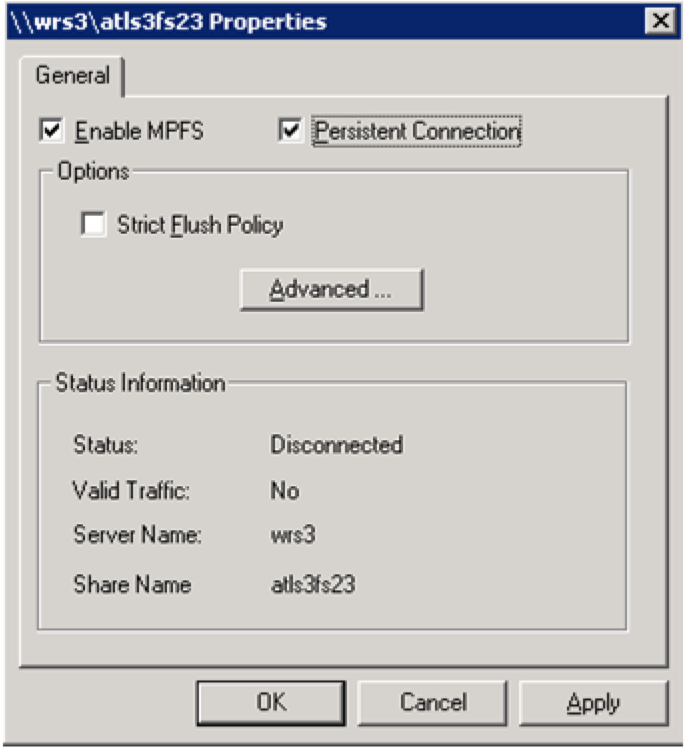


atls2fs10
atls2fs13
atls2fs14
atls2fs15
atls2fs20
atls2fs23
atls2fs24
atls2fs25
server2fs1

4. Right-click the share to be mapped and select **MPFS Volume Properties** from the drop-down menu.



5. In the **MPFS Properties** window, select **Enable MPFS** to enable MPFS on a share.



☑ Enable MPFS ☑ Persistent Connection

☐ Strict Flush Policy

Status: Disconnected

Valid Traffic: No

Server Name: wrs3

Share Name atls3fs23

Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**. Refer to the "Nexus Series Switches Setup Examples" and "EMC Connectrix B Setup Examples" chapters in the *Fibre Channel over Ethernet Case Studies TechBook* for detailed instructions on setting up configurations and switch interfaces in an FCoE environment.

5

Troubleshooting Basic FCoE and CEE Problems and Case Studies

This chapter provides several ways to troubleshoot basic FCoE and CEE problems. The first section provides basic troubleshooting concepts and techniques. The second section provides advanced troubleshooting topics and two case studies using different troubleshooting techniques, such as flowcharts.

Note: Fibre Channel over Ethernet case studies using Connectrix B, Nexus, Brocade, and HP can be found in the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Troubleshooting basic FCoE and CEE problems

There are several ways to troubleshoot basic FCoE and CEE problems. This section provides a few suggestions, including:

- "Process flow" on page 180
- "Documentation" on page 182
- "Creating questions" on page 182
- "Creating worksheets" on page 183
- "Log messages" on page 184
- "OSI layers" on page 186
- "FC layers" on page 187
- "Connectivity problems" on page 188
- "Physical interface status" on page 190
- "Interface errors" on page 193
- "MAC layer" on page 197
- "Understanding FCoE phases" on page 198
- "fcping and fctraceroute commands" on page 204
- "Upper layer protocol" on page 207

Two case studies are then provided in:

- "FCoE and CEE troubleshooting case studies" on page 208
  - "Case Study #1, Unable to access the LUNs/devices" on page 208
  - "Case Study #2, Unable to access a shared folder in the file server" on page 251

Process flow

As data centers grow larger and more complex, there is a greater chance of encountering network issues that can impede the entire infrastructure or degrade performance to unacceptable levels. Therefore, it is important to have a systematic and organized troubleshooting method in place and be able to apply it should the need arise.

A generally accepted troubleshooting model is shown in Figure 76. This model presents a process flow that can effectively guide data center and network support during troubleshooting tasks.

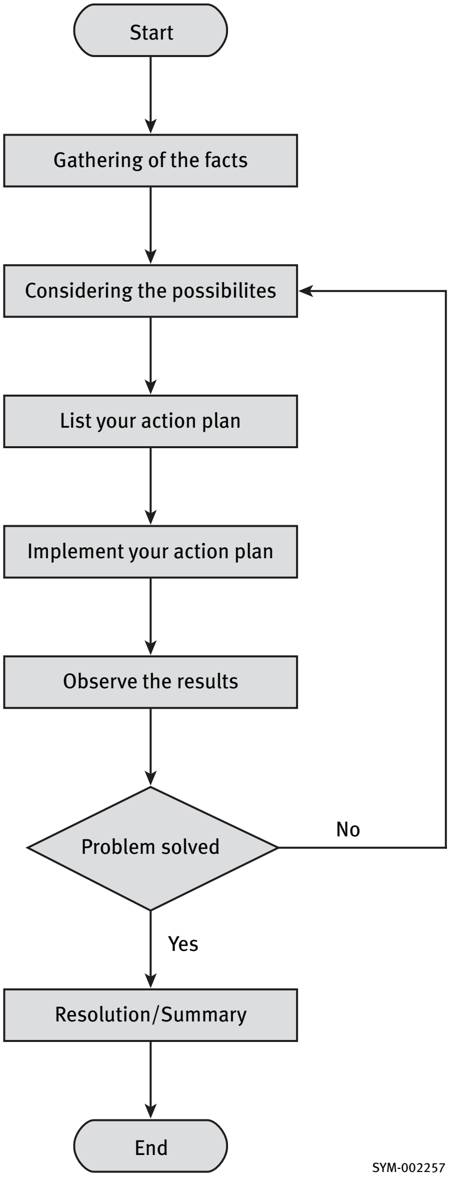


Figure 76 **Troubleshooting process flow**

Documentation

Documentation plays an important role in troubleshooting many data center network issues. Documentation provides the structure or basis to answer fundamental troubleshooting questions, such as:

- "What is the IP range of those devices connected to the switch?"
- "Where is this switch connected to?"
- "What settings did you use in configuring the interface?"
- "What spanning-tree cost did you use on port x?"

Remembering and rebuilding the network topology documentation during outages is a very difficult task, especially if certain Service Level Agreements (SLAs) need to be met. It is not a good practice to determine the network topology during the network downtime.

When using documentation, consider the accuracy of the content. In many cases, even when network and data center administrators are responsible for their own documentation, the documentation is not kept current. When you are trying to solve a complex network issue and the network devices are not accessible, or you are dealing with network-wide outages, unless the documentation is current, it can be useless.

Creating questions

Another way of isolating problems in an Ethernet network is to combine all the relevant facts and then address each suspected problem one at a time. Prepared troubleshooting questions can be used to check common issues and solutions. OSI layer and FC layer troubleshooting methodologies, discussed further in "OSI layers" on page 186 and "FC layers" on page 187 can be used as guides in creating probing questions.

Some examples of probing questions are:

- When did the problem start?
- Were there any (hardware/software/configuration) changes before the symptoms were observed?

Note: Hardware/software changes can be anything from adding a new network module, reseating the cards, rebooting the switch, upgrading the IOS, updating NIC drivers, or updating the host's patch. Anything that has been changed must be noted since network issues do not happen without cause.
Even if there have not been any apparent hardware failures, if everything had been working and now it is not, then something has to have been changed. This change could be planned, unplanned, or even caused by nature, like lightning or Electrostatic Discharge (ESD) caused by cosmic radiation.

- What is the topology/design of the Ethernet network and where are the devices connected to the switch?
- Do you have an accurate physical and logical map of the network?
- Have you identified the list of all the reported symptoms on this network problem?

Creating a list of probing questions can be useful in solving problems.

Creating worksheets

A worksheet can help in the troubleshooting process by using answers to prepared questions, such as those discussed in "Creating questions" on page 182 This worksheet can be customized as a guide or template to troubleshoot specific Ethernet network environment issues. Table 21 is an example of a worksheet that can be used as a customized guide for troubleshooting:

Table 21 Troubleshooting worksheet

Problem	Symptoms	Action Plans	Result (solved or not)	Comments
1)				
2)				
3)				

Data center and networking issues are most often represented by multiple symptoms. For instance, you may think you have a complex spanning tree problem because traffic is flowing in a way you think is wrong, or you may think you have some complicated Fibre Channel

ULP problem. However, the problem might be caused by a switchport interface failure. In other words, you can waste valuable time trying to analyze and fix a spanning tree or FLOGI problem, while the root cause may be just an interface hardware failure.

As you gain experience with your infrastructure, try to solve the issue by identifying the root cause and troubleshooting the lower layers first. Methods like OSI or FC layered troubleshooting and problem solving through questions and answers can be used to simplify the troubleshooting process (refer to "OSI layers" and "FC layers" on page 187). Although complicated failures occur, simple failures are far more common. Once a systematic approach to troubleshooting is applied, the root cause of the problem can be more easily found.

Log messages

One of the best practices in network management is maintaining the logs of significant events that occur on your network equipment. Device logging is automatically enabled on most of the 10 Gb CEE switches.

Device logging can also be configured to transmit logging information to a logging server/ TFTP file server to be used to evaluate at a later time. You might want this information for fault isolation and troubleshooting. More importantly, you can use the timestamps of the logs to gain valuable information. For example, the logs show the last configuration change or whether anyone performed any hardware or software changes on the device.

The number of log messages generated is equal to the configured logging level, as shown in Table 22. Logging level 7, or the debugging level, generates more log messages than a logging level 6, and so on. Logging level 7 is mainly used for troubleshooting, where the device logs all messages that are generated by the feature or hardware in question.

Table 22 Logging levels (page 1 of 2)

Level	Logging message
0	Emergencies
1	Alerts
2	Critical

Table 22 Logging levels (page 2 of 2)

Level	Logging message
3	Errors
4	Warnings
5	Notifications
6	Informational
7	Debugging

Hosts and storage arrays each have their own way of displaying event logs. Most modern network operating systems and CNA management utilities have ways of displaying port information such as status, counters, statistics, events, and errors. For example, storage arrays have management systems embedded in the product and some can be managed from specialized management software, such as EMC Ionix ControlCenter, which is capable of managing multiple devices. The concept is similar with network equipment. The information presented is relevant to the events, errors, and informational data for the events that are occurring on the device, such as port flapping. Timestamps display the time a particular event or alarm took place.

The way to display logs is the same in both the Nexus 5000 and MP-8000B switches as it is in Cisco switches. They use the same format of the **show logging** CLI command. Additionally, MP-8000B has an option to display logs in Fabric OS (FOS) using the **errshow** and **errdump** commands.

There are many network tools that can provide advance features for detecting, diagnosing, and fixing network problems. One tool is sFlow, a monitoring tool for a high-speed switched network. sFlow (RFC 3176) is an industry standard technology, not only used for fixing network problems, but also to provide performance improvement, accounting or billing for usage, and network security. For more information, refer to the sFlow website at http://www.sflow.org.

OSI layers

Fibre Channel over Ethernet, as the name implies, is an Ethernet layer 2 technology. Since it is an Ethernet technology, it can easily be mapped to layer 2 of the OSI reference model. As discussed in "OSI networking protocol" on page 75 the OSI model begins with the physical layer (layer 1) and ends with the application layer (layer 7). OSI layers depend on each other. For example, without layer 1, layer 2 will not function, layer 3 will not function, and so on.

As you develop troubleshooting skills, techniques like troubleshooting layer 1 and layer 2 as a pair will be beneficial and save you time troubleshooting the upper layers. You will discover that most issues occur on the first three layers of the OSI model:

- Layer 3: Network
- Layer 2: Data Link
- Layer 1: Physical

Table 23 lists some of the things to check and verify on each layers of OSI model.

Table 23 Verifying layers (page 1 of 2)

Layer	Verify
Application	Make sure applications are correct and installed properly. Use the correct applications when opening data. Check application compatibility with the Operating Systems, etc.
Presentation	Check encryption (VPN falls here), compression, and formatting. Check if the application is viewing the correct format that is supporting it.
Session	Check ULP errors like SQL, NFS, CIFS, etc. Using Network Analyzer would help to detect these issues. Also, check those logs from applications.
Transport	Check if there is UDP/TCP filtering (via ACLs or firewalls), QoS feature blocking, or rate-limiting particular traffic. Check UDP/TCP performance tuning (maximum number of TCP retransmissions, path MTUs, TCP retransmission timeout, TCP window size optimization, etc.).
Network	Check default gateway, static routes, dynamic routing, routing metrics, path cost, attributes. Also other things that would affect routing like IP addressing, IP ACLs, Route-maps, Filter-list, Distribute-list, Firewall policies, etc.

Table 23 Verifying layers (page 2 of 2)

Layer	Verify
Data Link	Check if the switch is learning the MAC addresses of the hosts connected. Ensure there's no security policy that is blocking specific MAC addresses or specific traffic types, i.e., FCoE and FIP frames. Check also if VLAN membership, native VLAN, and VLAN trunking encapsulation are configured properly. Also, check the keepalive timer or if keepalive is enabled. Make sure there is no configuration mismatch on both side of the links, i.e., speed, interface mode mismatch (access, trunk, converged) and trunk encapsulation type mismatch. Duplex mismatch is no longer an issue since CEE supports full duplex only. Check the spanning tree configurations as well.
Physical	Check the status of the physical cable, media/port connectors, and interface cards. Also, make sure cables are connected to the correct ports. Check the length and type of cable used. Use cable testers like copper/optical pulse tester, optical loss tester, copper/fiber certification tester. Make sure there is no unidirectional link (usually caused by undetected fiber or transceiver problem). This can be avoided if UDLD feature is turned ON on the switch.

FC layers

Since FCoE runs over Ethernet, the root cause for most physical and connectivity issues will be related to OSI layers 1 and 2. However, when troubleshooting native FC connectivity issues like FC addressing, classes of service, flow control, and frame-level errors, Fibre Channel layers FC-2 and FC-1 will be used. Problems with error detection and recovery using TimeOut Values (TOV) are included in this lower layer of FC, particularly FC-2. Upper layer protocol (ULP) and fabric services issues including FLOGI, PLOGI, PRLI, SCSI-FCP, are native to Fibre Channel. Therefore, in this scenario, you can use the five different layers of the Fibre Channel (FC) as a reference model for troubleshooting.

When dealing with connectivity issues and using the FC layers as a troubleshooting guide, try using the following information listed in Table 24. The order does not matter. The approach usually depends on the issue. For example, if the issue is that a host will not log in to a storage array, you would first check the zoning. However, if the issue is that a storage array is logging *host aborts* from a host, you would not look at zoning since the devices are talking to each other.

Table 24 Verifying checkpoints

Checkpoint	Verify
Port Configuration	Check to make sure the port is configured correctly, i.e., port state, port type and VSAN membership.
Port Error logs	Check to see if the host, storage ports (as well as ISL port, if involved in the path) are logging any errors.
Embedded Port logs	Check what is going on with the host and storage during the login process.
Zoning	Check the zoning to make sure host is zoned to storage array.
Name Server	Check to see if host and storage are logged and registered with the name server

Connectivity problems

To troubleshoot FCoE connectivity problems, you will need FC-2, FC-1, and FC-0 layers of the FC reference model, together with the physical layer 1 and data link layer 2 of the OSI model. As shown in Figure 77 on page 189, FCoE operates directly above Ethernet in the network protocol stack. Therefore, to troubleshoot FCoE, it is paramount to ensure there is no Ethernet layer 1 and layer 2 issues on the network.

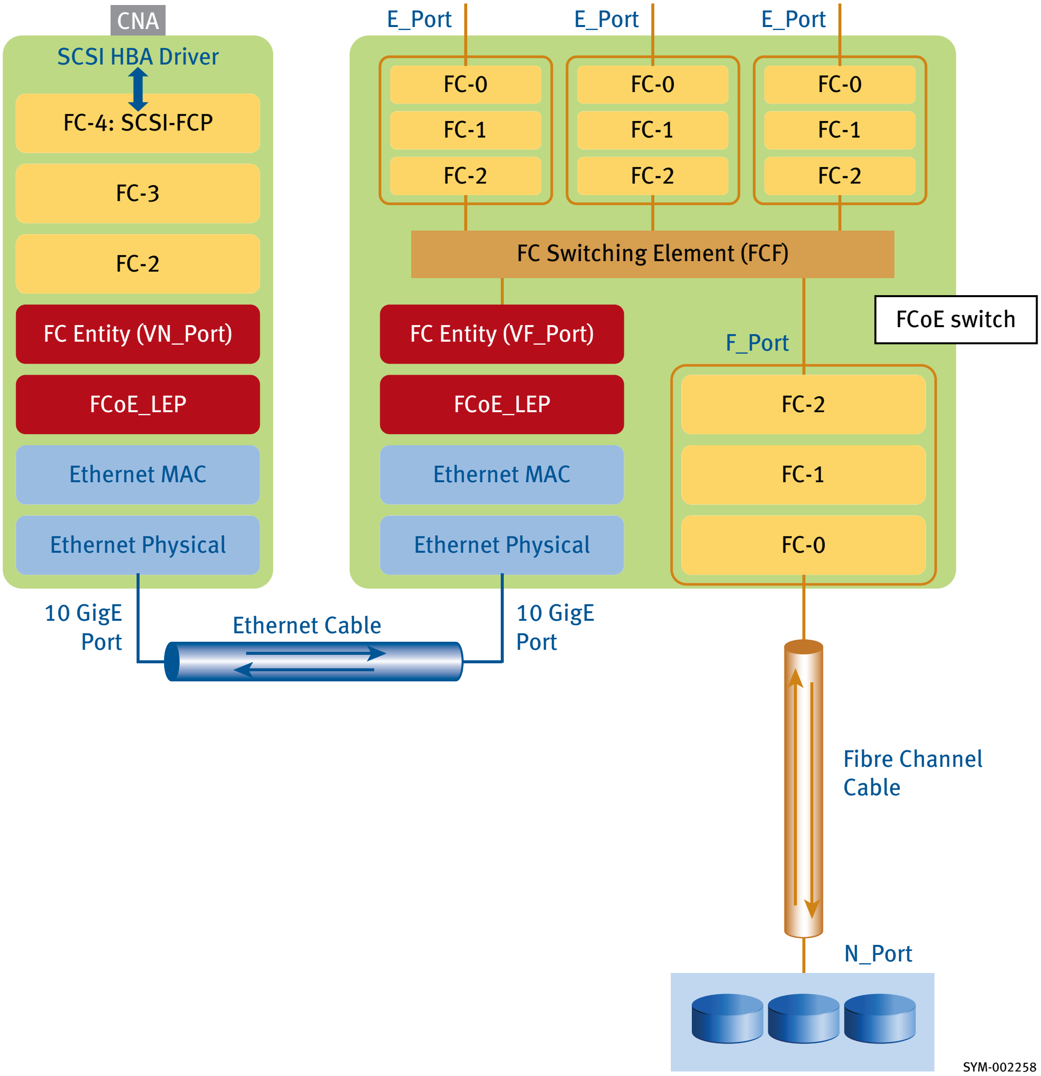


Figure 77 Data path through the FCoE layers

IMPORTANT

Before delving into higher-layer troubleshooting, it is crucial to ensure that lower layers are free and clear of any issues.

Physical interface status

When systematically troubleshooting, the most fundamental thing to check is the physical interface status. Both the Nexus Series and MP-8000B switches use the **show interface** <*intf type*> command to display the status of the interface. The output from these commands provides information as to whether link connectivity between the two end points (such as the link between the CNA port and the switch CEE port) is connected in both the data link and physical layer.

The two outputs in "Output example 1," next, and "Output example 2" on page 191, show that the physical interface is enabled by the administrator and is active. These outputs also indicate that both layer 1 and layer 2 are up, the interface was able to detect a signal, and the data link protocol was able to verify that there is a connected node on the other end of the link. When ports are in this condition, and assuming the port is in a forwarding state, the switch can then start sending and receiving data traffic through these ports.

Output example 1

```
Nexus 5020 # show interface ethernet 1/2
Ethernet1/2 is up
  Hardware: 1000/10000 Ethernet, address: 000d.ecb1.58c9 (bia 000d.ecb1.58c9)
  Description: test
  MTU 1500 bytes, BW 10000000 Kbit, DLY 10 usec,
     reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA
  Port mode is trunk
  full-duplex, 10 Gb/s, media type is 10g
  Beacon is turned off
  Input flow-control is off, output flow-control is off
  Rate mode is dedicated
  Switchport monitor is off
  Last link flapped 2d18h
  Last clearing of "show interface" counters never
  1 minute input rate 48 bits/sec, 0 packets/sec
  1 minute output rate 984 bits/sec, 1 packets/sec
  Rx
    184670 input packets 32481 unicast packets 22208 multicast packets
    129981 broadcast packets 48 jumbo packets 0 storm suppression packets
    16319456 bytes
  Tx
    1253778 output packets 1219722 multicast packets
    0 broadcast packets 48 jumbo packets
    91683067 bytes
    0 input error 0 short frame 0 watchdog
    0 no buffer 0 runt 0 CRC 0 ecc
    0 overrun  0 underrun 0 ignored 0 bad etype drop
    0 bad proto drop 0 if down drop 0 input with dribble
    0 input discard
    0 output error 0 collision 0 deferred
    0 late collision 0 lost carrier 0 no carrier
    0 babble
    0 Rx pause 0 Tx pause
  15 interface resets
```

Output example 2

The following output displays both the physical connectivity status and the data link protocol status. States other than "up" and "line protocol is up" indicate a physical connectivity issue.

```
MP-8000B #show interface tengigabitethernet 0/0
TenGigabitEthernet 0/0 is up, line protocol is up (connected)
Hardware is Ethernet, address is 0005.1e76.a024
    Current address is 0005.1e76.a024
Pluggable media present, Media type is sfp
    Wavelength is 850 nm
Interface index (ifindex) is 402653184
MTU 2500 bytes
LineSpeed: 10000 Mbit, Duplex: Full
Flowcontrol rx: on, tx: on
Last clearing of show interface counters: 6d22h16m
Queueing strategy: fifo
Receive Statistics:
    367310 packets, 446657560 bytes
    Unicasts: 347595, Multicasts: 19382, Broadcasts: 333
    64-byte pkts: 42, Over 64-byte pkts: 160034, Over 127-byte pkts: 152
    Over 255-byte pkts: 49, Over 511-byte pkts: 1771, Over 1023-byte pkts: 4503
    Over 1518-byte pkts(Jumbo): 200759
    Runts: 0, Jabbers: 0, CRC: 0, Overruns: 0
    Errors: 0, Discards: 0
Transmit Statistics:
    633925 packets, 156604746 bytes
    Unicasts: 0, Multicasts: 312891, Broadcasts: 0
    Underruns: 0
    Errors: 0, Discards: 0
Rate info (interval 299 seconds):
    Input 0.000000 Mbits/sec, 0 packets/sec, 0.00% of line-rate
    Output 0.000256 Mbits/sec, 0 packets/sec, 0.00% of line-rate
Time since last interface status change: 4d17h23m
```

The NX-OS **show interface brief** command and CMSH **show ip interface brief** command can be used to show summary port information of all ports on the switch. This command is useful when troubleshooting more than one port.

The following shows two example outputs from the Nexus 5020 and MP-8000B switches:

```
Nexus 5020 # show interface brief
-----------------------------------------------------------------------------
Interface  Vsan   Admin  Admin   Status          SFP    Oper  Oper   Port
                  Mode   Trunk                          Mode  Speed  Channel
                         Mode                                 (Gbps)
-----------------------------------------------------------------------------
fc2/1      1      auto   on      trunking         swl    TE      4    --
fc2/2      1      auto   on      sfpAbsent        --     --           --
fc2/3      1      auto   on      sfpAbsent        --     --           --
fc2/4      1      auto   on      sfpAbsent        --     --           --
fc2/5      1      auto   on      sfpAbsent        --     --           --
fc2/6      1      auto   on      down             swl    --           --
fc2/7      1      auto   on      sfpAbsent        --     --           --
fc2/8      1      auto   on      sfpAbsent        --     --           --

-----------------------------------------------------------------------------
Ethernet      VLAN   Type Mode   Status  Reason                   Speed     Port
Interface                                                                   Ch #
--------------------------------------------------------------------------------
Eth1/1        1      eth  access down    SFP not inserted            10G(D) --
Eth1/2        1      eth  trunk  up      none                        10G(D) --
Eth1/3        1      eth  access down    SFP not inserted            10G(D) --
Eth1/4        1      eth  access down    SFP not inserted            10G(D) --
Eth1/5        1      eth  access down    SFP not inserted            10G(D) --
Eth1/6        1      eth  access down    SFP not inserted            10G(D) --

<output truncated>

MP-8000B # show ip interface brief
-----------------------------------------------------------------------------
Interface                 IP-Address      Status                Protocol
=========                 ==========      ======                ========
TenGigabitEthernet 0/0    unassigned      up                     up
TenGigabitEthernet 0/1    unassigned      up                     down
TenGigabitEthernet 0/2    unassigned      up                     down
TenGigabitEthernet 0/3    unassigned      up                     down
TenGigabitEthernet 0/4    unassigned      up                     down
TenGigabitEthernet 0/5    unassigned      up                     down

<output truncated>
```

Ensure that the link is free from any frame corruption, high interface error rates, and other interface errors associated with layer 1 issues. Any of these errors can impede the operations of those link protocols like Bridge Protocol Data Units (BPDUs). If there are excessive interface errors, a certain number of consecutive BPDUs could be lost, resulting in a *blocking* port transitioning to the *forwarding* state. Usually, the causes of these interface errors are bad twinax or fiber cables, incorrect cable length, bad transceivers (such as SFP+, XFP problems), a faulty CNA, or a faulty switchport.

Interface errors

To check if there are any errors on the interface, use the **show interface** <*intf type*> command and look for any suspicious port errors. Table 25 provides descriptions of each field from the command.

Table 25 show interface command field descriptions (page 1 of 5)

MP-8000B	Nexus 5000	Field	Description
*	*	Ethernet is (up/is administratively down)	Indicates whether the interface hardware is currently active and if it has been taken down by an administrator.
*		line protocol is (up/down)	Indicates whether the software processes that handle the line protocol consider the line usable or if it has been taken down by an administrator.
*	*	Hardware	Hardware type (for example, 1000/10000/Ethernet) and MAC address.
*	*	Description	Alphanumeric string identifying the interface. This only appears if the **description** interface configuration command has been configured on the interface.
*	*	MTU	Maximum transmission unit of the interface.
	*	BW	Bandwidth of the interface in kilobits per second.
*		LineSpeed	Bandwidth of the interface in kilobits per second.
	*	DLY	Delay of the interface in microseconds.
	*	reliability	Reliability of the interface as a fraction of 255 (255/255 is 100 percent reliability), calculated as an exponential average over 5 minutes.
	*	txload, rxload	Load on the interface (in the transmit "tx" and receive "rx" directions) as a fraction of 255 (255/255 is completely saturated), calculated as an exponential average over 5 minutes.
	*	Encapsulation	Encapsulation method assigned to the interface.
*		Duplex: Full	Indicates the duplex mode for the interface. In CEE, the port is always Full-duplex
	*	Full-duplex	Indicates the duplex mode for the interface. In CEE, the port is always Full-duplex

Table 25 show interface command field descriptions (page 2 of 5)

MP-8000B	Nexus 5000	Field	Description
*	*	10 Gb/s or 1000Mbit	Speed of the interface
	*	Last link flapped	Number of days and hours since the last interface reset (flap) recorded by an interface.
*	*	Last clearing	Time at which the counters that measure cumulative statistics (such as number of bytes transmitted and received) shown in this report were last reset to zero. Note that variables that might affect routing (for example, load and reliability) are not cleared when the counters are cleared. A series of asterisks (***) indicates the elapsed time is too large to be displayed. In Nexus 5000, 0:00:00 indicates the counters were cleared more than 2^{31} ms (and less than 2^{32} ms) ago. In MP-8000B, 6d22h16m indicates the counters were cleared 6 days, 22 hours and 16 mins ago.
*		Rate info (interval 299 seconds) or 5 minute rate info	Average number of bits and packets transmitted per second in the last 5 minutes. The 5-minute input and output rates should be used only as an approximation of traffic per second during a given 5-minute period. These rates are exponentially weighted averages with a time constant of 5 minutes. A period of four time constants must pass before the average will be within 2 percent of the instantaneous rate of a uniform stream of traffic over that period.
	*	1 minute input rate, 1 minute output rate	Average number of bits and packets transmitted per second in the last minute. The 1-minute input and output rates should be used only as an approximation of traffic per second during a given 1-minute period. These rates are exponentially weighted averages with a time constant of 1 minute. A period of four time constants must pass before the average will be within 2 percent of the instantaneous rate of a uniform stream of traffic over that period.
	*	Input packets	Total number of error-free packets received by the system.
*		Receive Statistics: packets	Total number of error-free packets received by the system.
	*	Output packets	Total number of error-free packets transmitted by the system.
*		Transmit Statistics: packets	Total number of error-free packets transmitted by the system.
*	*	bytes	Total number of bytes, including data and MAC encapsulation, in the error-free packets received/ transmitted by the system.

Table 25 show interface command field descriptions (page 3 of 5)

MP-8000B	Nexus 5000	Field	Description
		broadcasts	Total number of broadcast packets received/ transmitted by the interface.
*	*	Runt/s	Number of packets that are discarded because they are smaller than the minimum packet size of the medium. For instance, any Ethernet packet that is smaller than 64 bytes is considered a runt.
	*	input errors	Includes runts, giants, no buffer, CRC, frame, overrun, and ignored counts. Other input-related errors can also cause the input errors count to be increased, and some datagrams may have more than one error; therefore, this sum may not balance with the sum of enumerated input error counts.
*		receive statistics: errors	Includes runts, giants, no buffer, CRC, frame, overrun, and ignored counts. Other input-related errors can also cause the input errors count to be increased, and some datagrams may have more than one error; therefore, this sum may not balance with the sum of enumerated input error counts.
*	*	CRC	Cyclic redundancy check generated by the originating LAN station or far-end device does not match the checksum calculated from the data received. On a LAN, this usually indicates noise or transmission problems on the LAN interface or the LAN bus itself. A high number of CRCs is usually the result of collisions or a station transmitting bad data.
	*	frame	Number of packets received incorrectly having a CRC error and a noninteger number of octets. On a LAN, this is usually the result of collisions or a malfunctioning Ethernet device.
*	*	overrun	Number of times the receiver hardware was unable to hand received data to a hardware buffer because the input rate exceeded the receiver's ability to handle the data.
*		jabbers	Jabber is described most often as a frame greater than the maximum of 1518 bytes with bad CRC. A jabbering NIC is often indicative of a hardware problem with a CNA or transceiver.

Table 25 show interface command field descriptions (page 4 of 5)

MP-8000B	Nexus 5000	Field	Description
	*	ignored	Number of received packets ignored by the interface because the interface hardware ran low on internal buffers. These buffers are different than the system buffers. Broadcast storms and bursts of noise can cause the ignored count to be increased.
	*	watchdog	Number of times the watchdog receive timer expired. Expiration happens when receiving a packet with a length greater than 2048 bytes.
*	*	underruns	Number of times that the transmitter has been running faster than the router can handle.
	*	output errors	Sum of all errors that prevented the final transmission of datagrams out of the interface being examined. Note that this may not balance with the sum of the enumerated output errors, as some datagrams may have more than one error and others may have errors that do not fall into any of the specifically tabulated categories.
*		transmit Statistics: errors	Sum of all errors that prevented the final transmission of datagrams out of the interface being examined. Note that this may not balance with the sum of the enumerated output errors, as some datagrams may have more than one error and others may have errors that do not fall into any of the specifically tabulated categories.
	*	collisions	Number of messages retransmitted because of an Ethernet collision. This is usually the result of an overextended LAN (Ethernet or transceiver cable too long, more than two repeaters between stations, or too many cascaded multiport transceivers). A packet that collides is counted only once in output packets.
	*	interface resets	Number of times an interface has been completely reset. This can happen if packets queued for transmission were not sent within several seconds. Interface resets can occur when an interface is looped back or shut down.
	*	input with dribble	Dribble bit error indicates that a frame is slightly too long. This frame error counter is incremented for informational purposes only; the switch accepts the frame.
	*	babbles	Transmit jabber timer expired.
	*	late collision	Number of late collisions. Late collision happens when a collision occurs after transmitting the preamble.

Table 25 show interface command field descriptions (page 5 of 5)

MP-8000B	Nexus 5000	Field	Description
	*	deferred	Number of times that the interface had to defer while ready to transmit a frame because the carrier was asserted.
	*	lost carrier	Number of times the carrier was lost during transmission.
	*	no carrier	Number of times the carrier was not present during the transmission.

Note: If examining performance issue in the MP-8000B switch, the **portperfshow** FOS command can be used to check the IO running on the FCoE ports. The output from the **porterrshow** FOS command does not provide the CEE port statistics for the FCoE ports; it only provides internal statistics for the Zeus to Condor2 connections.

Note: Most of the information in this table was found in the Cisco Command Lookup Tool, which can be located at http://www.cisco.com.

MAC layer

After ensuring that the most obvious physical problems do not exist, it is important to verify that there are no MAC layer issues. It is recommended that you check the MAC address table before performing any higher layer troubleshooting. FCoE will not work properly unless the following three MAC addresses are learned from the host:

- Physical MAC — This is the Burned In Address (BIA) of the CNA.
- Enode MAC — This is the MAC assigned by the CNA and used by FIP.
- FPMA — This is algorithmically derived from the FC-MAP (MAC 0EFC00) and FC_ID assigned during the FLOGI process. This is used for FCoE traffic. The **show fcoe database** command is used on the Nexus.

The command to check the learned MAC addresses on both the Nexus Series switches (NX-OS) and MP-8000B (CMSH) switches is **show mac-address-table**. Both static and dynamically-learned MAC addresses will be shown in the output of this command.

The following are examples of MAC address tables from both the Nexus Series and MP-8000B switches:

Note: In some instances, Cisco only displays Physical MAC and ENODE MAC in the **show mac-address-table command** output.

```
NX-5020 # show mac-address-table | include 1/40
1         00c0.dd10.22c2    dynamic 0         Eth1/40       <-- Physical  MAC
200       00c0.dd10.22c3    dynamic 0         Eth1/40       <-- Enode MAC

NX-5020 # show fcoe database | include 40
vfc40           0x230032        21:00:00:c0:dd:10:22:c3 00:c0:dd:10:22:c3

NX-5020 # show flog database | include 40
vfc40            1     0x230032  21:00:00:c0:dd:10:22:c3 20:00:00:c0:dd:10:22:c3

MP_8000B # show mac-address-table | include 0/8
1        0005.1e9a.a997    Dynamic  Active     Te 0/8        <--  Physical  MAC
1002     0005.1e9a.a995    Dynamic  Active     Te 0/8       <--  Enode MAC
1002     0efc.0001.1001    Dynamic  Active     Te 0/8       <--  FPMA MAC

MP_8000B # show mac-address-table | include 20
1        0005.1e9a.a9d7    Dynamic  Active     Te 0/20       <--  Physical  MAC
1002     0005.1e9a.a9d5    Dynamic  Active     Te 0/20      <--  Enode MAC
1002     0efc.0001.1c01    Dynamic  Active     Te 0/20      <--  FPMA
```

Understanding FCoE phases

To effectively troubleshoot FCoE connectivity and login issues, it is important to understand how the FCoE process works. Before actual storage traffic or I/O can operate, an initiator or FCoE host must establish a login session with the target. The three FCoE stages are discussed in this section.

- "FIP phase" on page 198
- "Fabric Login phase" on page 200
- "FC command phase" on page 203

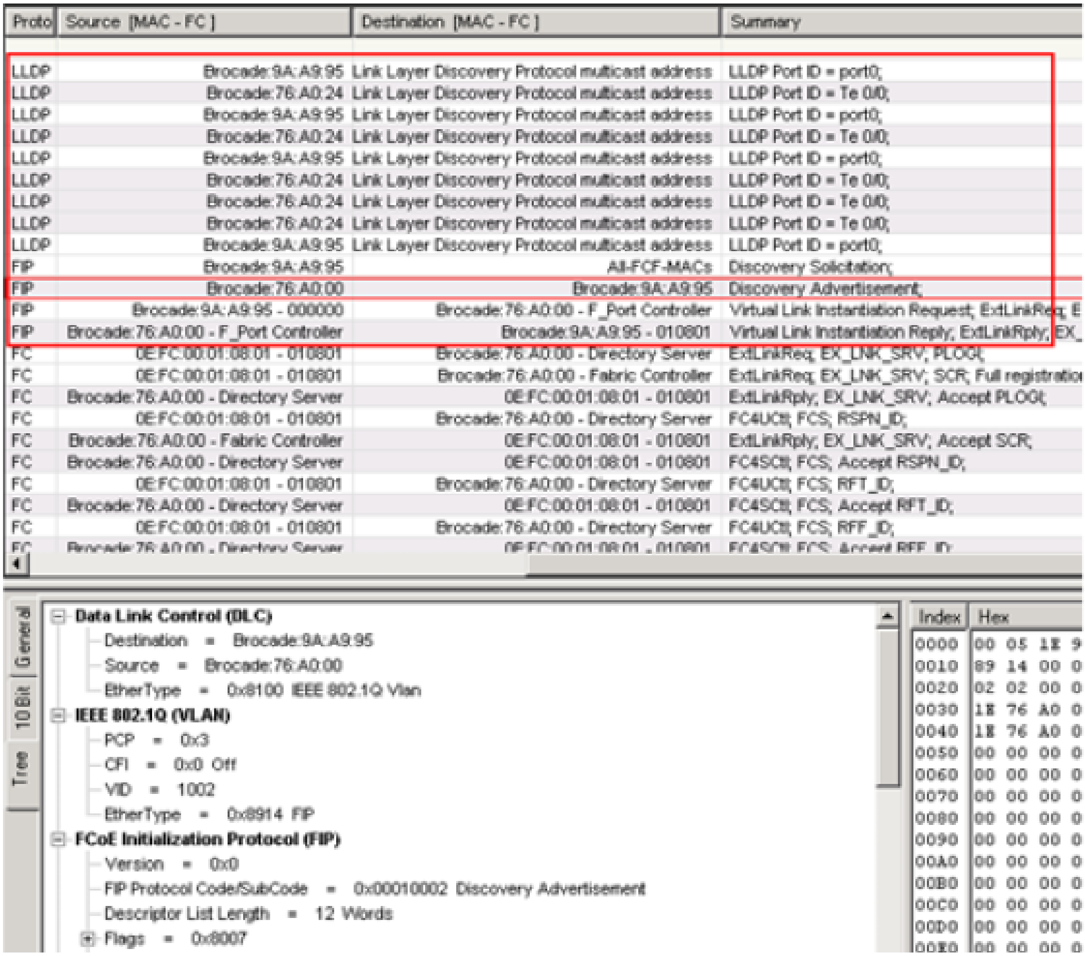
Phase 1 FIP phase

FCoE Initialization Protocol (FIP) is the protocol used to discover FCoE-capable nodes within the CEE network. Refer to "FCoE Initialization Protocol (FIP)" on page 44 for more information. Its role is to assign MAC address and negotiate capabilities.

The FIP phase starts at the discovery stage, wherein CNAs send multicast solicitation messages and FCoE switches reply with unicast advertisement messages. This phase discovers FCoE nodes and negotiates its capabilities.

The following must happen in this phase:

- **Exchange of TLVs**. The CNA and FCoE switch exchange capabilities. For an example, see the highlighted gray from the trace shown in Figure 78 on page 200.
- **Discovery Solicitatio**n. The CNA sends discovery solicitation. See the example from the trace shown in Figure 78.
- **FIP Advertisement**. The FCoE switch replies with FIP advertisement. See the example from the trace shown in Figure 78.



Proto	Source [MAC - FC]	Destination [MAC - FC]	Summary
LLDP	Brocade:9A:A9:95	Link Layer Discovery Protocol multicast address	LLDP Port ID = port0;
LLDP	Brocade:76:A0:24	Link Layer Discovery Protocol multicast address	LLDP Port ID = Te 0/0;
LLDP	Brocade:9A:A9:95	Link Layer Discovery Protocol multicast address	LLDP Port ID = port0;
LLDP	Brocade:76:A0:24	Link Layer Discovery Protocol multicast address	LLDP Port ID = Te 0/0;
LLDP	Brocade:9A:A9:95	Link Layer Discovery Protocol multicast address	LLDP Port ID = port0;
LLDP	Brocade:76:A0:24	Link Layer Discovery Protocol multicast address	LLDP Port ID = Te 0/0;
LLDP	Brocade:76:A0:24	Link Layer Discovery Protocol multicast address	LLDP Port ID = Te 0/0;
LLDP	Brocade:76:A0:24	Link Layer Discovery Protocol multicast address	LLDP Port ID = Te 0/0;
LLDP	Brocade:9A:A9:95	Link Layer Discovery Protocol multicast address	LLDP Port ID = port0;
FIP	Brocade:9A:A9:95	All-FCF-MACs	Discovery Solicitation;
FIP	Brocade:76:A0:00	Brocade:9A:A9:95	Discovery Advertisement;
FIP	Brocade:9A:A9:95 - 000000	Brocade:76:A0:00 - F_Port Controller	Virtual Link Instantiation Request; ExtLinkReq; E
FIP	Brocade:76:A0:00 - F_Port Controller	Brocade:9A:A9:95 - 010801	Virtual Link Instantiation Reply; ExtLinkRply; EX_
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Fabric Controller	ExtLinkReq; EX_LNK_SRV; SCR; Full registratio
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RSPN_ID;
FC	Brocade:76:A0:00 - Fabric Controller	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept SCR;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RSPN_ID;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RFT_ID;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFT_ID;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RFF_ID;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFF_ID;

```
⊟ Data Link Control (DLC)
    Destination  =  Brocade:9A:A9:95
    Source  =  Brocade:76:A0:00
    EtherType  =  0x8100  IEEE 802.1Q Vlan
⊟ IEEE 802.1Q (VLAN)
    PCP  =  0x3
    CFI  =  0x0  Off
    VID  =  1002
    EtherType  =  0x8914  FIP
⊟ FCoE Initialization Protocol (FIP)
    Version  =  0x0
    FIP Protocol Code/SubCode  =  0x00010002  Discovery Advertisement
    Descriptor List Length  =  12 Words
  ⊞ Flags  =  0x8007
```

```
Index  Hex
0000   00 05 1E 9
0010   89 14 00 0
0020   02 02 00 0
0030   1E 76 A0 0
0040   1E 76 A0 0
0050   00 00 00 0
0060   00 00 00 0
0070   00 00 00 0
0080   00 00 00 0
0090   00 00 00 0
00A0   00 00 00 0
00B0   00 00 00 0
00C0   00 00 00 0
00D0   00 00 00 0
00E0   00 00 00 0
```

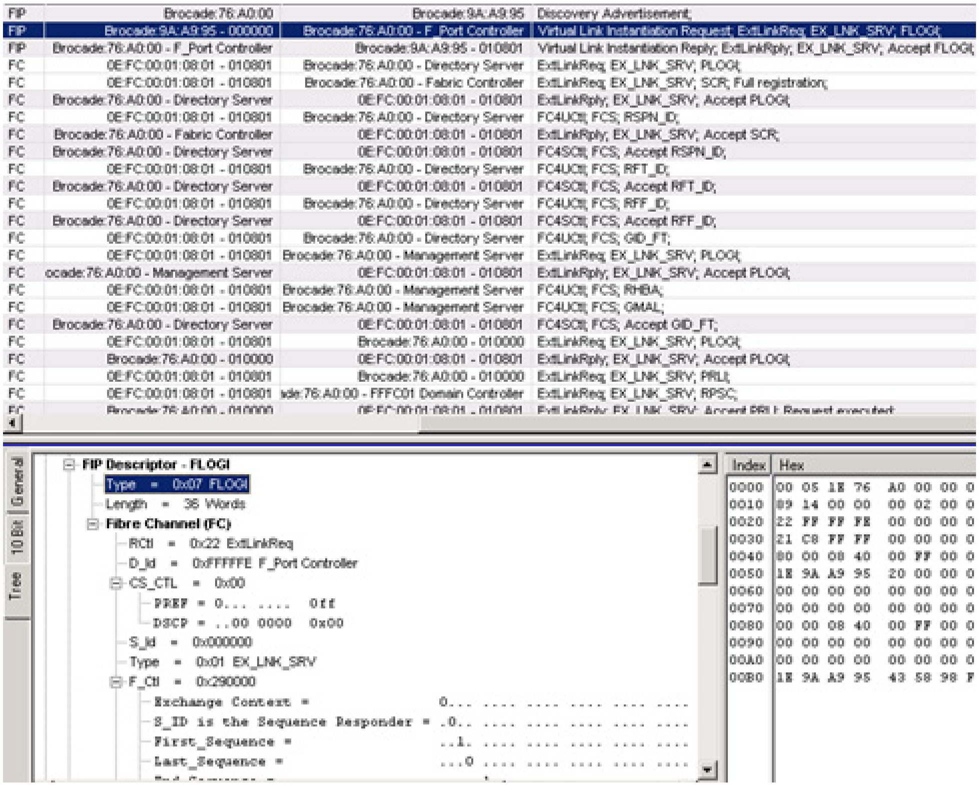
Figure 78 FIP Advertisement example

Phase 2 Fabric Login phase

In this phase, the CNA logs in to the fabric. The following lists what must happen in this phase:

- **FLOGI**. At fabric login, the attached device gets its 24-bit address. See the trace in Figure 79 on page 201 Notice that the FLOGI is carried inside by the FIP frame and not by the FCoE frame. This makes this message easy to intercept by intermediate switches.

- **LS_ACC**. If the FCoE switch accepts the FLOGI, it generates a LS_ACC frame, or an "Accept" message. This is carried inside by the FIP frame and not by the FCoE frame. LS_ACC is also encapsulated in FCoE frames for responses to other commands, such as State Change Registration (SCR, discussed further in this section). Once the attached device registers to the State Change, the FCoE switch will respond with LS_ACC. This "Accept" message is carried inside by the FCoE frame. See the examples from the trace shown in Figure 79.
- **PLOGI**. The device needs to log in to the Directory Server or Name Server. It is important to note that both the switch and CNA will perform PLOGI to each other. See the example from the trace shown in Figure 79. Figure 79 shows that the CNA is performing FLOGI. After that, both CNA and the switch perform PLOGI.



FIP	Brocade:76:A0:00	Brocade:9A:A9:95	Discovery Advertisement;
FIP	Brocade:9A:A9:95 - 000000	Brocade:76:A0:00 - F_Port Controller	Virtual Link Instantiation Request; ExtLinkReq; EX_LNK_SRV; FLOGI;
FIP	Brocade:76:A0:00 - F_Port Controller	Brocade:9A:A9:95 - 010801	Virtual Link Instantiation Reply; ExtLinkRply; EX_LNK_SRV; Accept FLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Fabric Controller	ExtLinkReq; EX_LNK_SRV; SCR; Full registration;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RSPN_ID;
FC	Brocade:76:A0:00 - Fabric Controller	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept SCR;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RSPN_ID;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RFT_ID;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFT_ID;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; RFF_ID;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFF_ID;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Directory Server	FC4UCtl; FCS; GID_FT;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Management Server	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	ocade:76:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Management Server	FC4UCtl; FCS; RHBA;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - Management Server	FC4UCtl; FCS; GMAL;
FC	Brocade:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept GID_FT;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	ExtLinkReq; EX_LNK_SRV; PRLI;
FC	0E:FC:00:01:08:01 - 010801	ade:76:A0:00 - FFFC01 Domain Controller	ExtLinkReq; EX_LNK_SRV; RPSC;
FC	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PRLI; Request executed;

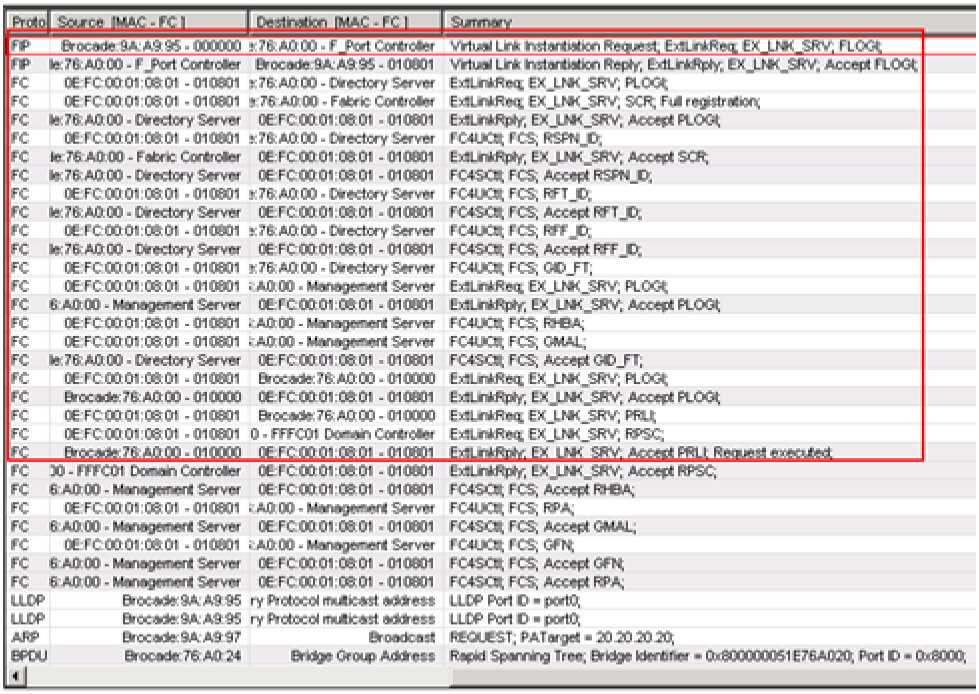
```
FIP Descriptor - FLOGI
   Type   =   0x07 FLOGI
   Length   =   36 Words
   Fibre Channel (FC)
      RCtl   =   0x22 ExtLinkReq
      D_Id   =   0xFFFFFE F_Port Controller
      CS_CTL   =   0x00
         PREF = 0... ....   Off
         DSCP = ..00 0000   0x00
      S_Id   =   0x000000
      Type   =   0x01 EX_LNK_SRV
      F_Ctl   =   0x290000
         Exchange Context =                    0... .... .... .... .... ....
         S_ID is the Sequence Responder =      .0.. .... .... .... .... ....
         First_Sequence =                      ..1. .... .... .... .... ....
         Last_Sequence =                       ...0 .... .... .... .... ....
```

```
Index  Hex
0000   00 05 1E 76   A0 00 00 0
0010   09 14 00 00   00 02 00 0
0020   22 FF FF FE   00 00 00 0
0030   21 C8 FF FF   00 00 00 0
0040   80 00 08 40   00 FF 00 0
0050   1E 9A A9 95   20 00 00 0
0060   00 00 00 00   00 00 00 0
0070   00 00 00 00   00 00 00 0
0080   00 00 08 40   00 FF 00 0
0090   00 00 00 00   00 00 00 0
00A0   00 00 00 00   00 00 00 0
00B0   1E 9A A9 95   43 58 98 F
```

Figure 79 Fabric Login phase, example 1

- **SCR**. State Change Registration. The attached device needs to register for state changes so that if there is a change in the fabric (such as a zoning change) that affects this device, then the device will be notified through the RSCN. See the example from the trace shown in Figure 80 on page 203.
- **GXX**. This is similar to the **get** command. It is usually a GID_PT or GID_FT. The **get** command is used to get a list of 24-bit addresses of the devices that are currently logged in to the fabric and that this device has access to or is zoned to. See the example GID_FT command from the trace shown in Figure 80.
- **RXXX**. One of the other commands that can be seen is the RXXX command, like the RSPN_ID and RFT_ID shown in Figure 80. These are register commands, where the attached device will register information (such as symbolic node name) with the switches' Name Server. What the device does depends on how the device driver is programmed. See the example from the trace shown in Figure 80.
- **PRLI**. Process Login is used by upper layers, like SCSI. PRLI is used to establish an upper level process of the node with an established upper level process of another node.

The traces in Figure 80 show the FLOGI, Accept FLOGI or LS_ACC, PLOGI, RSPN_ID, SCR, GID_FT, and PRLI as part of the Fabric Login Phase.

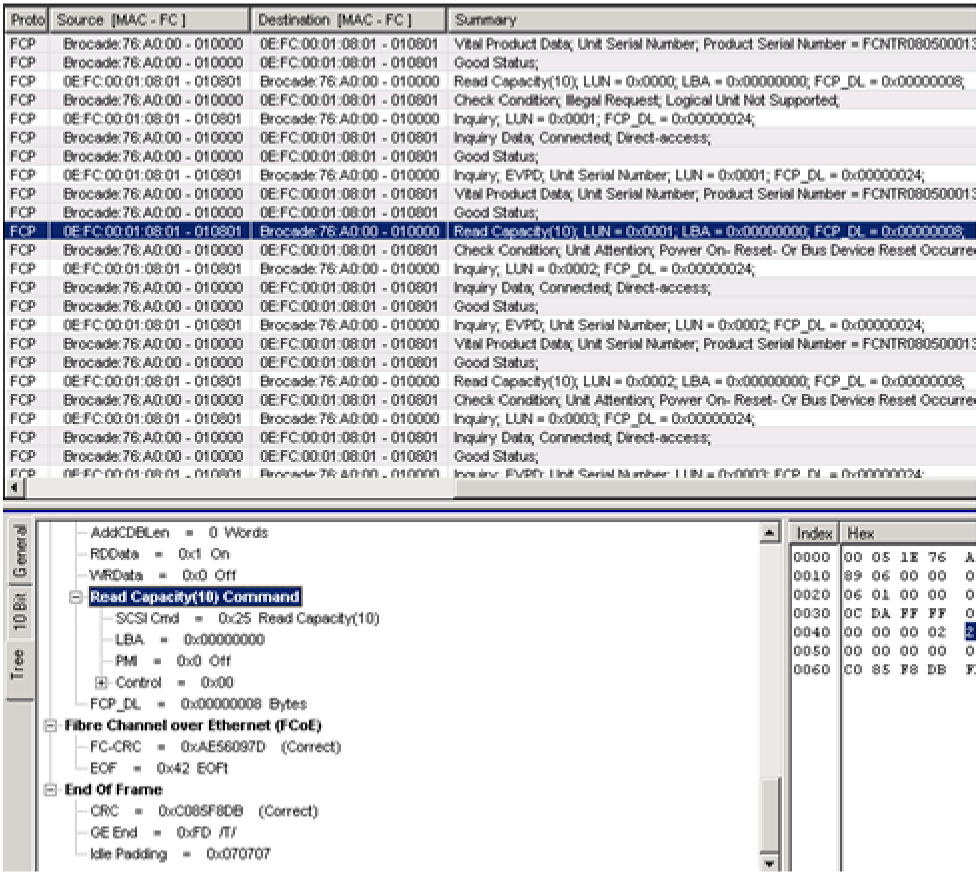


Proto	Source [MAC - FC]	Destination [MAC - FC]	Summary
FIP	Brocade:9A:A9:95 - 000000	e:76:A0:00 - F_Port Controller	Virtual Link Instantiation Request; ExtLinkReq; EX_LNK_SRV; FLOGI;
FIP	le:76:A0:00 - F_Port Controller	Brocade:9A:A9:95 - 010801	Virtual Link Instantiation Reply; ExtLinkRply; EX_LNK_SRV; Accept FLOGI;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Directory Server	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Fabric Controller	ExtLinkReq; EX_LNK_SRV; SCR; Full registration;
FC	le:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Directory Server	FC4UCtl; FCS; RSPN_ID;
FC	le:76:A0:00 - Fabric Controller	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept SCR;
FC	le:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RSPN_ID;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Directory Server	FC4UCtl; FCS; RFT_ID;
FC	le:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFT_ID;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Directory Server	FC4UCtl; FCS; RFF_ID;
FC	le:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RFF_ID;
FC	0E:FC:00:01:08:01 - 010801	e:76:A0:00 - Directory Server	FC4UCtl; FCS; GID_FT;
FC	0E:FC:00:01:08:01 - 010801	:A0:00 - Management Server	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	6:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	:A0:00 - Management Server	FC4UCtl; FCS; RHBA;
FC	0E:FC:00:01:08:01 - 010801	:A0:00 - Management Server	FC4UCtl; FCS; GMAL;
FC	le:76:A0:00 - Directory Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept GID_FT;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	ExtLinkReq; EX_LNK_SRV; PLOGI;
FC	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PLOGI;
FC	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	ExtLinkReq; EX_LNK_SRV; PRLI;
FC	0E:FC:00:01:08:01 - 010801	0 - FFFC01 Domain Controller	ExtLinkReq; EX_LNK_SRV; RPSC;
FC	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept PRLI; Request executed;
FC	00 - FFFC01 Domain Controller	0E:FC:00:01:08:01 - 010801	ExtLinkRply; EX_LNK_SRV; Accept RPSC;
FC	6:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RHBA;
FC	0E:FC:00:01:08:01 - 010801	:A0:00 - Management Server	FC4UCtl; FCS; RPA;
FC	6:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept GMAL;
FC	0E:FC:00:01:08:01 - 010801	:A0:00 - Management Server	FC4UCtl; FCS; GFN;
FC	6:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept GFN;
FC	6:A0:00 - Management Server	0E:FC:00:01:08:01 - 010801	FC4SCtl; FCS; Accept RPA;
LLDP	Brocade:9A:A9:95	ry Protocol multicast address	LLDP Port ID = port0;
LLDP	Brocade:9A:A9:95	ry Protocol multicast address	LLDP Port ID = port0;
ARP	Brocade:9A:A9:97	Broadcast	REQUEST; PATarget = 20.20.20.20;
BPDU	Brocade:76:A0:24	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x800000051E76A020; Port ID = 0x8000;

Figure 80 Fabric Login phase trace, example 2

Phase 3 FC command phase

Once the CNAs have gone through the Fabric Login phase, the FCoE host can now start sending regular FC frames using CEE as a transport, thus FCoE frames. In this phase, SCSI FCP data and commands can be seen. See the **Read Capacity(10)** command example in the trace shown in Figure 81 on page 204.



Proto	Source [MAC - FC]	Destination [MAC - FC]	Summary
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Vital Product Data; Unit Serial Number; Product Serial Number = FCNTR08050001
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Read Capacity(10); LUN = 0x0000; LBA = 0x00000000; FCP_DL = 0x00000008;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Check Condition; Illegal Request; Logical Unit Not Supported;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; LUN = 0x0001; FCP_DL = 0x00000024;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Inquiry Data; Connected; Direct-access;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; EVPD; Unit Serial Number; LUN = 0x0001; FCP_DL = 0x00000024;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Vital Product Data; Unit Serial Number; Product Serial Number = FCNTR08050001
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Read Capacity(10); LUN = 0x0001; LBA = 0x00000000; FCP_DL = 0x00000008;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Check Condition; Unit Attention; Power On- Reset- Or Bus Device Reset Occurre
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; LUN = 0x0002; FCP_DL = 0x00000024;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Inquiry Data; Connected; Direct-access;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; EVPD; Unit Serial Number; LUN = 0x0002; FCP_DL = 0x00000024;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Vital Product Data; Unit Serial Number; Product Serial Number = FCNTR08050001
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Read Capacity(10); LUN = 0x0002; LBA = 0x00000000; FCP_DL = 0x00000008;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Check Condition; Unit Attention; Power On- Reset- Or Bus Device Reset Occurre
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; LUN = 0x0003; FCP_DL = 0x00000024;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Inquiry Data; Connected; Direct-access;
FCP	Brocade:76:A0:00 - 010000	0E:FC:00:01:08:01 - 010801	Good Status;
FCP	0E:FC:00:01:08:01 - 010801	Brocade:76:A0:00 - 010000	Inquiry; EVPD; Unit Serial Number; LUN = 0x0003; FCP_DL = 0x00000024;

```
      AddCDBLen  =  0 Words
      RDData  =  0x1  On
      WRData  =  0x0  Off
   ⊟ Read Capacity(10) Command
        SCSI Cmd  =  0x25  Read Capacity(10)
        LBA  =  0x00000000
        PMI  =  0x0  Off
      ⊞ Control  =  0x00
      FCP_DL  =  0x00000008  Bytes
⊟ Fibre Channel over Ethernet (FCoE)
      FC-CRC  =  0xAE56097D   (Correct)
      EOF  =  0x42  EOFt
⊟ End Of Frame
      CRC  =  0xC085F8DB   (Correct)
      GE End  =  0xFD  /T/
      Idle Padding  =  0x070707
```

```
Index  Hex
0000   00 05 1E 76   A
0010   89 06 00 00   0
0020   06 01 00 00   0
0030   0C DA FF FF   0
0040   00 00 00 02   2
0050   00 00 00 00   0
0060   C0 85 F8 DB   F
```

Figure 81   FC command phase example

fcping and fctraceroute commands

Native Fibre Channel switches have an **fcping** command for connectivity testing. They also have an **fctrace** (fctraceroute) command to verify the path from the switch's port to the node's port. FCtrace also computes the interswitch hop latency. These commands are similar to the **ping** and **traceroute** commands used in the IP world. Nexus Series switches (NX-OS) and MP-8000B (FOS) support the **fcping** and **fctrace** diagnostics commands that can be used to test the link availability or connectivity between end points, that is, from a Nexus switch to a front-end port of the storage array. The following are some examples of these diagnostic commands.

Syntax Nexus 5000

```
fcping {device-alias|fcid|PWWN} <value> vsan <vsan number>
fctrace {device-alias|fcid|PWWN} <value> vsan <vsan number>
```

Note: Fabric Manager supports these diagnostic tools.

In this example, Host1 is zoned to a VNX series or CLARiiON port and it is configured under VSAN 667, as shown in Figure 82.

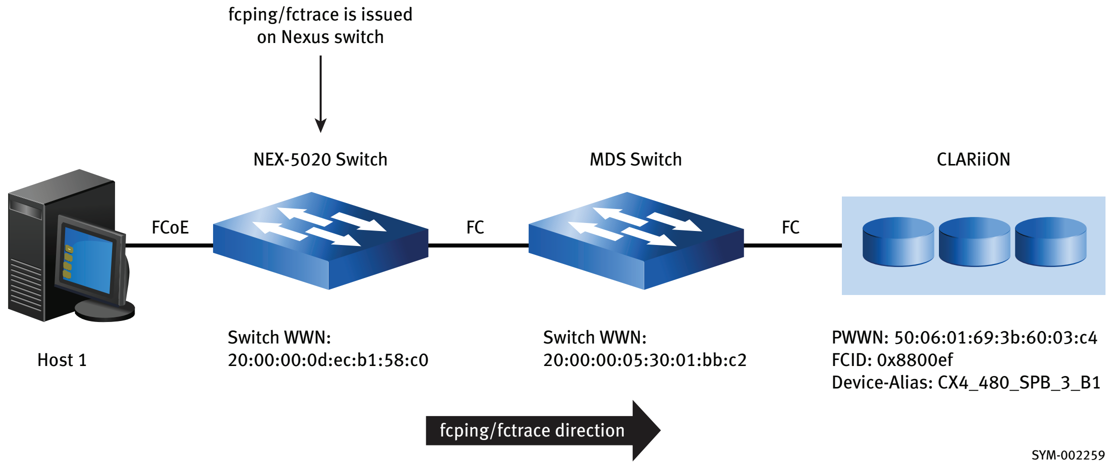


Figure 82 Nexus Series switch example

```
Nexus 5020 # fcping ?
  device-alias  Device-alias of the destination N-Port
  fcid          FC-id of the destination N-Port
  PWWN          PWWN of the destination N-Port

Nexus 5020# fcping PWWN 50:06:01:69:3b:60:03:c4 vsan 667
28 bytes from 50:06:01:69:3b:60:03:c4 time = 254 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 256 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 168 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 222 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 258 usec

5 frames sent, 5 frames received, 0 timeouts
Round-trip min/avg/max = 168/231/258 usec

Nexus 5020 # fcping device-alias CX4_480_SPB_3_B1 vsan 667
28 bytes from 50:06:01:69:3b:60:03:c4 time = 277 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 251 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 222 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 249 usec
28 bytes from 50:06:01:69:3b:60:03:c4 time = 232 usec
```

```
5 frames sent, 5 frames received, 0 timeouts
Round-trip min/avg/max = 222/246/277 usec

Nexus 5020 # fctrace ?
  device-alias  Device-alias of the destination N-Port
  fcid          FC-id of the destination N-Port
  PWWN          PWWN of the destination N-Port

Nexus 5020 # fctrace PWWN 50:06:01:69:3b:60:03:c4 vsan 667
Route present for : 50:06:01:69:3b:60:03:c4
20:00:00:0d:ec:b1:58:c0(0xfffcb3)
20:00:00:05:30:01:bb:32(0xfffc88)
20:00:00:05:30:01:bb:32(0xfffc88)

Nexus 5020 # fctrace fcid 0x8800ef vsan 667
Route present for :  0x8800ef
20:00:00:0d:ec:b1:58:c0(0xfffcb3)
20:00:00:05:30:01:bb:32(0xfffc88)
20:00:00:05:30:01:bb:32(0xfffc88)
```

MP-8000B

In the MP-8000B (FOS), **fcping** is executed the same way as on the Nexus Series switches, with the exception that you can **fcping** directly to PWWN or fcid without using the word "fcid" or "PWWN". The **fctrace** command is only supported in EMC Connectrix Manger Data Center Edition, or CMDCE, a GUI-based management suite.

```
fcping <fcid value| PWWN value>
fctrace - this can only be done on CMDCE.

MP-8000B: admin> fcping  21:00:00:c0:dd:10:28:bb;20
Destination:    21:00:00:c0:dd:10:28:bb

Pinging 21:00:00:c0:dd:10:28:bb [0x10801] with 12 bytes of data:
received reply from 21:00:00:c0:dd:10:28:bb: 12 bytes time:262 usec
received reply from 21:00:00:c0:dd:10:28:bb: 12 bytes time:152 usec
received reply from 21:00:00:c0:dd:10:28:bb: 12 bytes time:144 usec
received reply from 21:00:00:c0:dd:10:28:bb: 12 bytes time:143 usec
received reply from 21:00:00:c0:dd:10:28:bb: 12 bytes time:146 usec
5 frames sent, 5 frames received, 0 frames rejected, 0 frames timeout
Round-trip min/avg/max = 143/169/262 usec
MP-8000B: admin> fcping  0x010801
Destination:    0x10801
Pinging 0x10801 with 12 bytes of data:
received reply from 0x10801: 12 bytes time:334 usec
received reply from 0x10801: 12 bytes time:159 usec
received reply from 0x10801: 12 bytes time:143 usec
received reply from 0x10801: 12 bytes time:151 usec
received reply from 0x10801: 12 bytes time:144 usec
5 frames sent, 5 frames received, 0 frames rejected, 0 frames timeout
Round-trip min/avg/max = 143/186/334 usec
```

Upper layer protocol

The FC4 layer defines how the upper layer protocols (ULPs) map to the lower layers of Fibre Channel. It allows different protocols to be transported using the same physical port. These protocols are SCSI, HiPPI, ESCON, FICON, ATM, SONET, and IP.

Each of these protocols has specifications, which are responsible for defining how its data, command, status, sense information, and other protocol-specific information will be mapped into the FC frames using standards or defined formats. These specifications are handled by the CNA/HBA device drivers on both the initiator and target. Therefore, in a scenario where it is suspected that an FC-4 problem exists, checking the device drivers and firmware versions of your nodes is valuable.

If you believe you have found an Upper Layer Protocol problem, most likely the issue is a bug in the device driver, incompatibilities, or a hardware problem. When FC-4 troubleshooting, first ensure that there are no lower-layer problems, as they are the common culprit for causing upper-layer protocol (ULP) issues.

FCoE and CEE troubleshooting case studies

This section discusses two troubleshooting case studies that explain how to troubleshoot and resolve common issues encountered when setting up an FCoE/iSCSI environment:

- "Case Study #1, Unable to access the LUNs/devices" on page 208
- "Case Study #2, Unable to access a shared folder in the file server" on page 251

Each case study uses a flowchart to better illustrate the troubleshooting process.

Case Study #1, Unable to access the LUNs/devices

Problem definition

Unable to access the LUNs/devices being presented by the storage array.

Background

The host was not able to see the LUNs/devices presented by the storage array. The troubleshooting techniques used in this example are based on the concepts discussed in "Troubleshooting basic FCoE and CEE problems" on page 180. This example troubleshoots lower layer problems first, and then proceeds to higher layers, such as switch port configurations, storage configuration, zoning, host configuration, and so on.

Topology

A fabric can become complicated, such as full-mesh design, three-tiered designs with redundant connectivity and redundant switches on each level, or a SAN port channel configured between switches, to name just a few. It all comes down to the basic concept that a host is connected to a fabric (a switch or switches), and then this fabric is connected to a storage array. For the sake of simplicity, this example breaks down the converged FCoE and SAN topology into one Nexus 5020 switch and an MDS switch, as shown in Figure 83 on page 209.

In the FCoE environment shown in Figure 83, the customer has a CLARiiON CX4-480 with three LUNs assigned to the host.

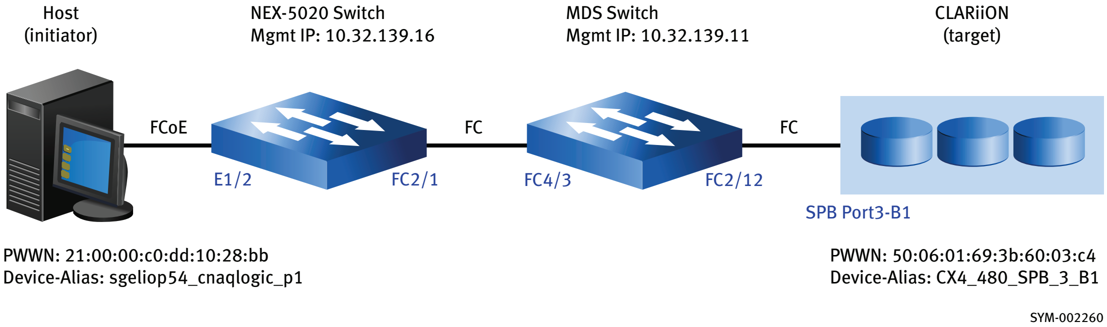


Figure 83 Case study #1 topology

This case study will analyze issues using the troubleshooting flowchart shown in Figure 84 on page 210. This flowchart shows how to proceed with various troubleshooting techniques that were discussed in "Troubleshooting basic FCoE and CEE problems" on page 180. This flowchart can better guide you to solving various issues. Examples are provided for each step in the flowchart.

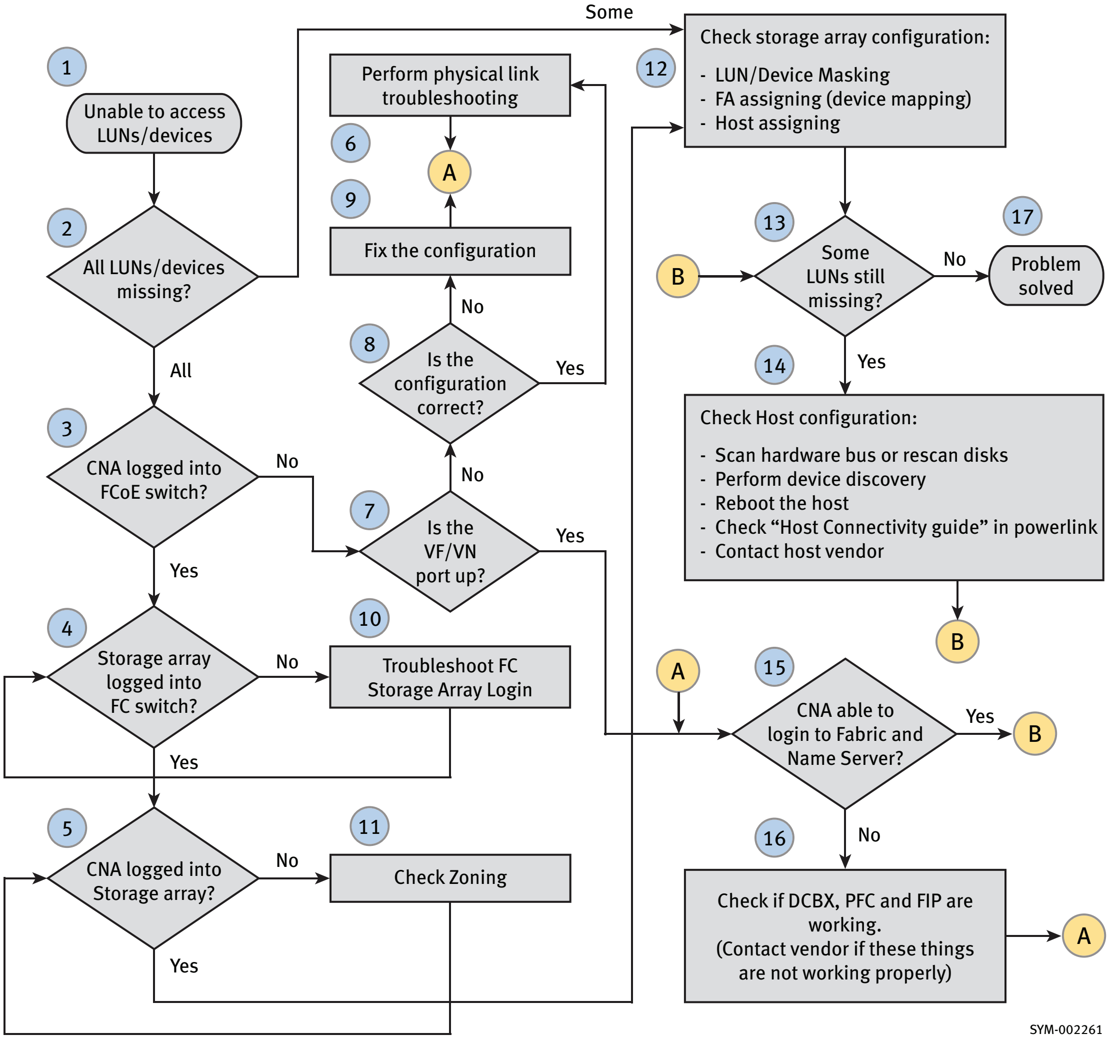

1
Unable to access LUNs/devices
2
All LUNs/devices missing?
Some
All
3
CNA logged into FCoE switch?
No
Yes
4
Storage array logged into FC switch?
No
Yes
5
CNA logged into Storage array?
No
Yes
6
Perform physical link troubleshooting
A
9
Fix the configuration
8
Is the configuration correct?
No
Yes
7
Is the VF/VN port up?
No
Yes
10
Troubleshoot FC Storage Array Login
11
Check Zoning
12
Check storage array configuration:
- LUN/Device Masking
- FA assigning (device mapping)
- Host assigning
13
B
Some LUNs still missing?
No
17
Problem solved
14
Yes
Check Host configuration:
- Scan hardware bus or rescan disks
- Perform device discovery
- Reboot the host
- Check "Host Connectivity guide" in powerlink
- Contact host vendor
B
A
15
CNA able to login to Fabric and Name Server?
Yes
B
16
No
Check if DCBX, PFC and FIP are working.
(Contact vendor if these things are not working properly)
A
SYM-002261

Figure 84 **Troubleshooting flowchart for case study #1**

Using the flowchart in Figure 84, each step will be further discussed in this section.

- "Flowchart step #1, Unable to access LUNs/devices" on page 211
- "Flowchart step #2, Are ALL LUNs/devices missing?" on page 211
- "Flowchart step #3, CNA logged in to FCoE switch?" on page 215

- "Flowchart step #4, Storage array logged in to FC switch?" on page 218
- "Flowchart step #5, CNA logged in to storage array?" on page 220
- "Flowchart step #6, Perform physical link troubleshooting" on page 224
- "Flowchart step #7, Is the VF/VN port up?" on page 227
- "Flowchart step #8, Is the configuration correct?" on page 229
- "Flowchart step #9, Fix the configuration" on page 234
- "Flowchart step #10, Troubleshoot FC storage array login" on page 237
- "Flowchart step #11, Check zoning" on page 239
- "Flowchart step #12, Check storage array configuration" on page 241
- "Flowchart step #13, Some LUNs still missing?" on page 245
- "Flowchart step #14, Check host configuration" on page 245
- "Flowchart step #15, CNA able to log in to fabric and name server?" on page 246
- "Flowchart step #16, Check if DCBX, PFC, and FIP are working" on page 247
- "Flowchart step #17, Problem solved" on page 251

Flowchart step #1, Unable to access LUNs/devices

In step 1, the problem is defined. In this example, the problem is that the customer is unable to access the LUNs/devices.

Flowchart step #2, Are ALL LUNs/devices missing?

Troubleshooting

1. Using disk management utility of your host, verify if ALL or SOME of the LUNs are missing. The diskpart or inq utility tool can be used on a Windows host.
2. Depending on the result of Step 1, follow the instructions in "Next steps" and proceed with the troubleshooting.

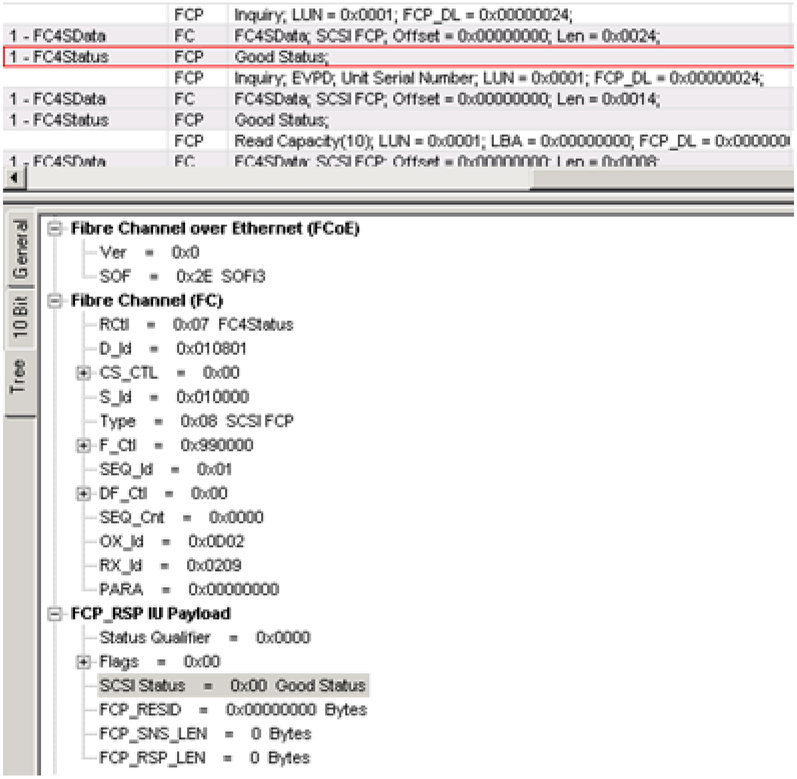
Example and interpretation of the results

The following example shows the output of the **inq** command on a Windows host when three LUNs from a VNX series or CLARiiON storage system are being presented to it. Notice that there are three

LUNs configured with RAID 5. The first RAID 5 LUN has 3 GB of storage, while the second and third have 1 GB of storage space each. Most of the time, the inq output will display LUN information such as the VEND, PROD, REV, or SER NUM that are visible to the operating system. These are the returned responses to the SCSI inquiry command, as shown in Figure 85 on page 213 and Figure 86 on page 214. The CAP, or capacity, information is returned in the SCSI read capacity command. If the answer on this flowchart symbol is **SOME**, you would see only one or two of the LUNs presented by the VNX series or CLARiiON device.

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.

.....

-------------------------------------------------------------------------
DEVICE             :VEND    :PROD            :REV   :SER NUM    :CAP(kb)
-------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE1 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE2 :DGC     :RAID 5          :0429  :07000097   :3145728
\\.\PHYSICALDRIVE3 :DGC     :RAID 5          :0429  :08000097   :1048576
\\.\PHYSICALDRIVE4 :DGC     :RAID 5          :0429  :09000097   :1048576
```

Figure 85 shows an example of a successful SCSI inquiry command.

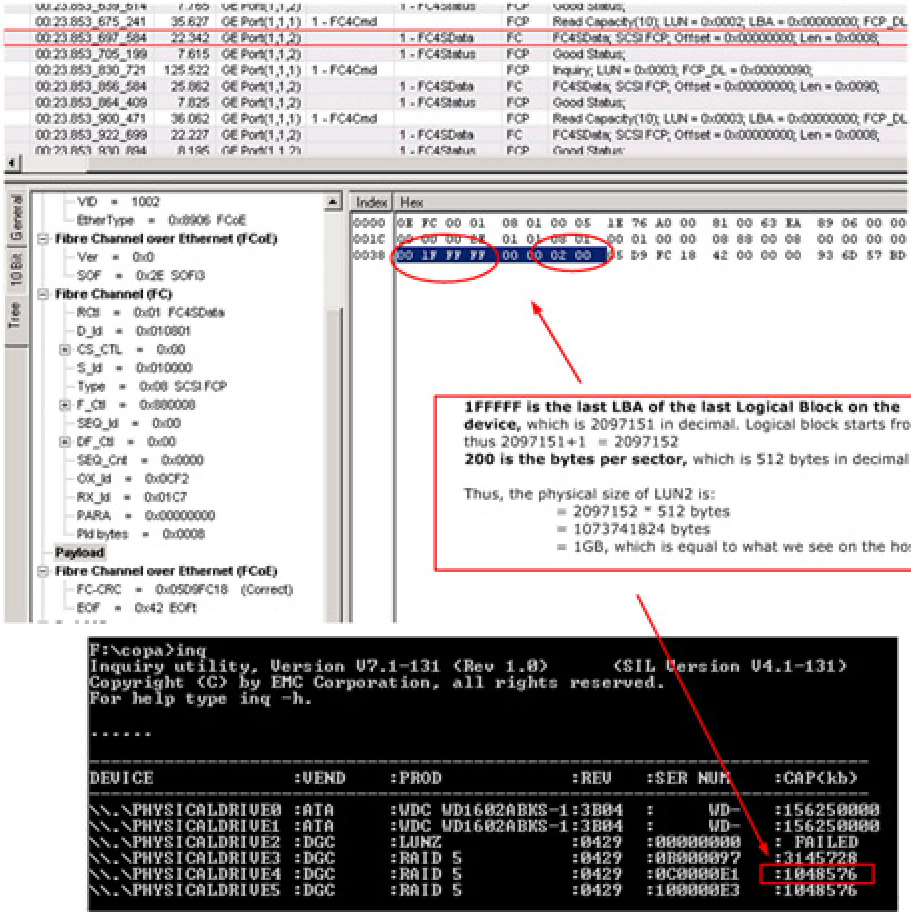


```
                  FCP    Inquiry; LUN = 0x0001; FCP_DL = 0x00000024;
1 - FC4SData      FC     FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0024;
1 - FC4Status     FCP    Good Status;
                  FCP    Inquiry; EVPD; Unit Serial Number; LUN = 0x0001; FCP_DL = 0x00000024;
1 - FC4SData      FC     FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0014;
1 - FC4Status     FCP    Good Status;
                  FCP    Read Capacity(10); LUN = 0x0001; LBA = 0x00000000; FCP_DL = 0x000000
1 - FC4SData      FC     FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0008;
```

```
⊟ Fibre Channel over Ethernet (FCoE)
   Ver  =  0x0
   SOF  =  0x2E  SOFi3
⊟ Fibre Channel (FC)
   RCtl  =  0x07  FC4Status
   D_Id  =  0x010801
 ⊞ CS_CTL  =  0x00
   S_Id  =  0x010000
   Type  =  0x08  SCSI FCP
 ⊞ F_Ctl  =  0x990000
   SEQ_Id  =  0x01
 ⊞ DF_Ctl  =  0x00
   SEQ_Cnt  =  0x0000
   OX_Id  =  0x0D02
   RX_Id  =  0x0209
   PARA  =  0x00000000
⊟ FCP_RSP IU Payload
   Status Qualifier  =  0x0000
 ⊞ Flags  =  0x00
   SCSI Status  =  0x00  Good Status
   FCP_RESID  =  0x00000000  Bytes
   FCP_SNS_LEN  =  0  Bytes
   FCP_RSP_LEN  =  0  Bytes
```

Figure 85 Successful SCSI Inquiry command example

Figure 86 shows an example of a successful SCSI Read Capacity command.



```
00:23.853_675_241   35.627  GE Port(1,1,1)  1 - FC4Cmd                 FCP  Read Capacity(10); LUN = 0x0002; LBA = 0x00000000; FCP_DL
00:23.853_697_584   22.342  GE Port(1,1,2)                 1 - FC4SData  FC   FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0008;
00:23.853_705_199    7.615  GE Port(1,1,2)                 1 - FC4Status FCP  Good Status;
00:23.853_830_721  125.522  GE Port(1,1,1)  1 - FC4Cmd                 FCP  Inquiry; LUN = 0x0003; FCP_DL = 0x00000090;
00:23.853_856_584   25.862  GE Port(1,1,2)                 1 - FC4SData  FC   FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0090;
00:23.853_864_409    7.825  GE Port(1,1,2)                 1 - FC4Status FCP  Good Status;
00:23.853_900_471   36.062  GE Port(1,1,1)  1 - FC4Cmd                 FCP  Read Capacity(10); LUN = 0x0003; LBA = 0x00000000; FCP_DL
00:23.853_922_699   22.227  GE Port(1,1,2)                 1 - FC4SData  FC   FC4SData; SCSI FCP; Offset = 0x00000000; Len = 0x0008;
00:23.853_930_894    8.195  GE Port(1,1,2)                 1 - FC4Status FCP  Good Status;
```

```
  VID  =  1002
  EtherType  =  0x8906 FCoE
Fibre Channel over Ethernet (FCoE)
  Ver  =  0x0
  SOF  =  0x2E SOFi3
Fibre Channel (FC)
  RCtl  =  0x01 FC4SData
  D_Id  =  0x010801
  CS_CTL  =  0x00
  S_Id  =  0x010000
  Type  =  0x08 SCSI FCP
  F_Ctl  =  0x880008
  SEQ_Id  =  0x00
  DF_Ctl  =  0x00
  SEQ_Cnt  =  0x0000
  OX_Id  =  0x0CF2
  RX_Id  =  0x01C7
  PARA  =  0x00000000
  Pld bytes  =  0x0008
Payload
Fibre Channel over Ethernet (FCoE)
  FC-CRC  =  0x05D9FC18  (Correct)
  EOF  =  0x42 EOFt
```

```
Index  Hex
0000   0E FC 00 01  08 01 00 05  1E 76 A0 00  01 00 63 EA  89 06 00 00
001C   00 00 00 1E  01 01 00 01  00 01 00 00  08 88 00 08  00 00 00 00
0038   00 1F FF FF  00 00 02 00  05 D9 FC 18  42 00 00 00  93 6D 57 BD
```

1FFFFF is the last LBA of the last Logical Block on the device, which is 2097151 in decimal. Logical block starts from 0, thus 2097151+1 = 2097152
200 is the bytes per sector, which is 512 bytes in decimal

Thus, the physical size of LUN2 is:
= 2097152 * 512 bytes
= 1073741824 bytes
= 1GB, which is equal to what we see on the host.

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.

......

------------------------------------------------------------------------------
DEVICE                 :VEND     :PROD              :REV    :SER NUM    :CAP(kb)
------------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE1 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE2 :DGC      :LUNZ             :0429   :00000000    :  FAILED
\\.\PHYSICALDRIVE3 :DGC      :RAID 5           :0429   :0B000097    :3145728
\\.\PHYSICALDRIVE4 :DGC      :RAID 5           :0429   :0C0000E1    :1048576
\\.\PHYSICALDRIVE5 :DGC      :RAID 5           :0429   :100000E3    :1048576
```

Figure 86 Successful SCSI Read Capacity command example

The following is another example of output when the **inq** command is issued on a Windows host. In this case, there were no VNX series or CLARiiON LUNs found when the command was issued; therefore, you would answer this flowchart symbol with **ALL** and proceed to the next step.

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.
```

```
.....

-------------------------------------------------------------------------
DEVICE             :VEND    :PROD            :REV   :SER NUM    :CAP(kb)
-------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE1 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
```

Next steps

On this flowchart step, the scope of the problem needs to be determined. The issue could be either:

- The customer is unable to access all VNX series or CLARiiON LUNs/devices.

  In this case, proceed with verifying if the CNA is logged in to the switch, which is depicted in flowchart step #3.

- The customer is able to see only some of the VNX series or CLARiiON LUNs/devices. In this case, proceed to VNX series or CLARiiON configuration troubleshooting, depicted in flowchart step #12.

Flowchart step #3, CNA logged in to FCoE switch?

Troubleshooting

Note: Other than CLI, you can use switch management software to verify whether the initiator device has logged in to the FCoE switch. In Brocade switches, you can use CMDCE, while in Cisco MDS switches you can use Fabric Manager.

In order to verify whether the CNA has logged to the Nexus 5000 switch, complete the following steps.

1. Before verifying if the host's CNA has logged to the FCoE switch, it is important to verify whether the FCoE switch is in fabric mode. This is the mode wherein the switch provides standard Fibre Channel switching capability, including FCoE. If the FCoE switch is not in the fabric mode, then it could be in NPV mode.

   - In the case of a Cisco FCoE switch, you can verify if the NPV feature is enabled by logging in to the Nexus 5000 switch and then issuing the **show feature** NX-OS command or the **show npv flogi-database** NX-OS command.

- In the case of a Brocade FCoE Switch, the switchshow output will include the string "Access Gateway Mode" in the switchMode: field.

2. Use the NX-OS **show flogi database** or **show fcns database** command to verify that the host is able to log into the Nexus 5000 switch while the switch is in fabric mode. Use the NX-OS **show npv flogi** command to verify that the host is able to log into the Nexus 5000 switch while the switch in NPV mode.

   In order to verify whether the CNA has logged to the MP-8000B, complete the following steps.

   a. Log in to MP-8000B switch using a valid username and password.

   b. Issue the FOS **fcoe --loginshow** or **nsshow** command**.**

3. Ensure that the host's PWWN is in the output of either of the above commands.

Example and interpretation of the results

Below are example outputs of the **show feature** and **show npv flogi-database** commands. If NPV is enabled, the output of the **show feature** command will display that the feature is enabled, as shown in the following example. The output of the **show npv flogi-table** command below shows that the switch is performing NPV. Notice that the CNA has logged in on a server interface of a Nexus 5020 switch.

```
Nexus 5020 # show feature
Feature Name           Instance   State
---------------------  --------   --------
fcsp                   1          disabled
tacacs                 1          disabled
port-security          1          disabled
fabric-binding         1          disabled
port_track             1          disabled
npiv                   1          disabled
lacp                   1          disabled
npv                    1          enabled
interface-vlan         1          disabled
private-vlan           1          disabled
udld                   1          disabled
vpc                    1          disabled
cimserver              1          disabled
fcoe                   1          enabled
fex                    1          enabled
```

```
Nexus 5020 # show npv flogi-table
-----------------------------------------------------------------------------
SERVER                                                                 EXTERNAL
INTERFACE VSAN FCID             PORT NAME               NODE NAME          INTERFACE
-----------------------------------------------------------------------------
 vfc1     1    0x0e0008 21:00:00:c0:dd:10:28:bb 20:00:00:c0:dd:10:28:bb fc2/1

<output truncated>
```

In the following examples, the first NX-OS CLI **show flogi database command** output shows the host's CNA is logged in to the Nexus 5020, while the second CLI **show fcns database** command output shows the Name Server, which shows the Directory Name Server that displays all the devices logged in to the entire fabric.

```
Nexus 5020 # show flogi database
-----------------------------------------------------------------------------
INTERFACE        VSAN    FCID            PORT NAME               NODE NAME
-----------------------------------------------------------------------------
vfc2             1      0xad0000  21:00:00:c0:dd:10:28:bb 20:00:00:c0:dd:10:28:bb
                            [sgeliop54_cnaqlogic_p1]

Total number of flogi = 1.

Nexus 5020 # show fcns database

VSAN 1:
--------------------------------------------------------------------------
FCID        TYPE  PWWN                    (VENDOR)        FC4-TYPE:FEATURE
--------------------------------------------------------------------------
0x0b03ef    N     50:06:01:69:3b:60:03:c4 (Clariion)      scsi-fcp:both
                  [CX4_480_SPB_3_B1]
0xad0000    N     21:00:00:c0:dd:10:28:bb (Qlogic)        scsi-fcp:init
                  [sgeliop54_cnaqlogic_p1]

Total number of entries = 2
```

In the MP-8000B switch, the command to verify that the initiator has logged in successfully is the **fcoe --loginshow** or **nsshow** FOS command, as shown in the following examples.

```
MP-8000B:admin> fcoe --loginshow
================================================================================
Port   Te port        Device WWN              Device MAC         Session MAC
================================================================================
8      Te 0/0    21:00:00:c0:dd:10:28:bb  00:c0:dd:10:28:bb  0e:fc:00:01:08:01


MP-8000B:admin> nsshow
{
 Type Pid    COS     PortName                NodeName                 TTL(sec)
 N    010801;      3;21:00:00:c0:dd:10:28:bb;20:00:00:c0:dd:10:28:bb; na
    FC4s: FCP
    NodeSymb: [33] "QLE8142 FW:v5.01.03 DVR:v9.1.8.17"
    Fabric Port Name: 20:08:00:05:1e:d8:fd:80
    Permanent Port Name: 20:08:00:05:1e:d8:fd:80
    Port Index: 8
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No
The Local Name Server has 1 entry }
```

Next steps

- If the initiator's PWWN from the above step is not listed, then proceed with flowchart step #7.
- If the CNA is logged in to the Nexus 5020 switch and all the VNX series or CLARiiON LUNs are still not visible, then proceed with flowchart step # 4.
- For more information on troubleshooting commands on the Nexus Series and MP-8000B switches, refer to “Troubleshooting the Nexus Series switches” section in the "Nexus Series Switches Setup Examples" chapter and the “ED-DCX-B” section of the "EMC Connectrix B Setup Examples" chapter of the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Flowchart step #4, Storage array logged in to FC switch?

Troubleshooting

Note: Other than CLI, you can use switch management software to verify whether the target device has logged in to the FC switch. In Brocade switches you can use CMDCE, while in Cisco MDS switches you can use Fabric Manager.

In order to verify whether the Storage Array has logged to the Cisco MDS switch, complete the following steps.

1. From the Cisco command line, issue the CLI **show flogi database** or **show fcns database** command to verify that the Storage Array is able to log into the Cisco MDS switch or into fabric.
2. Ensure that the Storage Array’s PWWN is in the output of either of the above commands.

In order to verify whether the Storage Array has logged to the Brocade FC switch, complete the following steps.

1. From the Brocade command line, issue the FOS **nsshow** command.
2. Ensure that the Storage Array's PWWN is in the output of the above command.

Example and interpretation of the results

In the following example, **show flogi database** output from a Cisco MDS switch shows that the PWWN of the VNX series or CLARiiON Storage system (50:06:01:69:3b:60:03:c4) was able to do a Fabric Login.

```
MDS-Switch # show flogi database interface fc 2/12
--------------------------------------------------------------------------------
INTERFACE         VSAN    FCID              PORT NAME                 NODE NAME
--------------------------------------------------------------------------------
fc2/12            1     0x0b03ef  50:06:01:69:3b:60:03:c4 50:06:01:60:bb:60:03:c4
                            [CX4_480_SPB_3_B1]

Total number of flogi = 1.
```

The same is happening in the show fcns database output from a Cisco MDS switch that follows.

```
MDS-Switch # show fcns database

VSAN 1:
--------------------------------------------------------------------------
FCID        TYPE  PWWN                    (VENDOR)        FC4-TYPE:FEATURE
--------------------------------------------------------------------------
0x0b03ef    N     50:06:01:69:3b:60:03:c4 (Clariion)      scsi-fcp:both
                  [CX4_480_SPB_3_B1]
0xad0000    N     21:00:00:c0:dd:10:28:bb (Qlogic)        scsi-fcp:init
                  [sgeliop54_cnaqlogic_p1]

<output truncated>
```

The same is happening in the **show fcns database** output from a Cisco MDS switch that follows.

```
MDS-Switch # show fcns database

VSAN 1:
--------------------------------------------------------------------------
FCID        TYPE  PWWN                    (VENDOR)        FC4-TYPE:FEATURE
--------------------------------------------------------------------------
0x0b03ef    N     50:06:01:69:3b:60:03:c4 (Clariion)      scsi-fcp:both
                  [CX4_480_SPB_3_B1]
0xad0000    N     21:00:00:c0:dd:10:28:bb (Qlogic)        scsi-fcp:init
                  [sgeliop54_cnaqlogic_p1]

<output truncated>
```

In the following example, **nsshow** output from a Brocade FC switch shows that the PWWN of the VNX series or CLARiiON storage system (50:06:01:69:3b:60:03:c4) was able to do Fabric Login.

```
Brocade_5000B:admin> nsshow
{
Type Pid    COS     PortName                NodeName                 TTL(sec)

<output truncated>

N    101500;      3;50:06:01:69:3b:60:03:c4;50:06:01:60:bb:60:03:c4; na
    FC4s: FCP [DGC     LUNZ            0429]
    Fabric Port Name: 20:15:00:05:1e:90:51:e9
    Permanent Port Name: 50:06:01:69:3b:60:03:c4
    Port Index: 21
    Share Area: No
    Device Shared in Other AD: No
    Redirect: No

<output truncated>
```

Next steps

- If it is found that the target is not able to log in, then proceed with troubleshooting the FC storage array login, depicted in flowchart step #10.
- At this stage, if it is discovered that the target device is able to log in to the fabric and all the VNX series or CLARiiON LUNs are still not visible, then proceed with verifying that the host's PWWN is logged in to the VNX series or CLARiiON device, as depicted in flowchart step #5.

Flowchart step #5, CNA logged in to storage array?

Troubleshooting

At this stage, you will verify that the initiator's PWWN is logged in to the VNX series or CLARiiON device. Complete the following steps to check whether the host was able to perform log in to the VNX series or CLARiiON device.

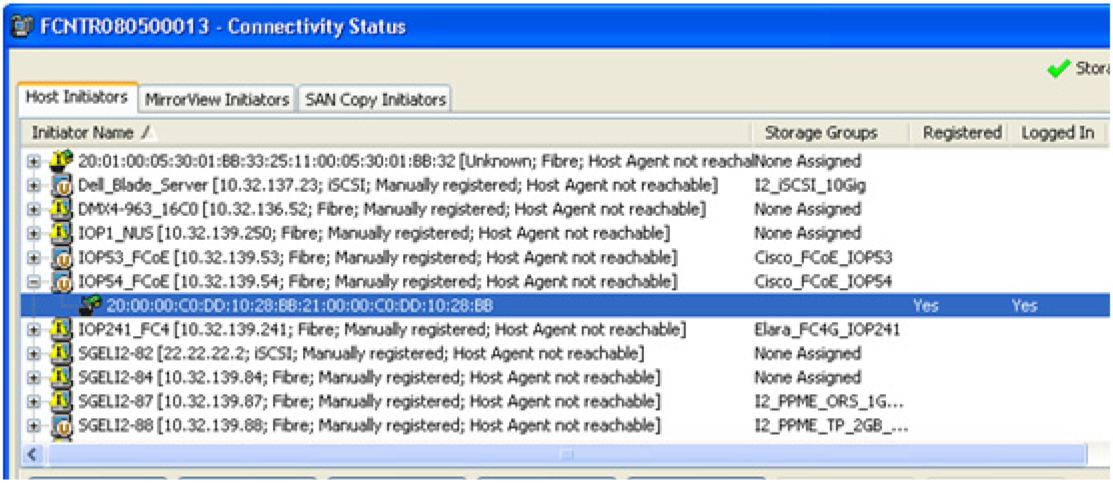
1. From Unisphere™/Navisphere® Manager, right-click the hostname of your VNX series or CLARiiON device.
2. Select **Connectivity Status.**
3. Ensure that the host was able to perform log in to the VNX series or CLARiiON device.

Note: Consult your Storage Array's technical documentation to verify host login if the target device is not a VNX series or CLARiiON device.

Example and interpretation of the results

As shown in Figure 87, the **Connectivity Status** window is accessed in VNX series or CLARiiON's Unisphere/Navisphere Manager. The host's WWN should appear in this table. If it is not, proceed to flowchart step #11.

If the initiator's PWWN is logged into the VNX series or CLARiiON and all the LUNs are still not visible, proceed to flowchart step #12. Figure 87 shows a successful host login to the VNX series or CLARiiON. The host's CNA address is WWN: 21:00:00:c0:dd:10:28:bb.



Initiator Name	Storage Groups	Registered	Logged In
20:01:00:05:30:01:88:33:25:11:00:05:30:01:88:32 [Unknown; Fibre; Host Agent not reachal	None Assigned		
Dell_Blade_Server [10.32.137.23; iSCSI; Manually registered; Host Agent not reachable]	I2_iSCSI_10Gig		
DMX4-963_16C0 [10.32.136.52; Fibre; Manually registered; Host Agent not reachable]	None Assigned		
IOP1_NUS [10.32.139.250; Fibre; Manually registered; Host Agent not reachable]	None Assigned		
IOP53_FCoE [10.32.139.53; Fibre; Manually registered; Host Agent not reachable]	Cisco_FCoE_IOP53		
IOP54_FCoE [10.32.139.54; Fibre; Manually registered; Host Agent not reachable]	Cisco_FCoE_IOP54		
20:00:00:C0:DD:10:28:BB:21:00:00:C0:DD:10:28:BB		Yes	Yes
IOP241_FC4 [10.32.139.241; Fibre; Manually registered; Host Agent not reachable]	Elara_FC4G_IOP241		
SGELI2-82 [22.22.22.2; iSCSI; Manually registered; Host Agent not reachable]	None Assigned		
SGELI2-84 [10.32.139.84; Fibre; Manually registered; Host Agent not reachable]	None Assigned		
SGELI2-87 [10.32.139.87; Fibre; Manually registered; Host Agent not reachable]	I2_PPME_ORS_1G...		
SGELI2-88 [10.32.139.88; Fibre; Manually registered; Host Agent not reachable]	I2_PPME_TP_2GB_...		

Figure 87 **Successful host login example**

If the host is not able to log in, ensure that there is no interswitch connectivity issue (TE connection) between the FCoE Nexus 5020 switch and the FC MDS switch.

Use the Cisco CLI command **show interface fc** <*slot/port*> to verify the connectivity and port status, as shown in an example in the following section. Ensure that the port is enabled and configured properly. Take notice the negotiated port type (TE in this case), port status, the port mode, port WWN, speed and even the allowed

VSANs on the TE port. You may want to do static configuration to bypass auto-negotiation and any dynamic processes to take place.

Note: In MP-8000B switch, the way to check E_Port status is through the FOS command **portshow** <*port* #>. An example is shown in the following section.

The following example shows how to check the interface status of a TE port in a Nexus 5000 switch. This command also applies to Cisco MDS switches. The highlighted information is the most useful and the one we usually need. Ensure that the negotiated port type is TE, port status is trunking, speed is detecting the correct speed and even the allowed VSANs. We may want to do static configuration to bypass auto-negotiation and any dynamic processes to take place.

```
Nexus 5020 # show int fc 2/1
fc2/1 is trunking
Hardware is Fibre Channel, SFP is short wave laser w/o OFC (SN)
Port WWN is 20:41:00:0d:ec:cf:98:80
Peer port WWN is 20:17:00:0d:ec:85:c9:00
Admin port mode is auto, trunk mode is on
snmp link state traps are enabled
Port mode is TE
Port vsan is 1
Speed is 4 Gbps
Transmit B2B Credit is 16
Receive B2B Credit is 16
Receive data field Size is 2112
Beacon is turned off
Trunk vsans (admin allowed and active) (1)
Trunk vsans (up) (1)
Trunk vsans (isolated) ()
Trunk vsans (initializing) ()
1 minute input rate 209488 bits/sec, 26186 bytes/sec, 51 frames/sec
1 minute output rate 3672584 bits/sec, 459073 bytes/sec, 240 frames/sec
220651299 frames input, 251694798448 bytes
0 discards, 0 errors
0 CRC, 0 unknown class
0 too long, 0 too short
497094335 frames output, 918842287368 bytes
0 discards, 0 errors
1 input OLS, 1 LRR, 0 NOS, 0 loop inits
1 output OLS, 1 LRR, 0 NOS, 0 loop inits
16 receive B2B credit remaining
16 transmit B2B credit remaining
0 low priority transmit B2B credit remaining
Interface last changed at Mon Dec 14 02:36:50 2009
```

In the MP-8000B switch, the way to check E_Port status is via FOS command **portshow** <*port* #>, as shown in the next example. The highlighted information is the most useful and the one you usually need. Ensure that the status is HEALTHY, the port state is Online, speed is detecting the correct speed, and port type is E_Port. You may want to do static configuration to bypass auto-negotiation and any dynamic processes to take place.

```
MP-8000B:admin> portshow 0
portName:
portHealth: HEALTHY
Authentication: None
portDisableReason: None
portCFlags: 0x1
portFlags: 0x10004903 PRESENT ACTIVE E_PORT G_PORT U_PORT LOGICAL_ONLINE LOGIN LED

portType: 17.0
POD Port: Port is licensed
portState: 1 Online
Protocol: FC
portPhys: 6 In_Sync portScn: 16 E_Port
port generation number: 6
state transition count: 3
portId: 010000
portIfId: 43020027
portWwn: 20:00:00:05:1e:76:7a:00
portWwn of device(s) connected:
Distance: normal
portSpeed: N4Gbps
LE domain: 0
FC Fastwrite: OFF
Interrupts: 0 Link_failure: 2 Frjt: 0
Unknown: 0 Loss_of_sync: 3 Fbsy: 0
Lli: 49 Loss_of_sig: 4
Proc_rqrd: 1900550 Protocol_err: 0
Timed_out: 0 Invalid_word: 2616763
Rx_flushed: 0 Invalid_crc: 0
Tx_unavail: 0 Delim_err: 0
Free_buffer: 0 Address_err: 0
Overrun: 0 Lr_in: 4
Suspended: 0 Lr_out: 3
Parity_err: 0 Ols_in: 1
2_parity_err: 0 Ols_out: 2
CMI_bus_err: 0
Port part of other ADs: No
```

Next steps

- At this stage, verify if the CNA was able to log in to the Storage Array. If the initiator's PWWN is logged into the VNX series or CLARiiON and all the LUNs are still not visible, proceed to flowchart step #12.
- If it is not logged in to the VNX series or CLARiiON, proceed to flowchart step #11.

Flowchart step #6, Perform physical link troubleshooting

Troubleshooting

This step is reached because you have verified that both sides are configured properly and have been tested and proven to interoperate with one another. If you are still seeing the VF/VN port as down, then it is likely that this is a hardware-related problem. To perform physical link/physical layer troubleshooting, follow the guidelines listed next:

- Verify each component of the link, the fiber, the media type used such as SFP+, the actual switchport, and the CNA card. You must ensure that each of these are working properly and with the correct specifications (for instance, a compatible SFP+ is used).
- Verify that you are using the correct fiber cabling type and that the cable is not exceeding the distance limitations.
- Always check whether or not the fiber is bad or if there is a unidirectional link. It is highly recommended that you have a spare working fiber cable should you need it.
- Ensure that the ports are working. Swapping ports can be done, if needed. Use a different CEE port on the switch, or a different CNA card if changing the cable does not solve the issue. If the issue is still unresolved, then try using a CNA from a known working FCoE host, installing it on the host in question. If this still does not solve the problem, then you know that the issue is on the host itself. At this time, you need to contact the host or server vendor.

Note: After troubleshooting the most obvious physical problems, it is important to verify that there are no MAC layer issues. Refer to "MAC layer" on page 197 for more information.

Example and interpretation of the results

The following are some examples of **show tech-support** options in the Nexus 5000 switch. Also shown is the example command outputs of **show tech-support fcoe** and **debug flogi all.**

```
Nexus 5020 # show tech-support ?
  <CR>
  >                Redirect it to a file
  >>               Redirect it to a file in append mode
  aaa              Display aaa information
  aclmgr           ACL commands
  adjmgr           Display Adjmgr information
  arp              Display ARP information
  ascii-cfg        Show ascii-cfg information for technical support personnel
  bootvar          Gather detailed information for bootvar troubleshooting
  brief            Display the switch summary
  callhome         Callhome troubleshooting information
  cdp              Gather information for CDP trouble shooting
  cfs              Gather detailed information for cfs troubleshooting
  cli              Gather information for parser troubleshooting
  clis             Gather information for CLI Server troubleshooting
  commands         Show commands executed as part of show tech-support commands
  details          Gather detailed information for troubleshooting
  device-alias     Show device-alias technical support information
  dhcp             Gather detailed information for dhcp troubleshooting
  ethport          Gather detailed information for ETHPORT troubleshooting
  fc               Get fibre channel related information
  fcdomain         Gather detailed information for fcdomain troubleshooting
  fcns             Show information for fcns technical support
  fcoe             Gather information for FCOE mgr trouble shooting
  fex              Gather detailed information for fex troubleshooting
  flogi            Gather detailed information for flogi troubleshooting
  fspf             Show information for fspf technical support
  ha               Gather detailed information for HA troubleshooting

<output truncated>
```

```
Nexus 5020 # show tech-support fcoe
******************** FCOE MGR tech-support start *****************
`show platform software fcoe_mgr event-history errors`
1) Event:E_DEBUG, length:86, at 300083 usecs after Tue Apr 20 07:39:28 2010
    [102] fcoe_mgr_vfc_ac_eval(2346): Bringing down VFC 1e000020 due to truly
       missing fka

2) Event:E_DEBUG, length:80, at 299787 usecs after Tue Apr 20 07:39:28 2010
    [102] fcoe_mgr_vfc_fka_expiry(3653): FKA Timer expired for VFC if_index
       1e000020

3) Event:E_DEBUG, length:86, at 200376 usecs after Tue Apr 20 07:35:59 2010
```

```
    [102] fcoe_mgr_vfc_ac_eval(2346): Bringing down VFC 1e000020 due to truly
       missing fka

4) Event:E_DEBUG, length:80, at 200103 usecs after Tue Apr 20 07:35:59 2010
    [102] fcoe_mgr_vfc_fka_expiry(3653): FKA Timer expired for VFC if_index
       1e000020

5) Event:E_DEBUG, length:86, at 750055 usecs after Tue Apr 20 02:33:40 2010
    [102] fcoe_mgr_vfc_ac_eval(2346): Bringing down VFC 1e000015 due to truly
       missing fka
<output truncated>
```

```
Nexus 5020 # debug flogi all
Nexus 5020 # 2010 Apr 21 03:17:46.464842 flogi: fs_demux: msg consumed by
   sdwrap_process msg
2010 Apr 21 03:17:46.465250 flogi: fu_fsm_execute_all: match_msg_id(0),
   log_already_open(0)
2010 Apr 21 03:17:46.465529 flogi: fu_fsm_execute_all: null fsm_event_list
2010 Apr 21 03:17:46.465805 flogi: fu_fsm_engine_post_event_processing
2010 Apr 21 03:17:46.466124 flogi: fu_mts_drop: ref 0x8192638 opc 182 payload
   0xb5725860
2010 Apr 21 03:17:46.466409 flogi: fu_fsm_engine_post_event_processing: mts msg
   MTS_OPC_DEBUG_WRAP_MSG(msg_id 1956529170) dropped
2010 Apr 21 03:17:46.466687 flogi:  end of while in fu_fsm_engine
2010 Apr 21 03:17:46.466966 flogi: begin fu_fsm_engine: line[2311]

<output truncated>
```

Next steps

During advance troubleshooting, there are some instances where the vendor requires network traces from the switch or fabric. The following are tools you can use for fabric service troubleshooting and for capturing traces:

- 10 G CEE Network Analyzer
- RMON (Remote Network Monitoring)
- SPAN (Switched Port Analyzer)
- show tech-support/ supportshow output from the switch
- Debugs from the switch
  - For the Nexus 5000: debug flogi all (NX-OS)
  - For the MP-8000B: debug -portlog (CMSH command) or portlogdump (FOS command)

Once the issue is resolved, then proceed to flowchart step #15.

Flowchart step #7, Is the VF/VN port up?

Troubleshooting

At this point, you need to verify that the virtual F_Port is up. It is important to verify that the CNA port (virtual N_Port) is also up. Follow the guidelines below in verifying status of the VF/VN ports.

- Use the CNA management suite to verify that the VN port is up. Since the CEE port is also detected by the host as a normal NIC card, it can be checked by using OS integrated tools, such as Windows' **Network Connections**, located in the **Control Panel**.
- The NX-OS CLI **show interface brief** command can be used to verify whether the link is up. You need to check not only the virtual F_Port, but also the physical CEE port to which the VF port is bound. Another command, **show interface fcoe**, can also provide information about the virtual F_Port, including the FCID, session MAC, and the PWWN of the connected host.
- In the Brocade MP-8000B switch, virtual F_Port status can be verified by using the FOS **portshow** command. You can also use the CMSH command **show ip interface brief** to verify the physical interface status. See the example in the following section.
- Check the CEE port for any layer 1 errors. If the CEE port is up but it is not logging into the fabric, there may be a dirty link, which could be caused by hardware issues. Refer to "Troubleshooting basic FCoE and CEE problems" on page 180 for more information.

Example and interpretation of the results

The following is an example output of the **show interface brief** command. Notice the CEE ports eth1/2 and vfc2 are up.

```
Nexus 5020 # show interface brief

<output truncated>

-----------------------------------------------------------------------------
Ethernet      VLAN   Type Mode   Status  Reason                   Speed     Port
Interface                                                                   Ch #
-----------------------------------------------------------------------------
Eth1/1        1      eth  access down    SFP not inserted            10G(D) --
Eth1/2        1      eth  trunk  up      none                        10G(D) --
Eth1/3        1      eth  access down    SFP not inserted            10G(D) --
Eth1/4        1      eth  access down    SFP not inserted            10G(D) --

<output truncated>
```

```
------------------------------------------------------------------------------
Interface  Vsan   Admin  Admin   Status          SFP    Oper  Oper   Port
                  Mode   Trunk                          Mode  Speed  Channel
                         Mode                                 (Gbps)
------------------------------------------------------------------------------
vfc2       1      F      on      up               --     F       auto -

<output truncated>
```

The following are example outputs from the MP-8000B switch. Notice that te0/0 is up (from CMSH command **show ip interface brief**) and the virtual F Port (port 8) is also up (as shown in FOS command **portshow** output). In FOS, notice that the virtual F Ports bound to the TenGigabitEthernet ports are called port8 to port31 (equivalent to TenGigabitEthernet port0/0 to TenGigabitEthernet port0/23). This is because the ports 0 to 7 are reserved to native FC ports.

```
MP-8000B # show ip interface brief
Interface                 IP-Address      Status                Protocol
=========                 ==========      ======                ========
TenGigabitEthernet 0/0    unassigned      up                     up
TenGigabitEthernet 0/1    unassigned      up                     down
TenGigabitEthernet 0/2    unassigned      up                     down
TenGigabitEthernet 0/3    unassigned      up                     down
TenGigabitEthernet 0/4    unassigned      up                     down

<output truncated>

MP-8000B:admin> portshow 8
portName:
portHealth: No Fabric Watch License

Authentication: None
portDisableReason: None
portCFlags: 0x1
portFlags: 0x24b03       PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE
   LOGIN NOELP LED ACCEPT FLOGI
portType:  17.0
POD Port: Port is licensed
portState: 1    Online
Protocol: FCoE
portPhys:  6    In_Sync         portScn:   32   F_Port
port generation number:    28
state transition count:    3

portId:    020800
portIfId:    43020028
portWwn:   20:08:00:05:1e:d8:ff:00
portWwn of device(s) connected:
        21:00:00:c0:dd:10:29:71
        20:08:00:05:1e:d8:ff:00
Distance:  normal
portSpeed: 10Gbps
```

Next steps

If the logical VF port is showing up as a port state, proceed to flowchart step #15. If it is down, then proceed to flowchart step #8 to verify the correctness of the configuration.

Flowchart step #8, Is the configuration correct?

Note: For more information, you can use the configuration guidelines discussed inthe “Nexus 5000 direct-connect topology” and “MP-8000B direct-connect topology” sections in the "Nexus Series Switches Setup Examples" chapter of the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**.

Troubleshooting

In this step, it is important to refer to the appropriate configuration guides. Verify the following by examining the running configuration of the FCoE switch and examining the configuration or settings of your CNA card.

Note: In both Nexus 5000 and MP-8000B switches the command to view running configuration is **show running-config**. In some cases, you need to examine the negotiated interface parameters using the **show interface** command.

- Verify that the virtual F_Port is bound to the ENode MAC of the CNA.
- Ensure that the VF_Port is configured with the correct port type, in this case, an F_Port.
- Verify that the CEE switchport speed is 10 Gb/s (the default). If not, reset this to 10 Gb/s.
- Make sure that the VLAN that this port is a member of is an FCF-capable VLAN.
- The CNA port must be configured for point-to-point and 10 Gb/s connection settings.
- Enable the port. This is an important, and often overlooked, step. Ensure that both ends (physical CEE and logical F_Ports and N_Ports) are enabled.

- If an ACL is configured on the switch, make sure that this ACL is not blocking any FCoE frames or blocking the source MAC itself.

Example and interpretation of the results

The following is an example of **show running-config** output from the Nexus 5020 switch. Notice that port is enabled and is assigned to an FCF-capable VLAN. It also shows that port doesn't have any ACL configuration. You will also see the **show interface** <*intf type*> output that shows the negotiated port type and port speed.

```
Nexus 5020 # show running-config

<output truncated>

vlan 200           -> this line and the next line show that VLAN200 is an
                        FCF-capable VLAN
  fcoe vsan 1

<output truncated>

interface vfc1
  bind interface Ethernet1/1 ->this line shows the binding of VF port vfc1 to
                                  physical port ethernet 1/1
  switchport description mapped_to_eth_1/1
  no shutdown  -> this line shows the port is enabled

<output truncated>

interface Ethernet1/1
  description test
  switchport mode trunk
  switchport trunk allowed vlan 1,200 -> this line shows the port assigned to
                                            FCF-capable VLAN200
  spanning-tree port type edge trunk
  spanning-tree bpduguard enable

<output truncated>

Nexus 5020 # show interface e1/16
Ethernet1/16 is up
  Hardware: 1000/10000 Ethernet, address: 000d.eccf.9897 (bia 000d.eccf.9897)
  Description: free
  MTU 1500 bytes, BW 10000000 Kbit, DLY 10 usec,
     reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA
  Port mode is trunk
  full-duplex, 10 Gb/s, media type is 1/10g -> this line shows the port has
                                                 negotiated 10G b/s of bandwidth
  Beacon is turned off
  Input flow-control is off, output flow-control is off
  Rate mode is dedicated
  Switchport monitor is off
  Last link flapped 2week(s) 4day(s)
  Last clearing of "show interface" counters never
  1 minute input rate 28125160 bits/sec, 4074 packets/sec
  1 minute output rate 47951184 bits/sec, 5792 packets/sec
  Rx
    3007391738 input packets 581544461 unicast packets 2425809597 multicast
         packets
    37680 broadcast packets 381701209 jumbo packets 0 storm suppression packets
    995874790733 bytes
  Tx
    571911480 output packets 19941993 multicast packets
    5344221 broadcast packets 257678461 jumbo packets
    590755838106 bytes
    0 input error 0 short frame 0 watchdog
    0 no buffer 0 runt 0 CRC 0 ecc
    0 overrun  0 underrun 0 ignored 0 bad etype drop
    0 bad proto drop 0 if down drop 0 input with dribble
    0 input discard
    0 output error 0 collision 0 deferred
    0 late collision 0 lost carrier 0 no carrier
    0 babble
    2425552689 Rx pause 1664 Tx pause
  28 interface resets

Nexus 5020 # show interface vfc 16
vfc16 is up
    Bound interface is Ethernet1/16
    FCF priority is 128
    Hardware is Virtual Fibre Channel
    Port WWN is 20:0f:00:0d:ec:cf:98:bf
    Admin port mode is F, trunk mode is on
    snmp link state traps are enabled
    Port mode is F, FCID is 0x220001   -> this line shows the port mode is F
    Port vsan is 1
    1 minute input rate 22652944 bits/sec, 2831618 bytes/sec, 26744 frames/sec
    1 minute output rate 48578192 bits/sec, 6072274 bytes/sec, 45728 frames/sec
      569292776 frames input, 838904749544 bytes
        0 discards, 0 errors
      535743448 frames output, 596842115540 bytes
        0 discards, 0 errors
    Interface last changed at Thu Apr  1 11:30:04 2010
```

In Nexus 5000, you can also use filters when examining the running configurations. The following example shows **show running-config interface** <*intf type*> outputs from the Nexus 5020 switch. This command will help you save time when looking for specific port configuration. You can also use **show vlan fcoe** to verify if the VLAN is FCF-capable.

```
Nexus 5020 # show running-config interface ethernet 1/1
version 4.1(3)N1(1)

interface Ethernet1/1
  description test
  switchport mode trunk
  switchport trunk allowed vlan 1,200
  spanning-tree port type edge trunk
  spanning-tree bpduguard enable

Nexus 5020 # show running-config interface vfc 1
version 4.1(3)N1(1)
interface vfc1
  bind interface Ethernet1/1
  switchport description mapped_to_eth_1/1
  no shutdown   -> this line shows the port is enabled


Nexus 5020 # show vlan fcoe
VLAN      VSAN      Status
--------  --------  --------
200       1         Operational -> this line and the next line show VLAN200 is
                                     an FCF-capable VLAN
```

The following is an example **show running-config** output from the MP-8000B switch. Notice that te0/0 is enabled and is configured with VLAN 1002, which is an FCF-capable VLAN. It also shows that port doesn't have any ACL configuration. You will also see the **show interface** <*intf type*> CMSH output and **portcfg** <*port* #> FOS output show the negotiated port speed and port mode, respectively.

```
MP-8000B #show running-config
!

<output truncated>

!
interface Vlan 1002
 fcf forward             -> this line shows VLAN1002 is an FCF-capable VLAN
!
interface TenGigabitEthernet 0/0
 switchport
 switchport mode converged
 vlan classifier activate group 1 vlan 1002
 no shutdown             -> this line shows the port is enabled
 spanning-tree edgeport
 spanning-tree guard root
 cee default             -> this is very important and sometimes overlooked
<output truncated>

MP-8000B #show interface tengigabitethernet 0/0
```

```
TenGigabitEthernet 0/0 is up, line protocol is up (connected)
Hardware is Ethernet, address is 0005.1ed8.ff24
    Current address is 0005.1ed8.ff24
Pluggable media present, Media type is sfp
    Wavelength is 850 nm
Interface index (ifindex) is 402653184
MTU 2500 bytes
LineSpeed: 10000 Mbit, Duplex: Full  -> this line shows the port has negotiated
                                         10 Gb/s of bandwidth
Flowcontrol rx: on, tx: on
Last clearing of show interface counters: 2w4d21h
Queueing strategy: fifo
Receive Statistics:
    1678842414 packets, 3316907939823 bytes
    Unicasts: 1678788904, Multicasts: 53494, Broadcasts: 16
    64-byte pkts: 63, Over 64-byte pkts: 101253124, Over 127-byte pkts: 269
    Over 255-byte pkts: 12, Over 511-byte pkts: 14341106, Over 1023-byte pkts:
   1977900
    Over 1518-byte pkts(Jumbo): 1561269940
    Runts: 0, Jabbers: 0, CRC: 0, Overruns: 0
    Errors: 0, Discards: 0
Transmit Statistics:
    307925623 packets, 30775762262 bytes
    Unicasts: 0, Multicasts: 1422814, Broadcasts: 2693327
    Underruns: 0
    Errors: 92, Discards: 0
Rate info (interval 299 seconds):
    Input 2.598144 Mbits/sec, 194 packets/sec, 0.03% of line-rate
    Output 0.140206 Mbits/sec, 86 packets/sec, 0.00% of line-rate
Time since last interface status change: 1w5d22h

MP-8000B:admin> portshow 8
portName:
portHealth: No Fabric Watch License

Authentication: None
portDisableReason: None
portCFlags: 0x1
portFlags: 0x24b03       PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE
   LOGIN NOELP LED ACCEPT FLOGI
portType:  17.0
POD Port: Port is licensed
portState: 1    Online
Protocol: FCoE
portPhys:  6    In_Sync         portScn:   32   F_Port  -> this line shows the
                                                            port mode is F
port generation number:    28
state transition count:    3

portId:    020800
portIfId:    43020028
portWwn:   20:08:00:05:1e:d8:ff:00
portWwn of device(s) connected:
        21:00:00:c0:dd:10:29:71
        20:08:00:05:1e:d8:ff:00
Distance:  normal
portSpeed: 10Gbps
```

Next steps

- If the configuration is correct and you are still not able to see ALL the LUNs, then proceed to flowchart step #6.
- If configuration is wrong, then proceed to flowchart step #9 to fix the configuration.

Flowchart step #9, Fix the configuration

Note: Use the configuration guidelines discussed in the “Nexus 5000 direct-connect topology”section in the "Nexus Series Switches Setup Examples" chapter and the “MP-8000B direct-connect topology” section in the "EMC Connectrix B Setup Examples" chapter of the *Fibre Channel over Ethernet (FCoE) Data Center Bridging (DCB) Case Studies TechBook* at http://elabnavigator.EMC.com, **Documents> Topology Resource Center**. as a reference. Once the configuration has been fixed, proceed to flowchart step #15. Troubleshooting

Most of the configuration issues are related to issues such as:

- A CEE interface that is not configured properly
- A virtual F_Port bound to the wrong CEE port
- A virtual F_Port has not yet been created
- The FCoE port may still not be part of the FCF-capable VLAN
- The Fibre Channel fabric is still forming

To check whether any of these issues exist, see the following examples from both the Nexus Series and MP-8000B switches.

Example and interpretation of the results

The Nexus 5000 command examples check that the virtual F_Port is bound to the correct physical port. It also shows that the FCoE port is assigned to an FCF-capable VLAN.

```
Nexus 5020 # show running-config interface ethernet 1/1
version 4.1(3)N1(1)

interface Ethernet1/1
  description test
  switchport mode trunk
  switchport trunk allowed vlan 1,200 -> this line shows the port assigned to
                                         FCF-capable VLAN200
  spanning-tree port type edge trunk
  spanning-tree bpduguard enable


Nexus 5020 # show running-config interface vfc 1
version 4.1(3)N1(1)

interface vfc1
  bind interface Ethernet1/1 1 ->this line shows the binding of VF port vfc1 to
                                  physical port ethernet 1/1
  switchport description mapped_to_eth_1/1
  no shutdown

Nexus 5020 # show vlan fcoe
VLAN      VSAN      Status
--------  --------  --------
200       1         Operational -> this line and the next line show VLAN200 is
                                   an FCF-capable VLAN
```

In the MP-8000B switch, the tengigabitethernet port is mapped automatically to the next FC port assignment, which is port 8. Therefore, port tengigabitethernet 0/0 is mapped to port 8, tengigabitethernet 0/1 is mapped to port 9, and so on, as shown next.

```
MP-8000B:admin> fcoe --cfgshow

User Port  Status       Port WWN              DeviceCount  Port  Type         MAC          VF_ID
=======================================================++++++++++++++++++++++++++++++++++=================
8          ENABLED   20:08:00:05:1e:76:a0:00      1       FCoE  VF-Port  00:05:1e:76:a0:00  128
9          ENABLED   20:09:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:01  128
10         ENABLED   20:0a:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:02  128
11         ENABLED   20:0b:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:03  128
12         ENABLED   20:0c:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:04  128
13         ENABLED   20:0d:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:05  128
14         ENABLED   20:0e:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:06  128
15         ENABLED   20:0f:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:07  128
16         ENABLED   20:10:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:08  128
17         ENABLED   20:11:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:09  128
18         ENABLED   20:12:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0a  128
19         ENABLED   20:13:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0b  128
20         ENABLED   20:14:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0c  128
21         ENABLED   20:15:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0d  128
22         ENABLED   20:16:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0e  128
23         ENABLED   20:17:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:0f  128
24         ENABLED   20:18:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:10  128
25         ENABLED   20:19:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:11  128
26         ENABLED   20:1a:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:12  128
27         ENABLED   20:1b:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:13  128
28         ENABLED   20:1c:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:14  128
29         ENABLED   20:1d:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:15  128
30         ENABLED   20:1e:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:16  128
31         ENABLED   20:1f:00:05:1e:76:a0:00      0       FCoE  VF-Port  00:05:1e:76:a0:17  128
```

The following are examples of commands to check in the Nexus 5020 and MP-8000B to see if the fabric is still forming.

```
Nexus 5020 # show topology

FC Topology for VSAN 1 :
--------------------------------------------------------------------------------
Interface  Peer Domain Peer Interface     Peer IP Address
--------------------------------------------------------------------------------
fc2/1  0x0b(11)            fc4/3  10.32.139.11 <-- sign the fabric is formed
                                                   already
FC Topology for VSAN 667 :
--------------------------------------------------------------------------------
Interface  Peer Domain Peer Interface     Peer IP Address
--------------------------------------------------------------------------------
fc2/1 0x88(136)            fc4/3  10.32.139.11 <-- sign the fabric is formed
                                                   already

MP-8000B :admin> switchshow

switchName:     MP-8000B
switchType:     76.7
switchState:    Online
switchMode:     Native
switchRole:     Subordinate
switchDomain:   1 (unconfirmed)  <-- sign the fabric is still forming
switchId:       fffc01
switchWwn:      10:00:00:05:1e:76:7a:00
zoning:         OFF
switchBeacon:   OFF

Area Port Media Speed State     Proto
==============================================
0    0    id     N4   Online      FC   E-Port  10:00:00:05:1e:90:18:6f "DS_5000B_13"
1    1    --     N8   No_Module   FC
2    2    --     N8   No_Module   FC
3    3    --     N8   No_Module   FC
4    4    id     N4   Online      FC   E-Port  10:00:00:05:1e:90:18:5d "DS_5000B_14"
                                           (upstream)
5    5    --     N8   No_Module   FC
6    6    --     N8   No_Module   FC
7    7    --     N8   No_Module   FC
8    8    --     10G  In_Sync    FCoE Disabled (switch not ready for F or L ports)
9    9    --     10G  In_Sync    FCoE Disabled (switch not ready for F or L ports)
10   10   --     10G  In_Sync    FCoE Disabled (switch not ready for F or L ports)
11   11   --     10G  In_Sync    FCoE Disabled (switch not ready for F or L ports)
```

The following are examples of commands to check in the Nexus 5020 and MP-8000B switches to see if an FCoE port is an FCF-capable VLAN:

```
Nexus 5020 # show vlan fcoe
```

```
VLAN      VSAN      Status
--------  --------  --------
200       1         Operational

MP-8000B # show vlan fcoe

VLAN     Name       State    Ports
                                 (u)-Untagged, (t)-Tagged
                                 (c)-Converged
====  =======       =====    ===+++++++++++++===============
1002    VLAN1002    ACTIVE   Te 0/0(c)
```

Next step

Once the configuration has been fixed, proceed to flowchart step #15.

Flowchart step #10, Troubleshoot FC storage array login

Troubleshooting

The following are some guidelines on troubleshooting FC Storage Array login. Additionally, different troubleshooting techniques can be used in this step. Refer to "FC layers" on page 187 for more information.

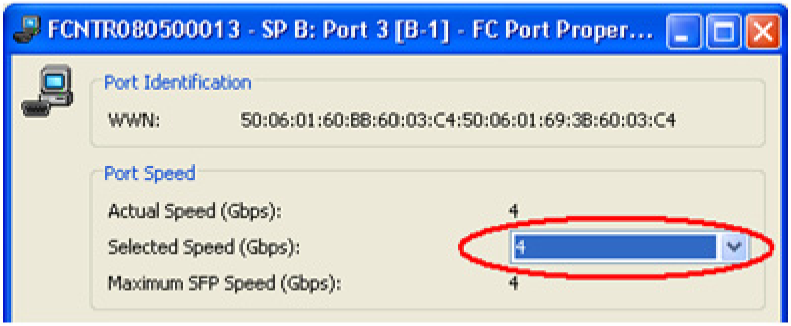
- Check when a storage array is not logging into the switch is the physical aspect, which is the FC-0 and FC-1 layer. This involves checking the status of cabling, SFP, type of cable used, or the node port (see if the switch FC port/VNX series or CLARiiON port is broken).
- The FC-2 layer could be the problem if there is a driver issue that could affect the Fibre Channel transport, especially on Informational Unit delivery or the operation between two nodes (in this case the VNX series or CLARiiON front-end N_Port to the MDS F_Port). For example, how are the exchanges, sequence, frames, encoding/decoding, and link-level protocols managed?
- It is very important to check the port configuration. Ensure that there is no mismatch of the port type and speed at both ends. If, in this example, the MDS port fc2/12 is set to E_Port, then this port would be unable to accept incoming FLOGI and PLOGI frames from the VNX series or CLARiiON, thus preventing the VNX series or CLARiiON to successfully log in. An E_Port will only accept ELPs (Exchange Link Parameters) from another E_Port.

- Excessive errors are usually related to physical issues, which can be resolved by replacing one of the components across the link, such as the cable, SFP, port, or even the switch itself, should the problem be the ASIC or controller.

Example and interpretation of the results

The following examples show how to set a 4 GB speed on both the VNX series or CLARiiON and MDS FC switch. Errors must be checked, such as CRCs, discards, and errors (see the following MDS switch example). In the MP-8000B or other Brocade switches, the **portstatsshow** <*port#*> and **portshow** <*port#*> FOS commands can be used to view the interface errors. Excessive errors are usually related to physical issues, which can be resolved by replacing one of the components across the link, such as the cable, SFP, port, or even the switch itself, should the problem be the ASIC or controller.

In Unisphere/Navisphere Manager, the port speed can be changed under the **Physical** menu by right-clicking the properties of the front-end port, as shown in Figure 88.



FCNTR080500013 - SP B: Port 3 [B-1] - FC Port Proper...

Port Identification

WWN: 50:06:01:60:BB:60:03:C4:50:06:01:69:3B:60:03:C4

Port Speed

Actual Speed (Gbps): 4

Selected Speed (Gbps): 4

Maximum SFP Speed (Gbps): 4

Figure 88 Changing the port speed in Unisphere/Navisphere Manager

The following are the different speed options in MDS switch using the **switchport speed ?** command:

```
MDS-Switch (config-if)# switchport speed ?
  1000 (no abbrev)  1000 Mbps speed
  2000 (no abbrev)  2000 Mbps speed
  4000 (no abbrev)  4000 Mbps speed
  8000 (no abbrev)  8000 Mbps speed
  auto              Auto negotiate speed
MDS-Switch (config-if)# switchport speed 4000
```

The following shows the FC port error on the switch seen using the **show interface** <*intf type*> command:

```
MDS-Switch# show interface fc2/12
fc2/12 is up
    Hardware is Fibre Channel, SFP is short wave laser w/o OFC (SN)
    Port WWN is 20:4c:00:05:30:01:bb:32
    Admin port mode is F, trunk mode is off
    snmp link state traps are enabled
    Port mode is F, FCID is 0x0b03ef
    Port vsan is 1
    Speed is 4 Gbps
    Rate mode is dedicated
    Transmit B2B Credit is 8
    Receive B2B Credit is 16
    Receive data field Size is 2112
    Beacon is turned off
    5 minutes input rate 0 bits/sec, 0 bytes/sec, 0 frames/sec
    5 minutes output rate 0 bits/sec, 0 bytes/sec, 0 frames/sec
      35479144 frames input, 58294732708 bytes
        0 discards, 0 errors
        40623 CRC,  0 unknown class
        0 too long, 0 too short
      25915665 frames output, 43385367848 bytes
        0 discards, 0 errors
      1 input OLS, 1 LRR, 0 NOS, 0 loop inits
      2 output OLS, 0 LRR, 0 NOS, 0 loop inits
      16 receive B2B credit remaining
      8 transmit B2B credit remaining
      8 low priority transmit B2B credit remaining
    Interface last changed at Thu Sep 24 02:05:56 2009
```

Next step

Once the Storage Array login has been resolved, proceed to flowchart step #4.

Flowchart step #11, Check zoning

Troubleshooting

Note: Other than CLI, you can use switch management software to verify the zoning configuration. In Brocade switches you can use CMDCE, while in Cisco MDS and Nexus 5000 switches you can use Fabric Manager.

In order to verify the zoning configurations in Nexus 5000 or Cisco MDS switches, complete the following steps.

1. From the Cisco command line, issue the CLI **show zoneset active** or **show zone active** command to verify that the zoning configuration is correct. You can also view the zoning configuration using the **show running-config** command.

2. Ensure that you are zoning the right initiator and target. When configuring the zoning, avoid manually typing the PWWNs to avoid typographical errors. Instead, copy and paste the desired PWWN that you want to zone.

In order to verify the zoning configurations in MP-8000B or other Brocade switches, complete the following steps.

1. Log in to the MP-8000B or Brocade FC switch and issue the FOS **zoneshow** command to verify that the zoning configuration is correct. You can also view the zoning configuration using the **cfgshow** command.

2. Ensure that you are zoning the right initiator and target. When configuring the zoning, avoid manually typing the PWWNs to avoid typographical errors. Instead, copy and paste the desired PWWN that you want to zone.

Example and interpretation of the results

In the following examples, you can verify that the initiator and target are properly zoned together. To confirm this, you can view the configuration using the **show running-config** command, or you can use the **show zoneset active** or **show zone active** command. The advantage of using the former is that it also shows if the devices are logged into the FLOGI database/Name Server.

```
Nexus 5020 # show zoneset active
zoneset name zoneset1 vsan 1
  zone name sgeliop54_cnaqlogic_p1_CX4_480_SPB_3_B1 vsan 1
  * fcid 0x0b03ef [PWWN 50:06:01:69:3b:60:03:c4] [CX4_480_SPB_3_B1]
  * fcid 0xad0000 [PWWN 21:00:00:c0:dd:10:28:bb] [sgeliop54_cnaqlogic_p1]
```

```
Nexus 5020 # show running-config zone vsan 1
version 4.1(3)N1(1a)
!Full Zone Database Section for vsan 1
zone name sgeliop54_cnaqlogic_p1_CX4_480_SPB_3_B1 vsan 1
    member PWWN 50:06:01:69:3b:60:03:c4
    member PWWN 21:00:00:c0:dd:10:28:bb

zoneset name zoneset1 vsan 1
    member sgeliop54_cnaqlogic_p1_CX4_480_SPB_3_B1

zoneset activate name zoneset1 vsan 1
```

In the MP-8000B, the **zoneshow** or **cfgshow** FOS commands can be used to verify the zoning configuration, and the **nsshow** or **switchshow** FOS commands can be used to verify whether the initiator and target that are zoned are logged in, as shown in the following examples.

```
MP-8000B:admin> zoneshow
Defined configuration:
 cfg:   my_zone Hostp1_SPBPort3B1
 zone:  Hostp1_SPBPort3B1
                21:00:00:c0:dd:10:28:bb; 50:06:01:60:bb:60:03:c4

Effective configuration:
 cfg:   my_zone
 zone:  Hostp1_SPBPort3B1
                21:00:00:c0:dd:10:28:bb
                50:06:01:60:bb:60:03:c4

MP-8000B:admin> cfgshow
Defined configuration:
 cfg:   my_zone Hostp1_SPBPort3B1
 zone:  Hostp1_SPBPort3B1
                21:00:00:c0:dd:10:28:bb; 50:06:01:60:bb:60:03:c4

Effective configuration:
 cfg:   my_zone
 zone:  Hostp1_SPBPort3B1
                21:00:00:c0:dd:10:28:bb
                50:06:01:60:bb:60:03:c4
```

Next step

Once the zoning has been verified correct, proceed to flowchart step #5.

Flowchart step #12, Check storage array configuration

Troubleshooting

To check the storage array configuration, complete the following steps:

1. From the host, if the LUNs are not all shown, check the LUN masking configuration in the VNX series or CLARiiON device.

2. Ensure that all the LUNs are assigned to the storage group created. In VNX series or CLARiiON device, LUN masking can be configured and verified in the **Storage Group Properties** window, **LUNs** tab, as shown in Figure 89 on page 243.

3. If the host is not able to see ALL the LUNs and you have gone through the previous steps, then verify the host assignment on the Storage Array. Ensure that the correct host was assigned to the LUNs selected.

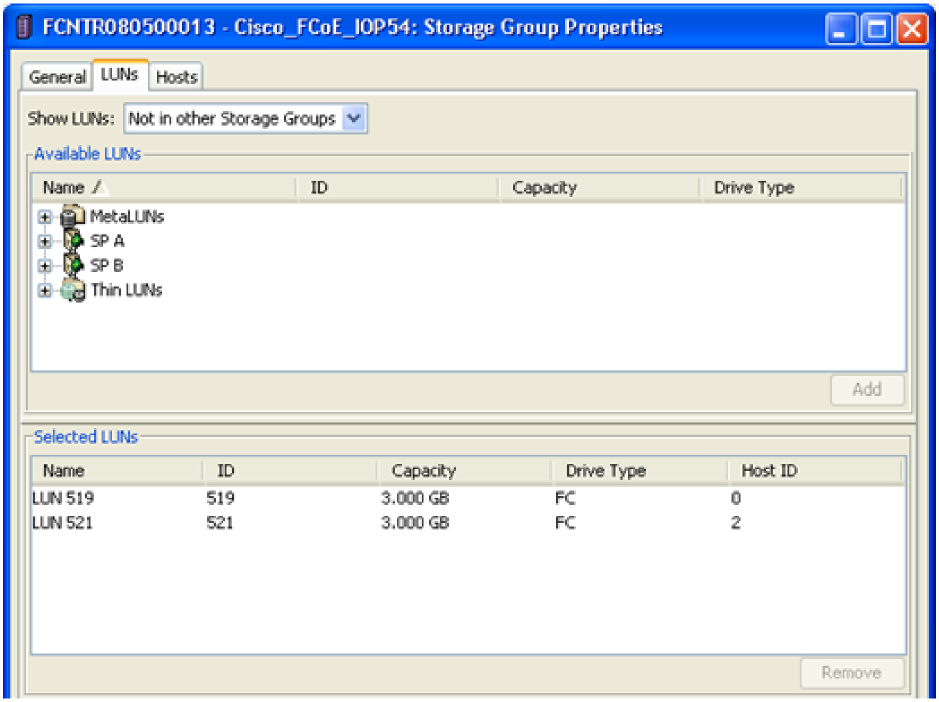
4. In a VNX series or CLARiiON device, host assignment can be configured and verified in the **Storage Group Properties** window, **Hosts** tab, as shown in the example in Figure 90 on page 244.

Example and interpretation of the results

You have verified that the host is able to see some of the LUNs that were configured in the VNX series or CLARiiON device. In the following example, one LUN is missing in the **inq** command output.

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.
.....
------------------------------------------------------------------------
DEVICE              :VEND     :PROD            :REV   :SER NUM    :CAP(kb)
------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA      :WDC WD1602ABKS-1:3B04  :     WD-    :156250000
\\.\PHYSICALDRIVE1 :ATA      :WDC WD1602ABKS-1:3B04  :     WD-    :156250000
\\.\PHYSICALDRIVE2 :DGC      :RAID 5          :0429  :07000097    :3145728
\\.\PHYSICALDRIVE4 :DGC      :RAID 5          :0429  :09000097    :3145728
```

In the example shown in Figure 89, only two LUNs on the Cisco_FCoE_IOP54 storage group are visible. This explains why only two LUNs are seen in the previous **inq** command issued on the host.



FCNTR080500013 - Cisco_FCoE_IOP54: Storage Group Properties

Show LUNs: Not in other Storage Groups

Available LUNs

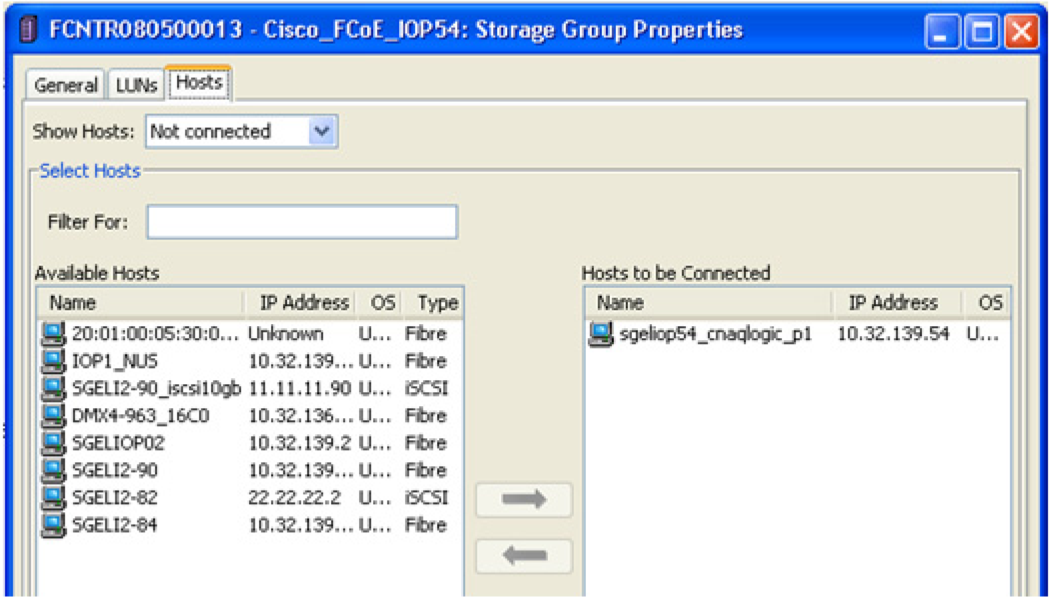
Name	ID	Capacity	Drive Type
⊞ MetaLUNs			
⊞ SP A			
⊞ SP B			
⊞ Thin LUNs			

Selected LUNs

Name	ID	Capacity	Drive Type	Host ID
LUN 519	519	3.000 GB	FC	0
LUN 521	521	3.000 GB	FC	2

Figure 89 Storage Group Properties window, LUNs tab

In the example shown in Figure 90, the host sgeliop54_cnaqlogic_p1 is assigned to the Cisco_FCoE_IOP54 storage group.



FCNTR080500013 - Cisco_FCoE_IOP54: Storage Group Properties

Show Hosts: Not connected

Select Hosts

Filter For:

Available Hosts

Name	IP Address	OS	Type
20:01:00:05:30:0...	Unknown	U...	Fibre
IOP1_NUS	10.32.139...	U...	Fibre
SGELI2-90_iscsi10gb	11.11.11.90	U...	iSCSI
DMX4-963_16C0	10.32.136...	U...	Fibre
SGELIOP02	10.32.139.2	U...	Fibre
SGELI2-90	10.32.139...	U...	Fibre
SGELI2-82	22.22.22.2	U...	iSCSI
SGELI2-84	10.32.139...	U...	Fibre

Hosts to be Connected

Name	IP Address	OS
sgeliop54_cnaqlogic_p1	10.32.139.54	U...

Figure 90 **Storage Group Properties window, Hosts tab**

Should you need to go further into troubleshooting the storage array and gather system and engineering level information, the following tools can be used for the VNX series or CLARiiON:

- SP_Collect (Storage processor collection tool)
- SPLAT (SP Log Analysis Tool)
- CAP2 (CLARiiON Array Properties)
- Admintool
- Psmtool (persistent storage manager tool)
- Ktcons (K10 trace console)
- Flarecons (flare console)

Next step

After completing this step, proceed to flowchart step #13.

Flowchart step #13, Some LUNs still missing?

Troubleshooting

To verify, if SOME or ALL LUNs are still missing, complete the following step.

Using the disk management utility of your host, verify if ALL or SOME of the LUNs are missing. Diskpart or inq utility tool can be used on a Windows host.

Example and interpretation of the results

To interpret the **inq** command, refer to "Flowchart step #2, Are ALL LUNs/devices missing?" on page 211.

Next steps

- Once the VNX series or CLARiiON configuration is updated, and all three LUNs are visible, then you can consider the issue resolved (flowchart step #17).
- If at this step you are still not seeing some of the LUNs, then you need to proceed to flowchart step #14.

Flowchart step #14, Check host configuration

Troubleshooting

The following are guidelines to check and troubleshoot your host:

- When performing troubleshooting actions on the host, your first impulse may be to reboot the host. Sometimes the problem can be solved by rebooting the host, but by doing this, you may never know what really happened or understand the real issue. This method does not capture all the necessary information that the host or server vendor may require to troubleshoot the issue, should it occur again.
- Before rebooting, check the host's ability to mount the LUNs/devices. Checking vendor documentation and release notes is highly recommended.
- Ensure that hardware bus rescan or device discovery has been tried.
- Different host troubleshooting tools can also be used for gathering OS system and storage array engineering information. Tools, such as EMCGrab and EMC Reports, can be used in conjunction with HEAT to check information relative to the latest *EMC Support Matrix* (ESM).

- Some technical documentation available through the EMC Online Support website at https://support.emc.com, such as the host configuration guides, can be used as a reference when dealing with host or OS level issues.
- If the issue is still unresolved, contact the host vendor.

Example and interpretation of the results

In the following example, a bus rescan is performed using the diskpart tool in a Windows host.

```
DISKPART> rescan

Please wait while DiskPart scans your configuration...

DiskPart has finished scanning your configuration.
```

Next step

Once the host has been checked and verified, proceed to flowchart step #13.

Flowchart step #15, CNA able to log in to fabric and name server?

Troubleshooting

At this step, you need to verify that the CNA is able to log in to the fabric and the Name Server. Use the NX-OS **show flogi database** and **show fcns database** commands to view the Fabric Login table and Name Server table, respectively.

Note: In MP-8000B, use the FOS **fcoe --loginshow** or **nsshow** command.

Example and interpretation of the results

To interpret the **inq** command, refer to "Flowchart step #3, CNA logged in to FCoE switch?" on page 215.

Next step

If the CNA's PWWN is visible, then proceed to flowchart step # 13. If it is not visible, proceed to flowchart step # 16.

Flowchart step #16, Check if DCBX, PFC, and FIP are working

Troubleshooting

Consider the following:

- The FCoE Initialization Protocol (FIP) allows the FCoE switch to discover and initialize FCoE-capable nodes in the Ethernet network. FIP or CEE-DCBX is the latest protocol supported on gen-2 CNAs. This case study uses a gen-2 CNA card, which means it is using FIP or CEE-DCBX on its FCoE entity discovery and initialization.

- To validate that FIP is working by design, check whether CEE-DCBX is enabled on both sides. This can be verified by checking if LLDP traffic is passing on the interface, as shown in the examples presented later in this section.

- You also want to make sure that there is no DCBX attribute mismatch. This can be checked by comparing the output of the NX-OS **show lldp interface ethernet 1/2** and **show lldp neighbors** commands. In the MP-800B switch, the command to check the LLDP TLVs is **show lldp interface** <*intf type*>. The LLDP TLV type should match on both ends.

- If some changes were made to the DCBX configuration, be aware that in some CNAs the change of settings takes affect only on the next outgoing LLDP packet. Therefore, in this scenario, you need to reset the CNA by disabling the port, then re-enabling it. This can also be done by clearing the virtual N_Port using the CNA management suite. First, check if TLVs are sent and received by the switch. Use the **show lldp traffic interface** <*intf type*> NX-OS command in the Nexus 5020. Use the **show lldp statistics interface** <*intf type*> CMSH command in the MP-8000B. You should see numbers on both Total frames in/out fields, as shown in the example presented later in this section.

- If something is showing up in either the error, discarded, or unrecognized TLV fields, then you need to verify what is happening at the protocol level. You can place an analyzer device between the host and the switch, which acts as a network tap. This provides a view of what is taking place during link initialization and before the FCoE node can log in. (Refer to "Process flow" on page 180 for more details.) You can also use debugging CLI tools, such as **debug dcbx all** and **debug flogi all**, which provide useful information when troubleshooting login problems. Verify the Priority Flow Control (PFC). To check if switchport is negotiating PFC with the CNA, use the **show interface** *<intf type>* **priority-flow-control** NX-OS command, an example of which is provided in the following section. By checking this, you can determine if the port is behaving as it should. By default, the Ethernet interfaces negotiate PFC capability with the connected CNA. You can override the negotiation result by force-enabling the PFC capability.

Example and interpretation of the results

The following is an example showing LLDP traffic on an interface:

```
Nexus 5020 # show lldp traffic interface ethernet 1/2
LLDP traffic statistics:

    Total frames out: 31961
    Total Entries aged: 139
    Total frames in: 31174
    Total frames received in error: 0
    Total frames discarded: 0
    Total TLVs unrecognized: 0
```

The following examples show how to check the LLDP information of an interface:

```
Nexus 5020 # show lldp interface ethernet 1/2
tx_enabled: TRUE
rx_enabled: TRUE
dcbx_enabled: TRUE

Port MAC address:  00:0d:ec:b1:58:c9

Remote Peers Information

Remote peer's MSAP: length 12 Bytes:
00    c0    dd    10    28    ba    00    c0    dd    10    28    ba

LLDP TLV's
LLDP TLV type:Chassis ID  LLDP TLV Length: 7
LLDP TLV type:Port ID  LLDP TLV Length: 7
LLDP TLV type:Time to Live  LLDP TLV Length: 2
LLDP TLV type:LLDP Organizationally Specific  LLDP TLV Length: 61
LLDP TLV type:END of LLDPDU  LLDP TLV Length: 0

Nexus 5020 # show lldp traffic interface ethernet 1/2
LLDP traffic statistics:

    Total frames out: 31961
    Total Entries aged: 139
    Total frames in: 31174
    Total frames received in error: 0
    Total frames discarded: 0
    Total TLVs unrecognized: 0
```

The following is an example of how to check the neighbor LLDP information:

```
Nexus 5020 # show lldp neighbors
LLDP Neighbors

Remote Peers Information on interface Eth1/2
Remote peer's MSAP: length 12 Bytes:
00    c0    dd    10    28    ba    00    c0    dd    10    28    ba

LLDP TLV's
LLDP TLV type:Chassis ID  LLDP TLV Length: 7
LLDP TLV type:Port ID  LLDP TLV Length: 7
LLDP TLV type:Time to Live  LLDP TLV Length: 2
LLDP TLV type:LLDP Organizationally Specific  LLDP TLV Length: 61
LLDP TLV type:END of LLDPDU  LLDP TLV Length: 0
```

The following are options available when doing FLOGI and DCBX debugging:

```
Nexus 5020 # debug dcbx ?
  all       Configure all debug flags of dc
  demux     Configure debugging of dcx message demux
  deque     Configure debugging of dcx message deque
  error     Configure debugging of dcx error
  event     Configure debugging of dcx FSM and Events
  ha        Configure debugging of dcx HA
  packets   Configure debugging of dcx packets
  trace     Configure debugging of dcx trace
  warning   Configure debugging of dcx warning

Nexus 5020 # debug flogi ?
  action    Configure debugging of flogi actions
  all       Configure all debug flags of flog
  bypass    Bypass some components in flogi execution
  demux     Configure debugging of flogi message demux
  error     Configure debugging of flogi error
  event     Configure debugging of flogi FSM and Events
  ha        Configure debugging of flogi HA
  init      Configure debugging of flogi adds, deletes and inits
  timers    Configure debugging of flogi message timers
  trace     Configure debugging of flogi trace
  warning   Configure debugging of flogi warning
```

In the following example, the PFC mode is set to negotiate PFC capability, the operation is *On*, and packets transmitted are 1597980. Even if you did not set up the PFC on the interface with the **priority-flow-control mode [auto | on]** command, the interface has a default behavior to negotiate PFC with CNA. If these are not working as expected, then contact the switch or CNA hardware vendor for further troubleshooting. It could be an issue with the driver, a bug on the switch, or a firmware or hardware problem.

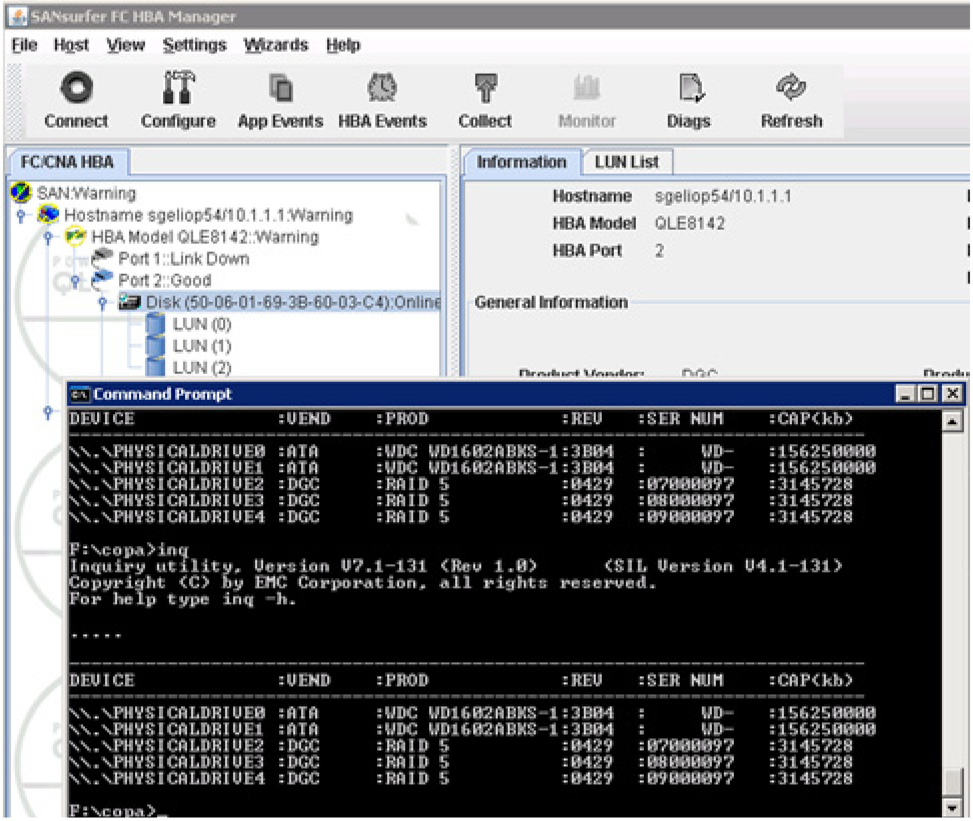
```
<Snip from the running configuration>

interface Ethernet1/2
  description test
  switchport mode trunk
  mac port access-group deny_fcoe
  spanning-tree port type edge trunk
  spanning-tree bpduguard enable

Nexus 5020 # show interface ethernet 1/2 priority-flow-control
===========================================================
Port               Mode Oper(VL bmap)  RxPPP      TxPPP
===========================================================

Ethernet1/2        Auto On  (8)       0          1597980
```

Next step

Once the DCBX, PFC and FIP have been examined, proceed to flowchart step # 15.

Flowchart step #17, Problem solved

You arrived at this step after you have verified that the issue is resolved. Using CNA management tools and the **inq** command, you can check if all the LUNs are visible, as shown in Figure 91.



Hostname sgeliop54/10.1.1.1
HBA Model QLE8142
HBA Port 2

```
DEVICE              :VEND     :PROD               :REV    :SER NUM    :CAP(kb)
--------------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE1 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE2 :DGC      :RAID 5            :0429   :07000097    :3145728
\\.\PHYSICALDRIVE3 :DGC      :RAID 5            :0429   :08000097    :3145728
\\.\PHYSICALDRIVE4 :DGC      :RAID 5            :0429   :09000097    :3145728

F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.

.....

--------------------------------------------------------------------------------
DEVICE              :VEND     :PROD               :REV    :SER NUM    :CAP(kb)
--------------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE1 :ATA      :WDC WD1602ABKS-1:3B04   :       WD-    :156250000
\\.\PHYSICALDRIVE2 :DGC      :RAID 5            :0429   :07000097    :3145728
\\.\PHYSICALDRIVE3 :DGC      :RAID 5            :0429   :08000097    :3145728
\\.\PHYSICALDRIVE4 :DGC      :RAID 5            :0429   :09000097    :3145728

F:\copa>_
```

Figure 91 **Verify LUNs are visible**

Case Study #2, Unable to access a shared folder in the file server

Problem definition

Unable to access a shared folder in the file server.

Background

In this example, Windows client #1 is not able to see the shared folder in the file server Host A. This file server is connected to another LAN (iSCSI LAN) to which the VNX series or CLARiiON storage device is connected. VNX series or CLARiiON LUNs are presented to the file server and are mapped as local drives, then the operating system's file server service is activated to provide CIFS (SMB) service. The troubleshooting techniques in this case study use the concepts discussed from "Troubleshooting basic FCoE and CEE problems" on page 180. A flowchart is provided to use as a guide in solving the issue (see Figure 93 on page 254).

Topology

In this environment, a multi-layer switch is used to simulate layered campus LAN design.

Note: Due to hardware availability limitations while this case study was being developed, the setup in this example is composed only of one Nexus 7000 switch, which serves as the redundant core and distribution layer. This setup is employed only for the purpose of demonstrating and discussing different network troubleshooting topics. In a production LAN, it is highly recommended that you have redundant distribution and core switches for better resiliency and scalability.

The Nexus 7000 switch is used as the collapsed core and distribution switch to provide layer 3 routing for these VLANs. Logical Switched Virtual Interfaces (SVIs) are used in the Nexus 7000 switch to provide a layer 3 interface and to serve as a default gateway for each node. The port channel links between the access switches and the Nexus 7000 switch are the layer 2 trunk ports that provide interswitch links and carry traffic from multiple VLANs.

One of the port channels is in blocking mode and will serve as a backup link should the primary link to the Nexus 7000 switch fail. Using a lower value bridge priority, the Nexus 7000 will be the root bridge in this environment.

Both the Nexus 5010 and MP-8000B switches are used as access switches and are also used to provide switch port connection to Windows client #1, Windows client #2, Windows client #3, and Host A (the file server). These switch ports are configured as VLAN70 or VLAN80, as depicted in Figure 92 on page 253.

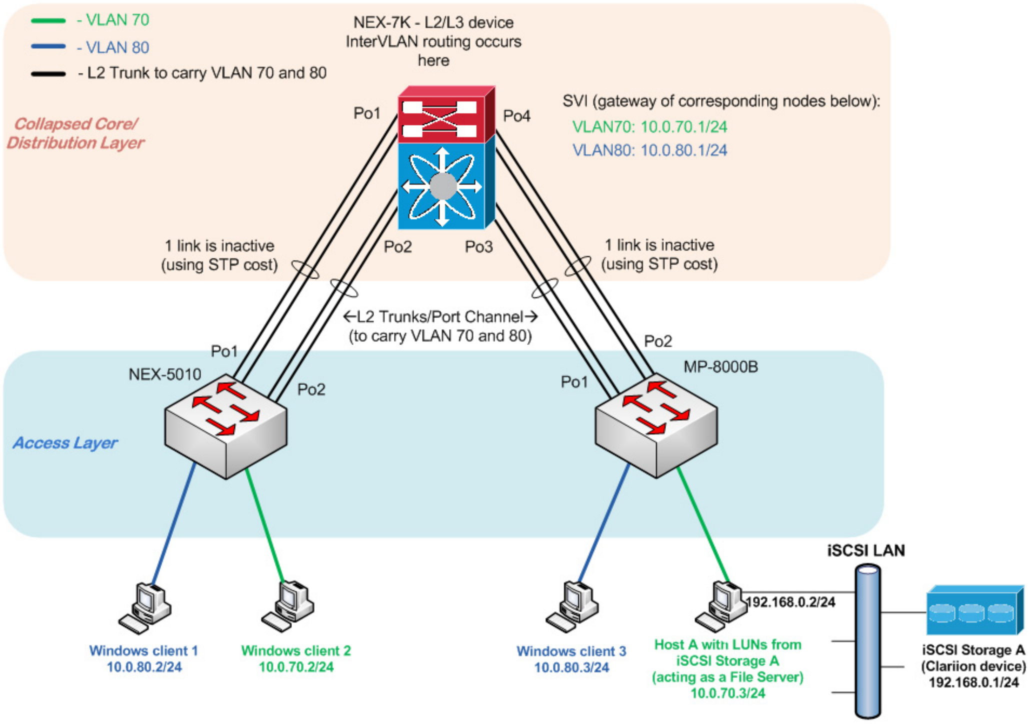


Figure 92 Case study #2 topology

This case study will analyze issues using the troubleshooting flowchart shown in Figure 93 on page 254. Examples are provided for each step in the flowchart.

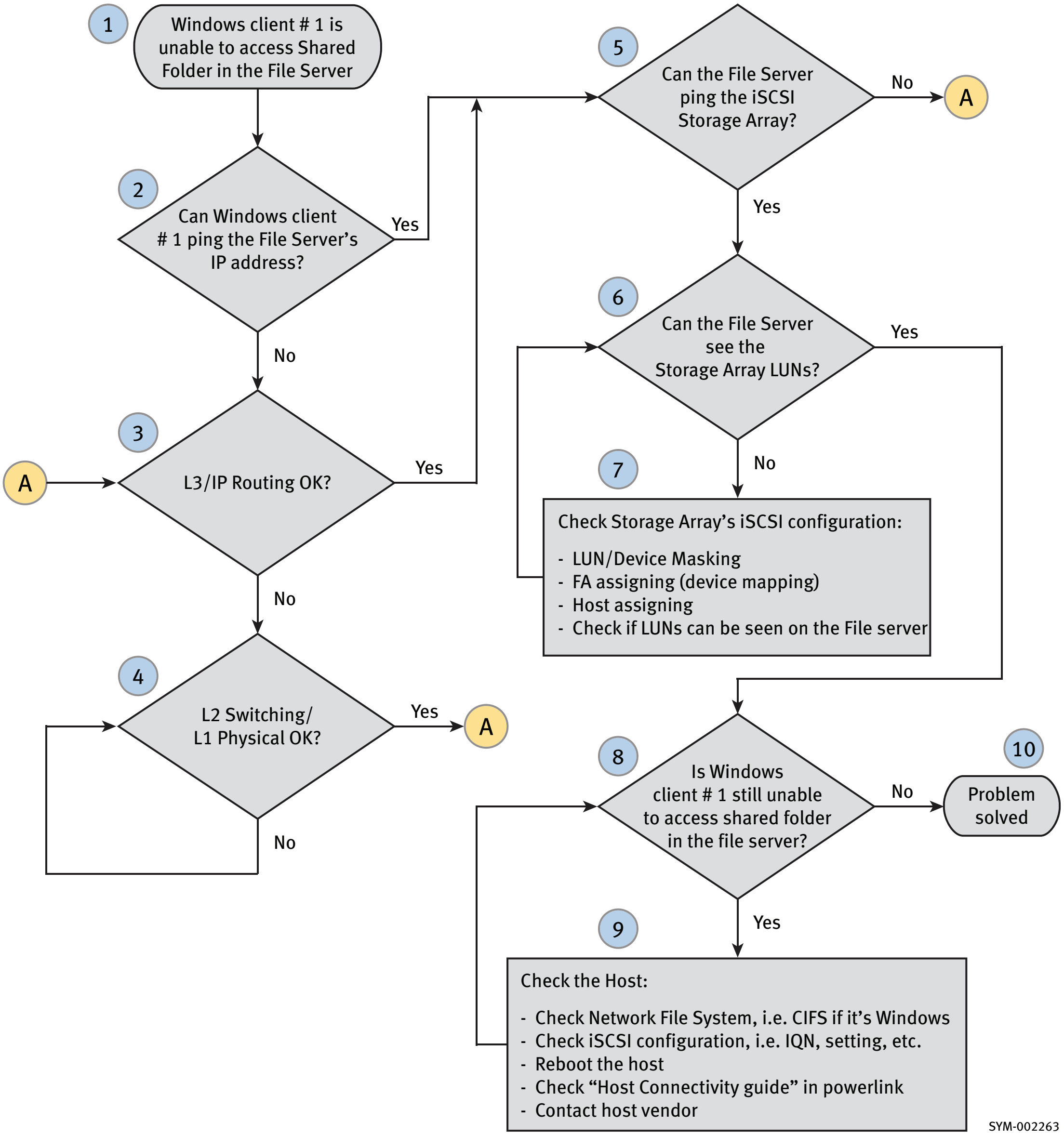

1
Windows client # 1 is unable to access Shared Folder in the File Server
2
Can Windows client # 1 ping the File Server's IP address?
Yes
No
3
A
L3/IP Routing OK?
Yes
No
4
L2 Switching/ L1 Physical OK?
Yes
A
No
5
Can the File Server ping the iSCSI Storage Array?
No
A
Yes
6
Can the File Server see the Storage Array LUNs?
Yes
No
7
Check Storage Array's iSCSI configuration:
- LUN/Device Masking
- FA assigning (device mapping)
- Host assigning
- Check if LUNs can be seen on the File server
8
Is Windows client # 1 still unable to access shared folder in the file server?
No
10
Problem solved
Yes
9
Check the Host:
- Check Network File System, i.e. CIFS if it's Windows
- Check iSCSI configuration, i.e. IQN, setting, etc.
- Reboot the host
- Check "Host Connectivity guide" in powerlink
- Contact host vendor
SYM-002263

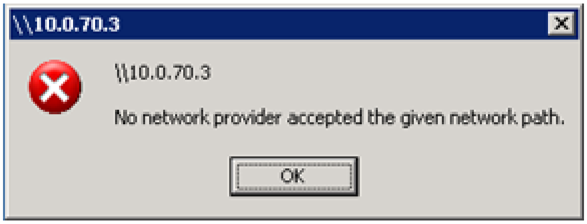
Figure 93 **Troubleshooting flowchart for case study #2**

Using the flowchart in Figure 93, each step will be further discussed in this section.

- "Flowchart step #1, Windows client #1 is unable to access the shared folder in the file server" on page 255
- "Flowchart step #2, Can Windows client #1 ping the file server's IP address?" on page 256
- "Flowchart step #3, L3/IP routing OK?" on page 257
- "Flowchart step #4, L2 switching/ L1 physical OK?" on page 263
- "Flowchart step #5, Can the file server ping the iSCSI storage array?" on page 278
- "Flowchart step #6, Can the file server see the storage array LUNs?" on page 280
- "Flowchart step #7, Check the storage array's iSCSI configuration" on page 280
- "Flowchart step #8, Is Windows client # 1 still unable to access the shared folder in the file server?" on page 284
- "Flowchart step #9, Check the host" on page 285
- "Flowchart step #10, Problem is solved" on page 290

Flowchart step #1, Windows client #1 is unable to access the shared folder in the file server

The first step in this troubleshooting scenario is to define the problem. The problem in this example is that the Windows client #1 (10.0.80.2) is unable to see the shared folder in the file server (Host A, 10.0.70.3). From Windows client #1, an error message will display if there is no access to the shared folder.



\\10.0.70.3

No network provider accepted the given network path.

Figure 94 Error message

Flowchart step #2, Can Windows client #1 ping the file server's IP address?

Troubleshooting

At this stage, you need to test the IP connectivity by using ICMP. Ping is one of the many tools within the ICMP protocol suite that network administrators can use to troubleshoot issues. Complete the following steps when pinging a device.

1. Initiate a ping command from Windows client #1 (10.0.80.2) to the file server (Host A, 10.0.70.3). Successful replies from the destination IP would look similar to the example in the following section.
2. Ensure that the *received* packets are equal to the *sent* packets.
3. If ICMP packets are not passing through and you see that interfaces are up, then it is worth checking the switch's L2/L3 interface configurations to see if any ACLs have been put in place. Also, check to make sure there are no firewalls installed between the devices and that the host in question is not running a local firewall to prevent any ICMP traffic from passing.

Example and interpretation of the results

In the following example, four (4) packets are sent and four (4) packets are received. Note the response time because packet delays can cause performance issues, especially when performing backup and recovery. In LAN environments, 30 ms or less is the ideal response time, whereas in WAN environments, 150 ms or less is the accepted value.

```
F:\Documents and Settings\Administrator>ping 10.0.70.3

Pinging 10.0.70.3 with 32 bytes of data:

Reply from 10.0.70.3: bytes=32 time<1ms TTL=127
Reply from 10.0.70.3: bytes=32 time<1ms TTL=127
Reply from 10.0.70.3: bytes=32 time<1ms TTL=127
Reply from 10.0.70.3: bytes=32 time<1ms TTL=127

Ping statistics for 10.0.70.3:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 0ms, Maximum = 0ms, Average = 0ms
```

Next steps

- If Windows client #1 (10.0.80.2) can't ping the file server (Host A, 10.0.70.3), then proceed with flowchart step # 3.
- If ping is successful and you are still unable to access the shared folder in the file server (Host A, 10.0.70.3), then proceed to flowchart step # 5.

Flowchart step #3, L3/IP routing OK?

Troubleshooting

In performing layer 3 troubleshooting, there are many things to consider, including the following:

- First, verify if the IP addressing is correct. Ensure that you have the correct IP address and subnet mask assigned to your nodes (see Figure 95 on page 259) as well as the correct layer 3 port assignment on your distribution/core switches. Use the **show ip interface brief** CLI command in the switches.
- If DHCP is used in addressing the L3 nodes, then verify that the correct fields are being assigned by the DHCP server. You can check this by comparing the IP fields (such as IP address range, subnet masks, and gateway) that were assigned by the network administrator in the DHCP server to the actual IP fields assigned by the DHCP server to the host, as shown in the example used shown later in this section.
- Check the default gateway (also known as next hop IP address). TCP/IP hosts always use a default gateway when the destination is not in the same network. If the default gateway of the host is configured improperly, the data will not be routed. All hosts need to point to a layer 3 device, such as a router or multi-layer switch (for example, the Nexus 7000) on the same network in order to be used as the default gateway.

  The **ping** and **traceroute** commands can be used to isolate default gateway problems. When the host uses a dynamic method of selecting a default gateway, there is a greater possibility that it may fail. This is why it is always a good practice to have a default gateway when an infrastructure involves more than one network. When the default gateway is not reachable, or is not forwarding the traffic from the host, the issue is most likely related to the configuration being incorrect. The default gateway should be on the same subnet where the host resides.

- Check that the routing information is correct on the routers in question. Ensure that the L3 devices (distribution/core switches) have routes for each of the subnets that are trying to communicate with each other. In this case study, each VLAN represents one subnet, therefore the L3 device Nexus 7000 must have routes for each VLAN (10.0.80.0/24 and 10.0.70.0/24) in its routing table.

  To view the routing table, use the **show ip route** command on the layer 3 device, as shown in an example in the following section.

- If the routing table has been verified and the L3 routing is still unresolved, a routing policy may have been put in place, which can affect the routing of the Nexus 7000. Verify whether there is an ACL, Route Map, Filter list, or Distribute list that could affect the routing. This can be accomplished by issuing one of the following three commands:

  - **show running-configuration**
  - **show running-config interface** <*intf type*>
  - **show access-lists** (or **show ip access-list)**
  - **show route-map** <*route-map name*>

  Once the network is free of any layer 3 issues, you should be able to ping from source IP to destination IP address, and vice versa. In this case study, you should be able to have successful ping replies between these source and destination IP pairs:

  ```
  Source: Windows Client #1 (10.0.80.2)
  Destination: Host A or File Server's 1st NIC
  (10.0.70.3)

  Source: Host A or File Server's 2nd NIC (192.168.0.2)
  Destination: CLARiiON FA port (192.168.0.1)
  ```

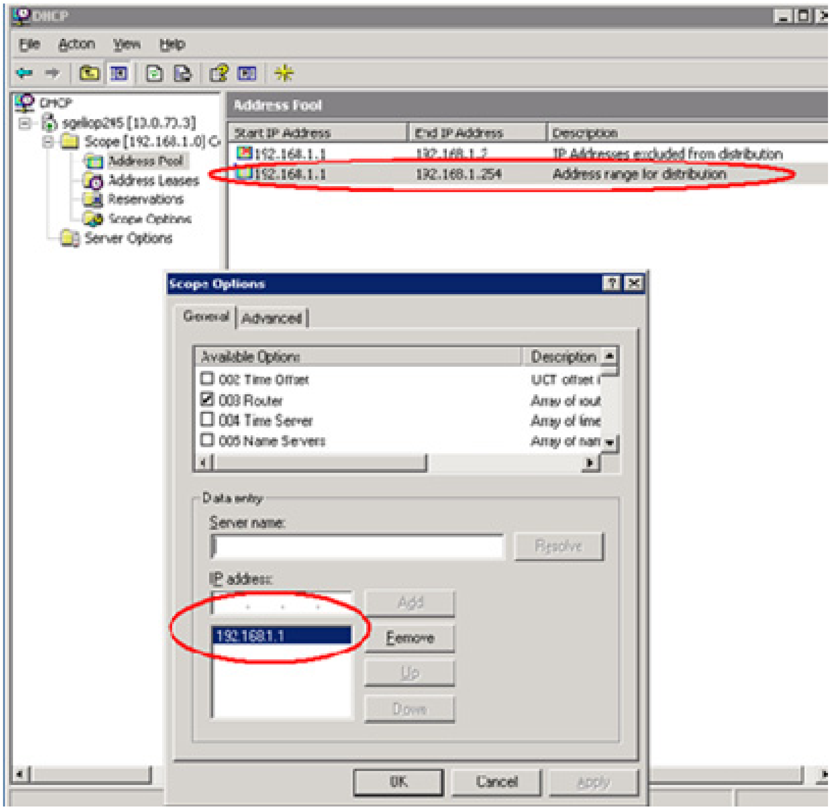
Example and interpretation of the results

To verify the statically/dynamically assigned IP fields of the host, use the **ipconfig /all** command in the DOS prompt, as shown next.

```
F:\Documents and Settings > ipconfig /all

Ethernet adapter Local Area Connection 6:
Connection-specific DNS Suffix  . :
   Description . . . . . . . . . . . : Brocade 10G Ethernet Adapter #2
   Physical Address. . . . . . . . . : 00-05-1E-9A-A9-97
   DHCP Enabled. . . . . . . . . . . : No
   IP Address. . . . . . . . . . . . : 10.0.70.3
   Subnet Mask . . . . . . . . . . . : 255.255.255.0
   Default Gateway . . . . . . . . . : 10.0.70.1
<output truncated>
```

To ensure the correct IP information assigned by the Network Administrator is correct on the DHCP server you will need to check the Address Pool.

On the DHCP server, go to Administrative Tools > DHCP > Address Pool to verify the IP address range being assigned matches what was originally assigned by the network administrator. To verify the default gateway is also being assigned correctly, select Scope Options and locate the router information (003). An example is shown in Figure 95.



Address Pool

Start IP Address	End IP Address	Description
192.168.1.1	192.168.1.2	IP Addresses excluded from distribution
192.168.1.1	192.168.1.254	Address range for distribution

Available Options	Description
☐ 002 Time Offset	UCT offset i
☑ 003 Router	Array of rout
☐ 004 Time Server	Array of time
☐ 005 Name Servers	Array of nam

IP address: 192.168.1.1

Figure 95 **Scope options example**

When troubleshooting a default gateway issue, **traceroute** plays a large role in verifying the number of routers that the packet traverses. The traceroute from Windows client #1, shown next, shows that there is only one router (10.0.80.1) to traverse in order to reach the destination (Host A, 10.0.70.3).

```
F:\Documents and Settings\Administrator>tracert 10.0.70.3

Tracing route to Host_A [10.0.70.3]
over a maximum of 30 hops:

  1    <1 ms    <1 ms    <1 ms  10.0.80.1
  2     3 ms    <1 ms    <1 ms  Host_A [10.0.70.3]

Trace complete.

 Nexus 7000 # show ip interface brief
IP Interface Status for VRF "default"
Interface            IP Address      Interface Status
Vlan70               10.0.70.1       protocol-up/link-up/admin-up
Vlan80               10.0.80.1       protocol-up/link-up/admin-up
```

The following is an example output of the **show ip route** command from Nexus 7000 multi-layer switch.

```
 Nexus 7000 # show ip route
IP Route Table for VRF "default"
'*' denotes best ucast next-hop       '**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

0.0.0.0/32, 1 ucast next-hops, 0 mcast next-hops
   *via Null0, [220/0], 3w4d, local, discard
10.0.70.0/24, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.70.1, Vlan70, [0/0], 1w4d, direct
10.0.70.0/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.70.0, Null0, [0/0], 1w4d, local
10.0.70.1/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.70.1, Vlan70, [0/0], 1w4d, local
10.0.70.3/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.70.3, Vlan70, [2/0], 1w4d, am
10.0.70.255/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.70.255, Vlan70, [0/0], 1w4d, local
10.0.80.0/24, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.80.1, Vlan80, [0/0], 1w4d, direct
10.0.80.0/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.80.0, Null0, [0/0], 1w4d, local
10.0.80.1/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.80.1, Vlan80, [0/0], 1w4d, local
10.0.80.2/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.80.2, Vlan80, [2/0], 1w4d, am
10.0.80.255/32, 1 ucast next-hops, 0 mcast next-hops, attached
   *via 10.0.80.255, Vlan80, [0/0], 1w4d, local
255.255.255.255/32, 1 ucast next-hops, 0 mcast next-hops
   *via sup-eth0, [0/0], 3w4d, local
 Nexus 7000 #
```

Notice from the previous output that both 10.0.80.0/24 and 10.0.70.0/24 routes are present and are connected through the layer 3 interface IP addresses. These IP addresses are also the default gateway of the Windows clients. Nodes on VLAN70 use the SVI IP address 10.0.70.1 as its default gateway, whereas nodes on VLAN80 use the SVI IP address 10.0.80.1 as its default gateway.

Switched Virtual Interface (SVI) is a layer 3 logical interface, which can be created and mapped to a layer 2 VLAN. This feature is only supported on multi-layer switches, such as the Nexus 7000. Note the word **direct** on the above routes. This shows that these routes are directly connected to the Nexus 7000 layer 3 device and have an administrative distance (AD) of 0. The AD is the metric used by the router to determine and measure the trustworthiness of the source of the route. Directly-connected routes are always preferred over statically- and dynamically-learned routes.

For more information about SVI or administrative distance and metrics, refer to Cisco documentation at http://www.cisco.com.

Routes can be learned by either *static* or *dynamic* routing.

- Static routing is accomplished by manually configuring any routes needed by entering the remote network and providing which neighboring path is needed to reach the remote network. Static routes are local to the router by default and are not advertised to neighboring routers unless configurations are added to do so.
- Dynamic routing is based on active routing protocols (such as RIP, EIGRP, or OSPF) where routers share route information with one another. Dynamic routing has an advantage over static routing, such as automatic route redundancy, which allows load sharing across multiple paths. Dynamic routing can become complicated on an enterprise and service provider environment. IGPs or Interior Routing Protocols, such as RIP, EIGRP, OSPF, and IS-IS can be used to connect large networks. Then, EGP, or Exterior Routing Protocols such as BGP, can be used to connect different autonomous systems or large networks that have IGPs inside.

This case study uses Inter VLAN routing (IVR). Since VLAN70 and VLAN80 are both connected on the same layer 3 switch, there is no need to use static or dynamic routing configurations. To view the IVR information use either the **show ip route** or **show ip route interface** <*vlan interface*> command, as shown in the next example.

```
 Nexus 7000 # show ip route interface vlan70
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

10.0.70.0/24, ubest/mbest: 1/0, attached
    *via 10.0.70.1, Vlan70, [0/0], 2w0d, direct
10.0.70.1/32, ubest/mbest: 1/0, attached
    *via 10.0.70.1, Vlan70, [0/0], 2w0d, local
10.0.70.3/32, ubest/mbest: 1/0, attached
    *via 10.0.70.3, Vlan70, [2/0], 1d18h, am
10.0.70.255/32, ubest/mbest: 1/0, attached
    *via 10.0.70.255, Vlan70, [0/0], 2w0d, broadcast


 Nexus 7000 # show ip route interface vlan80
IP Route Table for VRF "default"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]

10.0.80.0/24, ubest/mbest: 1/0, attached
    *via 10.0.80.1, Vlan80, [0/0], 2w0d, direct
10.0.80.1/32, ubest/mbest: 1/0, attached
    *via 10.0.80.1, Vlan80, [0/0], 2w0d, local
10.0.80.2/32, ubest/mbest: 1/0, attached
    *via 10.0.80.2, Vlan80, [2/0], 22:11:11, am
10.0.80.255/32, ubest/mbest: 1/0, attached
    *via 10.0.80.255, Vlan80, [0/0], 2w0d, broadcast
```

For more information about static and dynamic IP routing on the Nexus 7000, refer to Cisco documentation at http://www.cisco.com.

The following is an example of examining the Nexus 7000 if a routing policy has been put in place. It shows that there are no ACLs or Filters configured that could affect the routing operation of this multi-layer switch.

```
 Nexus 7000 # show running-config interface vlan70
version 4.1(4)

interface Vlan70
  no shutdown
  ip address 10.0.70.1/24
```

Next steps

- If L3 routing is ok, then proceed with flowchart step # 5.
- If not, proceed to flowchart step # 4.

Flowchart step #4, L2 switching/ L1 physical OK?

Troubleshooting

When performing L2/L1 troubleshooting, there are numerous things to consider, including the cabling, media type, spanning-tree protocol, VLAN membership, VLAN trunking encapsulation, ACLs (Ethertype and MAC ACLs), port-channel, interface mode mismatch (access, trunk, converged), speed mismatch, and duplex mismatch.

Use the following guidelines when performing Layer 1 and Layer 2 troubleshooting.

- Eliminate hardware problems

  At this step, the first thing you should eliminate is whether or not it is a hardware problem. To verify all components across the L2 path are free of any hardware issues, check the host, its NIC status, the access switch ports, the trunk ports, and the port-channel interfaces. From the host, the port will immediately detect connections once a working cable is connected and the access switchport is enabled.

  From the switch standpoint, you can verify the operational status by using the appropriate **show interface** <*intf type*> command. If one of the links is not showing up, swapping each component would speed up the connectivity troubleshooting process. For more information about layer 1 troubleshooting, refer to "OSI layers" on page 186.

- Ensure there is no MAC layer issue

  Verify if both nodes' MAC addresses have been learned by the switch by checking the MAC table. If MAC addresses are not learned, even if the access switch ports are active, you must ensure that there is no MAC ACL preventing any L2 communication between the host/storage array and the switch. These can be verified by using the following CLI commands:

  - For both the MP-8000B and Nexus 5010 switches, using either:

    show running-configuration, or
    show running-config interface <*intf type*>

  - For the MP-8000B switch:

    show mac access-group interface <*intf type*>

  - For the Nexus 5010 switch:

    show mac access-lists summary

IMPORTANT

Once you have ruled out hardware-related problems, look for any possible configuration mismatch and then validate the way the layer 2 protocols work.
If L2 ports and L2 protocols are configured properly (and you have already confirmed that there are no hardware-related problems), then the issue may be the way the switch firmware/OS is written. It could also be a software bug or an interoperability problem.

- ◆ Check the Spanning Tree protocol

The spanning tree reconfiguration can occur in less than one second with Rapid PVST+ (in contrast to 50 seconds with the default settings in the normal STP). Rapid PVST+ is the default STP mode on the Nexus Series switches, while MP-8000B switches use Rapid STP as its default mode. For more information about RSTP, refer to "Rapid Spanning Tree (802.1w)" on page 102. For more information about Rapid PVST+, refer to Cisco Nexus documentation at http://www.cisco.com.

When troubleshooting RSTP, you need to first verify that it is working and doing what is intended, such as forwarding/blocking the correct L2 ports. One way of checking if RSTP is working is by using the following switch CLI commands:

- For Nexus 5010:
  show spanning-tree summary
- For MP-8000B:
  show spanning-tree brief

Note: Most protocol level troubleshooting, such as Spanning-Tree Protocol (STP), port-channel, and VLANs, involves switches, since edge ports do not need layer 2 protocol updates, such as BPDUs. Controlling layer 2 protocol updates on edge ports can be enabled by configuring **portfast** and **bpduguard** on access ports.

When troubleshooting Spanning Tree, it is advisable to verify different aspects of STP from the actual switch. Ideally, the first thing to discover is the current root bridge and its location in the network. You might also want to see the bridge ID of the switch to which you are connected in order to see how it participates in STP.

To see the STP bridge ID and bridge priority of the switch, use the NX-OS **show spanning-tree vlan** <*vlan id*> command in the Nexus Series switches and the CMSH **show spanning-tree brief** command in the MP-8000B, as shown in the following section.

If the actual root bridge should *not* be the root bridge, there are a few commands that can be issued to force a specific switch to become a root bridge. In the Cisco switch, the following NX-OS command can force the switch priority to become the lowest number, thereby making it the root bridge:

spanning-tree vlan <*vlan id*> **root primary**

If you want to delve deeper into troubleshooting the issue, you can use an analyzer device and verify if BPDUs are sent and received by the switches. For more information, refer to "BPDUs" on page 94. Figure 99 shows that the BPDUs are sent every two seconds, which is the default on both the Nexus 5010 and MP-8000B switches.

Debug commands can also be used on both Nexus 5010 and MP-8000B switches.

- For Nexus 5010, **debug spanning-tree bpdu_rx** can be used for incoming BPDUs and **debug spanning-tree bpdu_tx** for outgoing BPDUs.
- For MP-8000B, **debug spanning-tree bpdu rx** can be used for incoming BPDUs and **debug spanning-tree bpdu tx** for outgoing BPDUs.

Examples of these commands are shown in the following section.

- Check the Port-Channel

  Another layer 2 aspect you should check is the port-channel. If, for some reason, port-channels are not forming and you have verified that the interfaces are up and working, check the channel-group configuration. If you are setting up the bundled link using the dynamic way (LACP), pay attention to the correct combination of the configuration on both ends of the link. For more details on LACP, refer to "Link Aggregation Control Protocol (LACP)" on page 107.

  To verify the channel-group configuration, use the interface filter command, **show run interface** <*intf type*>, as shown in the following example. Since LACP frames are sent in multicast, check whether the port is receiving multicast LACP frames from the other end of the link. Use the **show interface** <*intf type*>

counters command and note the InMcastPkts and OutMcastPkts values. These should be showing incremental values as you enable the ports to form a dynamic port channel. Examples are provided in the next section.

Note that if you configure an active-passive pair for dynamic port-channel, the first one that will send the multicast LACP frame is the port configured as *active*. In this case, the active port is the Nexus 7000. Therefore, from the captured trace shown in Figure 101 on page 276, you will see that the ethernet 1/4 (with source MAC 0023.ebd6.c173) of the Nexus 7000 sent the LACP frame first, then the ethernet 1/4 (with source MAC 000d.ecb1.58cb) of the Nexus 5010 responded accordingly.

- Verify VLAN configuration

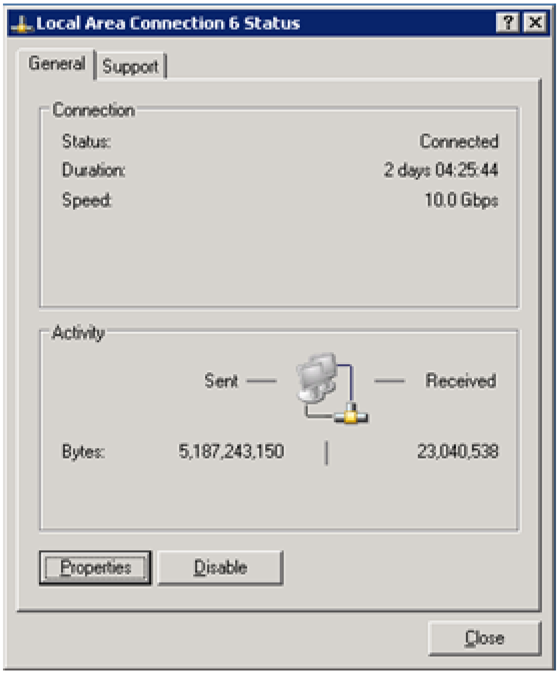
  It is important to verify the VLAN configuration. You must ensure that the access port where your host is connected is assigned to the correct VLAN number. You also need to verify that this VLAN is configured on the switch. Use the CLI **show vlan brief** command to verify if the VLAN is already created. Examples are provided in the following section.

  You also want to make sure that L2 trunk ports are configured to allow the VLAN number assigned to the access port. This can be verified using the **show interface** <*intf type*> **switchport** command, as shown in the following example. In the current Nexus Series and MP-8000B switches, the only supported trunking encapsulation is dot1Q, which is already enabled by default.

  Also verify that the Native Trunking VLAN is consistent. It is extremely important that switches have the same Native VLAN. If they do not match, the Trunk will not come up. See the example in the following section to examine the Native Trunking VLAN and how to change it if it is not consistent on both ends of the switch.

Example and interpretation of the results

As shown in Figure 96 on page 267, the **Local Area Connection Status** dialog box shows whether the port is connected or not. It also shows the detected speed, as well the received frames, which are depicted in bytes.


Local Area Connection 6 Status

General | Support

Connection

Status: Connected

Duration: 2 days 04:25:44

Speed: 10.0 Gbps

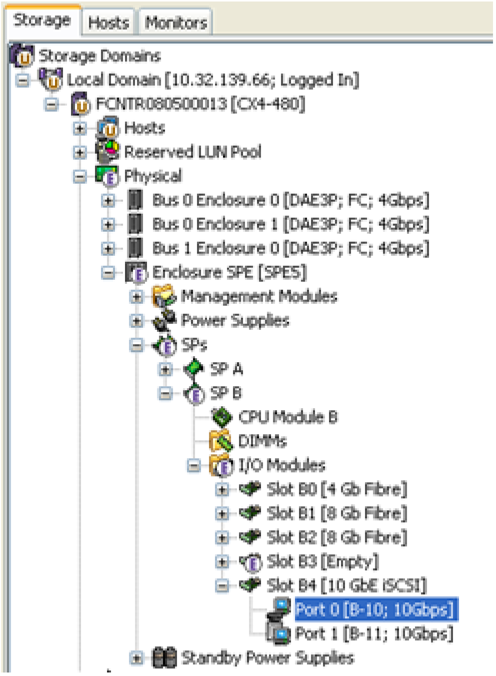
Activity

Sent — — Received

Bytes: 5,187,243,150 | 23,040,538

Figure 96 **Local Area Connection 6 Status dialog box**

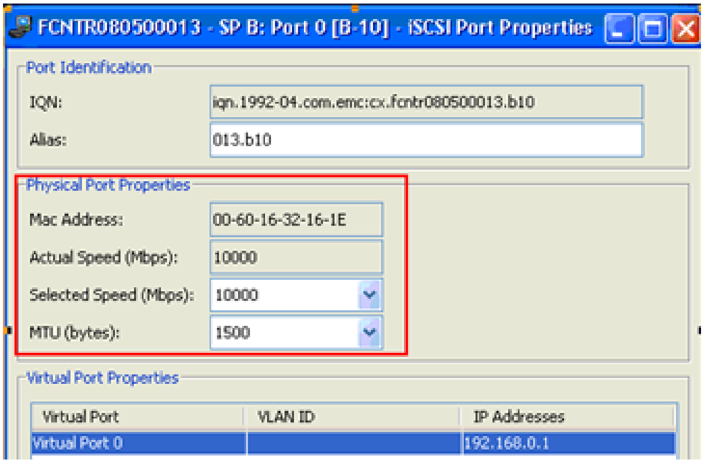
From the storage system (VNX series or CLARiiON) perspective, confirm whether there is an active link between the file server and the VNX series or CLARiiON, as shown in Figure 97 on page 268.



```
Storage Domains
  Local Domain [10.32.139.66; Logged In]
    FCNTR080500013 [CX4-480]
      Hosts
      Reserved LUN Pool
      Physical
        Bus 0 Enclosure 0 [DAE3P; FC; 4Gbps]
        Bus 0 Enclosure 1 [DAE3P; FC; 4Gbps]
        Bus 1 Enclosure 0 [DAE3P; FC; 4Gbps]
        Enclosure SPE [SPE5]
          Management Modules
          Power Supplies
          SPs
            SP A
            SP B
              CPU Module B
              DIMMs
              I/O Modules
                Slot B0 [4 Gb Fibre]
                Slot B1 [8 Gb Fibre]
                Slot B2 [8 Gb Fibre]
                Slot B3 [Empty]
                Slot B4 [10 GbE iSCSI]
                  Port 0 [B-10; 10Gbps]
                  Port 1 [B-11; 10Gbps]
          Standby Power Supplies
```

Figure 97 Storage tab, confirming active link

The Unisphere/Navisphere Manager's tool shows that the VNX series or CLARiiON port has detected the 10 GE speed of the link, as illustrated in Figure 98 on page 269.



FCNTR080500013 - SP B: Port 0 [B-10] - iSCSI Port Properties

Port Identification

IQN: iqn.1992-04.com.emc:cx.fcntr080500013.b10

Alias: 013.b10

Physical Port Properties

Mac Address: 00-60-16-32-16-1E

Actual Speed (Mbps): 10000

Selected Speed (Mbps): 10000

MTU (bytes): 1500

Virtual Port Properties

Virtual Port	VLAN ID	IP Addresses
Virtual Port 0		192.168.0.1

Figure 98 **Physical Port Properties**

In the following example, there is no MAC ACL configured on the port-channel1 interface. Therefore, the switch is not performing any security filter on this port.

```
Nexus 5010 # show running-config interface port-channel 1
version 4.1(3)N1(1a)

interface port-channel1
  switchport mode trunk
  speed 10000
```

In the following examples for Nexus 5010 and MP-8000B, you can see that RSTP is working:

```
Nexus 5010 # show spanning-tree summary
Switch is in rapid-pvst mode
Root bridge for: VLAN0070, VLAN0080
Port Type Default                        is disable
Edge Port [PortFast] BPDU Guard Default  is disabled
Edge Port [PortFast] BPDU Filter Default is disabled
Bridge Assurance                         is enabled
Loopguard Default                        is disabled
Pathcost method used                     is short

Name                     Blocking Listening Learning Forwarding STP Active
---------------------- -------- --------- -------- ---------- ----------
VLAN0001                      1         0        0          1          2
VLAN0070                      0         0        0          2          2
VLAN0080                      0         0        0          3          3
---------------------- -------- --------- -------- ---------- ----------
3 vlans                       1         0        0          6          7
```

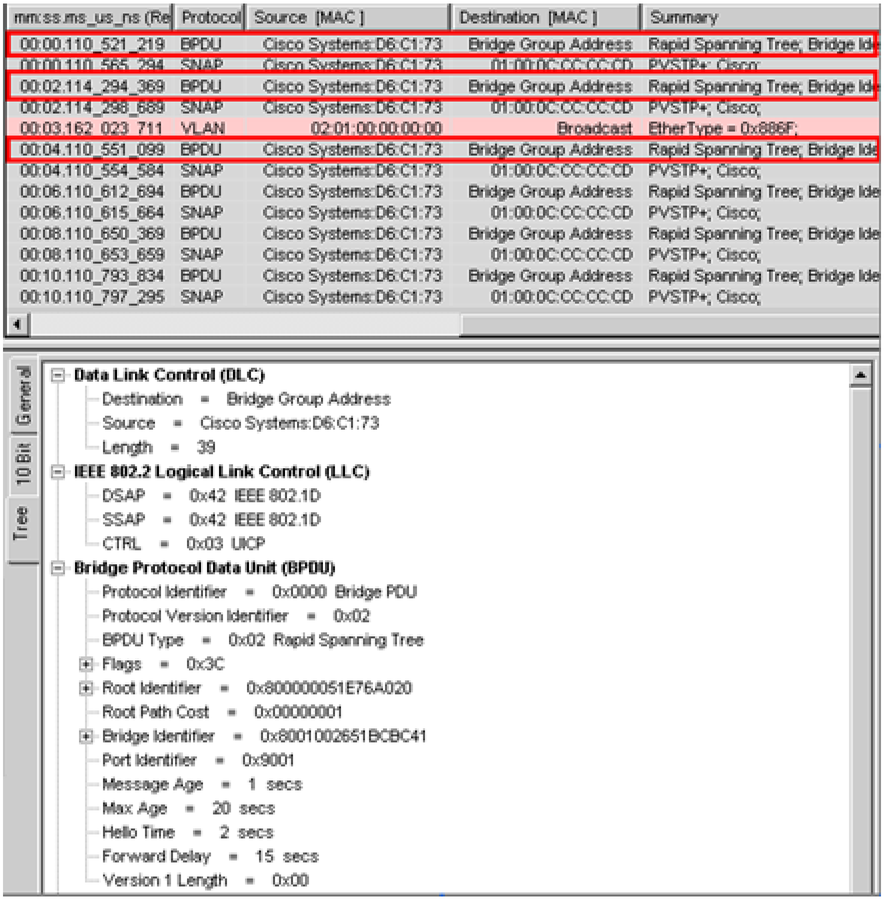
```
MP-8000B # show spanning-tree brief

 Spanning-tree Mode: Rapid Spanning Tree Protocol

      Root ID      Priority 32768
                   Address  0005.1e76.a020
                   Hello Time 2, Max Age 20, Forward Delay 15

      Bridge ID    Priority 32768
                   Address  0005.1e76.a020
                   Hello Time 2, Max Age 20, Forward Delay 15, Tx-HoldCount 6
                   Migrate Time 3 sec
<output truncated>
```

Using the following output examples, the root bridge is the Nexus 7000. The output shows that the switch with a bridge address of 0026.51bc.bc41 has a bridge priority of 24646, which is the lowest of the three switches. To see the STP bridge ID and bridge priority of the switch, use the NX-OS **show spanning-tree vlan** <*vlan id*> command in the Nexus Series switches and the CMSH **show spanning-tree brief** command in the MP-8000B, as shown in the following examples. For more information, refer to "Spanning Tree Protocol (STP)" on page 92.

```
 Nexus 7000 # show spanning-tree vlan 70

VLAN0070
  Spanning tree enabled protocol rstp
  Root ID    Priority    24646 --> same as the bridge priority of this switch,
                                   telling this switch is the root bridge
             Address     0026.51bc.bc41
             This bridge is the root
             Hello Time  2   sec   Max Age 20 sec   Forward Delay 15 sec

  Bridge ID  Priority    24646  (priority 24576 sys-id-ext 70)   --> bridge
                                                   priority   of this switch
             Address     0026.51bc.bc41       --> bridge ID of this switch
             Hello Time  2   sec   Max Age 20 sec   Forward Delay 15 sec
<output truncated>
```

```
MP-8000B # show spanning-tree brief

 Spanning-tree Mode: Rapid Spanning Tree Protocol

      Root ID      Priority 24577
                   Address  0026.51bc.bc41
                   Root Path Cost 2000
                   Root Port Id 8801 (Po 1)
                   Hello Time 2, Max Age 20, Forward Delay 15
      Bridge ID    Priority 32768 --> bridge priority of this switch
                   Address  0005.1e76.a020 --> bridge ID of this switch
                   Hello Time 2, Max Age 20, Forward Delay 15, Tx-HoldCount 6
                   Migrate Time 3 sec
<output truncated>
```

Figure 99 shows that the BPDUs are sent every two seconds, which is the default on both the Nexus 5010 and MP-8000B switches.



mm:ss.ms_us_ns (Re	Protocol	Source [MAC]	Destination [MAC]	Summary
00:00.110_521_219	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:00.110_565_294	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
00:02.114_294_369	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:02.114_298_689	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
00:03.162_023_711	VLAN	02:01:00:00:00:00	Broadcast	EtherType = 0x886F;
00:04.110_551_099	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:04.110_554_584	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
00:06.110_612_694	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:06.110_615_664	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
00:08.110_650_369	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:08.110_653_659	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
00:10.110_793_834	BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Ide
00:10.110_797_295	SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;

```
⊟ Data Link Control (DLC)
    Destination  =  Bridge Group Address
    Source  =  Cisco Systems:D6:C1:73
    Length  =  39
⊟ IEEE 802.2 Logical Link Control (LLC)
    DSAP  =  0x42  IEEE 802.1D
    SSAP  =  0x42  IEEE 802.1D
    CTRL  =  0x03  UICP
⊟ Bridge Protocol Data Unit (BPDU)
    Protocol Identifier  =  0x0000  Bridge PDU
    Protocol Version Identifier  =  0x02
    BPDU Type  =  0x02  Rapid Spanning Tree
  ⊞ Flags  =  0x3C
  ⊞ Root Identifier  =  0x800000051E76A020
    Root Path Cost  =  0x00000001
  ⊞ Bridge Identifier  =  0x8001002651BCBC41
    Port Identifier  =  0x9001
    Message Age  =  1  secs
    Max Age  =  20  secs
    Hello Time  =  2  secs
    Forward Delay  =  15  secs
    Version 1 Length  =  0x00
```

Figure 99 **BPDU information**

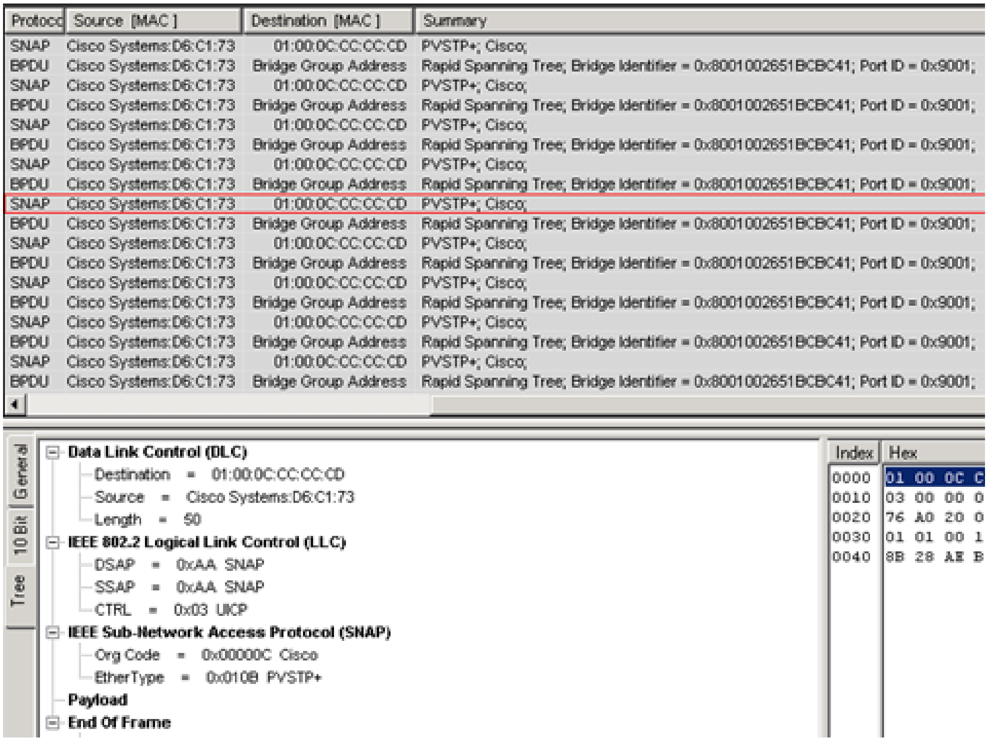
Debugging commands can also be used to examine if BPDUs are sent and received by the switch. Additional options to enable the debug commands on specific interfaces or specific VLANs are available.

For example:

```
Nexus 5010 # debug spanning-tree bpdu_rx ?
  <CR>
  interface  Enter interface
  tree       Spanning tree instance

Nexus 5010 # debug spanning-tree bpdu_rx tree ?
  <1-4095>  Enter tree ID (MST use 1-based tree id)

Nexus 5010 # debug spanning-tree bpdu_rx
interface   tree
Nexus 5010 # debug spanning-tree bpdu_rx
Nexus 5010 # 2009 Nov 20 06:29:28.622947 stp: BPDU RX: vb 1 vlan 0, ifi 0x5000000
           (mgmt0)
2009 Nov 20 06:29:29.586777 stp: BPDU RX: vb 1 vlan 1, ifi 0x16000001
   (port-channel2)
2009 Nov 20 06:29:29.586831 stp: BPDU Rx: Received BPDU on vb 1 vlan 1 port
   port-channel2 pkt_len 60 bpdu_len 36 netstack flags 0x00ed enc_type ieee
2009 Nov 20 06:29:29.586910 stp: RSTP(1): msg on port-channel2
2009 Nov 20 06:29:29.586935 stp:        : repeated-designated on port alternate
2009 Nov 20 06:29:29.586958 stp: RSTP(1): port-channel2 repeated msg
2009 Nov 20 06:29:29.586981 stp: RSTP(1): port-channel2 rcvd info remaining 6
2009 Nov 20 06:29:29.587095 stp: BPDU RX: vb 1 vlan 1, ifi 0x16000001
   (port-channel2)
2009 Nov 20 06:29:29.587122 stp: BPDU Rx: Received BPDU on vb 1 vlan 1 port
   port-channel2 pkt_len 64 bpdu_len 42 netstack flags 0x00ed enc_type sstp
2009 Nov 20 06:29:29.587170 stp: BPDU Rx: Dropping redundant SSTP packet received
   on port port-channel2 vlan VLAN0001
2009 Nov 20 06:29:29.587240 stp: BPDU RX: vb 1 vlan 1, ifi 0x16000000
   (port-channel1)
2009 Nov 20 06:29:29.587267 stp: BPDU Rx: Received BPDU on vb 1 vlan 1 port
   port-channel1 pkt_len 60 bpdu_len 36 netstack flags 0x00ed enc_type ieee
2009 Nov 20 06:29:29.587310 stp: RSTP(1): msg on port-channel1
2009 Nov 20 06:29:29.587334 stp:         : repeated-designated on port root
2009 Nov 20 06:29:29.587356 stp: RSTP(1): port-channel1 repeated msg
2009 Nov 20 06:29:29.587378 stp: RSTP(1): port-channel1 rcvd info remaining 6
2009 Nov 20 06:29:29.587446 stp: BPDU RX: vb 1 vlan 1, ifi 0x16000000
   (port-channel1)
2009 Nov 20 06:29:29.587473 stp: BPDU Rx: Received BPDU on vb 1 vlan 1 port
   port-channel1 pkt_len 64 bpdu_len 42 netstack flags 0x00ed enc_type sstp
2009 Nov 20 06:29:29.587518 stp: BPDU Rx: Dropping redundant SSTP packet received
   on port port-channel1 vlan VLAN0001
2009 Nov 20 06:29:30.700193 stp: BPDU RX: vb 1 vlan 0, ifi 0x5000000 (mgmt0)
2009 Nov 20 06:29:31.586058 stp: BPDU RX: vb 1 vlan 1, ifi 0x16000001
   (port-channel2)
2009 Nov 20 06:29:31.586113 stp: BPDU Rx: Received BPDU on vb 1 vlan 1 port
   port-channel2 pkt_len 60 bpdu_len 36 netstack flags 0x00ed enc_type ieee

MP-8000B # debug spanning-tree bpdu ?
  rx  Receive
  tx  Transmit
```

```
MP-8000B # debug spanning-tree bpdu tx ?
  all         All Interface
  interface   Interface information

MP-8000B # debug spanning-tree bpdu rx interface ?
  port-channel        Port-channel interface
  tengigabitethernet  TenGigabit Ethernet interface
```

If you see that BPDUs are sent and received by the switches, you should also be able to see that STP convergence is taking place accordingly. If you place an analyzer between two switches, you should see an exchange of STP information and that root bridge election is taking place. Figure 100 shows the actual STP frames. For more information about STP convergence or STP path cost, refer to “Spanning Tree Protocol (STP)” on page 92.



Protocc	Source [MAC]	Destination [MAC]	Summary
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;
SNAP	Cisco Systems:D6:C1:73	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Cisco Systems:D6:C1:73	Bridge Group Address	Rapid Spanning Tree; Bridge Identifier = 0x8001002651BCBC41; Port ID = 0x9001;

```
Data Link Control (DLC)
    Destination  =  01:00:0C:CC:CC:CD
    Source  =  Cisco Systems:D6:C1:73
    Length  =  50
IEEE 802.2 Logical Link Control (LLC)
    DSAP  =  0xAA  SNAP
    SSAP  =  0xAA  SNAP
    CTRL  =  0x03  UICP
IEEE Sub-Network Access Protocol (SNAP)
    Org Code  =  0x00000C  Cisco
    EtherType  =  0x010B  PVSTP+
Payload
End Of Frame
```

```
Index  Hex
0000   01 00 0C C
0010   03 00 00 0
0020   76 A0 20 0
0030   01 01 00 1
0040   8B 28 AE B
```

Figure 100 STP frames

The following is an example of the **show run interface** <*intf type*> command. This command is used to verify the channel-group configuration.

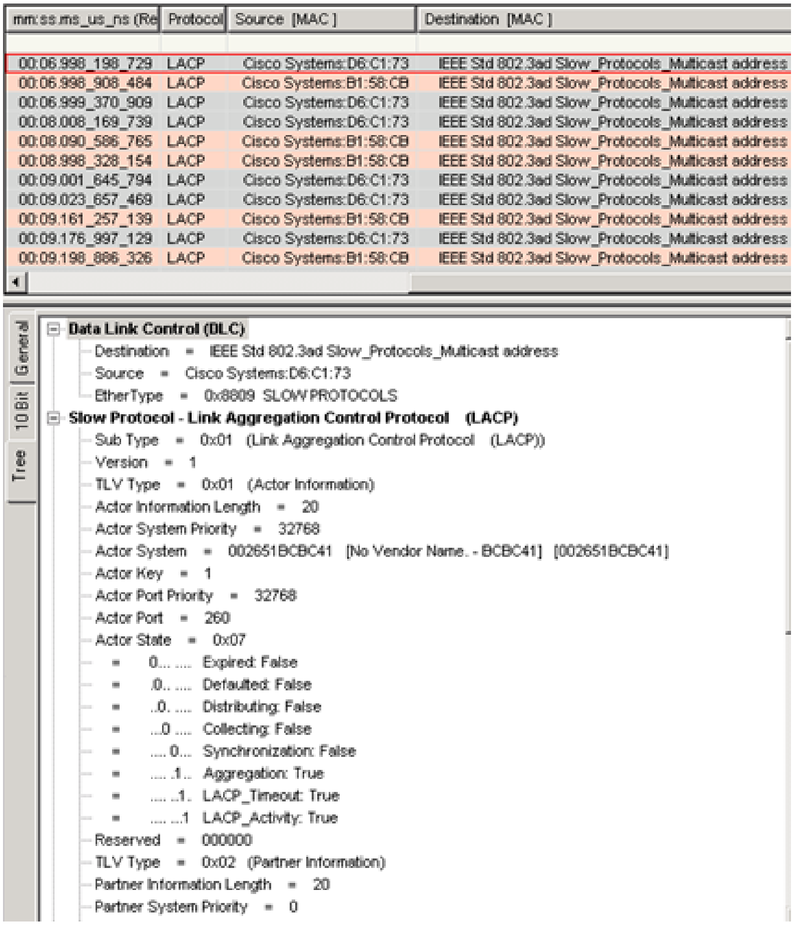
```
 Nexus 7000 # show run int ethernet 1/3
version 4.1(4)

interface Ethernet1/3
  switchport
  switchport mode trunk
  channel-group 2 mode active
  no shutdown

 Nexus 7000 # show run int ethernet 1/4
version 4.1(4)

interface Ethernet1/4
  switchport
  switchport mode trunk
  channel-group 2 mode active
  no shutdown

Nexus 5010 # show run int ethernet 1/3
version 4.1(3)N1(1a)

interface Ethernet1/3
  switchport mode trunk
  channel-group 2 mode passive

Nexus 5010 # show run int ethernet 1/4
version 4.1(3)N1(1a)

interface Ethernet1/4
  switchport mode trunk
  channel-group 2 mode passive
```

The **show interface** <*intf type*> **counters** command output example below shows the received multicast LACP frames (notice the InMcastPkts and OutMcastPkts values) from the other end of the link. These should be showing incremental values as you enable the ports to form a dynamic port channel.

```
Nexus 5010 # show interface port-channel 1 counters

-----------------------------------------------------------------------------
Port                InOctets       InUcastPkts       InMcastPkts       InBcastPkts
-----------------------------------------------------------------------------
Po1                 50303243            317163            322005       8344
-----------------------------------------------------------------------------
Port               OutOctets      OutUcastPkts      OutMcastPkts      OutBcastPkts
-----------------------------------------------------------------------------
Po1               5433936846           3632010            296852       18243
```

Note that if you configure an active-passive pair for dynamic port-channel, the first one that will send the multicast LACP frame is the port configured as *active*. Based from above port-channel configuration, you will see that the active port ethernet 1/4 (with source MAC 0023.ebd6.c173) of the Nexus 7000 sent the LACP frame first, and then the passive port ethernet 1/4 (with source MAC 000d.ecb1.58cb) of the Nexus 5010 responded accordingly, as shown in the captured trace in Figure 101 on page 276.

```
EX-7K # show interface ethernet 1/4 | i bia
  Hardware: 10000 Ethernet, address: 0023.ebd6.c173 (bia 0023.ebd6.c173)

Nexus 5010 # show interface ethernet 1/4 | i bia
  Hardware: 1000/10000 Ethernet, address: 000d.ecb1.58cb (bia 000d.ecb1.58cb)
```



mm:ss.ms_us_ns (Re	Protocol	Source [MAC]	Destination [MAC]
00:06.998_198_729	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:06.998_908_484	LACP	Cisco Systems:B1:58:CB	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:06.999_370_909	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:08.008_169_739	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:08.090_586_765	LACP	Cisco Systems:B1:58:CB	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:08.998_328_154	LACP	Cisco Systems:B1:58:CB	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:09.001_645_794	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:09.023_657_469	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:09.161_257_139	LACP	Cisco Systems:B1:58:CB	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:09.176_997_129	LACP	Cisco Systems:D6:C1:73	IEEE Std 802.3ad Slow_Protocols_Multicast address
00:09.198_886_326	LACP	Cisco Systems:B1:58:CB	IEEE Std 802.3ad Slow_Protocols_Multicast address

```
⊟ Data Link Control (DLC)
    Destination  =  IEEE Std 802.3ad Slow_Protocols_Multicast address
    Source  =  Cisco Systems:D6:C1:73
    EtherType  =  0x8809  SLOW PROTOCOLS
⊟ Slow Protocol - Link Aggregation Control Protocol   (LACP)
    Sub Type  =  0x01   (Link Aggregation Control Protocol   (LACP))
    Version  =  1
    TLV Type  =  0x01   (Actor Information)
    Actor Information Length  =  20
    Actor System Priority  =  32768
    Actor System  =  002651BCBC41   [No Vendor Name. - BCBC41]   [002651BCBC41]
    Actor Key  =  1
    Actor Port Priority  =  32768
    Actor Port  =  260
    Actor State  =  0x07
        =    0... ....  Expired: False
        =    .0.. ....  Defaulted: False
        =    ..0. ....  Distributing: False
        =    ...0 ....  Collecting: False
        =    .... 0...  Synchronization: False
        =    .... .1..  Aggregation: True
        =    .... ..1.  LACP_Timeout: True
        =    .... ...1  LACP_Activity: True
    Reserved  =  000000
    TLV Type  =  0x02   (Partner Information)
    Partner Information Length  =  20
    Partner System Priority  =  0
```

Figure 101   LACP frames

From the following example of the **show vlan brief** command, you can see the VLANs that are created in the switch, as well as the VLAN assignment to each access ports.

```
MP-8000B # show vlan brief

VLAN     Name               State    Ports
                                     (u)-Untagged, (t)-Tagged
                                     (c)-Converged
====  ====            =====      ======================
1        default            ACTIVE   Te 0/4(c) Te 0/5(c) Te 0/6(c)
                                     Te 0/7(c) Te 0/9(c) Te 0/10(c)
                                     Te 0/11(c) Te 0/12(c) Te 0/13(c)
```

```
                                 Te 0/14(c) Te 0/15(c) Te 0/17(c)
                                 Te 0/18(c) Te 0/19(c) Te 0/21(c)
                                 Te 0/22(u) Te 0/23(u) Po 1(t)
                                 Po 2(t)
10      VLAN0010         ACTIVE  Te 0/16(u) Te 0/20(u) Po 1(t)
                                 Po 2(t)
70      VLAN0070         ACTIVE  Te 0/8(u) Po 1(t) Po 2(t)
80      VLAN0080         ACTIVE  Po 1(t) Po 2(t)
```

You can also use **show interface** <*intf type*> **switchport** command (as shown next) to make sure that L2 trunk ports are configured to allow the VLAN number assigned to the access port. The example below shows that VLAN 1, VLAN 70, and VLAN 80 are allowed to cross the L2 trunk port, port-channel 1.

```
Nexus 5010 # show interface port-channel 1 switchport
Name: port-channel1
  Switchport: Enabled
  Switchport Monitor: Not enabled
  Operational Mode: trunk
  Access Mode VLAN: 1 (default)
  Trunking Native Mode VLAN: 1 (default)
  Trunking VLANs Enabled: 1-3967,4048-4093
  Administrative private-vlan primary host-association: none
  Administrative private-vlan secondary host-association: none
  Administrative private-vlan primary mapping: none
  Administrative private-vlan secondary mapping: none
  Administrative private-vlan trunk native VLAN: none
  Administrative private-vlan trunk encapsulation: dot1q
  Administrative private-vlan trunk normal VLANs: none
  Administrative private-vlan trunk private VLANs: none
  Operational private-vlan: none
```

In the following example, you can see the native VLAN is VLAN 1. Assuming no one changes the default values, the expected output will be similar to that shown in the next example, which shows VLAN 1 as the native VLAN.

```
Nexus 5010 # show interface port-channel 1 switchport
Name: port-channel1
  Switchport: Enabled
  Switchport Monitor: Not enabled
  Operational Mode: trunk
  Access Mode VLAN: 1 (default)
  Trunking Native Mode VLAN: 1 (default)
  Trunking VLANs Enabled: 1-3967,4048-4093
  Administrative private-vlan primary host-association: none
  Administrative private-vlan secondary host-association: none

<output truncated>
```

If the Native VLAN is different than the one on connected switch (usually it is VLAN 1), then you need to change it back to the correct VLAN. To change the native VLAN of a trunk port, use the CLI **switchport trunk native vlan** <*vlan#*> command, as shown in the next example. Ensure that both ends of the trunk port are configured with the same native VLAN.

```
Nexus 5010 (config-if)# int port-channel 2
Nexus 5010 (config-if)# switchport trunk native vlan 1
```

The following is the error message that will display if one end of the trunk is using a different native VLAN number:

```
Nexus 5010 # show interface port-channel 1 switchpo2009 Nov 20 07:39:15 Nexus 5010
   %STP-2-BLOCK_PVID_PEER: Blocking port-channel2 on VLAN0001. Inconsistent peer
   vlan.
2009 Nov 20 07:39:15 Nexus 5010 %STP-2-BLOCK_PVID_LOCAL: Blocking port-channel2
   on VLAN0002. Inconsistent local vlan.
```

IMPORTANT

It is extremely important to have native VLAN on trunk ports because this VLAN carries control protocols such as STP, VTP, CDP, LACP, and so on.

Next step

Once Layer 1 and Layer 2 have been verified as okay, then proceed to flowchart step #3.

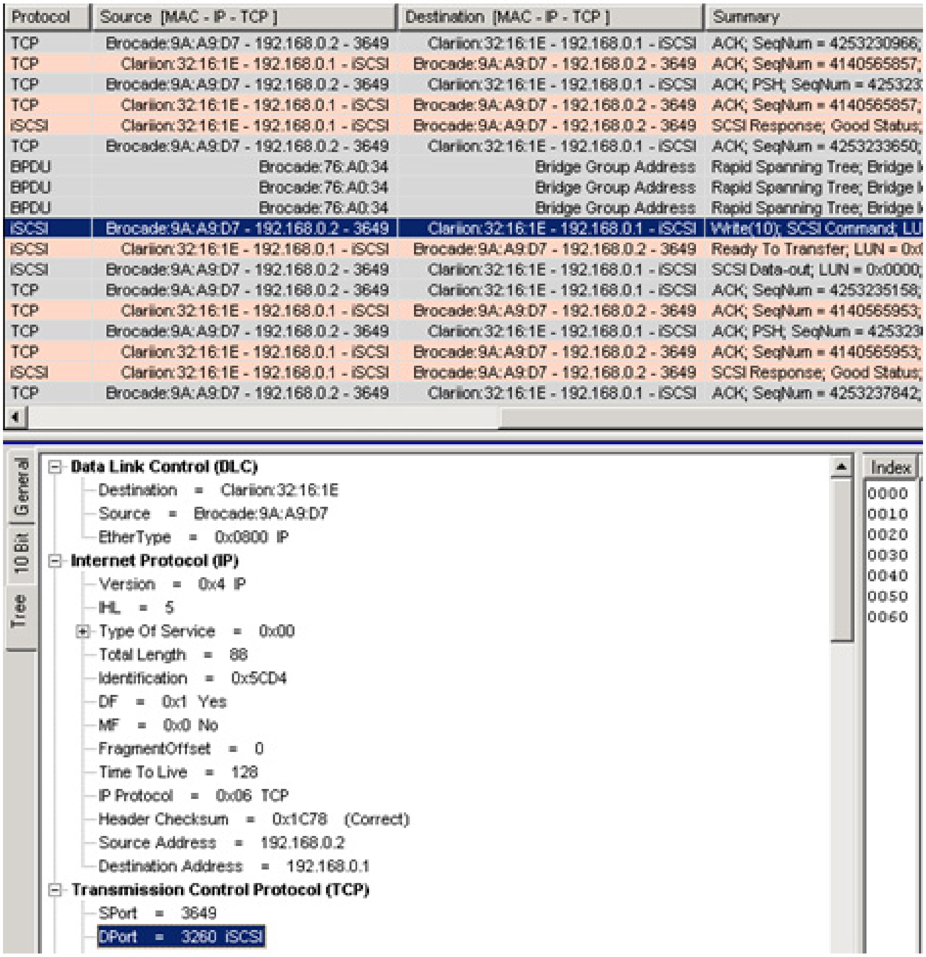
Flowchart step #5, Can the file server ping the iSCSI storage array?

Troubleshooting

From the iSCSI LAN, verify that the file server 192.168.0.2 (note that this is also the secondary LAN IP of Host A) can ping the VNX series or CLARiiON IP address 192.168.0.1. Part of this step requires that you verify whether iSCSI traffic is passing across the L2 path (the path from file server (192.168.0.2) to the VNX series or CLARiiON (192.168.0.1), and vice versa. You can verify this by checking if iSCSI TCP ports (typically TCP ports 860 and 3260) are passing through across the link, as shown in the trace captures shown in Figure 102 on page 279. Ensure there is no ACL or firewall (stand-alone or host-based) blocking these ports.

Example and interpretation of the results

The trace in Figure 102 shows that iSCSI traffic is allowed to pass in the network.



Protocol	Source [MAC - IP - TCP]	Destination [MAC - IP - TCP]	Summary
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; SeqNum = 4253230966;
TCP	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	ACK; SeqNum = 4140565857;
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; PSH; SeqNum = 425323
TCP	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	ACK; SeqNum = 4140565857;
iSCSI	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	SCSI Response; Good Status;
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; SeqNum = 4253233650;
BPDU	Brocade:76:A0:34	Bridge Group Address	Rapid Spanning Tree; Bridge I
BPDU	Brocade:76:A0:34	Bridge Group Address	Rapid Spanning Tree; Bridge I
BPDU	Brocade:76:A0:34	Bridge Group Address	Rapid Spanning Tree; Bridge I
iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Write(10); SCSI Command; LU
iSCSI	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Ready To Transfer; LUN = 0x0
iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	SCSI Data-out; LUN = 0x0000;
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; SeqNum = 4253235158;
TCP	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	ACK; SeqNum = 4140565953;
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; PSH; SeqNum = 425323
TCP	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	ACK; SeqNum = 4140565953;
iSCSI	Clariion:32:16:1E - 192.168.0.1 - iSCSI	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	SCSI Response; Good Status;
TCP	Brocade:9A:A9:D7 - 192.168.0.2 - 3649	Clariion:32:16:1E - 192.168.0.1 - iSCSI	ACK; SeqNum = 4253237842;

```
⊟ Data Link Control (DLC)
    Destination  =  Clariion:32:16:1E
    Source  =  Brocade:9A:A9:D7
    EtherType  =  0x0800  IP
⊟ Internet Protocol (IP)
    Version  =  0x4  IP
    IHL  =  5
  ⊞ Type Of Service  =  0x00
    Total Length  =  88
    Identification  =  0x5CD4
    DF  =  0x1  Yes
    MF  =  0x0  No
    FragmentOffset  =  0
    Time To Live  =  128
    IP Protocol  =  0x06  TCP
    Header Checksum  =  0x1C78  (Correct)
    Source Address  =  192.168.0.2
    Destination Address  =  192.168.0.1
⊟ Transmission Control Protocol (TCP)
    SPort  =  3649
    DPort  =  3260  iSCSI
```

Figure 102 iSCSI traffic

For a ping example, refer to "Flowchart step #2, Can Windows client #1 ping the file server's IP address?" on page 256.

Next steps

- If ping is successful, proceed with flowchart step #6.
- If the file server cannot ping the VNX series or CLARiiON IP address, then go to flowchart step #3.

Flowchart step #6, Can the file server see the storage array LUNs?

Troubleshooting

At this step, verify whether the file server can see the VNX series or CLARiiON LUNs. The diskpart or inq utility tool can be used on a Windows host.

Example and interpretation of the results

Using the **inq** command, the following example shows that VNX series or CLARiiON LUNs are not present in the file server (Host A, 10.0.70.3).

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.

.....

-------------------------------------------------------------------------
DEVICE             :VEND    :PROD            :REV   :SER NUM    :CAP(kb)
-------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE1 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
```

Next steps

- If you do not see the LUNs, then proceed to flowchart step #7.
- If LUNs are visible on the file server (Host A, 192.168.0.2), then proceed to flowchart step #8.

Flowchart step #7, Check the storage array's iSCSI configuration

Troubleshooting

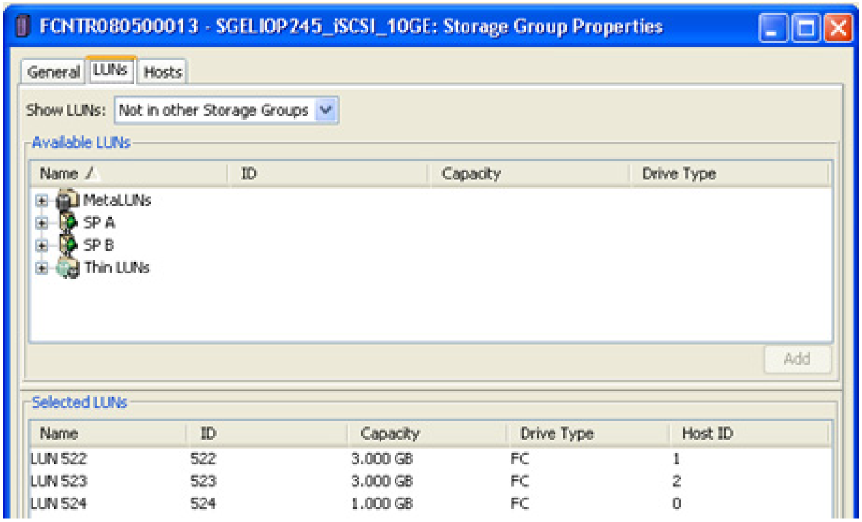
You arrived at this step because you are still unable to access the shared folder on the file server (Host A), even though Windows client #1 (10.0.80.2) can ping the file server (10.0.70.3) and the file server (192.168.0.2) can ping the VNX series or CLARiiON iSCSI port (192.168.0.2).

At this stage, you want to confirm that the storage array configuration has been put in place and is configured correctly. Follow the guidelines listed next to examine and troubleshoot Storage Array's iSCSI configuration.

1. Check the LUN masking configuration. Ensure that you already assigned all the LUNs you need to the storage group you created.
2. Verify the host assignment. Ensure that you assigned the correct host to the LUNs that you selected. In VNX series or CLARiiON, this can be configured and verified in the same window shown in Figure 103 under the **Hosts** tab, as shown in Figure 104 on page 282.

Example and interpretation of the results

Figure 103 shows three LUNs assigned to the SGELIOP245_iSCSI_10GE storage group. This verifies that LUN masking was already configured.

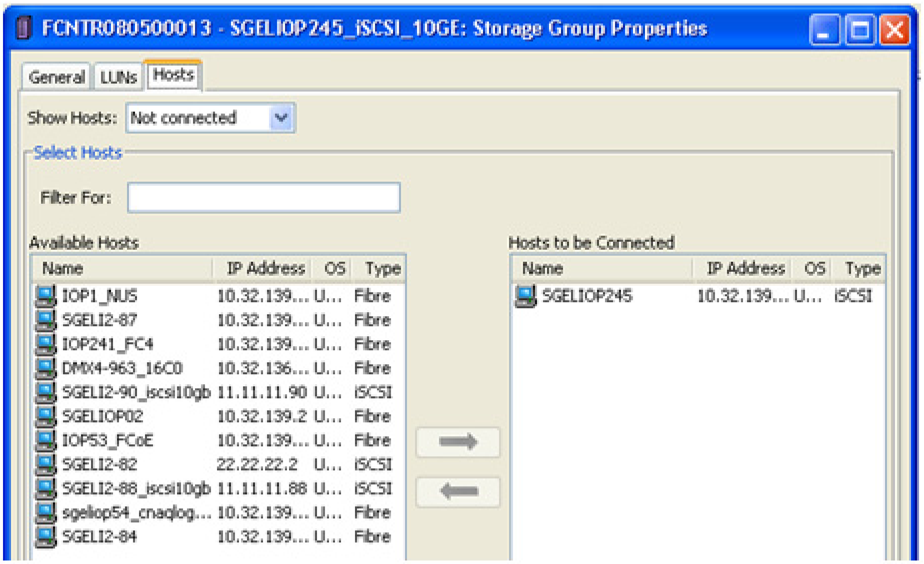


Show LUNs: Not in other Storage Groups

Name	ID	Capacity	Drive Type	Host ID
LUN 522	522	3.000 GB	FC	1
LUN 523	523	3.000 GB	FC	2
LUN 524	524	1.000 GB	FC	0

Figure 103 Verify LUNs

Figure 104 shows the host SGELIOP245 is assigned to the SGELIOP245_iSCSI_10GE storage group you created.



FCNTR080500013 - SGELIOP245_iSCSI_10GE: Storage Group Properties

Show Hosts: Not connected

Select Hosts

Filter For:

Available Hosts

Name	IP Address	OS	Type
IOP1_NUS	10.32.139...	U...	Fibre
SGELI2-87	10.32.139...	U...	Fibre
IOP241_FC4	10.32.139...	U...	Fibre
DMX4-963_16C0	10.32.136...	U...	Fibre
SGELI2-90_iscsi10gb	11.11.11.90	U...	iSCSI
SGELIOP02	10.32.139.2	U...	Fibre
IOP53_FCoE	10.32.139...	U...	Fibre
SGELI2-82	22.22.22.2	U...	iSCSI
SGELI2-88_iscsi10gb	11.11.11.88	U...	iSCSI
sgeliop54_cnaqlog...	10.32.139...	U...	Fibre
SGELI2-84	10.32.139...	U...	Fibre

Hosts to be Connected

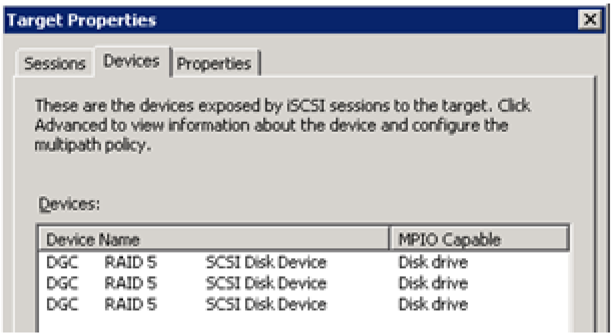
Name	IP Address	OS	Type
SGELIOP245	10.32.139...	U...	iSCSI

Figure 104 Verify host assignment

If everything is correct on the VNX series or CLARiiON side, you should be seeing the LUNs presented to the file server. If using the **inq** command, the output should look similar to the following example:

```
F:\copa>inq
Inquiry utility, Version V7.1-131 (Rev 1.0)      (SIL Version V4.1-131)
Copyright (C) by EMC Corporation, all rights reserved.
For help type inq -h.

.....

-------------------------------------------------------------------------
DEVICE             :VEND    :PROD            :REV   :SER NUM    :CAP(kb)
-------------------------------------------------------------------------
\\.\PHYSICALDRIVE0 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE1 :ATA     :WDC WD1602ABKS-1:3B04  :     WD-   :156250000
\\.\PHYSICALDRIVE2 :DGC     :RAID 5          :0429  :0C0000E1   :1048576
\\.\PHYSICALDRIVE3 :DGC     :RAID 5          :0429  :0A000097   :3145728
\\.\PHYSICALDRIVE4 :DGC     :RAID 5          :0429  :0B000097   :3145728
```

This information can also be seen using the iSCSI initiator tool in the Control Panel, under the connected target details devices, as shown in Figure 105.



These are the devices exposed by iSCSI sessions to the target. Click Advanced to view information about the device and configure the multipath policy.

Devices:

Device Name	MPIO Capable
DGC RAID 5 SCSI Disk Device	Disk drive
DGC RAID 5 SCSI Disk Device	Disk drive
DGC RAID 5 SCSI Disk Device	Disk drive

Figure 105 Target Properties, Devices tab

If you need to delve deeper into storage array troubleshooting and want to gather system and engineering level information, the following are some of the troubleshooting tools that can be used for the VNX series or CLARiiON:

- SP_Collect (Storage processor collection tool)
- SPLAT (SP Log Analysis Tool)
- CAP2 (CLARiiON Array Properties)
- Admintool
- Psmtool (persistent storage manager tool)
- Ktcons (K10 trace console)
- Flarecons (flare console)

Next step

Once the Storage Array's iSCSI configuration is confirmed correct, proceed to flowchart step #6.

Flowchart step #8, Is Windows client # 1 still unable to access the shared folder in the file server?

Troubleshooting

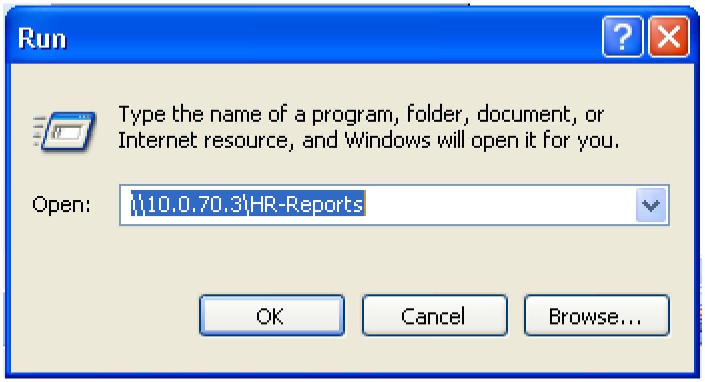
After you have completed one or all the previous steps, you need to again verify whether you can or cannot access the shared folder in the file server (Host A, 10.0.70.3). The file share can be accessed using your host's file manager like Windows Explorer in Windows. You can also type the shared folder path in the **Run** option in Windows. An example is provided in the following section.

Example and interpretation of the results

In Windows, accessing shared folder path can be done by putting the exact path in the **Run** option from **Start**.

\\File-Server-IP\shared-folder-name

Figure 106 shows an example on how to access a shared folder name HR-Reports in a Windows file server address 10.0.70.3.



Type the name of a program, folder, document, or Internet resource, and Windows will open it for you.

Open: \\10.0.70.3\HR-Reports

Figure 106 Run option

Next steps

- If the Windows client # 1 can access the shared folder in the File Server then the problem is fixed, as shown in flowchart step #10.
- If you are still unable to see the shared folder, then proceed with flowchart step #9.

Flowchart step #9, Check the host

Troubleshooting

At this stage, you need to verify that Host A's configuration and settings are correct. Follow the guidelines below when examining the Host A (file server).

- When performing host troubleshooting, first check the file server status. Ensure that the file server service is ON by completing the following steps:

  a. From **Start**, select **Run**.

  b. Type the **services.msc** command.

  c. Double-click the server service in the services list.

     The **Server Properties** dialog box displays, as shown in Figure 107 on page 287.

- In order to share the storage over the corporate network, you must ensure that Host A has performed login to the target device and that it is able to see the VNX series or CLARiiON LUNs, as shown in Figure 108 on page 288.

- If security features such as CHAP or IPsec are used on this iSCSI setup, ensure that the configuration parameters are correct, not only on the host, but also on the VNX series or CLARiiON.

- Also check the IQN name. The Microsoft iSCSI initiator service will automatically choose an IQN name based on the Windows computer and domain name and the microsoft.com domain name address. If the Windows computer or domain name is changed, then the IQN name will also change. However, an IQN name can be specifically changed to use a fixed value instead of the generated IQN name. If the administrator specifies a fixed IQN name, that name must be maintained as world wide unique. For more information about Microsoft iSCSI features and iSCSI initiator troubleshooting, see the *Microsoft iSCSI Software Initiator 2.x Users Guide* at http://www.microsoft.com.

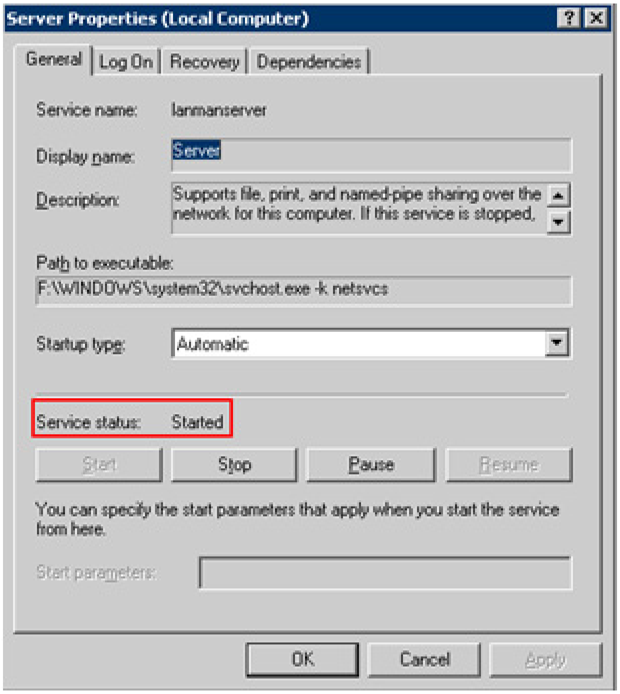
- Once the application layer has been checked and you still have a problem, verify whether the CIFS (SMB) traffic is passing through the corporate network. Ensure there is no firewall (standalone or host-based) blocking the UDP/TCP ports used by the CIFS service. Figure 109 on page 289 shows that the CIFS (SMB) packets are not blocked and are successfully passing from Host (10.0.70.3) to Windows client #1 (10.0.80.2) with a source port

445/destination port 1556 for SMB. For more information about CIFS, refer to online resources at http://msdn.microsoft.com and http://technet.microsoft.com.

- Also check the host's ability to mount the LUNs/devices. Again, vendor documentation and release notes are vital to troubleshooting.
- Ensure that hardware bus rescan or device discovery was tried. If the issue is still unresolved, rebooting the host can sometimes resolve the issue.
- Additionally, different host troubleshooting tools can be used for gathering OS system and storage array engineering information. Tools like EMCGrab and EMC Reports can be used in conjunction with HEAT to check information relative to the latest *EMC Support Matrix* (ESM). Some technical documentation available on the EMC Online Support website at https://support.emc.com, such as the *iSCSI Server Setup Guide for Windows*, can be used as a reference when dealing with host or OS level issues.
- If the issue is still unresolved, contact the host vendor.

Example and interpretation of the results

Figure 107 shows that file server service is enabled, which means that CIFS is working.



Server Properties (Local Computer)

Service name: lanmanserver

Display name: Server

Description: Supports file, print, and named-pipe sharing over the network for this computer. If this service is stopped,

Path to executable:
F:\WINDOWS\system32\svchost.exe -k netsvcs

Startup type: Automatic

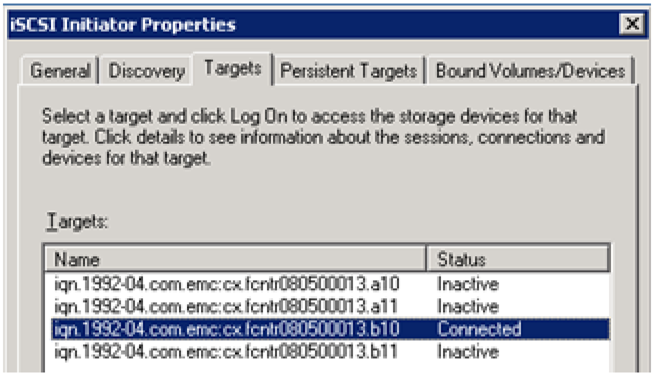
Service status: Started

You can specify the start parameters that apply when you start the service from here.

Start parameters:

Figure 107 Verify Service status

Figure 108 shows that Host A has performed login to the iSCSI target device and it was able to see VNX series or CLARiiON LUNs.



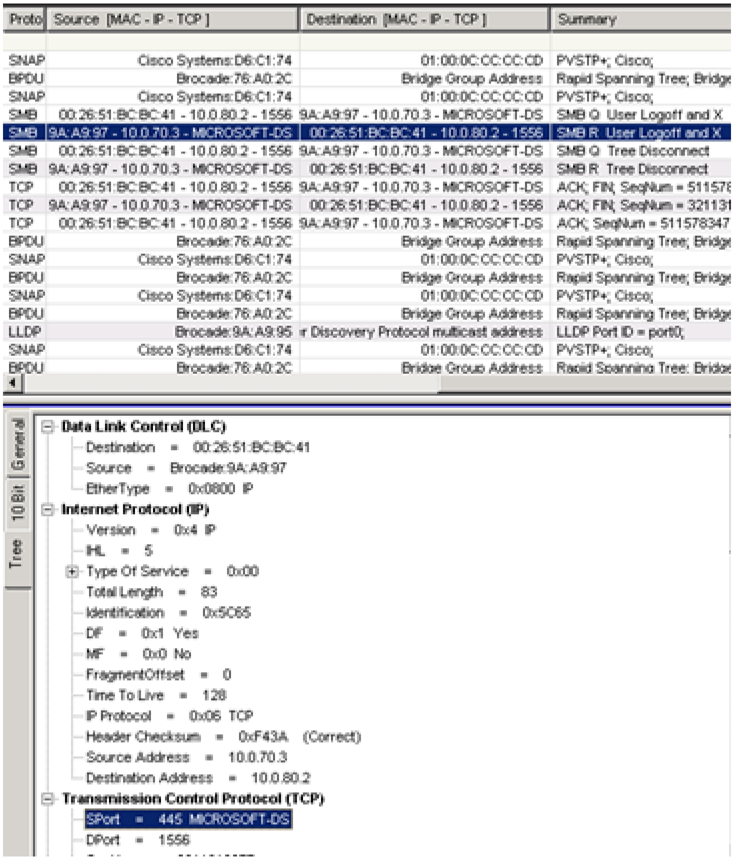
Select a target and click Log On to access the storage devices for that target. Click details to see information about the sessions, connections and devices for that target.

Targets:

Name	Status
iqn.1992-04.com.emc:cx.fcntr080500013.a10	Inactive
iqn.1992-04.com.emc:cx.fcntr080500013.a11	Inactive
iqn.1992-04.com.emc:cx.fcntr080500013.b10	Connected
iqn.1992-04.com.emc:cx.fcntr080500013.b11	Inactive

Figure 108 iSCSI Initiator Properties dialog box, Targets tab

Figure 109 shows that the CIFS (SMB) packets are not blocked and are successfully passing from Host (10.0.70.3) to Windows client #1 (10.0.80.2) with a source port 445/destination port 1556 for SMB.



Proto	Source [MAC - IP - TCP]	Destination [MAC - IP - TCP]	Summary
SNAP	Cisco Systems:D6:C1:74	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Brocade:76:A0:2C	Bridge Group Address	Rapid Spanning Tree; Bridge
SNAP	Cisco Systems:D6:C1:74	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
SMB	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	SMB Q User Logoff and X
SMB	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	SMB R User Logoff and X
SMB	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	SMB Q Tree Disconnect
SMB	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	SMB R Tree Disconnect
TCP	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	ACK; FIN; SeqNum = 511578
TCP	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	ACK; FIN; SeqNum = 321131
TCP	00:26:51:BC:BC:41 - 10.0.80.2 - 1556	9A:A9:97 - 10.0.70.3 - MICROSOFT-DS	ACK; SeqNum = 511578347
BPDU	Brocade:76:A0:2C	Bridge Group Address	Rapid Spanning Tree; Bridge
SNAP	Cisco Systems:D6:C1:74	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Brocade:76:A0:2C	Bridge Group Address	Rapid Spanning Tree; Bridge
SNAP	Cisco Systems:D6:C1:74	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Brocade:76:A0:2C	Bridge Group Address	Rapid Spanning Tree; Bridge
LLDP	Brocade:9A:A9:95	r Discovery Protocol multicast address	LLDP Port ID = port0;
SNAP	Cisco Systems:D6:C1:74	01:00:0C:CC:CC:CD	PVSTP+; Cisco;
BPDU	Brocade:76:A0:2C	Bridge Group Address	Rapid Spanning Tree; Bridge

```
⊟ Data Link Control (DLC)
    Destination  =  00:26:51:BC:BC:41
    Source  =  Brocade:9A:A9:97
    EtherType  =  0x0800  IP
⊟ Internet Protocol (IP)
    Version  =  0x4  IP
    IHL  =  5
  ⊞ Type Of Service  =  0x00
    Total Length  =  83
    Identification  =  0x5C65
    DF  =  0x1  Yes
    MF  =  0x0  No
    FragmentOffset  =  0
    Time To Live  =  128
    IP Protocol  =  0x06  TCP
    Header Checksum  =  0xF43A  (Correct)
    Source Address  =  10.0.70.3
    Destination Address  =  10.0.80.2
⊟ Transmission Control Protocol (TCP)
    SPort  =  445  MICROSOFT-DS
    DPort  =  1556
```

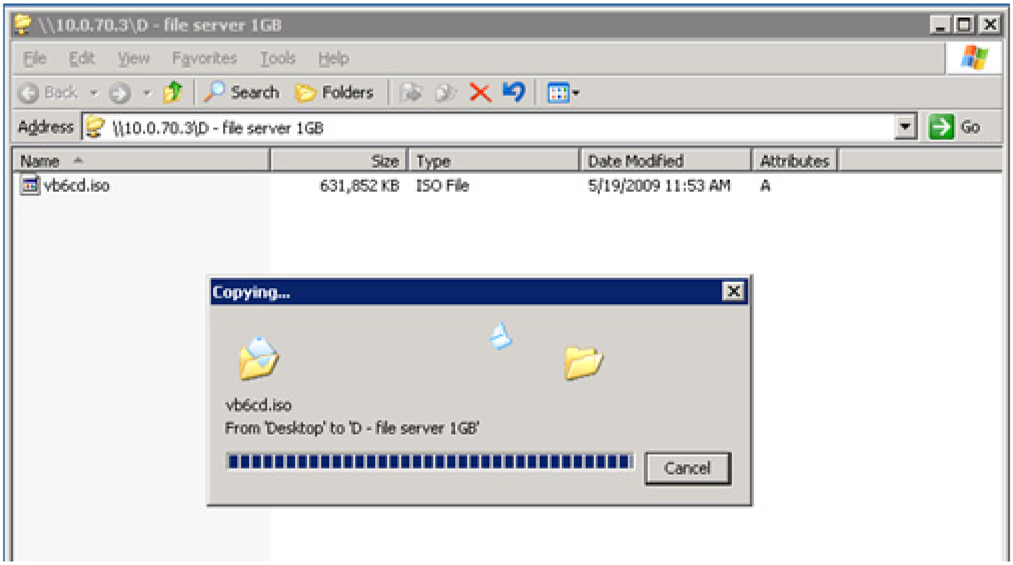
Figure 109 Verify CIFS (SMB) traffic

Next step

Once the Host A (the file server) has been examined and verified that configuration and settings are correct, then proceed again to flowchart step #8.

Flowchart step #10, Problem is solved

You arrived at this step because you have verified that the issue is resolved. Using the Windows file server service, you can verify if the file transfer to the file server is successful, as shown in Figure 110.



\\10.0.70.3\D - file server 1GB

Address \\10.0.70.3\D - file server 1GB

Name	Size	Type	Date Modified	Attributes
vb6cd.iso	631,852 KB	ISO File	5/19/2009 11:53 AM	A

Copying...

vb6cd.iso
From 'Desktop' to 'D - file server 1GB'

Figure 110 Verify transfer is successful

Glossary

This glossary contains terms related to EMC products and EMC networked storage concepts.

access control A service that allows or prohibits access to a resource. Storage management products implement access control to allow or prohibit specific users. Storage platform products implement access control, often called LUN Masking, to allow or prohibit access to volumes by Initiators (HBAs). *See also* "persistent binding" and "zoning."

active domain ID The domain ID actively being used by a switch. It is assigned to a switch by the principal switch.

active zone set The active zone set is the zone set definition currently in effect and enforced by the fabric or other entity (for example, the name server). Only one zone set at a time can be active.

agent An autonomous agent is a system situated within (and is part of) an environment that senses that environment, and acts on it over time in pursuit of its own agenda. Storage management software centralizes the control and monitoring of highly distributed storage infrastructure. The centralizing part of the software management system can depend on agents that are installed on the distributed parts of the infrastructure. For example, an agent (software component) can be installed on each of the hosts (servers) in an environment to allow the centralizing software to control and monitor the hosts.

alarm An SNMP message notifying an operator of a network problem.

any-to-any port connectivity

A characteristic of a Fibre Channel switch that allows any port on the switch to communicate with any other port on the same switch.

application

Application software is a defined subclass of computer software that employs the capabilities of a computer directly to a task that users want to perform. This is in contrast to system software that participates with integration of various capabilities of a computer, and typically does not directly apply these capabilities to performing tasks that benefit users. The term application refers to both the application software and its implementation which often refers to the use of an information processing system. (For example, a payroll application, an airline reservation application, or a network application.) Typically an application is installed “on top of” an operating system like Windows or LINUX, and contains a user interface.

application-specific integrated circuit (ASIC)

A circuit designed for a specific purpose, such as implementing lower-layer Fibre Channel protocols (FC-1 and FC-0). ASICs contrast with general-purpose devices such as memory chips or microprocessors, which can be used in many different applications.

arbitration

The process of selecting one respondent from a collection of several candidates that request service concurrently.

ASIC family

Different switch hardware platforms that utilize the same port ASIC can be grouped into collections known as an ASIC family. For example, the Fuji ASIC family which consists of the ED-64M and ED-140M run different microprocessors, but both utilize the same port ASIC to provide Fibre Channel connectivity, and are therefore in the same ASIC family. For inter operability concerns, it is useful to understand to which ASIC family a switch belongs.

ASCII

ASCII (American Standard Code for Information Interchange), generally pronounced [aeski], is a character encoding based on the English alphabet. ASCII codes represent text in computers, communications equipment, and other devices that work with text. Most modern character encodings, which support many more characters, have a historical basis in ASCII.

audit log

A log containing summaries of actions taken by a Connectrix Management software user that creates an audit trail of changes. Adding, modifying, or deleting user or product administration values, creates a record in the audit log that includes the date and time.

authentication Verification of the identity of a process or person.

B

backpressure The effect on the environment leading up to the point of restriction. See "congestion."

BB_Credit *See* "buffer-to-buffer credit."

beaconing Repeated transmission of a beacon light and message until an error is corrected or bypassed. Typically used by a piece of equipment when an individual Field Replaceable Unit (FRU) needs replacement. Beaconing helps the field engineer locate the specific defective component. Some equipment management software systems such as Connectrix Manager offer beaconing capability.

BER *See* "bit error rate."

bidirectional In Fibre Channel, the capability to simultaneously communicate at maximum speeds in both directions over a link.

bit error rate Ratio of received bits that contain errors to total of all bits transmitted.

blade server A consolidation of independent servers and switch technology in the same chassis.

blocked port Devices communicating with a blocked port are prevented from logging in to the Fibre Channel switch containing the port or communicating with other devices attached to the switch. A blocked port continuously transmits the off-line sequence (OLS).

bridge A device that provides a translation service between two network segments utilizing different communication protocols. EMC supports and sells bridges that convert iSCSI storage commands from a NIC-attached server to Fibre Channel commands for a storage platform.

broadcast Sends a transmission to all ports in a network. Typically used in IP networks. Not typically used in Fibre Channel networks.

broadcast frames Data packet, also known as a broadcast packet, whose destination address specifies all computers on a network. *See also* "multicast."

buffer

Storage area for data in transit. Buffers compensate for differences in link speeds and link congestion between devices.

buffer-to-buffer credit

The number of receive buffers allocated by a receiving FC_Port to a transmitting FC_Port. The value is negotiated between Fibre Channel ports during link initialization. Each time a port transmits a frame it decrements this credit value. Each time a port receives an R_Rdy frame it increments this credit value. If the credit value is decremented to zero, the transmitter stops sending any new frames until the receiver has transmitted an R_Rdy frame. Buffer-to-buffer credit is particularly important in SRDF and Mirror View distance extension solutions.

C

Call Home

A product feature that allows the Connectrix service processor to automatically dial out to a support center and report system problems. The support center server accepts calls from the Connectrix service processor, logs reported events, and can notify one or more support center representatives. Telephone numbers and other information are configured through the Windows NT dial-up networking application. The Call Home function can be enabled and disabled through the Connectrix Product Manager.

channel

With Open Systems, a channel is a point-to-point link that transports data from one point to another on the communication path, typically with high throughput and low latency that is generally required by storage systems. With Mainframe environments, a channel refers to the server-side of the server-storage communication path, analogous to the HBA in Open Systems.

Class 2 Fibre Channel class of service

In Class 2 service, the fabric and destination N_Ports provide connectionless service with notification of delivery or nondelivery between the two N_Ports. Historically Class 2 service is not widely used in Fibre Channel system.

Class 3 Fibre Channel class of service

Class 3 service provides a connectionless service without notification of delivery between N_Ports. (This is also known as datagram service.) The transmission and routing of Class 3 frames is the same as for Class 2 frames. Class 3 is the dominant class of communication used in Fibre Channel for moving data between servers and storage and may be referred to as “Ship and pray.”

Class F Fibre Channel class of service
Class F service is used for all switch-to-switch communication in a multiswitch fabric environment. It is nearly identical to class 2 from a flow control point of view.

community
A relationship between an SNMP agent and a set of SNMP managers that defines authentication, access control, and proxy characteristics.

community name
A name that represents an SNMP community that the agent software recognizes as a valid source for SNMP requests. An SNMP management program that sends an SNMP request to an agent program must identify the request with a community name that the agent recognizes or the agent discards the message as an authentication failure. The agent counts these failures and reports the count to the manager program upon request, or sends an authentication failure trap message to the manager program.

community profile
Information that specifies which management objects are available to what management domain or SNMP community name.

congestion
Occurs at the point of restriction. See "backpressure."

connectionless
Non dedicated link. Typically used to describe a link between nodes that allows the switch to forward Class 2 or Class 3 frames as resources (ports) allow. *Contrast with* the dedicated bandwidth that is required in a Class 1 Fibre Channel Service point-to-point link.

Connectivity Unit
A hardware component that contains hardware (and possibly software) that provides Fibre Channel connectivity across a fabric. Connectrix switches are example of Connectivity Units. This is a term popularized by the Fibre Alliance MIB, sometimes abbreviated to connunit.

Connectrix management software
The software application that implements the management user interface for all managed Fibre Channel products, typically the Connectrix -M product line. Connectrix Management software is a client/server application with the server running on the Connectrix service processor, and clients running remotely or on the service processor.

Connectrix service processor
An optional 1U server shipped with the Connectrix -M product line to run the Connectrix Management server software and EMC remote support application software.

Control Unit In mainframe environments, a Control Unit controls access to storage. It is analogous to a Target in Open Systems environments.

core switch Occupies central locations within the interconnections of a fabric. Generally provides the primary data paths across the fabric and the direct connections to storage devices. Connectrix directors are typically installed as core switches, but may be located anywhere in the fabric.

credit A numeric value that relates to the number of available BB_Credits on a Fibre Channel port. *See*“buffer-to-buffer credit”.

D

DASD Direct Access Storage Device.

default Pertaining to an attribute, value, or option that is assumed when none is explicitly specified.

default zone A zone containing all attached devices that are not members of any active zone. Typically the default zone is disabled in a Connectrix M environment which prevents newly installed servers and storage from communicating until they have been provisioned.

Dense Wavelength Division Multiplexing (DWDM) A process that carries different data channels at different wavelengths over one pair of fiber optic links. A conventional fiber-optic system carries only one channel over a single wavelength traveling through a single fiber.

destination ID A field in a Fibre Channel header that specifies the destination address for a frame. The Fibre Channel header also contains a Source ID (SID). The FCID for a port contains both the SID and the DID.

device A piece of equipment, such as a server, switch or storage system.

dialog box A user interface element of a software product typically implemented as a pop-up window containing informational messages and fields for modification. Facilitates a dialog between the user and the application. Dialog box is often used interchangeably with window.

DID An acronym used to refer to either Domain ID or Destination ID. This ambiguity can create confusion. As a result E-Lab recommends this acronym be used to apply to Domain ID. Destination ID can be abbreviated to FCID.

director

An enterprise-class Fibre Channel switch, such as the Connectrix ED-140M, MDS 9509, or ED-48000B. Directors deliver high availability, failure ride-through, and repair under power to insure maximum uptime for business critical applications. Major assemblies, such as power supplies, fan modules, switch controller cards, switching elements, and port modules, are all hot-swappable.

The term director may also refer to a board-level module in the VMAX that provides the interface between host channels (through an associated adapter module in the VMAX) and VMAX disk devices. (This description is presented here only to clarify a term used in other EMC documents.)

DNS

See “domain name service name.”

domain ID

A byte-wide field in the three byte Fibre Channel address that uniquely identifies a switch in a fabric. The three fields in a FCID are domain, area, and port. A distinct Domain ID is requested from the principal switch. The principal switch allocates one Domain ID to each switch in the fabric. A user may be able to set a Preferred ID which can be requested of the Principal switch, or set an Insistent Domain ID. If two switches insist on the same DID one or both switches will segment from the fabric.

domain name service name

Host or node name for a system that is translated to an IP address through a name server. All DNS names have a host name component and, if fully qualified, a domain component, such as *host1.abcd.com*. In this example, *host1* is the host name.

dual-attached host

A host that has two (or more) connections to a set of devices.

E

E_D_TOV

A time-out period within which each data frame in a Fibre Channel sequence transmits. This avoids time-out errors at the destination Nx_Port. This function facilitates high speed recovery from dropped frames. Typically this value is 2 seconds.

E_Port

Expansion Port, a port type in a Fibre Channel switch that attaches to another E_Port on a second Fibre Channel switch forming an Interswitch Link (ISL). This link typically conforms to the FC-SW standards developed by the T11 committee, but might not support heterogeneous inter operability.

edge switch Occupies the periphery of the fabric, generally providing the direct connections to host servers and management workstations. No two edge switches can be connected by interswitch links (ISLs). Connectrix departmental switches are typically installed as edge switches in a multiswitch fabric, but may be located anywhere in the fabric

Embedded Web Server A management interface embedded on the switch's code that offers features similar to (but not as robust as) the Connectrix Manager and Product Manager.

error detect time out value Defines the time the switch waits for an expected response before declaring an error condition. The error detect time out value (E_D_TOV) can be set within a range of two-tenths of a second to one second using the Connectrix switch Product Manager.

error message An indication that an error has been detected. *See also* "information message" *and* "warning message."

Ethernet A baseband LAN that allows multiple station access to the transmission medium at will without prior coordination and which avoids or resolves contention.

event log A record of significant events that have occurred on a Connectrix switch, such as FRU failures, degraded operation, and port problems.

expansionport *See* "E_Port."

explicit fabric login In order to join a fabric, an Nport must login to the fabric (an operation referred to as an FLOGI). Typically this is an explicit operation performed by the Nport communicating with the F_port of the switch, and is called an explicit fabric login. Some legacy Fibre Channel ports do not perform explicit login, and switch vendors perform login for ports creating an implicit login. Typically logins are explicit.

F

FA Fibre Adapter, another name for a VMAX Fibre Channel director.

F_Port Fabric Port, a port type on a Fibre Channel switch. An F_Port attaches to an N_Port through a point-to-point full-duplex link connection. A G_Port automatically becomes an F_port or an E-Port depending on the port initialization process.

fabric One or more switching devices that interconnect Fibre Channel N_Ports, and route Fibre Channel frames based on destination IDs in the frame headers. A fabric provides discovery, path provisioning, and state change management services for a Fibre Channel environment.

fabric element Any active switch or director in the fabric.

fabric login Process used by N_Ports to establish their operating parameters including class of service, speed, and buffer-to-buffer credit value.

fabric port A port type (F_Port) on a Fibre Channel switch that attaches to an N_Port through a point-to-point full-duplex link connection. An N_Port is typically a host (HBA) or a storage device like VMAX or VNX series.

fabric shortest path first (FSPF) A routing algorithm implemented by Fibre Channel switches in a fabric. The algorithm seeks to minimize the number of hops traversed as a Fibre Channel frame travels from its source to its destination.

fabric tree A hierarchical list in Connectrix Manager of all fabrics currently known to the Connectrix service processor. The tree includes all members of the fabrics, listed by WWN or nickname.

failover The process of detecting a failure on an active Connectrix switch FRU and the automatic transition of functions to a backup FRU.

fan-in/fan-out Term used to describe the server:storage ratio, where a graphic representation of a 1:n (fan-in) or n:1 (fan-out) logical topology looks like a hand-held fan, with the wide end toward n. By convention fan-out refers to the number of server ports that share a single storage port. Fan-out consolidates a large number of server ports on a fewer number of storage ports. Fan-in refers to the number of storage ports that a single server port uses. Fan-in enlarges the storage capacity used by a server. A fan-in or fan-out rate is often referred to as just the n part of the ratio; For example, a 16:1 fan-out is also called a fan-out rate of 16, in this case 16 server ports are sharing a single storage port.

FCP *See* "Fibre Channel Protocol."

FC-SW The Fibre Channel fabric standard. The standard is developed by the T11 organization whose documentation can be found at T11.org. EMC actively participates in T11. T11 is a committee within the InterNational Committee for Information Technology (INCITS).

fiber optics The branch of optical technology concerned with the transmission of radiant power through fibers made of transparent materials such as glass, fused silica, and plastic.

Either a single discrete fiber or a non spatially aligned fiber bundle can be used for each information channel. Such fibers are often called optical fibers to differentiate them from fibers used in non-communication applications.

fibre A general term used to cover all physical media types supported by the Fibre Channel specification, such as optical fiber, twisted pair, and coaxial cable.

Fibre Channel The general name of an integrated set of ANSI standards that define new protocols for flexible information transfer. Logically, Fibre Channel is a high-performance serial data channel.

Fibre Channel Protocol A standard Fibre Channel FC-4 level protocol used to run SCSI over Fibre Channel.

Fibre Channel switch modules The embedded switch modules in the back plane of the blade server. See "blade server" on page 293.

firmware The program code (embedded software) that resides and executes on a connectivity device, such as a Connectrix switch, a VMAX Fibre Channel director, or a host bus adapter (HBA).

F_Port Fabric Port, a physical interface within the fabric. An F_Port attaches to an N_Port through a point-to-point full-duplex link connection.

frame A set of fields making up a unit of transmission. Each field is made of bytes. The typical Fibre Channel frame consists of fields: Start-of-frame, header, data-field, CRC, end-of-frame. The maximum frame size is 2148 bytes.

frame header Control information placed before the data-field when encapsulating data for network transmission. The header provides the source and destination IDs of the frame.

FRU Field-replaceable unit, a hardware component that can be replaced as an entire unit. The Connectrix switch Product Manager can display status for the FRUs installed in the unit.

FSPF Fabric Shortest Path First, an algorithm used for routing traffic. This means that, between the source and destination, only the paths that have the least amount of physical hops will be used for frame delivery.

G

gateway address In TCP/IP, a device that connects two systems that use the same or different protocols.

gigabyte (GB) A unit of measure for storage size, loosely one billion (10^9) bytes. One gigabyte actually equals 1,073,741,824 bytes.

G_Port A port type on a Fibre Channel switch capable of acting either as an F_Port or an E_Port, depending on the port type at the other end of the link.

GUI Graphical user interface.

H

HBA *See* "host bus adapter."

hexadecimal Pertaining to a numbering system with base of 16; valid numbers use the digits 0 through 9 and characters A through F (which represent the numbers 10 through 15).

high availability A performance feature characterized by hardware component redundancy and hot-swappability (enabling non-disruptive maintenance). High-availability systems maximize system uptime while providing superior reliability, availability, and serviceability.

hop A hop refers to the number of InterSwitch Links (ISLs) a Fibre Channel frame must traverse to go from its source to its destination. Good design practice encourages three hops or less to minimize congestion and performance management complexities.

host bus adapter A bus card in a host system that allows the host system to connect to the storage system. Typically the HBA communicates with the host over a PCI or PCI Express bus and has a single Fibre Channel link to the fabric. The HBA contains an embedded microprocessor with on board firmware, one or more ASICs, and a Small Form Factor Pluggable module (SFP) to connect to the Fibre Channel link.

I

I/O *See* “input/output.”

in-band management Transmission of monitoring and control functions over the Fibre Channel interface. You can also perform these functions out-of-band typically by use of the ethernet to manage Fibre Channel devices.

information message A message telling a user that a function is performing normally or has completed normally. User acknowledgement might or might not be required, depending on the message. *See also* “error message” *and* “warning message.”

input/output (1) Pertaining to a device whose parts can perform an input process and an output process at the same time. (2) Pertaining to a functional unit or channel involved in an input process, output process, or both (concurrently or not), and to the data involved in such a process. (3) Pertaining to input, output, or both.

interface (1) A shared boundary between two functional units, defined by functional characteristics, signal characteristics, or other characteristics as appropriate. The concept includes the specification of the connection of two devices having different functions. (2) Hardware, software, or both, that links systems, programs, or devices.

Internet Protocol *See* “IP.”

interoperability The ability to communicate, execute programs, or transfer data between various functional units over a network. Also refers to a Fibre Channel fabric that contains switches from more than one vendor.

interswitch link (ISL) Interswitch link, a physical E_Port connection between any two switches in a Fibre Channel fabric. An ISL forms a hop in a fabric.

IP Internet Protocol, the TCP/IP standard protocol that defines the datagram as the unit of information passed across an internet and provides the basis for connectionless, best-effort packet delivery service. IP includes the ICMP control and error message protocol as an integral part.

IP address A unique string of numbers that identifies a device on a network. The address consists of four groups (quadrants) of numbers delimited by periods. (This is called *dotted-decimal* notation.) All resources on the network must have an IP address. A valid IP address is in the form *nnn.nnn.nnn.nnn*, where each *nnn* is a decimal in the range 0 to 255.

ISL Interswitch link, a physical E_Port connection between any two switches in a Fibre Channel fabric.

K

kilobyte (K) A unit of measure for storage size, loosely one thousand bytes. One kilobyte actually equals 1,024 bytes.

L

laser A device that produces optical radiation using a population inversion to provide light amplification by stimulated emission of radiation and (generally) an optical resonant cavity to provide positive feedback. Laser radiation can be highly coherent temporally, spatially, or both.

LED Light-emitting diode.

link The physical connection between two devices on a switched fabric.

link incident A problem detected on a fiber-optic link; for example, loss of light, or invalid sequences.

load balancing The ability to distribute traffic over all network ports that are the same distance from the destination address by assigning different paths to different messages. Increases effective network bandwidth. EMC PowerPath software provides load-balancing services for server IO.

logical volume A named unit of storage consisting of a logically contiguous set of disk sectors.

Logical Unit Number (LUN) A number, assigned to a storage volume, that (in combination with the storage device node's World Wide Port Name (WWPN)) represents a unique identifier for a logical volume on a storage area network.

M

MAC address Media Access Control address, the hardware address of a device connected to a shared network.

managed product A hardware product that can be managed using the Connectrix Product Manager. For example, a Connectrix switch is a managed product.

management session Exists when a user logs in to the Connectrix Management software and successfully connects to the product server. The user must specify the network address of the product server at login time.

media The disk surface on which data is stored.

media access control *See* "MAC address."

megabyte (MB) A unit of measure for storage size, loosely one million (10^6) bytes. One megabyte actually equals 1,048,576 bytes.

MIB Management Information Base, a related set of objects (variables) containing information about a managed device and accessed through SNMP from a network management station.

multicast Multicast is used when multiple copies of data are to be sent to designated, multiple, destinations.

multiswitch fabric Fibre Channel fabric created by linking more than one switch or director together to allow communication. *See also* "ISL."

multiswitch linking Port-to-port connections between two switches.

N

name server (dNS) A service known as the distributed Name Server provided by a Fibre Channel fabric that provides device discovery, path provisioning, and state change notification services to the N_Ports in the fabric. The service is implemented in a distributed fashion, for example, each switch in a fabric participates in providing the service. The service is addressed by the N_Ports through a Well Known Address.

network address A name or address that identifies a managed product, such as a Connectrix switch, or a Connectrix service processor on a TCP/IP network. The network address can be either an IP address in dotted decimal notation, or a Domain Name Service (DNS) name as administered on a customer network. All DNS names have a host name component and (if fully qualified) a domain component, such as *host1.emc.com*. In this example, *host1* is the host name and *EMC.com* is the domain component.

nickname A user-defined name representing a specific WWxN, typically used in a Connectrix -M management environment. The analog in the Connectrix -B and MDS environments is alias.

node The point at which one or more functional units connect to the network.

N_Port Node Port, a Fibre Channel port implemented by an end device (node) that can attach to an F_Port or directly to another N_Port through a point-to-point link connection. HBAs and storage systems implement N_Ports that connect to the fabric.

NVRAM Nonvolatile random access memory.

O

offline sequence (OLS) The OLS Primitive Sequence is transmitted to indicate that the FC_Port transmitting the Sequence is:

a. initiating the Link Initialization Protocol

b. receiving and recognizing NOS

c. or entering the offline state

OLS *See* "offline sequence (OLS)".

operating mode Regulates what other types of switches can share a multiswitch fabric with the switch under consideration.

operating system Software that controls the execution of programs and that may provide such services as resource allocation, scheduling, input/output control, and data management. Although operating systems are predominantly software, partial hardware implementations are possible.

optical cable A fiber, multiple fibers, or a fiber bundle in a structure built to meet optical, mechanical, and environmental specifications.

OS *See* "operating system."

out-of-band management Transmission of monitoring/control functions outside of the Fibre Channel interface, typically over ethernet.

oversubscription The ratio of bandwidth required to bandwidth available. When all ports, associated pair-wise, in any random fashion, cannot sustain full duplex at full line-rate, the switch is oversubscribed.

P

parameter A characteristic element with a variable value that is given a constant value for a specified application. Also, a user-specified value for an item in a menu; a value that the system provides when a menu is interpreted; data passed between programs or procedures.

password (1) A value used in authentication or a value used to establish membership in a group having specific privileges. (2) A unique string of characters known to the computer system and to a user who must specify it to gain full or limited access to a system and to the information stored within it.

path In a network, any route between any two nodes.

persistent binding Use of server-level access control configuration information to persistently bind a server device name to a specific Fibre Channel storage volume or logical unit number, through a specific HBA and storage port WWN. The address of a persistently bound device does not shift if a storage target fails to recover during a power cycle. This function is the responsibility of the HBA device driver.

port (1) An access point for data entry or exit. (2) A receptacle on a device to which a cable for another device is attached.

port card Field replaceable hardware component that provides the connection for fiber cables and performs specific device-dependent logic functions.

port name A symbolic name that the user defines for a particular port through the Product Manager.

preferred domain ID An ID configured by the fabric administrator. During the fabric build process a switch requests permission from the principal switch to use its preferred domain ID. The principal switch can deny this request by providing an alternate domain ID only if there is a conflict for the requested Domain ID. Typically a principal switch grants the non-principal switch its requested Preferred Domain ID.

principal downstream ISL The ISL to which each switch will forward frames originating from the principal switch.

principal ISL The principal ISL is the ISL that frames destined to, or coming from, the principal switch in the fabric will use. An example is an RDI frame.

principal switch In a multiswitch fabric, the switch that allocates domain IDs to itself and to all other switches in the fabric. There is always one principal switch in a fabric. If a switch is not connected to any other switches, it acts as its own principal switch.

principal upstream ISL The ISL to which each switch will forward frames destined for the principal switch. The principal switch does not have any upstream ISLs.

product (1) Connectivity Product, a generic name for a switch, director, or any other Fibre Channel product. (2) Managed Product, a generic hardware product that can be managed by the Product Manager (a Connectrix switch is a managed product). Note distinction from the definition for "device."

Product Manager A software component of Connectrix Manager software such as a Connectrix switch product manager, that implements the management user interface for a specific product. When a product instance is opened from the Connectrix Manager software products view, the corresponding product manager is invoked. The product manager is also known as an Element Manager.

product name A user configurable identifier assigned to a Managed Product. Typically, this name is stored on the product itself. For a Connectrix switch, the Product Name can also be accessed by an SNMP Manager as the System Name. The Product Name should align with the host name component of a Network Address.

products view The top-level display in the Connectrix Management software user interface that displays icons of Managed Products.

protocol (1) A set of semantic and syntactic rules that determines the behavior of functional units in achieving communication. (2) A specification for the format and relative timing of information exchanged between communicating parties.

R

R_A_TOV *See* "resource allocation time out value."

remote access link The ability to communicate with a data processing facility through a remote data link.

remote notification The system can be programmed to notify remote sites of certain classes of events.

remote user workstation A workstation, such as a PC, using Connectrix Management software and Product Manager software that can access the Connectrix service processor over a LAN connection. A user at a remote workstation can perform all of the management and monitoring tasks available to a local user on the Connectrix service processor.

resource allocation time out value A value used to time-out operations that depend on a maximum time that an exchange can be delayed in a fabric and still be delivered. The resource allocation time-out value of (R_A_TOV) can be set within a range of two-tenths of a second to 120 seconds using the Connectrix switch product manager. The typical value is 10 seconds.

S

SAN *See* "storage area network (SAN)."

segmentation A non-connection between two switches. Numerous reasons exist for an operational ISL to segment, including interop mode incompatibility, zoning conflicts, and domain overlaps.

segmented E_Port E_Port that has ceased to function as an E_Port within a multiswitch fabric due to an incompatibility between the fabrics that it joins.

service processor *See* "Connectrix service processor."

session *See* "management session."

single attached host A host that only has a single connection to a set of devices.

small form factor pluggable (SFP)
An optical module implementing a shortwave or long wave optical transceiver.

SMTP
Simple Mail Transfer Protocol, a TCP/IP protocol that allows users to create, send, and receive text messages. SMTP protocols specify how messages are passed across a link from one system to another. They do not specify how the mail application accepts, presents or stores the mail.

SNMP
Simple Network Management Protocol, a TCP/IP protocol that generally uses the User Datagram Protocol (UDP) to exchange messages between a management information base (MIB) and a management client residing on a network.

storage area network (SAN)
A network linking servers or workstations to disk arrays, tape backup systems, and other devices, typically over Fibre Channel and consisting of multiple fabrics.

subnet mask
Used by a computer to determine whether another computer with which it needs to communicate is located on a local or remote network. The network mask depends upon the class of networks to which the computer is connecting. The mask indicates which digits to look at in a longer network address and allows the router to avoid handling the entire address. Subnet masking allows routers to move the packets more quickly. Typically, a subnet may represent all the machines at one geographic location, in one building, or on the same local area network.

switch priority
Value configured into each switch in a fabric that determines its relative likelihood of becoming the fabric's principal switch.

T

TCP/IP
Transmission Control Protocol/Internet Protocol. TCP/IP refers to the protocols that are used on the Internet and most computer networks. TCP refers to the Transport layer that provides flow control and connection services. IP refers to the Internet Protocol level where addressing and routing are implemented.

toggle
To change the state of a feature/function that has only two states. For example, if a feature/function is *enabled*, toggling changes the state to *disabled*.

topology Logical and/or physical arrangement of switches on a network.

trap An asynchronous (unsolicited) notification of an event originating on an SNMP-managed device and directed to a centralized SNMP Network Management Station.

U

unblocked port Devices communicating with an unblocked port can log in to a Connectrix switch or a similar product and communicate with devices attached to any other unblocked port if the devices are in the same zone.

Unicast Unicast routing provides one or more optimal path(s) between any of two switches that make up the fabric. (This is used to send a single copy of the data to designated destinations.)

upper layer protocol (ULP) The protocol user of FC-4 including IPI, SCSI, IP, and SBCCS. In a device driver ULP typically refers to the operations that are managed by the class level of the driver, not the port level.

URL Uniform Resource Locater, the addressing system used by the World Wide Web. It describes the location of a file or server anywhere on the Internet.

V

virtual switch A Fibre Channel switch function that allows users to subdivide a physical switch into multiple virtual switches. Each virtual switch consists of a subset of ports on the physical switch, and has all the properties of a Fibre Channel switch. Multiple virtual switches can be connected through ISL to form a virtual fabric or VSAN.

virtual storage area network (VSAN) An allocation of switch ports that can span multiple physical switches, and forms a virtual fabric. A single physical switch can sometimes host more than one VSAN.

volume A general term referring to an addressable logically contiguous storage space providing block IO services.

VSAN Virtual Storage Area Network.

W

warning message An indication that a possible error has been detected. *See also* "error message" *and* "information message."

World Wide Name (WWN) A unique identifier, even on global networks. The WWN is a 64-bit number (XX:XX:XX:XX:XX:XX:XX:XX). The WWN contains an OUI which uniquely determines the equipment manufacturer. OUIs are administered by the Institute of Electronic and Electrical Engineers (IEEE). The Fibre Channel environment uses two types of WWNs; a World Wide Node Name (WWNN) and a World Wide Port Name (WWPN). Typically the WWPN is used for zoning (path provisioning function).

Z

zone An information object implemented by the distributed Nameserver(dNS) of a Fibre Channel switch. A zone contains a set of members which are permitted to discover and communicate with one another. The members can be identified by a WWPN or port ID. EMC recommends the use of WWPNs in zone management.

zone set An information object implemented by the distributed Nameserver(dNS) of a Fibre Channel switch. A Zone Set contains a set of Zones. A Zone Set is activated against a fabric, and only one Zone Set can be active in a fabric.

zonie A storage administrator who spends a large percentage of his workday zoning a Fibre Channel network and provisioning storage.

zoning Zoning allows an administrator to group several devices by function or by location. All devices connected to a connectivity product, such as a Connectrix switch, may be configured into one or more zones.

Index