DCNID

Data Center Network Infrastructure Design

Volume 2

Version 2.0

Student Guide

Text Part Number: 97-2680-01

https://t.me/learningnets

Table of Contents

Volume 2

Cisco Nexus 5000 Series Switches ***3-1***

Overview 3-1
Module Objectives 3-1

Fibre Channel over Ethernet **3-3**

Overview 3-3
Objectives 3-3
Introduction 3-4
Data Center Landscape 3-4
Data Center: Unified I/O with FCoE 3-5
FCoE: Storage Targets 3-6
Unified I/O—DCE Fabric with FCoE 3-7
10 GE Drivers in the Data Center 3-8
Converged Network Adapter 3-10
Converged Adapter with Enhanced Ethernet 3-11
Fibre Channel over Ethernet 3-12
Fibre Channel over Ethernet 3-13
PCI Express—PCIe 3-14
FCoE Frame Format 3-15
Decoded FCoE Packet (Wireshark) 3-17
FCoE Protocol Stack 3-18
Enhanced Ethernet Features: Ethernet Extensions Used in FCoE Networks 3-19
Data Center Enhanced Ethernet Features 3-19
I/O Consolidation with FCoE—Benefits 3-21
Fibre Channel over Ethernet 3-22
Lossless Fibre Channel: BB_Credits 3-23
Enhanced Ethernet: PFC 3-27
Enhanced Ethernet: Data Center Bridging Capability eXchange Protocol 3-28
Discovery—DCBCXP 3-29
DCBCXP Capabilities 3-30
Backward Congestion Notification: For Persistent Congestion 3-31
Enhanced Ethernet: BCN/QCN Congestion Management 3-32
Enhanced Ethernet: Delayed Drop for Transient Congestion 3-33
Delayed Drop 3-34
Enhanced Ethernet: COS-based BW Management 3-35
L2 Multipathing 3-36
Evolution of Ethernet Physical Media 3-37
Fibre Channel Protocol 3-38
Fibre Channel Layers 3-38
FCoE Mapping 3-39
Standard Fabric Services 3-40
Fibre Channel Classes of Service 3-41
Frames and Sequences 3-42
Fibre Channel Frame 3-43
Fibre Channel Frame Details 3-44
Fibre Channel Frame (Decoded Inside FCoE) 3-45
Fibre Channel over Ethernet 3-46
Fibre Channel and FCoE Addresses 3-47
FCoE Switch: Functional Diagram 3-48
Consolidated Network Adapter: Fibre Channel and Ethernet 3-49
FCoE Architecture: Day One 3-50
FCoE Architecture: The Big Picture 3-51
Fibre Channel and FCoE Address Mapping 3-52
Fibre Channel and FCoE Address Mapping—Step 1 3-53
Fibre Channel and FCoE Address Mapping—Step 2 3-54
Fibre Channel and FCoE Address Mapping—Step 3 3-55
Fibre Channel and FCoE Address Mapping—Last Step 3-56

https://t.me/learningnets

Summary 3-57
Appendix 3-58
FCoE Architecture 3-59
MAC Addressing 3-62
FCoE Discovery 3-65
Appendix Summary 3-68
Cisco Nexus 5000 Series Switches Overview 3-69
Overview 3-69
Objectives 3-69
Cisco Nexus 5000 Series Switch Product Family 3-70
Cisco Nexus 5000 Series Switch Family 3-70
SFP+ for Cisco Nexus 5000 Series Switch 3-71
Front View and Rear View 3-72
Cisco Nexus 5000 Series Switch: Data Center Class 3-73
Cisco Nexus 5000 Series Switch Features 3-74
Cisco Nexus Operating System (Cisco NX-OS) 3-75
Data Center Solution Components 3-76
Access Switch 3-77
Data Center Ethernet (DCE) 3-78
Fibre Channel over Ethernet 3-79
Virtualization 3-80
Unified SAN Fabric 3-81
Fibre Channel N-Port Virtualizer 3-82
N-Port Virtualizer 3-83
Base Configuration 3-85
Rack Server Deployment 3-86
Where Cisco Nexus 5000 Series Switches Fits 3-87
Current Data Center Environment 3-88
Top-of-Rack Consolidated I/O 3-89
10Gb Aggregation and I/O Consolidation 3-90
End-of-Row: 10Gb Aggregation 3-91
10Gb Aggregation 3-92
Top-of-Rack Consolidated I/O 3-93
Blade Center Environment 3-94
Cisco Nexus 5000 Series Switches Management Tools 3-95
Comprehensive Management Tools 3-95
Fibre Channel and Ethernet Management 3-96
SNMP Capabilities 3-97
XML/NetConf API 3-98
NX-OS 4.0 XML/NetConf Capabilities 3-99
Cisco Fabric Manager Suite 3-100
Cisco Fabric Manager 3-101
Device Manager 3-104
Web Services 3-105
Cisco Nexus 5000 Series Switches Licensing 3-106
Summary 3-108
Data Center Design Models 4-1
Overview 4-1
Module Objectives 4-1
Data Center Design Overview 4-3
Overview 4-3
Objectives 4-3
Multi-Tier Model Design 4-4
The Enterprise Composite Network Model 4-11
Summary 4-23

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Application Design 4-25
Overview 4-25
Objectives 4-25
Application Architecture Tiers 4-26
Cisco Application Optimization Delivery Products 4-37
ACE Appliance/Module Impact on Data Center Design 4-42
Network Topologies with Cisco ACE Appliance/Module 4-67
Cisco ACE Modes 4-67
Intra-Data Center Redundancy 4-92
Multi-Data Center Redundancy 4-96
GSLB Overview 4-98
Resolving a DNS Query 4-98
Server Load Balancing vs. Global Server Load Balancing 4-100
Route Health Injection 4-101
Summary 4-106
Appendix A 4-107
Server Virtualization 4-108
Server Clusters 4-113
Cluster Types 4-120
Cluster Interconnects 4-125
Server Farms and Load Balancing 4-130
WAN Optimization and Branch Consolidation 4-141

Data Center Strategy ***5-1***

Overview 5-1
Module Objectives 5-1

Data Center Strategy **5-3**

Overview 5-3
Objectives 5-3
Data Center Driver 5-4
Data Center Network Architecture 5-10
Access, Aggregation, Core, Server Farm, POD, Zone 5-10
Cisco Data Center Virtual Switching Technology 5-13
VSS and VDC 5-13
Virtual Blade Switching (VBS) 5-14
Server Virtualization Switching (SVS) 5-20
Unified Fabric and Unified I/O 5-21
Data Center Vision: A Unified Fabric 5-21
Data Center Ethernet 5-26
Data Center Sizing 5-37
Server Farm—Types and Sizes 5-37
Server Farms and Data Center Network Layers 5-39
Data Center Sizing 5-42
Summary 5-46

https://t.me/learningnets

Data Center Network Design ***6-1***

Overview 6-1
Module Objectives 6-1

Data Center Network Design **6-3**

Overview 6-3
Objectives 6-3
Data Center Network Design 6-4
Service Design 6-8
Layer 2 and Layer 3 Access Layer Design 6-31
Access Layer Looped Triangle Failure Analysis 6-39
Access Layer Looped Topologies 6-44
Failure Analysis in Looped Square Design 6-47
Data Center Loop-Free Design 6-51
Minimal Layer 2 Design 6-53
Data Center New Topologies 6-56
Virtual Switching Systems 6-57
Classic Design with VSS, VRF and VDC 6-83
Routing Design 6-101
Enhanced L2 Topology 6-113
Layer 2 Extensions for Data Center Interconnect 6-121
Nexus 7000 Series Switches 6-131
EEM in N-PE: Regular Operation 6-140
EEM in N-PE: P Node Failure 6-141
EEM in N-PE: P Node Recovering 6-142
EEM in N-PE: P Node Is Operational (After Probing Delay) 6-143
Reference Design Summary 6-144
Summary 6-145

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Module 3

Cisco Nexus 5000 Series Switches

Overview

This module introduces concepts and basic characteristics of Fibre Channel over Ethernet (FCoE) protocol and of the associated Cisco Nexus 5000 Series Switches.

Module Objectives

Upon completing this module, you will be able to describe Fibre Channel networks, frames, and flow control; explain the various protocols used with Fibre Channel; and identify the standard fabric services and their well-known addresses. This ability includes being able to meet these objectives:

- Describe how FCoE operates within Storage Area Network (SAN) and LAN environments
- Identify the benefits provided by the Cisco Nexus 5000 Series Switches in the most common user situations and explain how the switch works to provide these benefits
- Describe the features and functions of the Cisco Nexus 5000 Series Switches platform
- Explain how to position the Nexus 5000 Series Switches in the data center

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 1

Fibre Channel over Ethernet

Overview

This lesson describes the concept of the Fibre Channel over Ethernet (FCoE) protocol.

Objectives

Upon completing this lesson, you will be able to meet these objectives:

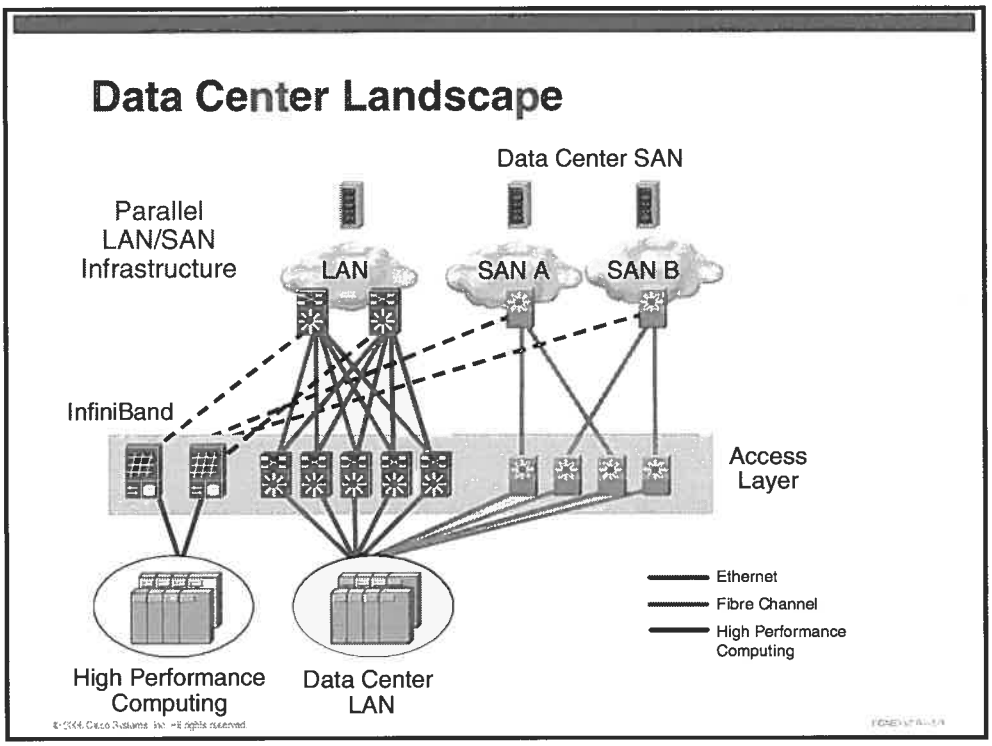
- Explain the requirements in the data center that make the FCoE protocol necessary
- Explain how FCoE is used for I/O consolidation within the data center
- Describe the Ethernet extensions used in FCoE networks
- Explain the basic functions and operations of the FCoE protocol
- Describe the basic components and operations of the Fibre Channel protocol
- Describe the Fibre Channel and FCoE address mapping

https://t.me/learningnets

Introduction

This topic describes the FCoE protocol basics and related features.

Data Center Landscape

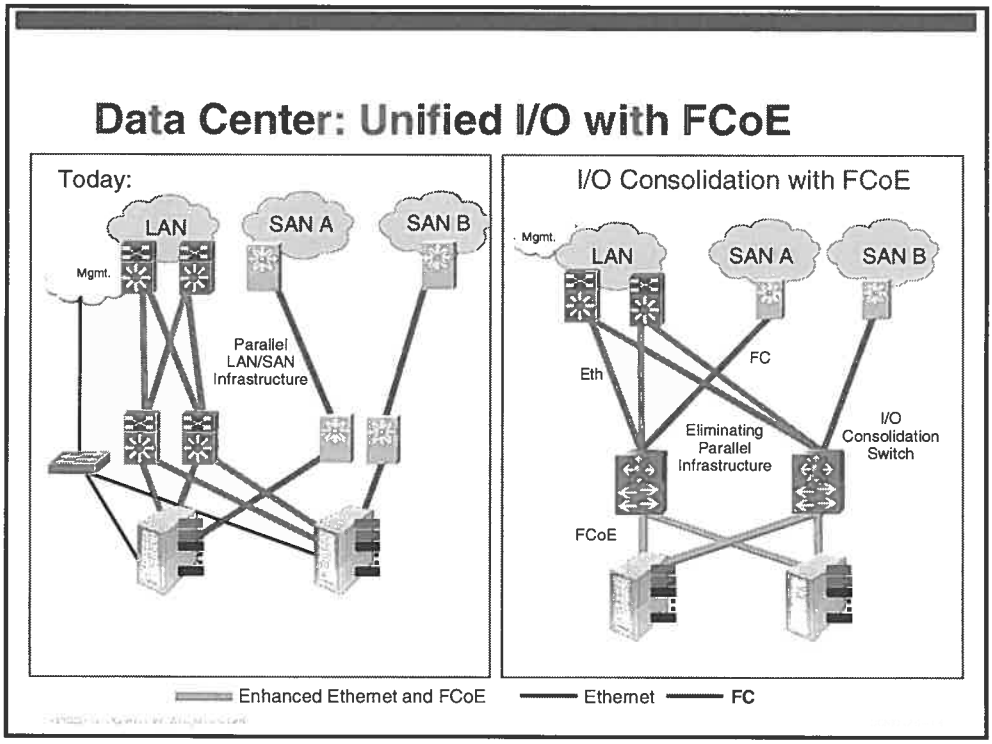


The parallel LAN/Storage Area Networking (SAN) infrastructure illustrated in the figure makes inefficient use of network infrastructure. The requirement for unified LAN/SAN fabric in the data center can be served by the Fibre Channel over Ethernet (FCoE) protocol.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center: Unified I/O with FCoE



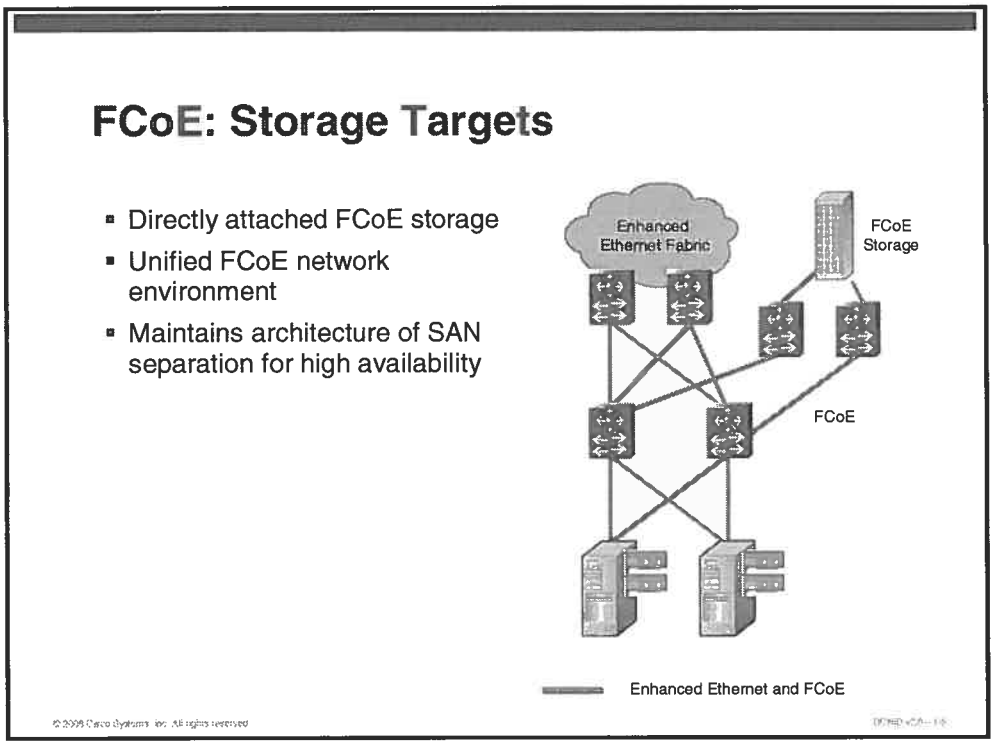
The parallel infrastructure (LAN and SAN) used today is an inefficient use of network infrastructure.

FCoE technology coexists with Fibre Channel fabric, allowing IT organizations to implement incrementally, while reaping business benefits including the following:

- Fewer interface cards per server
- Reduction in data center cabling
- Fewer access-layer switches
- Flexible performance management
- Power and cooling savings
- Fibre Channel infrastructure
- Seamless management integration

https://t.me/learningnets

FCoE: Storage Targets

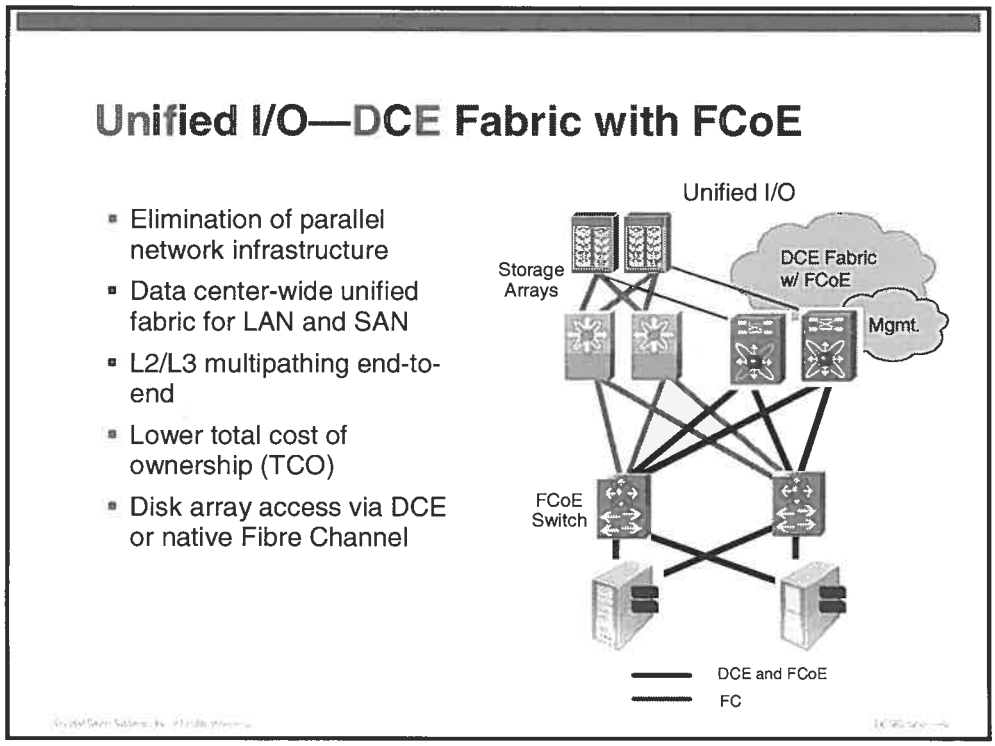


This topology represents an alternative to the heterogeneous fabrics: an end-to-end FCoE fabric connecting servers with storage subsystems.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

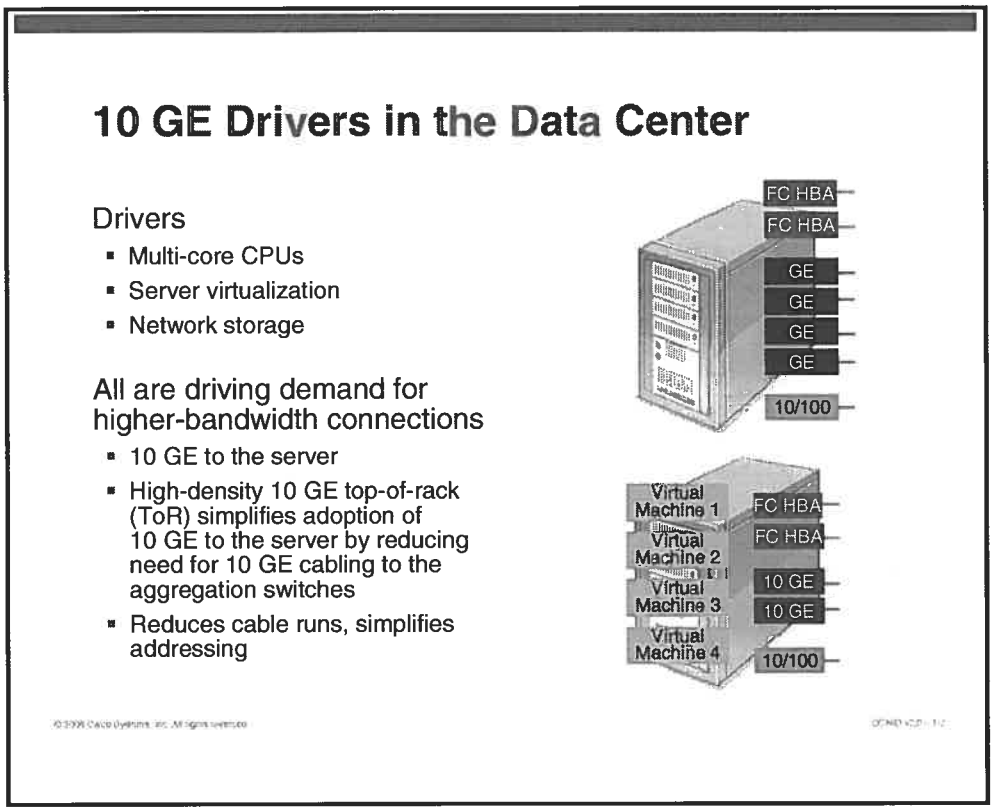
Unified I/O—DCE Fabric with FCoE



In the future, FCoE can be passed across the data center aggregation layer along with the TCP/IP Ethernet traffic using DCE of FCoE ports on the Cisco Nexus 7000 Series switching modules.

https://t.me/learningnets

10 GE Drivers in the Data Center



Multi-core CPUs allow faster execution of multiple applications on the same processor. Multi-socket, multi-core server technology supports higher workload levels, which, in turn, demand greater network throughput.

Virtualization allows the consolidation of multiple applications on a server, thus driving the bandwidth requirement. Server virtualization enables workload consolidation, which contributes, additionally, to network throughput demands.

The growing demand for network storage is driving the network bandwidth demand. Increasing use of network storage requires higher bandwidth between servers and storage.

10 Gigabit Ethernet (10 GE) is the enabling technology for I/O convergence because it offers the bandwidth and underlying enhancements to support multiple I/O streams on a single cable. The rapid adoption of 10 GE is driven by rapid growth in server performance, virtualization, and storage demand.

FCoE allows the use of a single cabling infrastructure within server racks. This technology simplifies network topology while reducing cabling cost and complexity, eliminating half of the I/O adapter cards in a rack, reducing power and cooling overhead—all while improving bandwidth by leveraging 10 GE.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Parallel LAN/SAN infrastructure: Inefficient use of network infrastructure

10 GE to the Server:

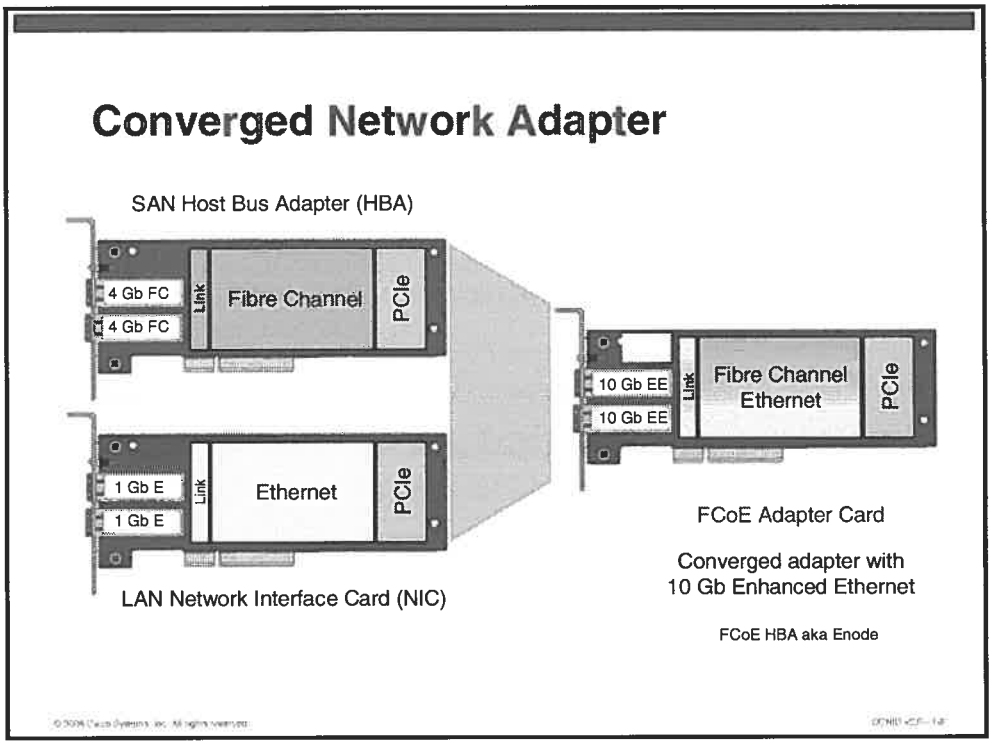
- Current 10 GE implementations are end-of-row due to lack of high-density top-of-rack 10 GE
- High-density 10 GE top-of-rack simplifies adoption of 10 GE to the server by reducing need for 10 GE cabling to the aggregation switches
- Reduces cable runs, simplifies MACs

Conceptually, FCoE can be broken down into three components:

- Encapsulation of a native Fibre Channel frame into an Ethernet frame
- Extension of Ethernet to become a lossless fabric
- Replacing of a Fibre Channel link with MAC addresses in a lossless Ethernet fabric

https://t.me/learningnets

Converged Network Adapter



A converged (Fibre channel + Ethernet) network adapter. A single adapter with the 10 Gb enhanced Ethernet basic can provide all the same functionality over a single wire. The value-proposition is sound, that is, the ability for the FCoE adapter to be able to take advantage of the same driver stacks as the SAN.

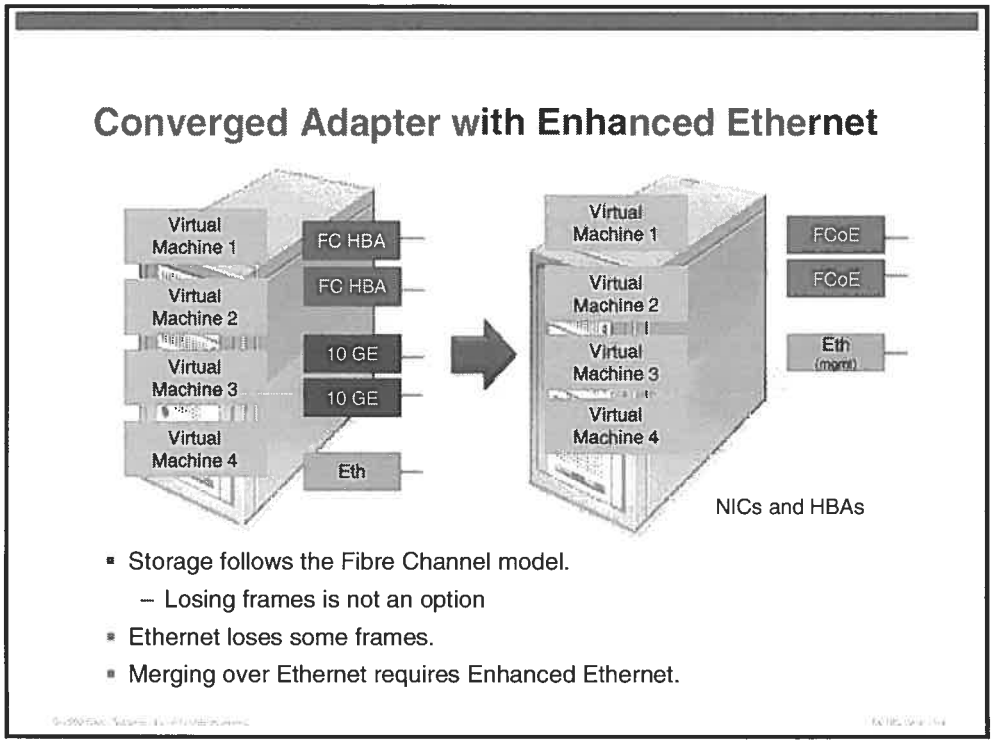
Converged Network Adapters (CNAs) allow storage administrators to manage storage networks as if each of their servers is configured with Fibre Channel host bus adapters.

CNAs help to further simplify FCoE deployment, allowing FCoE to be incorporated without disrupting current data center management practices, software, or the roles of network and storage administrators.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Converged Adapter with Enhanced Ethernet



With the simplified FCoE topology, what formerly took a minimum of four interfaces per server—two network interface cards (NICs) and two host bus adapters (HBAs)—now requires only two FCoE adapters per server.

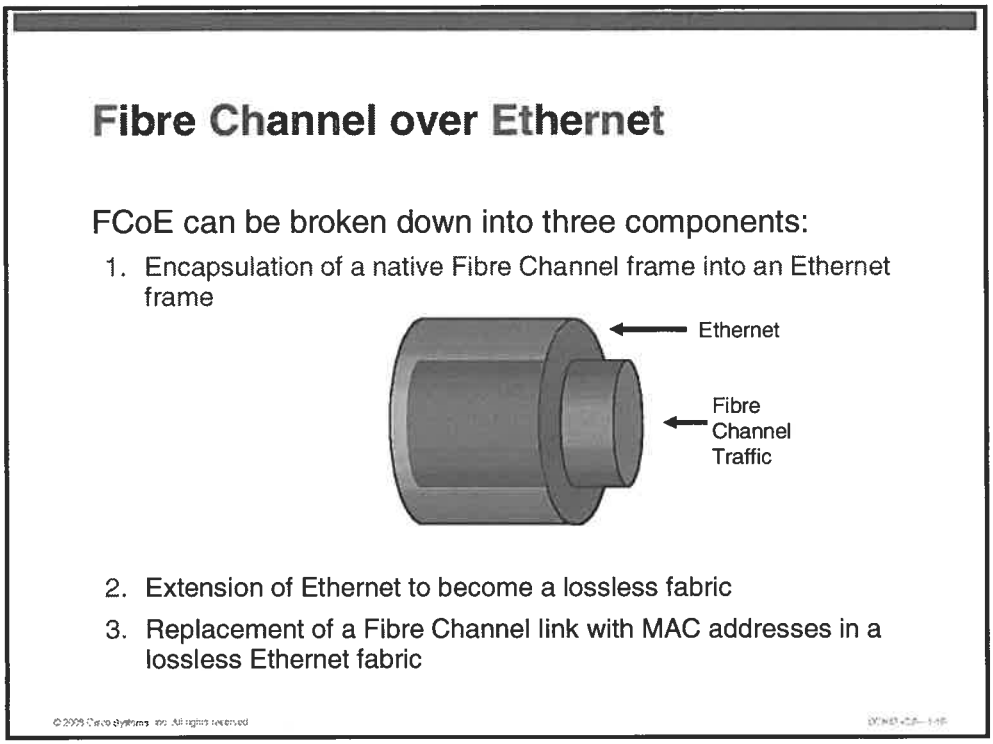
This technology eliminates a large number of Fibre Channel HBAs and allows IT organizations to realize more fully their vision of connecting every server to their central pool of storage. FCoE helps to extend data center SANs to servers whose expansion capability was insufficient, or where the cost of the Fibre Channel HBAs was prohibitive. Now virtually every data center server can leverage the benefits of centrally managed storage.

FCoE helps to extend data center SANs to servers whose expansion capability was insufficient, or where the cost of the Fibre Channel HBAs was prohibitive. Now virtually every data center server can leverage the benefits of centrally managed storage.

Because the redundant Fibre Channel HBAs and NICs can be replaced by two redundant CNAs, the total number of server adapters is reduced from five to three.

https://t.me/learningnets

Fibre Channel over Ethernet

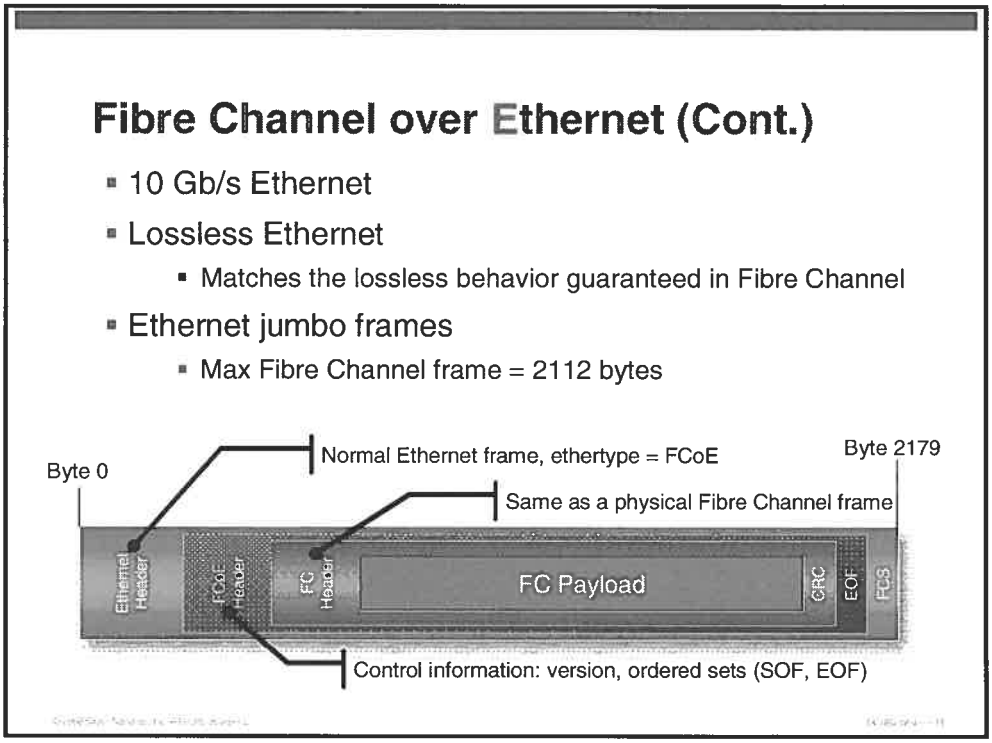


An Ethernet network that supports FCoE must meet three basic characteristics listed in the figure in order to transport Fibre Channel traffic.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel over Ethernet



FCoE is an extension of Fibre Channel (and its operating model) onto a lossless Ethernet fabric. FCoE maintains the Fibre Channel operation model, which provides seamless connectivity between the two networks.

FCoE positions Fibre Channel as the storage networking protocol of choice and extends the reach of Fibre Channel throughout the data center to all servers. Fibre Channel frames are encapsulated into Ethernet frames with no fragmentation, eliminating the need for higher-level protocols to re-assemble packets.

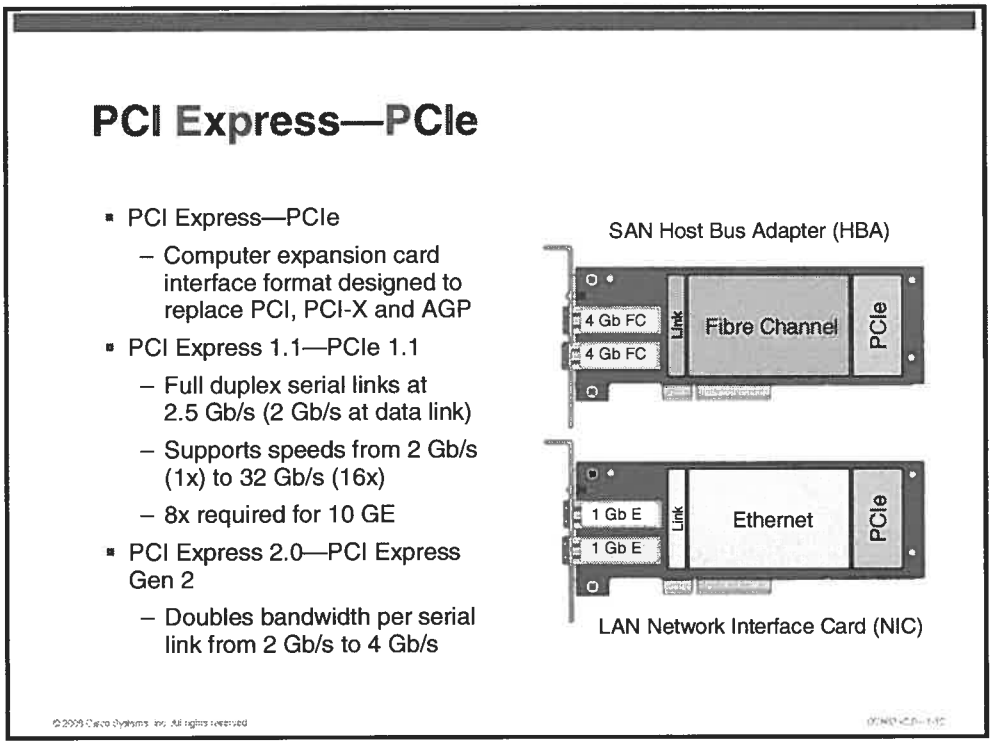
Fibre Channel overcomes the distance and switching limitations inherent in SCSI. Fibre Channel carries SCSI as its higher-level protocol. SCSI does not respond well to lost frames, which can result in significant delays when recovering from a loss. Because Fibre Channel carries SCSI, it inherits the requirement for an underlying lossless network.

FCoE transports native Fibre Channel frames over an Ethernet infrastructure, allowing existing Fibre Channel management modes to stay intact. One FCoE prerequisite is for the underlying network fabric to be lossless.

Frame size is a factor in FCoE. A typical Fibre Channel data frame has a 2112 byte payload, a header and Frame Check Sequence (FCS). A classical Ethernet frame is typically 1.5 KB or less. To maintain good performance, FCoE must utilize jumbo frames (or the 2.5 KB "baby jumbo") to prevent a Fibre Channel frame from being split into two Ethernet frames.

https://t.me/learningnets

PCI Express—PCIe



PCI Express (PCI-E or PCIe), is a computer expansion card interface format designed to replace PCI, PCI-X and AGP. PCIe 1.1 uses full duplex serial links at 2.5 Gb/s (2 Gb/s at the datalink).

- PCI Express supports speeds from 2 Gb/s (1x) to 32 Gb/s (16x).
- 8x is required for 10 GE.
- Products are currently shipping.

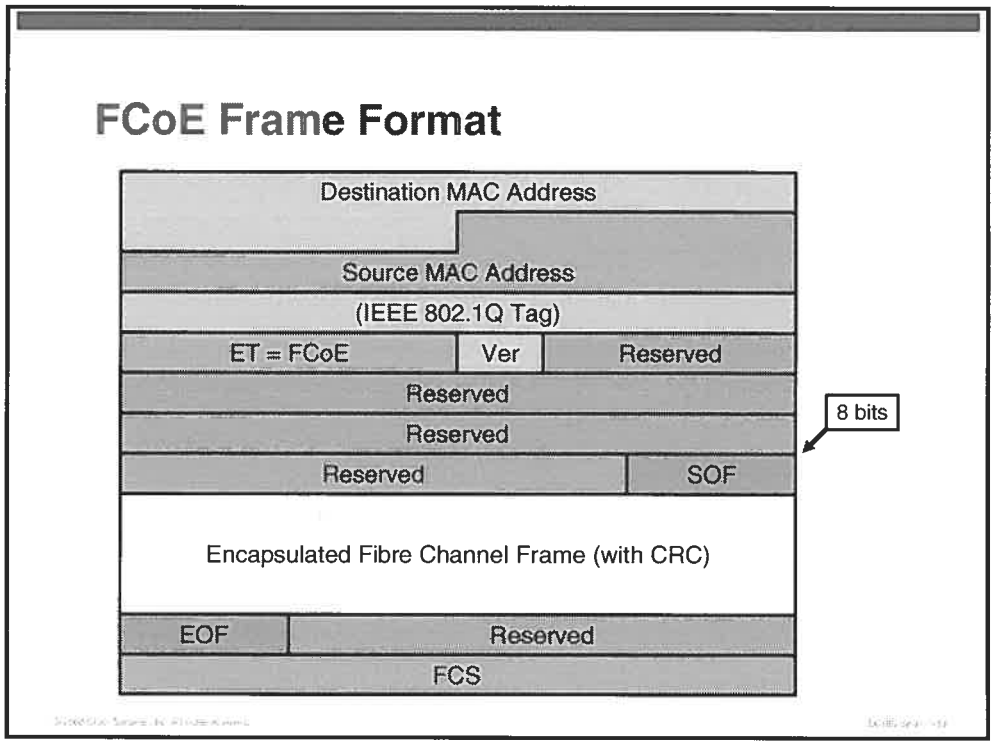
PCIe 2.0 (known as PCIe Gen 2) doubles the bandwidth per serial link from 2 Gb/s to 4 Gb/s.

- Specification has been available since January 2007.
- Products are becoming available.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

FCoE Frame Format



Format

FCoE encapsulates a Fibre Channel frame within an Ethernet frame.

The first 48 bits in the frame are used to specify the destination MAC address and the next 48 bits specify the source MAC address. The 32 bit IEEE 802.1Q Tag provides the same function as for virtual LANs, allowing multiple virtual networks across a single physical infrastructure. FCoE has its own Ethertype value as designated by the next 16 bits, followed by the 4 bit version field. The next 100 bits are reserved and are followed by the 8 bit start of frame and then the actual Fibre Channel frame. The 8 bit end of frame delimiter is followed by 24 reserved bits. The frame ends with the final 32 bits dedicated to the FCS function that provides error detection for the Ethernet frame.

The encapsulated Fibre Channel frame comprises the original 24 byte Fibre Channel header and the data that is being transported (including the Fibre Channel CRC). The CRC is a cyclical redundancy check that is used for error detection. The Fibre Channel header is maintained so that when a traditional Fibre Channel Storage Area Network is connected to an FCoE-capable switch, the frame is de-encapsulated and handed off seamlessly. This capability enables FCoE to integrate with existing Fibre Channel SANs without the need of a gateway.

Frame size is also a factor in FCoE. A typical Fibre Channel data frame has a 2112 byte payload, a header and FCS. A classical Ethernet frame is typically 1.5 KB or less. To maintain good performance, FCoE must utilize jumbo frames (or the 2.5 KB "baby jumbo") to prevent a Fibre Channel frame from being split into two Ethernet frames.

https://t.me/learningnets

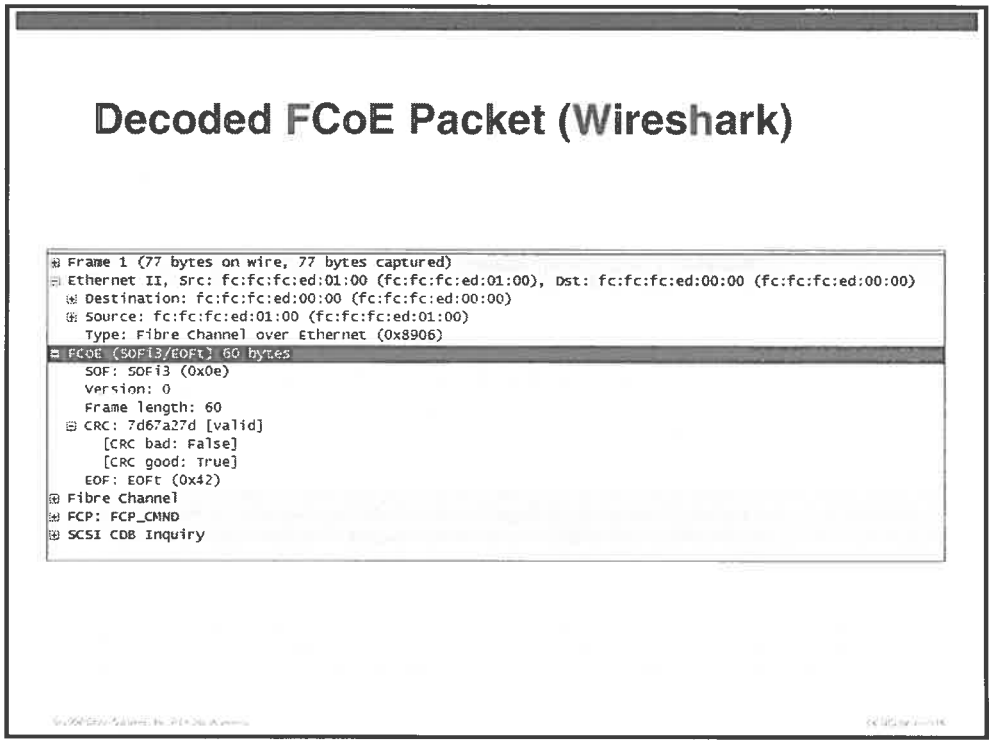
Security

Using IEEE 802.1Q Tags, Ethernet can be configured with multiple virtual LANs (VLANs) that partition the physical network into multiple separate and secure virtual networks. Using VLANs, FCoE traffic can be separated from IP traffic so that the two domains are isolated and one network cannot be used to view traffic on the other.

© 2008 Cisco Systems, Inc.

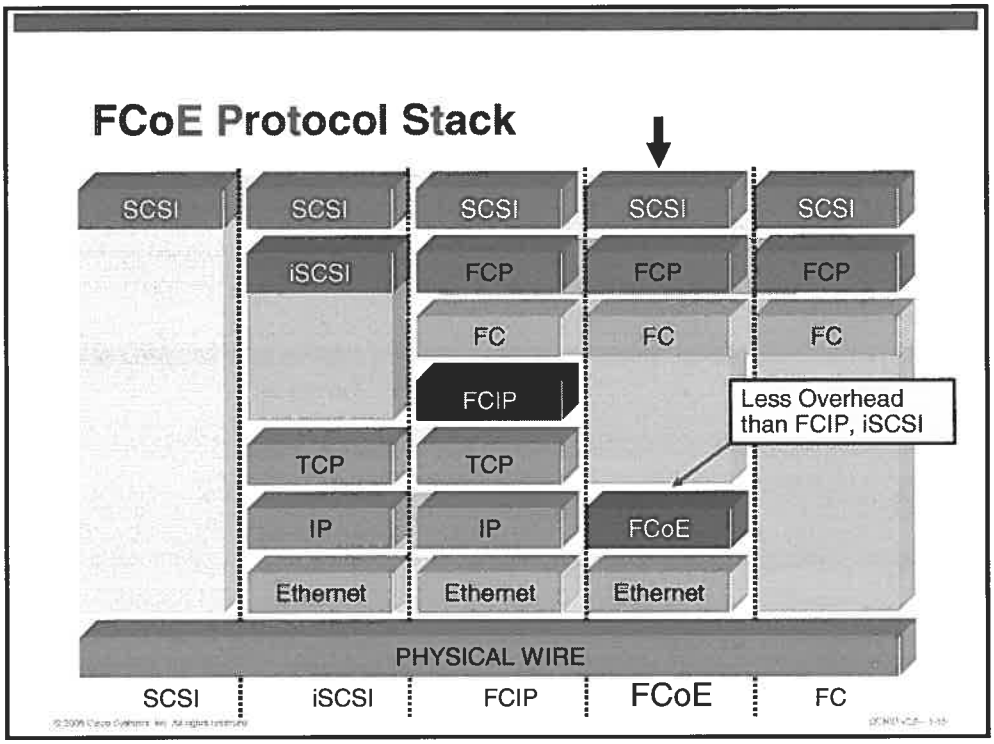
https://t.me/learningnets

Decoded FCoE Packet (Wireshark)



https://t.me/learningnets

FCoE Protocol Stack



Protocol Stack Comparison

The Small Computer Systems Interface (SCSI) is a mechanism that provides a low-overhead, high-performance parallel interface that is efficient in handling storage traffic within a chassis. Fibre Channel overcomes the distance and switching limitations inherent in SCSI. Fibre Channel carries SCSI as its higher-level protocol. SCSI does not respond well to lost frames, which can result in significant delays when recovering from a loss. Because Fibre Channel carries SCSI, it inherits the requirement for an underlying lossless network.

The ability of FCoE to work seamlessly with existing infrastructure makes it an evolutionary technology, one that data centers can deploy at the pace and to the extent that best serves their needs. FCoE allows IP and Fibre Channel network traffic to be carried over existing FCoE-aware drivers, NICs, and switches, allowing the use of a single cabling infrastructure within server racks. This technology simplifies network topology while reducing cabling cost and complexity, eliminating half of the I/O adapter cards in a rack, reducing power and cooling overhead—all while improving bandwidth by leveraging 10 Gbps Ethernet. As FCoE-enabled storage systems become available, data centers can implement a fully converged fabric, reaching from servers to storage, using FCoE-aware switches.

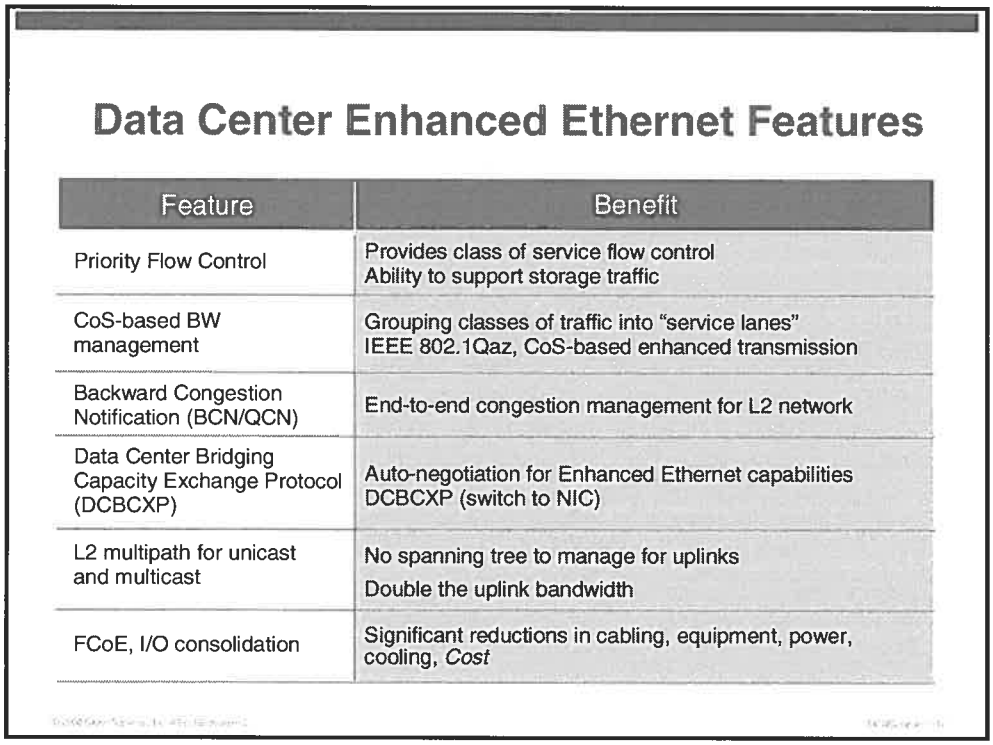
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enhanced Ethernet Features: Ethernet Extensions Used in FCoE Networks

This topic describes the Data Center Ethernet (DCE) or Enhanced Ethernet protocol features and characteristics as compared to the FCoE protocol.

Data Center Enhanced Ethernet Features



Extensions to the Ethernet specification give 10 Gb/s Ethernet the flexibility to support transmission mechanisms beyond Internet Protocol, including FCoE.

Classical Ethernet manages congestion by dropping packets, and higher-level, connection-based protocols recover from packet losses. The IEEE 802.3x PAUSE mechanism transforms Ethernet into a lossless fabric, allowing it to emulate Fibre Channel operation.

While the IEEE 802.3x PAUSE mechanism operates on a port level, Priority Flow Control operates on a "per-priority class" basis. The IEEE 802.1p Priority Flow Control mechanism further enhances classical Ethernet, enabling it to transport multiple traffic types simultaneously over a common, enhanced Ethernet fabric. With Priority Flow Control, Ethernet can support traffic including FCoE, high-performance computing, and classical Ethernet.

Priority Groups, implemented utilizing the IEEE 802.1Q specification, allows even greater QoS control among different "lanes" on the same physical cable. Priority Groups allow storage traffic to be managed as a group, with configurable QoS guarantees such as latency and bandwidth. This feature further enhances Ethernet's ability to accommodate Fibre Channel storage traffic on a common 10 Gb/s Ethernet fabric.

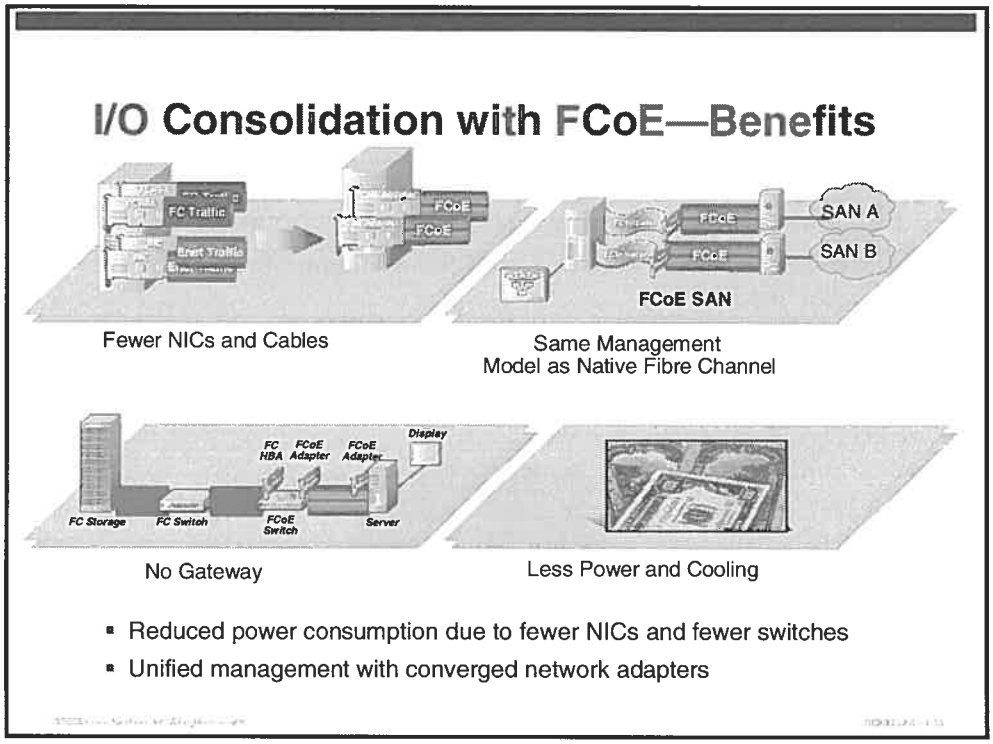
https://t.me/learningnets

Congestion management: In addition to the lossless fabric enabled by the PAUSE mechanism, a large network requires end-to-end congestion management. The IEEE 802.1Qau specification is being developed to support end-to-end congestion notification. With IEEE 802.1Qau, when congestion occurs within the interior of the network, traffic sources at the edges are instructed to throttle transmission, thus reducing congestion.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

I/O Consolidation with FCoE—Benefits



I/O convergence in the data center is rapidly becoming a reality. I/O consolidation supports SAN and LAN on a single network.

FCoE is an example of how the enhanced features of 10 GE can be leveraged for I/O consolidation in the data center.

10 GE is the enabling technology for I/O convergence because it offers the bandwidth and the underlying enhancements to support multiple I/O streams on a single cable. The rapid adoption of 10 Gbps Ethernet is driven by rapid growth in server performance, virtualization, and storage demand.

FCoE hosted on 10 Gb/s Ethernet extends the reach of Fibre Channel networks enabling connectivity from virtually every data center server to a centralized pool of storage.

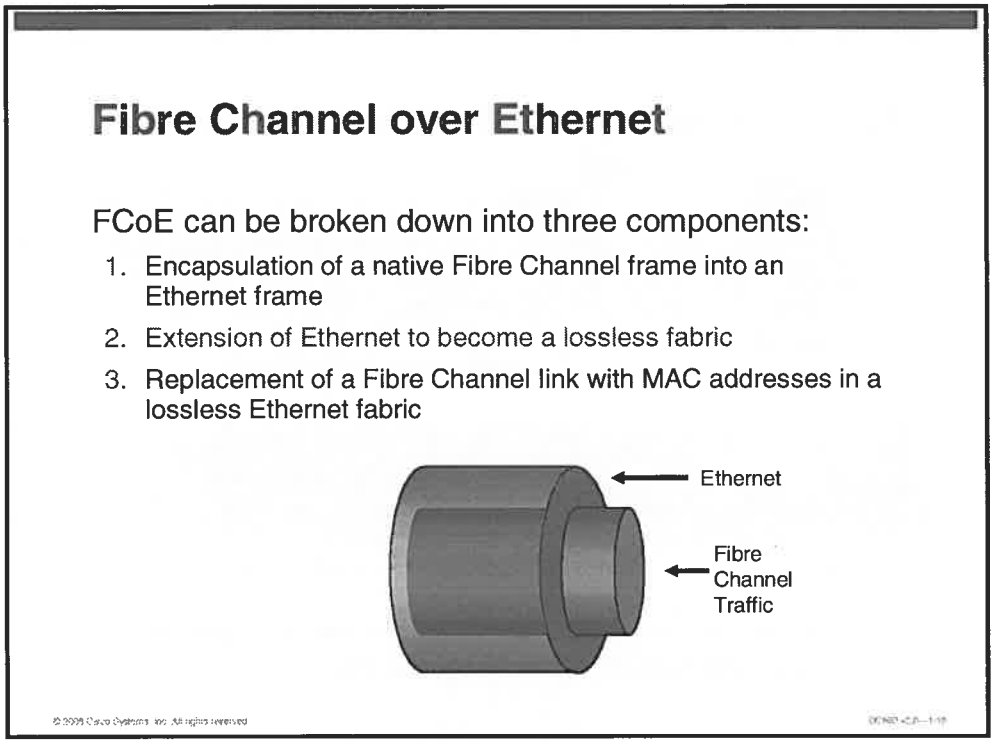
Unified management with CNAs: A new class of CNAs further helps to simplify FCoE deployment, allowing FCoE to be incorporated without disrupting current data center management practices, software, or the roles of network and storage administrators.

The reduced power consumption that comes from using fewer NICs and fewer switches provides relief to organizations that are at the limits of their data center power and cooling abilities.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel over Ethernet



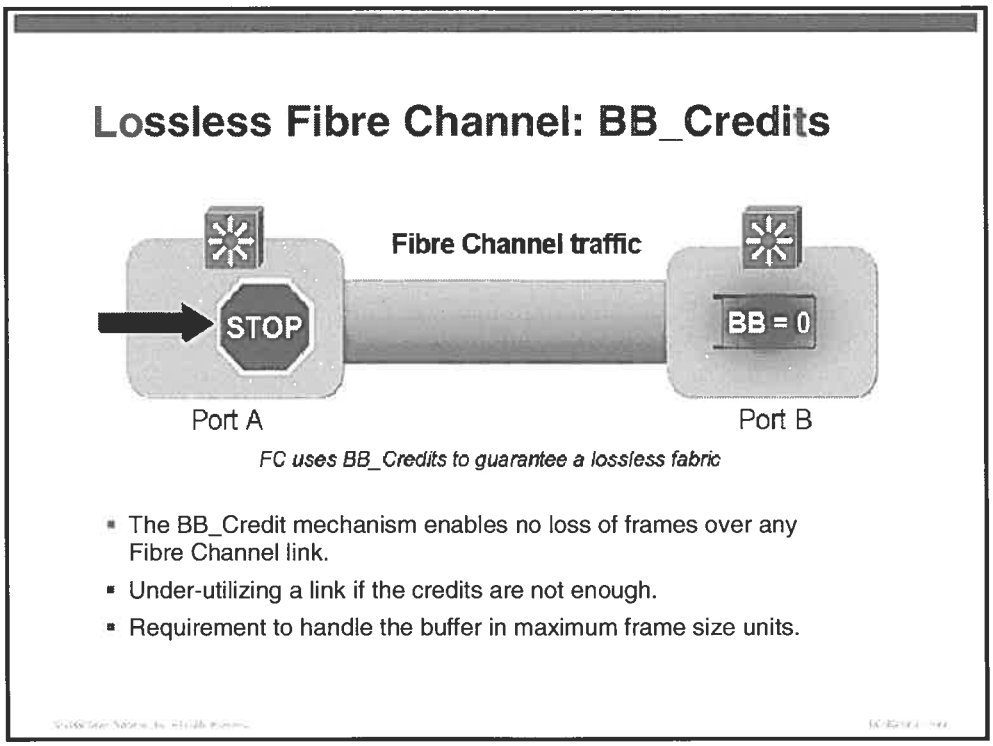
FCoE is encapsulated over Ethernet with the use of a dedicated Ethertype, 0x8906. A single 4-bit field (version) satisfies the IEEE sub-type requirements. The other bits in the frame specifically, the source mac address, destination mac address, VLAN tags, SOF and EOF are all encoded as specified in RFC 3643.

Reserved bits are present to guarantee that the FCoE frame meets the minimum length requirement of Ethernet. Inside the encapuslated Fibre Channel frame, the frame header is retained so as to allow connecting to a storage network by passing on the Fibre Channel frame directly after de-encapsulation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lossless Fibre Channel: BB_Credits



Conceptually, FCoE can be broken down into three components:

1. Encapsulation of a native Fibre Channel frame into an Ethernet frame
2. Extension of Ethernet to become a lossless fabric
3. Replacing of a Fibre Channel link with MAC addresses in a lossless Ethernet fabric

https://t.me/learningnets

Ethernet Pausing

Ethernet Pause Frame is a standard Ethernet Frame:

- Not VLAN tagged
- Applies to LINK, not per VLAN or per Priority
- EtherType = 0x8808
- OPC code=0x0001
- Pause Time field
- Reports time the link needs to pause the transmission
- Un-pause the link
- Frame with Pause Time=0

Note:

- Ethernet assumes receiver has availabile buffers unless Pause is received
- Fibre Channel B2B credit mechanism knows the number of receivers buffers

There are three possible causes of frame drop in Ethernet networks:

1. Frame Errors (CRC)

   The probability of frame errors in Fibre Channel and Ethernet are the same.

2. Collisions

   Frames cannot be transmitted due to collisions on wired media.10GE is full duplex and does not have collisions.

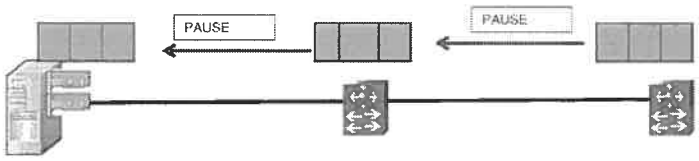
3. Congestion in Switch that causes buffer overflow

   Buffer overflows are avoided in Fibre Channel using a B2B credit mechanism. The Ethernet equivalent functionality is accomplished using PAUSE Frames in Ethernet IEEE 802.3, Annex 31B.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

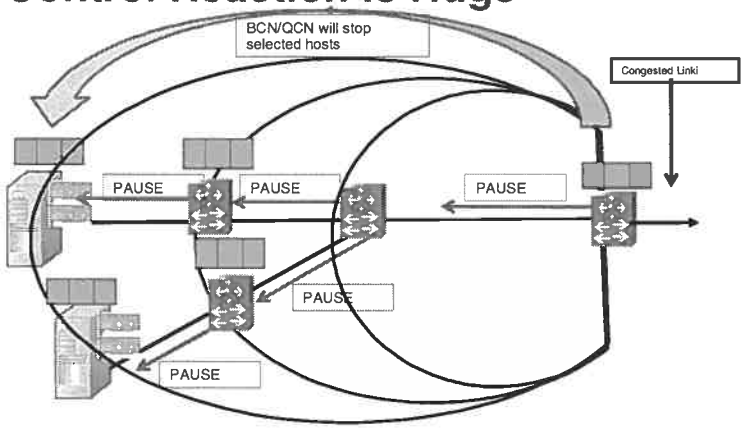
Ethernet Pause Propagation



- The goal is to suspend transmission of frames so that the receiver will not drop them due to congestion.
- When the receiveing Queue reaches the threshold, the Switch sends Pause back to the sender.
- The Pause is propagated back to the servers using the same congested link.
- Fibre Channel/FCoE works better with losseless fabric.
- Short TCP flows work better with lossless fabric since the first and last frames in the flow are protected.
- Otherwise, TCP relies on losses to adjust Window size.
  - Therefore, PFC is needed.

The previous figure shows how Ethernet PAUSE propagates through the network.

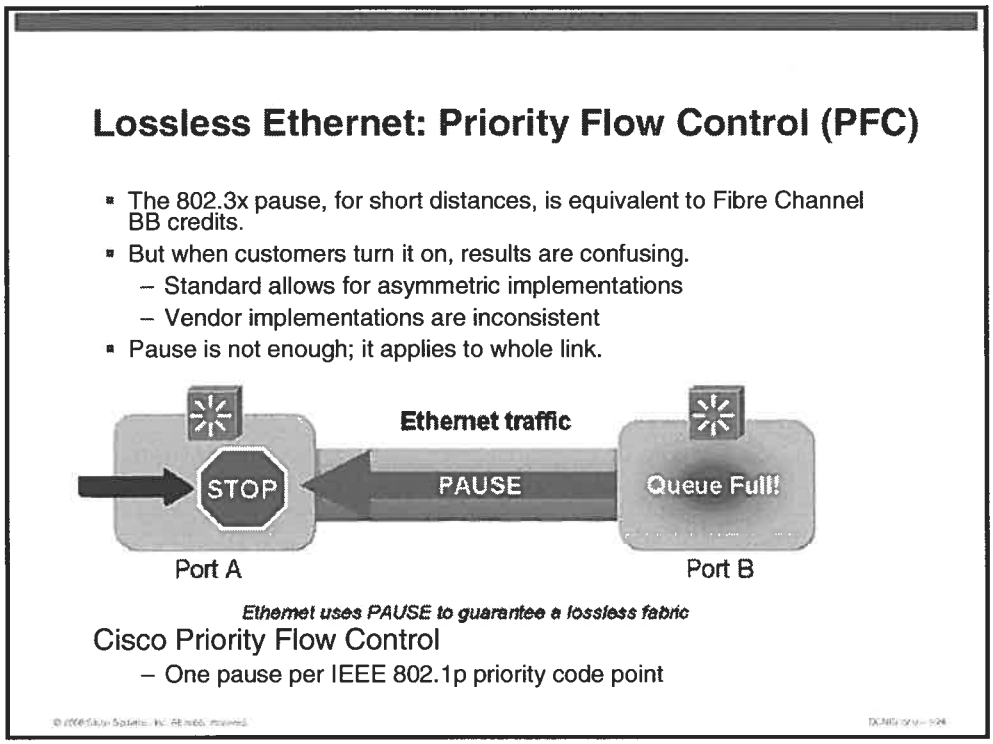
Ethernet Pausing/Fibre Channel B2B Flow Control Reaction is Huge



Priority flow control will stop only the problematic source, whereas Ethernet pausing will stop all servers in the back utilizing the same congested ISL. BCN/QCN propagates faster then Ethernet Pausing.

BCN improves reaction to congestion because it stops selected source of congestion, and propagates faster than Ethernet pausing.

https://t.me/learningnets



Priority Flow Control for Lossless Ethernet

While the IEEE 802.3x PAUSE mechanism operates on a port level, Priority Flow Control (PFC) operates on a "per-priority class" basis. The IEEE 802.1p PFC mechanism further enhances classical Ethernet, enabling it to transport multiple traffic types simultaneously over a common, enhanced Ethernet fabric. With PFC, Ethernet can support traffic including FCoE, high-performance computing, and classical Ethernet.

One of the challenges with passing Fibre Channel frames over Ethernet is that Fibre Channel provides a lossless transport. Congestion in a lossless network is an issue that must be managed. Traditional Fibre Channel manages congestion through the use of a link-level, credit-based flow control that guarantees no loss of a frame under normal conditions. Typical Ethernet, coupled with TCP/IP, uses a packet drop flow control mechanism. Packet drop flow control is not lossless, making it unacceptable for use with storage traffic. Fortunately, classical Ethernet has a PAUSE capability so that a busy receive port can send a control frame to the transmit port requesting a PAUSE in transmission. Use of this optional IEEE 802.3x Ethernet feature enables Fibre Channel traffic to pass on an Ethernet network in a lossless fabric.

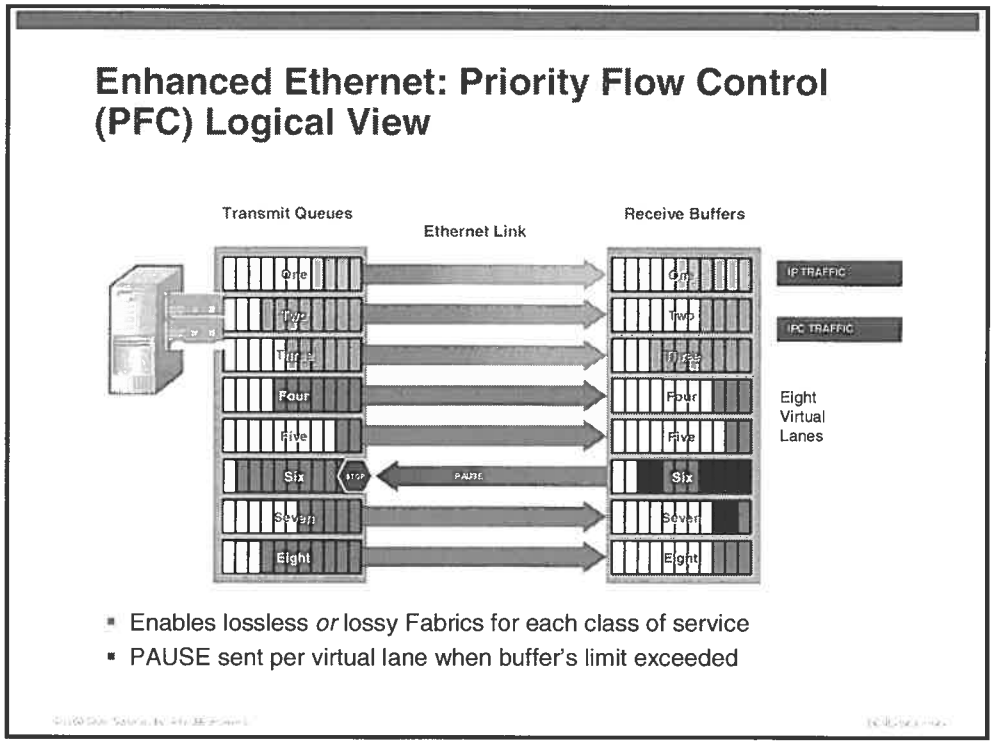
Cisco PFC

- One PAUSE per IEEE 802.1p priority code point
- Public domain (no standard activity yet)
- Embraced by many network gear vendors

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enhanced Ethernet: PFC



IEEE 802.3x PAUSE provides no drop flow control.

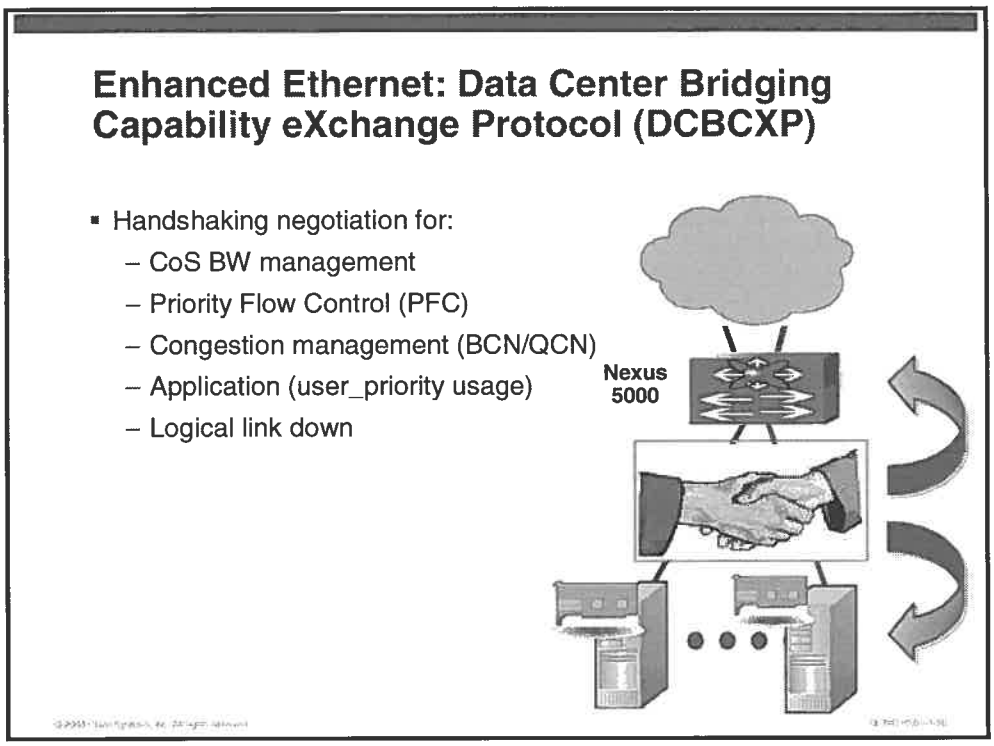
- The 802.3x frame is an actual Ethernet control frame.
- This feature transforms Ethernet to a lossless fabric.

PFC is a fine-grain mechanism of flow control.

- PFC enables PAUSE functionality on traffic mapped to a specific system class (CoS).
- PFC allows PAUSE of FCoE and Ethernet traffic separately.

https://t.me/learningnets

Enhanced Ethernet: Data Center Bridging Capability eXchange Protocol

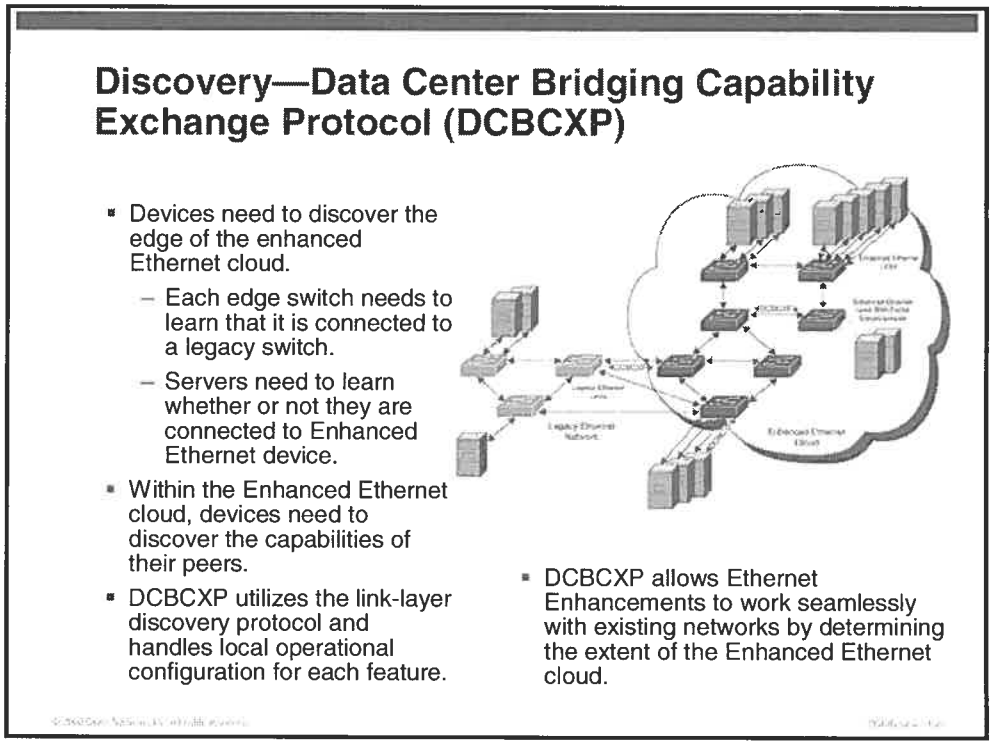


This figure illustrates Data Center Bridging Capability eXchange Protocol (DCBCXP).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Discovery—DCBCXP



DCBCXP allows Ethernet Enhancements to work seamlessly with existing networks by determining the extent of the Enhanced Ethernet cloud.

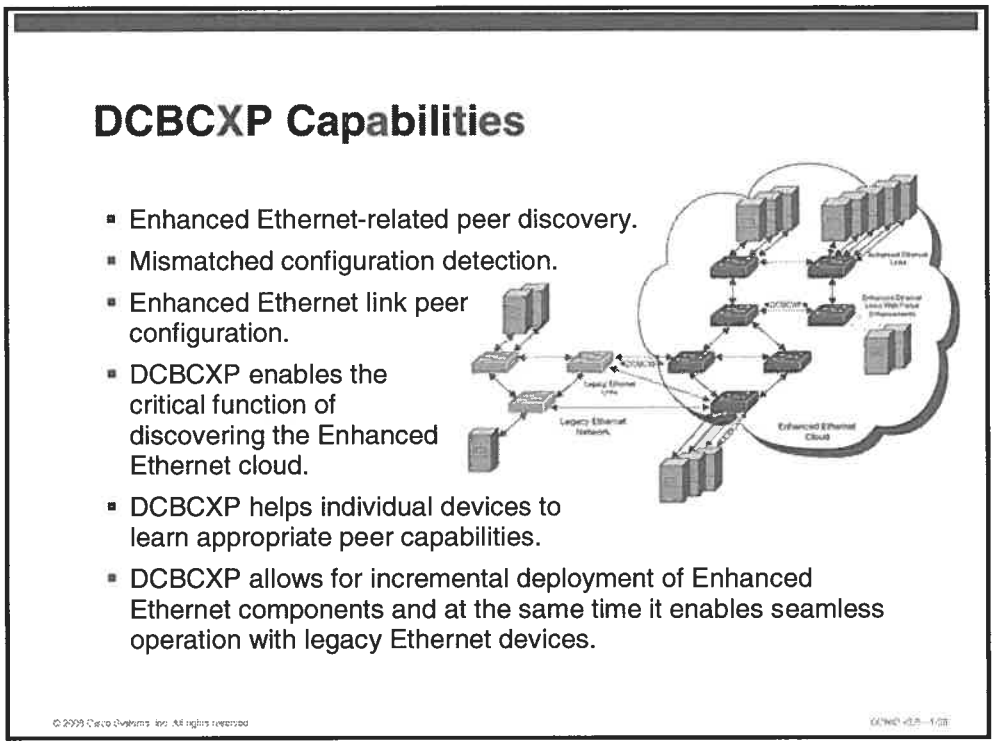
The enhancements presented so far enable I/O consolidation on an Ethernet network. For these enhancements to work seamlessly with existing Ethernet deployments, a management protocol that dynamically discovers the capabilities of its peers is needed. DCBCXP provides the needed configuration and discovery functionality.

Consider the typical Ethernet network depicted previously. Multiple devices of varying capabilities are connected to each other. In such a network there are two distinct scenarios where dynamic configuration management is required. First, devices need to discover the edge of the enhanced Ethernet cloud. This means that each edge switch needs to learn that it is connected to a legacy switch. Also, servers need to learn whether or not they are connected to Enhanced Ethernet device. Second, within the Enhanced Ethernet cloud, devices must discover the capabilities of their peers.

DCBCXP utilizes the link-layer discovery protocol and defines new Type-Length Values (TLVs) for capability exchange parameters. The protocol state machines handle local operational configuration for each feature by comparing and synchronizing with the peer's feature settings. The protocol provides the following capabilities (see next section, "DCBCXP Capabilities").

https://t.me/learningnets

DCBCXP Capabilities



The protocol provides the following capabilities:

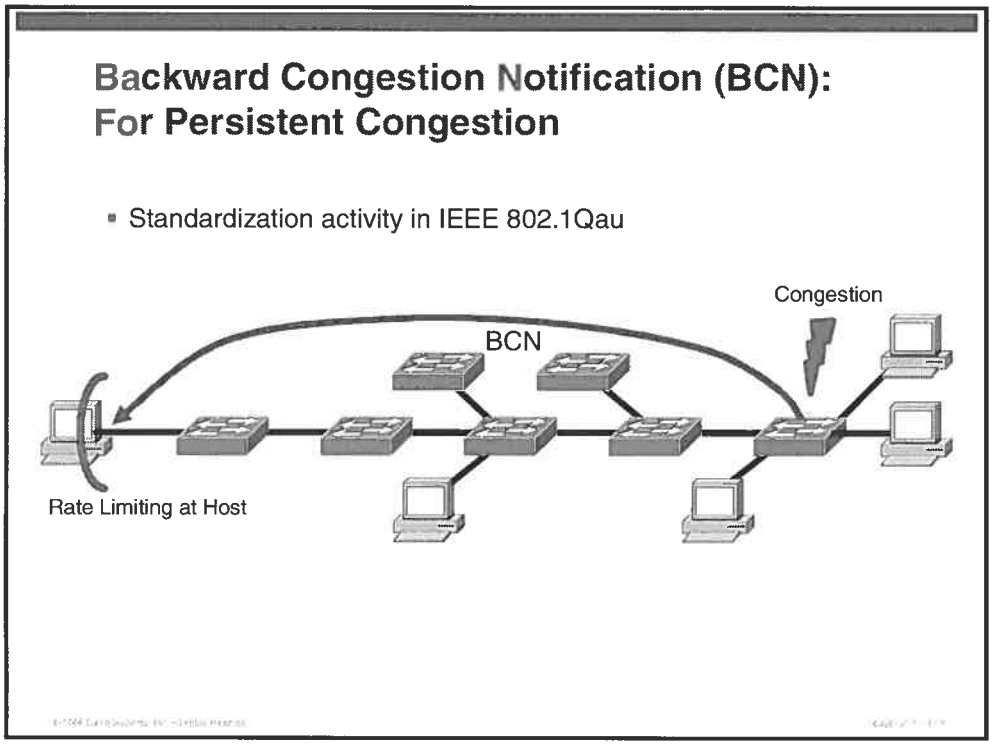
- Enhanced Ethernet-related peer discovery allows Enhanced Ethernet devices to discover whether a peer device supports particular features such as Congestion Notification or Priority Flow control. Appropriate action can be based on the compatible features.
- Mismatched configuration detection can determine if the bandwidth groups and bandwidth allocations on both ends of the link are the same. This information can be used to notify the device manager in case of a conflict.
- Enhanced Ethernet link peer configuration can be used by switch devices to push configuration parameters uplink to Server adapters or the reverse.

DCBCXP enables the critical function of discovering the Enhanced Ethernet cloud. It helps individual devices to learn appropriate peer capabilities. Furthermore, DCBCXP allows for incremental deployment of Enhanced Ethernet components and, at the same, time it enables seamless operation with legacy Ethernet devices.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Backward Congestion Notification: For Persistent Congestion



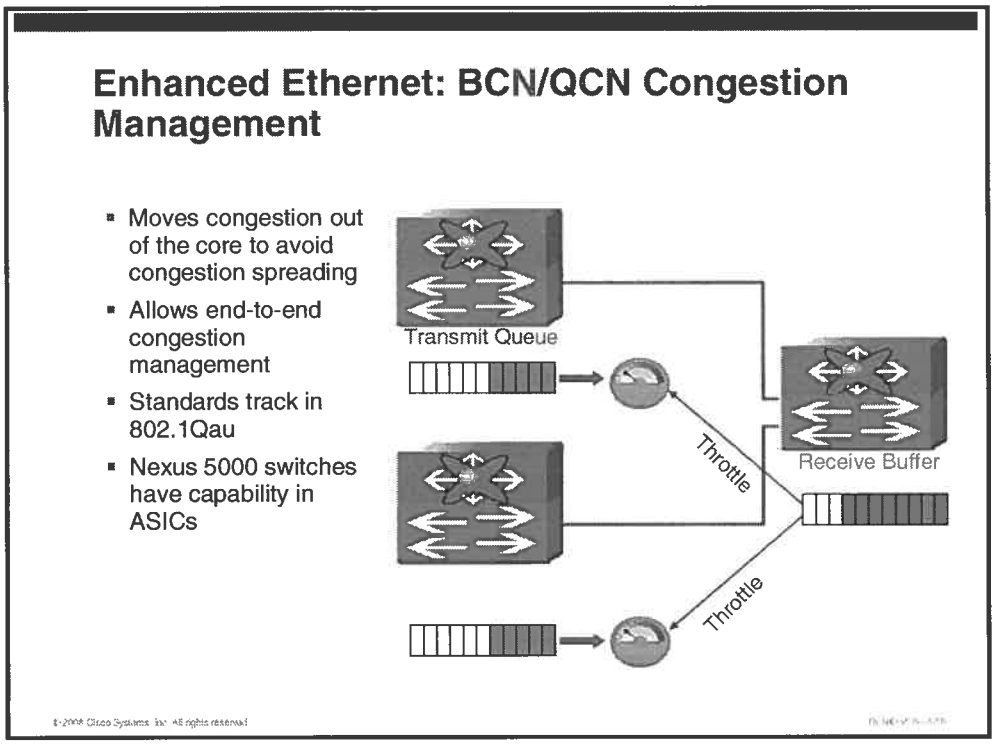
This figure illustrates Backward Congestion Notification (BCN).

In addition to the lossless fabric enabled by the PAUSE mechanism, a large network requires end-to-end congestion management. The IEEE 802.1Qau specification is being developed to support end-to-end congestion notification. With IEEE 802.1Qau, when congestion occurs within the interior of the network, traffic sources at the edges are instructed to throttle transmission, thus reducing congestion.

Congestion notification is a form of traffic management that pushes congestion to the edge of a network by notifying upstream rate limiters to shape the traffic causing congestion. The Cisco Quantized Congestion Notification (QCN) proposal model defines a two-point architecture with a congestion point and a reaction point. In this architecture, congestion is measured at the congestion point and rate-limiting, or back pressure, is imposed on the reaction point to shape traffic and reduce the impact of congestion.

https://t.me/learningnets

Enhanced Ethernet: BCN/QCN Congestion Management

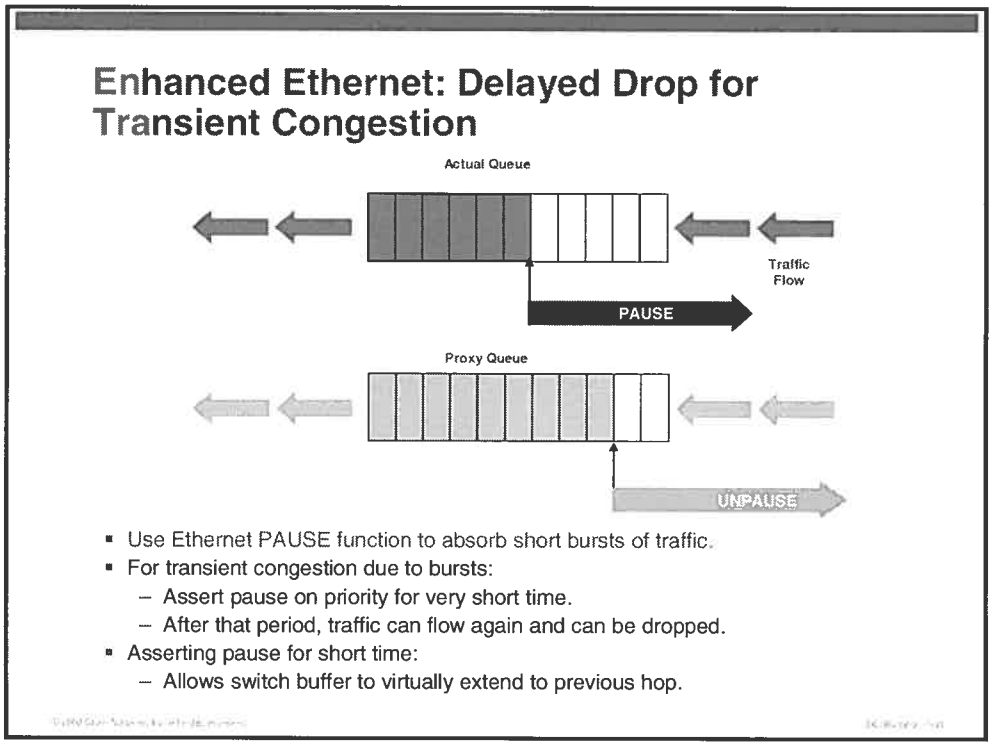


The Cisco Nexus 5000 Series Switches have capability in ASICs, in future OS releases.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

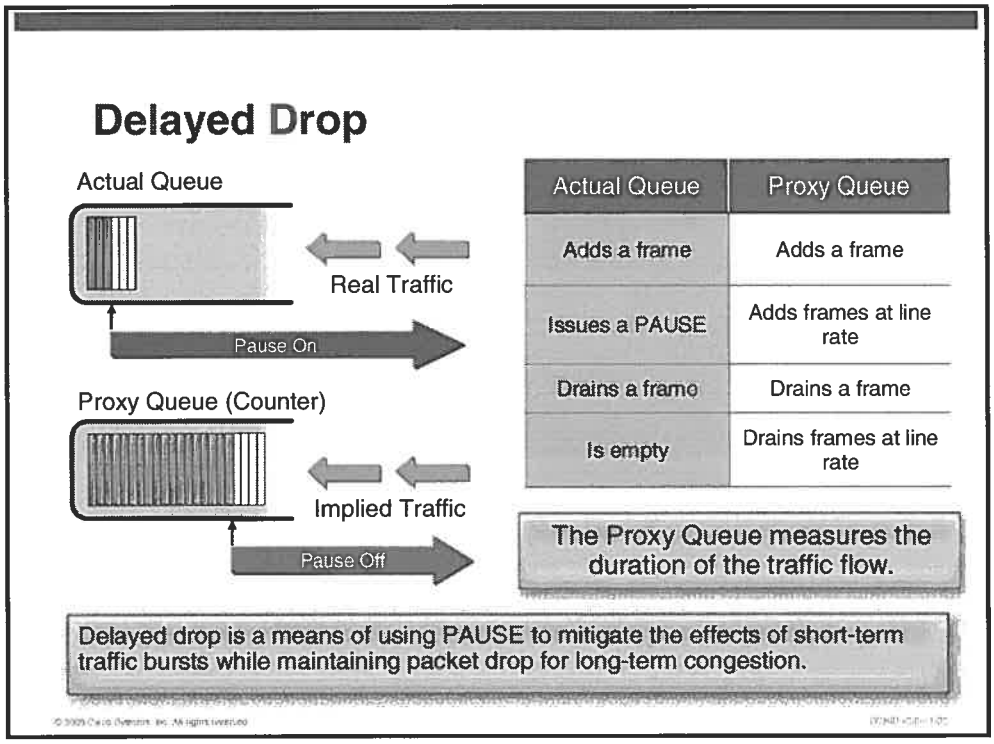
Enhanced Ethernet: Delayed Drop for Transient Congestion



Delayed Drop is an enhancement. It uses the PAUSE mechanism to reduce packet drop on short-lived traffic bursts while triggering upper-layer congestion control through packet drops to handle long-term congestion. This technology allows congestion to spread only for short-term bursts and alleviates it for long-term congestion. Delayed Drop can be individually enabled on each user priority.

https://t.me/learningnets

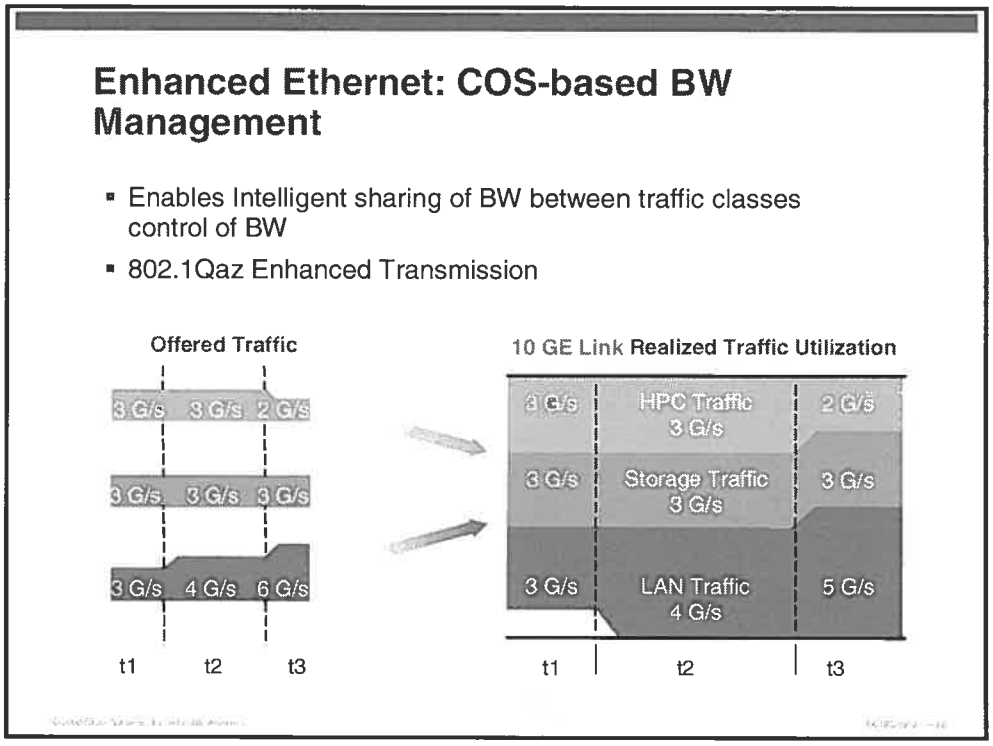
Delayed Drop



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enhanced Ethernet: COS-based BW Management



CoS-based Bandwidth Management (IEEE 802.1Qaz)

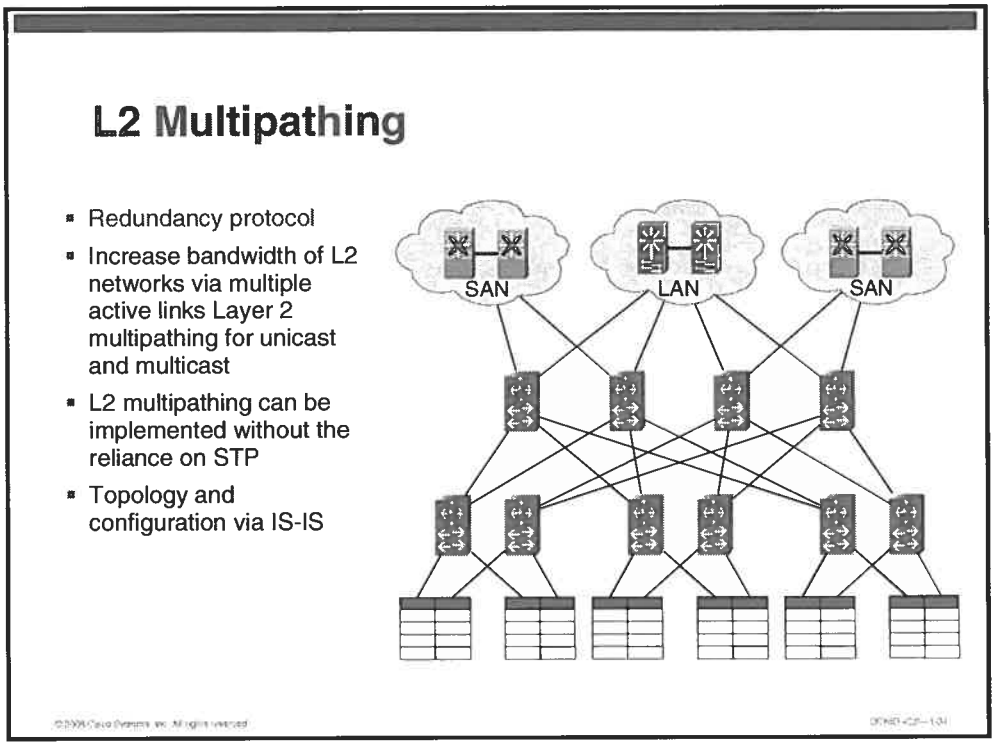
- Most customers want simple, consistent QoS.
- IEEE 802.1p defines Strict Priority link scheduling.
  - All vendors support some WRR/DWRR/etc variant
    - Configuration and interworking vendor dependent
- Proposal for HW efficient two-level DWRR with strict priority punch-through.
  - Consistent behavior and configuration across network elements
  - Project expected soon in IEEE 802.1Qaz

QoS Highlights

- Modular QoS CLI-compliant
- Virtual Output Queuing
- Flexible queue management
- Class Based Lossless Data Path
- Newer schemes of packet drop and buffer management
- Priority Flow Control

https://t.me/learningnets

L2 Multipathing



Layer 2 multipathing can be implemented without the reliance on Layer 2 redundancy protocols, such as Spanning Tree Protocol.

Increase bandwidth of L2 networks via multiple active links: Common in IP networks, switch fabrics, and so on in release 2, Ethernet End Host Virtualizer.

- TRILL in IETF
  - Encapsulate frames in transport header
    - Unicast frames target an edge RBridge
    - Multicast frames target a multicast fabric

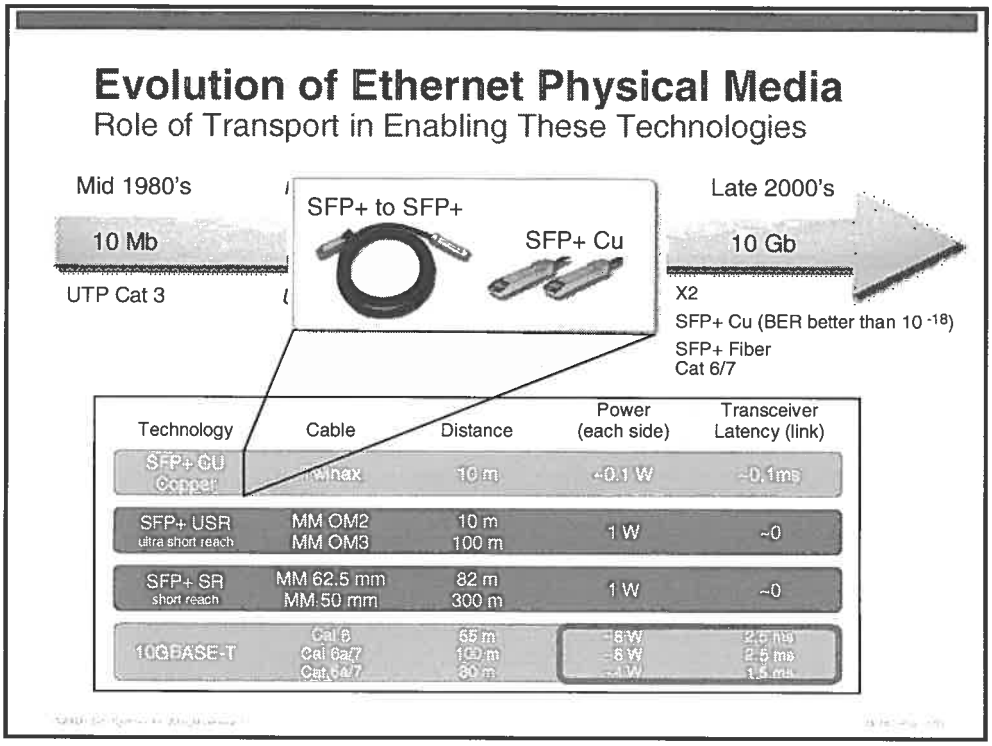
Layer 2 multipathing is available in post-FCS.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Evolution of Ethernet Physical Media

Role of Transport in Enabling these Technologies



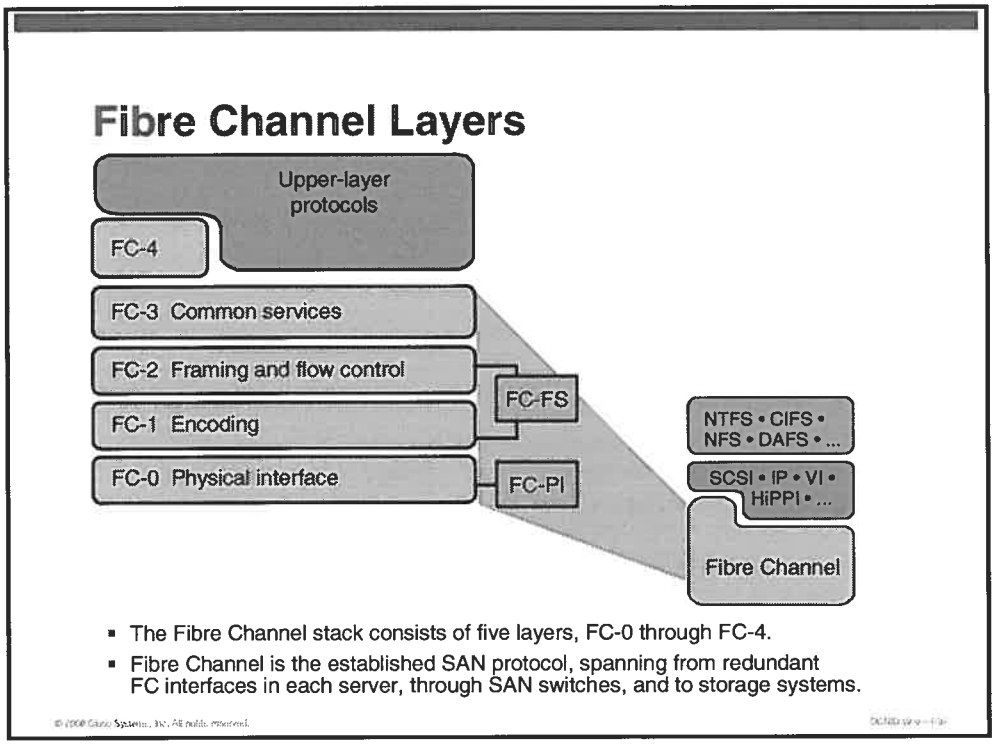
Server virtualization created the need to look into the network virtualization.

https://t.me/learningnets

Fibre Channel Protocol

This topic describes the Fibre Channel protocol.

Fibre Channel Layers



Most large data centers have existing SANs that are supported by Fibre Channel fabric.

The Small Computer Systems Interface (SCSI) is a mechanism that provides a low-overhead, high-performance parallel interface that is efficient in handling storage traffic within a chassis.

Fibre Channel was developed to overcome the distance and switching limitations inherent in SCSI.

Fibre Channel carries SCSI as its higher-level protocol.

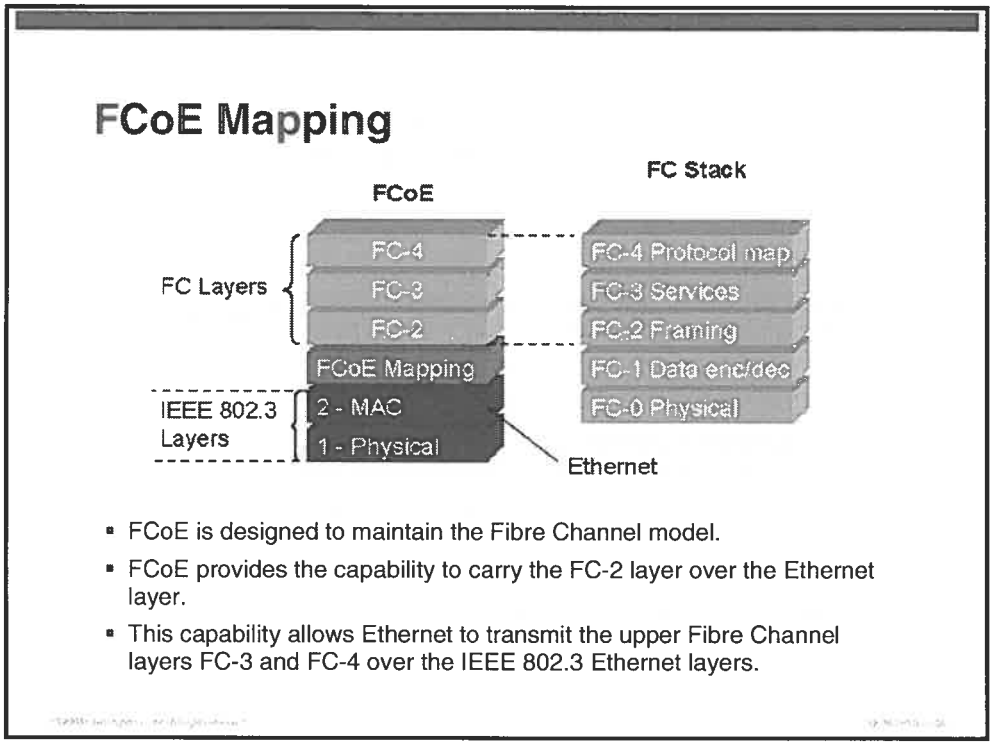
Fibre Channel consists of five layers:

1. FC0 physical layer, which includes cables, fiber optics, connectors, pin outs.
2. FC1 data link layer, which implements the 8b/10b encoding and decoding of signals.
3. FC2 network layer, defined by the FC-PI-2 standard, consists of the core of Fibre Channel, and defines the main protocols.
4. FC3 common services layer, a thin layer that could eventually implement functions like encryption or RAID.
5. FC4 Protocol Mapping layer. Layer in which other protocols, such as SCSI, are encapsulated into an information unit for delivery to FC2.

© 2008 Cisco Systems, Inc.

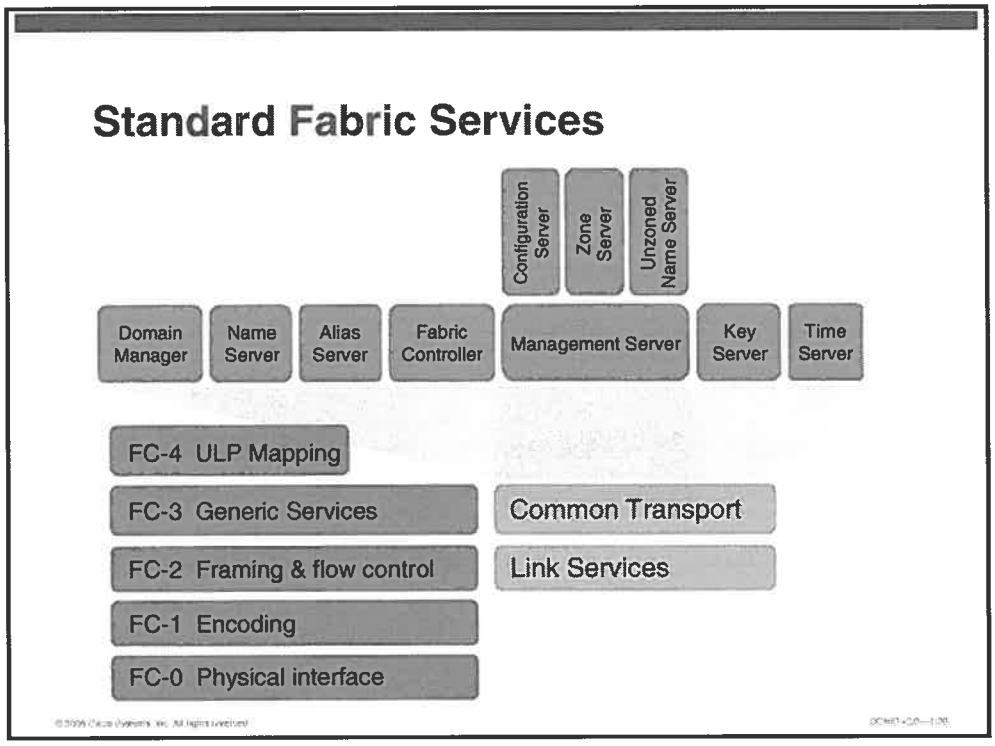
https://t.me/learningnets

FCoE Mapping



https://t.me/learningnets

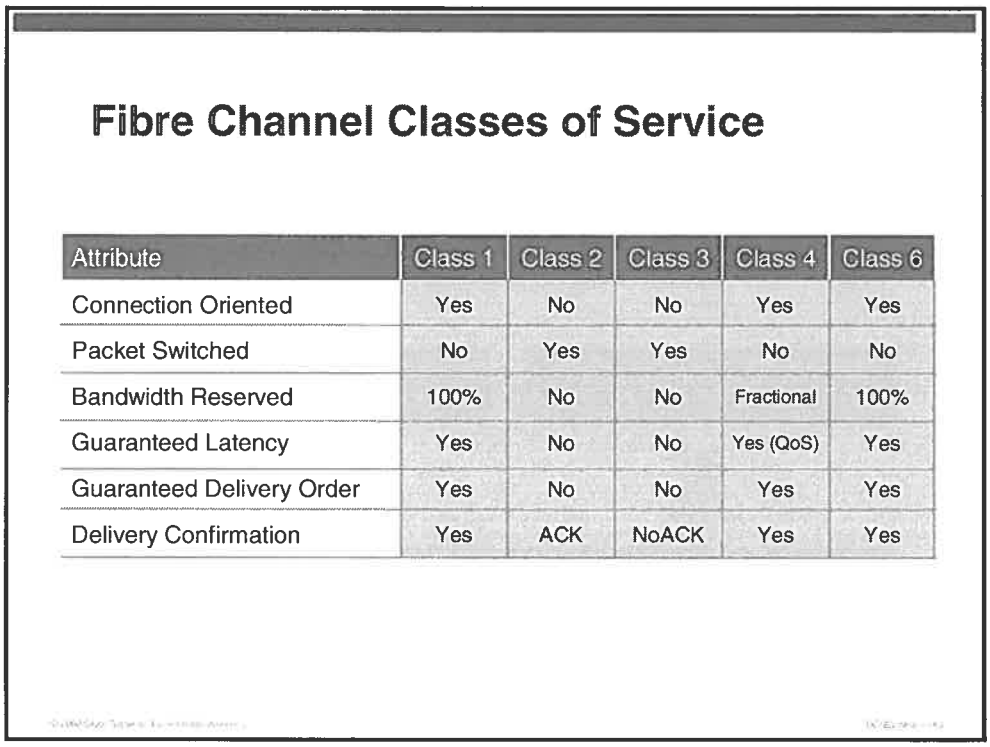
Standard Fabric Services



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

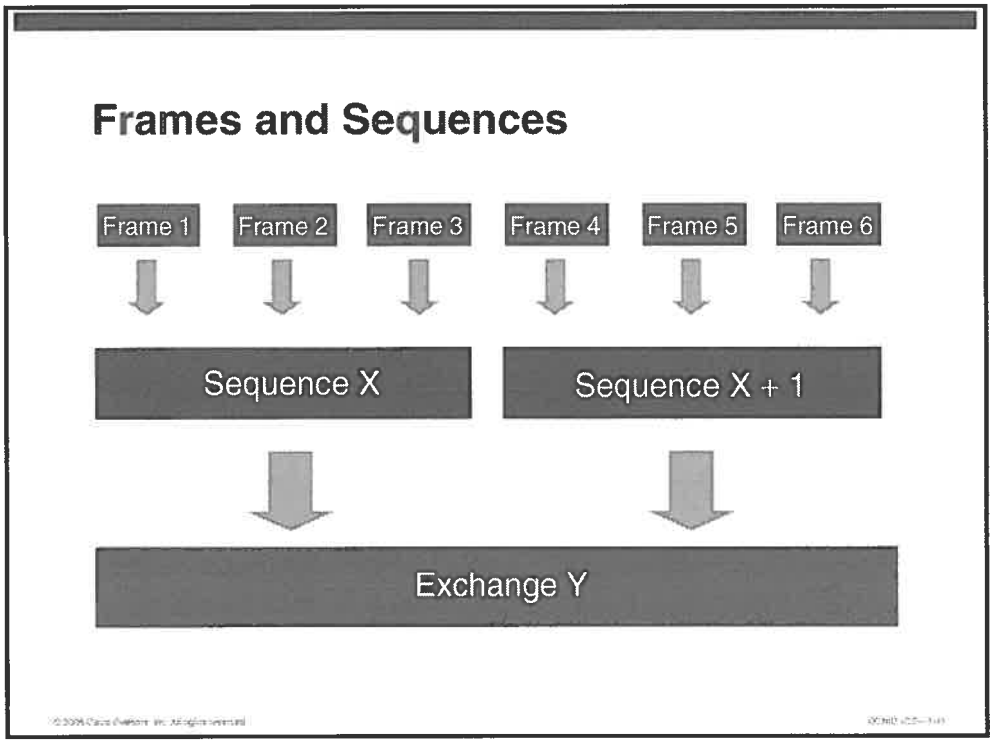
Fibre Channel Classes of Service



Attribute	Class 1	Class 2	Class 3	Class 4	Class 6
Connection Oriented	Yes	No	No	Yes	Yes
Packet Switched	No	Yes	Yes	No	No
Bandwidth Reserved	100%	No	No	Fractional	100%
Guaranteed Latency	Yes	No	No	Yes (QoS)	Yes
Guaranteed Delivery Order	Yes	No	No	Yes	Yes
Delivery Confirmation	Yes	ACK	NoACK	Yes	Yes

https://t.me/learningnets

Frames and Sequences



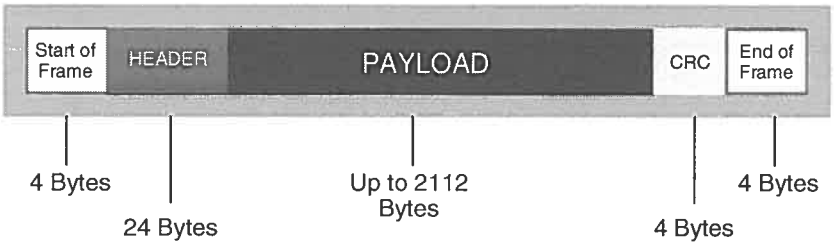
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel Frame

Fibre Channel Frame

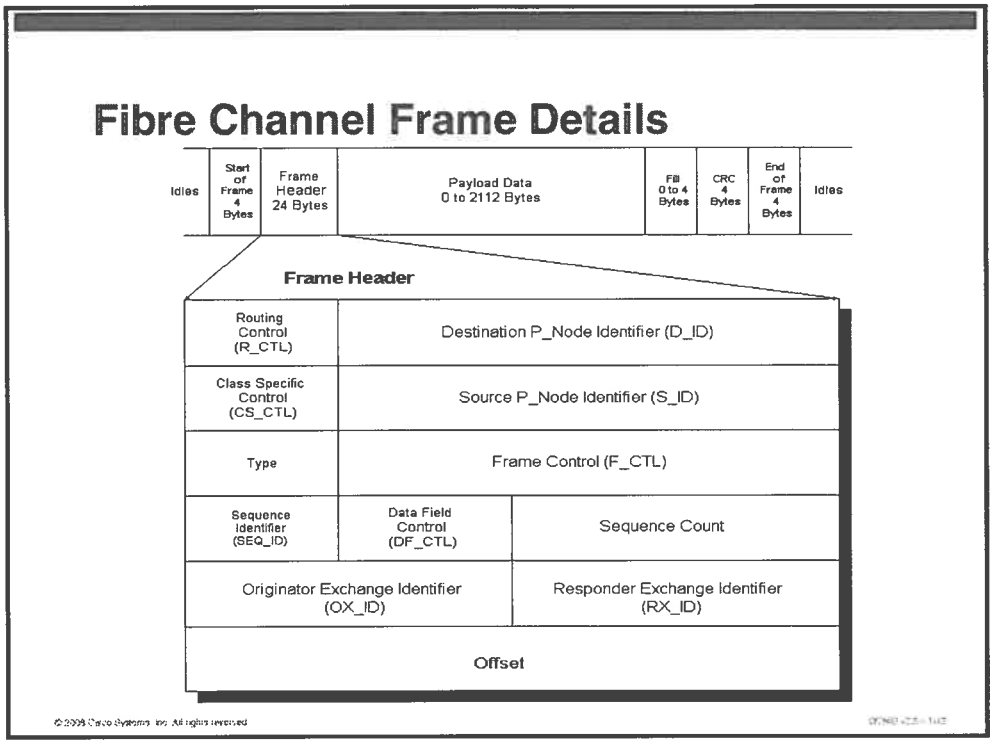
Fibre Channel frames can have up to 2128 bytes.



Fibre Channel Frame Format

https://t.me/learningnets

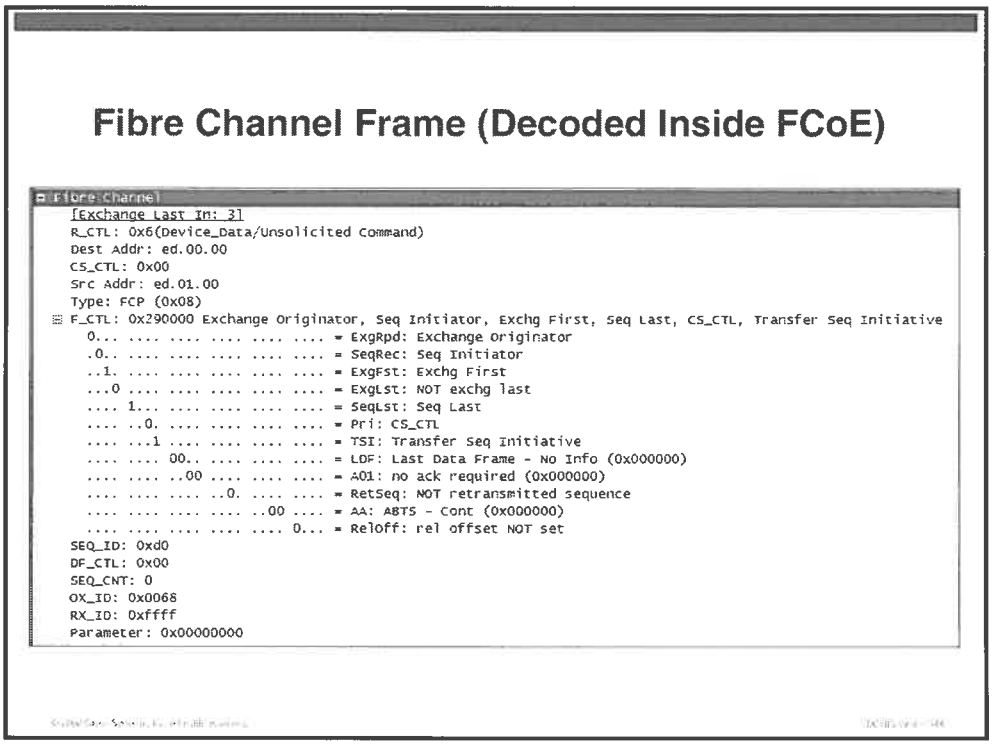
Fibre Channel Frame Details



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel Frame (Decoded Inside FCoE)



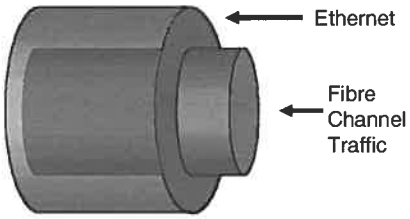
https://t.me/learningnets

Fibre Channel over Ethernet

Fibre Channel over Ethernet

FCoE can be broken down into three components:

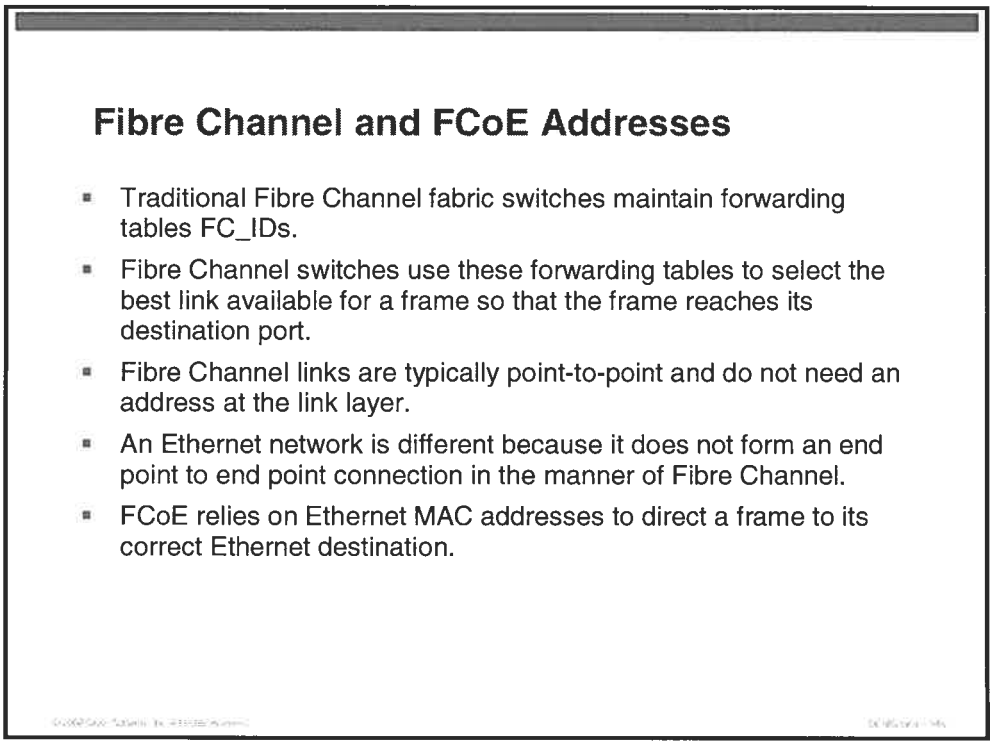
1. Encapsulation of a Native Fibre Channel Frame into an Ethernet frame
2. Extension of Ethernet to become a lossless fabric
3. Replacement of a Fibre Channel link with MAC addresses in a lossless Ethernet fabric



© 2008 Cisco Systems, Inc. All rights reserved.

https://t.me/learningnets

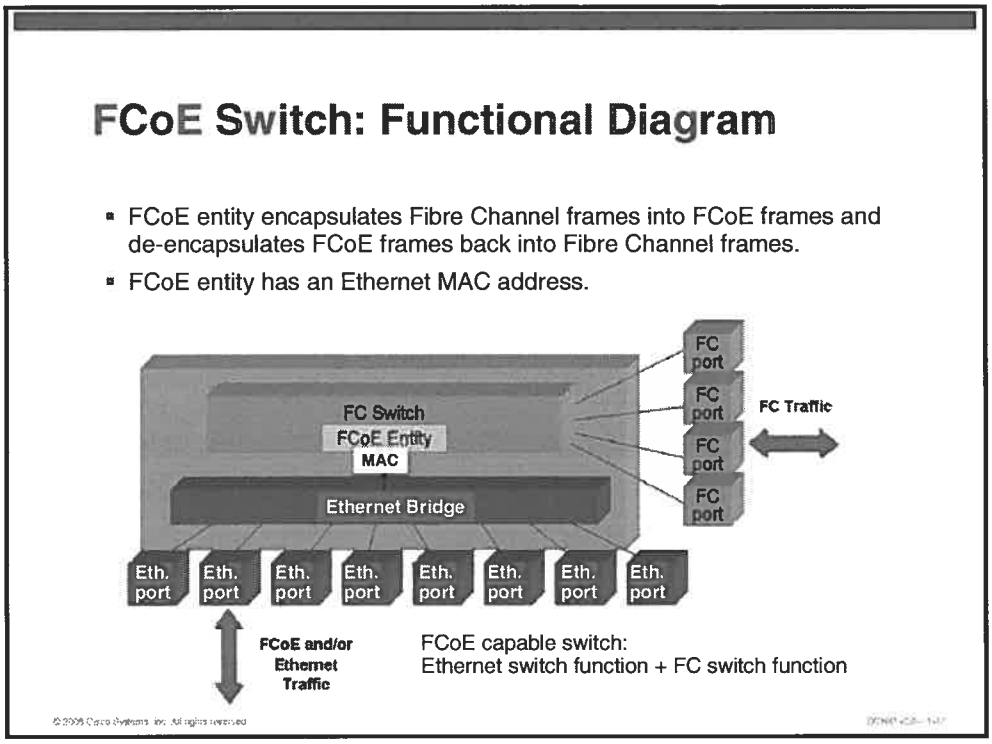
Fibre Channel and FCoE Addresses



The first two fields of the FCoE frame are MAC addresses.

https://t.me/learningnets

FCoE Switch: Functional Diagram



Fibre Channel traffic flows into one end of the FCoE entity and FCoE traffic flows out the other end.

The FCoE Entity has an Ethernet MAC address that will be used as a source or destination address when FCoE frames are transmitted through an Ethernet fabric.

In the preceding figure, a Fibre Channel frame enters from the green FC ports to the right and into the Fibre Channel switch, where the entity encapsulates the Fibre Channel frame into an FCoE frame. The FCoE frame is then sent to an Ethernet Bridge, where it is distributed to the appropriate Ethernet port on the bottom of the diagram.

The reverse is also true: FCoE traffic will flow from the red ports, through the Ethernet Bridge until it is de-encapsulated inside of the FCoE Entity, where it passes through the FC switch to the appropriate FC port.

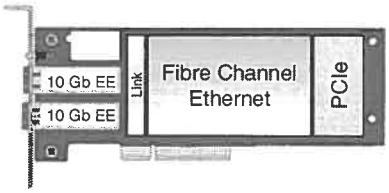
This is just one example of how a switch can be FCoE-enabled. Different vendors may choose to implement FCoE in different ways.

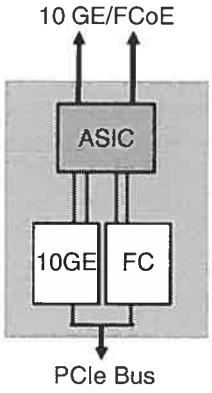
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Consolidated Network Adapter: Fibre Channel and Ethernet

Consolidated Network Adapter: Fibre Channel and Ethernet

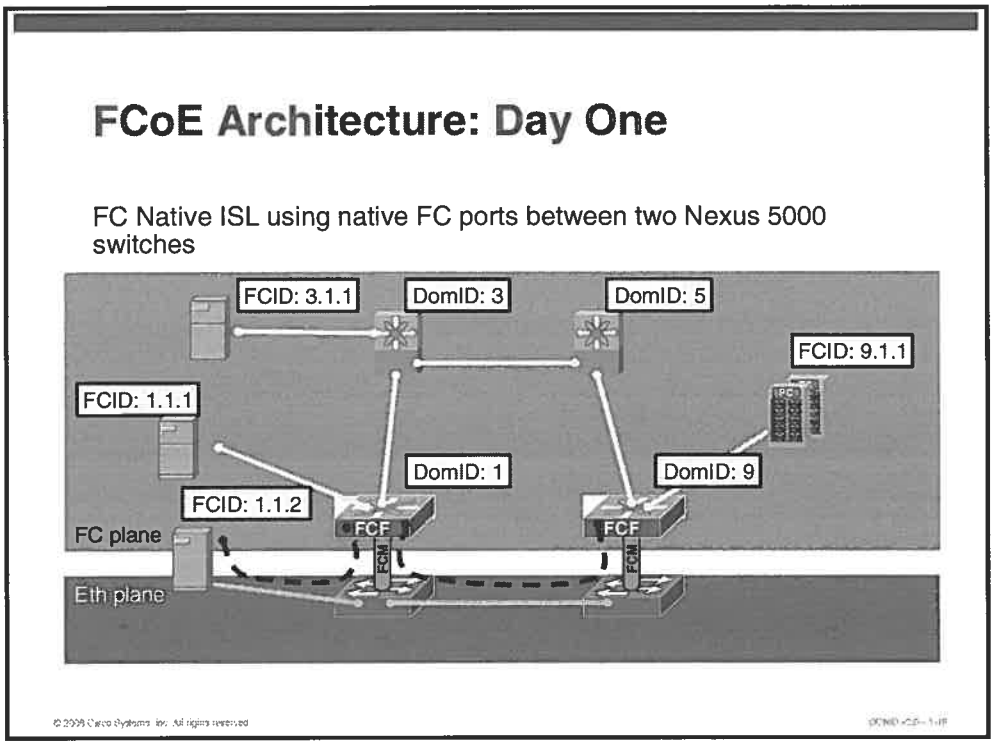




- Off-the-shelf NIC and HBA ASICs from Qlogic, Emulex
- Dual 10 GE/FCoE ports
- Replaces multiple adapters per server
- Consolidates 10 GE and Fibre Channel on single interface
- Minimum disruption in existing environments

https://t.me/learningnets

FCoE Architecture: Day One

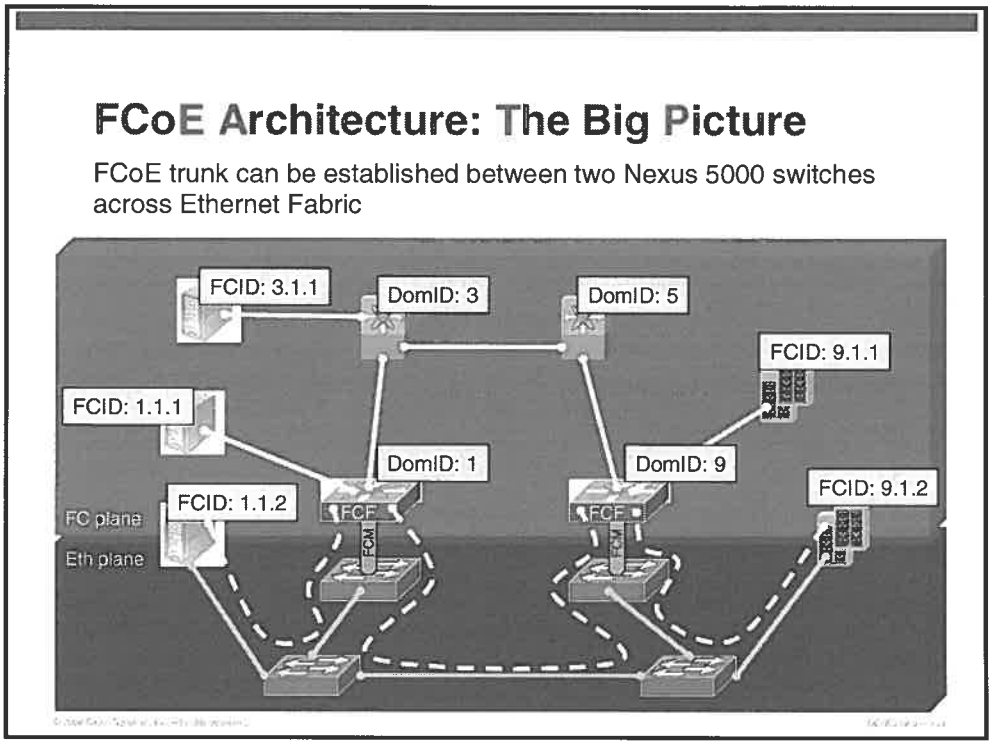


Fibre Channel Native ISL using native Fibre Channel ports between two Nexus 5000 Series switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

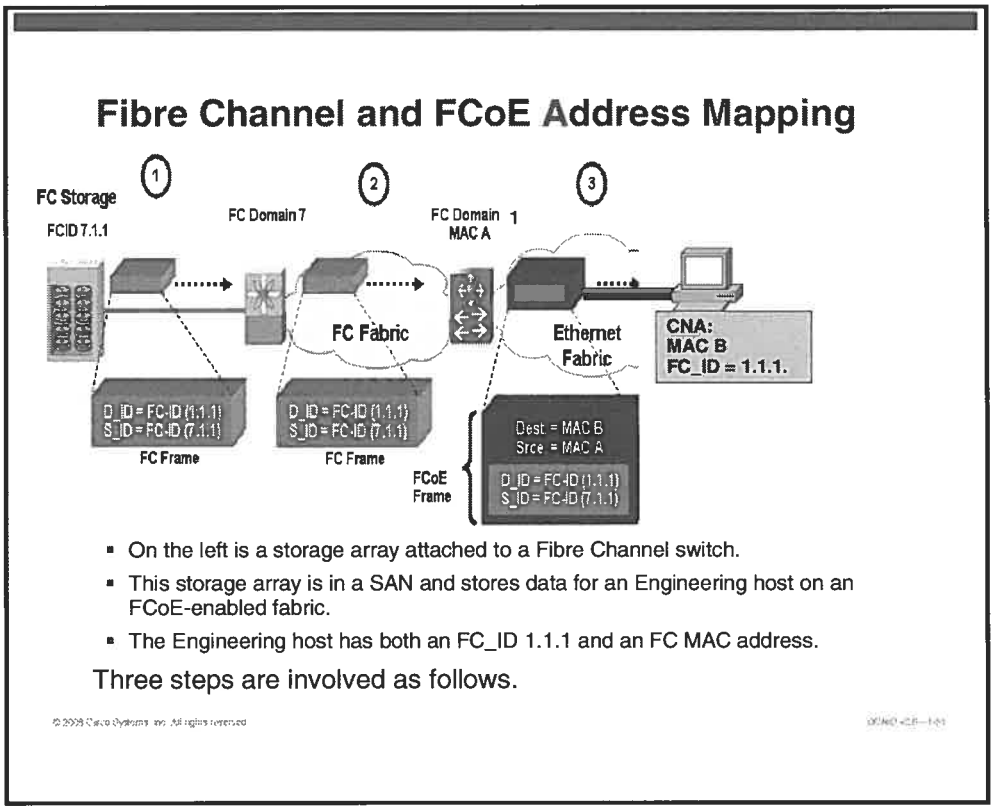
FCoE Architecture: The Big Picture



FCoE trunk can be established between two Nexus 5000 Series switches across Ethernet Fabric.

https://t.me/learningnets

Fibre Channel and FCoE Address Mapping



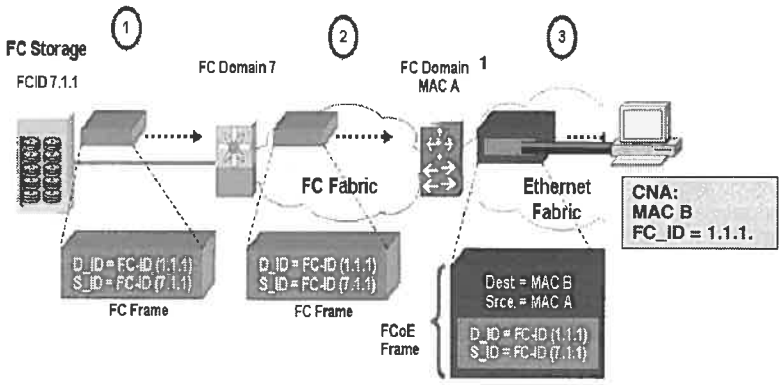
This slide demonstrates how Fibre Channel addresses are mapped to MAC addresses in an FCoE network. FCoE address mapping to traditional Fibre Channel addressing is also shown. Topologies will vary depending on an implementation requirement and the FCoE products deployed. While the topologies will change, the addressing essentials will remain constant.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel and FCoE Address Mapping—Step 1

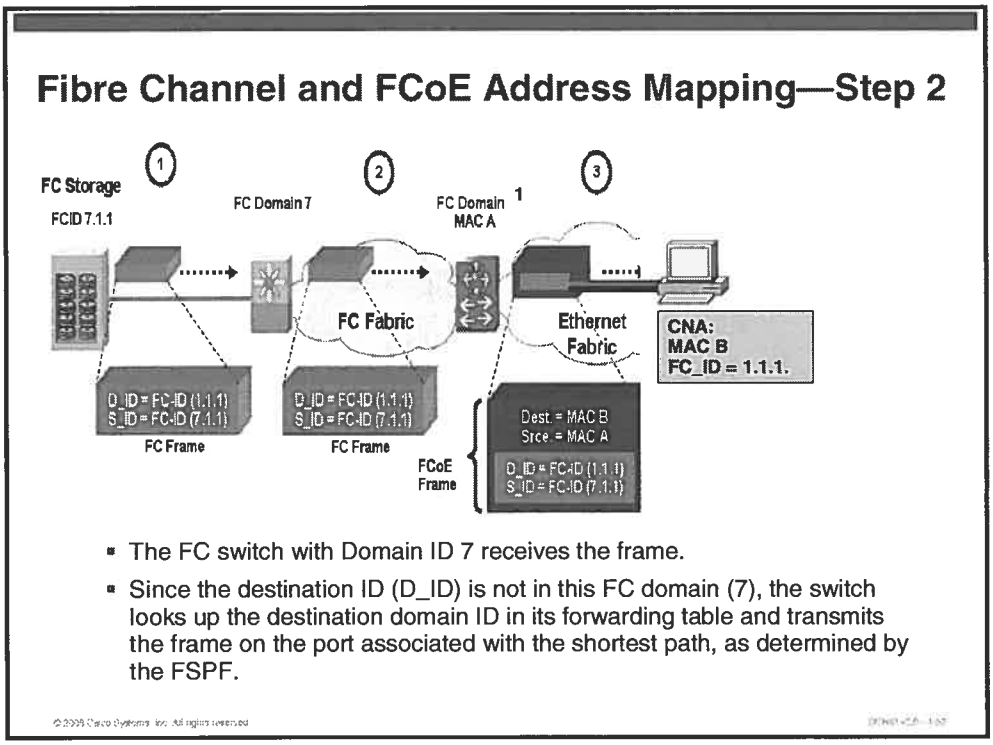
Fibre Channel and FCoE Address Mapping—Step 1



- The FC N_port on the storage array sends out the FC frame, which includes the Destination FC_ID (D_ID = 1.1.1) and the Source FC_ID (S_ID = 7.1.1) in the header.
- For simplicity, only the header information is displayed in the figure's frame graphic.

https://t.me/learningnets

Fibre Channel and FCoE Address Mapping—Step 2



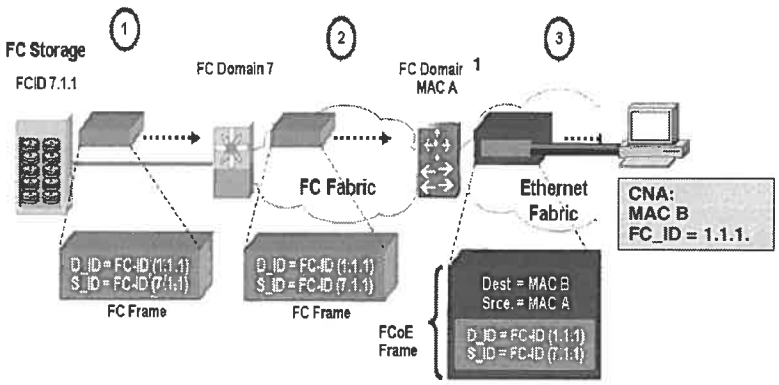
FSPF stands for Fabric Shortest Path First algorithm.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fibre Channel and FCoE Address Mapping—Step 3

Fibre Channel and FCoE Address Mapping—Step 3

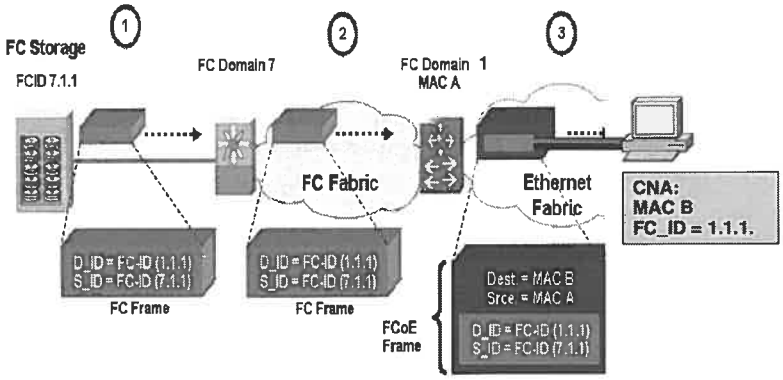


- The FCoE frame arrives at the Domain 1 switch and the switch determines that the FC frame destination is within its domain (Domain 1).
- Then the frame is transmitted locally across the FCoE mapper and Ethernet-switched to the appropriate exit Ethernet port, to the Host with MAC address B.

https://t.me/learningnets

Fibre Channel and FCoE Address Mapping—Last Step

Fibre Channel and FCoE Address Mapping—Last Step

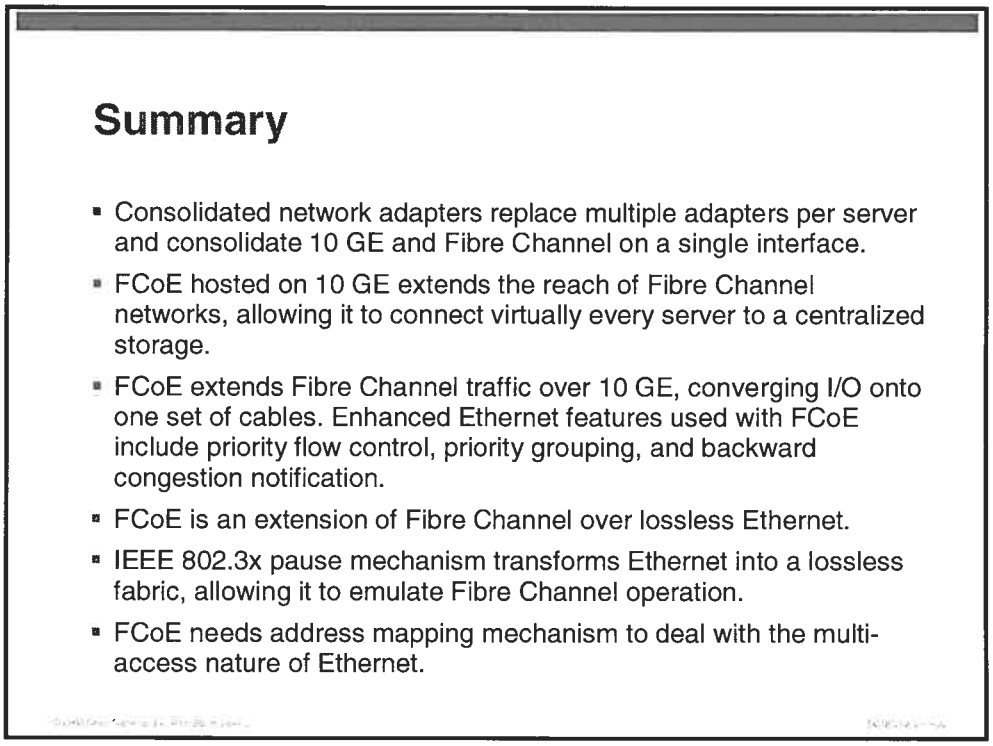


- When the frame is received by the FCoE host bus adapter with MAC address B, the FCoE frame is de-encapsulated and the FC frame accepted by the Engineering host with FC_ID 1.1.1.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary



In Fibre Channel Fabric architecture there is one Universal MAC address per Fibre Channel module. There is a dedicated MAC address for each Fibre Channel-ID.

FCoE needs a discovery protocol to keep the FLOGI processing unchanged and deal with the multi-access nature of Ethernet.

https://t.me/learningnets

Appendix

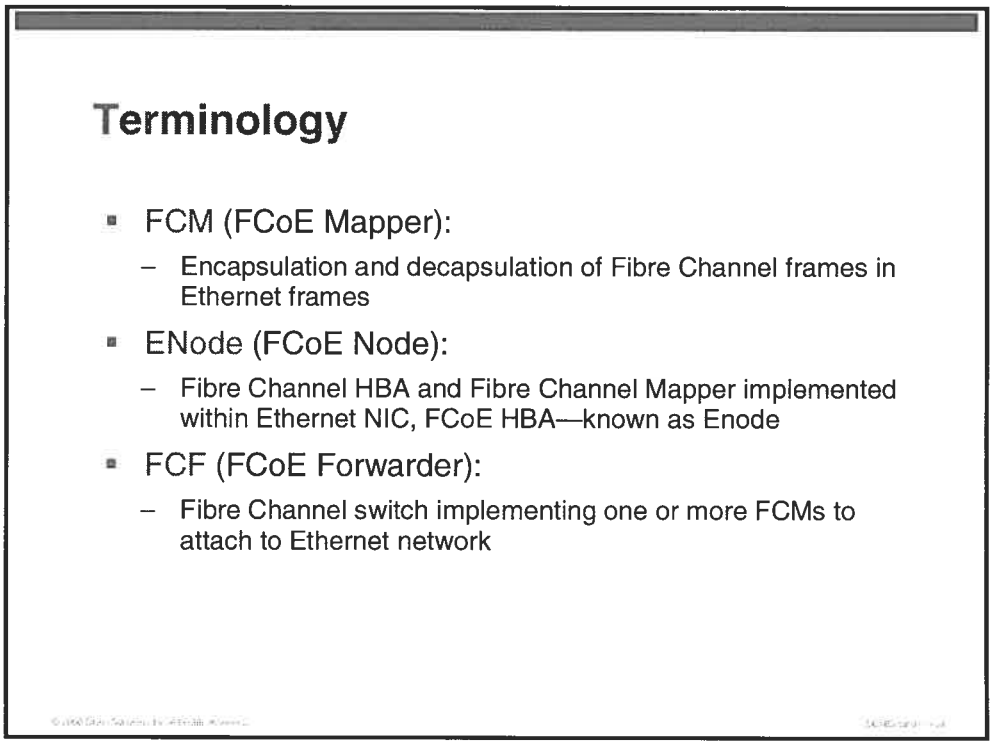
The appendix topics for Module 3 Lesson 1 include:

- FCoE Architecture
- MAC Addressing
- FCoE Discovery

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

FCoE Architecture



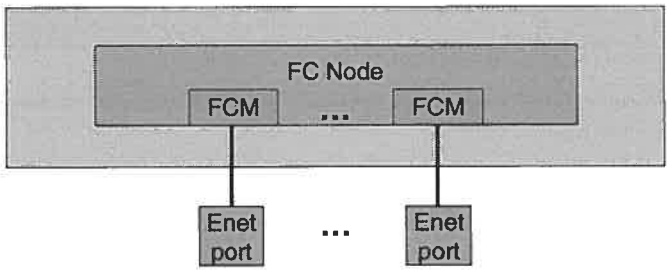
FCoE can be broken down into three components:

- Encapsulation of a Native Fibre Channel Frame into an Ethernet frame
- Extension of Ethernet to become a lossless fabric
- Replacing of a Fibre Channel link with MAC addresses in a lossless Ethernet fabric

https://t.me/learningnets

ENode Architecture

One MAC address per FCM



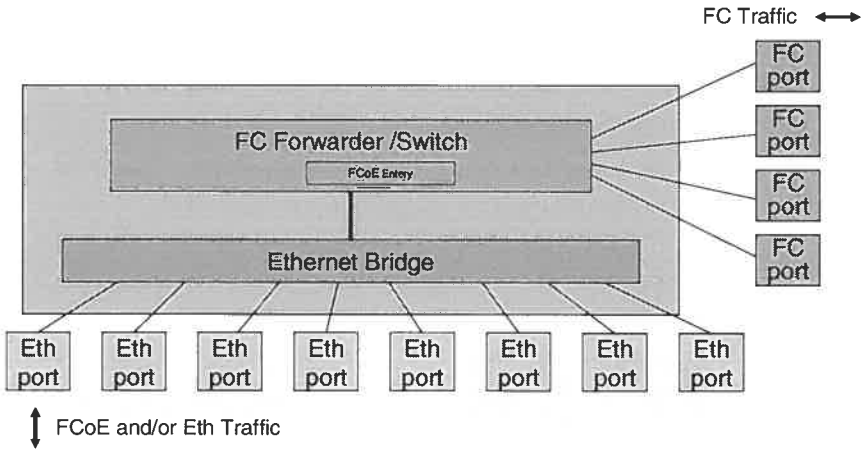
As with native Ethernet packets, FCoE uses Media Access Control (MAC) addressing to transfer packets between individual network hops. Fibre Channel addressing requires endpoint to endpoint knowledge. FCoE implementation contains mechanisms for mapping and resolving Fibre Channel endpoint addresses to Ethernet MAC addresses.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

FCoE Switch: Functional Diagram

One Universal MAC address per Fibre Channel module (FCM)



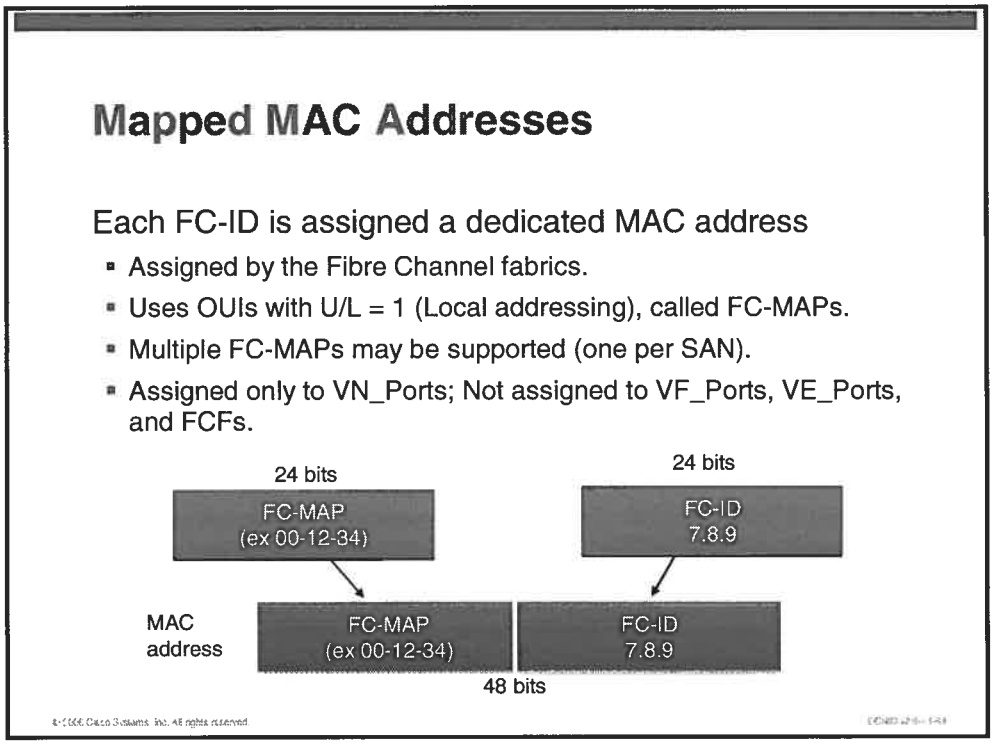
Inside the device in this figure are both an Ethernet switch function and a Fibre Channel switch function. Note the FCoE entity within the Fibre Channel switch. This entity encapsulates Fibre Channel frames into FCoE frames and de-encapsulates FCoE frames back into Fibre Channel frames. Fibre Channel traffic flows into one end of the FCoE entity and FCoE traffic flows out the other end.

FCoE can be broken down into three components:

- Encapsulation of a Native Fibre Channel Frame into an Ethernet frame
- Extension of Ethernet to become a lossless fabric
- Replacing of a Fibre Channel link with MAC addresses in a lossless Ethernet fabric

https://t.me/learningnets

MAC Addressing



As with native Ethernet packets, FCoE uses MAC addressing to transfer packets between individual network hops. Fibre Channel addressing requires endpoint to endpoint knowledge.

FCoE implementation contains mechanisms for mapping and resolving Fibre Channel endpoint addresses to Ethernet MAC addresses.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

FCoE Mapped MAC Addresses

MAC addresses of the form "FC-MAP || FC-ID"

- The FC-MAP value (the message's most significant 24 bits of the MAC address) is provided by the fabric in the FCoE Advertisement.
- The FC-ID value (the least significant 24 bits of the MAC address) is provided by the fabric in the FLOGI LS_ACC.

Not used by the FCoE Discovery protocol and by FLOGI/FDISC Requests and LS_ACCs

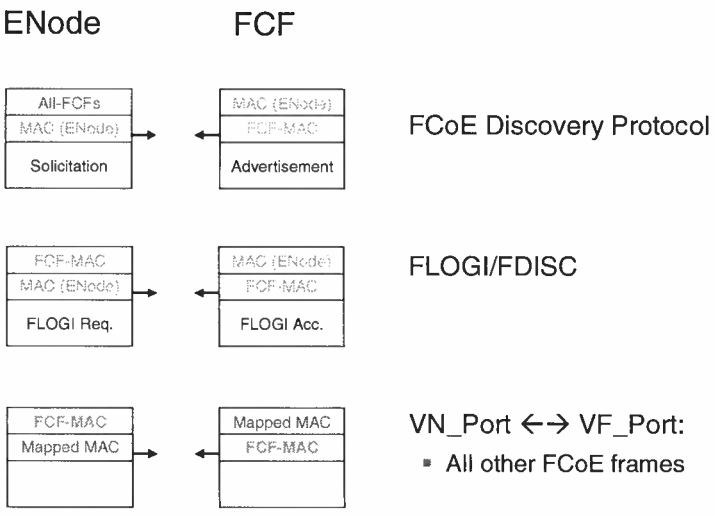
- The FCoE Discovery protocol and FLOGI/FDISC use the burned-in MAC addresses of the involved lossless Ethernet MACs.

Assigned only to VN_Ports

- Not assigned to VF_Ports, VE_Ports, and FCFs.

This definition supersedes any other previous one.

FCoE Mapped MAC Addresses: Usage



https://t.me/learningnets

Address Usage

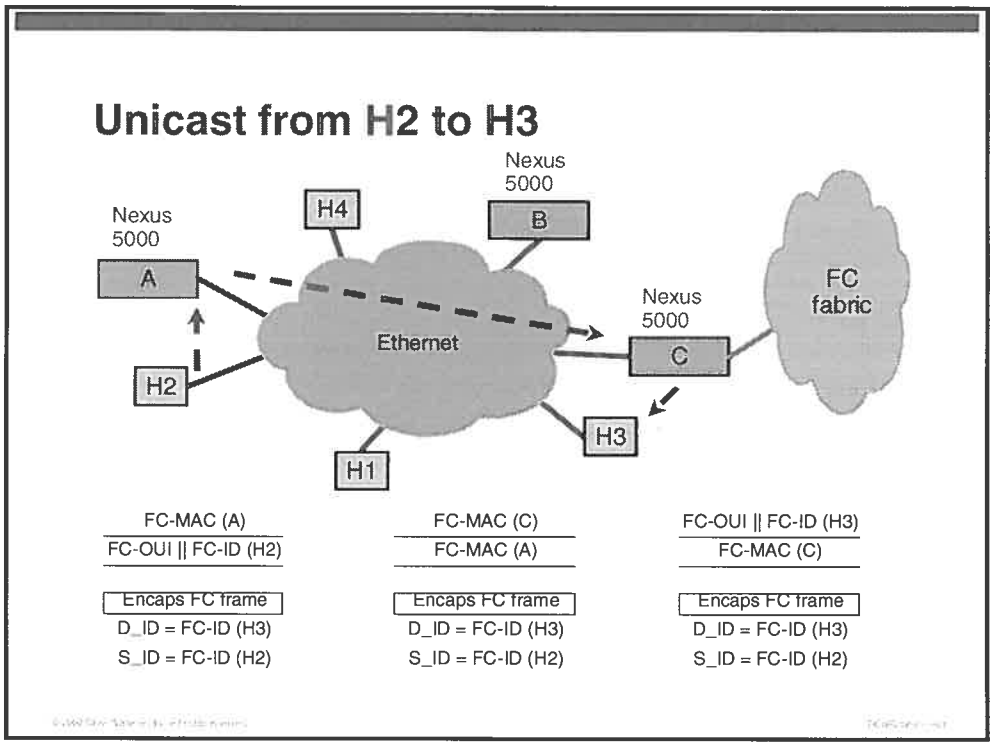
VE_Port to VE_Port	DA = Not mapped SA = Not mapped
VF_Port to VN_Port	DA = Mapped SA = Not mapped
VN_Port to VF_Port	DA = Not mapped SA = Mapped
VN_Port to VN_Port	DA = Mapped SA = Mapped

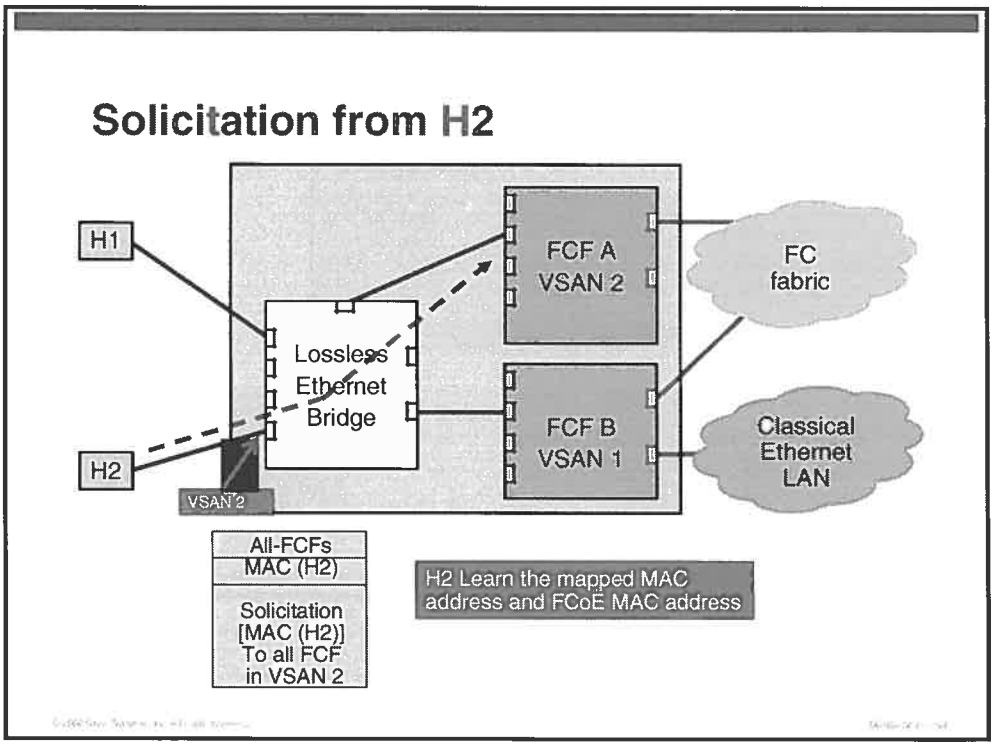
MAC address pair also encodes type of communication

© 2008 Cisco Systems, Inc. All rights reserved.

https://t.me/learningnets

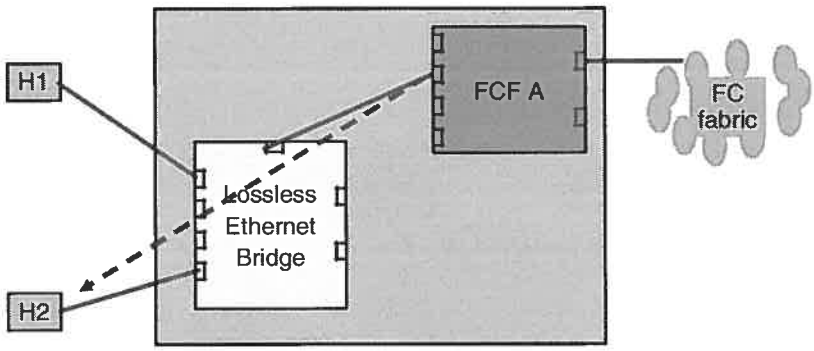
FCoE Discovery





https://t.me/learningnets

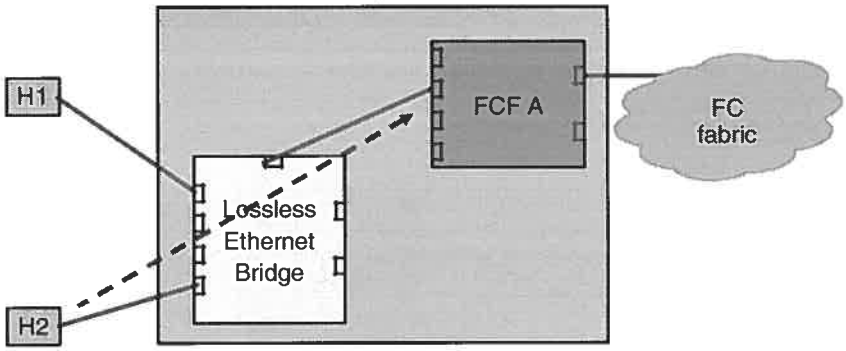
Advertisements from A



MAC (H2)
FCF-MAC (A)
Advertisement
[FCF-MAC (A), FC-OUI value]

H2 Learn the mapped MAC address and FCoE MAC address

FLOGI/FDISC Requests



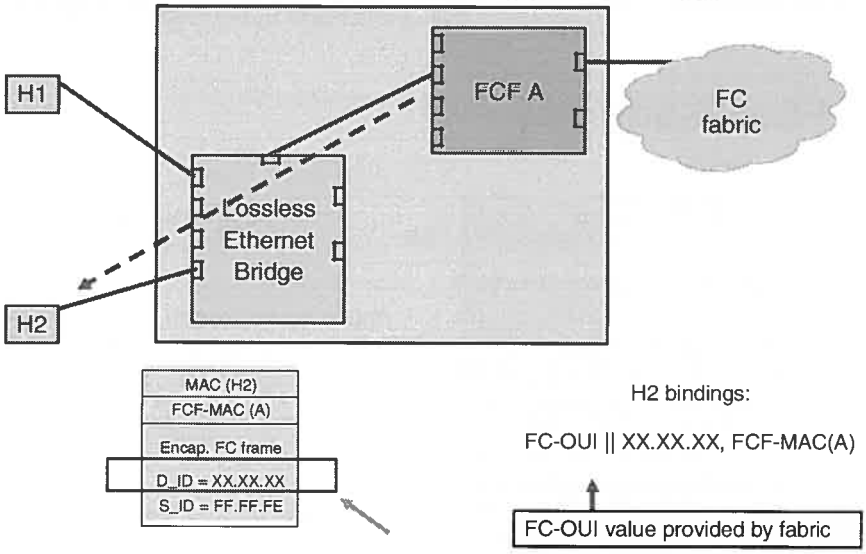
FCF-MAC (A)
MAC (H2)
Encap. FC frame
D_ID = FF.FF.FE
S_ID = 00.00.00

Source address (SA) of all FLOGI/FDISC requests is universal MAC address of station

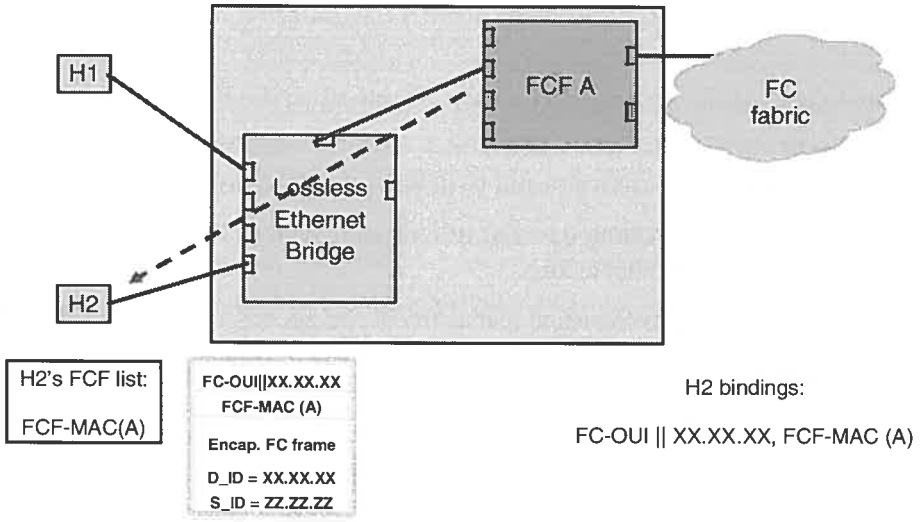
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

FLOGI/FDISC Accept in Reply to H2

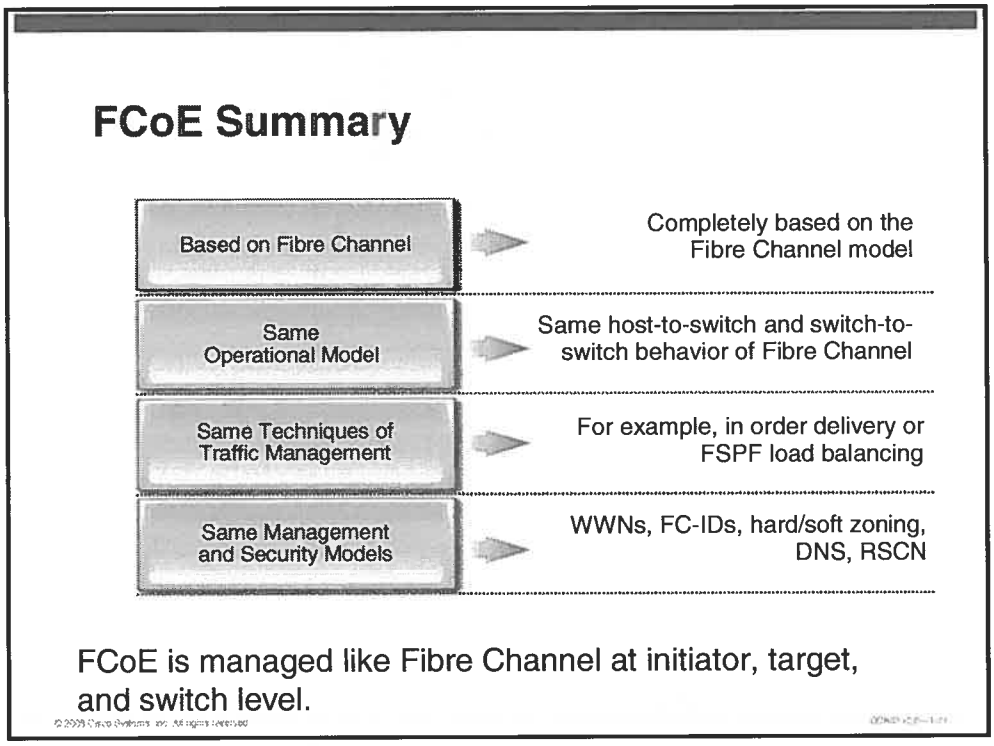


Subsequent FCoE Frame



https://t.me/learningnets

Appendix Summary



There are several advantages to FCoE, foremost of which is that it is managed as Fibre Channel at the host and switch level.

FCoE:

- Easy to understand because it is completely based on the Fibre Channel model.
- Because FCoE uses the same operational model as Fibre Channel, administrators can expect the same host-to-switch and switch-to-switch behavior of Fibre Channel.
- FCoE uses the same techniques of traffic management as Fibre Channel, such as in order delivery or FSPF load balancing.
- FCoE also implements the same management and security models as Fibre Channel:
  - WWNs
  - Fibre Channel-IDs
  - Hard/soft zoning
  - DNS
  - RSCN

Essentially, FCoE is aligned with the FC-BB-4 model.

Note The alignment described here guarantees transparent support for all Fibre Channel advanced features (VSANs, IVR/IFR, and security).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 2

Cisco Nexus 5000 Series Switches Overview

Overview

This lesson describes the Cisco Nexus 5000 Series Switches. The topics discussed will be:

- Cisco Nexus 5000 Series switch product portfolio
- Cisco Nexus 5000 Series switch features and functions
- Where the Cisco Nexus 5000 Series switch fits in the data center

Objectives

Upon completing this lesson, you will be able to meet these objectives:

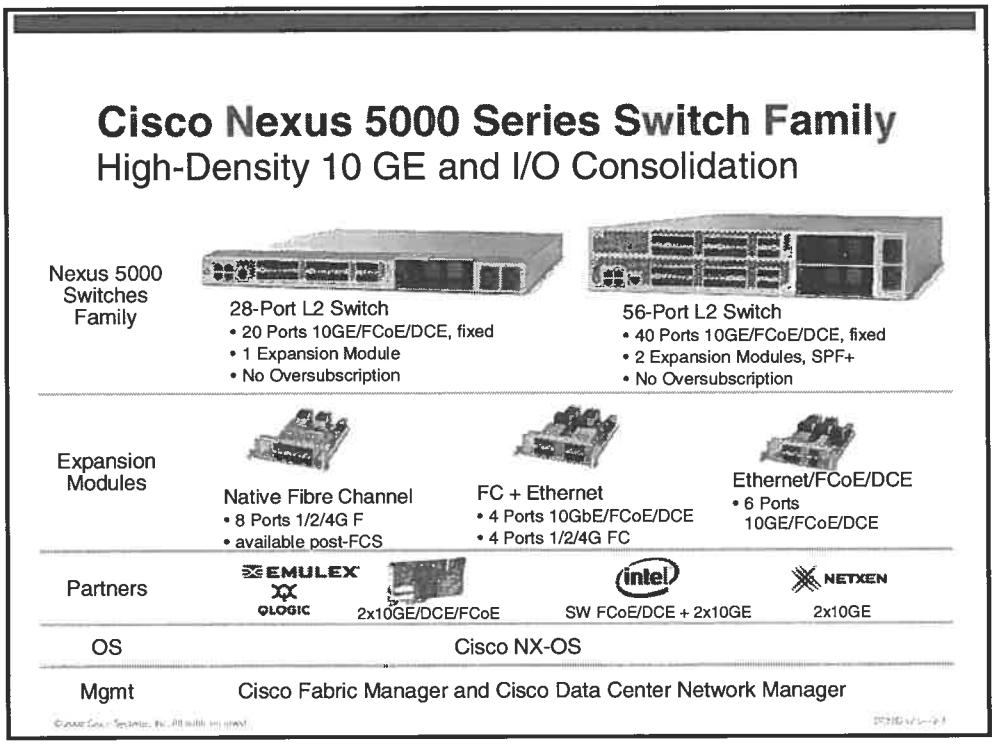
- Describe the Cisco Nexus 5000 Series switch products
- Explain where the Cisco Nexus 5000 Series switch fits best in the data center
- Describe how the Cisco Nexus 5000 Series switch is used for I/O consolidation
- Explain the features and functions provided by the Cisco Nexus 5000 Series switch
- Identify the management tools available for managing the Cisco Nexus 5000 Series switch

https://t.me/learningnets

Cisco Nexus 5000 Series Switch Product Family

This topic describes the Cisco Nexus 5000 Series switch product family and related features.

Cisco Nexus 5000 Series Switch Family



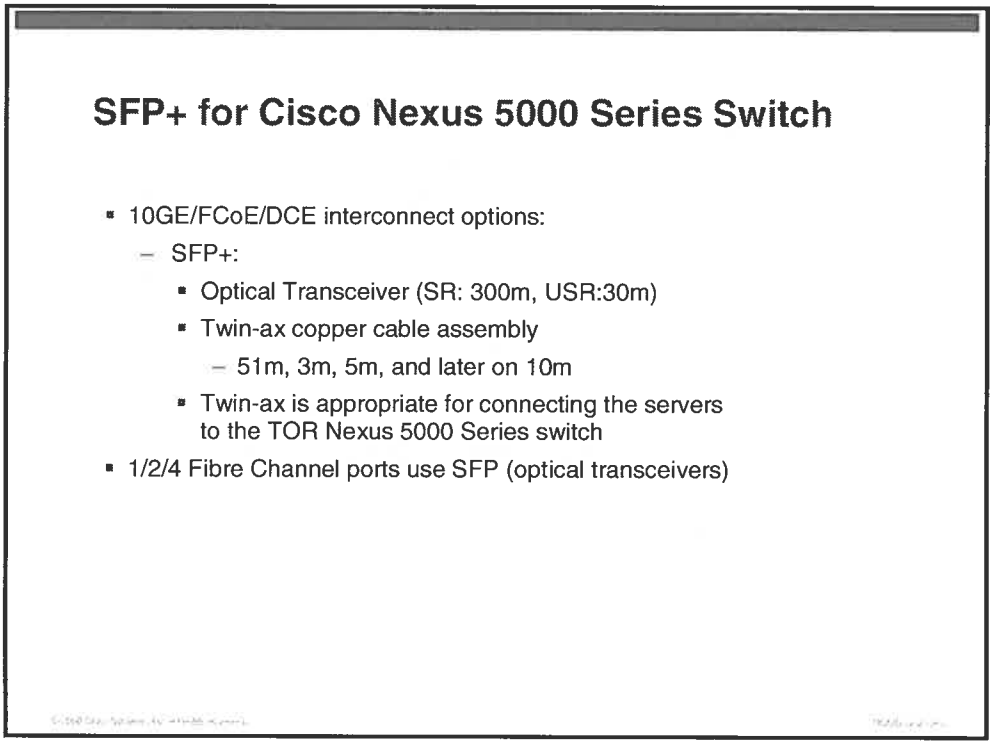
Fibre Channel over Ethernet (FCoE) Adapters via Partners: Emulex, Qlogic, Intel. All the Ethernet ports on the switch support 10 GE, FCoE and DCE. The first 16 fixed ports on 56-port switch can also be 1 GE port if needed.

Ports in Cisco Nexus 5020 Series Switches (1-16 fixed ports) will have future suport for 1 GE.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

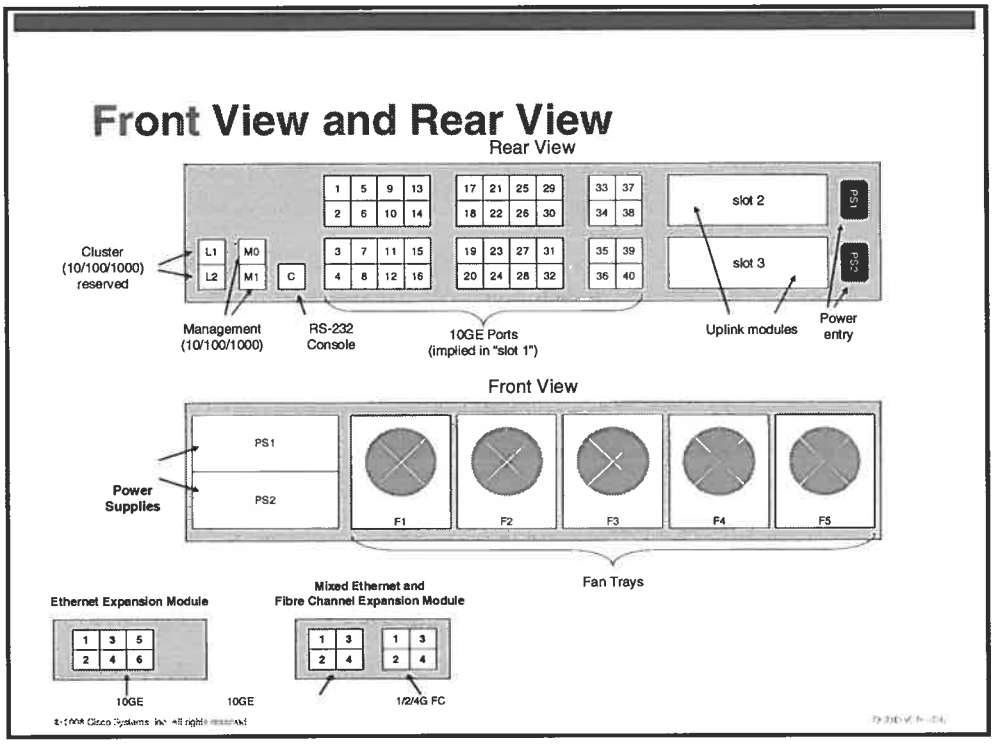
SFP+ for Cisco Nexus 5000 Series Switch



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Front View and Rear View



The Cisco Nexus 5020 is a two rack-unit (2RU), 10 Gigabit Ethernet, Cisco Data Center Ethernet, and FCoE 1/2/4 Gbps Fibre Channel switch built to provide 1.04 terabits per second (Tbps) throughput with very low latency. It has 40 fixed 10 Gigabit Ethernet, Cisco Data Center Ethernet, and FCoE Small Form Factor Pluggable Plus (SFP+) ports.

Two expansion module slots can be configured to support up to 12 additional 10 Gigabit Ethernet, Cisco Data Center Ethernet, and FCoE SFP+ ports, up to 16 Fibre Channel switch ports, or a combination of both. The switch has a serial console port and an out-of-band 10/100/1000 Mbps Ethernet management port. The switch is powered by 1+1 redundant, hot-pluggable power supplies, and 4+1 redundant, hot-pluggable fan modules to provide highly reliable front-to-back cooling.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 5000 Series Switch: Data Center Class

Cisco Nexus 5000 Series Switch: Data Center Class

- Front to back cooling
- Simplified serviceability
  - Front servicing
  - Reduce impact to other devices
- High availability
  - Hot plug power supplies and fans
  - Fans N+1 redundant
  - PSUs N+N grid redundant
  - NX-OS modular operating system
- Simplified cabling
  - Ports in back
- Power: 180-264 VAC
  SFP+CU - down, optical - up
  - 2RU: 750W worst case, ~ 450W nominal
- Low latency
  - 3.2µs (port to port) with services turned on
  - Consistent for all packet sizes
  - Cut-through architecture

Max Port Densities
Up to 52X10GE Line Rate
Up to 16X1/2/4G FC + 40X10GE



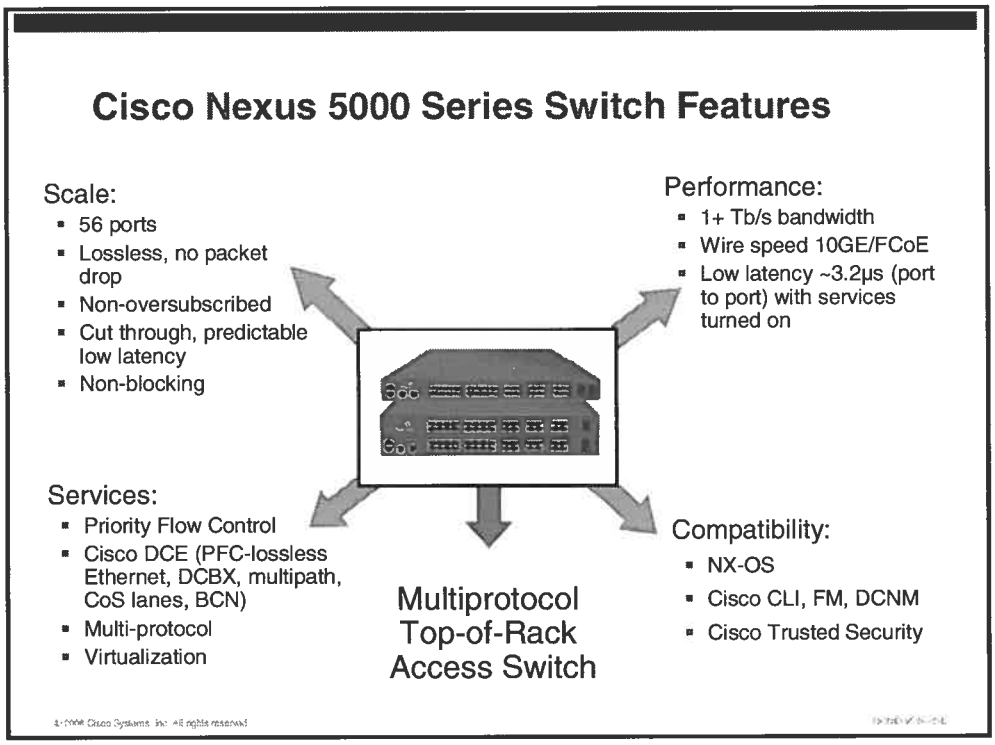
* FC ports are SFP

Cisco NX-OS modular operating system has a SAN-OS origin.

VNtag Server Virtualization is scheduled in the future.

https://t.me/learningnets

Cisco Nexus 5000 Series Switch Features



- **Management**: Cisco Fabric Manager and Cisco Data Center Network Manager (DCNM)
- **OS**: Cisco NX-OS
  - NX-OS—Data Center Operating System class:
    - Superior operational efficiency
    - Pervasive security (CTS: Cisco Trusted Security)
    - Continuous system operation
    - Future-proof architecture
    - Purpose-built for data center
- **Performance**:
  - Wire-speed cut-through switch
  - Up to 56 non-blocking 10 GE ports
  - Consistent latency across all packet sizes
  - Low latency ~3.2µs (port to port) with services turned on

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus Operating System (Cisco NX-OS)

Nexus Operating System (Cisco NX-OS)

Superior Operational Efficiency
- Graceful system operations
- Programmable APIs (XML)
- Comprehensive management and instrumentation

Pervasive Security
- Roles-based administration

Highly Available:
Continuous System Operation
- Seamless integration of features and fixes
- Infrastructure for ISSU
  - Protected and restartable processes
  - Process survivability
  - In service software upgrades
- Modular design



Purpose Built for Data Center

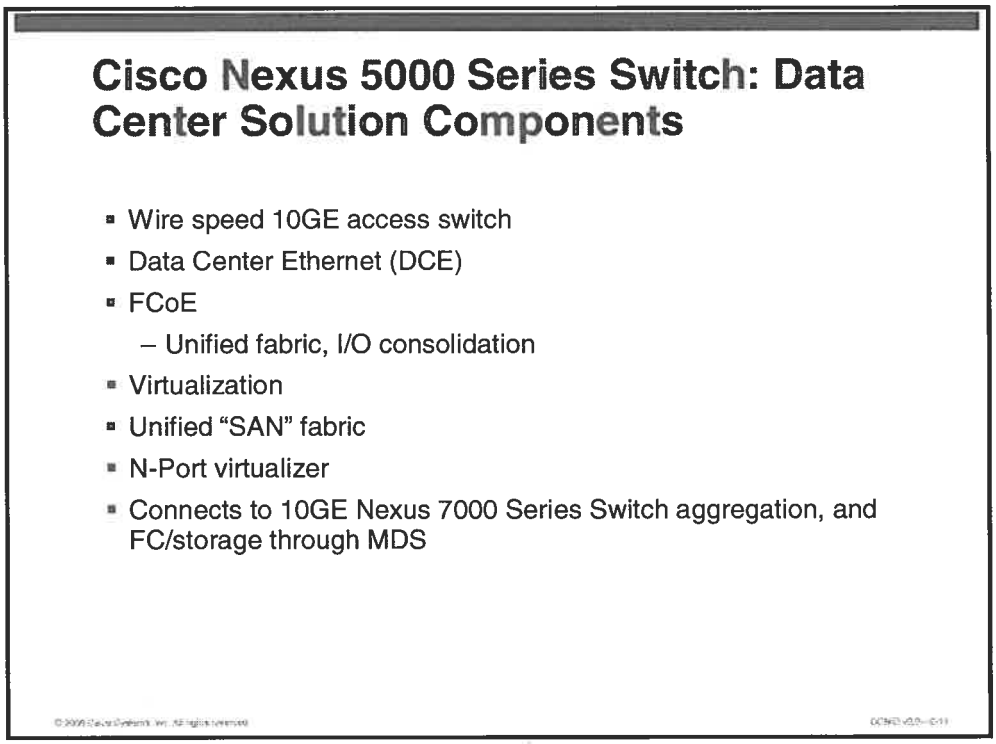
Infrastructure for ISSU (features in future release)

- Protected and restartable processes
- Process survivability
- In Service Software Upgrades

https://t.me/learningnets

Data Center Solution Components

This topic describes the data center solution components.



The entire product will go through standard Cisco channels (including 10 GE, Fibre Channel, and FCoE license upgrades). Cisco is targeting OSMs and OEMs that traditionally handle storage. Cisco is seeking qualification from partners and expects that they will also sell the product. OSM can sell the system only with FCoE turned on, so it will not be competitive in an Ethernet-only deal.

Competition: HP Virtual Connect, merchant silicon switch vendors, and Brocade

Solves major problems identified in data centers: low costs, power, cooling

Ethernet Host Virtualizer (future): No spanning tree to manage for uplinks, double the uplink bandwidth

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

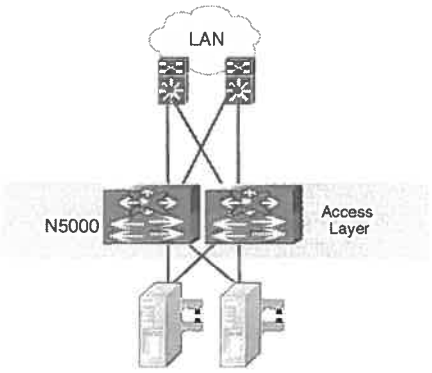
Access Switch

Cisco Nexus 5000 Series Switch Data Center Solution Components: Access Switch

- Where it fits

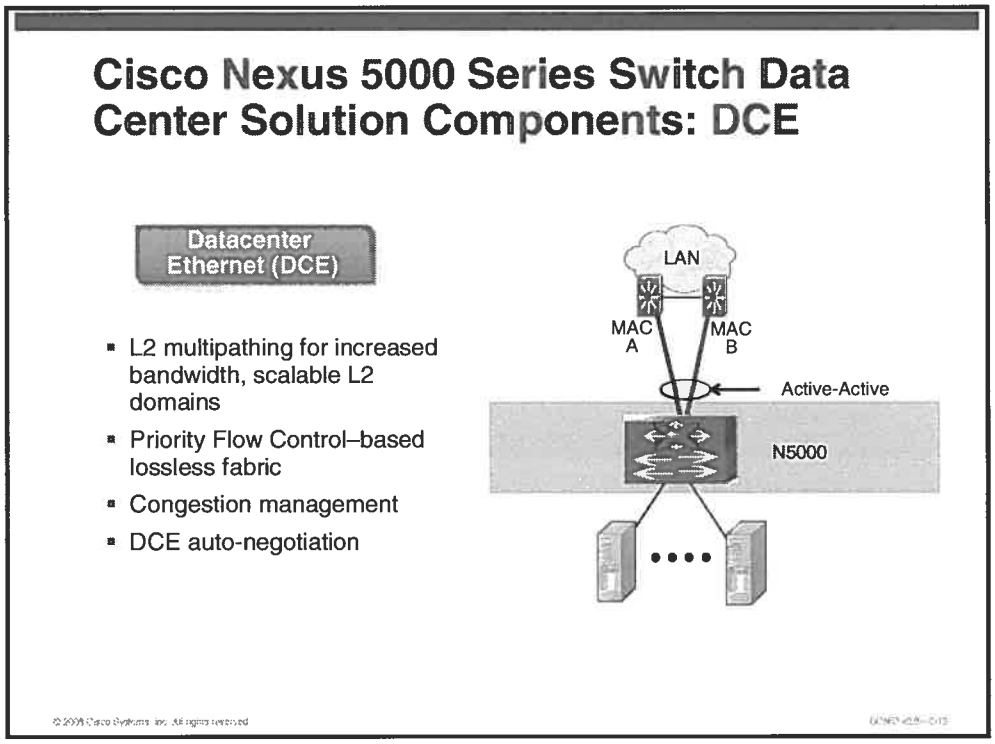
Wire-speed 10GE Access Switch

- 10GE L2 non-blocking, wire-speed switch
- Low latency < 3.2 ms
- Lossless networking
- Unified fabric



https://t.me/learningnets

Data Center Ethernet (DCE)



With DCE and FCoE you can now fully consolidate Storage Area Networking (SAN), LAN and cluster, and high-performance cluster traffic onto one infrastructure.

DCE allows starting consolidating certain traffic types besides traditional LAN traffic such as low-latency traffic types of high-performance computing applications, and cluster applications. It is the infrastructure by which Fibre Channel can be put over the unified fabric.

Data Center Ethernet

- Enhanced Ethernet for data center applications
- FCoE
- Priority Flow Control (Priority Pause)
- CoS-based bandwidth management (IEEE 802.1Qaz)
- Congestion management
- Layer 2 multipathing

© 2008 Cisco Systems, Inc.

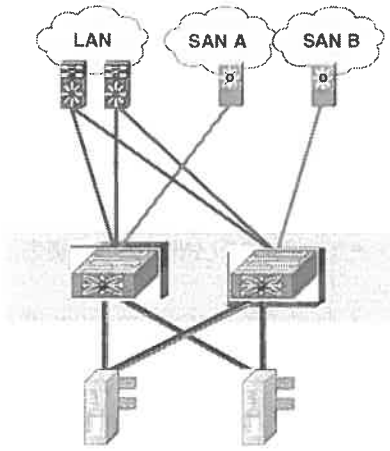
https://t.me/learningnets

Fibre Channel over Ethernet

Cisco Nexus 5000 Series Switch Data Center Solution Components: FCoE

FCoE

- Unified fabric for LAN, SAN, HPC/IPC
- Enables Fibre Channel connectivity across more servers
- Fewer switches, fewer points of management
- Significant cable and adapter reduction
- Lower cost, power, cooling



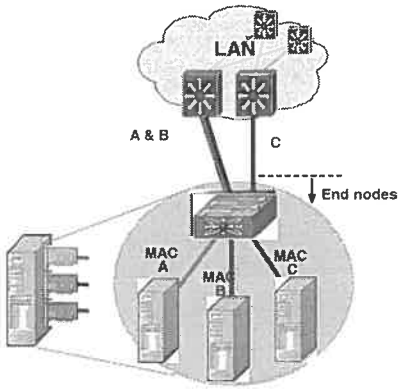
https://t.me/learningnets

Virtualization

Cisco Nexus 5000 Series Switch Data Center Solution Components: Virtualization

Virtualization Networking

- VM-optimized network services
- End-port virtualization
- Unified I/O with 10Gb and FCoE
- Delivers increased, security operational agility

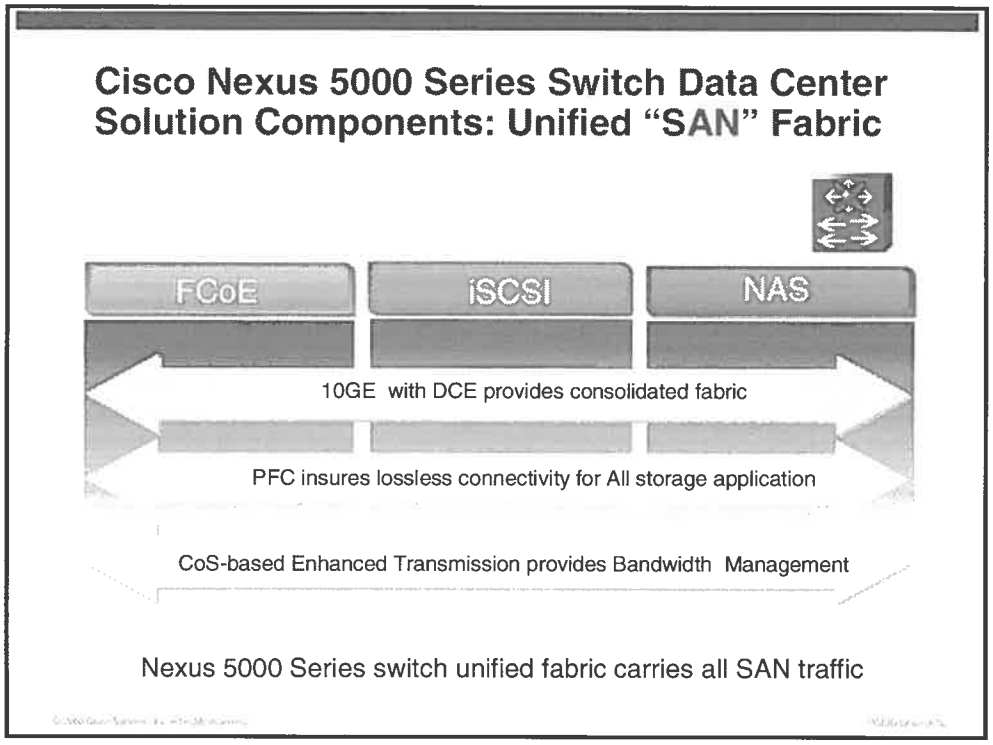


VM Optimized Services: Switch has the ability to recognize virtual machines that are running on the servers.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Unified SAN Fabric

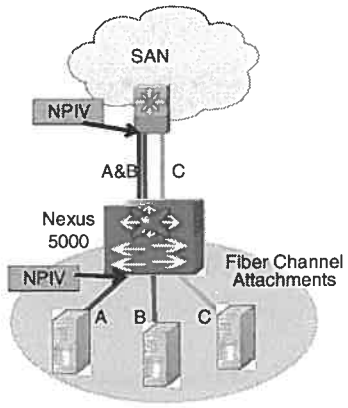


https://t.me/learningnets

Fibre Channel N-Port Virtualizer

Cisco Nexus 5000 Series Switch Data Center Solution Components: Fibre Channel N-Port Virtualizer

- No Fibre Channel domain required on switch
- Provides physical port level virtualization of multiple Fibre Channel end nodes to one F_Port off a Fibre Channel switch
  - Nexus 5000 Series switch operates in N_Port Proxy Mode (not in Fibre Channel switch mode)
  - Simplifies multi-vendor interoperation
- Simplifies management
- Used in conjunction with NPIV on Nexus 5000 Series switch and MDS



FC N_Port Virtualizer

Extends limits of Fibre Channel domain IDs. By virtue of technically not being in Fibre Channel Switch Mode, it does not burn Fibre Channel domain IDs and thus extends the domain IDs.

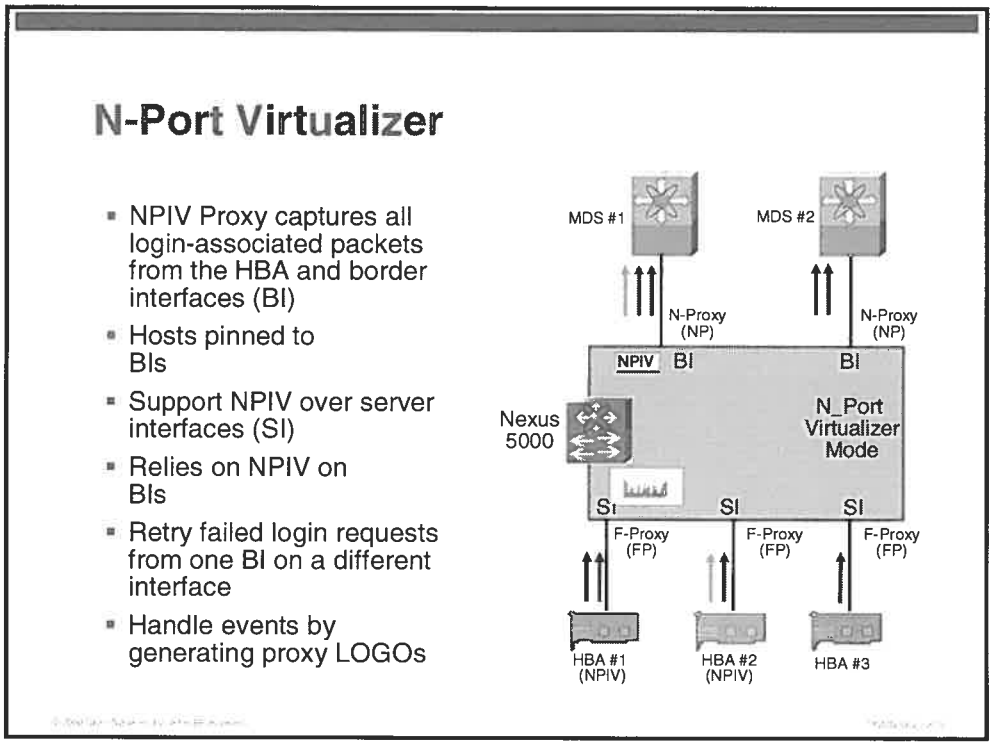
Note on NPIV:

- Multiple N_ports share a single F_port.
  - First N_port uses fabric login (FLOGI).
  - Subsequent N_ports use FDISC.
- Each N_port is known based on its World Wide Name (WWN).
  - Each N_port receives its own FCID.
  - Zoning is applied independently.
- All N_ports share the BB_credits negotiated at FLOGI.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

N-Port Virtualizer



The N Port Identifier Virtualization (NPIV)-proxy module in the TORS provides the proxy functionality of distributing FLOGI requests from the servers over the available external Fibre Channel interfaces. The Fibre Channel HBAs in the servers and the Fibre Channel switch upstream to the TORS think that they are directly connected to each other using a physical cable. Note that the NPIV-proxy functionality allows NPIV to be used between the TORS and the switch even if some or all HBAs implement just basic N-port functionality.

In this document, the interface that is connected to the HBA is referred to as Server Interface, and the interface that is connected to the Fibre Channel switch is referred to as External interface.

The main responsibilities of the NPIV-proxy Module F-Proxy Mode, or FP Mode, and the interface that is connected to the Fibre Channel switch that is said to be in N-Proxy Mode, or NP Mode, are to:

- Capture all the login-associated packets like FLOGI, LOGO, LS_ACC, and LS_RJT from the servers and external interfaces that are enabled for N_Port or NPIV functionality
- Perform intelligent load balancing of login sessions coming from servers over the set of external interfaces in such a way that the logins are uniformly distributed
- Support NPIV functionality over server interfaces, to support HBAs that are NPIV capable
- Support NPIV on external interfaces, to support Fibre Channel switches that support NPIV logins
- If a login request on one external interface fails, provide the retry functionality for sending the login request on a different external interface

https://t.me/learningnets

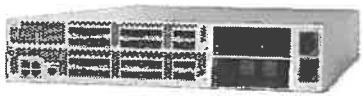
- Take care of events like external interfaces, the server interfaces and VSANs going up and down by generating Proxy LOGOs
- Provide high availability via process restart and switchover

© 2008 Cisco Systems, Inc.

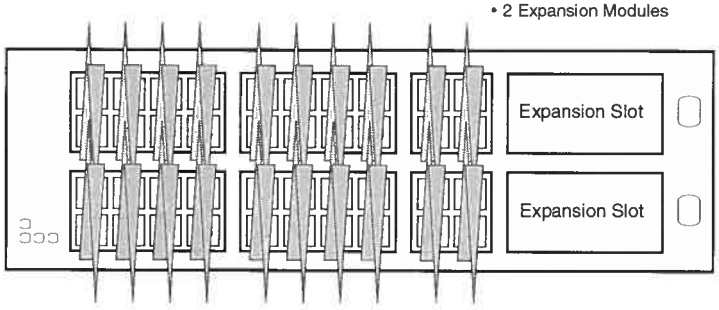
https://t.me/learningnets

Base Configuration

Cisco Nexus 5000 Series Switch—Base Configuration



56-Port L2 Switch
- 40 Ports 10GE/FCoE/DCE, fixed
- 2 Expansion Modules



40 x 10G Base Ethernet Ports + 2 Expansion Modules

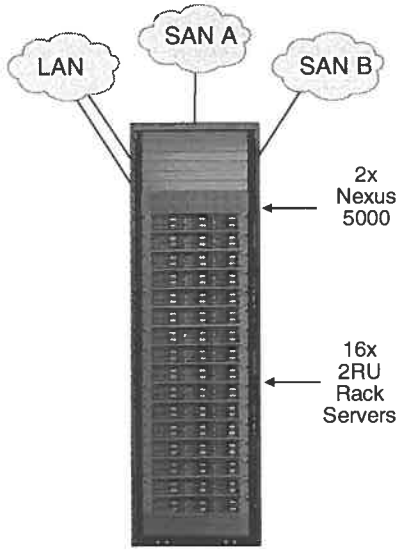
https://t.me/learningnets

Rack Server Deployment

Cisco Nexus 5000 Series Switch—Rack Server Deployment

Example:

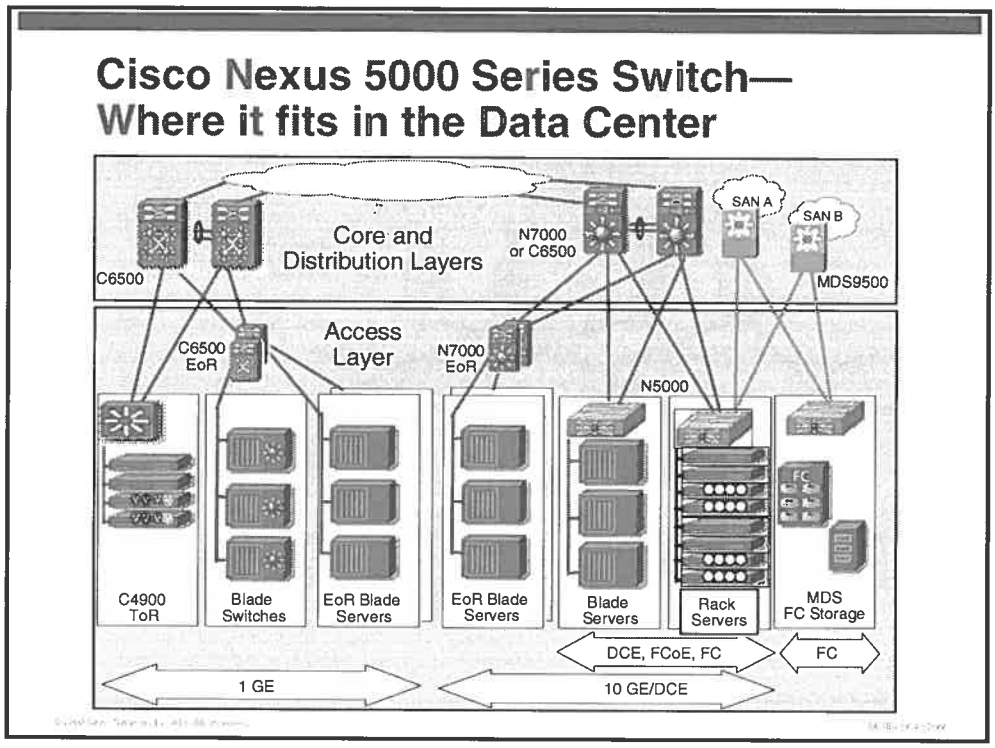
- Redundant 28-port server switch configuration
  - 2RU total rack space
  - 28 available ports
- 16 2RU servers per rack
- 20 10-GE ports used
  - 16 connections per switch
  - 4 10-GE uplinks
- 8 1/2/4G Fibre Channel uplinks



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Where Cisco Nexus 5000 Series Switches Fits



Regardless of its position—core, distribution, or access—the Cisco Data Center switching portfolio addresses all data center design options. Cisco is delivering end-to-end unified fabric architecture: Cisco Data Center product and solution portfolio, and Cisco Nexus 5000 Series switches products.

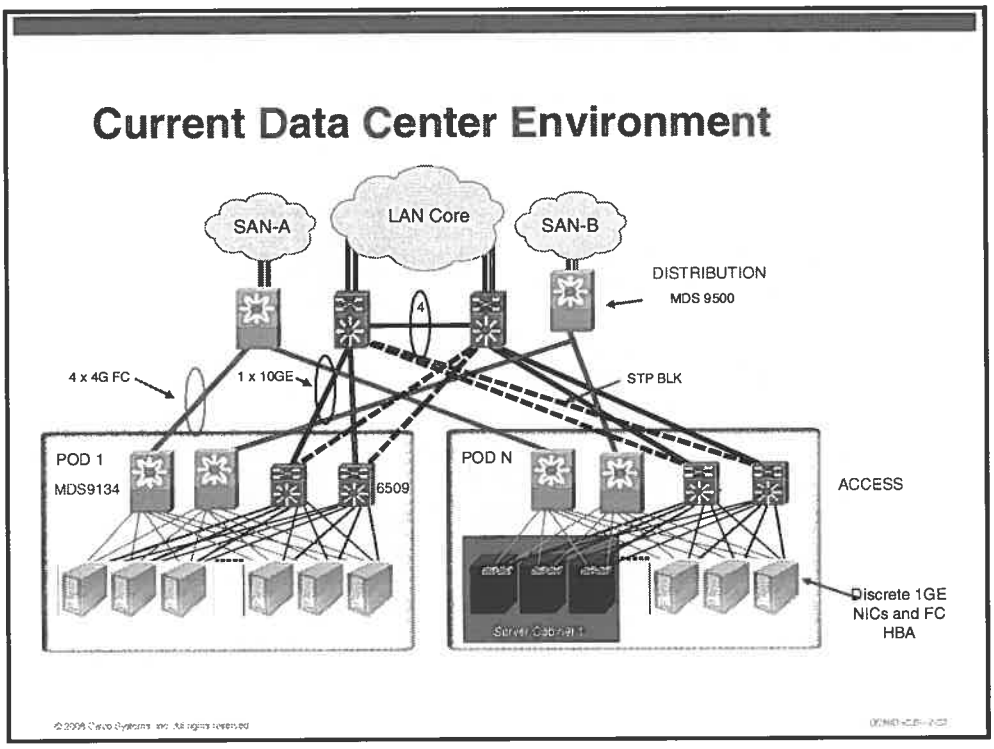
Benefits:

- Lower CapEx/OpEx
- Reduce carbon footprint
- Operational efficiency: wire once, walk away
- Application performance and resiliency

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

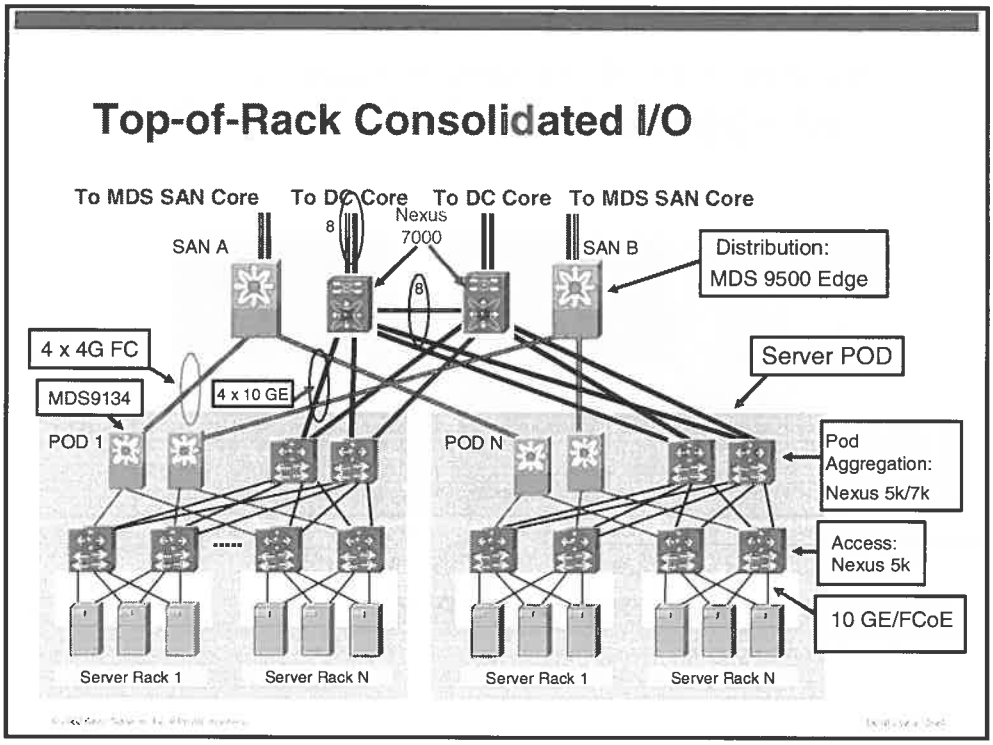
Current Data Center Environment



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Top-of-Rack Consolidated I/O

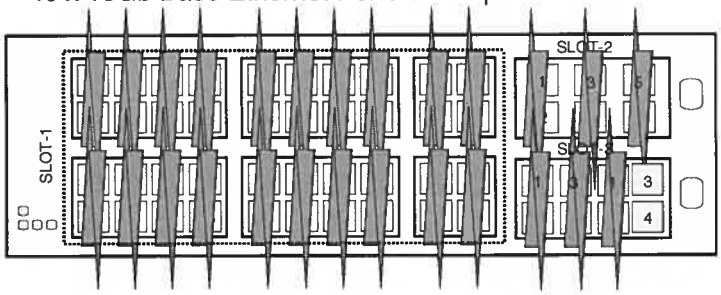


https://t.me/learningnets

10Gb Aggregation and I/O Consolidation

Cisco Nexus 5000 Series Switch—10Gb Aggregation and I/O Consolidation

40 x 10Gb Base Ethernet Ports + 2 Expansion Modules



50 x Line-rate 10Gb Line Ethernet Ports

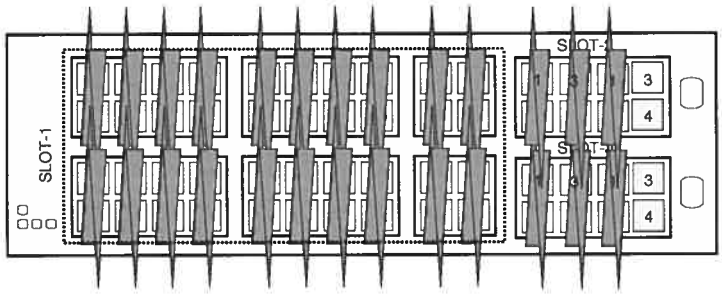
4 x 1/2/4Gb Fibre Channel Ports

© 2008 Cisco Systems, Inc. All rights reserved.

DCNID v2.0—7-25

Cisco Nexus 5000 Series Switch—10Gb Aggregation and I/O Consolidation (cont.)

40 10Gb Base Ethernet Ports + 2 Expansion Modules



48 Line-rate 10Gb Line Ethernet Ports

8 1/2/4Gb Fibre Channel Ports

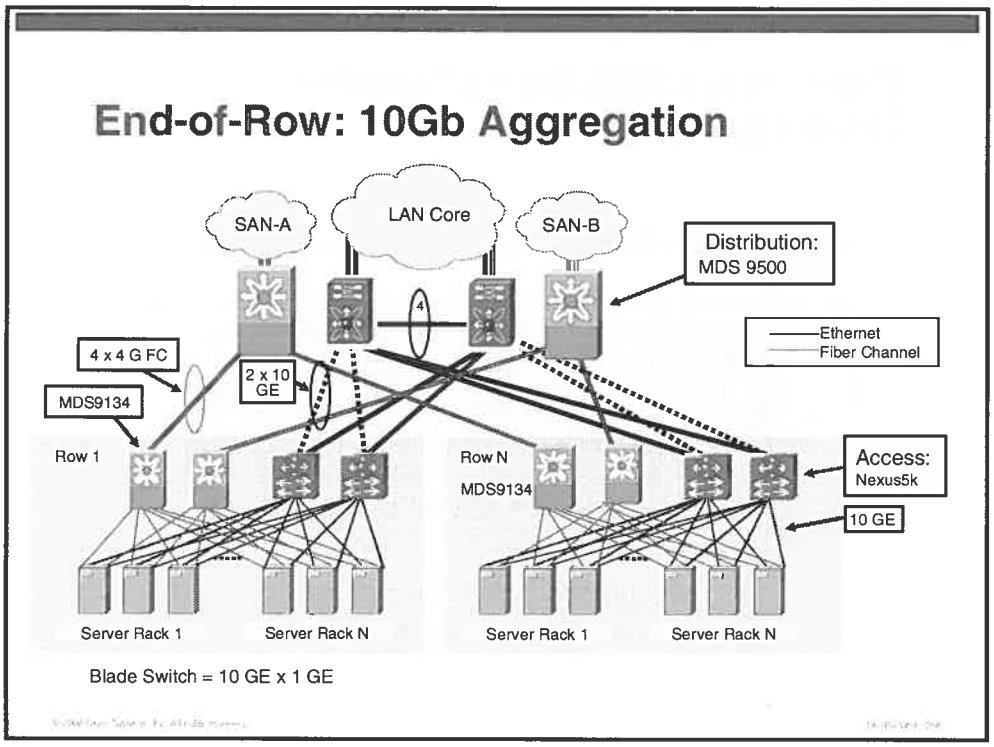
© 2008 Cisco Systems, Inc. All rights reserved.

DCNID v2.0—7-26

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

End-of-Row: 10Gb Aggregation



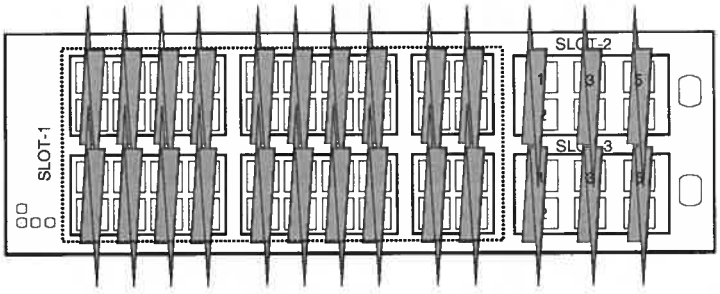
The Cisco Nexus 5000 Series switches enable end-of-row 10Gb aggregation topology.

https://t.me/learningnets

10Gb Aggregation

Cisco Nexus 5000 Series Switch—10Gb Aggregation

40 x 10Gb Base Ethernet Ports + 2 Expansion Modules



52 Line-rate 10Gb Ethernet Ports

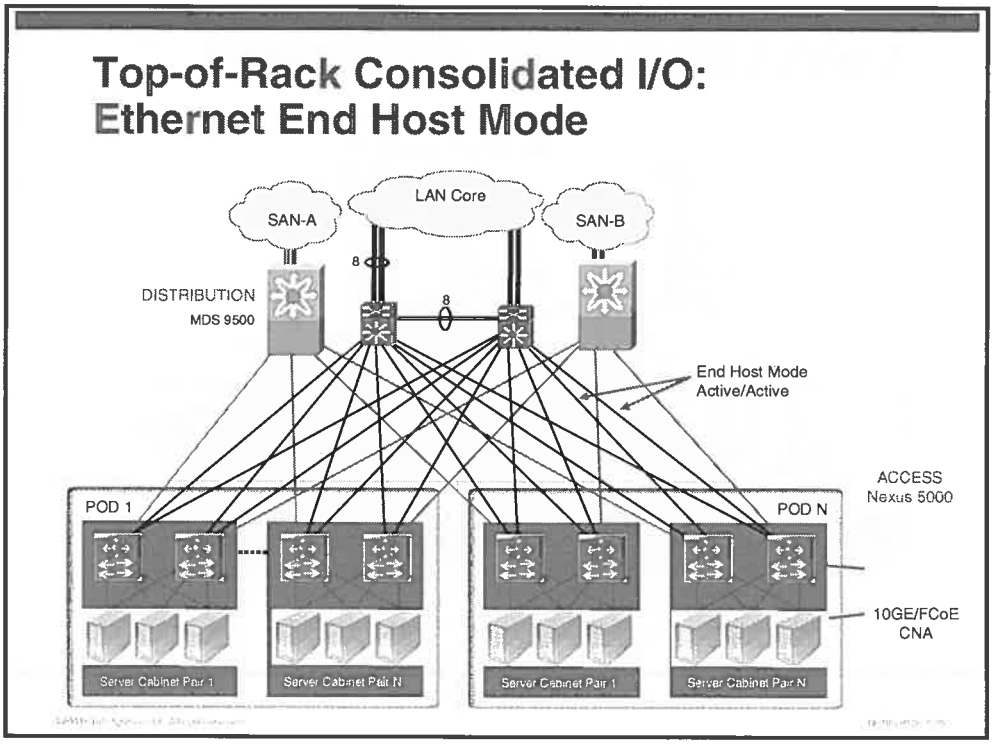
0 Fibre Channel Ports

© 2008 Cisco Systems, Inc. All rights reserved.

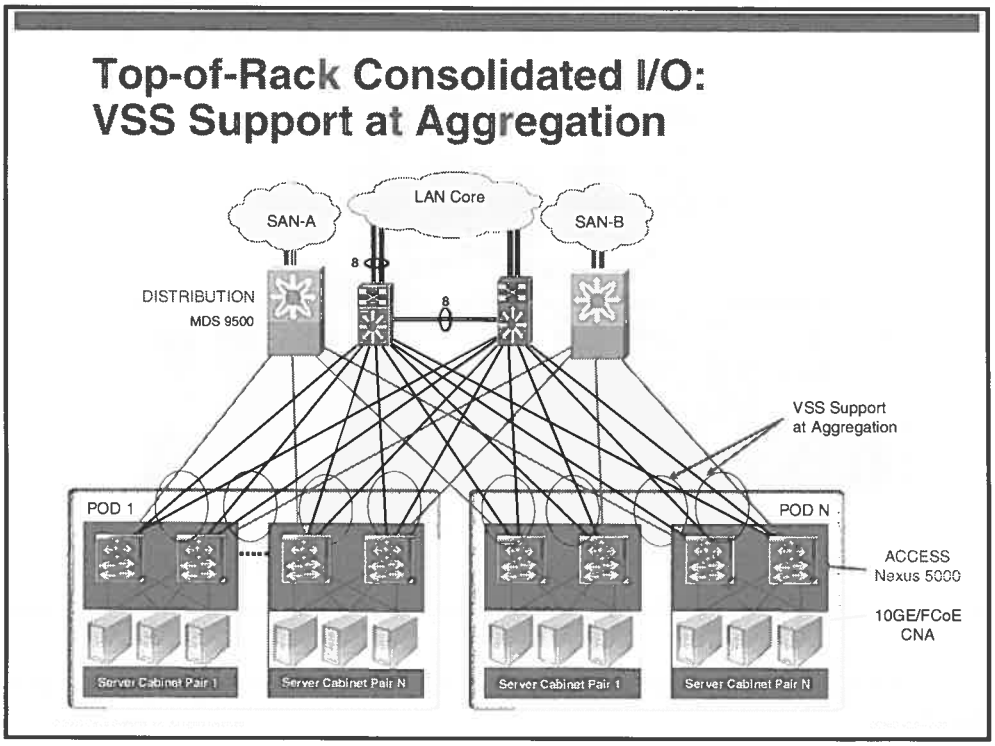
https://t.me/learningnets

Top-of-Rack Consolidated I/O

Ethernet End Host Mode



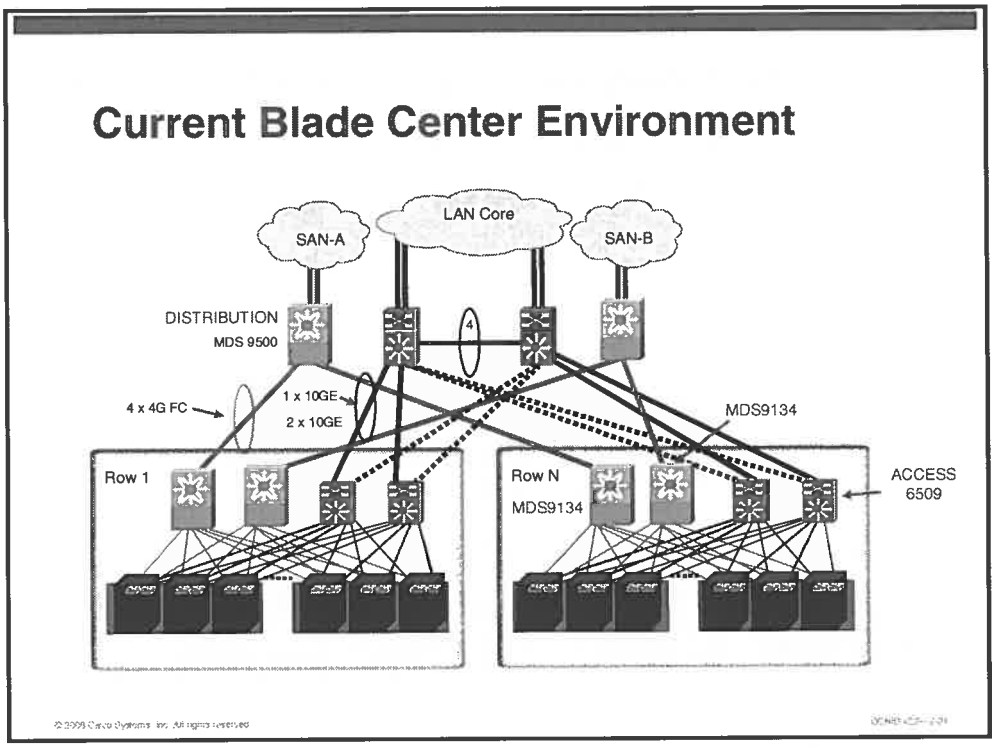
VSS Support at Aggregation

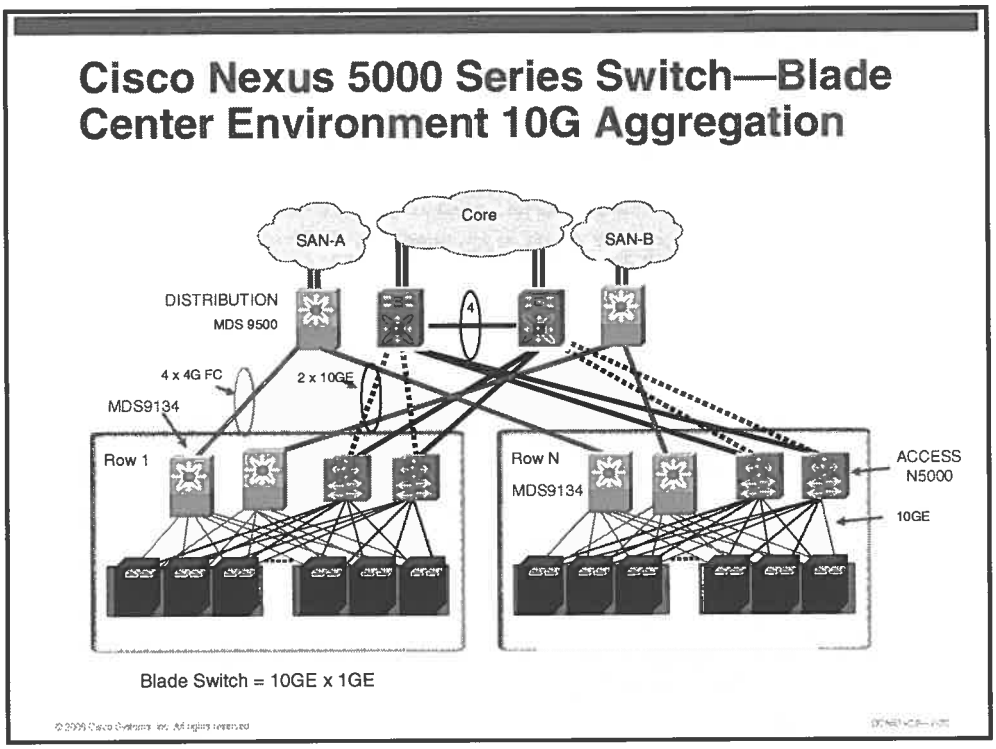


© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Blade Center Environment





16 Racks in a row; 3 uplinks per rack; 48 server-facing interfaces in Oregon; 4 network-facing interfaces

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 5000 Series Switches Management Tools

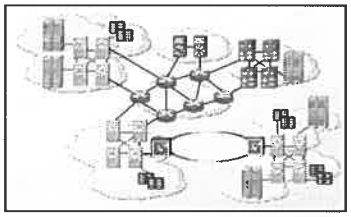
This topic describes managing Cisco Nexus 5000 Series Switches using Cisco Device Manager and Cisco Fabric Manager

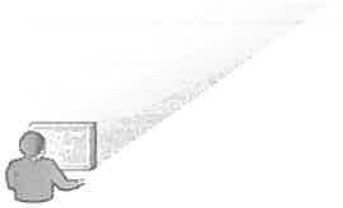
Comprehensive Management Tools

Comprehensive Management Tools

- SNMP-based Management
  Cisco Fabric Manager/
  Device manager
- XML-API based Management
  DCNM will be suported
- Comprehensive management for the Nexus 5000 Series switch, featuring:
  - CLI structured like Cisco IOS® CLI
  - Cisco Fabric Manager
- Role-based access control
- Standard, open interfaces:
  - SNMP, CIM XML (SMI-S)

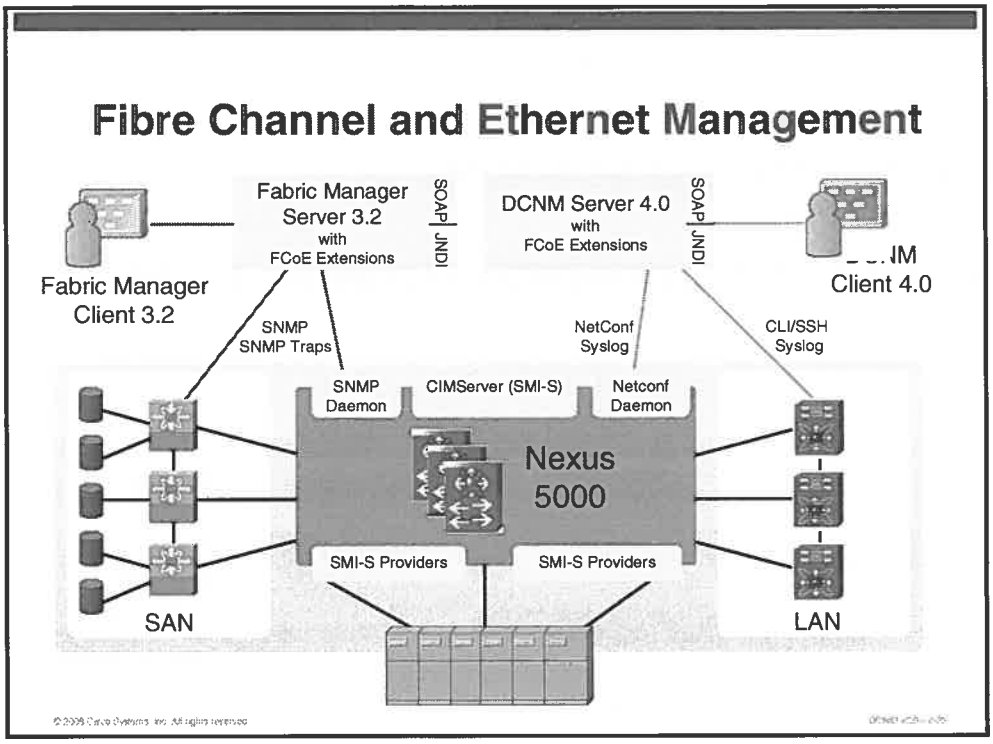
Comprehensive Management





https://t.me/learningnets

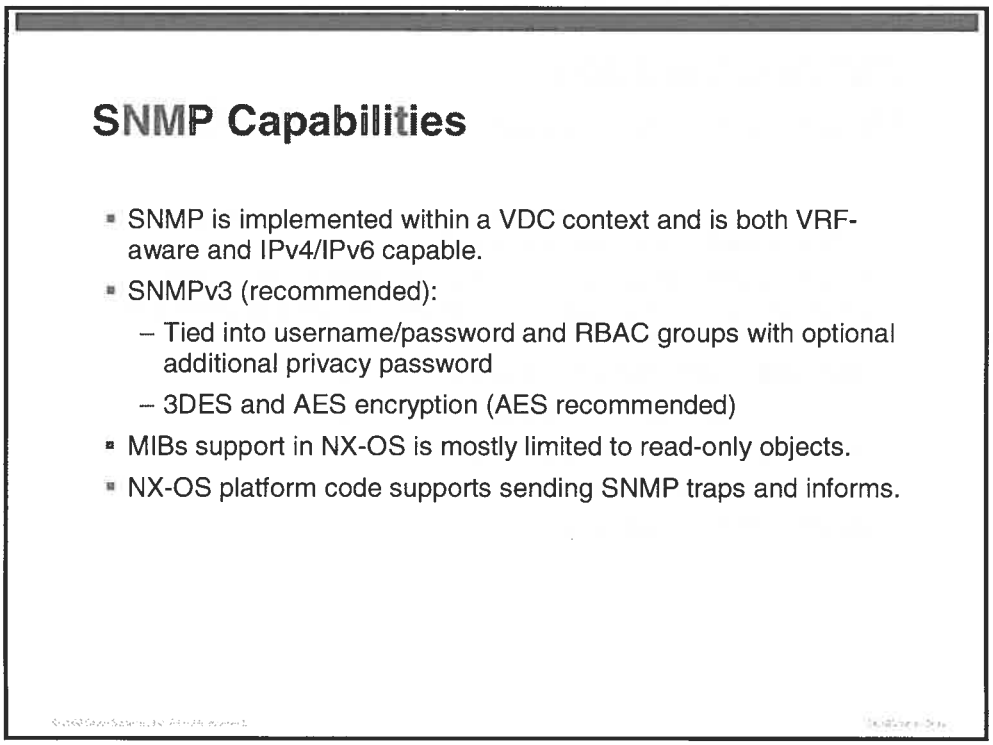
Fibre Channel and Ethernet Management



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

SNMP Capabilities



https://t.me/learningnets

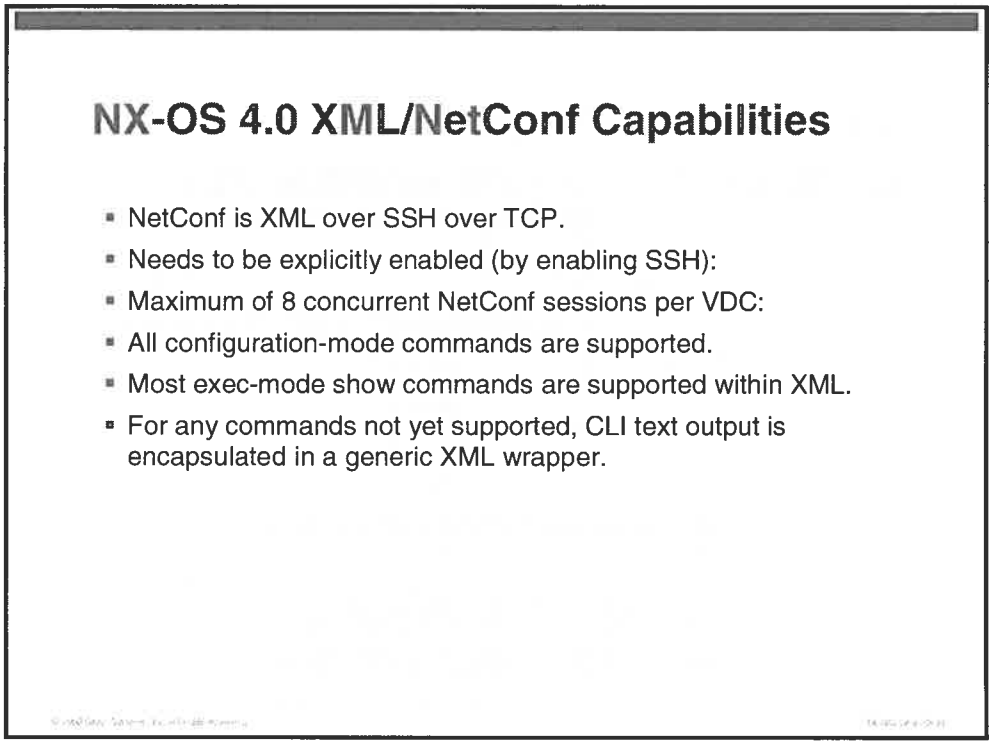
XML/NetConf API: Remedy for the Shortcomings of SNMP

- CLI scraping was still necessary because many statistics and counters were not available in MIBs.
- SNMP typically does not scale. UDP-based get next/response (even with bulk gets) causes excessive CPU load, and it is difficult for a client to collect many statistics simultaneously.
- Many MIBs suffer from 32-bit counter rollover.
  - At 10 GbE it takes <4 seconds to roll over.
- No guarantee that interfaces always keep their index:
- Binary protocols are difficult to extend.
- Reading MIBs is not easy.

© 2008 Cisco Systems, Inc. All rights reserved.

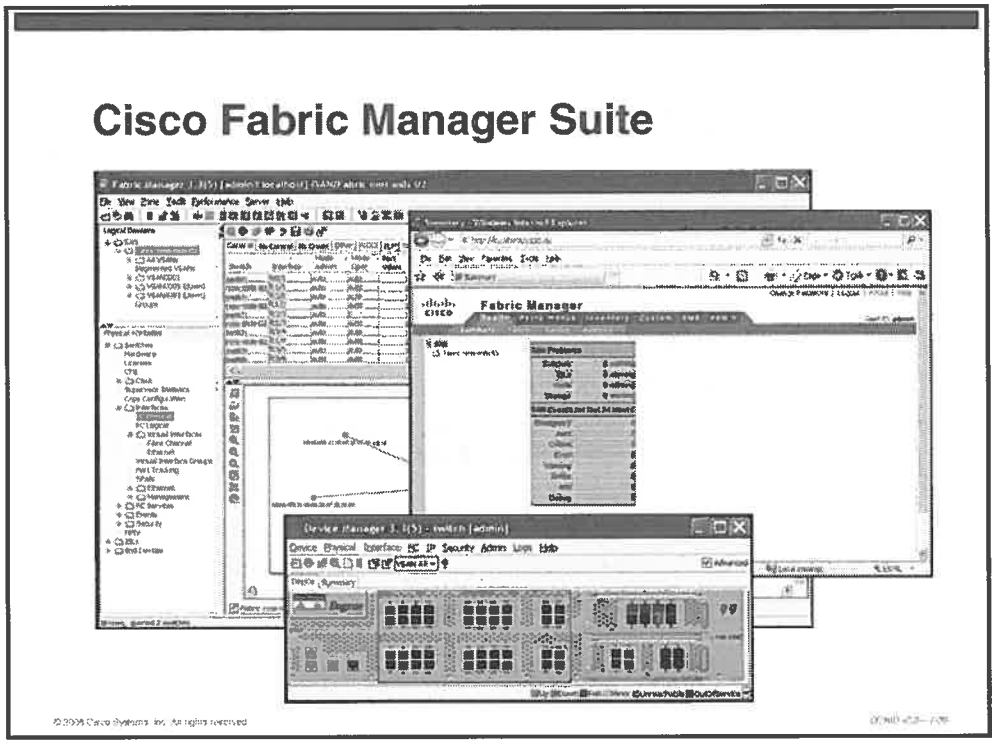
https://t.me/learningnets

NX-OS 4.0 XML/NetConf Capabilities



https://t.me/learningnets

Cisco Fabric Manager Suite



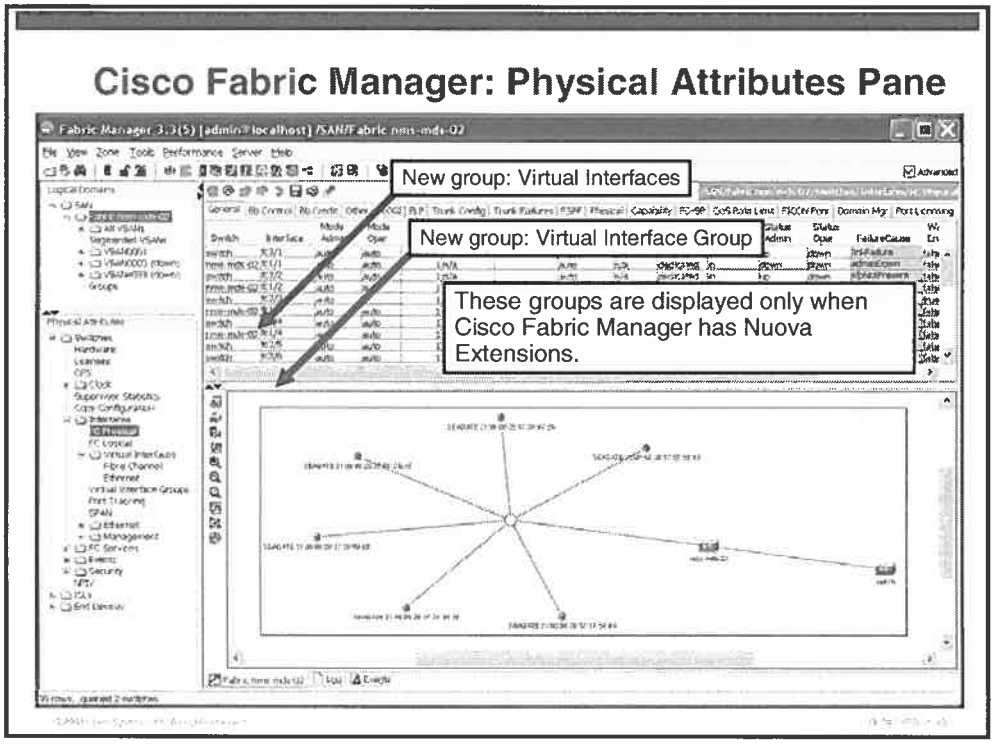
Monitoring an FCoE Network with Cisco Fabric Manager.

© 2008 Cisco Systems, Inc.

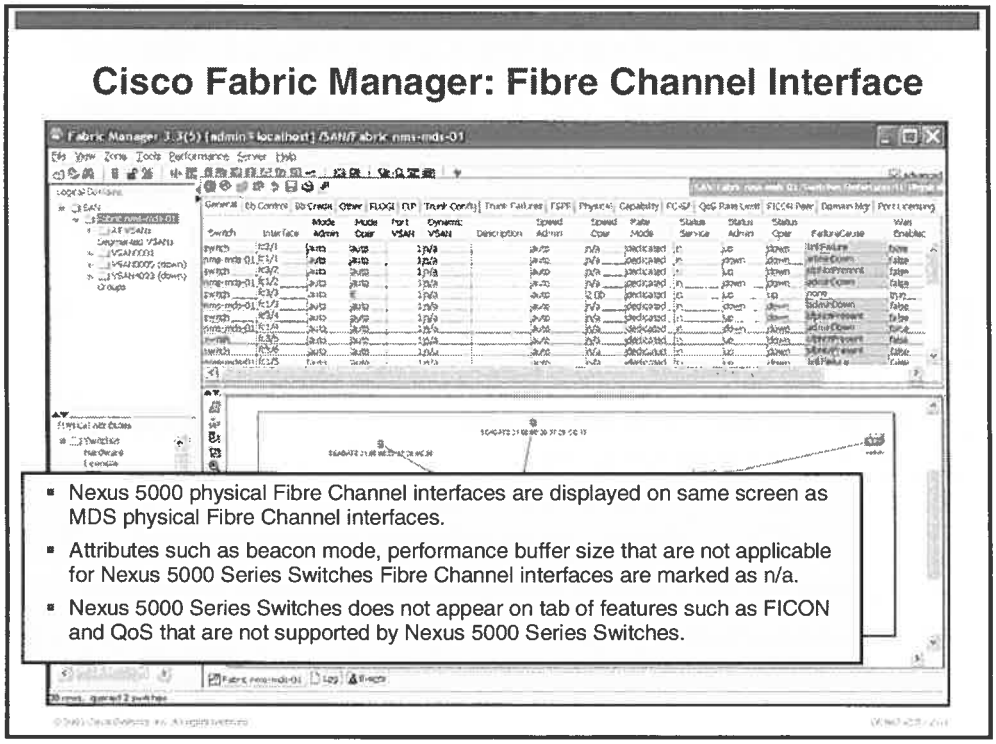
https://t.me/learningnets

Cisco Fabric Manager

Physical Attributes Pane

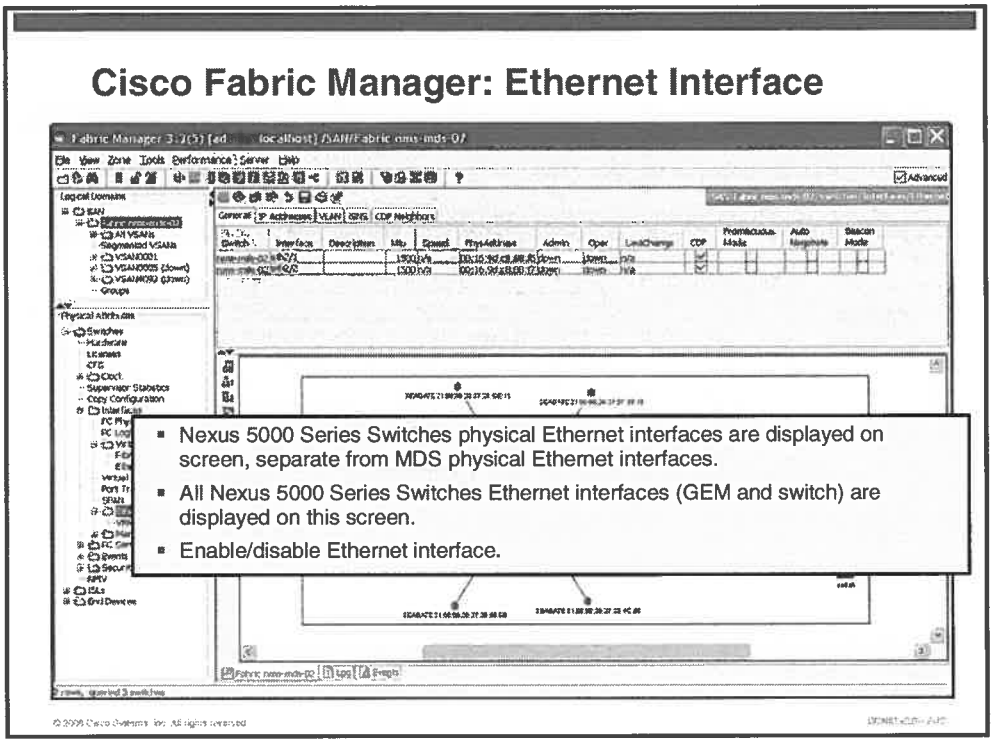


Fibre Channel Interface

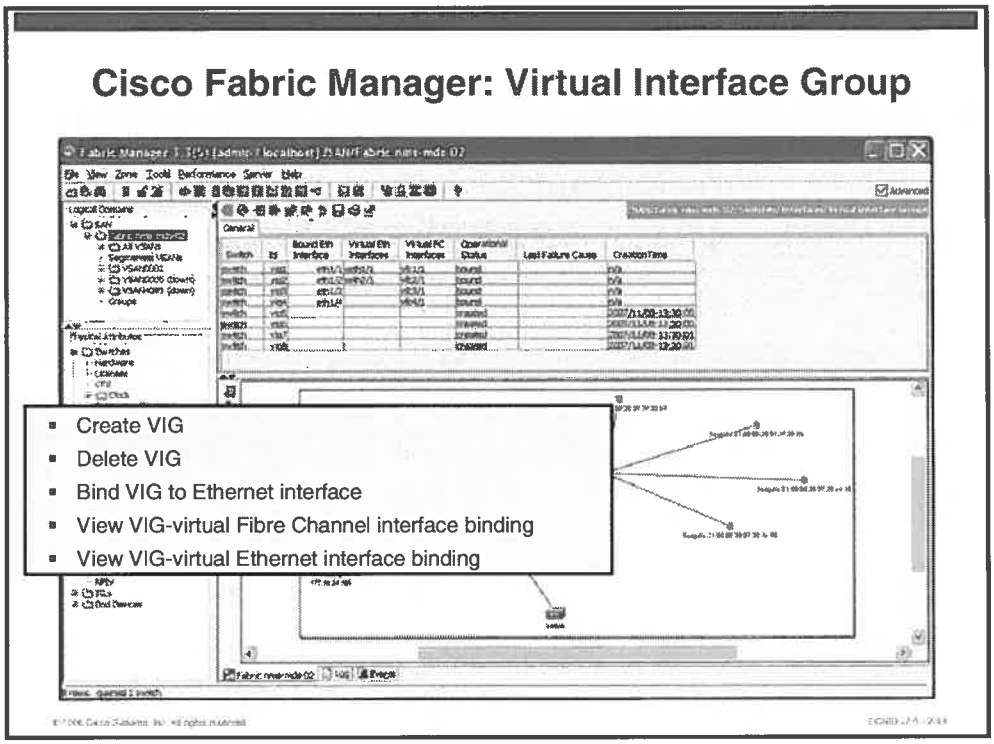


https://t.me/learningnets

Ethernet Interface



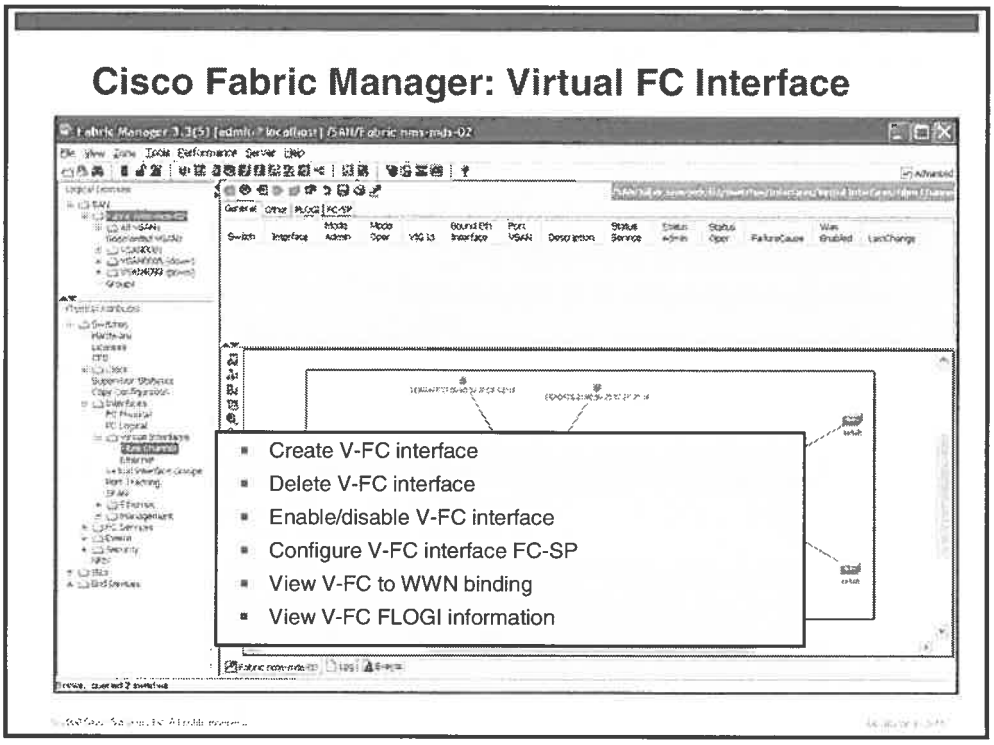
Virtual Interface Group



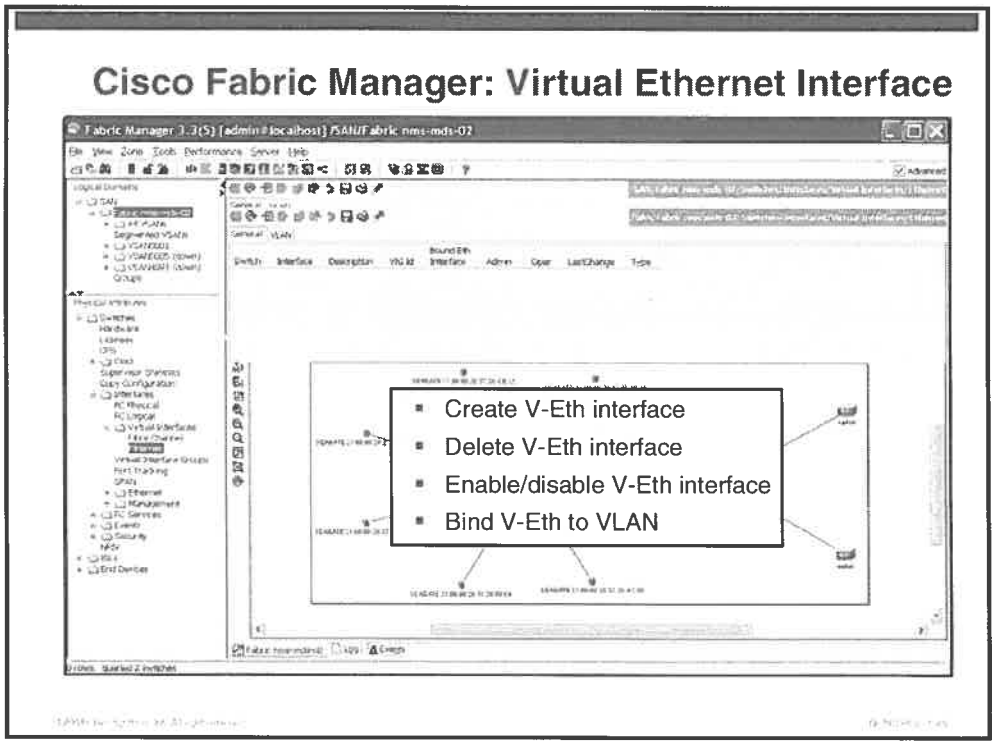
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Virtual Fibre Channel Interface



Virtual Ethernet Interface

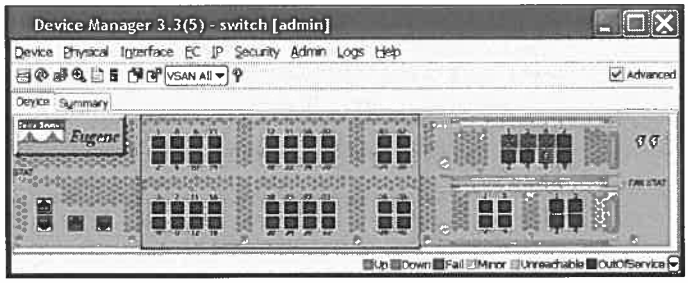


https://t.me/learningnets

Device Manager

Device Manager

- Manages single device at one time
- Device-centric rather than fabric-centric
- Pure Java application running on Windows and Linux
- Distributed with part of NXOS, resides on switch flash
- Java Web Start-based installation



Monitoring an FCoE Network with Cisco Fabric Manager.

© 2008 Cisco Systems, Inc.

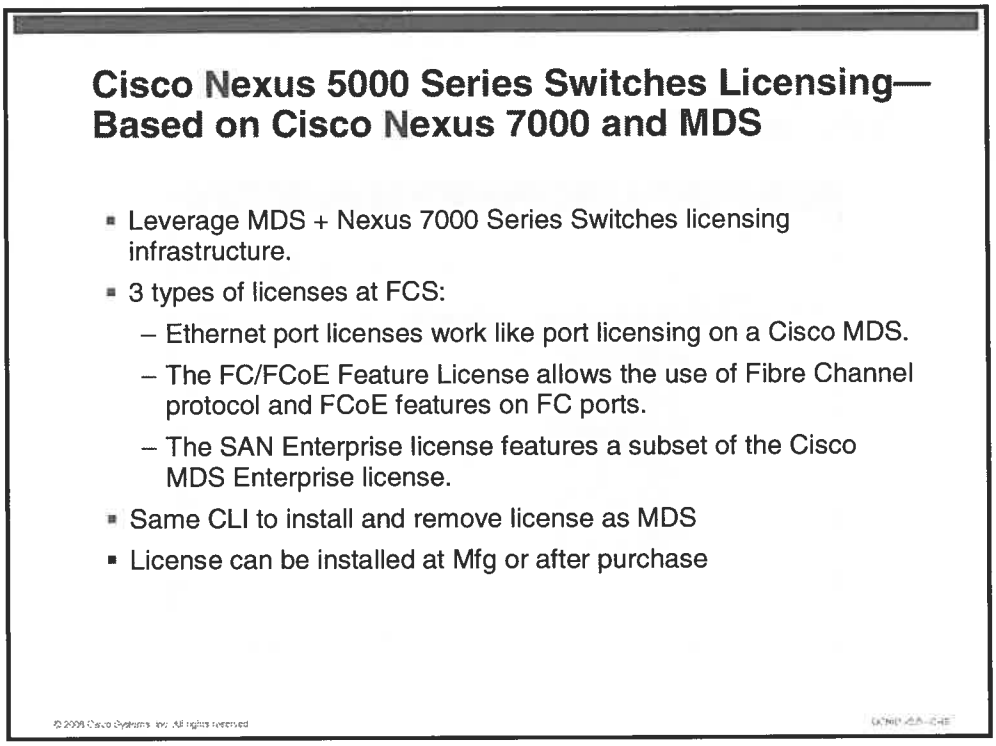
https://t.me/learningnets

Web Services



https://t.me/learningnets

Cisco Nexus 5000 Series Switches Licensing



Feature-Based Licenses

- **Base Services Package**:

  This package is included with the switch hardware at no additional charge. It includes all Ethernet and system features, except features explicitly listed in the storage Services Package.

- **Storage Services Package**:

  — N5020-SS includes the following services for each NX5020 system:

  - Native Fibre Channel
  - FCoE
  - NPV
  - Fibre Channel Port Security
  - Fabric Binding

- **Advanced Services Package:**

  This package will be available in a future release.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

License Installation

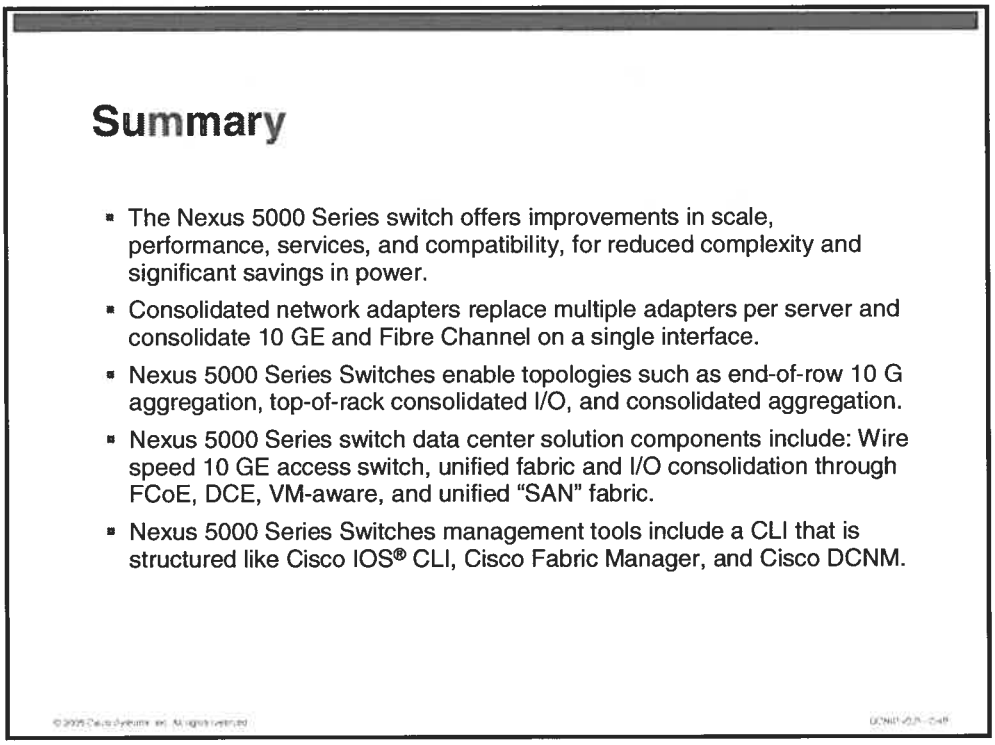
You can obtain a factory-installed license (applies only to new device orders). You can obtain factory-installed licenses for a new Cisco NX-OS device. To obtain a factory-installed license, follow these steps:

1. Contact your reseller or Cisco representative and request this service.
2. Obtain the host ID from the proof-of-purchase document for future use.
3. Start to use the device and the licensed features.

All Cisco Nexus 5000 Series Switches licenses are factory-installed. Manual installation is not required.

https://t.me/learningnets

Summary



- Connects to 10 GE Ethernet module and Cisco Nexus 7000 Series Switches aggregation, and Fibre Channel/storage through Multilayer Director Switch (MDS).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Module 4

Data Center Design Models

Overview

This module explains the data center network design strategy at component, network and architecture levels.

Module Objectives

Upon completing this module, you will be able to analyze the layers of data center networking architecture. This ability includes being able to meet these objectives:

- Explain the architecture of data center networks based on the Enterprise Composite Network Model (ECNM) and the High-Speed Data Module (HDM)
- Discuss the impact of data center application design on the network architecture

https://t.me/learningnets

© 2008 Cisco Systems, Inc.
https://t.me/learningnets

Lesson 1

Data Center Design Overview

Overview

This lesson describes and reviews the evolution of the data center design models, including the Enterprise Composite Network Model (ECNM) and the multi-tier model.

Objectives

Upon completing this lesson, you will be able to discuss the evolution of the data center design models.

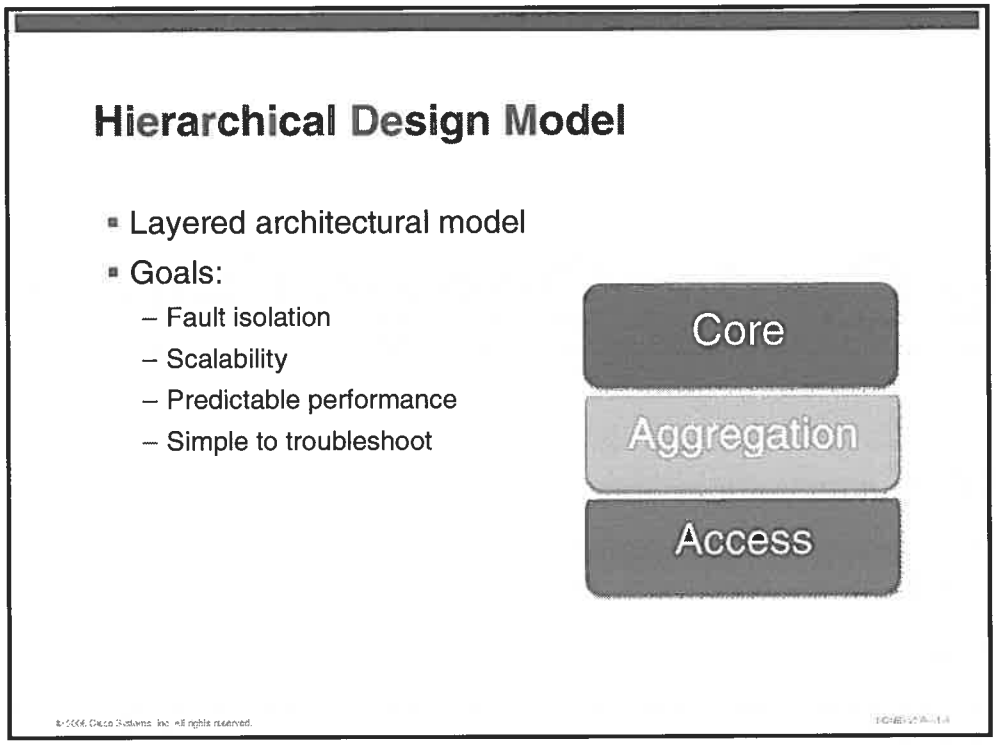
This ability includes being able to meet these objectives:

- Identify the layers in the data center multi-tier model
- Review the ECNM to provide a starting point for the data center design

https://t.me/learningnets

Multi-Tier Model Design

This topic discusses the components of the multi-tier model design model.



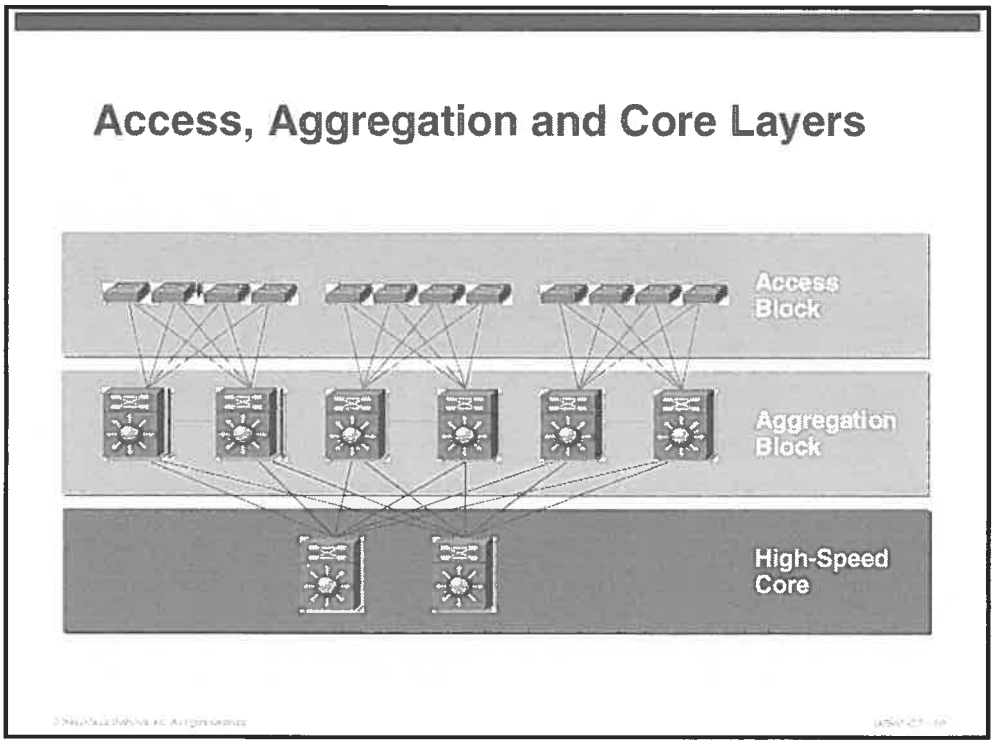
The Hierarchical Design Model (HDM) was created to facilitate network designs that were scalable and that provided consistent design and performance. Both small and large enterprise campuses require a highly available and secure, intelligent network infrastructure to support business solutions such as voice, video, wireless, and mission-critical data applications. The use of hierarchical design principles provides the foundation for implementing campus networks that meet these requirements. The hierarchical design uses a building block approach, leveraging a high speed routed core network layer to which are attached multiple independent distribution blocks. The distribution blocks comprise two layers of switches: the actual distribution nodes that act as aggregators and the wiring closet access switches. The hierarchical design segregates the functions of the network into these separate building blocks to provide for flexibility, scalability and fault isolation.

The HDM divides the network design into three layers. The functionality of each layer is defined to give the network designer guidance for the features to be supported at each layer, which in turn guides the device-selection process.

A hierarchical network design distributes networking functions at each layer through a layered organization. Modular designs are made out of building blocks. Modules can be added or removed without redesigning the network. A modular design is also easier to grow and troubleshoot. The Cisco multilayer design is an example of a modular hierarchical design model. The key elements of the structured hierarchy are the core, distribution, and access layer in a network.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

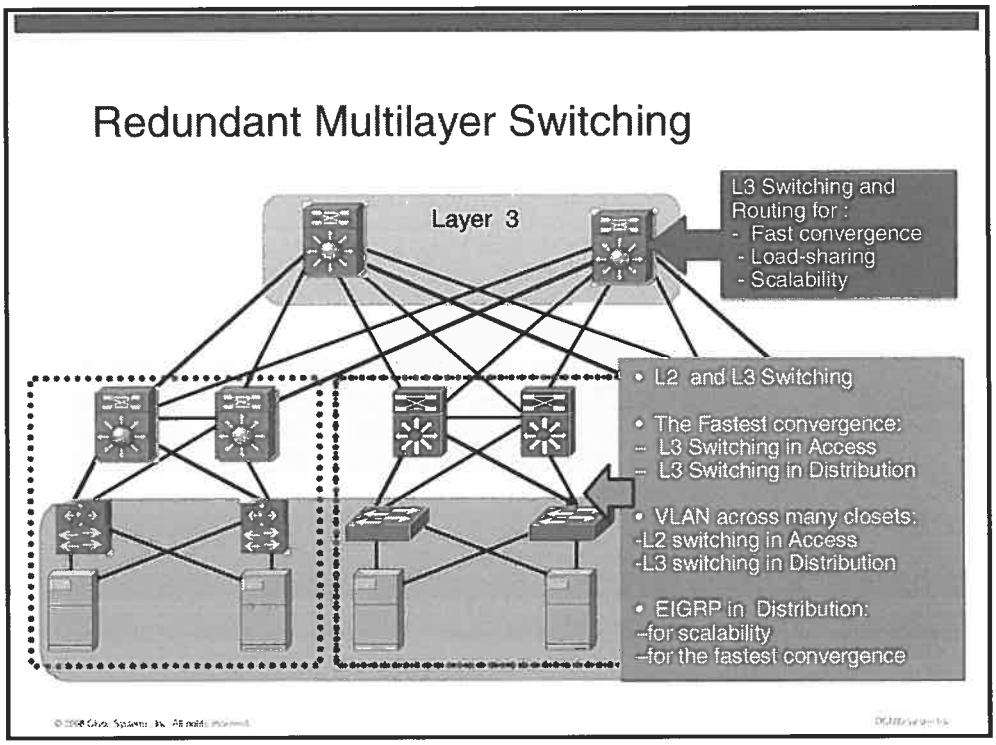


The first layer of the HDM is the access layer, also often referred to as the network edge. The access layer is the point at which devices that use the network for data transfer connect to the network. Different types of ports are required at the access layer depending on the type of device that is to be attached to the network. WANs have ports for communication circuits, such as T-1s and T-3s as the access ports, while LANs have copper or fiber Ethernet ports.

The second layer of the HDM is the aggregation layer, also called the distribution layer. The aggregation layer devices aggregate the traffic from the access layer devices. The aggregation layer is usually the layer in which network usage and performance management policies are controlled through the use of policy-driven features, such as access lists and quality of service (QoS) controls. The distribution block provides for policy enforcement and access control, route aggregation, and the demarcation between the layer 2 subnet (VLAN) and the rest of the Layer 3 routed network.

The core layers of the network provide for high capacity transport between the attached aggregation building blocks. The core is a high-speed switching backbone designed with reliability, redundancy, and the ability to switch packets as fast as possible. Early design guidelines for the core layer emphasized the ability of the core to switch packets fast and cautioned against doing any packet manipulation or filtering. Modern switch devices, such as the Cisco Nexus 7000 Series Switches and the Cisco Catalyst 6500 Series Switches, are capable of making routing and switching decisions at wire speed, relaxing some of the limitations on functions applied at the core layer while maintaining network performance.

https://t.me/learningnets



The preceding figure shows a more modern design. Although the technology and capabilities of this network far surpass the earliest network design, the hierarchical design model is still used. Access layer devices are still implemented with Ethernet switches or DCE Ethernet switches, such as the Cisco Nexus 5000 Series switch. The distribution and core layers are now implemented with redundant pairs of multilayer switches, such as Nexus 7000 Series or Catalyst 6500 Series switches, capable of making Layer 2 switching and Layer 3 routing decisions at wire speed.

Each building block within the network leverages appropriate switching technologies to best meet the element's architecture. The core layer of the network utilizes Layer 3 switching (routing) to meet its need for scalability, load sharing, fast convergence, and high speed capacity.

Each distribution block utilizes a combination of Layer 2 and Layer 3 switching to provide for the appropriate balance of policy and access controls, availability, and flexibility in subnet allocation and VLAN usage.

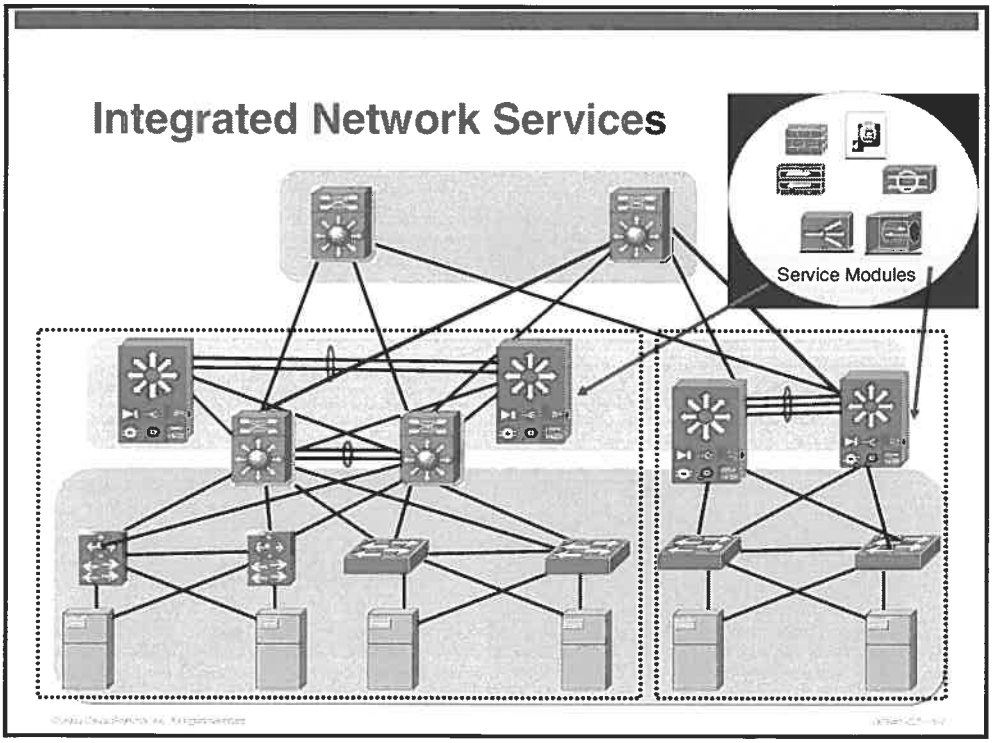
For those campus designs requiring greater flexibility in subnet usage (for instance, there may be situations where VLANs must span multiple wiring closets), distribution block designs using Layer 2 switching in the access layer and Layer 3 switching at the distribution layer will provide the best balance for the distribution block design.

For campus designs requiring the fastest convergence, distribution block designs using Layer 3 switching in the access layer (routed access) in combination with Layer 3 switching at the distribution layer will provide for the best balance of functionality.

For those networks using a routed access (Layer 3 access switching) within their distribution blocks, we recommend Enhanced Interior Gateway Routing Protocol (EIGRP) be used as the campus Interior Gateway Protocol (IGP). Using EIGRP will provide for faster convergence and better scalability than a design using Open Shortest Path First (OSPF) or Routing Information Protocol (RIP). It will also provide a more reliable and manageable network than a design constructed with static routes.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

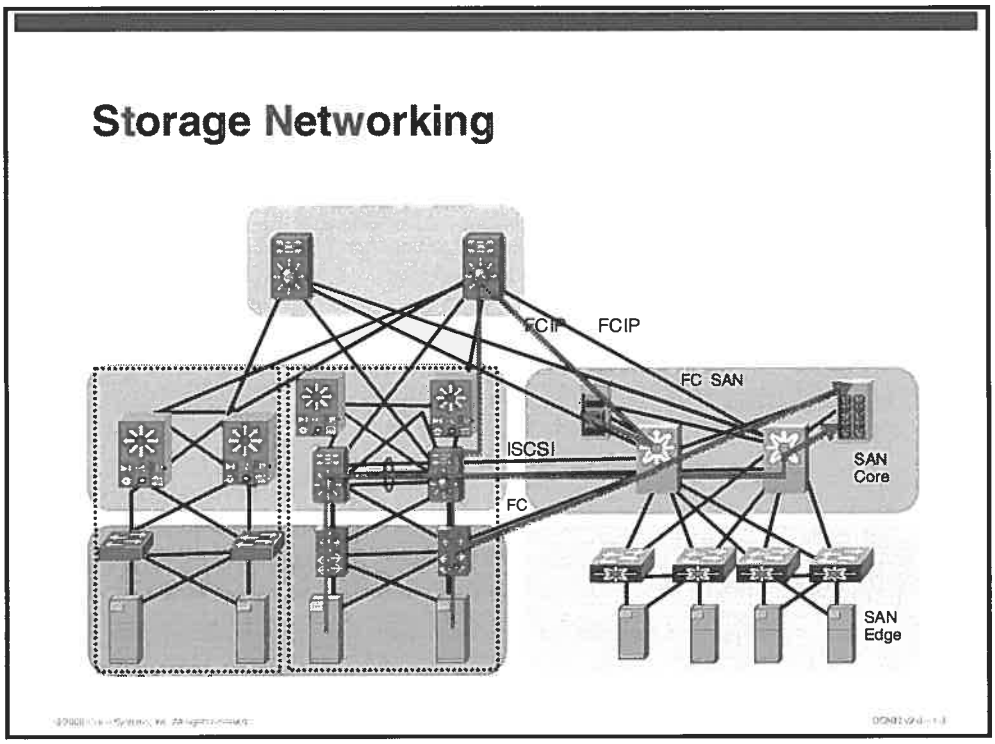


As networks have grown more complex, the proliferation of network appliances that perform functions like firewalling, load-balancing, intrusion detection (IDS), intrusion prevention (IPS), virtual private networking (VPN), and network monitoring has grown increasingly difficult to manage. Cisco pioneered the approach of integrating many of these network services into the switch chassis.

A modern 3-tier network design integrates appropriate network services into the aggregation layer, evolving the function of the aggregation layer into a service layer. Services can be integrated using the Catalyst 6500 Series switch with integrated Network services modules in the distribution layer or using a Cisco Catalyst Layer 3 Series switch with external services modules. When access layer traffic is heterogeneous, for example DCE (Data Center Ethernet) traffic (carrying Fibre Channel and data traffic), and is passed up to the aggregation layer, then the Cisco Nexus 7010 Series switch is needed in the aggregation layer, with an attached Catalyst 6500 Series switch for the network services integration.

The Nexus 7010 Series switch provides the highest throughput and availability, as well as a converged networking solution. The Catalyst 6500 Series switches with integrated services modules can be enabled in Virtual Switching System (VSS) mode. The Nexus 7010 Series switch enables access layer servers from different workgroups to communicate. For example, it can aggregate Fibre Channel over Ethernet (FCoE) traffic from the access layer and pass it in the form of the classical Ethernet to the core or to the access layer.

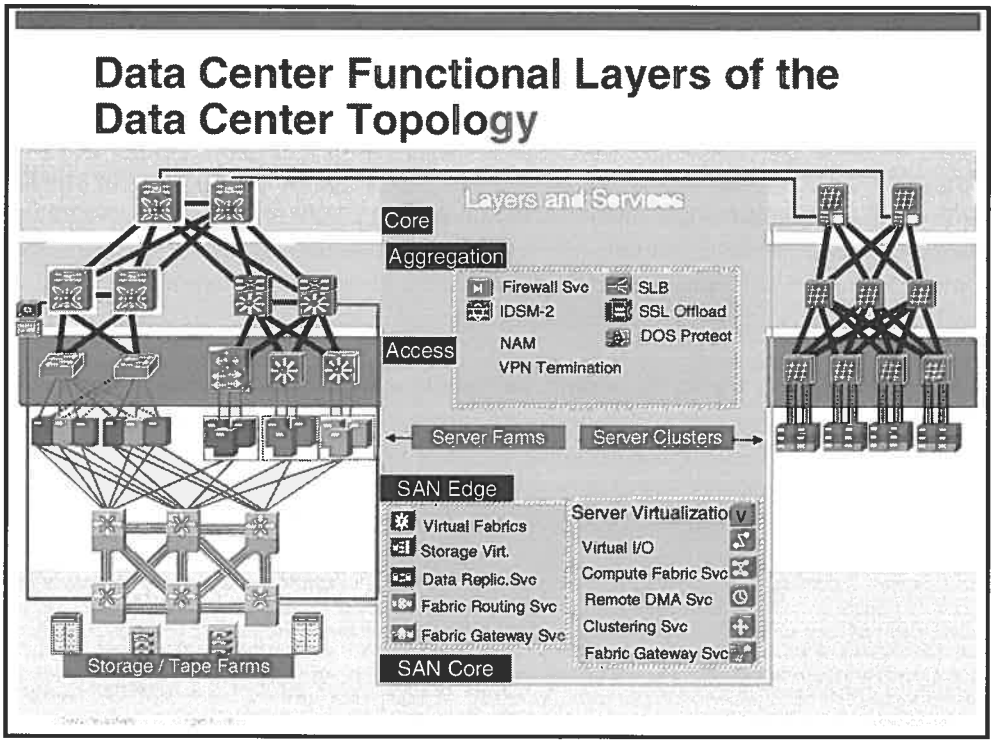
https://t.me/learningnets



While storage networks have been traditionally separate from the IP network, the back-end storage and front-end data networks can now converge with the Cisco Nexus 7000 and Nexus 5000 Series switches using DCE protocol, storage-over-IP protocols like Cisco MDS Fiber Channel over IP (FCIP) and Internet Small Computer Systems Interface (iSCSI). Today, servers in the access layer can establish an FCoE connection across the Nexus 5000 Series switch in the access layer providing the iSCSI or FCIP tunnel to a Cisco MDS Director switch in the storage core. An iSCSI or MDS tunnel can be established across the data center core network. When a native Fibre Channel connection is needed to the Cisco MDS storage core, the server can use an FCoE link to the Nexus 5000 Series switch and then a Fibre Channel connection to the Cisco MDS Director switch in order to read/write to storage arrays or tape devices.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The traditional hierarchical model when empowered with integrated services provides the foundation for virtualization, segmentation and data center operation automation. The data center infrastructure is central to the IT architecture, from which all content is sourced or passes through. Proper planning of the data center infrastructure design is critical, and performance, resiliency, and scalability need to be carefully considered. Another important aspect of the data center design is flexibility in quickly deploying and supporting new services or so called network scalability. Designing a flexible architecture that has the ability to support new applications in a short timeframe can result in a significant competitive advantage. Such a design requires solid initial planning and thoughtful consideration in the areas of port density, access layer uplink bandwidth, true server capacity, and oversubscription, to name just a few. The data center network design is based on a proven layered approach, which has been tested and improved over the past several years in some of the largest data center implementations in the world. The layered approach is the basic foundation of the data center design that seeks to improve scalability, performance, flexibility, resiliency, and maintenance. The preceding figure shows the basic data center layered design. The layers of the data center design are the core, aggregation, and access layers. These layers are referred to extensively throughout this guide and are briefly described as follows:

Core Layer

The core layer provides the high-speed packet switching backplane for all flows going in and out of the data center. The core layer provides connectivity to multiple aggregation modules and provides a resilient Layer 3 routed fabric with no single point of failure. The core layer runs an interior routing protocol, such as OSPF or EIGRP, and load balances traffic between the campus core and aggregation.

https://t.me/learningnets

Aggregation Layer Modules

Aggregation layer modules provide important functions, such as service module integration, Layer 2 domain definitions, spanning tree processing, and default gateway redundancy. Server-to-server multi-tier traffic flows through the aggregation layer and can use services, such as firewall and server load balancing, to optimize and secure applications. The smaller icons within the aggregation layer switch in the preceding figure represent the integrated service modules. These modules provide services, such as content switching, firewall, Secure Socket Layer (SSL) offload, intrusion detection, network analysis, and more.

Access Layer

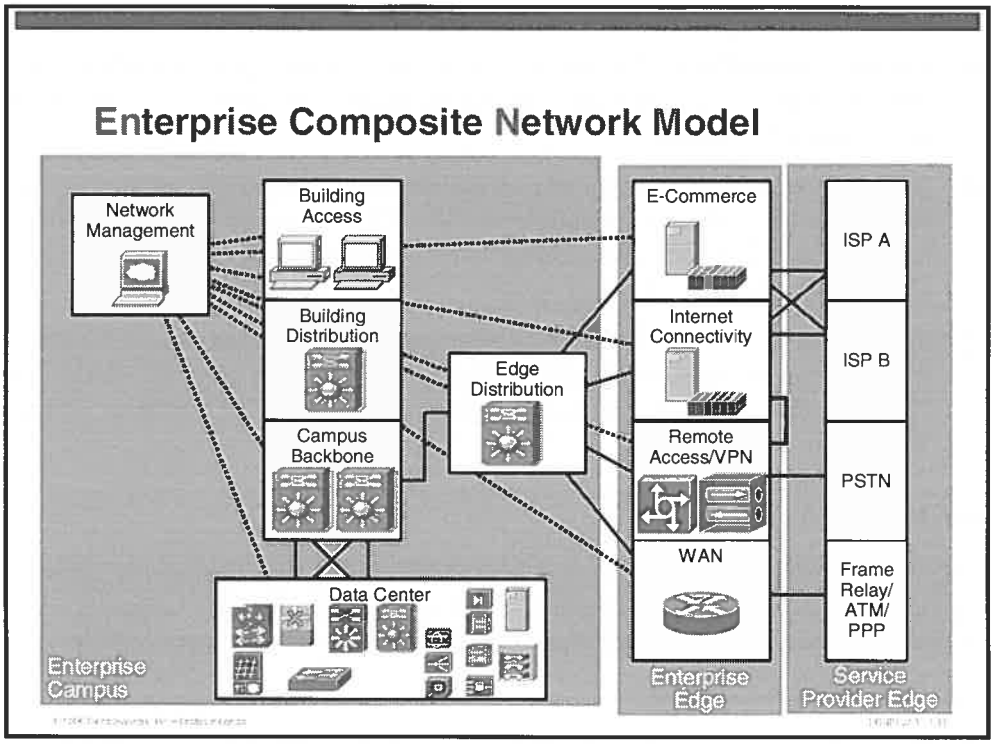
The Access layer is where the servers physically attach to the network. The server components consist of 1RU servers, blade servers with integral switches, blade servers with pass-through cabling, clustered servers, and IBM mainframes with Open Systems Adapters (OSA) adapters. The access layer network infrastructure consists of modular switches, fixed configuration 1 or 2RU switches, and integral blade server switches. Switches provide both Layer 2 and Layer 3 topologies, fulfilling the various servers' broadcast domain or administrative requirements.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

The Enterprise Composite Network Model

This topic describes the Enterprise Composite Network Model (ECNM), which is used as a starting point for the data center design.



The ECNM divides an enterprise-wide network design into functional areas. Each functional area consists of several modules, each of which is addressed separately from a design perspective.

The preceding figure shows an overview of the modules contained in a corporate network.

Enterprise Campus

The enterprise campus functional area contains all the network resources within a single campus that might span multiple buildings. Multiple building access and building distribution modules are designed much as they were using the HDM, one for each building.

The campus backbone performs the same functions as the core layer in the HDM. The server farm module contains all the campus servers and the network equipment used to attach the servers to the campus backbone. An additional network management module provides both in-band and out-of-band management of all network devices within the campus or enterprise.

The edge distribution module connects the campus network to larger networks that connect to the rest of the enterprise network at other campuses or to external users and networks.

https://t.me/learningnets

Enterprise Edge

The enterprise edge functional area contains the following modules, which connect the campus to resources off campus:

- **E-commerce:** The e-commerce module contains the systems that provide e-commerce access to enterprise clients via the Internet.
- **Internet connectivity:** The Internet connectivity module provides Internet access for systems on the campus and also handles incoming connections from the Internet that are not e-commerce related.
- **Remote access/VPN:** The remote access/VPN module provides attachment to the external phone network and the termination point for VPN connections coming in from dial-up or Internet users. The remote access/VPN module provides access to the internal enterprise network for authorized users.
- **WAN:** The WAN module connects the campus network to private WAN resources, typically used to implement a private enterprise long-haul network between campuses. The WAN module is designed with the assumption that the enterprise owns or controls the WAN to which it attaches.

Service Provider Edge

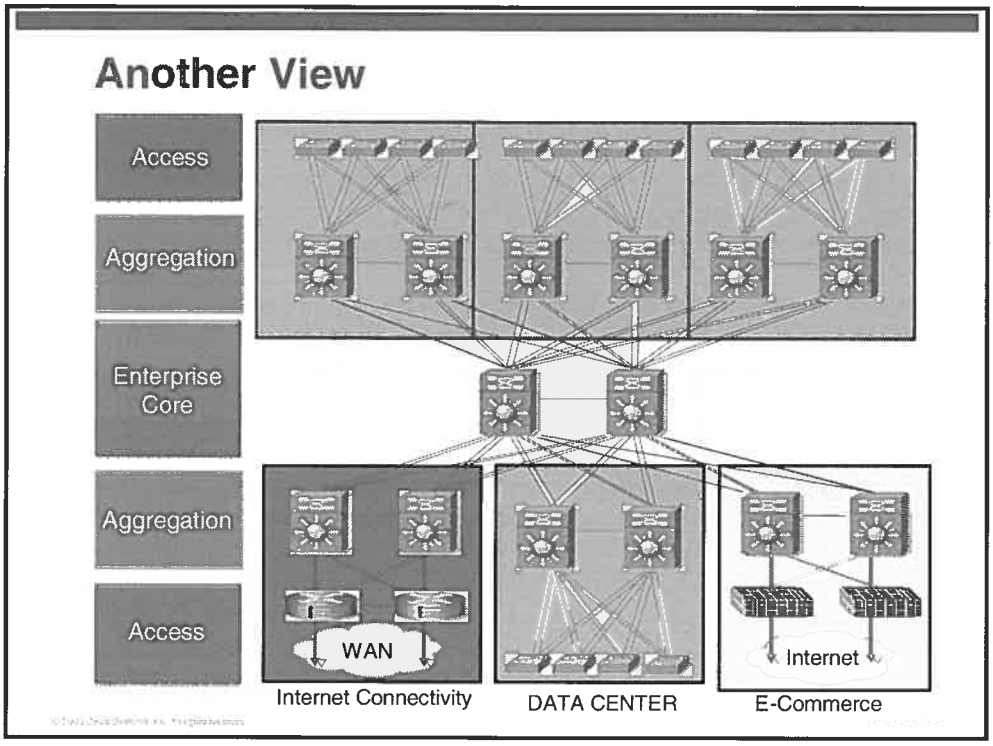
The service provider edge functional area consists of modules that are not owned or managed by the enterprise. Instead, the modules of the service provider edge represent attachment to communications services purchased from outside telecommunications vendors.

The modules in this area are:

- **ISP:** ISP A and ISP B are redundant connections to the public Internet that are purchased through different organizations.
- **Public switched telephone network (PSTN):** The PTSN module provides transit for dial-up connections.
- **Frame/Relay/ATM/PPP:** This service provider module furnishes communications circuits that are dedicated to the enterprise network and are used to connect campuses together.

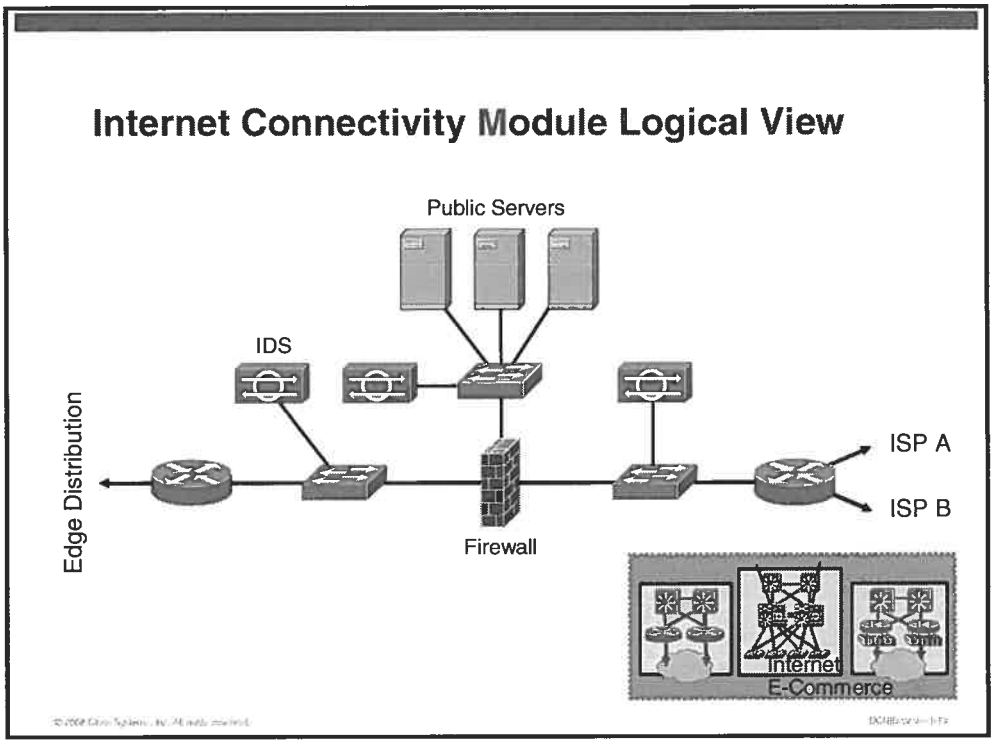
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The preceding figure shows a view that maps some of the ECNM modules onto the core-aggregation-access model.

https://t.me/learningnets



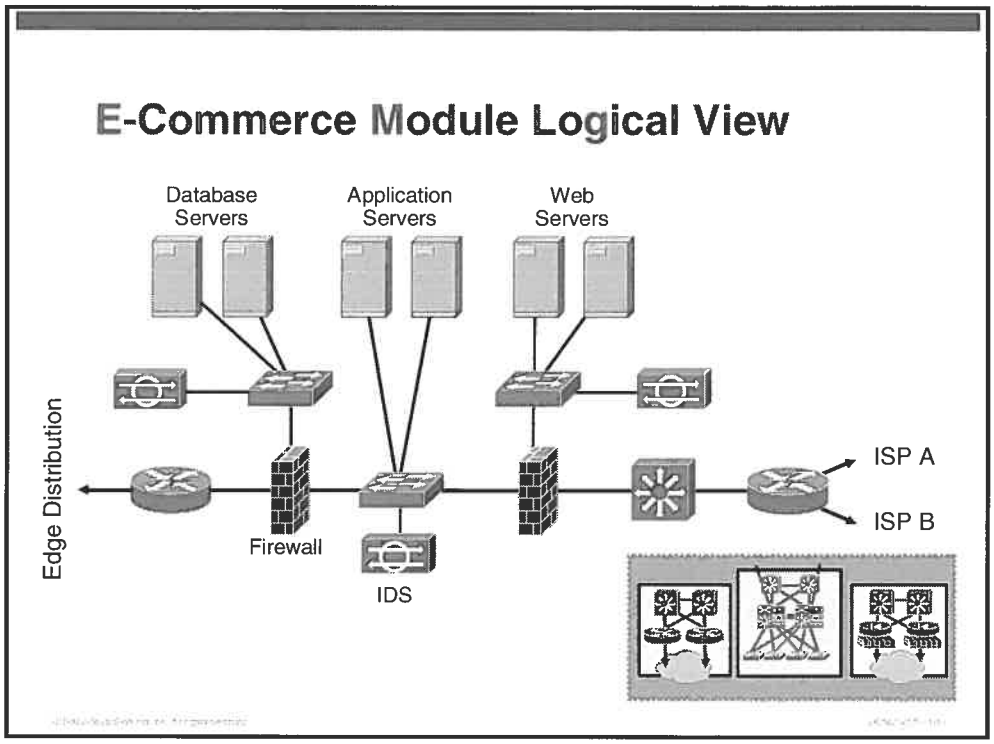
The Internet connectivity module, illustrated in the preceding figure, provides connectivity to the Internet for all functions except e-commerce. This module supports traffic for all incoming connections that are not e-commerce related, and all outgoing connections that are used by the enterprise to access resources on the Internet.

The major components of the Internet connectivity module are:

- **Public servers:** These servers include the corporate e-mail servers, authoritative external Domain Name System (DNS) servers that also relay internal DNS requests and public web servers that supply information about the enterprise.
- **Security devices:** The security devices in this module control communications through the various levels of security in the Internet connectivity module and include firewalls and IDSs.
- **Edge routers:** The edge routers provide Layer 3 connectivity between the Internet connectivity module, the Internet, and the enterprise edge module. The edge routers can also be configured to perform basic filtering as part of the security policy.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

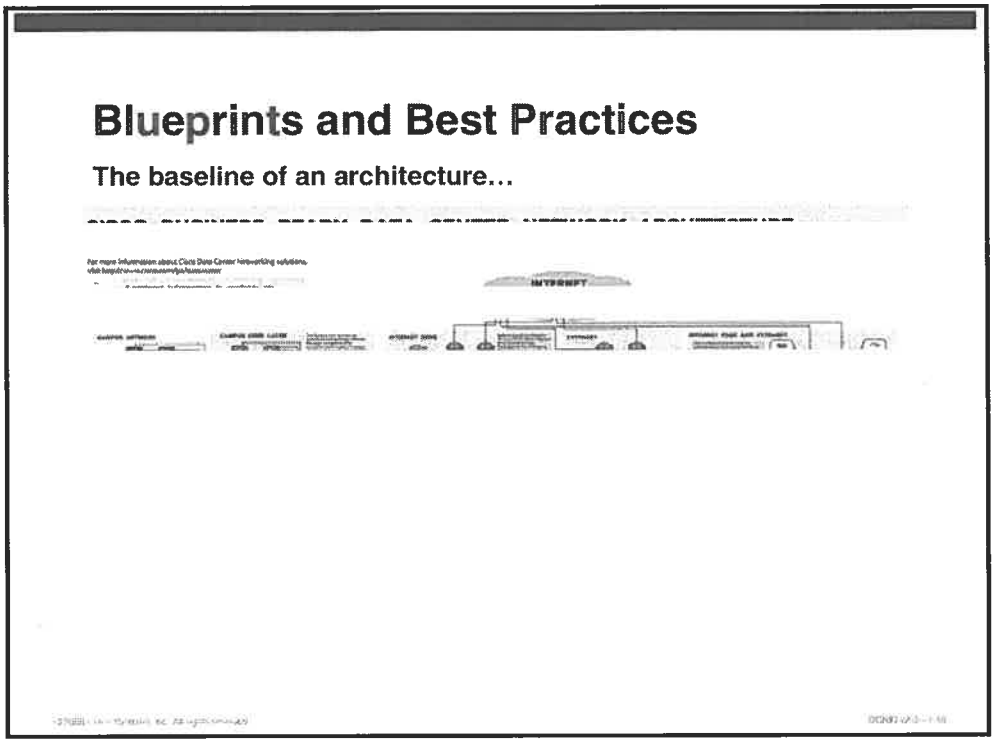


The e-commerce module, shown in the preceding figure, provides all the components necessary to host an e-commerce presence on the Internet. The e-commerce module supports all of the traffic related to e-commerce activities.

The major components of the e-commerce module are:

- **Web servers:** The web servers act as the primary interface to the e-commerce customer and provide static resources and the templates that are used to present dynamic resources that are built by back-end servers.
- **Applications servers:** The application servers support e-commerce applications including online catalog browsing, ordering, transaction processing, and customer status queries. The application servers build dynamic content for web interaction, which is then placed along with static content according to templates served from the web servers.
- **Database servers:** The database servers store the database of information that is queried and updated by applications running on the application servers.
- **Security devices:** The security devices analyze and control traffic flow among the various security levels in the e-commerce module. Security devices such as firewalls and IDS are common in the e-commerce module.
- **Edge routers:** The edge routers provide connectivity to the Internet and to the edge distribution module and can be configured to perform basic filtering and security functions.

https://t.me/learningnets



Cisco offers design guides and reference architectures that are available to you and your customers on the Cisco.com website. These documents represent many years of experience and expertise in the Enterprise Solutions Engineering team. You will find them valuable aids to reduce your time to deployment of a single integrated network architecture. These guides and architectures have been tried and tested, offering a highly resilient framework for your mission-critical data center environments. They have been tested as generic Cisco infrastructures and with Cisco partners such as IBM and EMC.

Internet Edge Module

Internet edge module supports several functions that include the following:

- Securing the enterprise data center network
- Controlling Internet access from the intranet
- Controlling the access to the Internet server farms

Internet Server Farms

The Internet edge server farm faces the Internet. This implies that users accessing the server farms are located on the Internet and that they use the Internet to reach the Internet server farm. Internet server farms support business-to-consumer services that are based on web browsers. The core business is based on the Internet presence or on the Internet e-commerce.

There are two types of internet server farms:

- Dedicated server farm
- DMZ server farm

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

The Dedicated Internet server farm supports core business functions and is based on the Internet Presence or the Internet commerce. The second type of the Internet server farm is the **DMZ server farm,** which is built to support the Internet-based applications, in addition to the Internet access from the enterprise users. This is called a DMZ server farm since it is located in the DMZ zone, which has different security and access restrictions from the Internet dedicated server farm. Security requirements are very stringent because they are aimed to protect the server farms from external users while keeping the enterprise network safe.

Behind the Internet edge there is Campus network, intranet server farm (Campus primary and secondary data center) and Private data center WAN.

Data Center Campus Server Farm

Data center intranet server farms resemble the Internet server farm in the ease of access, yet they are available only to the enterprise internal users. It is also referred to as an intranet server farm. An intranet server farm includes most of the enterprise-critical computing resources that support business processes and internal applications. The data center server farm is connected to the data center Campus core switches that form a portion of the enterprise backbone and provide the connectivity between the Internet edge and the data center private WAN.

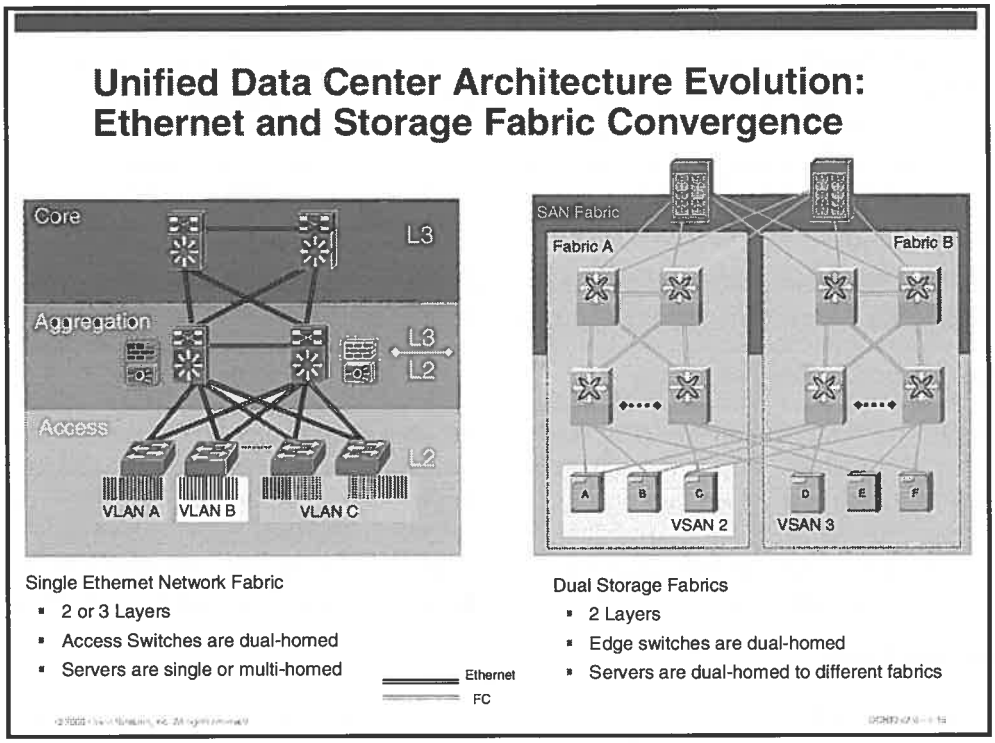
Extranet Server Farm

From the functional perspective, the extranet server farm sits between the Internet and data center campus server farms. Extranet server farms continue the trend to use the web-based applications but they are accessed only by the selected group of users such as partners that are neither intranet nor Internet. The purpose of the extranet server farm is to provide server farm services to the trusted external end users. Business partners have access to the subset of business applications but are restricted from accessing the rest of the enterprise network. Access from the extranet to the intranet is prevented or highly secured. Scalability, availability and security are the major design concerns of the extranet server farm.

The server farm design guidelines do not apply equally to all types of the server farms.

© 2008 Cisco Systems, Inc.

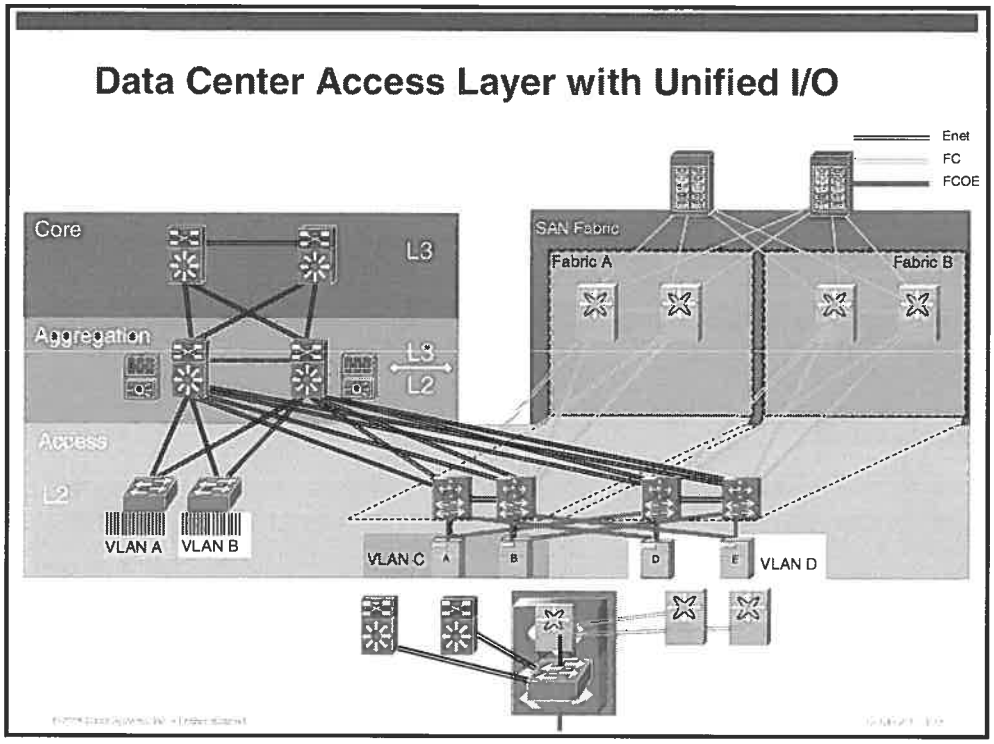
https://t.me/learningnets



The starting point for the data center Ethernet and storage fabric convergence are two separate Ethernet and Fibre Channel storage networks, which are slowly converging on the common data center network using unified I/O protocols.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



New data center access layers consist of Cisco Catalyst switches providing 10Gb classical Ethernet connections and Nexus 5000 Series switches providing unified 10Gps Ethernet connections: transporting classical Ethernet and FCoE.

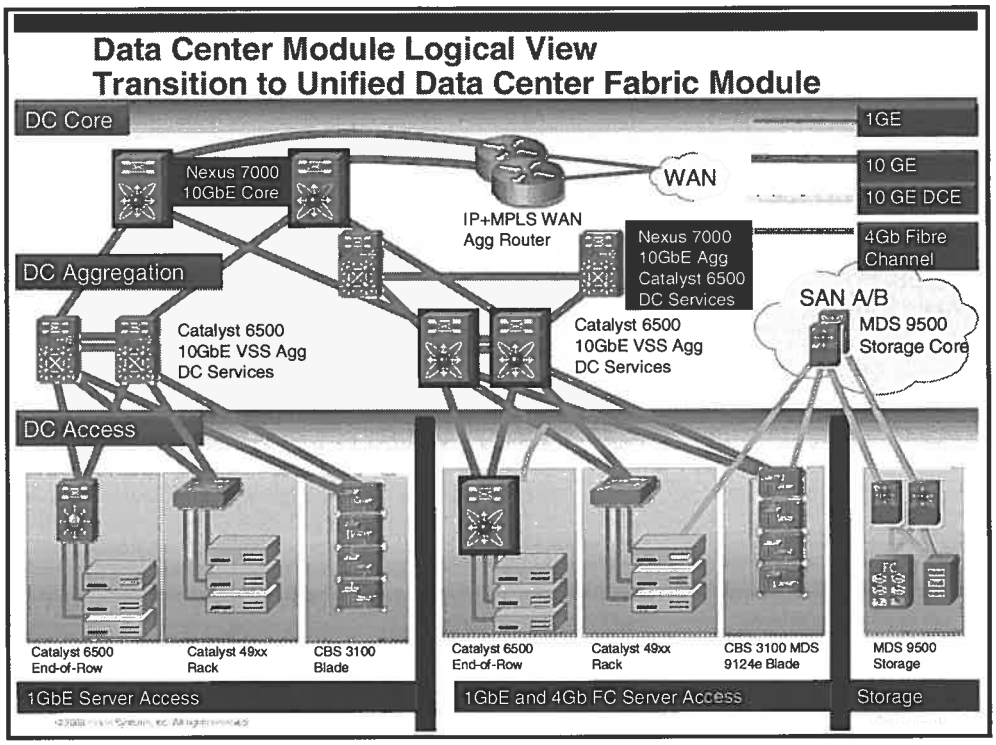
Cisco Nexus 5000 Series switches can operate simultaneously as classic Catalyst Layer 2 switch and like Cisco MDS switches on Fibre Channel ports (Fibre Channel trunking). FCoE connection is supported only between the Server and the Nexus 5000 Series switch. The Nexus 5000 Series switch supports 1, 2, or 4Gb Fibre Channel connections to Cisco MDS-based SANs.

FCoE connections currently are *not* supported between Nexus 5000 switches; therefore, Fibre Channel ISL cannot be extended from Fibre Channel SAN across Nexus 5000 fabric back to Fibre Channel SAN.

The data center access layer is now built of classical Catalyst Ethernet switches and Nexus 5000 Series switches.

© 2008 Cisco Systems, Inc.

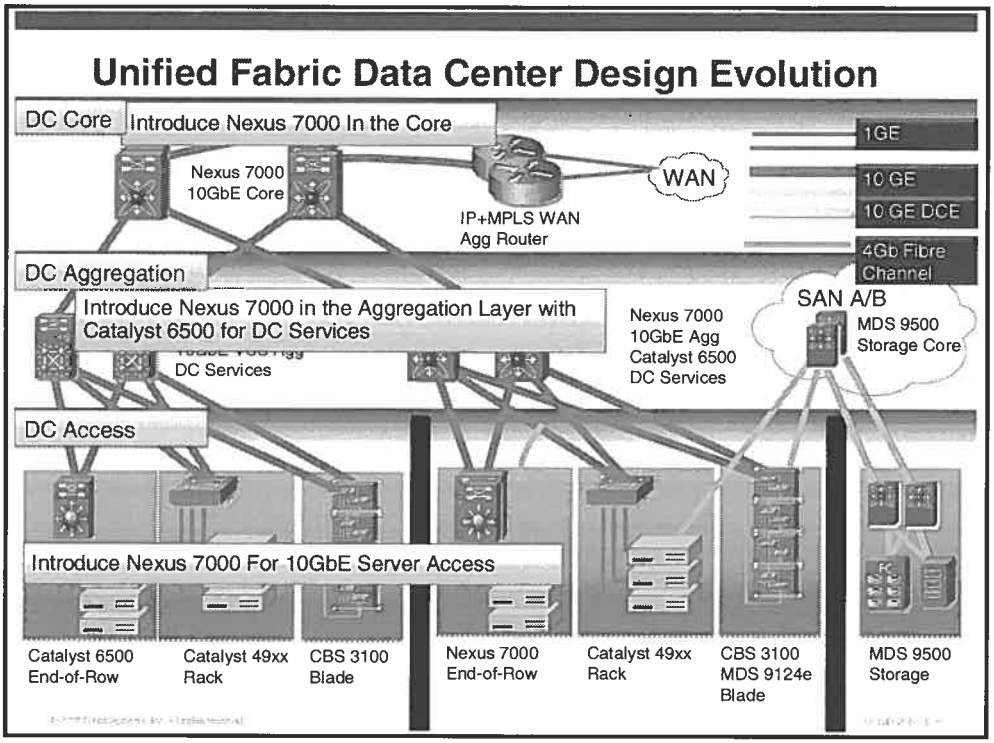
https://t.me/learningnets



The data center module is providing the redundant data center application server access to the storage devices, and client access to the application servers. The data center module offers heterogeneous 1Gb or 10Gb Ethernet or unified server access to the Catalyst 6500 Series switches, Nexus 7000 Series switches, Nexus 5000 Series switches, Catalyst 4900M Series switches or the Cisco Catalyst Blade switches. The data center module design adds elements to the standard design to increase the functionality. Added elements are services modules or device-performing network services, including the server load balancing (SLB) and SSL offload, security services and application acceleration, to enhance the performance, reliability, and availability of the web, application, and database servers. The data center module supports different server application design, such as hierarchical, multi-tier application design, as discussed in the next lesson.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



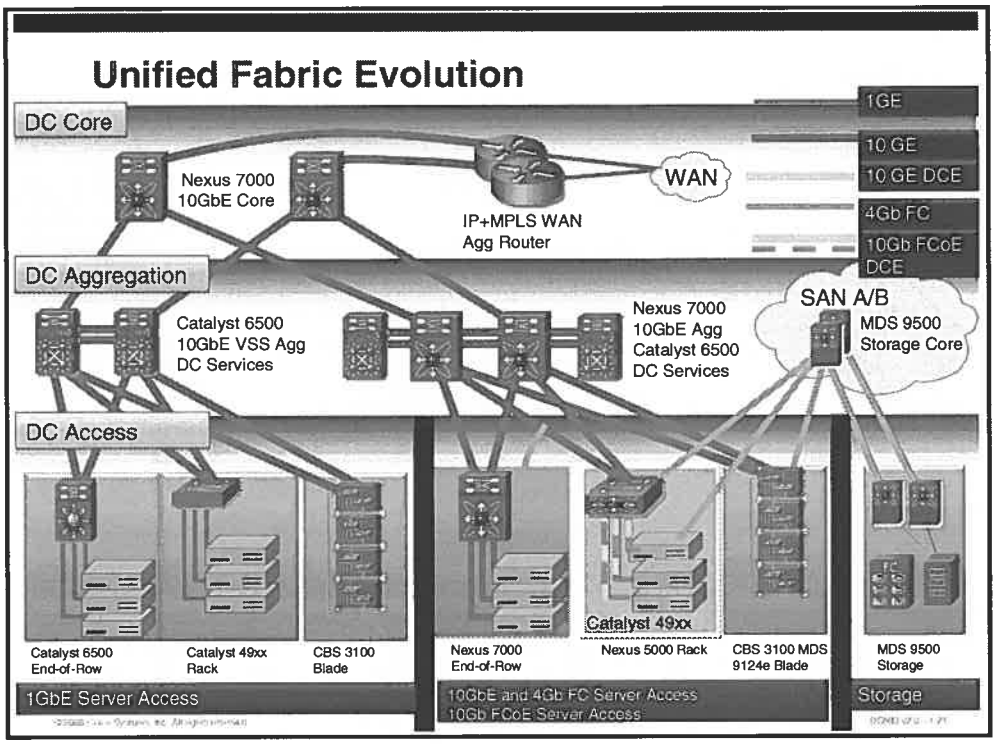
The Cisco Nexus 7000 Series switches can be introduced in the data center aggregation layer where the density requirements are high, when hundreds of 10Gpbs ports are needed and where high performance and throughput is needed and when unified I/O will be sent across the aggregation layer. Catalyst 6500 Series switches can still be deployed in that environment, especially when service modules are needed since Nexus 7000 Series switches do not support them. When both services modules—unified I/O and high performance—are needed, the Nexus 7000 Series switches and Catalyst 6500 Series switches can both co-exist in the aggregation layer as shown in the preceding figure.

The Cisco Nexus 7000 Series switch can also be introduced as part of the high densities 10 Gigabit Ethernet end-of-row access.

FCoE unified I/O host access is provided by the Nexus 5000 Series switch. The unified I/O channel data will split to the MDS SAN environment and to the Nexus 7000 Series switch environment, or to the Catalyst 6500 Series switch environment for LAN access. Nexus 5000 Series switches represent the top-of-rack access for server I/O consolidation with FCoE.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The preceding figure shows the introduction of the Nexus 5000 Series switch for the rack server unified I/O access and for the server I/O consolidation with FCoE. The Nexus 5000 Series switch replaces the Catalyst 4900 family of switches in the access layer in this example, providing the server Ethernet connection, thereby eliminating the need for the separate Fibre Channel connection to the MDS fabric. Currently, the unified FCoE I/O is extended up to the access layer and does not yet reach MDS SAN or the data center aggregation layer.

The MDS SAN will support FCoE I/O modules in order to spread the unified I/O out up to the MDS SAN island. MDS will therefore forward the native Fibre Channel traffic to the storage back end.

This figure represents the Cisco Nexus 7000 DCE I/O module to interconnect the Ethernet server access and the MDS SAN using FCoE unified I/O or to unify the data center server access to the MDS SAN and to the Ethernet fabric. The Nexus 7000 DCE I/O modules are assumed to be present in the data center access layer in order to accomplish this task. The DCE module should be available for Nexus 7000 Series switches, and when the Nexus FCoE blade switches are available, they will enable the FCoE/DCE to be extended up to the MDS SAN core or to the MDS SAN edge layer.

Network-wide unified fabric capabilities with multipathing in the core will be possible when DCE I/O modules with the some software differences from the prior phase to this phase will be introduced later on in the aggregation and in the data center core layer.

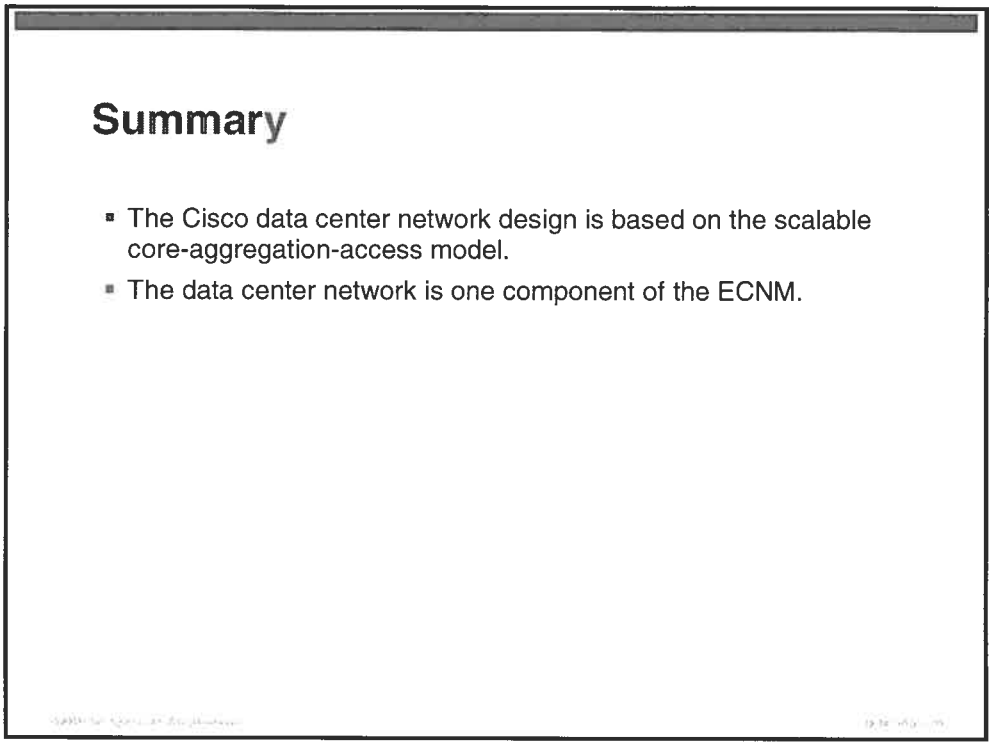
The next phase of data center design will enable the aggregation layer to be populated with the DCE I/O modules. FCoE connections are still supported up to the MDS switch. At the beginning, the FCoE targets will be seen in the data center SAN but the majority of our customers are very likely going to utilize the existing Fibre Channel targets. The MDS switches play an integral role in the overall evolution of the data center unified fabrics since the FCoE will be supported by the Fibre Channel switches. More and more, routers are becoming appliances that support different application services these days. The consolidation and virtualization advantages had been introduced by services modules in the Catalyst 6500 series switch in the past, although introduction of Nexus 7000 Series switches made the idea of external appliances appealing to the customer again. Nexus 7000 Series switches are not deployed to support service modules.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

This topic summarizes the key points that were discussed in this lesson.



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 2

Data Center Application Design

Overview

This lesson discusses the impact of data center application design on the network architecture.

Objectives

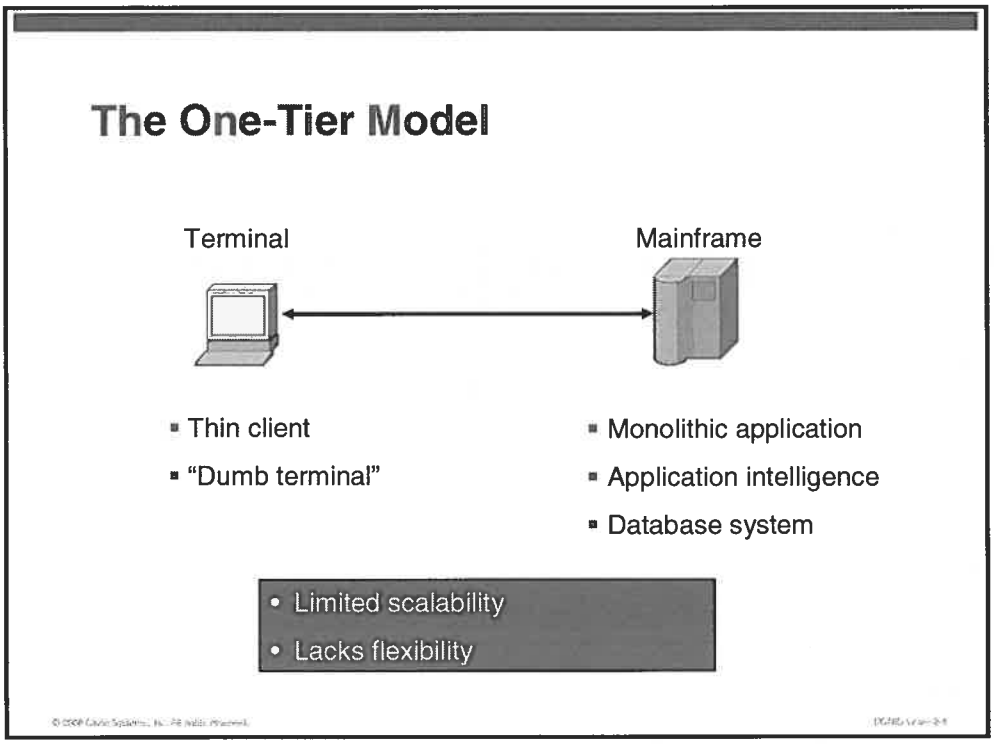
Upon completing this lesson, you will be able to discuss the impact of data center application design on the network architecture. This ability includes being able to meet these objectives:

- Describe the network as it is seen by the application designer
- Explain the impact of server virtualization on the network design
- Explain the impact of the Cisco ACE appliance/module server load balancing (SLB), content offloading, caching and security functions along with the appliance virtualization on the data center design
- Explain the purpose of global server load balancing (GSLB) in a multi-site data center design
- Explain the impact of WAN application optimization and branch consolidation on the network design
- Explain the impact of server clustering on the network design
- Explain the requirements for cluster interconnects
- Explain the concepts of server farms and SLB
- Explain the impact of content caching and content offloading on the network design
- Define SOA and describe its overall impact on the network design
- Describe how to make recommendations for a sample enterprise

https://t.me/learningnets

Application Architecture Tiers

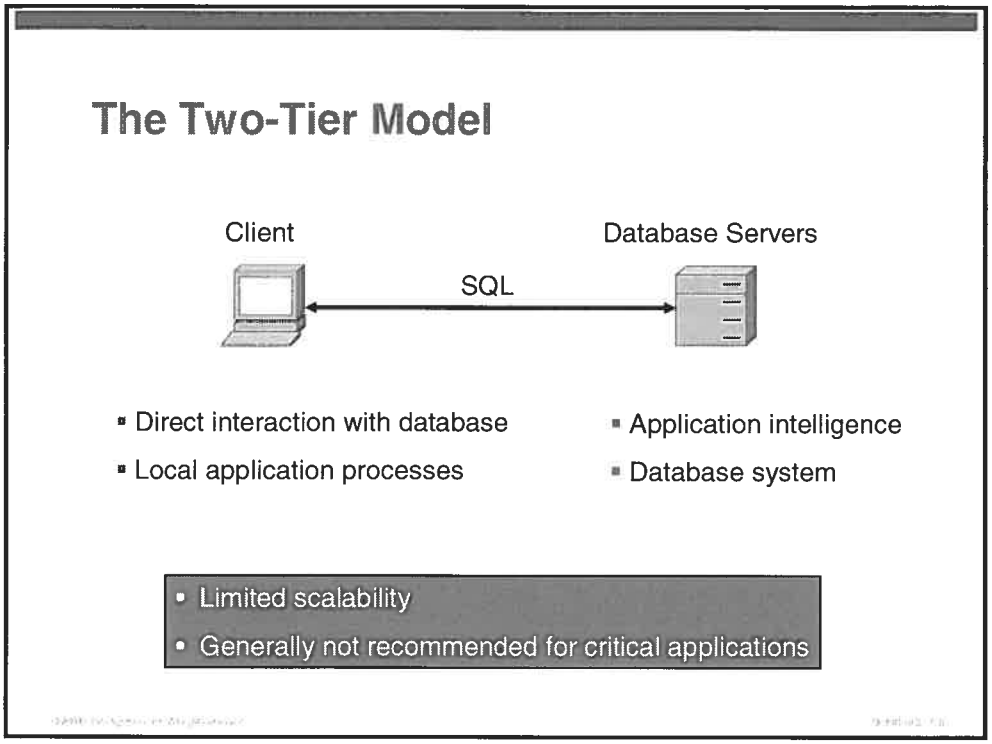
This topic describes and looks at the application architecture tiers as seen by the application designer.



This model is the classic "dumb terminal" situation, where the client has very little intelligence and mostly only screen refreshes like characters, pixels, and so on. The old green screen is a classic example of this model.

© 2008 Cisco Systems, Inc.

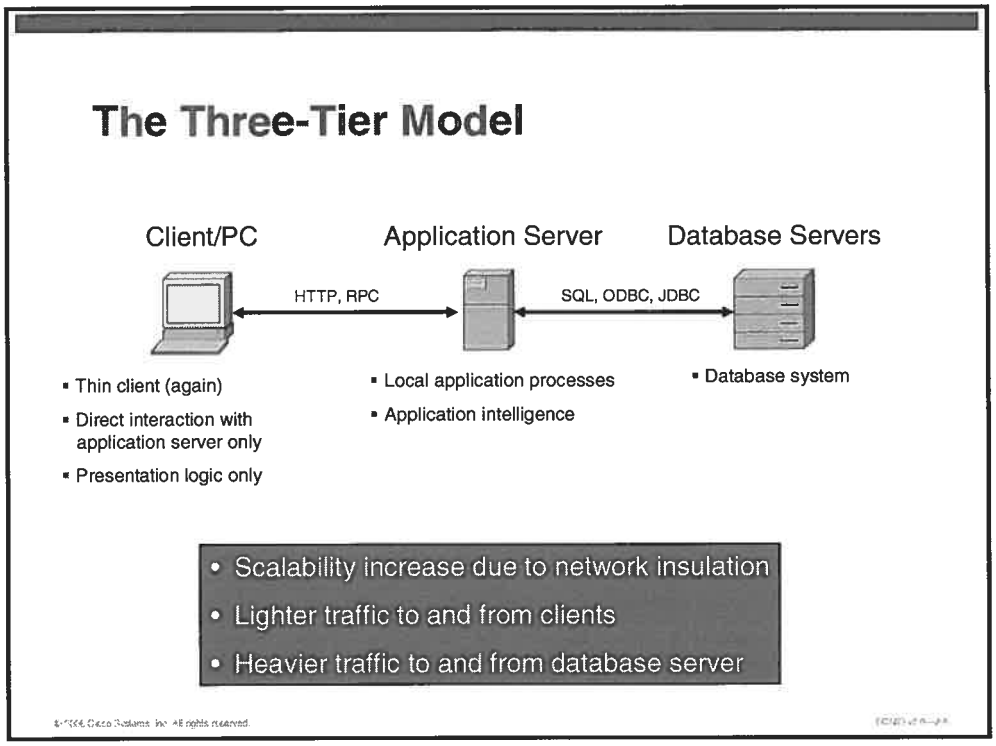
https://t.me/learningnets



This model describes a client-server where the client has some processing power via an application-specific engine that resides on the PC. This model is distinctly different from the previous one-tier model because from the application's perspective, some logic has been distributed among two machines or tiers.

The two-tier model's limitation is scalability. It is limited by the number of connections the database server can handle. Also, there are no built-in limitations or business logic limiting how the client can query the database. A single client could dominate the database server's resources with certain queries. Thus, this model is generally not recommended for critical applications.

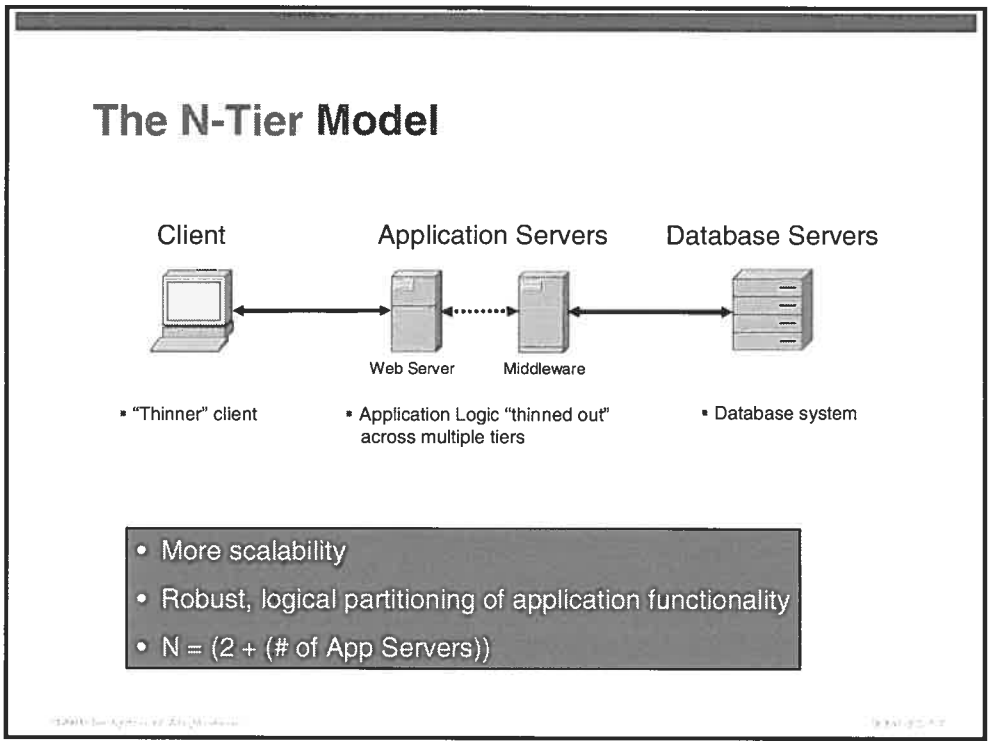
https://t.me/learningnets



This model depicts the separation of the database server on the back end. The three-tier model is a distributed system characterized by clusters of autonomous functionality across multiple tiers. This model is more scalable and more resilient than the two-tier model. Business logic can be applied in the application server tier to limit the extent and scope of queries from the clients. In this way, the application designers can better manage loads on the system.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



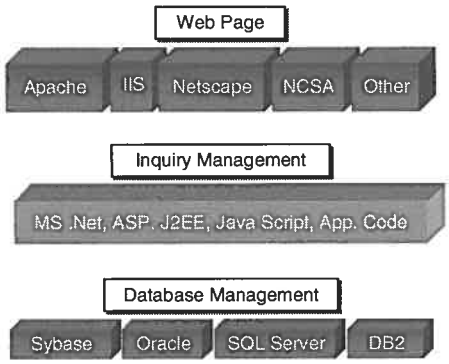
The N-tier model describes the unbound set of potential application tiers. The number of tiers determines how scalable, portable, and manageable an application can be. As applications grow, different services supporting that application can grow independently as needed.

https://t.me/learningnets

The Web Services Model

Tasks performed by multiple hosts, each with a specific role:

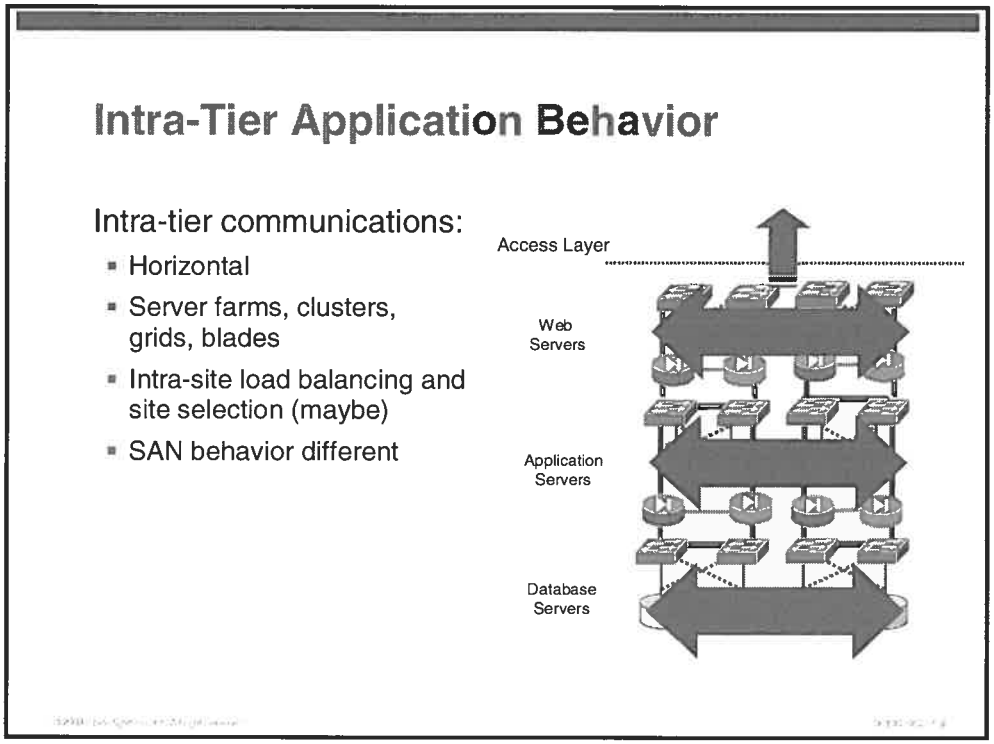
- Web server
  - Serve pages
  - Static vs. dynamic content
- Application server
  - Implements business logic
  - Manipulates data
- Database server
  - Accesses data store
  - Retrieves and writes efficiently and quickly



© 2008 Cisco Systems, Inc. All rights reserved.

The web services model is an example of an N-tier system. Some examples of applications that might use this model include customer relationship management (CRM) or enterprise resource planning (ERP) applications.. Web servers on the front end handle individual user requests. These requests can be load-balanced to the web servers using a network-based solution. The web servers communicate directly with the application tier and the application tier communicates with the database tier on the back end. Business logic can now be applied against the user requests. For instance, important users (large customers, executives, and so on) can be provided with priority service.

https://t.me/learningnets



Single-Tier Application Behavior

Server farms represent single tiers in the N-tier model, if not physically, then at least logically (usually, both). In any case, the types of clustering that occur at each tier must be understood so that the network can provide the proper resources and services to the application, such as common VLANs, PVLANs, access control lists (ACLs), and so on.

Though most network based functionality is currently aimed at the web server environment, many applications currently use a proprietary form of application server clustering technology that precludes Cisco from performing any hardware-based load-balancing or clustering assistance. This is also true in the database tier.

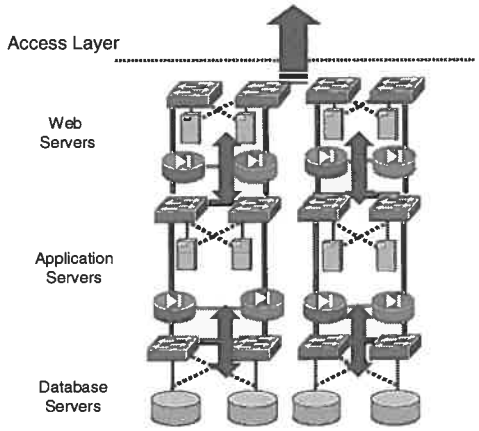
Storage area networks (SANs) are unique and are discussed later.

https://t.me/learningnets

Inter-Tier Application Behavior

Inter-tier communications:

- Vertical
  - Front end to application
  - Application to back end
  - Client to application
  - Client to back end
- Security
- Protocol transparency



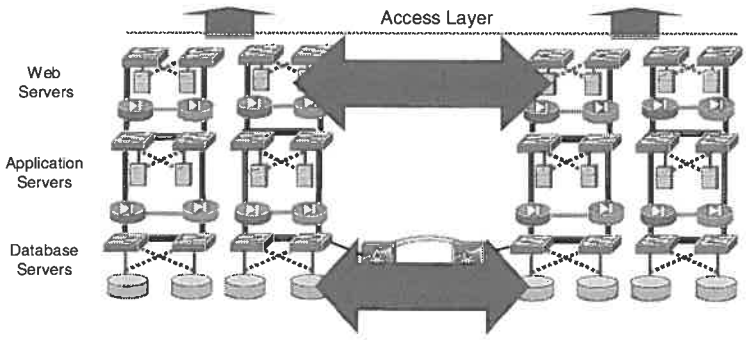
© 2006 Cisco Systems, Inc. All rights reserved.

Clients communicate directly with web servers, and sometimes with application and database servers too. Web servers communicate with application servers. Application servers communicate with database servers. Database servers communicate with and across SANs.

This vertical communication mandates that the data center network must also be designed to accommodate inter-tier communications.

https://t.me/learningnets

Multi-Site Application Behavior

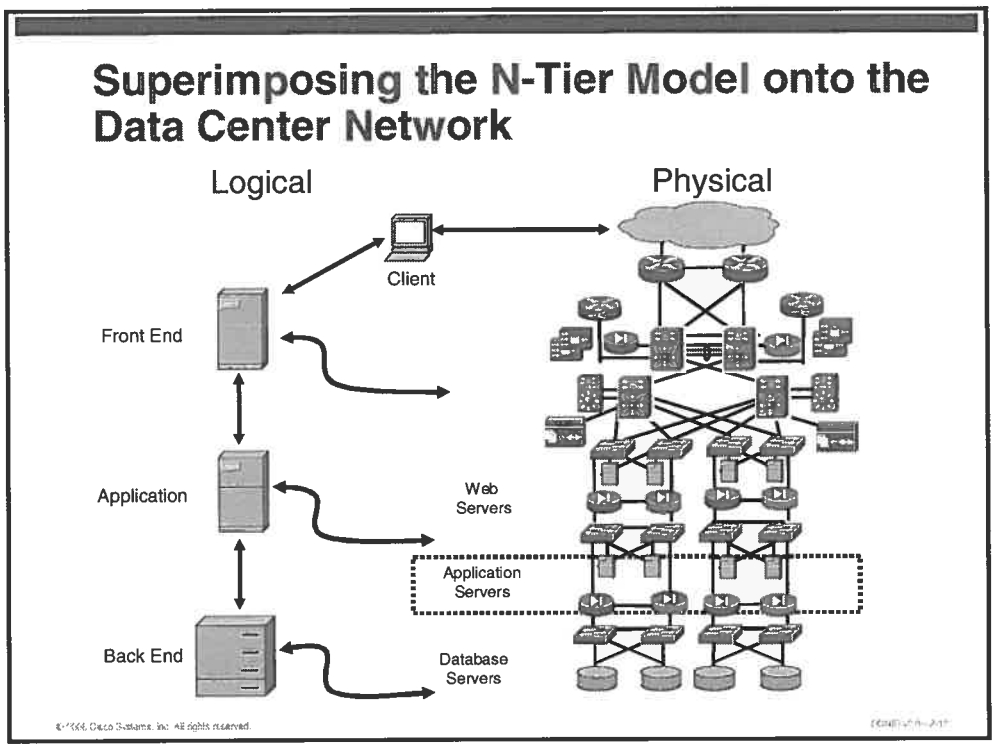


Inter-site requirements:

- Site selection and load balancing
- DR/BC (Disaster Recovery/Business Continuance)
- Synchronous/asynchronous transactions

Distributed data centers offer globally distributed application type services, which implies that the communications required between these sites must be understood. Communication includes the server cluster hearbeats, server storage synchronous/asynchronous storage replication, and backup and failover functions between the primary and the secondary data center site.

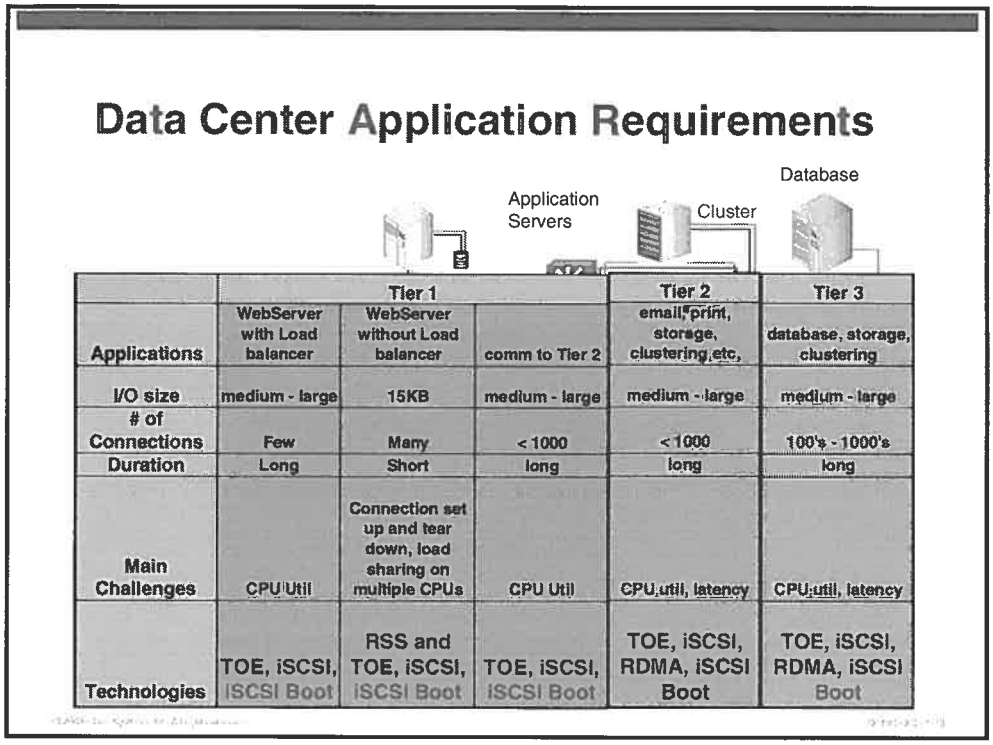
https://t.me/learningnets



Data centers are being built a certain way as a direct result of the types of services that these N-tier applications are demanding.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Servers and storage devices impose certain network requirements on the data center network such as effficient data, storage transport, efficient clustering, and support for data center management functions.

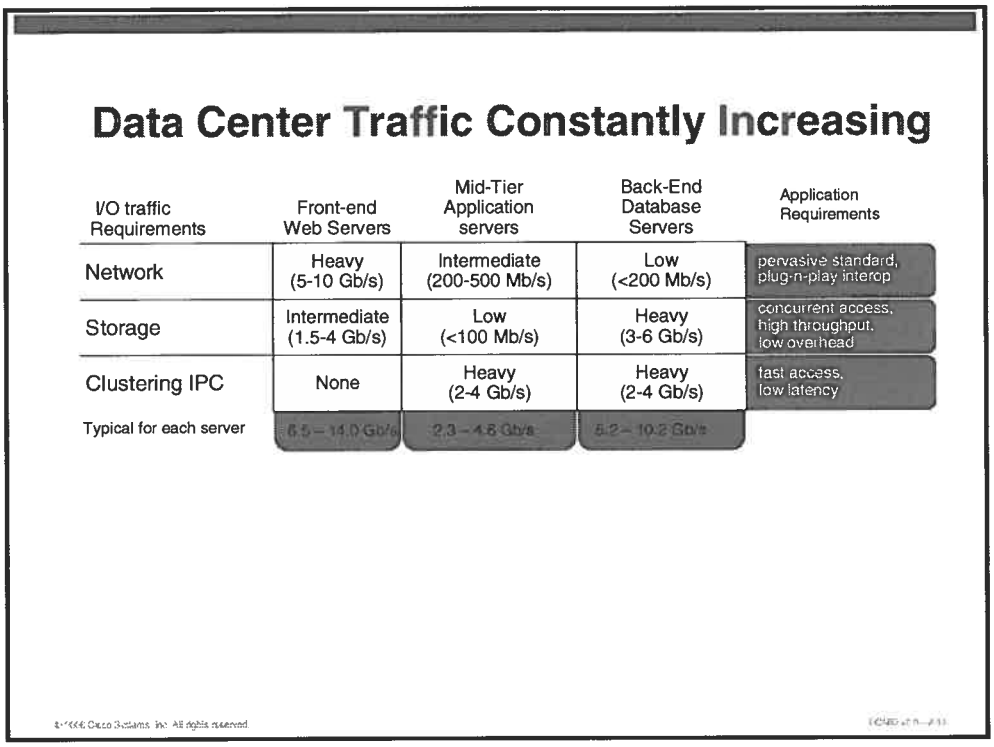
Server applications require acceleration of the data and storage transport such as TCP accelearation, iSCSI, and clustering accelerartion using RDMA.

Application requirements include more transactions per server, higher transactoin rates and suport for larger messages (for example, e-mail) today.

The preceding table reports the data center application characteristics of servers in each data center application tier, such as:

- The technologies that are used
- The application I/O profile
- The number of connections that are established
- The duration of connections
- The CPU and latency isuess for each tier

https://t.me/learningnets



Data center traffic of each application tier is increasing significantly every year, and scaling process of the 3-tier fabric infrastructure is expensive and cumbersome.

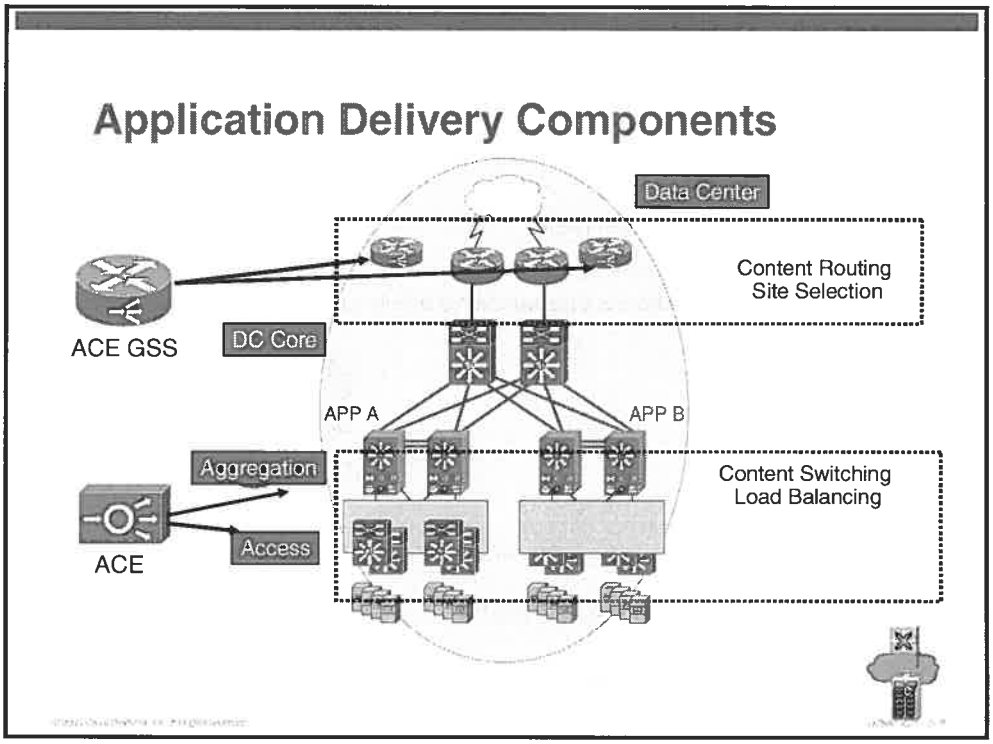
Server density complicates connections to three fabrics. Data center server application requirements triggered the need for unifying data center fabrics. Unified fabrics are the right solution that meets new data center application requirements.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Application Optimization Delivery Products

This topic identifies the Cisco products providing application delivery optimization for the data center design.



The figure illustrates the application delivery components in the data center network. At the content routing layer, site selection is provided by global site selection (GSS). At the content switching layer, load balancing is provided by the ACE module/appliance. The ACE appliance is deployed in the data center access layer as well.

Note Without GSS, use DNS round robin or IP anycast for a non-intelligent device solution.

https://t.me/learningnets

Content Routing

- Provides global redundancy among the redundant data center sites
- Used for site selection in global server load balancing (GSLB)
- Different site selection technologies:
  - HTTP redirect
  - Route health injection (RHI)
  - DNS-based solution
- DNS-based solution = Global Site Selector (GSS)



ACE GSS

© 2008 Cisco Systems, Inc. All rights reserved.

The figure introduces the content routing role and functions in the data center.

https://t.me/learningnets

Cisco ACE Content Switching

- Also referred to as Server Load Balancing (SLB) the group of Servers or Cache (WAE) farms
- TCP connection must be served by the same server unless TCP is split across members of server cluster
- Session persistence ensures that many TCP/UDP connections are served by the same server
- Provides local application/server access redundancy within the single data center
- Can be deployed in conjunction with GSS to provide:
  - Load sharing of client requests among servers between data centers
  - Disaster recovery
  - Redundancy between data centers

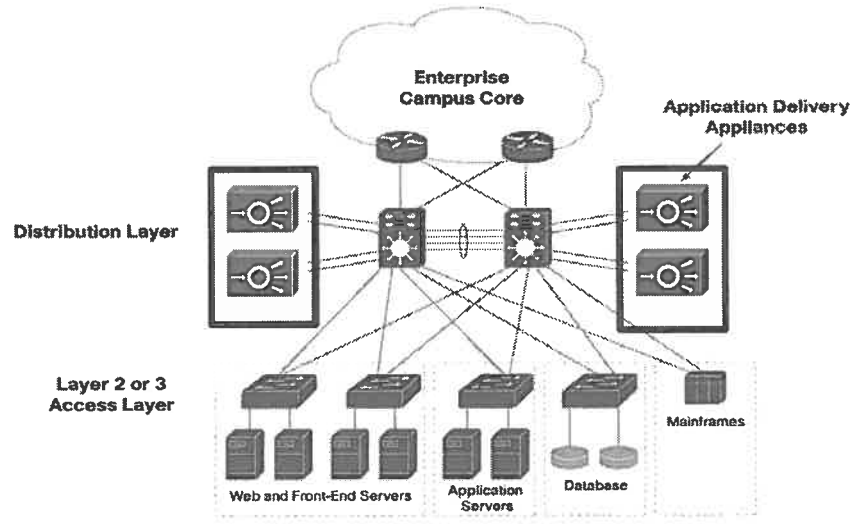


ACE

The figure introduces the content switching role and functions in the data center.

https://t.me/learningnets

Cisco ACE Deployment—Data Center Aggregation Layer

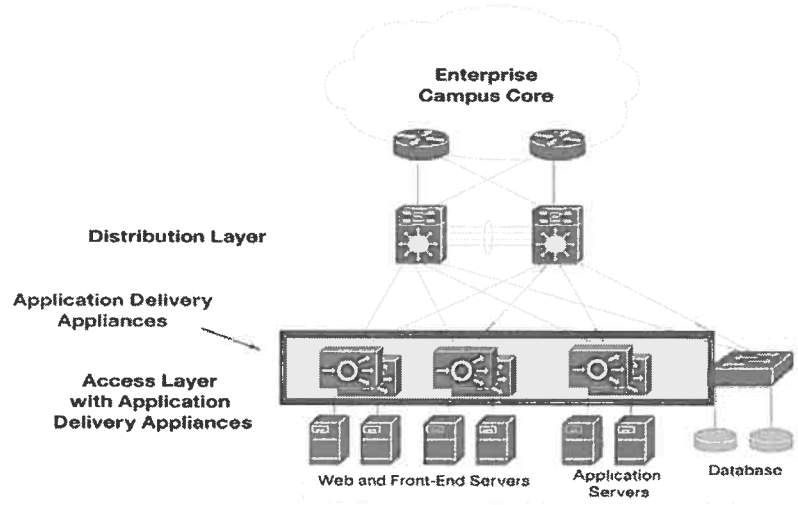


ACE server load balancing can be deployed in the data center distribution layer using the Cisco Catalyst 6500 Series Switches and ACE services modules or using an external ACE appliance. The ACE appliance can be connected to the Nexus 7000 Series Switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco ACE Deployment—Data Center Access Layer

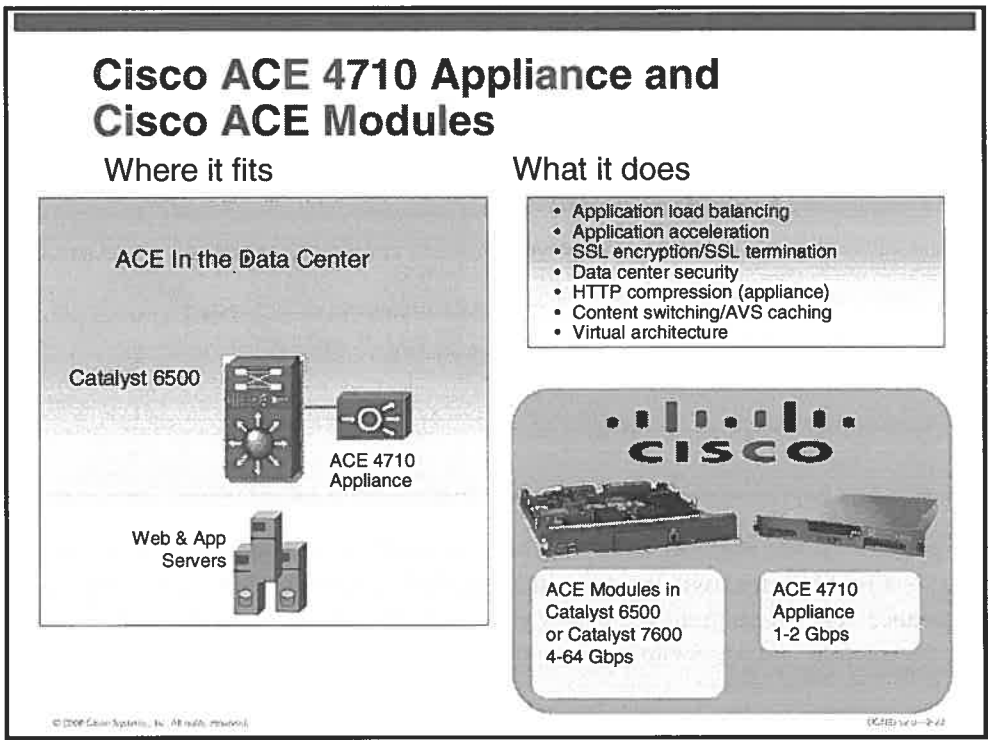


Cisco ACE server load balancing can be deployed in the data center access layer using the Catalyst 6500 Series Switches with integrated ACE services modules or using an external ACE appliance. The ACE appliance can be connected to the Cisco Nexus 7000 Series Switches or to the Nexus 5000 Series Switches.

https://t.me/learningnets

ACE Appliance/Module Impact on Data Center Design

This topic discusses the impact of the ACE appliance and ACE module functions on the overall data center design.



The Cisco ACE family includes highly scalable application switching modules for the Catalyst 6500 Series switches and Cisco 7600 Series Routers. The Cisco ACE Service modules consolidate a broad range of services in one device.

ACE is an asymmetrical solution, which means that it is fully transparent from the perspective of the local and remote client systems as well as the server and applications. Cisco states that as a standards-based technology, ACE can be integrated with standards-compliant solutions from other vendors for functions such as firewalling and Secure Socket Layer (SSL) processing offloading.

Cisco ACE Application Control Engine Application switch technology is now also available in a convenient appliance form factor for those deployments requiring a standalone platform from a Catalyst 6500 Series switch. The Cisco ACE 4710 Appliance uses the full range of Cisco application delivery technology, including Layers 4 and 7 content switching, extensive application security, and server offload for SSL processing and TCP processing, as well as an expanded set of application acceleration capabilities. Used to increase the performance of data center-based web applications, the ACE 4710 Appliance uses asymmetric compression, FlashForward, and other patented techniques to reduce the volume of transferred data and improve application response times to end users, thus providing a better overall end-user experience.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Other ACE appliance/module functions are:

- Reflexive access lists, TCP and IP normalization, protocol fix-ups
- HTTP inspection, RFC compliance, control on headers and payload
- Modular Policy CLI for integrated features configuration
- Configuration rollback for quick error recovery and testing
- Integrated SSL termination, with support for back-end SSL
- TCP reuse for HTTP (same as TCP multiplexing or TCP offload)
- CLI objects name completion
- HTTP compression (appliance only)
- Web application and acceleration (appliance only)

The ACE family also includes the Cisco Application Control Engine (ACE) Extensible Markup Language (XML) Gateway. This appliance extends the core ACE function set to the Extensible Markup Language (XML) and web services, including security, scalability, and acceleration. To help ensure data center availability, the ACE application delivery solution is integrated with the Cisco market-leading data center availability system, Cisco Application Control Engine (ACE) Global Site Selector (GSS).

ACE GSS provides intelligent failover between data centers to help ensure business continuity and provide full Domain Name System (DNS) and Dynamic Host Configuration Protocol (DHCP) support, as well as protection from the DNS-based distributed denial-of-service (DDoS) attacks.

To meet business requirements, application service must deliver scalability, high availability, and optimal performance to all users across the extended enterprise. By delivering 16 Gbps of raw throughput and 4 million bidirectional connections, the ACE solution can handle the largest production environments and be partitioned to be shared across different constituents. High availability is provided in a variety of active-active configurations, including across geographically dispersed data centers.

ACE delivers high application performance including high throughput in Layers 2 through 7, highly scalable SSL acceleration and offload, TCP reuse or connection optimization, latency mitigation, and compression, thereby reducing the burden on servers and significantly enhancing the delivery of applications. In addition, this solution provides graphical views of application performance metrics including end-user response times, helping to quickly identify and troubleshoot application bottlenecks. The result is more efficient use of network and server resources, which lowers costs for the entire application infrastructure as well as deployment and support costs for every application.

ACE Intelligent Proxy Architecture

By default, many application-delivery products perform full TCP and application proxy even for basic SLB. As a result, the performance of these products is significantly lower when basic features such as source IP persistence and dynamic load-balancing algorithms (such as the "least-connections" algorithm) are enabled. However, the ACE Application switch is designed to intelligently perform full proxy only when needed in order to maximize performance and scalability. In particular, the ACE can automatically select one of the two following mechanisms to handle connections that are based on load-balancing policies:

https://t.me/learningnets

- **Layer 4 SLB**:

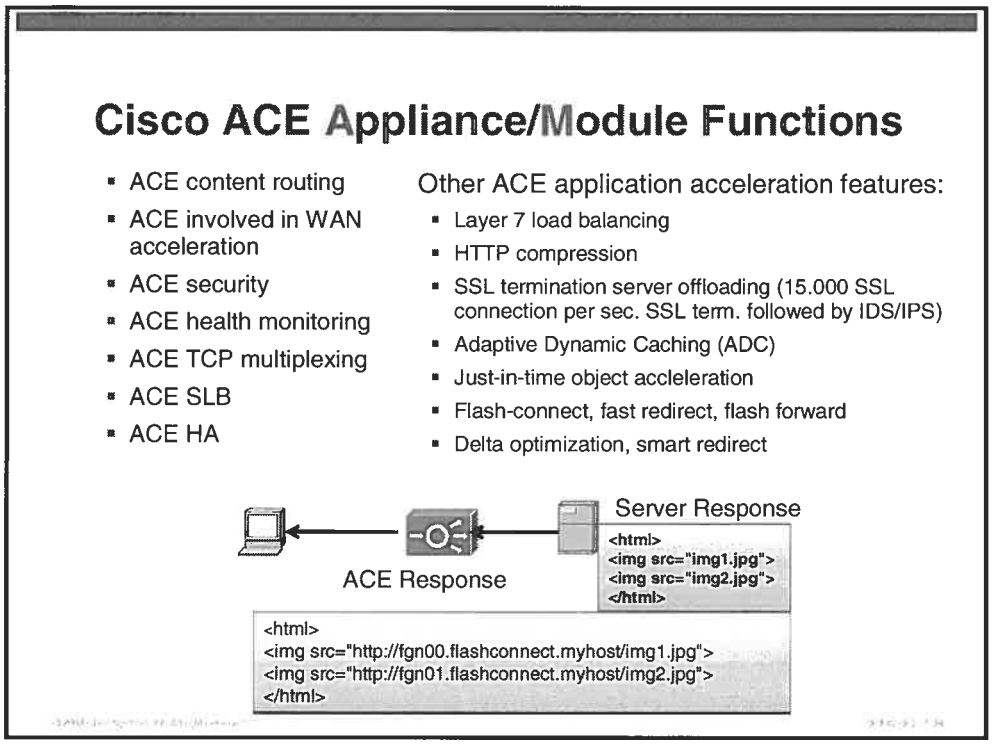
  In many customer deployments today, many server load-balancing decisions still rely on basic information about the client and services that they request. Usually these load-balancing decisions are made at the IP and TCP layers (IP or MAC address, TCP, or User Datagram Protocol port numbers) that do not require a SLB device to proxy TCP connections. Hence, the ACE does not proxy the TCP connections for SLB policies based on Layers 2 through 4 information; in fact, proxying connections, by default, provide no added benefit. Server load balancers that do proxy merely add complexity and exhibit poor performance and scalability.

- **Delayed binding for Layer 7 SLB**:

  As applications and offered services become more complex, there is an increasing need to provide load-balancing decisions at layers above the transport layer (TCP layer)—the point at which Layer 7 load balancing applies. Layer 7 load balancing requires proxying of TCP connections to inspect packet payloads and identify headers and fields to enable intelligent load balancing of user requests. The ACE Application switch is the only product that can proxy, unproxy, or reproxy TCP connections as needed based on load-balancing policies and features that you configure. For example, some features such as HTTP URL-based load balancing or cookie-based session persistence require ACE to parse HTTP headers or other Layers 5 through 7 data and terminate TCP connections, as any proxy would do. The ACE "unproxies" TCP connections after it analyzes the data that is required to make the appropriate load-balancing decision. This approach, also called "delayed binding," accelerates processing for subsequent data packets, hence improving performance for your system. At the same time, the ACE can "reproxy" connections whenever a feature requires it. For example, HTTP 1.1 connection persistence requires parsing additional client requests sent at a later time over the same TCP connection. The ACE can switch the connection from unproxy to reproxy mode to look for this additional information. The ability of ACE to switch individual connections between proxy and unproxy mode offers a significant performance advantage in real-world deployments.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



ACE SSL Termination and Initiation

ACE is designed specifically for high performance SSL and performs this function in hardware, providing up to 15,000 SSL connections per second. By terminating SSL on the ACE module, the traffic can be passed in the clear so that security mechanisms such as intrusion detection and intrusion prevention and web application security can be employed before delivery to the server. After these services have been applied, ACE can re-encrypt the traffic and forward it to the server. Alternatively, the SSL termination process can be entirely offloaded from the server for a significant performance boost of up to 10 times in some cases. When SSL session reuse is available on ACE, there will be a middle ground where traffic is still encrypted all the way to the server but at a lower processing cost.

ACE Content Routing

ACE can look deep into the HTTP header to make forwarding decisions that are based on other variables besides server load and availability. By directing requests to servers tuned for certain types of content, web browser or language support, the server environment can be further optimized.

ACE WAN Acceleration

In high latency WAN environments, response time for web transactions can be unacceptably slow. ACE provides a scalable way to transparently move server-destined transactions to acceleration devices such as Cisco Wide Area Application Services (WAAS) to optimize the session before forwarding traffic to the real servers. This ability results in better session response time for clients while providing high availability and scalability for the optimization engines themselves. ACE replaces the WCCPv2, policy-based routing (PBR) or inline methods for optimized data redirection to the Cisco Wide-Area Application Engine (WAE).

ACE End-to-end Health Monitoring

Through the use of advanced probes, ACE can provide health checks that extend beyond the web/application servers themselves to also calculate the availability of all critical path

https://t.me/learningnets

resources before sending requests to a server. For example, if the database is down, ACE can forward the request to a backup server farm or sorry server rather than sending it to a web server.

ACE TCP Reuse/Multiplexing

Instead of setting up a new TCP connection for every flow destined to a server, Cisco ACE can multiplex multiple flows across a single connection. This ability reduces the processing load on the server, allowing it to support more simultaneous connections.

ACE High Availability

Because Cisco ACE maintains the state of each connection and sticky entry in the standby context, failover is seamless and connections remain intact. This means that service is not impacted during maintenance or failures in the network.

ACE SLB Consolidation

While a pair of web dispatchers must be dedicated to each SAP application, all of these functions can be combined onto a single pair of ACE modules or contexts.

ACE Appliance/Module Functions

- **SSL termination and SSL processing server offloading**
- **ACE Smart Redirect**: Some applications automatically redirect client browsers from one page to another using HTML meta tags. HTML meta tags page redirection can cause poor download times because the browser issues IMS for each embedded object on the redirected page. Smart redirect enables the ACE automatically and transparently to convert HTML meta tag redirections into more efficient HTTP header-based redirections.

  Features of smart redirect:

  — Convert meta refresh to HTTP header redirect

  — Parses for the meta refresh tag

  — Identify by IMS requests for FlashForward objects

- **ACE HTTP compression**: HTTP Compression algorithms include deflate and zip. A default compression method is used.

  ACE appliance/module compression client request processing:

  - Parse for the following headers:
    - Accept-Encoding (includes deflate and gzip) from the client
    - User-Agent (browser decompression and compression software) must be supported
  - Determine compression support:
    - Compression method listed in the Accept-Encoding client message that is supported by the client browser is supported
  - Modify compression method:
    - Remove the Accept -Encoding: xxx header
    - Insert Accept- Encoding: Identity header

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

ACE appliance/module compression response processing:

- Parse for the following headers:
  - Content-Length (length of returned content)
  - Content-Type (type of resource)
- Determine compression support:
  - Length over minimum
  - MIME type allowed
  - No Cache-Control: No-Transform header
- Compress response:
  - Remove Content-Length header
  - Add Transfer-Encoding Chunked header

ACE appliance compression configuration example

```
parameter-map type http COMPRESSION_PARMS
  compress mimetype text/*
  compress minimum-size 512
policy-map type loadbalance first-match slb-logic
  class class-default
  serverfarm ext-servers
  compress default-method gzip
class-map match-all external-vip
  2 match virtual-address 198.133.219.25 tcp eq www
policy-map multi-match client-vips
  class external-vip
    loadbalance vip inservice
    loadbalance policy slb-logic
    appl-parameter http advanced-options COMPRESSION_PARMS
```

- **ACE Just-in-time object acceleration**: Just-in-time object acceleration is used to accelerate non-cacheable embedded objects to accelerate application response time. Just-in-time object acceleration applies to the following conditions:
  - Dynamic HTML content that is larger than the maximum condensable page size (250 KB)
  - Content that is marked by the origin server as expired or not cacheable

  Just-in-time object acceleration functions as follows:
  - The browser requests an object, and the ACE fetches it on behalf of the browser request.
  - The ACE constructs and inserts an ETag (entity tag) header by using an MD5 hash of the content, and then it sends the object to the browser. The ETag is a request header that you can use to identify different versions of the same object.
  - All subsequent requests from the browser include this ETag header, which the ACE compares with a recomputed MD5 hash of the latest origin server content as follows:
    - If the content has not changed, then the ACE returns a 304 (Not Modified) response and no data
    - If the content has changed, the ACE retrieves the new content and resets the ETag.

https://t.me/learningnets

It is recommended to use compression along with the Just-in-time object acceleration. Dynamic ETag is MD5 hash of object. Just-in-time object acceleartion converts 200 response to 304 response (when possible) and is used on objects or pages. Cache headers might need to be explicitly modified with Just-in-time object acceleartion but does not work with compression.

- **ACE Flash-connect**: FlashConnect dynamically renames embedded objects by adding a prefix and changing the hostname, making the objects appear to reside on different hosts although they might all reside on a single host. FlashConnect makes the browser open separate connections to the origin server for each object, which increases the network performance because the objects are retrieved in parallel, rather than serially.

  FlashConnect:

  — Makes the browser believe that it is communicating with multiple web servers

  — Rewrites object references in pages

  — Renames objects (as a transparent proxy)

  — Does **not** work with SSL

- **ACE Fast Redirect:** Fast redirect intercepts HTTP 302 responses from the origin server and makes a second request on behalf of the client for the redirect URL, fetches it, and sends it to the client. This applies only to redirects within the same domain.

  Steps of a Fast Redirect:

  Step 1 Client requests URL.

  Step 2 Server responds with HTTP 302 Redirect to a server in the same domain as the origin server.

  Step 3 ACE fast redirects a request to the server in the origin server's response.

  Step 4 The redirection server responds to the ACE.

  Step 5 The ACE forwards the content from the redirection server seamlessly to the client.

  Fast Redirect characteristics:

  — Prevents 302 from being sent to client

  — Retrieves redirect URL

  — Combines all headers that are sent on 302 response with 200 response

  — Identifies by 302 in FgnStatLog or client trace tool

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **ACE Flash Forward:** FlashForward is applied to static web objects. The ACE renames and caches static objects internally and alters the embedded objects HTTP headers to extend the length of time they are valid in the client browser cache. It also modifies the URL references in the parent HTML code sent to the client browser to refer to the renamed objects so that the client requests only objects known to be new or modified at the time of the HTML request. FlashForward technology results in significantly accelerated page downloads by the client and reduced upstream traffic that is associated with object validation requests.

  FlashForward works as follows (after web objects have been cached by the ACE, which is triggered by a client accessing the URL or a similar URL):

  — On the first visit:

  - The client requests the URL.
  - The ACE proxies the request to the origin server.
  - The server responds to the request normally and the ACE intercepts the response.
  - The ACE responds with the HTML container page, which has been modified to reference the new objects' references. If delta optimization is also configured, the HTML container page will be a dynamic HTML page with an embedded JavaScript:
    - Embedded objects that are referenced in HTML container pages are:
      - Served with expiry date in the future:
        - The default is over 20 years.
      - The object name is appended with an MD5 hash based on the object as well as a version number:
        - For example, `header.gif` could become `header_vtmmi14xg2fvmlkxsxuk0ty1xd_AVS_V01.gif`.
  - The client receives the response and parses the HTML container page for objects to request.
  - The client requests the objects from the server.
  - The ACE intercepts the request and forwards the transformed request objects (minus FlashForward alterations) from the server (according to the refresh rate configuration).
  - Assuming that the object in the ACE cache is fresh, the ACE responds to the client with the requested object.

  — On subsequent visits:

  - The client requests the URL.
  - The ACE proxies the request origin server.
  - The server creates and delivers the HTML page to the ACE.

https://t.me/learningnets

- The ACE parses the HTML for all the references to embedded objects and checks if the objects are cached locally on the ACE. If the object is cached locally, the ACE issues an HTTP IMS request to the server:
  - If the server responds with an HTTP 304 "Not Modified" message, the ACE uses its previously created object reference in the HTML container document.
  - If the server responds with an HTTP 200 "OK" message (and the modified object), the ACE caches the object, renames it, changes the expiry date, and adds the new object name to the HTML container document.
- The ACE sends the modified server HTML container document to the client (or an update script if delta optimization is used).
- The client parses the HTML container document for objects:
  - Any previously cached FlashForward objects that are referenced will not have an HTTP GET issued because they have a long expiration date.
  - Any new FlashForward objects that are referenced will be fetched from the ACE.
- Embedded objects that are referenced in HTML container pages are served with expiration date in the future.
- On second visit, the browser does not send a GET request for objects in cache if the current date and time is not greater than the object expiration date, which reduces the total number of HTTP requests for subsequent visits to the same page.
- Benefits:
  - Decreased page download time
  - Decreased network congestion
  - Decreased number of requests to origin server

— An embedded object can be FlashForwarded if the following conditions are met:

- Container page referencing the embedded object must match an application class that includes OptimizationPolicy FlashForward.
- Reference to an embedded object must match an application class with OptimizationPolicy FlashForwardObject.
- Embedded object must not be referenced in a Java script.
- Embedded object must have been previously requested and currently present in the AVS cache.
- Embedded object must not have been explicitly marked noncacheable by the original server. These directives allow control over caching headers sent by client or server: RequestCachePolicy, ResponseCachePolicy, RequestHeader, and ResponseHeader.
- Reference to an embedded object must not change with each request of the container page.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **ACE Smart Redirect:** Some applications automatically redirect client browsers from one page to another using HTML meta tags. HTML meta tags page redirection can cause poor download times because the browser issues IMS for each embedded object on the redirected page. Smart redirect enables the ACE automatically and transparently to convert HTML meta tag redirections into more efficient HTTP header-based redirections.

  Features of smart redirect:

  — Convert meta refresh to HTTP header redirect

  — Parses for the meta refresh tag

  — Identify by IMS requests for FlashForward objects

- **ACE Delta optimization:** The Delta optimization methodology includes the following steps:

  — The client request is proxied by the ACE to the server and the response is intercepted by the ACE.

  — The ACE responds with the dynamic HTML container page in a JavaScript (in this example the original file is 70 KB).

  — If the client requests the page again, the request is again proxied to the server by the ACE, but this time when the ACE receives the response it compares it with the original response and builds a delta of the two.

  — The ACE creates a JavaScript from the delta and sends that to the client to execute and reassemble the HTML container document (in this example, the delta file is only 2 KB).

  Delta optimization is applied to dynamic HTML (uncacheable) documents. The delta optimization feature allows the ACE to cache dynamically created content on the client browser and then on subsequent visits to send the differences or deltas using dynamic HTML:

  — Dynamically generated pages are not cacheable.

  — Delta optimization calculates and sends the difference between two visits to a dynamic HTML page.

  — Delta optimization creates a JavaScript base file that contains the entire first visit to the page.

  — Delta optimization uses a JavaScript on the delta page to reassemble the entire page.

  — Benefits of delta optimization include:

    - Reduced bandwidth usage
    - Reduced page download times
    - Compatible with other optimizations

  — Browsers that are supported must be configurable.

  — Browsers must support cookies, JavaScript, and caching (automatically checked).

  — HTML must use only single-byte character sets.

https://t.me/learningnets

- Certain HTML elements are not condensable:
  - IFRAME
  - SCRIPT
  - OBJECT
- Two modes of Delta optimization exist:
  - DeltaOptimize
  - PerUserDeltaOptimize
- Delta is generated against a base file, which is shared among all users:
  - Can contain information that all users should not see
  - Use PerUserDeltaOptimize to create individual base files for each user
- Cacheable for 30 days
- Compressible with certain browser types
- Delat optimization is canonical:
  - Base files are created based on the URL without query parameters:
    - The same URL can generate different pages based on query parameters or cookies.
    - Several different URLs might generate pages that are similar enough to warrant a single base file. Use CanonicalUrl to create base files based on URL query parameters:
      - Reduces the overall size of the deltas and increases performance

Delta optimization is applied dynamically to pages. Not all content is delta encoded. The following conditions must be met for the page to be successfully delta optimized:

- Client's browser must match those listed in useragent.conf.
- Client's browser must support JavaScript.
- Client's browser must send the FGNCDN cookie set to the JavaScript.
- Delta optimize must be turned on at the global level in fgn.conf with the **DeltaOptimize On** setting.
- URL must match an application class that has OptimizationPolicy that includes DeltaOptimize.
- Page must be larger than 1024 bytes and be adjustable using MinCondensablePage.
- Page must be smaller than 250,000 bytes and be adjustable using MaxCondensablePage.

The following conditions must also be met for the page to be successfully delta optimized:

- Page must not return a ContentType listed in mimetypes.conf with the directive NoDeltaOptimizeMimeType.
- Page should be dynamic html and is not cacheable. Set **DeltaOptimizeCacheableContent On** to allow delta optimization of static content.
- Page must use single-byte character set (iso-8859-1).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- Page should contain less than five UTF-8 characters. This is configurable using **UTF8Detection** and **UTF8Threshold**.
- Size of the delta must not exceed 50 percent of the base file size. This is tunable by adjusting RebaseDeltaPercent.
- Appscope must not have selected the request for pass-through measurement.

- **ACE Adaptive Dynamic Caching (ADC):** Adaptive dynamic caching (ADC) allows the ACE to answer requests for dynamic or personalized information, reducing the burden on servers and databases. ADC enables the ACE to cache dynamic content such as active server pages (ASP) scripts. ADC features are:
  - **Cache parameterization:** A response can be differentiated by more than the URL and its query parameters. Cookie values, HTTP header values, and the HTTP method that is used also can be used as parameters. Cache parameterization can be applied to both static and dynamic caching.
  - **Expanded Expiration Rules:** Time-to-Live (TTL) of cached content can be based on server load—for example, increasing the TTL when the server load increases.
  - **Delta Cache:** Stores the delta content in a cache (a memory-only object) when the original HTML content is in the dynamic cache and a rebase has not occurred.

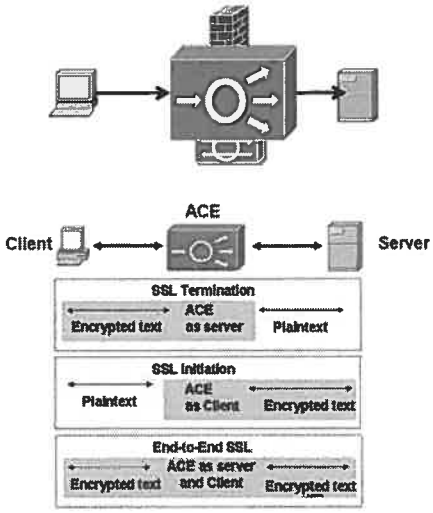
  ADC can be configured for use with multiple or single users. Cache dynamically generates noncacheable pages, for which TTL can be based on server load. It works in conjunction with other optimizations and identifies pages with larger server latency. ACE must, as part of the ADC process, decide which parameters uniquely identify the page and define it using the respective Cache Parameter. It must also decide valid shelf life.

https://t.me/learningnets

Cisco ACE Security Features

ACE Security:

- SSL initiation/termination/server offloading
- Stateful firewall ACL
- L3 ACL filtering
- L7 DPI
- L4–L7 packet inspection
- Protocol complaince inspection
- Application Inspection
- NAT/PAT
- Blocks anomalous signatures traffic to protect against DoS, identity theft, data thef application disruption, and fraud
- Secures XML applications, Web-services, and intra-application communications



ACE is designed to serve as a last line of defense for servers and applications in data centers. The Cisco ACE Family of application switches performs deep packet inspection and blocks malicious attacks. The ACE family of application switches provides both data center firewall and application-layer firewall. The data center firewall protects against protocol and denial-of-service (DoS) attacks and encrypts mission-critical content. The Cisco ACE application-layer firewall protects against both known and unknown threats and against "Day Zero" attacks. An integrated firewall enables IT professionals to comprehensively secure high-value applications in the data center. ACE secures mission-critical applications; protects against identity theft, data theft, application disruption, and fraud; and defends applications and transactions from targeted attacks by professional hackers.

The Cisco ACE Web Application Firewall feature performs both Layer 3 access-control filtering and Layer 7 deep packet inspection (DPI) to identify anomalous signatures that could cause DoS. The ACE Web Application Firewall feature prevents day-zero attacks by identifying and blocking suspicious traffic for which there is no known malicious signature already identified and stored in a database to match. ACE also secures XML applications, web services, and intra-application communications application.

ACE also supports SSL encryption and standard network security features such as Network Address Translation (NAT) and Port Address Translation (PAT).

ACE supports stateful firewall ACLs in conjunction with application and protocol inspection engines that use the Layers 4 through 7 packet inspection capabilities of content switching. For example, ACE performs HTTP deep packet inspection of the header, URLs, and payload. ACE also performs protocol compliance inspection, filtering, and fix-up for data center protocols such as Real-Time Streaming Protocol (RTSP), DNS, FTP, and Internet Control Message Protocol (ICMP). These capabilities allow the ACE to provide IDS- and IPS-type protection from protocol attacks and DoS attacks, plus an additional level of protection against day-zero attacks.

Security remains a priority concern for nearly all IT organizations. An important feature of ACE is that it blocks anomalous signatures in network traffic targeted at server software that

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

might otherwise result in DoS. In this way, ACE protects against identity theft, data theft, application disruption, and fraud.

This integrated firewall capability minimizes the space, capital expense, management, and possible performance implications associated with having to install a separate security device between data center switches and servers to manage user access and to identify and control malware. Building layers of security protection into all key network junctures is a security best practice that Cisco recommends.

ACE SSL processing charateristics:

- Digital encryption technologies are used to provide secure HTTP communications.
- SSL support on the ACE can be used to provide SSL offload, back-end SSL encryption, or end-to-end encryption.
- PKI resources such as key pairs and certificates are required to support SSL.
- SSL services on the ACE module are provided through SSL proxies.
- SSL session reuse allows a previous SSL session ID to be used in a new TCP connection.
- SSL queue delay configures the ACE to queue data from the server before encrypting it for return to the client.
- In a standard SSL implementation, the server authenticates itself to the clients by sending an X509 certificate. This gives the clients some assurance that the server is who it says it is.

As XML traffic increases in the data centers, an application switching solution must also provide XML security. The Cisco Application Control Engine (ACE) Extensible Markup Language (XML) Gateway appliance secures XML applications, web services, and intra-application communications. When the Cisco XML ACE Gateway is deployed inline, stateful ACLs limit the traffic flows allowed between clients and servers. Access control prevents worms and intruders from arbitrarily scanning and attacking vulnerable ports that may be open on server platforms but not necessary for the application itself. Locating access control in the network limits vulnerabilities, even when the servers are not completely hardened.

https://t.me/learningnets

Can Cisco ACE Be Considered a Firewall?

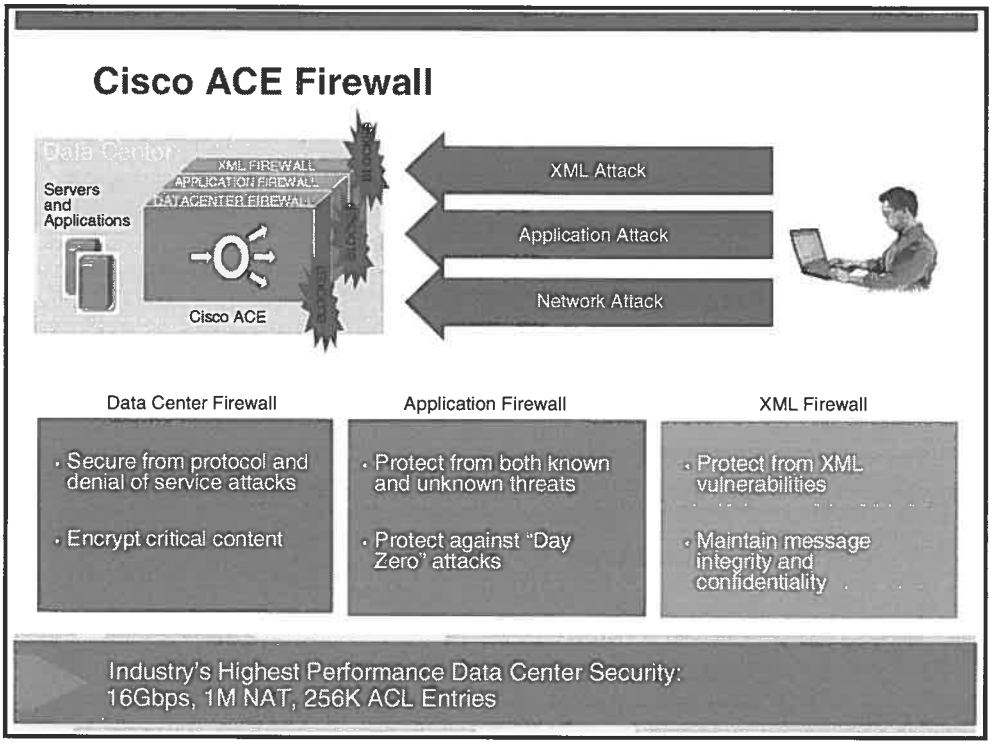
This is a case-by-case decision.

- Analyze customer's security requirements.
- Review the ACE security capabilities (next figure).
- Ensure the ACE security capabilities > Customer requirements.
- Understand that there are no plans to bring Cisco ACE to parity with the FWSM / ASA product line. The Cisco ACE focus is on Layer 7 (HTTP; Web App Security; XML).

© 2008, Cisco Systems, Inc. All rights reserved.

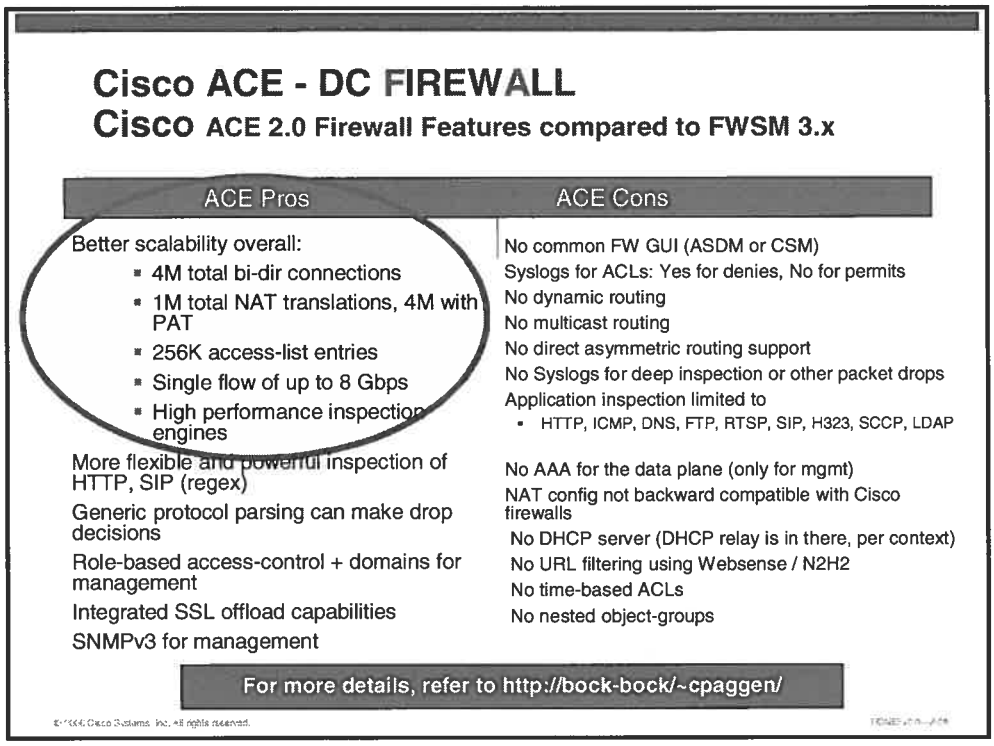
You must evaluate the customer's situation, analyze the customer's needs, and understand what ACE offers in terms of security. One important point is that Cisco is not trying to make the ACE to function as a firewall service module. The purpose of the ACE family is to offer integrated security.

https://t.me/learningnets



ACE is a last line of defense in front of the server—not to compete with FWSM, PIX or ASA.

https://t.me/learningnets



If the customer requires high performance NAT or scalable ACLs, 256,000 access control elements can be supported by ACE.

© 2008 Cisco Systems, Inc.

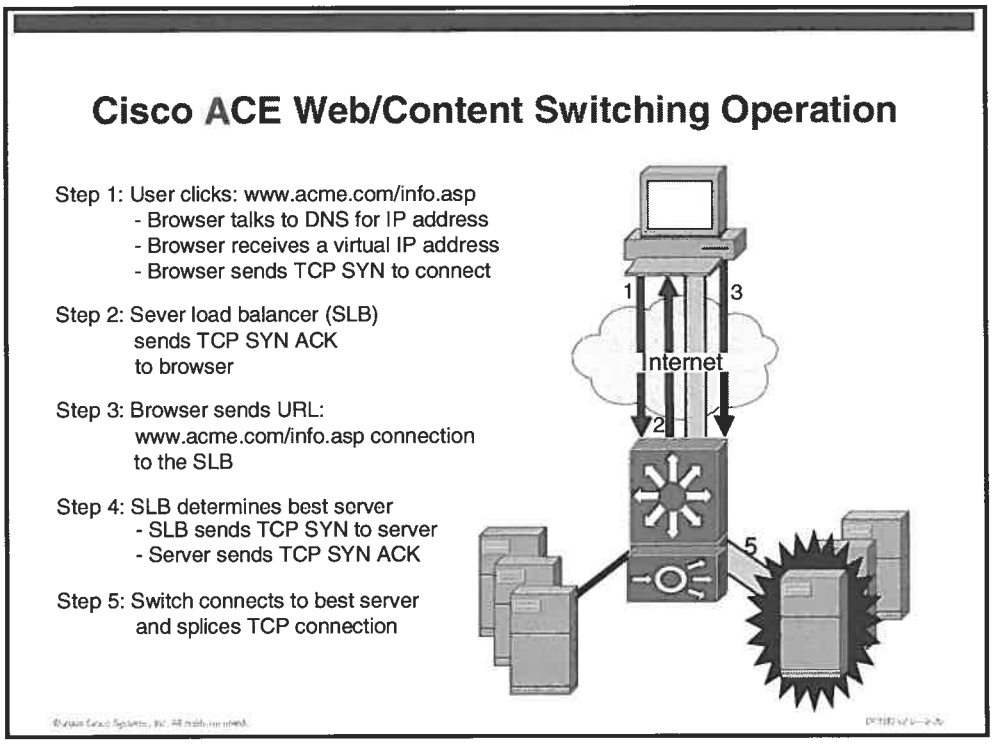
https://t.me/learningnets

Cisco ACE Module Performance and Scalability

	ACE 4710	ACE Module
Throughput	1/2Gbps	4/8/16Gbps
L4 Perf - CPS	120,000	345,000
L7 Perf - CPS	40,000	133,000
SSL Concurrent Conns	100.000	4M?
SSL – TPS (transaction per second)	7500	15,000
SSL Bulk Encryption	1 Gbps	4 Gbps
Virtualized Contexts	20	250
NAT / ACE	1M / 40K	1M / 256K
Compression Speed (Gzip/deflate)	Up to 1Gbps	NA
Application Accelerated Conns (flash forward, etc.)	10,000	NA

This figure shows a performance comparison in the ACE products.

https://t.me/learningnets



Content Switching Networks functions include the following functions:

- Web switching
  - Flow-based web traffic management
  - Classification based on content/URL
- Layer 4 switching
  - Per-packet forwarding and application
  - Classification by IP address/TCP port
- IP Routers, Layer 3 switching
  - Per-packet forwarding, low level QoS
  - Based on MPLS, IP Diff-svcs
- Layer 2 switching
  - Ethernet/PoS(frames)

The following information discusses web switching as it relates to the ISO model, including the level of intelligence that each layer enables and the role that each layer plays in the Internet.

Layer 2

Layer 2 switching in an enterprise network is switched Ethernet, but in the Internet, it can be frames or cells running over SONET/WDM physical networks. Asynchronous Transfer Mode (ATM) is one option to carry IP traffic in fixed sized cells, or Packet over SONET can carry IP frames directly in variable length frames.

Layer 3

Layer 3 is the traditional domain of the IP router, which is optimized for packet-by-packet forwarding of connection-less IP datagrams. At this level, core routers can spend a limited

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

amount of resources classifying traffic and applying policies to traffic, especially at OC-192. The mechanisms available include IP Diff-svcs, which provides a small number of relative priority classes for each packet, based on the Type of Service (ToS) bits that are found in the IP header. Multiprotocol Label Switching (MPLS) is another technique to apply limited policy on a packet-by-packet basis.

Layer 4

Layer 4 SLBs simply extend intelligence to the TCP/UP layer, which enables them to classify packets based on the port number (that is, TCP: port 80 for HTTP, port 23 for telnet, port 119 for NNTP, port 21 for FTP Control; and UDP: port 53 for DNS, port 20 FTP Data). They may be able to apply policies using this information, and basic load balancers can use a unique IP address/port pair to make simple server selection decisions. NAT is a Layer 4 function, and can map external IP address/port to a private IP address/port.

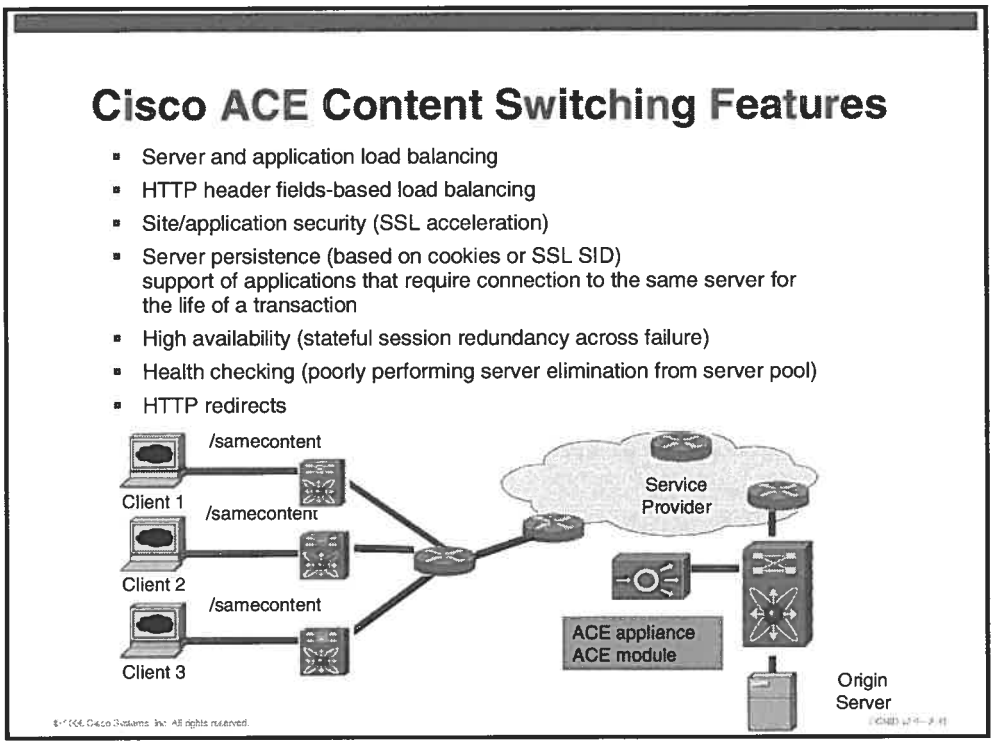
Layer 5

Layer 5 is the session layer. On the web this means HTTP, and by looking at the HTTP header, an SLB can make routing and policy decisions based on the content being requested, using the **host tag/URL/cookie**. (See Content Switching Features, in the following section, for more details)

Layer 7

A Layer 7 switch uses the same intelligence as a web switch. Layer 7 is the Application layer in the original ISO model and can be indicated by these file extensions: .HTML, .CGI, or .RAM. Using this information, an SLB can make decisions to select optimized servers or caches for particular applications. Although this is extended functionality, the key information is still learned from inspecting the HTTP header as described in Layer 5.

https://t.me/learningnets



Cisco Content Switching formerly provided a feature-rich solution for today's web sites. These features are now supported by the Cisco ACE Appliance and the Cisco ACE module.

Server & Application Load Balancing: By configuring content rules, the Content Switch provides full URL-based switching.

HTTP Header Load Balancing: Enables the Content Switch to inspect incoming content requests for HTTP header fields. This inspection allows the Content Switch to make load balancing decisions that are based on the HTTP header field information and then direct content requests to the servers that are designed to handle the type of content being requested.

Site & Application Security: Integrated SSL Acceleration ensures that user data is encrypted without the management and processing burden of server based SSL

Server Persistence: In support of applications that require connection to the same server for the life of a transaction, you can configure users to stick to a particular server employing cookies or SSL session ID.

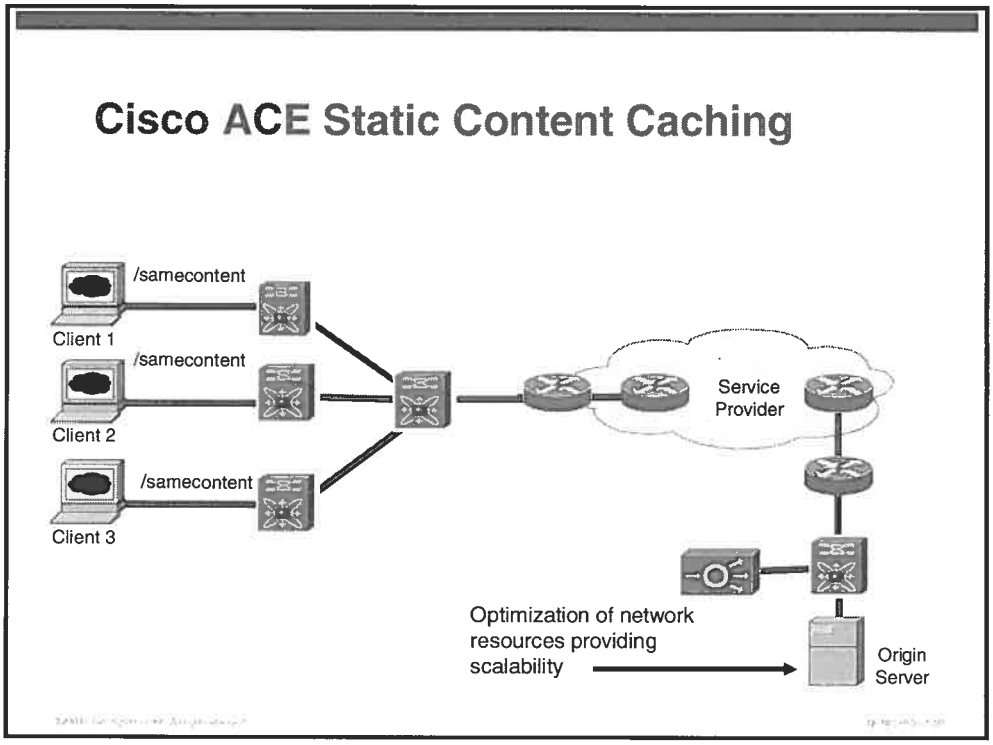
High Availability: Stateful session redundancy provides uninterrupted application access, even through hardware failures.

Health Checking: Standard and custom synthetic transactions automatically eliminate poorly performing servers or applications from the pool of available resources.

HTTP Redirects: The Content Switch supports HTTP redirect services and redirect strings to send users to an alternative location or domain.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

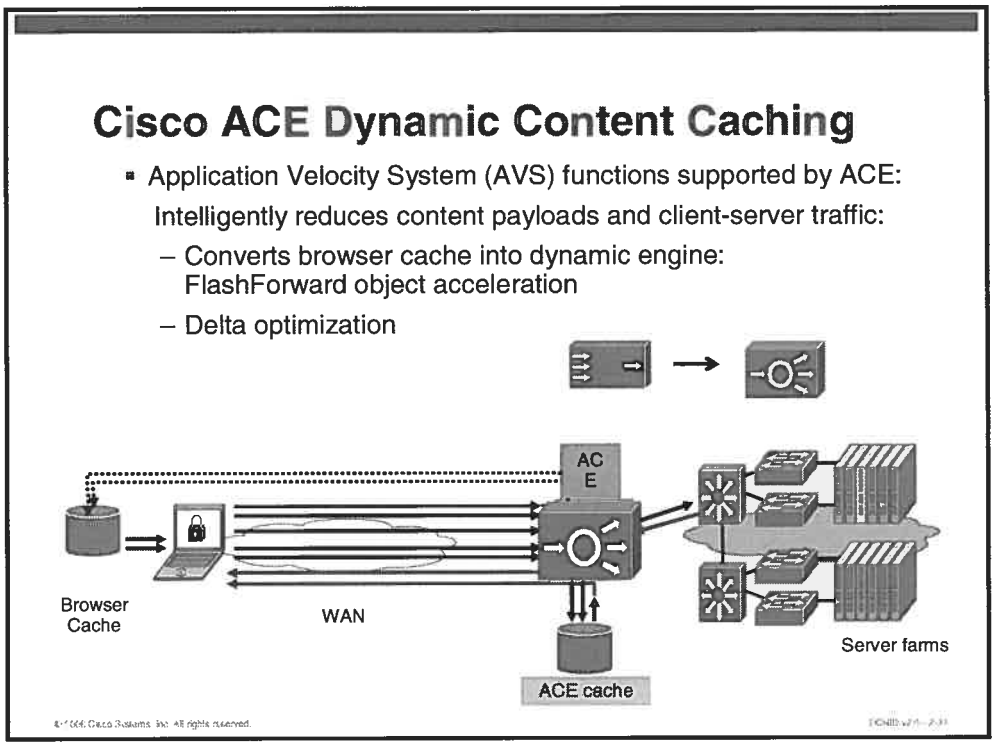


One cost-effective solution to reducing web wait is to push content out to the edges of the Internet and closer to end users. The combination of the Cisco Content Engines and WCCP-enabled routers delivers powerful solutions for accelerating content delivery and optimizing WAN bandwidth usage by pushing content to the edges of a network. In the Network Caching Services solution, Cisco WCCP-enabled Routers and Content Engines accelerate content delivery, optimize WAN bandwidth usage, and deliver rapid payback for any enterprise or Service Provider.

Network caching is the saving and storing of information locally. Copies of recently requested content are stored temporarily on a content engine in locations topologically close to the client. The content is readily available to be reused for subsequent client requests for the same content. Network caching offers:

- Improved response time
- Reduced Internet congestion
- Conservation of network bandwidth
- Optimization of network resources

https://t.me/learningnets



The Cisco Application Velocity System (AVS) is an intelligent cache system that is integrated with Cisco ACE load balancer. AVS intelligently reduces content payloads and client-server traffic by:

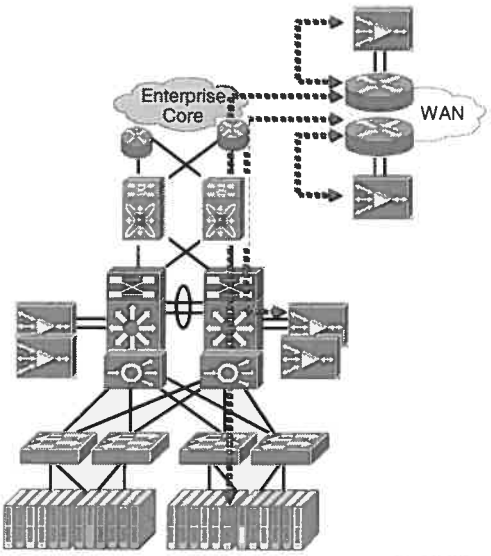
- Using the client's browser cache as a dynamic caching engine. With AVS FlashForward object acceleration, embedded objects referenced in HTML container pages are served with "Expires:" tags, which sets expiry in the future. On the next visit, the browser will not send GET for objects in cache if the current date and time are not greater than the object expiry date. This reduces the total number of HTTP requests for subsequent visits to the same page.
- Delta optimization calculates and sends only the difference between two visits to a dynamic HTML page. This results in reduced bandwidth usage and reduced page download times.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

The Impact of Caching

- Overall reduced traffic flow through data center network...
- ...but traffic profiles change as applications evolve
  - Shift from static web content to dynamic web applications



Caching in the data center can be accomplished by the WAE in combination with Cisco ACE appliance/module SLB. It results in an overall reduction in traffic flow through the data center network, but traffic profiles can change as applications evolve. For example, the trend today is a shift from static web content to dynamic web applications. This reduces the effectiveness of traditional cache devices.

https://t.me/learningnets

CSS 11500 / Cisco ACE 4710 Appliance Comparison

	CSS 11501-S-C	CSS 11503	CSS 11506	ACE 4710 *
Server Load Balancing	Yes	Yes	Yes	Yes
SSL Offload	Yes	Yes	Yes	Yes
Virtualization	No	No	No	Yes
Max Throughput	1 Gbps	2 Gbps	4 Gbps	2 Gbps
Layer 4 CPS	7,000	22,000	44,000	120,000
Layer 7 CPS	3,500	10,000	20,000	40,000
Concurrent Connections	150,000	550,000	1.1 M	1 M
SSL TPS	1,400	2,800	5,600	10,000
SSL Bulk Encryption	256 Mbps	512 Mbps	1 Gbps	1 Gbps
SSL Concurrent Connections	40,000	80,000	160,000	100,000
Max Compression	256 Mbps	512 Mbps	1 Gbps	1 Gbps

© 2006 Cisco Systems, Inc. All rights reserved

* ACE 4710 performance numbers are "target" at this time.

A quick comparison chart on how the Cisco ACE appliance compares to the Cisco CSS Content Services Switches family.

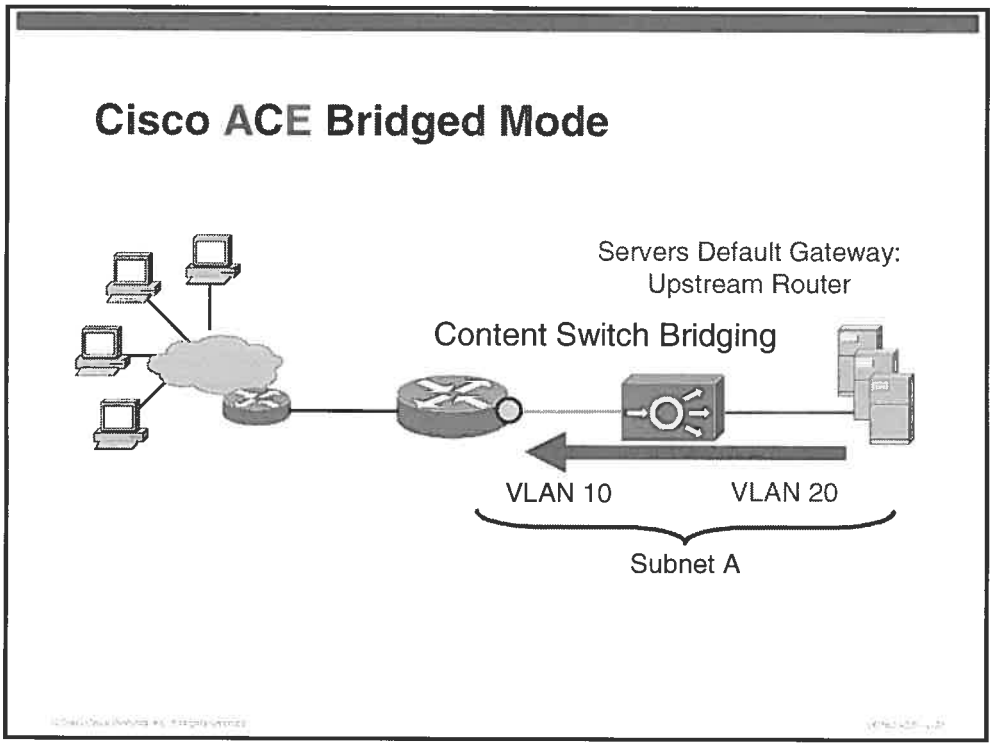
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Network Topologies with Cisco ACE Appliance/Module

The objective of this topic is to describe possible Cisco ACE Module or Cisco ACE appliance deployment topologies including routed, bridge, and one-arm modes, as well as direct server return possible deployment topologies.

Cisco ACE Modes

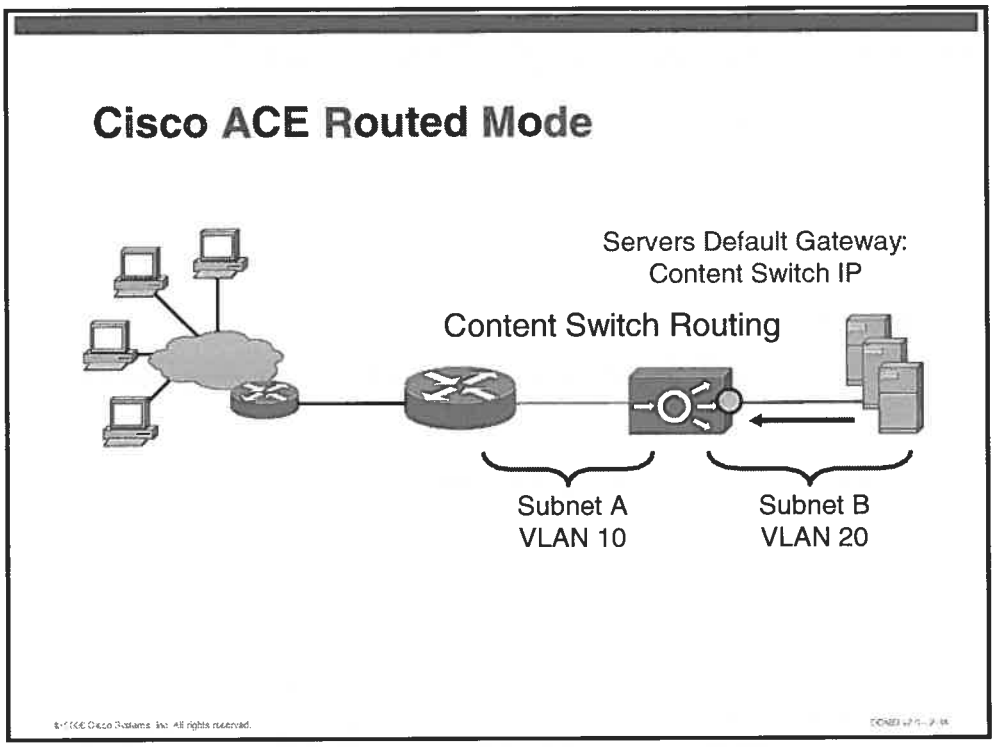


The Cisco ACE module can be configured in bridge mode. In this mode, the client and server VLANs are part of the same IP subnet, as shown in the preceding figure. The Cisco ACE module uses an Address Resolution Protocol (ARP) table to track which VLAN contains what physical devices.

In the figure, VLAN 10 is used as the client-side VLAN, while VLAN 20 is the server-side VLAN. The same IP subnet is used on both VLANs. The physical port attached to the upstream router is assigned to VLAN 10. Physical ports connected to the servers are assigned to VLAN 20. The servers in a bridge mode environment are configured to use the IP address of the upstream router interface as their default gateway.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

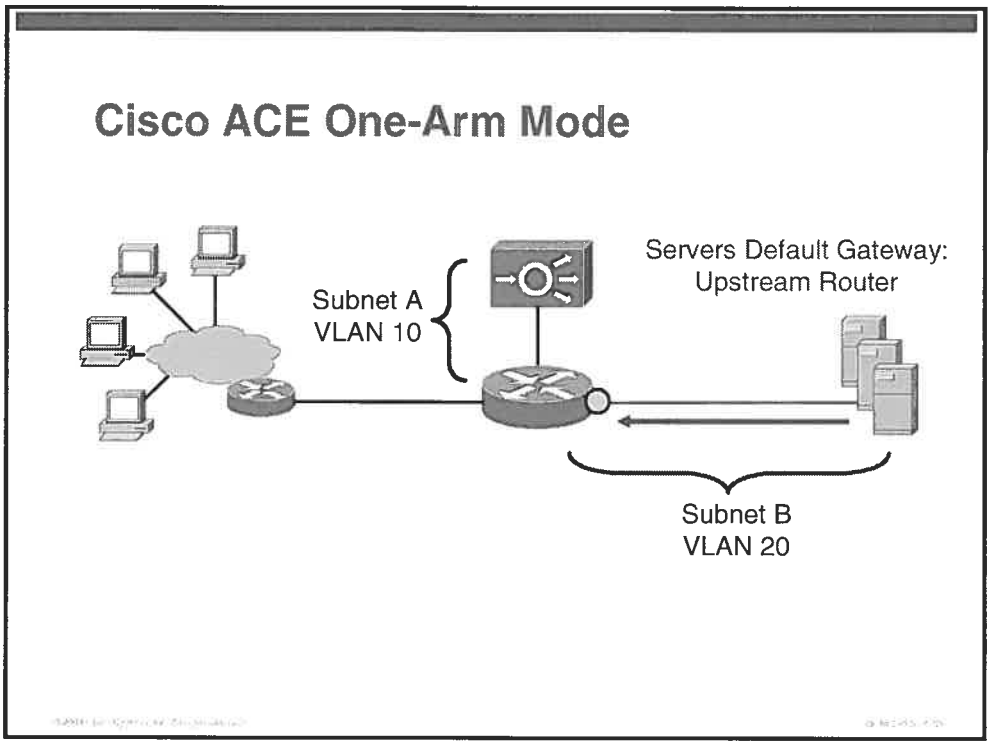


The Cisco ACE module can be configured in routed mode. In this mode, the client and server VLANs are part of different IP subnets, as shown in the preceding figure. This breaks the old rule of 1 VLAN per subnet and 1 subnet per VLAN.

In the figure, VLAN 10 is configured as the client-side VLAN, and VLAN 20 is the server-side VLAN. Different IP subnets are associated with each VLAN. The physical port attached to the upstream router is assigned to VLAN 10. Physical ports connected to the servers are assigned to VLAN 20. The servers in a router mode environment are configured to use the IP address of the ACE as their default gateway.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The one-arm mode removes the Cisco ACE from a position directly in the transit path for all traffic to the server farms. This configuration has the advantage that the Cisco ACE does not need to process traffic that is not affected by Cisco ACE features. In the preceding figure, VLAN 10 is used for traffic between the Cisco ACE and the Multilayer Switch Feature Card (MSFC), and VLAN 20 is used for traffic to the server farms. A VLAN 10 interface is configured on the MSFC, and an IP address from subnet A is configured on the Cisco ACE. Additional IP addresses from subnet A are used to configure the virtual server IP addresses. A VLAN 20 interface is configured on the Multilayer Switch Feature Card (MSFC) and is used by the servers as their default gateway.

Note A Cisco ACE in one-arm mode has only one VLAN.

Return traffic that is generated by the servers in response to load-balanced requests is still needed by the Cisco ACE for full functionality. Getting this traffic to flow through the Cisco ACE is more complicated than with an inline configuration. There are two ways to address this situation:

- Source NAT
- Policy-based routing (PBR)

Source NAT

Source Network Address Translation (SNAT) is configured by creating a pool of IP addresses. Client IP addresses are translated to IP addresses from the client pool. These translated addresses are used as the source address in the source packet that is sent to the server.

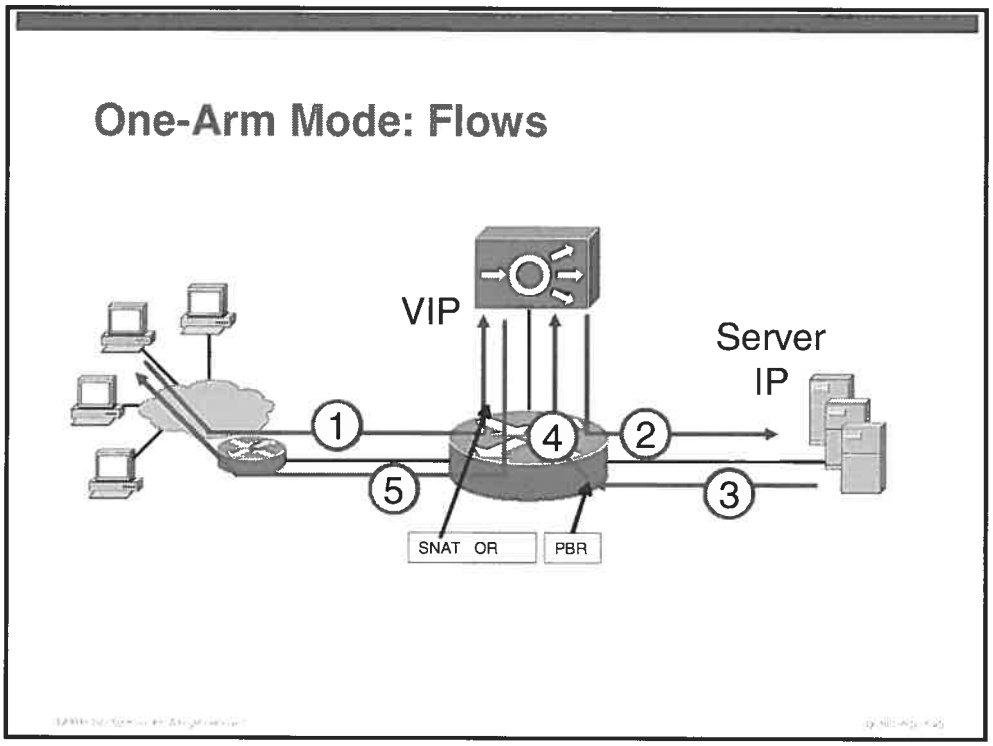
https://t.me/learningnets

Policy-Based Routing

PBR is a router feature available on Cisco IOS-based routers, including the Catalyst 6500 MSFC. PBR allows the router to be configured to select a next hop for a packet based on a configured policy. This policy overrides the routing decision that would have been made by consulting the routing database. A routing policy is attached to the ingress interface on the router. Access lists can be used to limit the traffic to which the policy is applied. For example, web responses being sent to clients can be load balanced and redirected via policy-based routing, while SNMP responses from the servers are routed normally.

© 2008 Cisco Systems, Inc.

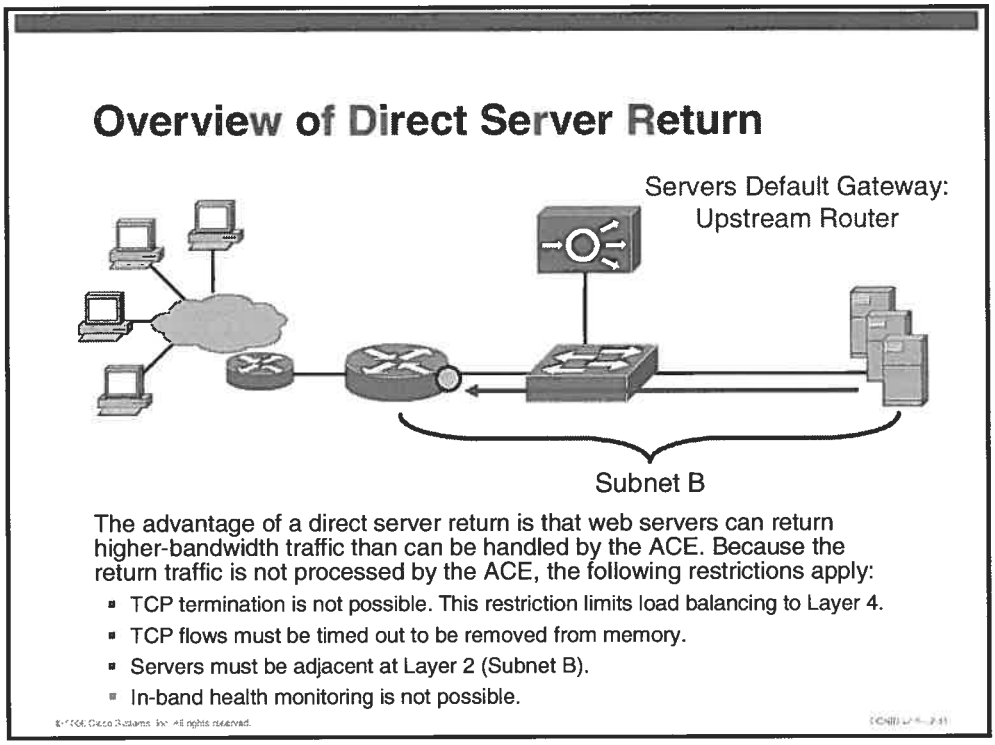
https://t.me/learningnets



Traffic flow for load-balanced requests is shown in the figure. Packets are handled as follows:

- Traffic from the client to the VIP is routed normally by the MSFC.
- Traffic from the Cisco ACE to the server is routed normally by the MSFC. If SNAT is used, the source IP address is in the client NAT pool. Otherwise, the source IP address remains the client IP address.
- Traffic from the server is returned to the MSFC because the MSFC is the server default gateway.
- If SNAT is used, the destination IP address in the server response is routed normally to the Cisco ACE. If SNAT is not used, PBR must be used on the MSFC interface that is used as the server default gateway. The policies configured must match any traffic being sent in response to a load balanced request. The IP address specified for the Cisco ACE is set as the next hop address by PBR.
- Traffic from the Cisco ACE to the client is routed normally by the MSFC. If SNAT is used, the Cisco ACE has translated the destination IP address from the NAT pool IP address to the client IP address. If PBR is used, the Cisco ACE does not need to modify the destination IP address because the client IP address is already in the packet.

https://t.me/learningnets



A variation of one-arm mode is a direct server return. The figure shows the architecture of this variation.

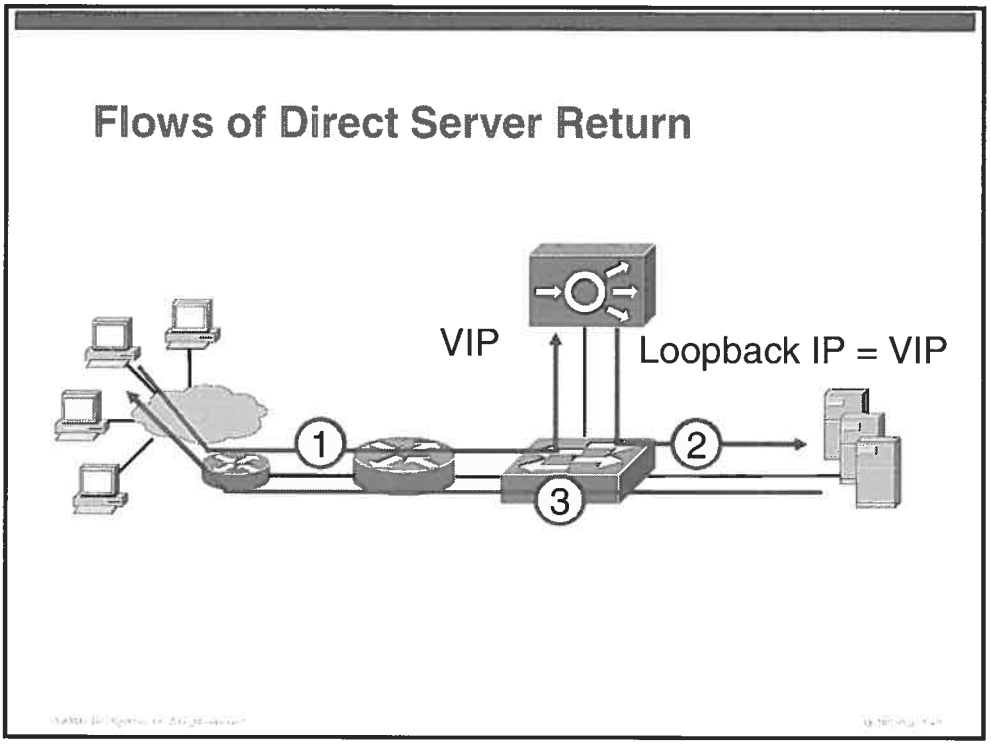
The Cisco ACE and the servers are placed in the same VLAN and IP subnet. An interface on that VLAN is defined on the MSFC and is the default gateway for the Cisco ACE and the servers. NAT is turned off for the server destination address. Return traffic does not flow through the Cisco ACE but returns directly to the client.

The advantage of a direct server return is that web servers can return higher-bandwidth traffic than can be handled by the Cisco ACE. Because the return traffic is not processed by the Cisco ACE, the following restrictions apply:

- TCP termination is not possible. This restriction limits load balancing to Layer 4.
- TCP flows must be timed out to be removed from memory.
- Servers must be adjacent at Layer 2.
- In-band health monitoring is not possible.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

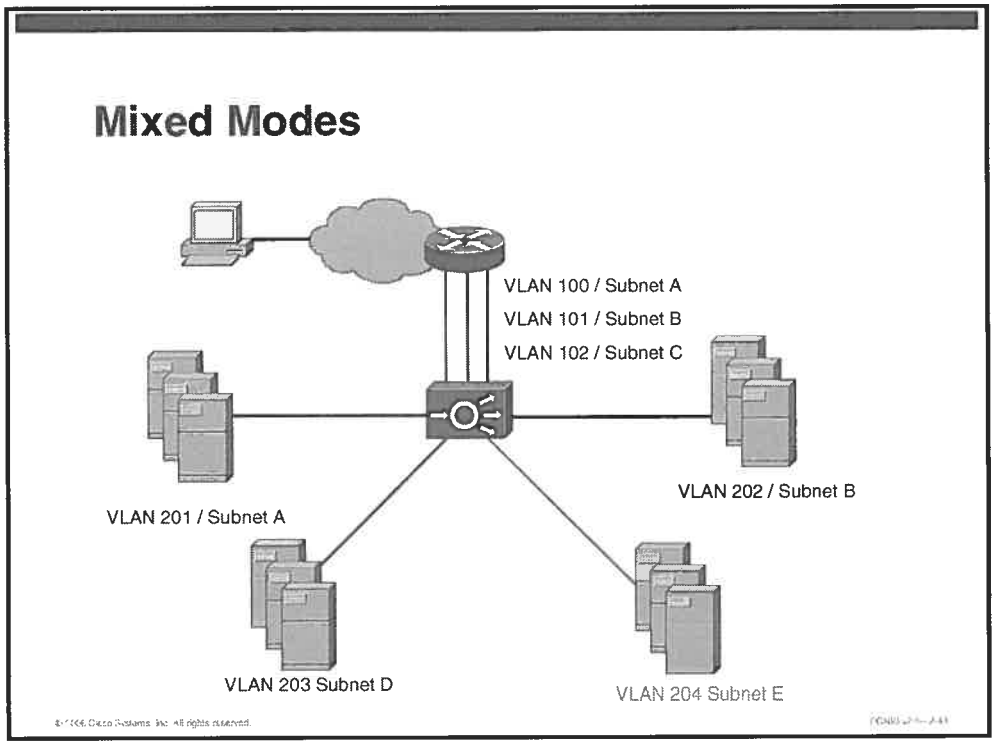


The traffic flow for load-balanced requests is shown in the figure. The packets are handled as follows:

- Incoming client requests are routed to the server VLAN. The packet is switched to the Cisco ACE.
- The Cisco ACE rewrites the Layer 2 destination MAC address and returns the packet to the switch processor. The packet is switched to the server. The server uses a loopback interface configured with the VIP address so that the server accepts a packet destined for the VIP.
- The server responds directly to the client. This traffic is routed normally because the MSFC is the default gateway for the server.

When no more traffic is generated by the client on this TCP connection, the connection goes idle. After the idle timeout, the Cisco ACE removes the connection from its session table.

https://t.me/learningnets



The Cisco ACE can handle multiple pairs of VLANs and mixed modes. The figure shows one Cisco ACE handling several VLANs. The following mode configurations are possible:

- Subnet A bridged between VLAN 100 and VLAN 201
- Subnet B bridged between VLAN 101 and VLAN 202
- Subnet C on VLAN 102 routed to Subnet D on VLAN 203 or Subnet E on VLAN 204

Note	The bridging versus routing mode selection is not an extensive module configuration option. Rather, it is driven by the interface configuration within the Cisco ACE module.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco ACE Module Scalable Performance

ACE Module Performance Options Configuration	Throughput (Gbps)	Maximum Layer 4 Setups and Teardowns per sec
4 ACE Modules with 16-Gbps license	64	1,392,000
3 ACE Modules with 16-Gbps license	48	1,044,000
2 ACE Modules with 16-Gbps license	32	696,000
ACE Modules with 16-Gbps license	16	348,000
ACE Modules with 8-Gbps license	8	348,000
ACE Modules with 4-Gbps license	4	348,000

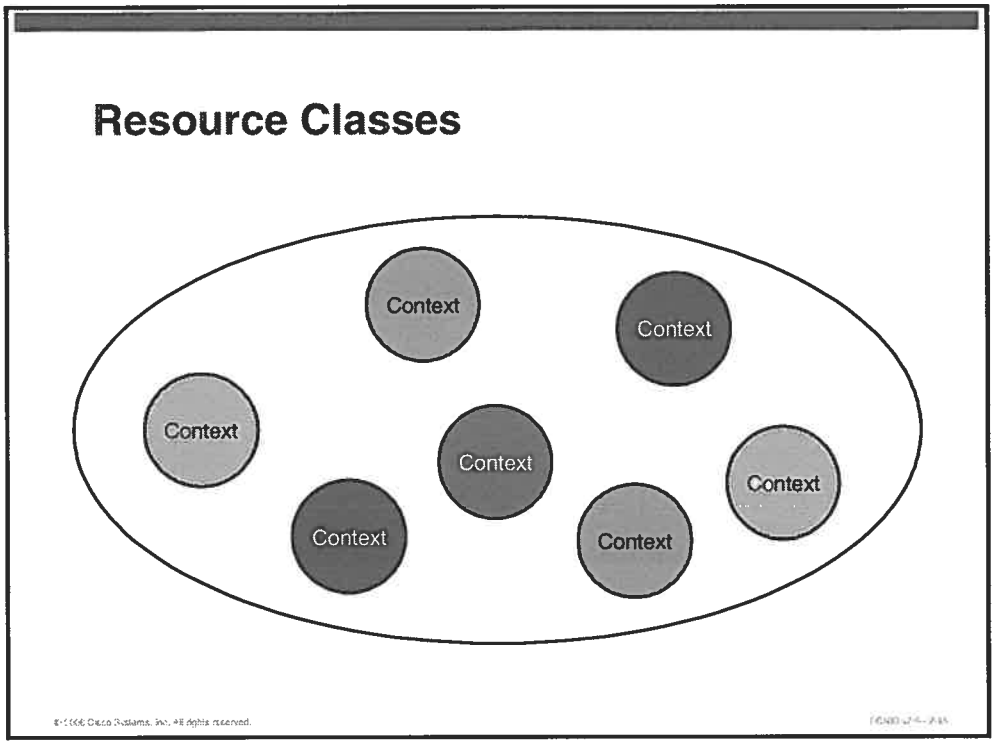
The Cisco ACE module uses a single interconnect to the Cisco Catalyst 6500 Series switch fabric to support up to 16-Gbps throughput. It has five major hardware components:

- Parallel multicore network processors to handle data-plane traffic
- An application-specific field programmable gate array (FPGA) to distribute traffic across two network processors
- A dual-core processor to handle management (control-plane) traffic
- A dedicated cryptographic processor to handle SSL encryption and decryption
- Two optional daughter cards to support advanced application services such as web application security and application acceleration

Protocol inspection and fix-up of the most popular data center protocols

- ACLs with up to 256,000 access control elements to selectively allow traffic between ports
- SYN cookies
- TCP connection state tracking
- Virtual connection state for UDP
- SEQ number randomization
- TCP header validation
- TCP window size checking
- Unicast Reverse Path Forwarding (URPF) checking at session establishment

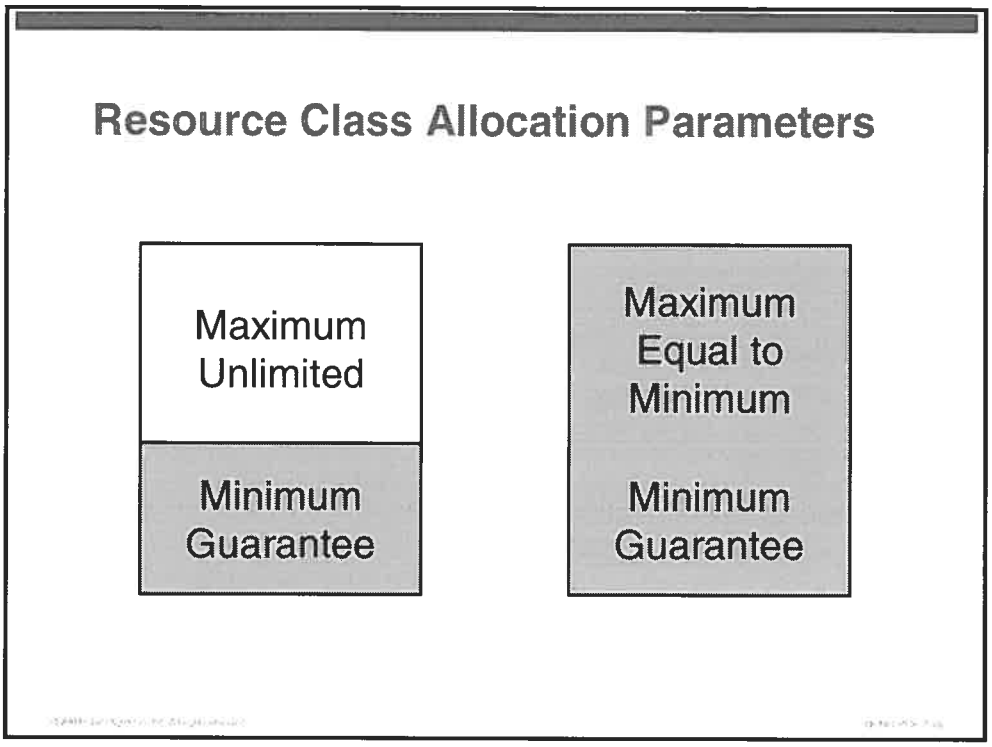
https://t.me/learningnets



Resource allocation controls are defined in resource classes. Each context is then assigned to a single resource class. If no resource classes are defined, every context is a member of a resource class named default, which has unlimited access to system resources. A maximum of 100 resource classes can be configured.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

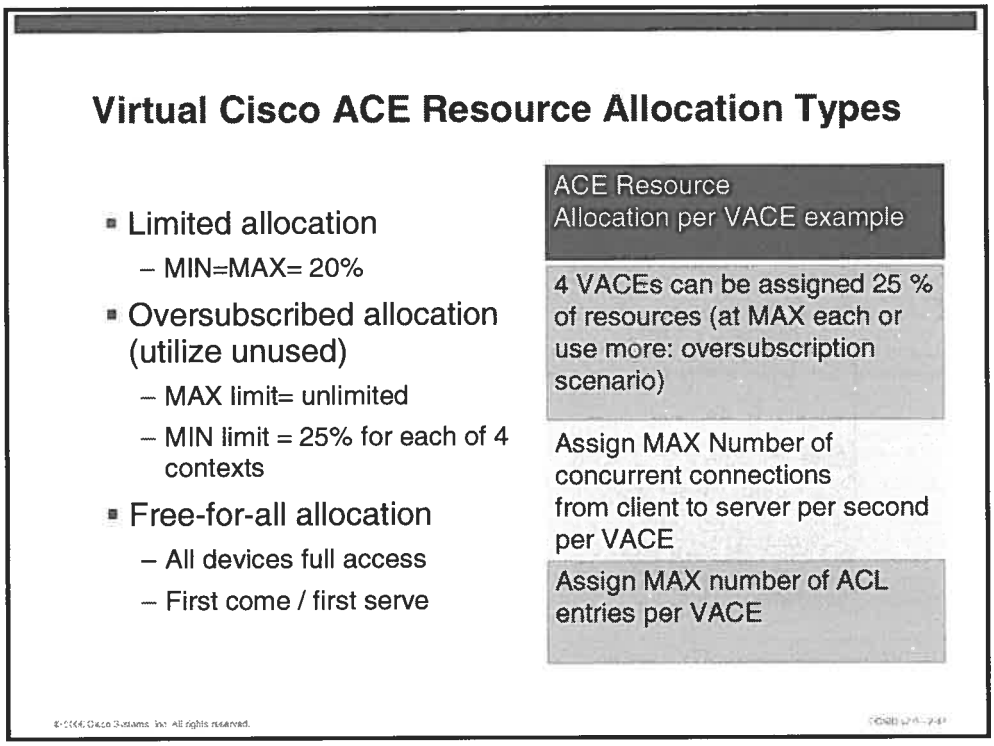


Individual resource allocation controls specify a minimum and a maximum resource allocation, which is configured as a percentage of overall system resources. Minimum resource allocations constitute a guaranteed resource allocation. Maximum resource allocations can either be specified as unlimited or as equal to the minimum.

The resource allocations defined in a resource class are applied to each context in the resource class. For example, a resource class that specifies a 20 percent minimum resource allocation and that has four contexts results in 80 percent of system resources allocated.

© 2008 Cisco Systems, Inc.

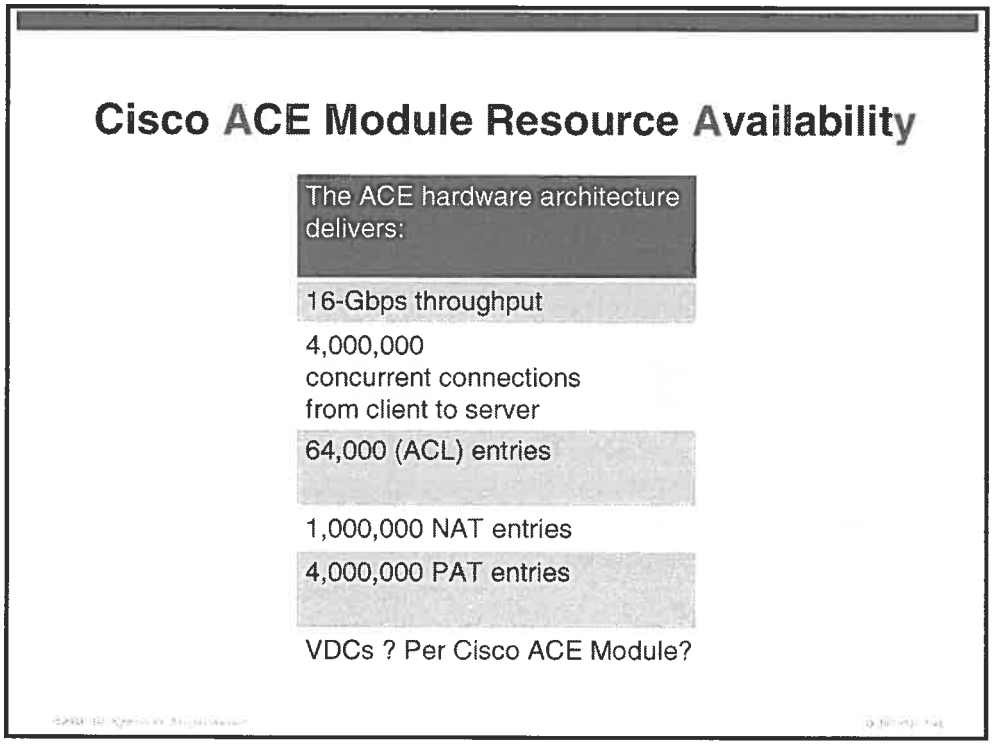
https://t.me/learningnets



The Cisco ACE application switch allows you to limit and manage various resources such as bandwidth, connections per second, management connections per second, SSL connections, and access lists for each virtual device. You can allocate minimum and maximum amounts of each resource or all resources in terms of percentages for each virtual device. In addition, the Cisco ACE provides flexibility to allocate its resources to a virtual device in three different ways: limited, oversubscription, and free-for-all. To help ensure resources for an application, the Cisco ACE allows you to limit minimum and maximum amounts of all Cisco ACE resources or each resource for a virtual device. For example, you can limit the Cisco ACE resources to 20 percent for an application hosted on a virtual device by setting the virtual-device minimum and maximum resource limit to 20 percent. The Cisco ACE also supports oversubscription by setting the maximum resource limit to unlimited, allowing an application hosted on a virtual device to use the Cisco ACE resources that are allocated but are not consumed by other virtual devices. As an oversubscription example, you can partition a physical Cisco ACE device into four virtual devices, and allocate each virtual device a minimum of 25 percent of the resources and a maximum of an unlimited amount of the resources. In this example, the Cisco ACE allows an application hosted on each of the virtual devices to consume more than the allocated 25 percent of Cisco ACE resources if any of the other three virtual devices are not consuming their minimum allocated 25 percent. In the free-for-all default setup, the Cisco ACE permits all virtual devices to have full access to all of the resources on a first-come, first-served basis

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Cisco ACE application switch module data plane consists of four of the five main hardware components. Because it uses multicore network processors, the Cisco ACE module provides powerful performance for application-delivery services. Moreover, this hardware-based application-delivery module maintains good performance levels compared to application-delivery devices based on general-purpose processors when most commonly used features are selected. You do not have to trade off features to obtain a desired level of performance with the Cisco ACE module.

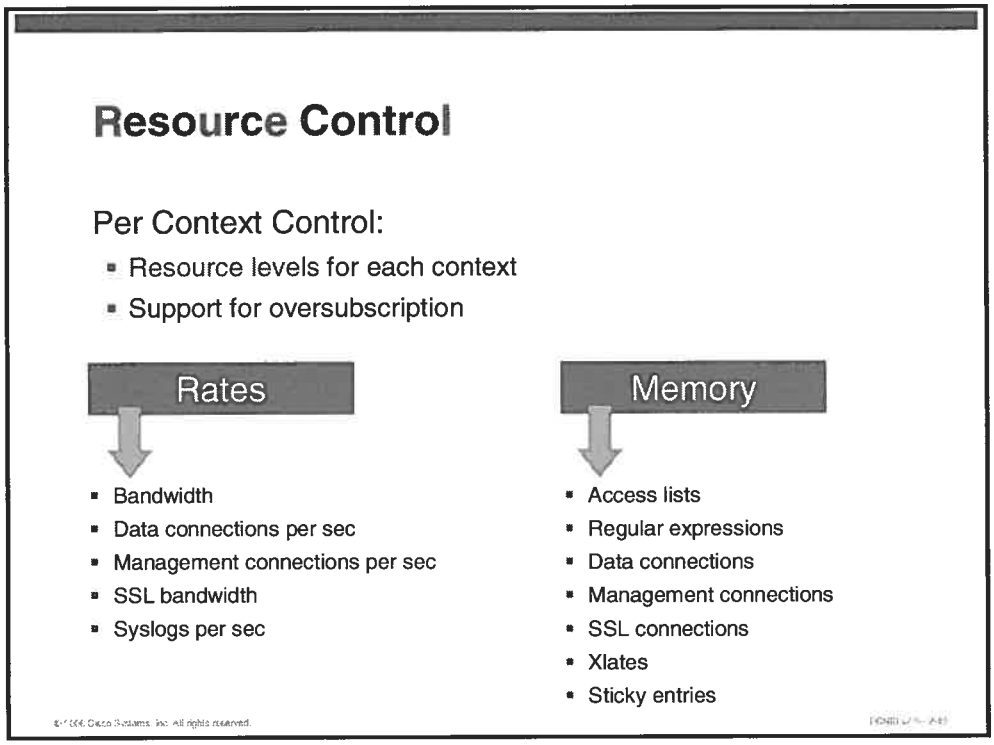
The Cisco ACE hardware architecture delivers:

- 16-Gbps throughput
- 4,000,000 concurrent connections from client to server
- 64,000 access-control-list (ACL) entries
- 1,000,000 Network Address Translation (NAT) entries
- 4,000,000 Port Address Translation (PAT) entries.

These performance numbers are difficult to achieve on software-based application-delivery devices.

Scalability up to 64-Gbps throughput: The Cisco ACE application switch is the only application-delivery product in the industry that provides linear scalability up to 64-Gbps throughput. If you need more than the 16-Gbps throughput provided by a single Cisco ACE module, you can deploy up to four additional modules within the same Cisco Catalyst 6500 Series Switch chassis without adding cables or rack space. Miercom, an independent network test vendor, verified tests results showing that four Cisco ACE modules in a Cisco Catalyst 6500 Series Switch chassis can deliver close to the theoretical limit of 64 Gbps without reaching any chassis or backplane limitation. The Cisco ACE modular architecture reduces the operational and data center power, space, and cooling costs compared to appliance-based solutions. It can also help you consolidate your server-based application delivery devices, resulting in more effective use of your application server infrastructure.

https://t.me/learningnets

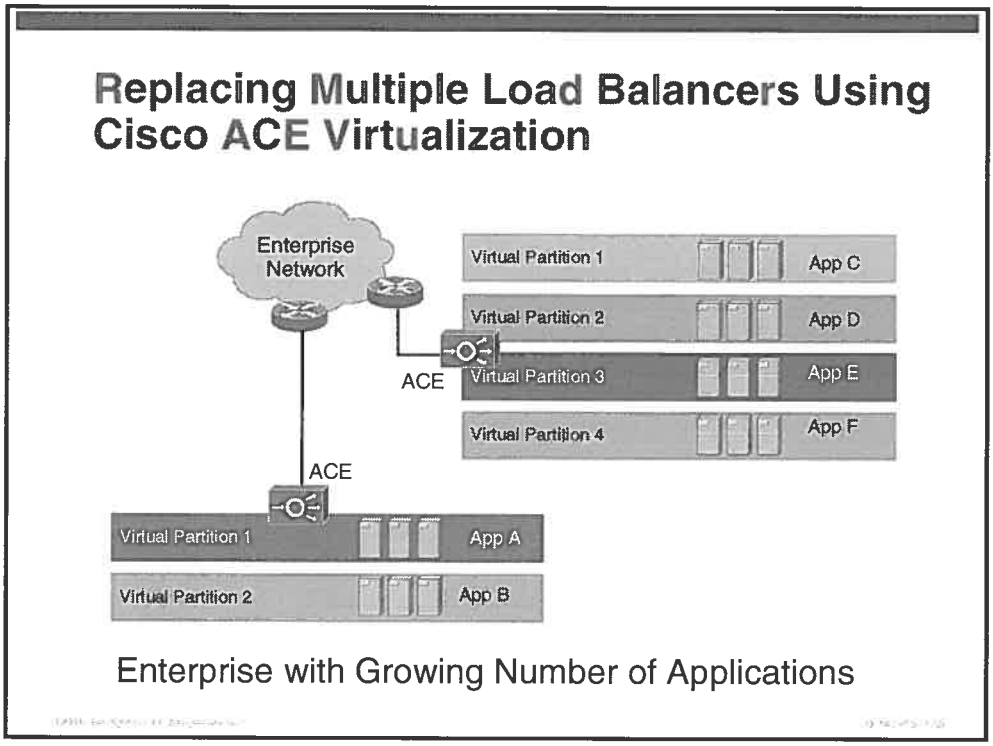


Cisco ACE hardware resources are allocated to individual contexts under the control of resource-level controls configured in the Admin context. The resources that can be managed fall into two categories and are listed in the preceding figure. The rate-based resources are amounts of data or number of events per second. The memory-based resources control the storage for different kinds of state and configuration objects, which are stored in memory.

Note	The memory available on an appliance-wide basis for each type of state and configuration data is pre-allocated.

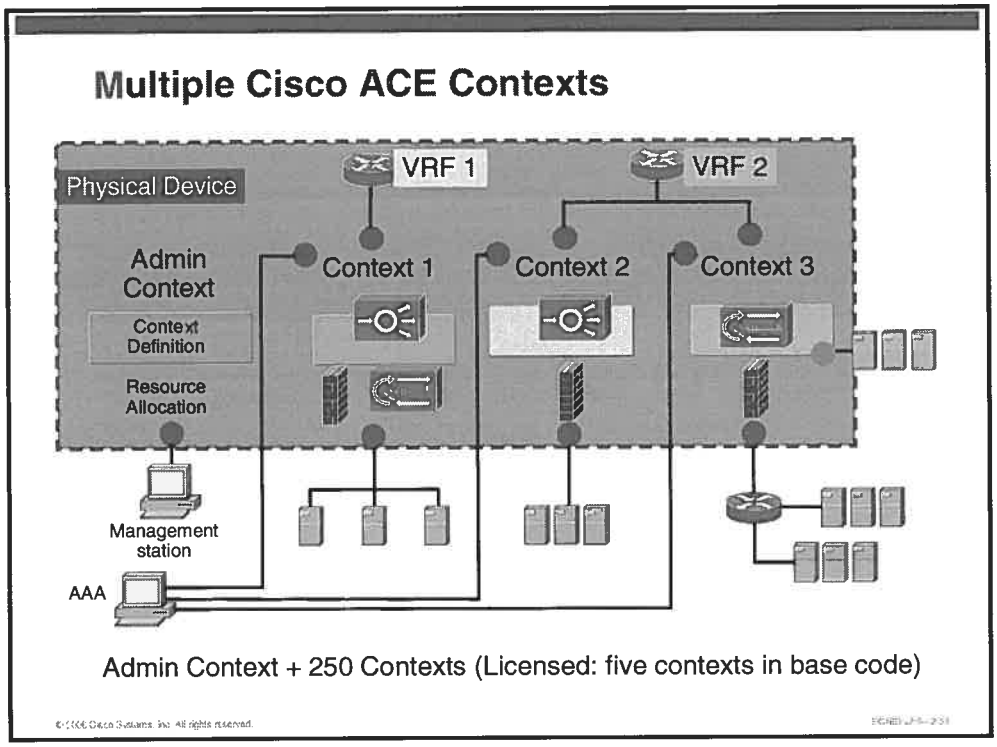
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Multiple Cisco ACE contexts can be used to provide the same load-balancing functions needed by the previous scenario. Additional applications can be added without the purchase of additional hardware if capacity is available.

https://t.me/learningnets



Cisco ACE Virtualization and Network Segmentation

The Cisco ACE Appliance supports virtualization through the extension of the logic to the application delivery space of the Layers 2 and 3 VLANs and VRF instances that the Cisco Catalyst 6500 Series natively supports. It is simple to map Cisco ACE virtual devices to VLANs and VRF instances, thus associating a separate network instance on the Cisco Catalyst 6500 Supervisor Engine with a completely independent application delivery instance.

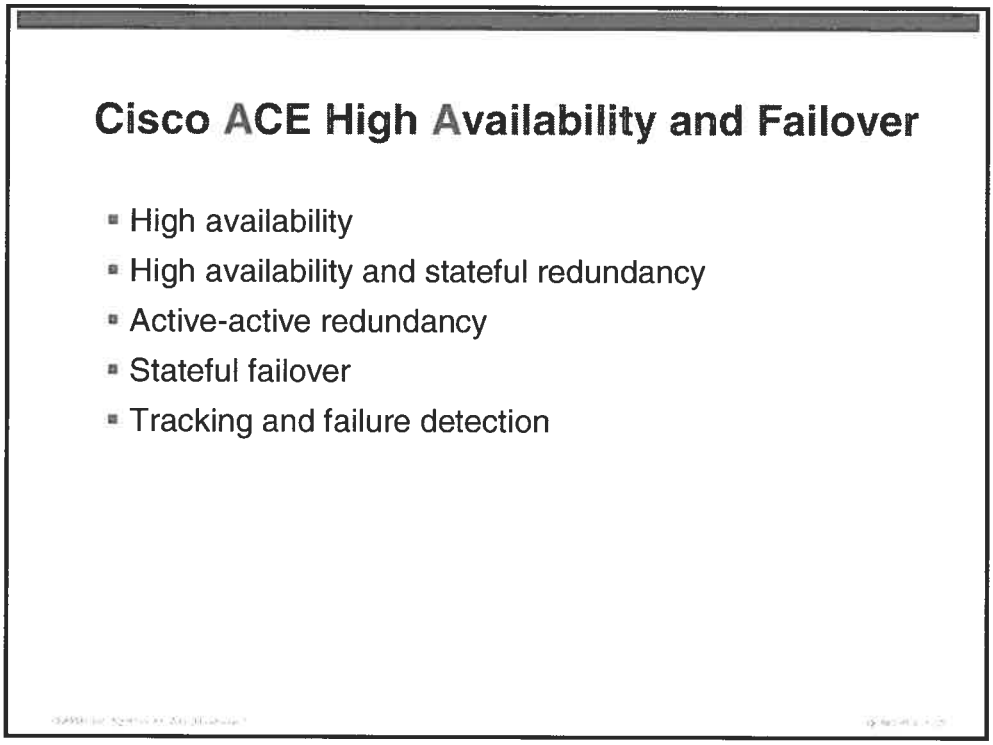
Note	The RHI feature is aware of VRF instances, so Cisco ACE virtual devices can inject and remove routes directly from VRF routing tables in the Cisco Catalyst 6500 Supervisor Engine.

Each virtual device can be dedicated to a set of applications, to an organization within the enterprise, or to a customer in a hosted environment. Overlapping IP addresses are supported and each virtual device benefits from independent network management and policies and also from a dedicated virtual routing instance with full Cisco IOS routing protocol support.

Network resources can be dedicated to a single context or shared between contexts as shown in the preceding diagram. By default, a context named Admin is created by the Cisco ACE module. This context cannot be removed or renamed. Additional contexts and the resources to be allocated to each context are defined in the configuration of the Admin context. The number of contexts that can be configured is controlled by licensing on the Cisco ACE module. The base code allows five contexts to be configured while licenses are available, which expand the virtualization possible to 250 contexts. The Admin context does not count to the licensed limit on the number of contexts.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



High Availability

Cisco ACE Modules can be deployed in redundant mode either within the same Cisco Catalyst 6500 Series chassis or, more commonly, in separate chassis. In both cases, the Cisco ACE Module supports the capability to preserve existing connections even after a module failure or shutdown, by synchronizing flows and client-to-server associations (stateful failover).

The Cisco ACE failover mechanism allows each virtual device to fail over independently, according to user-configurable rules that provide the flexibility to respond rapidly in a variety of failure scenarios. The basic heartbeat between the two modules can be sent at 100 milliseconds (ms) intervals, achieving a complete failover in a fraction of a second.

Using an internal secure communication channel, the Cisco Catalyst 6500 Supervisor Engine can immediately inform the Cisco ACE Module of link and interface failures or Hot Standby Router Protocol (HSRP) group state changes, thus enabling each Cisco ACE virtual device to take immediate action on events that the Cisco Catalyst 6500 Supervisor Engine is constantly monitoring. The Cisco ACE Module can then complement that information with health monitoring probes (including scripts) to check the status of upstream gateways or critical devices, thus providing a comprehensive set of simple and fast failover mechanisms that can adapt to the most complex scenarios.

High availability and stateful redundancy: Many mission-critical applications require transparent and sub-second failover if an application-delivery device becomes unresponsive. The Cisco ACE application switch is designed to deliver robust stateful redundancy at the physical and virtual device levels. The stateful redundancy feature of Cisco ACE can enhance your experience by helping to ensure that network services and applications are available regardless of device failure.

Virtual device redundancy: You can configure the Cisco ACE application switch to provide redundancy between two Cisco ACE modules in the same Cisco Catalyst 6500 Series Switch, in the same Cisco 7600 Series Router chassis, or in two physically separate chassis. The Cisco ACE also supports redundancy between two appliance form-factor devices. Compared to traditional solutions, the Cisco ACE supports high availability at both physical and virtual

https://t.me/learningnets

device levels. The Cisco ACE application switch can be partitioned into up to 250 virtual devices, each with its own configuration files, resources, and management interfaces. Each virtual device has all the capabilities of the actual physical device, and each virtual device is independent and isolated so that it appears to be a unique physical device from the viewpoint of the network and the network administrator. The Cisco ACE application switch provides the flexibility to configure redundancy only for selected virtual devices. Configurations in one virtual device do not affect configurations in other virtual devices. As a result, virtual partitioning provides an innovative way of protecting a set of services configured in several virtual devices from accidental mistakes, or from malicious configurations, made in another virtual device. A configuration failure on Cisco ACE is limited to the scope of the virtual device in which it was created. A failure in one virtual device has no effect on other virtual devices in the Cisco ACE, maximizing uptime for critical applications, especially when Cisco ACE is deployed in a redundant high-availability configuration.

Active-active redundancy: The Cisco ACE application switch supports a flexible active-active redundancy configuration between two physical devices. This setup allows you to distribute workload between both physical Cisco ACE devices rather than using only one of the devices in active mode.

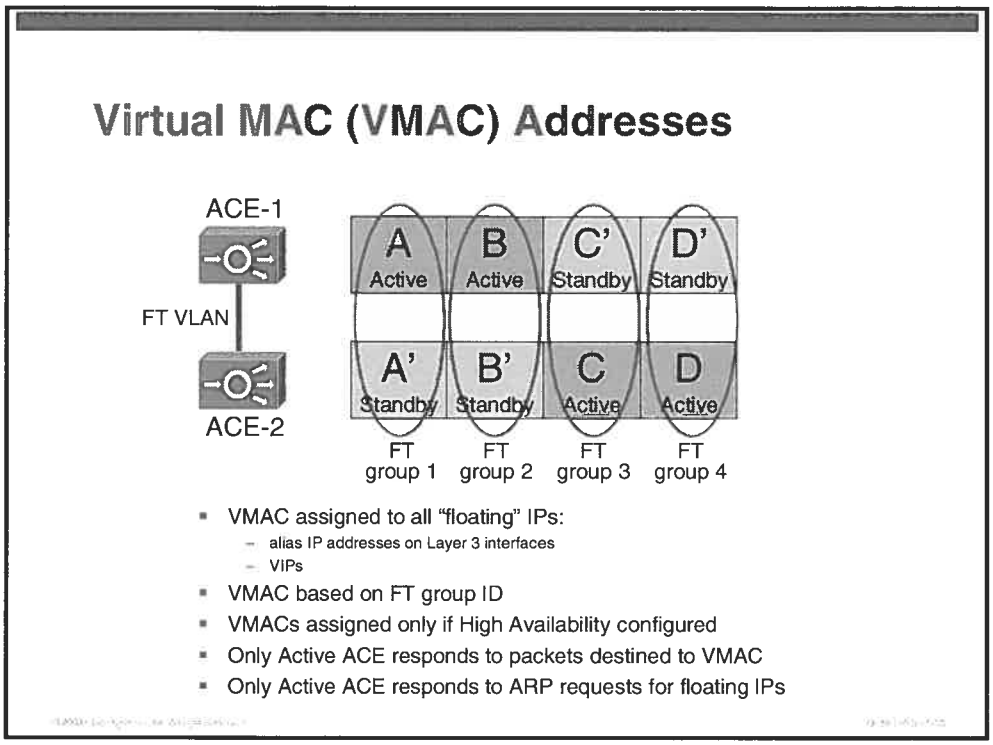
Stateful failover: The Cisco ACE application switch replicates connection flows on the active virtual device to the standby virtual device. The replicated connection flows contain all of the flow-state information necessary for the standby virtual device to take over the flow if the active virtual device becomes unresponsive. The former active virtual device transitions to a standby state to fully back up the current flows to the newly active virtual device. With this feature, end-user applications do not need to reconnect to maintain the same network session during the failover and failback. Thus, the Cisco ACE stateful failover and failback redundancy capability provides true transparent failover for applications regardless of device failure.

Tracking and failure detection: The Cisco ACE Application switch supports tracking and failure detection of several network items, including servers and VLANs, and performs *transparent switchover* from an active Cisco ACE physical or virtual device to a standby Cisco ACE device if the tracked network resources become unresponsive. All active flows that exist at the time of the switchover continue uninterrupted on the new active Cisco ACE device. When the failed tracked network device comes back up, the Cisco ACE application switch evaluates the priority between active and standby Cisco ACE devices. If the resulting priority of the standby Cisco ACE device is greater than the priority of the active Cisco ACE device, the Cisco ACE performs *transparent switchover* to the original active Cisco ACE device.

The Cisco ACE can be configured to track several network devices such as gateways or hosts, interfaces, and Hot Standby Router Protocol (HSRP) groups. For example, the Cisco ACE can track the HSRP group and perform transparent switchover from an active Cisco ACE device to a standby Cisco ACE device if the HSRP group is unresponsive. In this scenario, the ability of the Cisco ACE application switch to track and switch over reduces traffic on the Inter-Switch Link [ISL] between two distribution layer switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

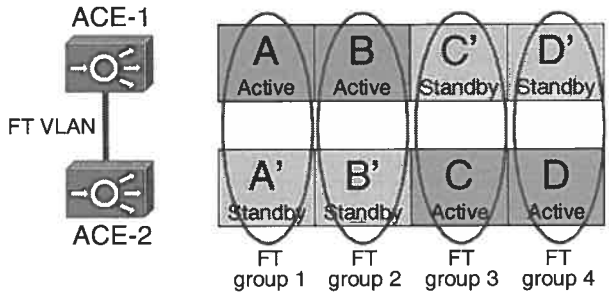


Virtual MAC (VMAC) addresses are assigned to IP addresses which will be moved from one Cisco ACE module to another in case of a failover. These IP addresses are often referred to as "floating" IP addresses and consist of all alias IP addresses configured on Layer 3 interfaces and all VIPs. The VMAC assigned is based on the FT group id which allows redundancy support for VLANs which are shared among multiple contexts. VMACs are assigned only if fault tolerance is configured.

The active Cisco ACE module will respond to packets that are destined to a VMAC and will answer ARP requests for the floating IP addresses. The standby Cisco ACE will block packets destined to a VMAC and does not respond to ARP requests for the floating IP addresses.

https://t.me/learningnets

Failover Actions



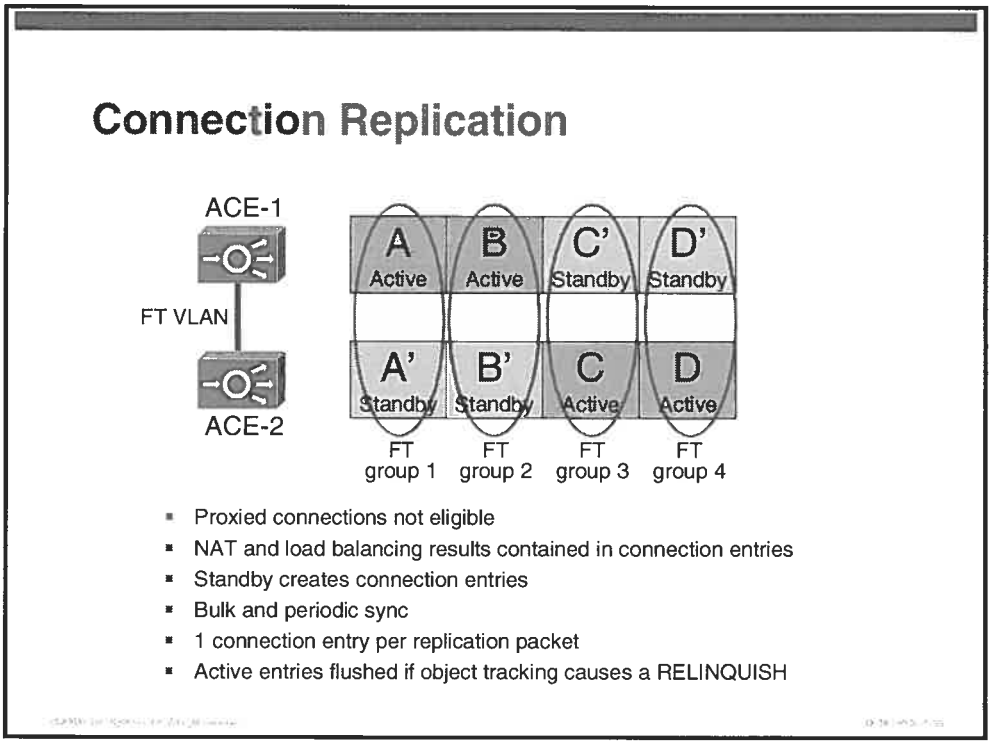
New active Cisco ACE module:

- Sends dummy multicast for every VLAN/VMAC pair
- Requests Layer 2 table purge on the supervisor
- Sends gratuitous ARPs for alias and VIP addresses
- Sends ARP request for gateway in bridged mode

© 2004, Cisco Systems, Inc. All rights reserved.

A failover of a Cisco ACE context to the redundant Cisco ACE module might require that switches attached to VLANs to which the Cisco ACE is connected move the MAC address to a different interface. To expedite these changes the Cisco ACE module takes several actions. A dummy multicast is sent out on every VLAN on which a VMAC has been assigned. This multicast packet uses the VMAC as the source MAC address. The Cisco ACE module also sends an out-of-band request to the supervisor requesting that the supervisor purge entries from its Layer 2 table. Gratuitous ARPs are generated for all floating IP addresses. If the Cisco ACE module has interfaces in bridged mode it will also generate an ARP request for the gateway IP address and bridge the ARP response.

https://t.me/learningnets

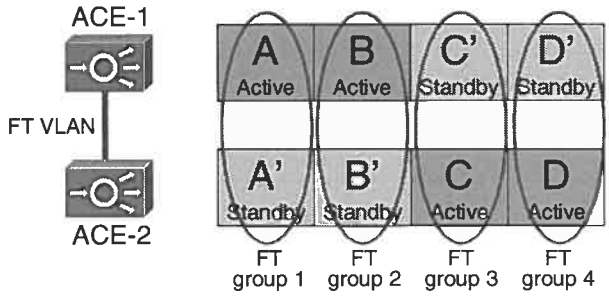


Connection Replication (CR) sends connection entries from the active Cisco ACE module to the standby. Connection entries for proxied connections are not eligible for replication. No special replication is necessary for NAT xlates or load balancing state information because all of the required information can be derived from the connection entries. The standby Cisco ACE module creates connection entries as it receives them via replication, giving it the full state information needed to continue processing after a failover. CR contains both Bulk and Periodic replication processes which both send 1 connection entry per replication packet.

If a Cisco ACE module transitions from active to standby as a result of object tracking, it will flush all of its connection entries and allow the new Active Cisco ACE module to bulk sync its connection entry table back to the original module.

https://t.me/learningnets

Connection Replication Statistics



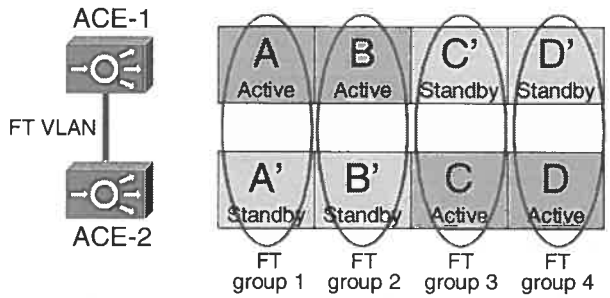
- show np <1 | 2> me-stats –sicm | inc replicate
- show conn
- show rserver
- show serverfarm

Connection replication statistics for each NP can be displayed with the **show np <1 | 2> me-stats –sicm** command. The results of connection replication can also be seen on the standby module through the use of the **show conn**, **show rserver**, and **show serverfarm** commands.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Sticky Replication

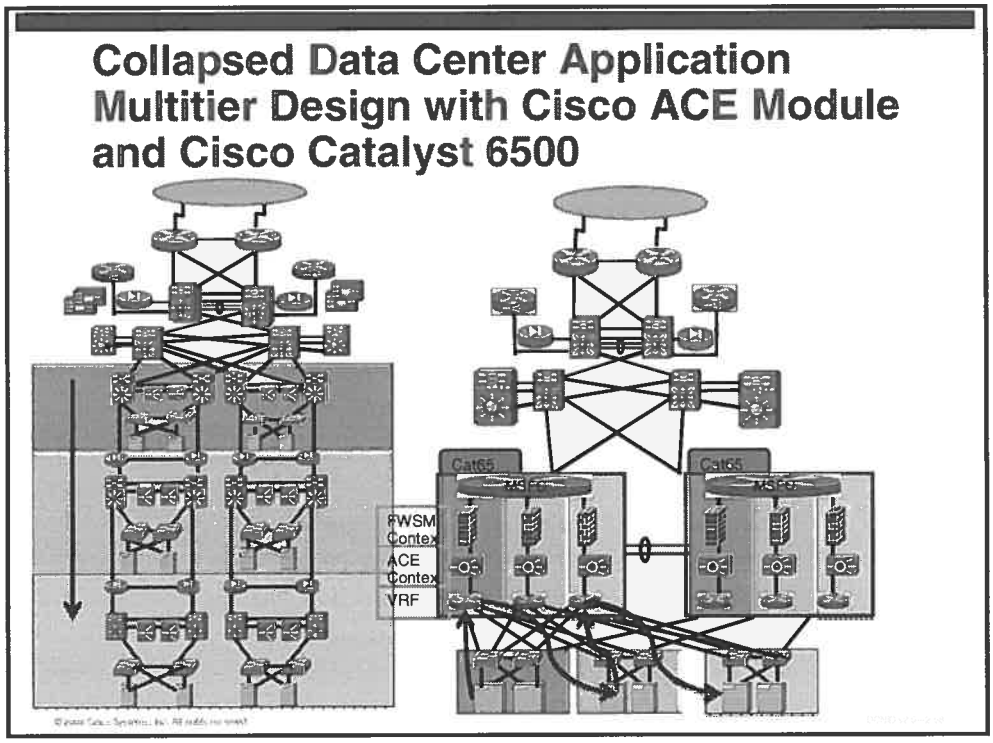


- Independent from connection replication
- Enabled in the sticky group configuration
- Bulk and periodic sync processes
- Multiple sticky entries per replication packet

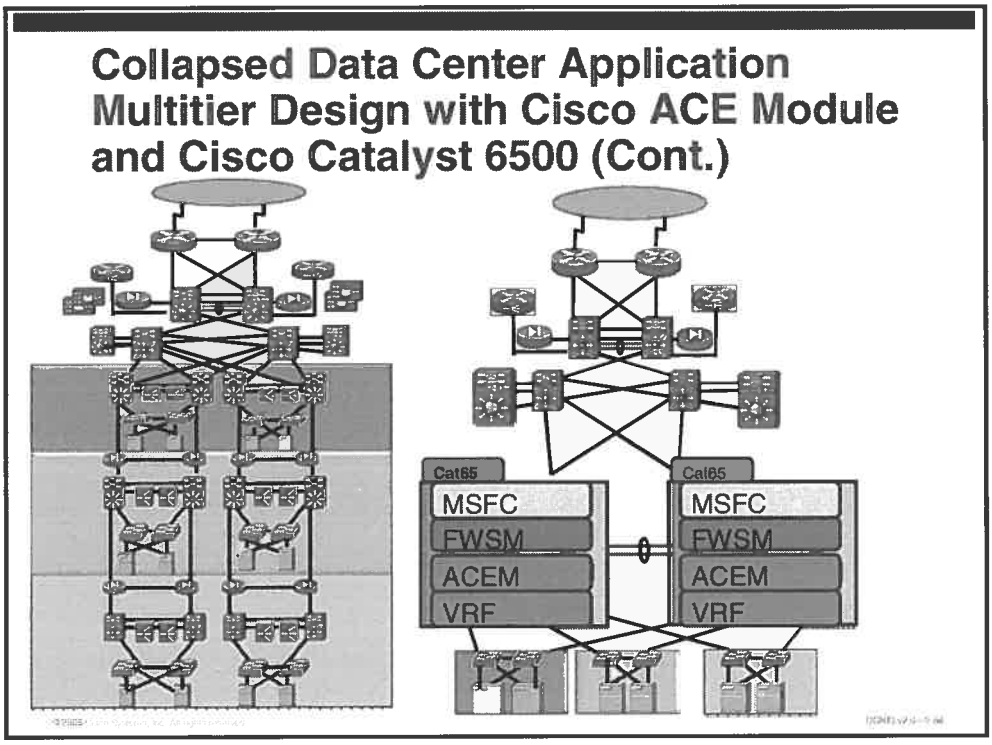
Replication of sticky database entries is handled by a separate process. Sticky database entries are also synchronized with both bulk and periodic sync processes. Replication of sticky database entries must be enabled in the sticky group configuration. Multiple sticky entries are transmitted in each sticky replication packet.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

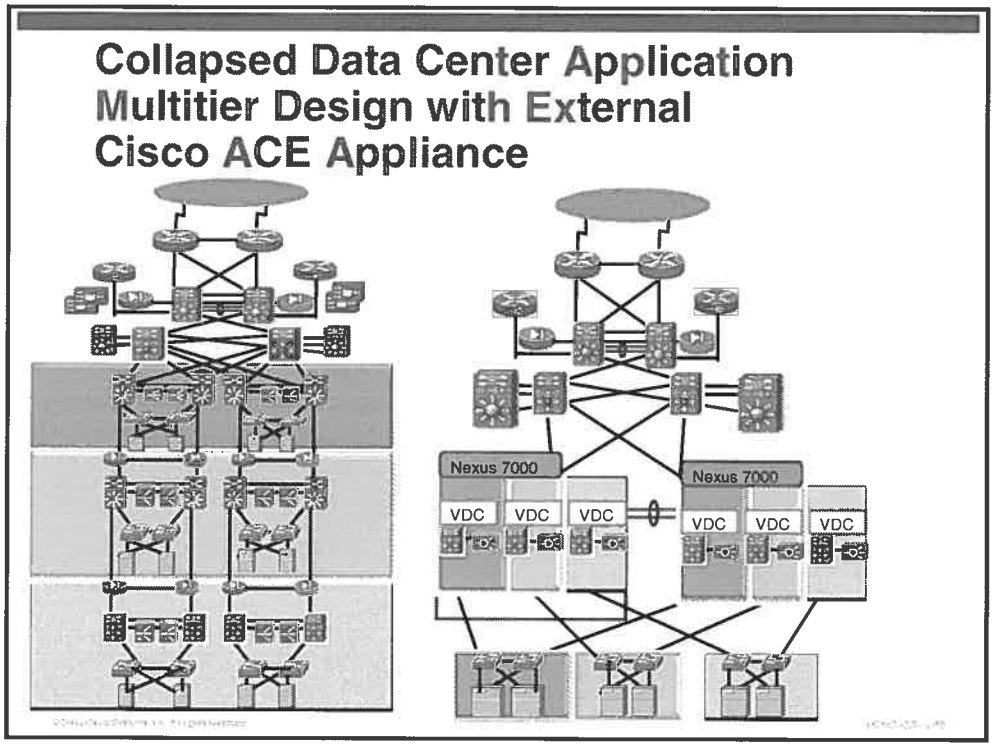


The 3-tier application networking infrastructure on the left requires many discrete appliances. The right side of the figure illustrates the benefits of consolidating the components on a common Catalyst 6500 Series switch with integrated FWSM and the Cisco ACE services module. Every application tier has a respective virtual Cisco ACE, virtual FWSM context, and VRF instance, consolidated on the common Catalyst 6500 Series switch. The integrated solution is easier to manage and less expensive to maintain and support.



© 2008 Cisco Systems, Inc.

https://t.me/learningnets



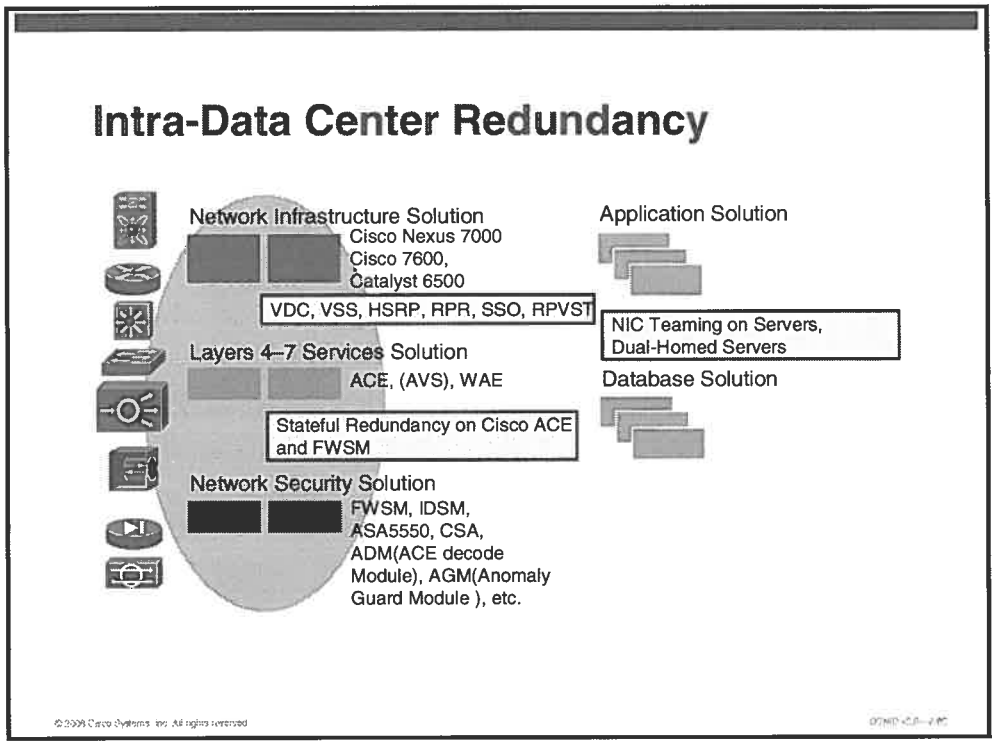
All three application tier networking infrastructure on the left are now consolidated on common Catalyst 6500 and Nexus 7000 Series switches with external Cisco ACE appliances operating in one-arm mode. Each application tier has an associated virtual Cisco ACE context, virtual FWSM context and VRF instance or VDC instance consolidated on the common Catalyst 6500 Series switch or Nexus 7000 series switch. For High Availability (HA), both active-active or active-standby designs are possible.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Intra-Data Center Redundancy

This topic describes the redundancy design requirements within the data center.



The figure describes intra-data center redundancy and high availability options.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

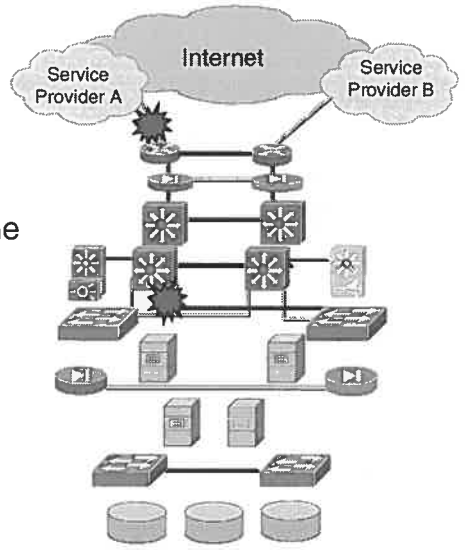
Network Failures

ISP failure:

- Dual ISP connections
- Multiple ISP

Connection failure within the

- Etherchannel, ECMP
- Multiple route paths



The figure illustrates network failures at Layers 1 through 3.

There are many different network connections in a network to consider:

- Connections to ISPs
- Connections within and between subnetworks
- Route convergence
- Network saturation

There are also many types of networks in a data center:

- Client access
- Management
- Device heartbeat
- Data replication

Network layer redundancy consideration:

- **Multiple ISP with dual connections**: client access public network
- **Inter-device network**: private heartbeat and data transfer network
- **Redundancy built into network**: Etherchannel, multiple route paths
- Routing protocol tuning, broadcast storm prevention, spanning-tree design, QoS

https://t.me/learningnets

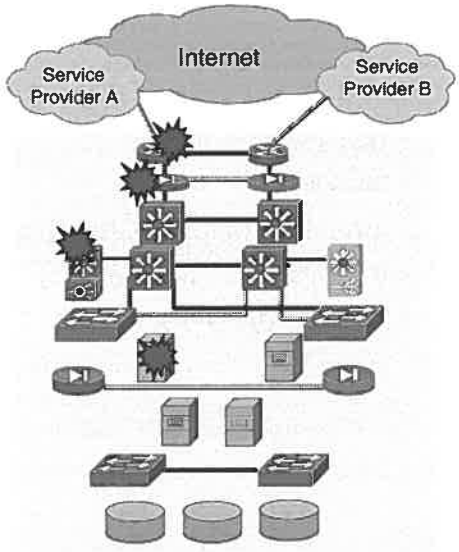
Device Failures

Routers, Switches, Content Switches, FWs:

- HSRP
- VRRP

Hosts:

- HA cluster
- LB server farm
- NIC teaming
- VSS
- VDC



Possible Physical Failures

Network device:

- Routers
- Switches
- Hubs
- Ports
- Load balancers
- Application firewalls

Host device:

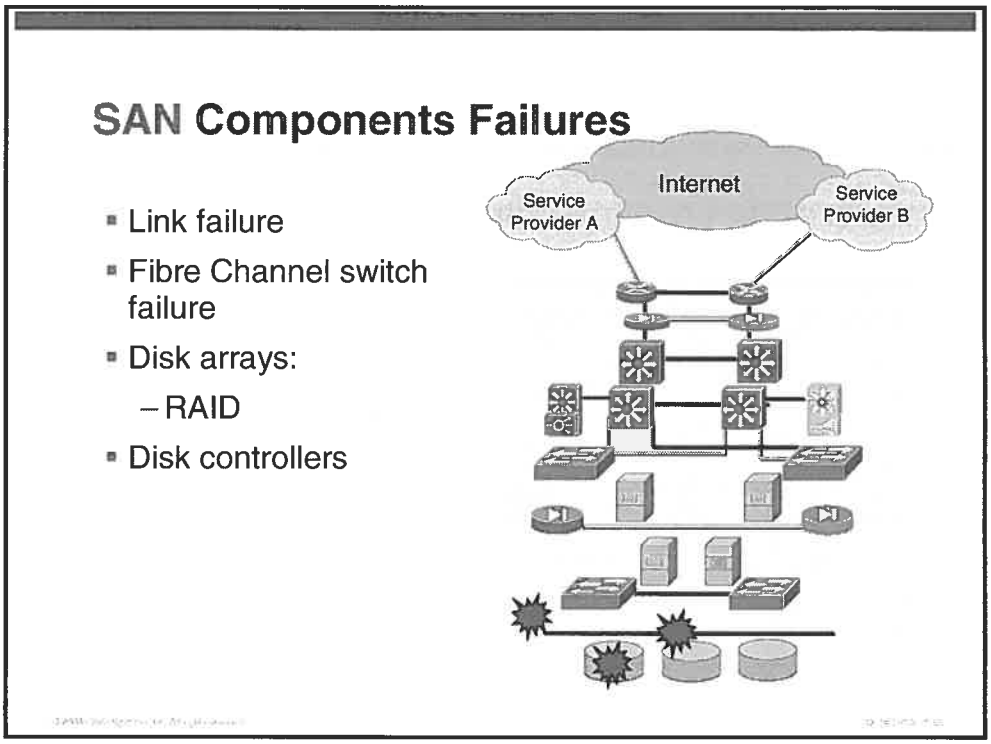
- Server hardware
- NICs
- Hard drives
- Processors
- Server software
- Application failure
- OS failure or vulnerability

Device redundancy considerations:

- **Routers**: HSRP
- **Firewall and SLB**: Active-active, active-standby redundancy with VRRP
- **Host**: High-availability cluster

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure illustrates locations for failures in a SAN.

Potential SAN failures:

- Disk arrays
- Controllers
- Host bus adapter (HBA)

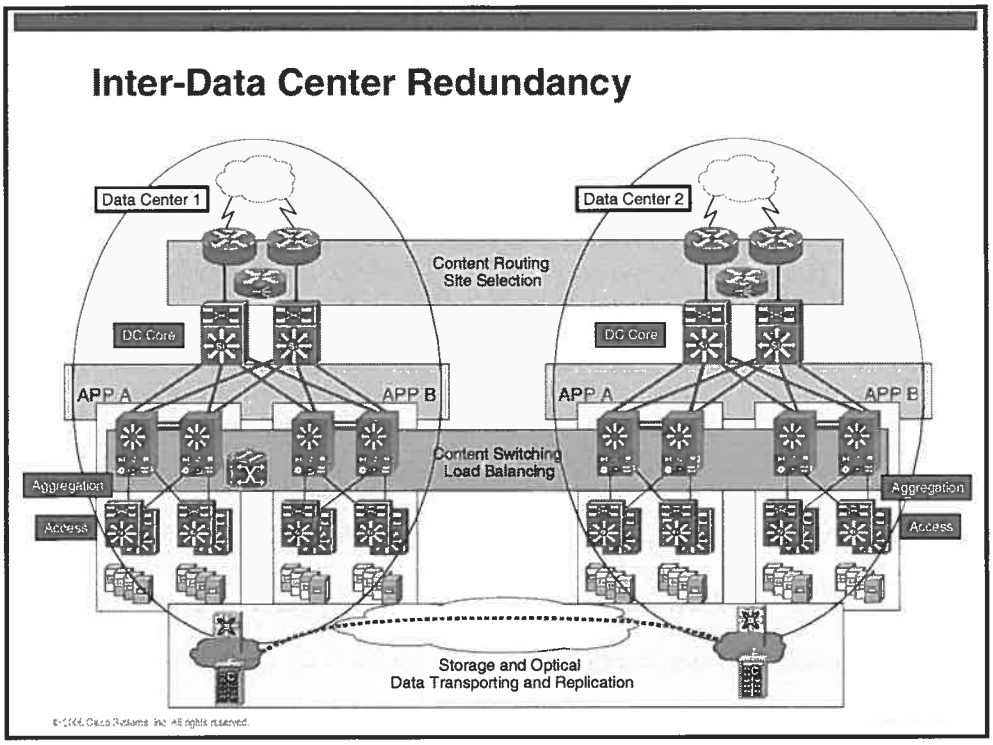
SAN considerations to provide fault tolerance:

- Redundant Array of Independent Disks (RAID) for protection of the drives
- Redundant storage frames with synchronize replication (mirroring)
- Hardware and software solutions from different vendors to allow devices, and components in the storage controller itself, to be swapped out as necessary without losing access to the data itself

https://t.me/learningnets

Multi-Data Center Redundancy

This topic describes the design challenges associated with multi-data center redundancy.



The figure illustrates the advantages of multiple data centers.

What is a distributed data center?

- Multiple geographically located data centers providing the same or similar service function for end users around the world, through different ISP connections, is the optional internal IP network.

Advantages of distributed data centers:

- Avoid another type of single point of failure (after redundant ISP connections, router HSRP and server farm).
- High-availability business continuity is the ability of a business to adapt, change, and continue with its service when confronted with various outside impacts, mitigating the impact of a disaster.
- You provide not only disaster recovery in the case of an unfortunate event, but also load balancing of all your resources, and overall better use of global resources.
- Scalability is increased application, user base, traffic amount.
- For everyday business, providing users with more efficient traffic flow and better response time when compared with a single data center serving global users.
- User proximity versus data center proximity.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Other non-network-related failures:

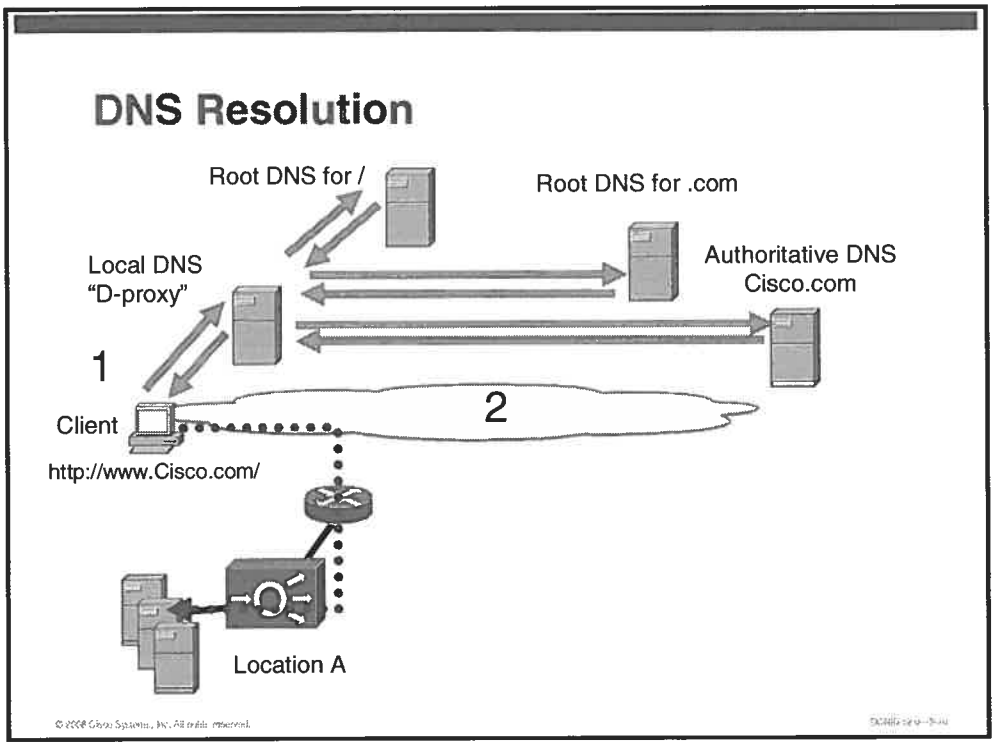
- Natural disaster
- Power outage
- Tornado
- Hurricane
- Earthquake
- Flooding
- Fire
- Terrorism
- War

https://t.me/learningnets

GSLB Overview

The objective of this topic is to describe global server load balancing (GSLB).

Resolving a DNS Query



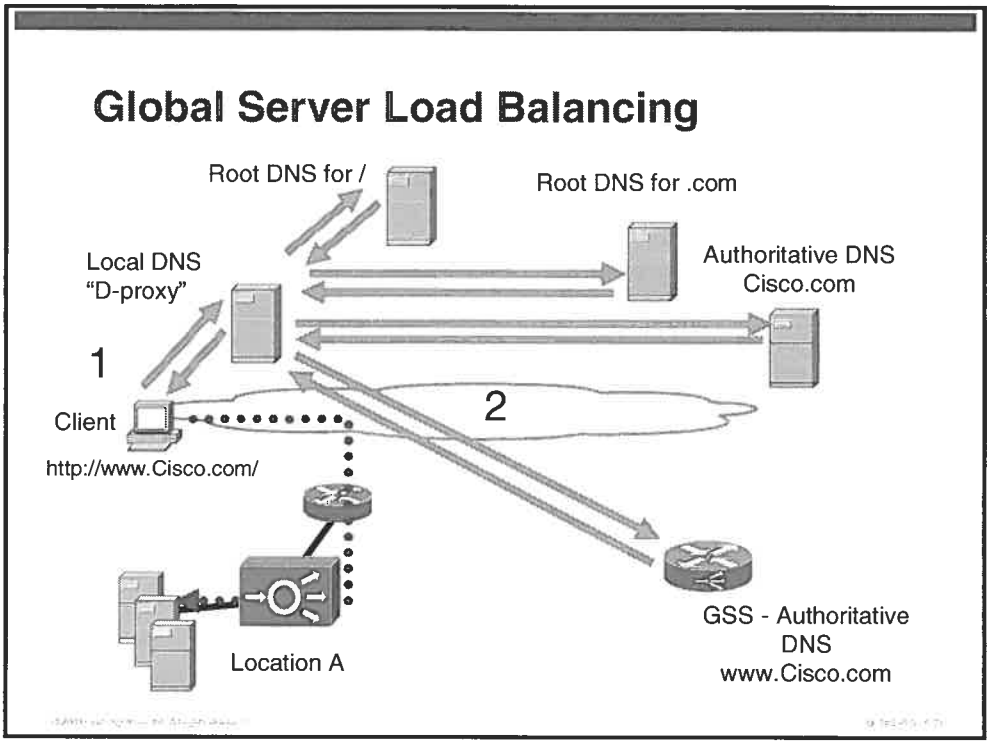
The following steps summarize the process of resolving a simple DNS query:

1. The client sends a request to their local DNS server. The local DNS server is sometimes referred to as the DNS proxy.
2. The DNS proxy uses a hierarchy of DNS servers to find the IP address requested.
3. The response to the query is ultimately returned to the DNS Proxy with either a single IP or a statically ordered list of possible addresses, at which point it is now cached. It is now the DNS Proxy who replies to the clients with the cached information.
4. The client now chooses the destination IP from the ordered list. The client proceeds sequentially trying the next if the current destination is unavailable.

The large dotted line in the preceding figure shows the resulting traffic flow.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



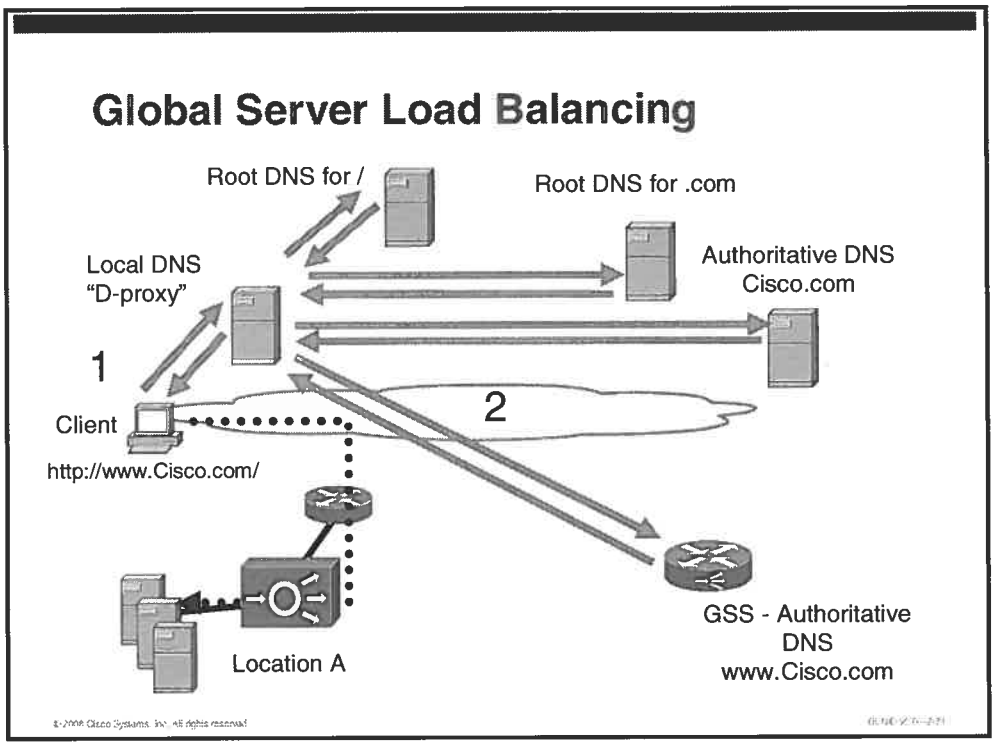
As the figure illustrates, the following steps summarize the process of resolving a DNS query when GSLB is used:

1. The client sends a request to their local DNS server. The local DNS server is sometimes referred to as the DNS proxy.
2. The DNS proxy uses a hierarchy of DNS servers to find the IP address requested; eventually, a request is sent to the GSS. In the figure, the GSLB function has responded with the address of the virtual IP (VIP) at location A, resulting in the client connection.
3. With GSLB the response to the query is ultimately returned to the DNS Proxy with either a single IP or a dynamically ordered list of possible addresses. This allows the response to vary, creating a distribution of traffic flow. It is now the DNS Proxy who replies to the clients with the cached information. DNS proxies usually cache DNS results for a period of time. This time-to-live (TTL) is returned with every DNS response. The GSS is configured to return DNS records with a short TTL. During this TTL interval, any connections from clients that use the same DNS proxy will use the cached destination.

The large dotted line in the figure shows the resulting traffic flow.

https://t.me/learningnets

Server Load Balancing vs. Global Server Load Balancing



What is the difference between SLB and GSLB?

Server load balancing (SLB) provides the company with the means to load balance between servers, serving same content at the same site.

Global server load balancing (GSLB) provides the company with the means to load balance client requests between two data centers serving the same content, and then to load balance within that site, between servers, serving the same content.

By definition, GSLB balances user requests to available SLBs (VIPs) or hosts that are hosted at different locations.

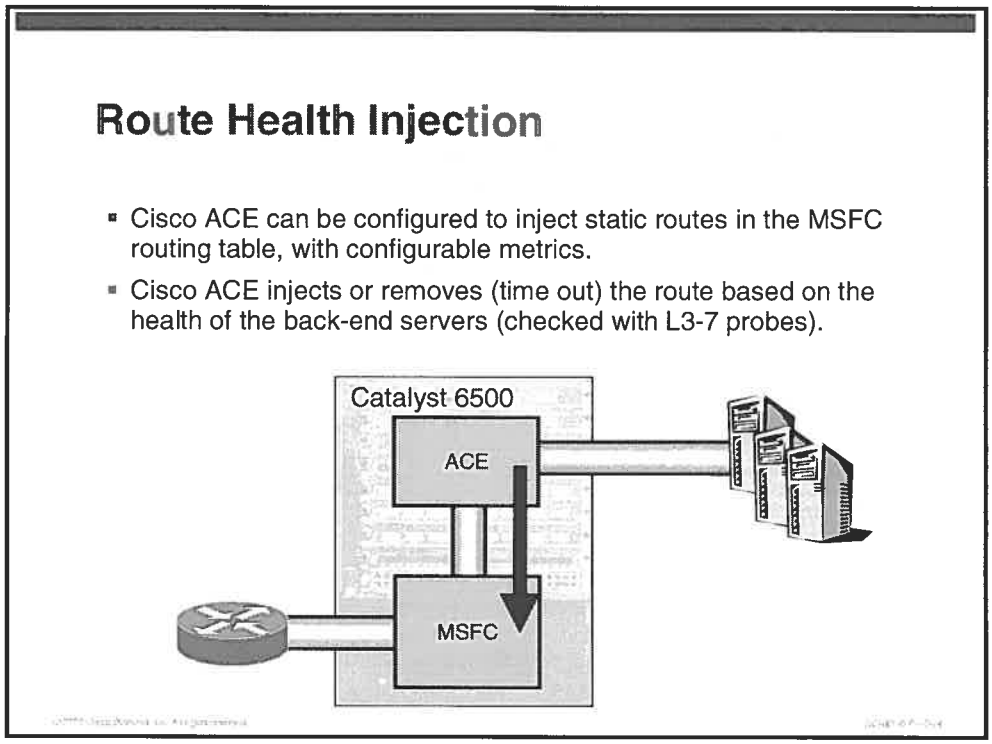
Note	Typically, the locations are geographically dispersed and GSLB is used either in a disaster recovery (DR) design or a site load balancing design where users are directed to different locations based on a specific request load balancing algorithm, or by a proximity discovery method.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Route Health Injection

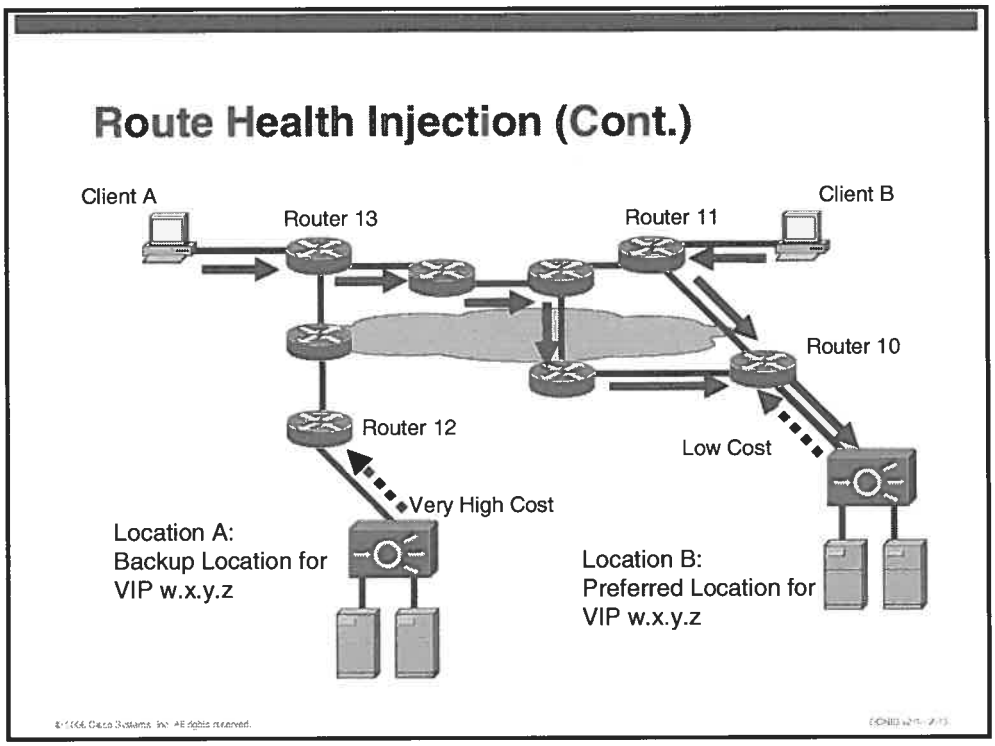
This topic explains Route Health Injection.



Route health injection (RHI) gives the Cisco ACE module the capability to inject static host routes into the routing table in the base IOS on the Cisco Catalyst 6500 Series chassis. These advertisements are sent out-of-band from the Cisco ACE module directly to the Catalyst 6500 Series Supervisor. The IOS image on the supervisor then takes the information from the RHI advertisement and creates a static route in its routing table. Both the Cisco ACE module and the IOS are VPN routing and forwarding (VRF) aware, and the routes advertised by RHI can therefore be put into appropriate VRF routing tables. On the Catalyst 6500 Series, this requires IOS SXF5 or later.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

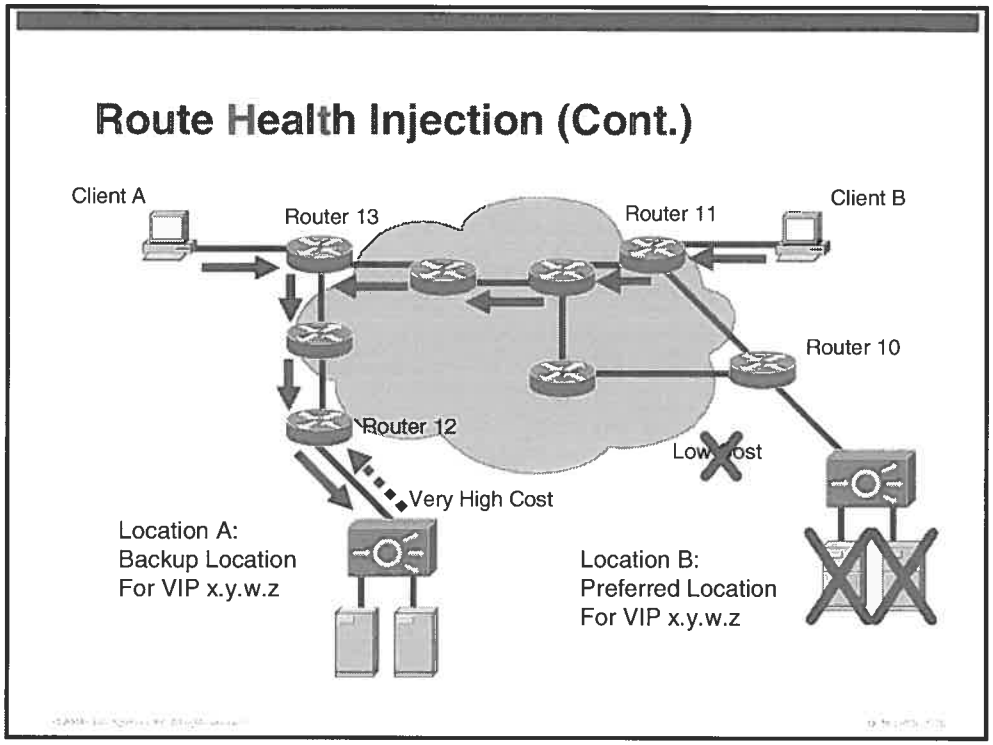


The route health injection feature gives the Cisco ACE the ability to advertise host routes to the Multilayer Switch Feature Card (MSFC) acting as the upstream router. These routes can then be propagated throughout the enterprise using any available routing protocol.

The network diagram illustrates the use of RHI to provide access to a redundant site. Location B is the preferred location for traffic to a particular VIP. RHI advertises, with low cost, the VIP from Location B. Location A announces itself as a backup site by also using RHI to advertise the same VIP. Location A uses a high enough cost for this advertisement that traffic goes to Location B first.

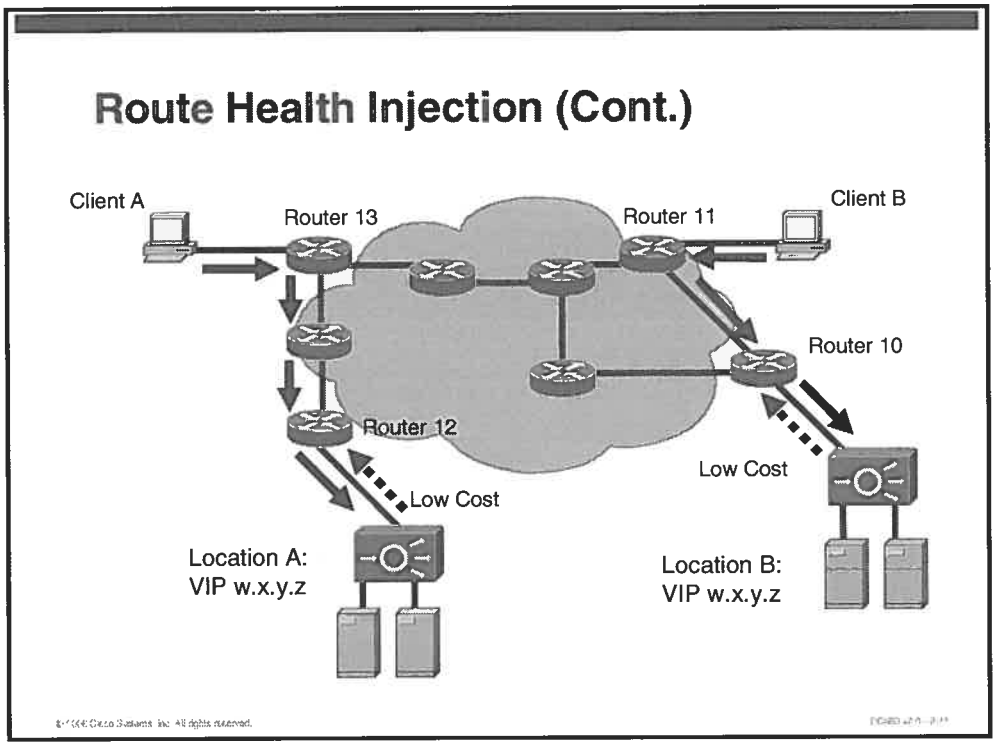
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The network diagram shows how RHI handles a failure of the primary site. Clients are able to reach the backup VIP at Location A as soon as the routing protocols involved in the network converge. In general, this convergence happens very quickly. Note that connections that were in progress to Location B when Location B failed are lost; however Location A accepts new connections very quickly.

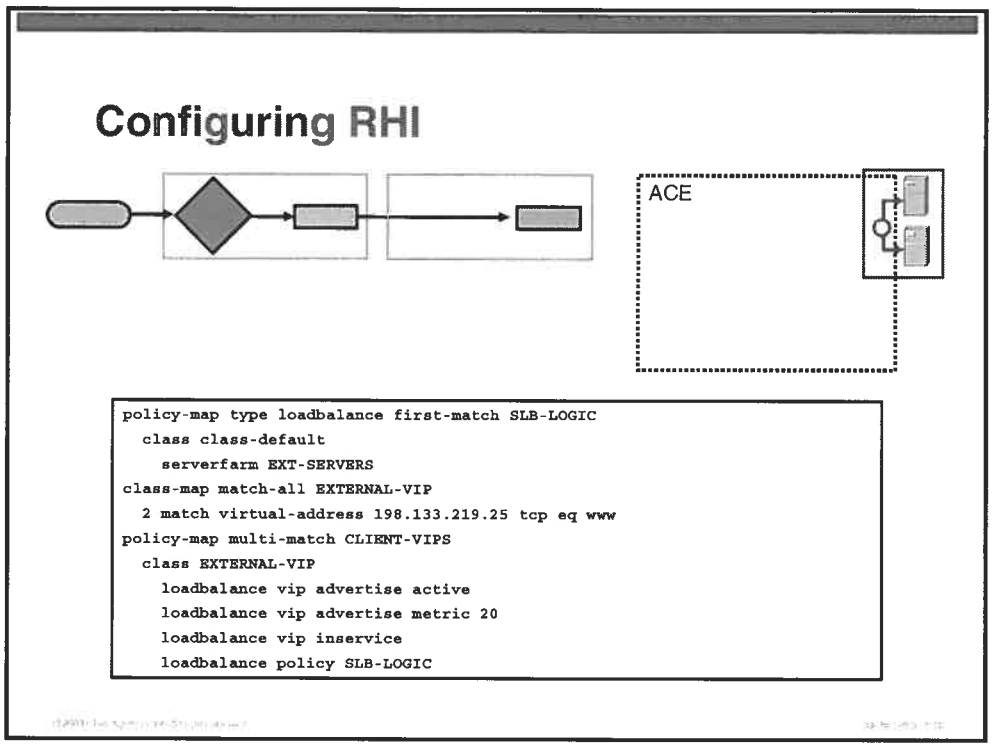
https://t.me/learningnets



RHI is also used to provide load-balanced service based on the proximity of the client to one of the server farms. The network diagram shows both locations advertising the same VIP via RHI. Routing functions in the network direct client requests to the server farm that is the closest to the client. In the case of a failure of either location, the routing protocols in the network quickly converge and the remaining location receives all client requests.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

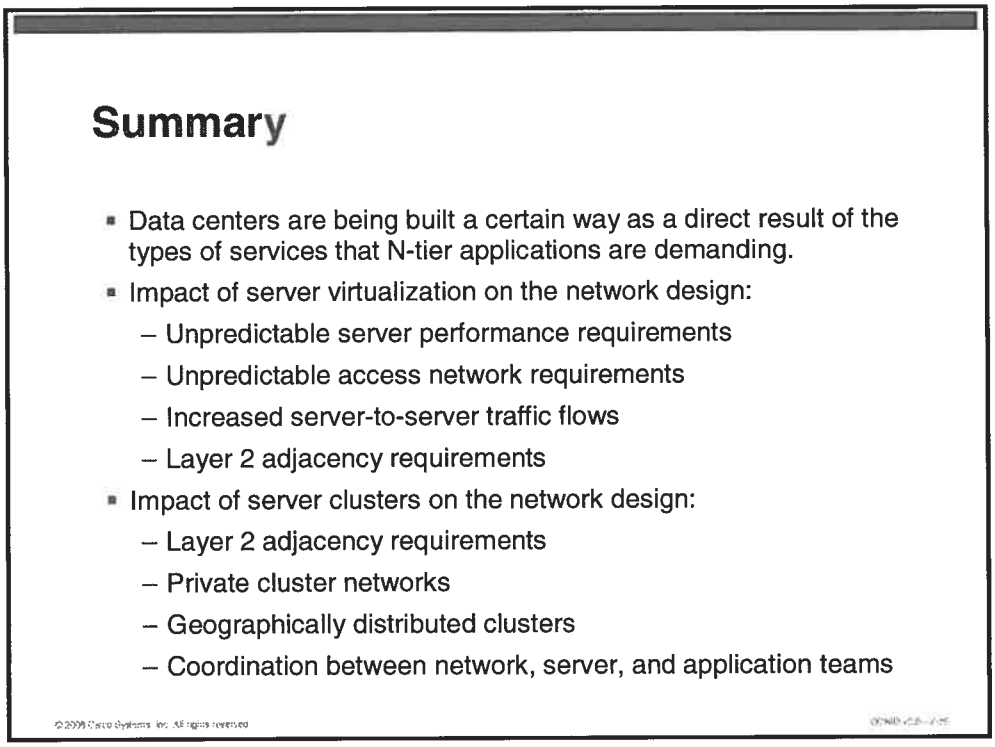


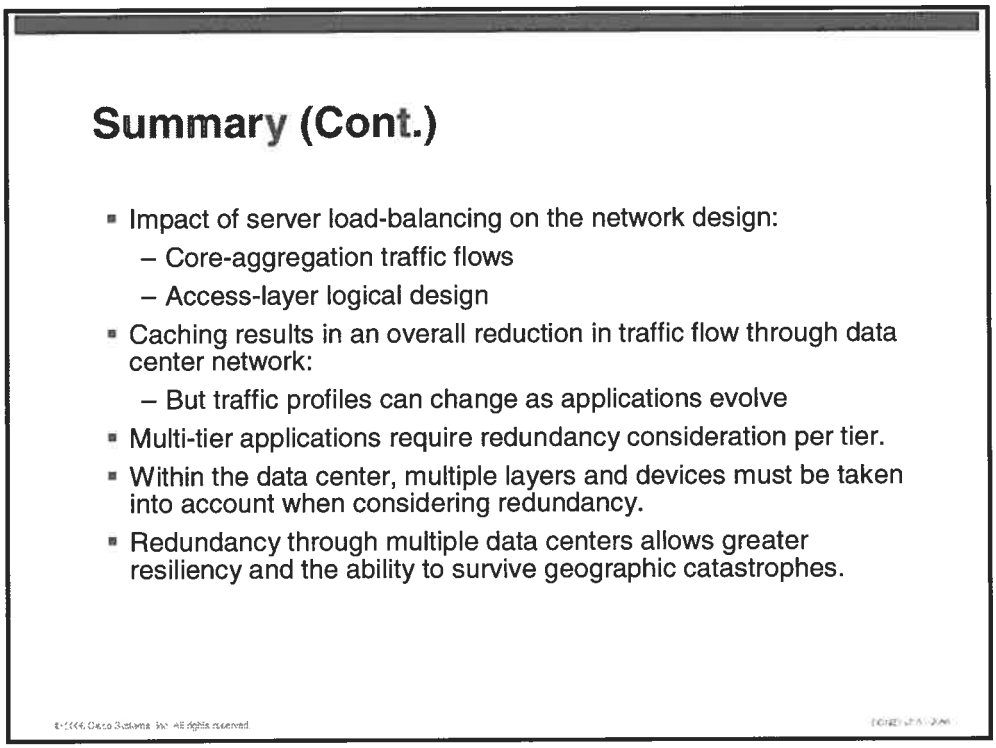
RHI is configured by adding the **loadbalance vip advertise** [*active*] as an action statement in the Layers 3 and 4 policy-map. The **active** keyword is an optional keyword. If **active** is specified, the route is advertised only if there is at least one active server in the server farm. Without the **active** keyword, the Cisco ACE module will advertise the route at all times. The metric advertised with the route can be specified with the **loadbalance vip metric** *metric_number* statement.

https://t.me/learningnets

Summary

This topic summarizes the key points that were discussed in this lesson.





© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Appendix A

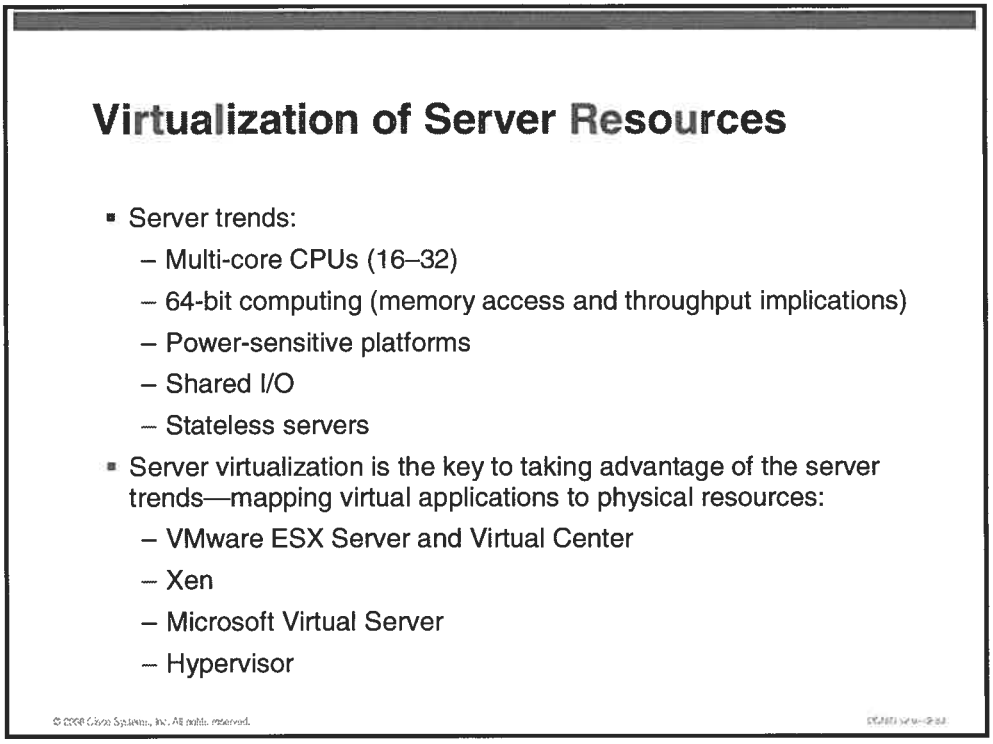
The appendix topics for Module 4 Lesson 2 include:

- Server Clusters
- Cluster Types
- Cluster Interconnects
- Server Farms and Load Balancing
- WAN Optimization and Branch Consolidation

https://t.me/learningnets

Server Virtualization

This topic discusses the concept of server virtualization and the impact of virtualization on the network design.



Trends in data center server deployments include:

- Multi-core CPUs (16–32)
- 64-bit computing (memory access and throughput implications)
- Power-sensitive platforms
- Shared I/O
- Stateless servers

Server virtualization is the key to taking advantage of the server trends—mapping virtual applications to physical resources. The following are examples of virtual server products:

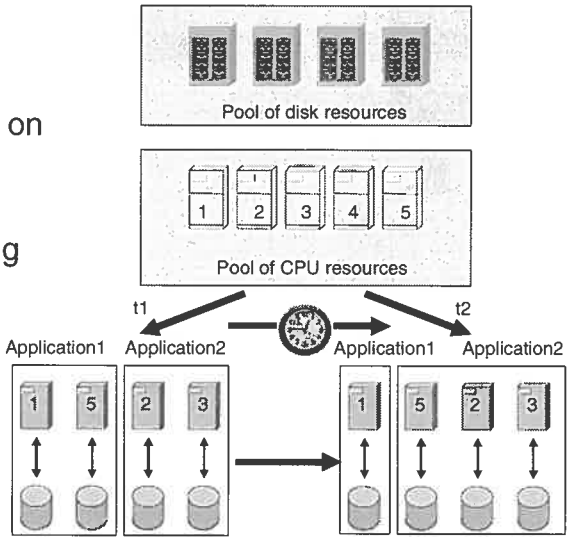
- VMware ESX Server and Virtual Center
- Xen
- Microsoft Virtual Server
- Hypervisor

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Utility Computing: Dynamic Provisioning

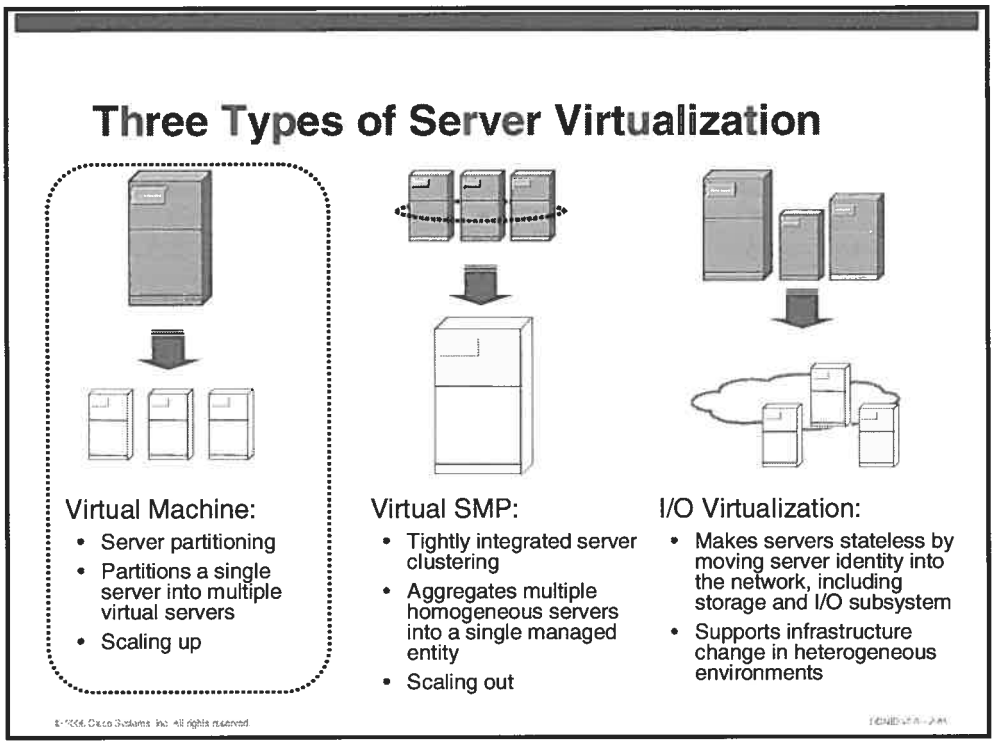
- Pools of hardware resources
- Provisioned based on user demand
- Requires:
  - Resource pooling
  - Virtualization
  - Dynamic provisioning



Utility computing refers to a pool of hardware resources that can be provisioned based on user demand. To add capacity to a given application on demand requires:

- Resource pooling
- Virtualization
- Dynamic provisioning

https://t.me/learningnets



There are three basic types of server virtualization:

- With the virtual machine (VM) approach, a single large server is partitioned into multiple virtual servers. This approach is the oldest, and is still in wide use today on IBM mainframe systems as well as open-systems platforms. This is a scaling up approach to compute power scalability. Examples include VMWare, XEN, and Microsoft Virtual Server. The main value of this type of solution is higher server utilization—scaling up.
- The virtual SMP approach combines servers together into a single managed entity. This approach is a tightly integrated cluster of homogeneous servers that are managed as a single virtual entity. This is a scaling out approach to compute power scalability. Examples include Virtual Iron and clusters. The main value of this approach is scaling mission critical applications on commodity hardware—scaling out.
- I/O virtualization is the most recent approach to server virtualization. In this approach, the server identity is moved onto the network, effectively making the servers stateless. The virtualization is managed by a server fabric switch that sits between the servers and the network. One of the key advantages of this approach is that it works with heterogeneous commodity servers. This is the approach that is used by the Cisco Server Fabric Switch and VFrame products. The main value of this solution is simplifying infrastructure change in heterogeneous environments.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

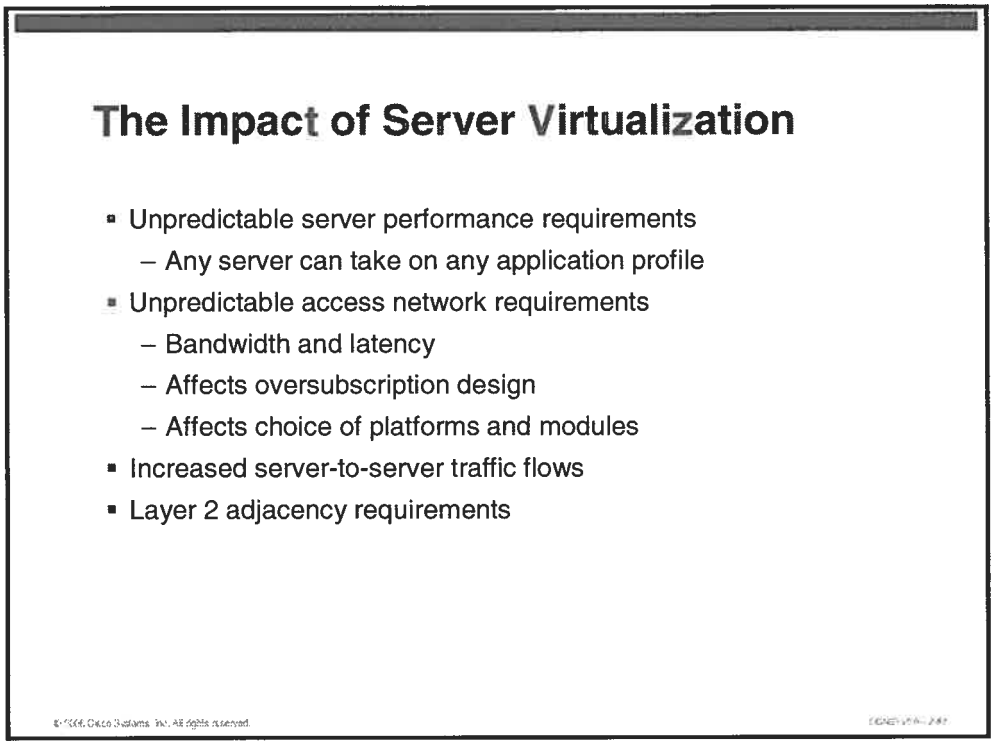
Virtual Machines

Example products for x86 architecture:

- VMware ESX Server
- Microsoft Virtual Server
- XEN open source project

Software products such as VMware ESX server, Microsoft Virtual Server, and the open source project known as XEN allow virtualization of the x86 platforms to varying degrees.

https://t.me/learningnets



The deployment of server virtualization can impact the network design in several ways:

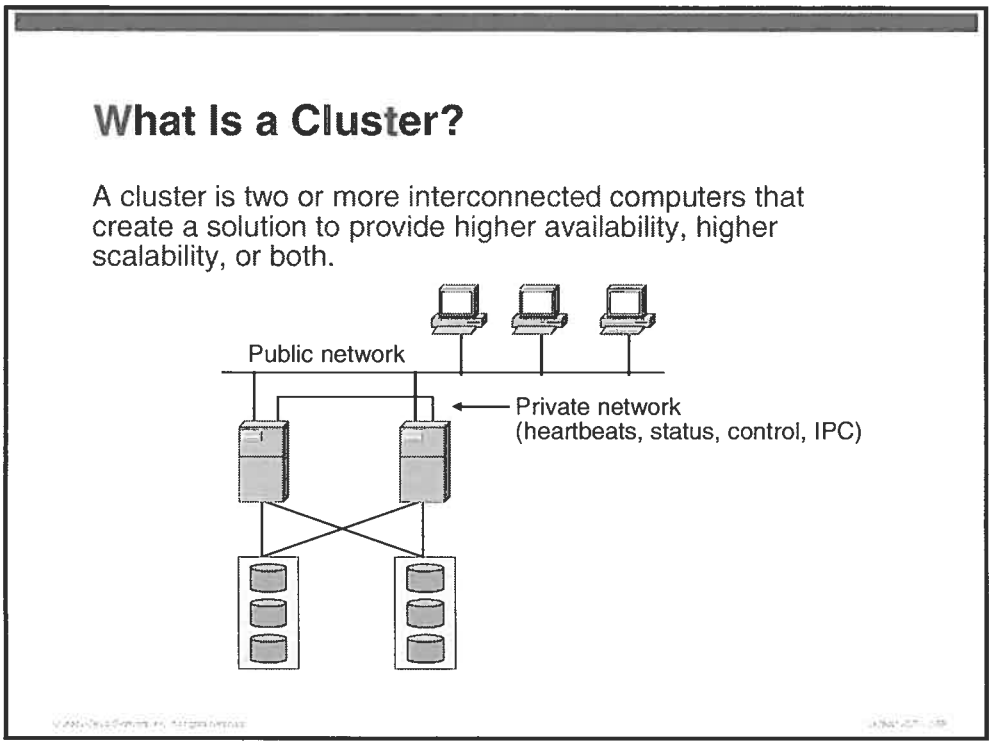
- **Unpredictable server performance requirements:** Any server can take on any application profile at any time.
- **Unpredictable access network requirements:** As applications move from server to server, the bandwidth and latency that are required by each server changes. Traffic profiles can change unpredictably. The unpredictable nature of virtual server traffic profiles affects oversubscription design and the choice of platforms and modules.
- **Increased server-to-server traffic flows:** Techniques for migrating virtual machines, such as VMWare ESX VMotion, can cause spikes of server-to-server traffic, which is not usually seen in a traditional access layer.
- **Layer 2 adjacency requirements:** Virtual machines have specific Layer 2 adjacency requirements.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Server Clusters

This topic discusses the basic characteristics of server clusters and the impact of clustering on the network design.



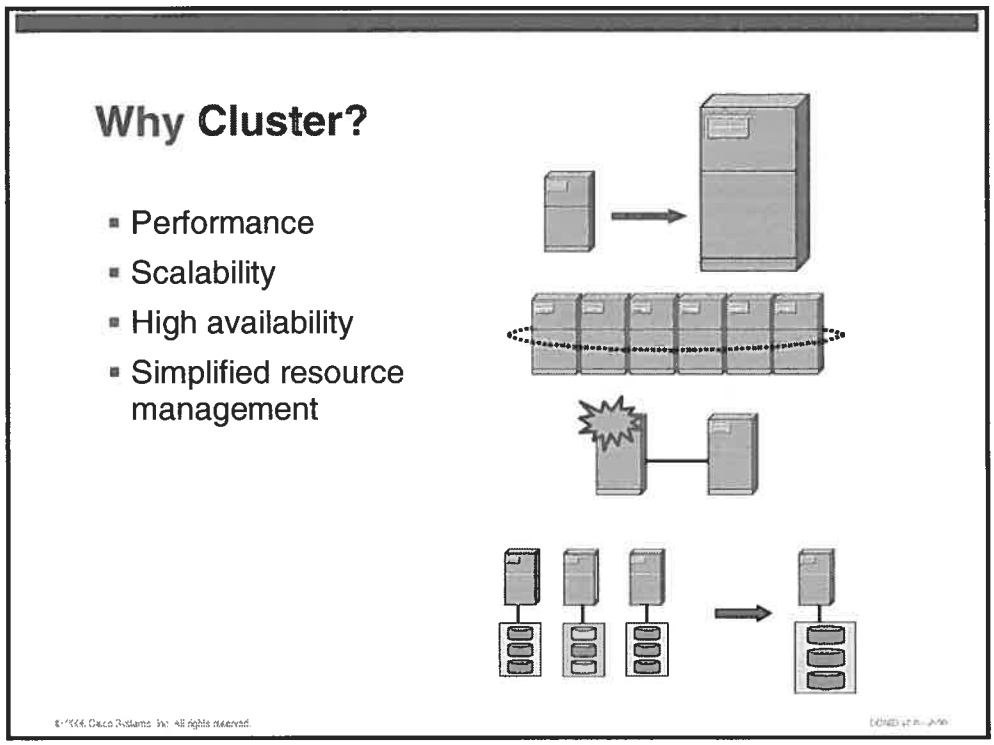
A cluster consists of two or more interconnected computers that create a solution to provide higher availability, higher scalability, or both. The main goal of a computer cluster is to provide scalable computational power to applications by utilizing multiple commodity computers (up to thousands of nodes, or servers), which are networked together to function as a single, large monolithic computer.

The common goal of server clusters is to combine multiple physical servers to appear as a unified system through special software and network interconnects to achieve improved load distribution and higher performance.

In addition to public network connections, cluster nodes typically also connect to a private network. The private network is used for intra-cluster communications, such as heartbeats, status and control messages, and inter-process communication (IPC). Because the bandwidth and latency requirements for a private network can be high, the private network is often a physically separate network, so that traffic on each network does not interfere with the other network.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Scalability

As applications grow, clustering greatly helps to scale them non-disruptively. In the past, when an application that is used a single server and then outgrew its capability, the application was ported to a larger server that could handle the load. This practice is both disruptive, and in a datacenter with hundreds of applications, nearly impossible to manage. Using clusters, you simply add more computing resources to the existing cluster and rebalance the load.

Performance

Enterprise developers are requesting network staff to provide predictable bandwidth with low latency for both custom and off-the-shelf applications. Developers are also turning to clusters because the processing power of clusters has a direct impact on time-to-market for many applications. For example:

- The time that it takes to have large animated film segments distributed on a network, rendered and returned to the artists, directly influences the time that it takes for the film to get to the market.
- Financial trending analysis can require a large CPU compute capacity. How quickly a financial firm can determine a particular trend provides a huge competitive advantage for the firm and its investment clients.

Predictable bandwidth with low latency enhances performance of both custom and off-the-shelf applications. For example:

- Database performance (lock times, parallel scans)
- Financial trending analysis (large computations)
- Animation rendering (large computations)
- Search engines (fast parallel lookup + content insertion)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

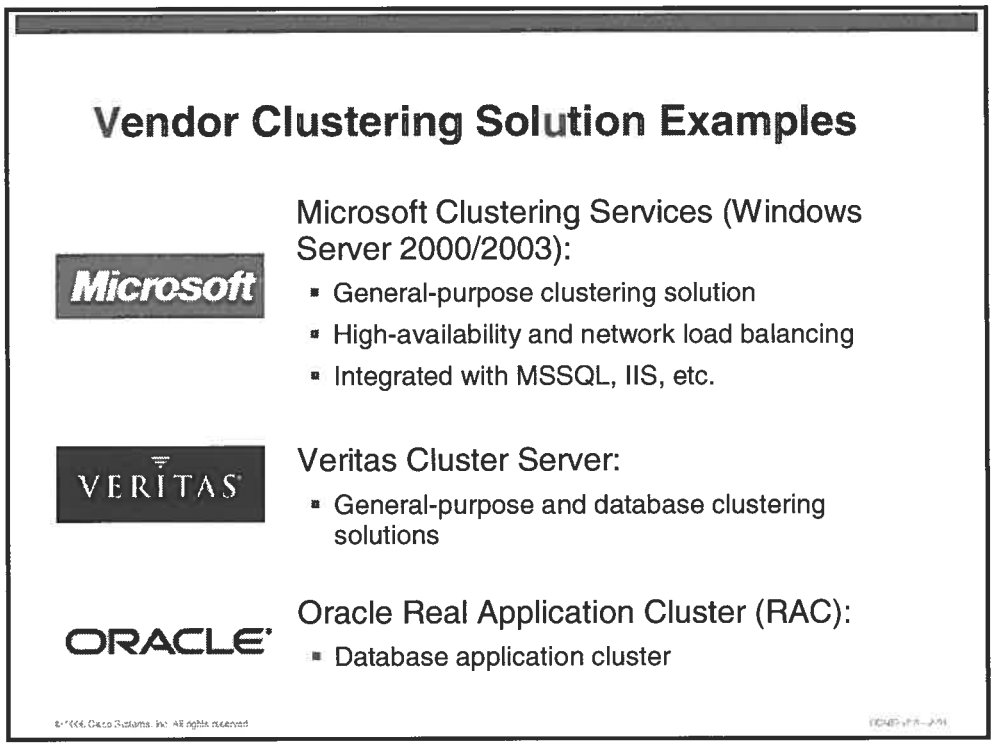
High Availability

Many critical applications cannot afford to go down. Clustering provides a simple way to ensure application uptime. This can be provided by either Active-Active or Active-Passive deployments. A main advantage of using clustering for high availability is that it can take the place of other expensive resiliency strategies. For instance, to ensure uptime, servers are often dual-connected to their Ethernet switches as well as their storage resources. This requires large servers with multiple network connections and expensive components, such as two HBAs, for the SAN connection. With clustering, you could deploy low-cost, single connection servers using Internet Small Computer Systems Interface (iSCSI) and have failover occur at the server level.

Simplifying Resources

Some applications require scalable computing power, but also need to draw upon similar storage resources. You could duplicate each server's storage for every other server, but this would waste storage resources. One solution is to cluster servers together and use a common file system or SAN file system to allow servers to share common storage access.

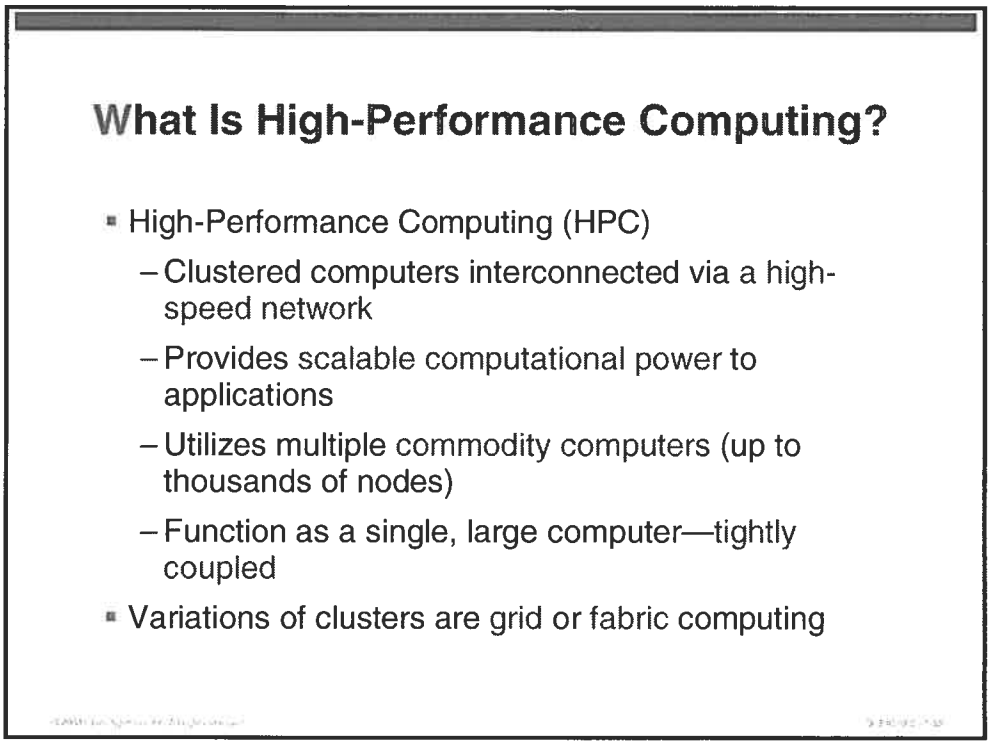
https://t.me/learningnets



There are many cluster vendors and solutions available for a variety of purposes. This slide shows a small sample of the types of clusters on the market.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



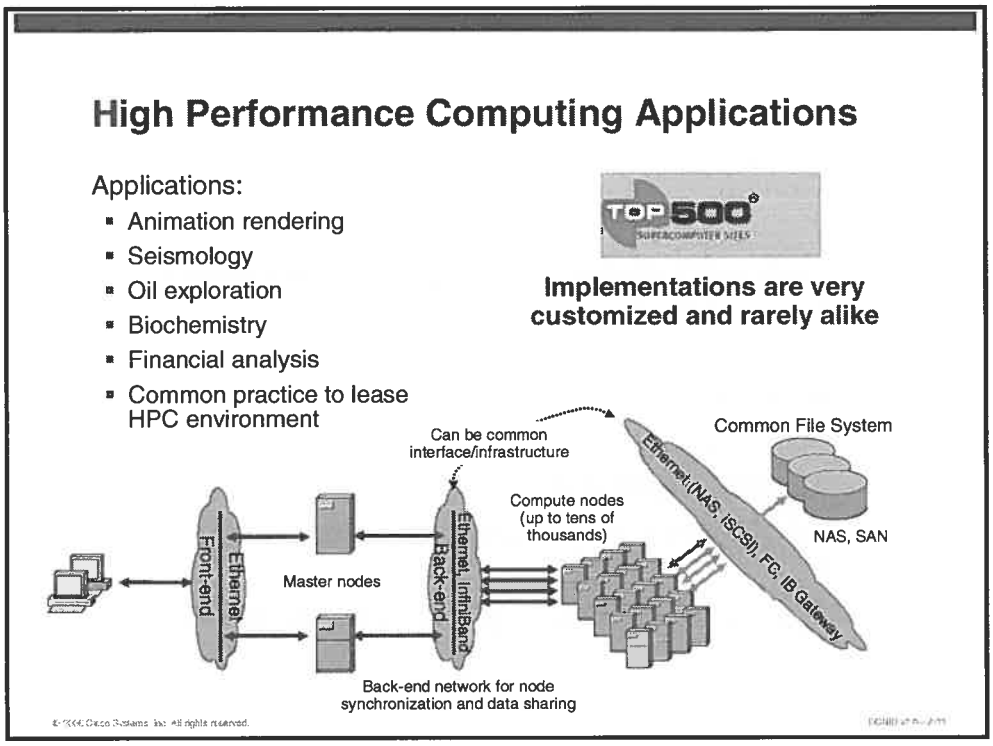
The term High-Performance Computing (HPC) is used to describe a type of clustering where the clustered computers are interconnected via a high-speed network. The main goal of a cluster is to provide scalable computational power to applications. Variations of clusters include grid or fabric computing.

Measuring Performance

In computing, the term FLOPS is an abbreviation of Floating point Operations per Second. This is used as a measure of a computer's performance, especially in scientific calculations that make heavy use of floating point calculations. Computing devices exhibit an enormous range of performance levels in floating-point applications, so performance is often measured in units such as megaFLOPS (MFLOPS, approximately 1,000,000 FLOPS), gigaFLOPS (GFLOPS, approximately 1000 MFLOPS), teraFLOPS (TFLOPS, approximately 1000 GFLOPS), and even petaFLOPS (PFLOPS, approximately 1000 TFLOPS).

A cheap, but modern desktop computer using, for example, a Pentium 4 or Athlon 64 CPU, typically runs at a clock frequency in excess of 2 GHz and provides computational performance in the range of a few GFLOPS. Usually, computer systems that operate at or above the teraFLOP level are counted as HPC-computers.

https://t.me/learningnets



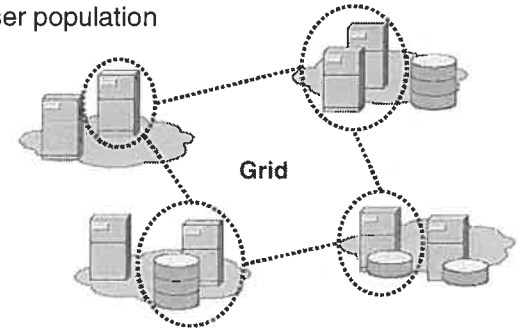
HPC clusters tend to have customized implementations built around the specific application requirements. These applications generally require high performance implementations with the ability to compute nodes of up to tens of thousands of devices.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

What Is a Grid?

- Sometimes considered a type of very loosely coupled cluster
- Large-scale and widely distributed
- Require decentralized management
- Comprise numerous heterogeneous resources
- Typically a transient user population



The concept of grid computing has been interpreted in many ways. A grid generally refers to a pool of heterogeneous compute resources that may be under the control of several administrative authorities. A grid may be made up of several pools of resources—storage, compute, and so on—and may be formed of multiple geographically distributed clusters. From an external perspective, grid middleware allows resources within the grid to be scheduled to perform compute tasks.

Grids can be thought of as very loosely coupled clusters, although grids differ from traditional clusters in several ways:

- Clusters must contain heterogeneous servers running the same operating system. Grids can run multiple operating systems.
- Clusters are connected by private intra-cluster networks, usually requiring low latency and deterministic throughput. Grids can be connected by public networks, even the Internet.
- Clusters consist of a known population of well-connected nodes. Grids can consist of a transient population of widely distributed nodes.

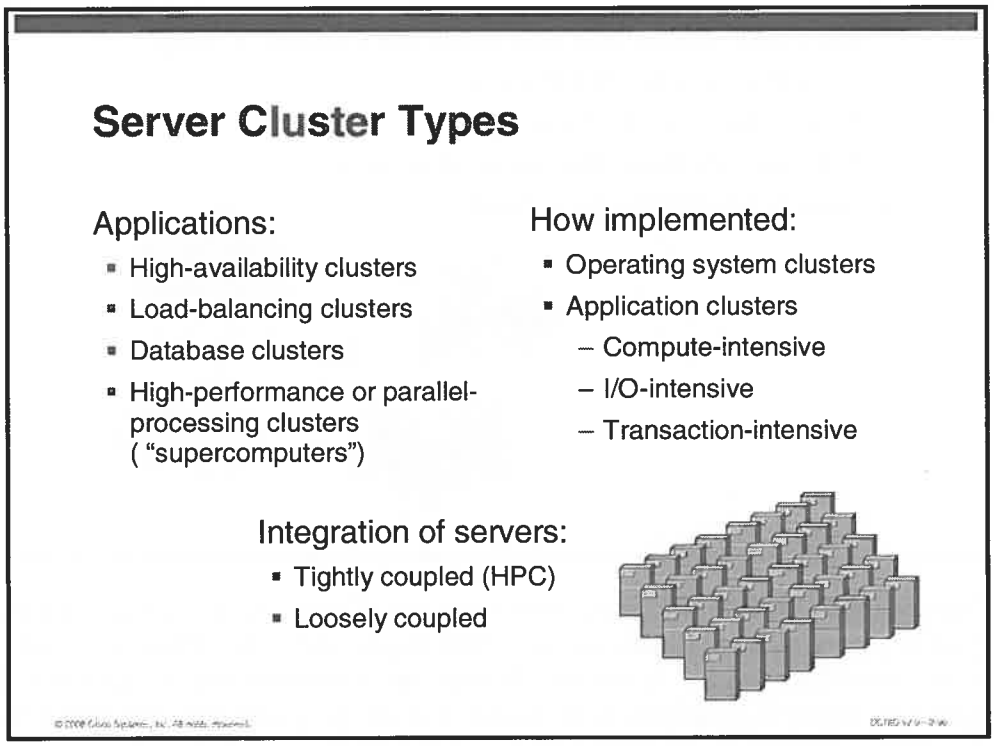
Many enterprises are looking toward grid computing to enable sharing of information, such as part of an extended supply chain management agreement, or for research and development purposes. In this context grid is often associated with utility computing, where compute resources can be scaled up or down—according to demand, and where compute power can be purchased on a $-per-CPU, per-hour basis. Grids are also associated with scavenger-type applications such as Folding@home and Seti@home, which use spare CPU cycles on distributed PCs to perform computations.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cluster Types

The objective of this topic is to describe the different types of clusters found in a data center environment.



The term cluster, loosely defined, generally refers to a collection of multiple compute resources under the control of a single administrative authority. Although the cluster is made up of multiple compute resources, from an external perspective, the cluster appears to be a single system. Clusters are typically geographically bound due to latency or administrative reasons. The term cluster is used to describe many different types of collaborative operations between servers.

There are different ways to categorize server clusters.

Applications

There are four basic applications for clustering:

- High-availability clusters provide fault-tolerance by permitting an application to fail over to another node when a given node fails. High-availability clusters can be implemented as two-node pairs.
- Network load-balancing clusters provide scalability by load-balancing traffic across many servers. Web server farms are a common example of load-balancing clusters.
- Database clusters apply the benefits of high-availability and load-balancing to database environments.
- High-performance or parallel-processing clusters are specialized clusters that are customized to perform a single task, such as complex simulation modelling.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

High-Availability Clusters

A high availability cluster provides non-stop, or as near non-stop as is possible, application availability. There are several strategies that may be employed to achieve high availability, but in general one server is the master node and one or more servers provide back-up services if the master node fails. This type of cluster typically uses interserver heartbeats that periodically exchange node health information. High-availability clusters are not inherently CPU or network-bandwidth intensive, and are not latency sensitive.

Load-Balanced Clusters

A load-balanced cluster may be used to scale the performance of an application by dividing user sessions among servers within the server farm. If the resources in the server farm become exhausted, additional capacity can be added without the users being aware of the changes. Again, there are various strategies that can be employed; for example, the Cisco Content Services Module may be used to load-balance sessions across multiple servers. It should be noted that load-balancing clusters can also provide high availability if application level persistence—for example, cookies or Java session persistence—is used. Load-balancing clusters are not inherently CPU or network-bandwidth intensive because adding more servers can address capacity problems, and are not latency sensitive.

Transactional Application Clusters

Transactional applications such as decision support systems typically refer to database clusters. Clustering improves transactional performance by dividing database queries among several database servers. The database servers use some form of middleware—for example Oracle RAC—to distribute the query to perform the database lookup. It should be noted that transactional clusters are neither HPC nor high availability/load-balancing clusters, even though most implementations have inherent load-balancing and high availability capabilities. Transactional clusters are not inherently CPU or bandwidth intensive and adding more servers can address capacity problems; however, low latency and high bandwidth are required to maintain database record consistency (cache synchronization, and so on) and to accommodate the size of database response, respectively.

How Clustering Is Implemented

Clustering can be implemented at different levels of the system:

- Operating system clusters are general-purpose clusters that provide clustered resources to any application that runs on the cluster. UNIX/LINUX Parallel Server cluster, Microsoft Cluster Server, and VERITAS Cluster Server are all examples of operating system clusters. Generally, applications must be modified to take full advantage of the cluster's capabilities, but some limited value may be derived from running standard off-the-shelf applications on general-purpose clusters.
- In an application cluster, the clustering functions are implemented in the application, rather than the underlying operating system. Database clusters are the most common example of application clusters.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

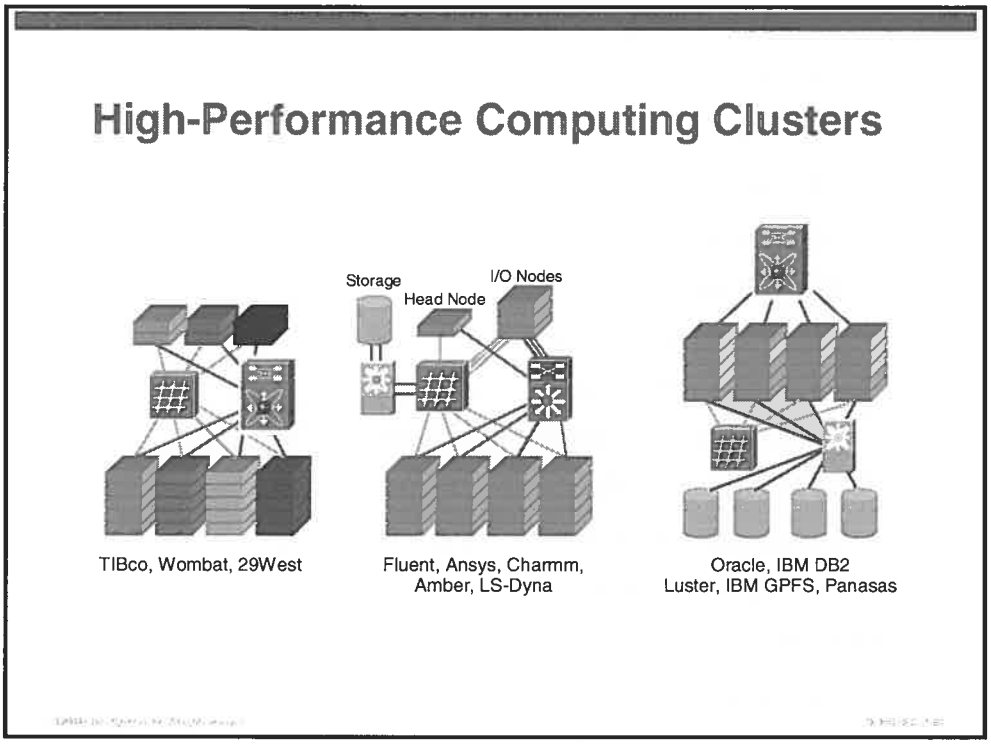
Integration of Servers

Another way to categorize clusters is by how closely coupled the servers are in the cluster.

- In a loosely coupled cluster, the servers might share storage, but they do not share a common system memory space.
- In a tightly coupled cluster, the servers share a common system memory space. Shared-memory clusters are able to interact more closely and rapidly. This type of cluster requires extremely low-latency interconnects to facilitate rapid transfers of data directly from the system memory of one node to another node.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



HPC clusters refer to two or more computers that are used to solve a single problem. HPC clusters consist of a single master node and multiple slave nodes that are interconnected using a high-performance network. Depending upon the application, HPCs can be divided into three subcategories:

- **Parametric execution:** Parametric execution is used for applications that cannot be parallelized, and consequently the application runs on a single compute node. What parametric execution leverages is that, although the application cannot share information between nodes, if different input data is sent to different nodes running the same application, the nodes can compute their information in parallel, thereby speeding up the operation. Because parametric execution does not exchange messages between nodes, a parametric execution HPC cluster is latency insensitive, but may be bandwidth intensive depending upon the application.
- **Loosely coupled applications**: Loosely coupled applications require minimal interaction with other cluster nodes because they can perform the entire computation autonomously. Loosely coupled clusters are generally latency insensitive and CPU intensive, and require moderate to high bandwidth network interconnects. Other commonly encountered classifications are massively parallel, embarrassingly parallel and nearly embarrassingly parallel. These applications share much same characteristics as loosely coupled applications, although there are variations on sensitivity to latency and bandwidth consumed.
- **Tightly coupled applications:** Tightly coupled applications require information to be periodically exchanged with other nodes—either all nodes or subsets of nodes—in the form of messages. Tightly coupled applications are generally latency sensitive and CPU intensive, and generate bursty, unpredictable traffic patterns as data is exchanged. Because time spent communicating is time spent not processing data, tightly coupled applications require low latency and high bandwidth network interconnects.

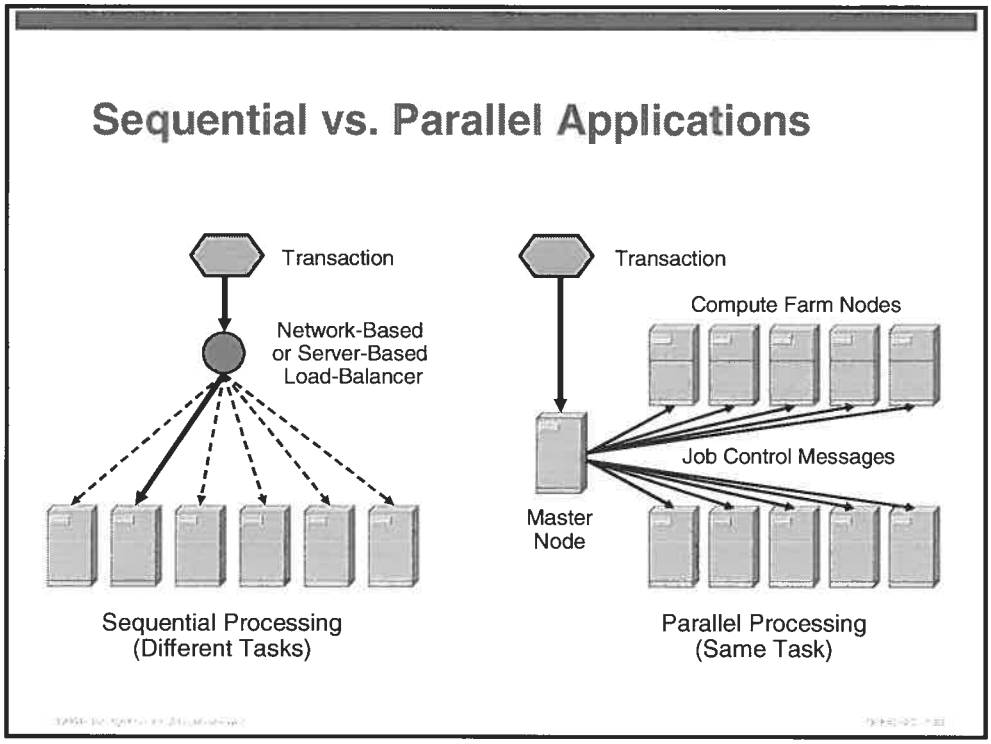
https://t.me/learningnets

There are three major categories of HPC application clusters.

1. Compute intensive:
   - Seismic analysis
   - Linear algebra
   - Aeronautical engineering
   - 3D rendering
   - Weather simulation
   - Medical image analysis
   - Bioinformatics – protein/gene/drug analysis
   - Human genome study
   - Nuclear particle analysis
   - Weapons development
2. I/O intensive:
   - Data mining
   - Demographics analysis
   - Satellite imagery assembly
   - Regression analysis
   - Bioinformatics pharmacology studies
3. Transaction intensive:
   - Financial/banking
   - HR
   - DNA sequencing
   - Oracle and DB2-based

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Another way to categorize a cluster is by how the cluster manages its workload:

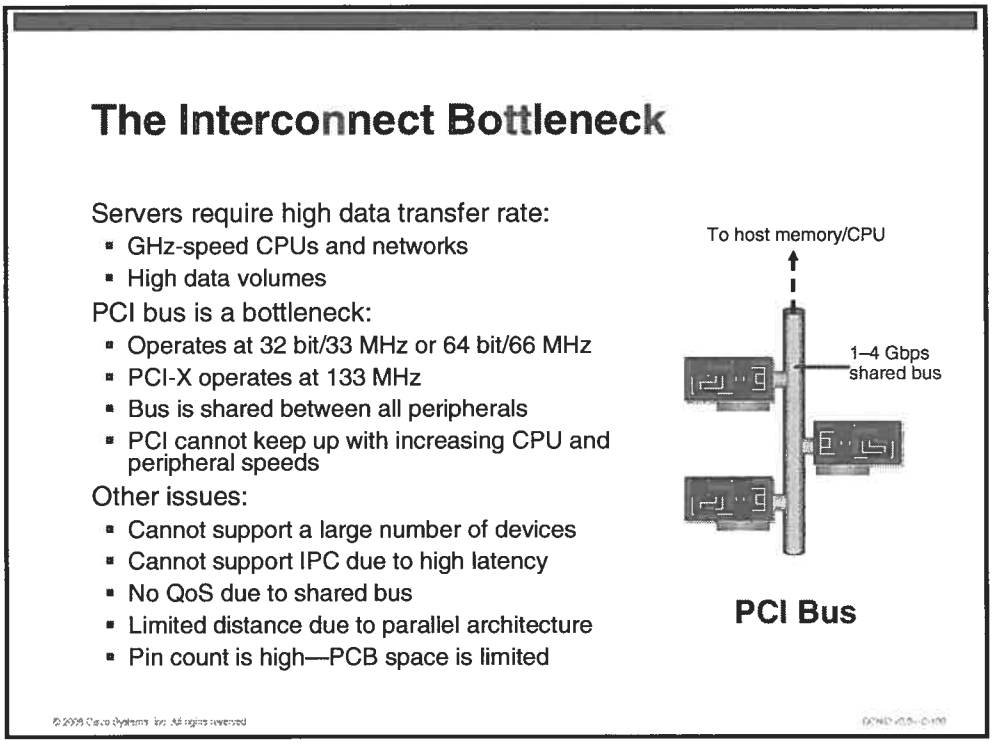
- In a sequential processing cluster, each node effectively functions as a separate system that processes a discrete task. For example, a complex database query would be processed by a single node. A network- or server-based load-balancing function distributes tasks between the nodes.
- In a parallel processing cluster, the nodes can work cooperatively to process a single task. A complex database query may be processed by multiple nodes in a parallel processing environment. A master node is responsible for scheduling and distributing tasks, and for collecting the responses from each node and presenting the final result to the requestor.

This distinction can become important in network design. For example, when you calculate the bandwidth requirements for a cluster, the master node in a parallel-processing cluster might be the only node that requires significant public-network traffic. The private network carries the IPC traffic between cluster nodes, and it is this network that is the focus of the cluster network design. Conversely, in a sequential processing cluster, like a web server farm, the private network might have lower bandwidth requirements.

https://t.me/learningnets

Cluster Interconnects

This topic explains the requirements for cluster interconnects.



PCI is a shared-bandwidth, connection-oriented parallel bus that is used in the vast majority of servers today. This architecture imposes some serious limitations.

Standard PCI buses operate at rates from 33 MHz to 66 MHz, providing from 1 to 4 Gbps of bandwidth, while the next-generation PCI-X bus operates at 133 MHz and provides 8.5 Gbps. All devices on the bus—including the host memory controller, the video card, the storage I/O adapter, and the network I/O adapter—must share this bandwidth. Given that data center servers are based on gigabit-speed CPUs and regularly use gigabit-speed networks like Gigabit Ethernet and Fibre Channel to transmit and receive large volumes of data, it is clear that the PCI bus cannot keep up with the demand for I/O.

Other critical problems with PCI are:

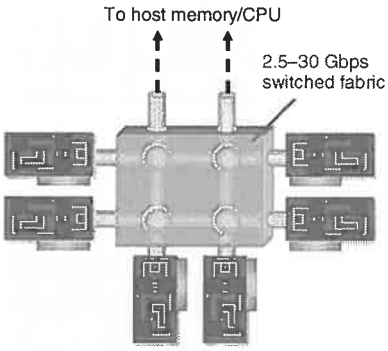
- Because the PCI bus is a shared bus, it cannot support large numbers of devices or provide quality of service (QoS).
- High latency makes the PCI bus unsuitable for interprocess communication (IPC) between clustered servers.
- The parallel architecture of the PCI bus strictly limits the length of PCI links to the centimeter range.
- PCI's high pin count complicates printed circuit board (PCB) design and consumes chassis space.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

What Is InfiniBand?

- InfiniBand is a high-speed, low-latency switched communications technology that is used to interconnect servers, storage and networks.
- Routable protocol based on IPv6.
- Standards-based:
  - InfiniBand Trade Association
  - http://www.infinibandta.org
- The Cisco Server Fabric Switch (SFS) uses InfiniBand to create a high-speed, low-latency server fabric.



InfiniBand Fabric

InfiniBand is a next-generation device I/O bus architecture. It replaces the traditional shared-bandwidth PCI bus with a scalable switched interconnect. Because InfiniBand uses a serial network architecture, it can support higher bandwidth, longer distances, and more devices than traditional I/O buses. InfiniBand provides attached devices with fast, direct, low-latency access to the system memory bus to support distributed computing architectures.

InfiniBand blurs the line between the concepts of bus and network. It can be used as the internal device I/O bus in a server, but it also allows external devices to connect directly to the system's memory bus. The InfiniBand protocol is a routable protocol that is based on the IP architecture. InfiniBand uses IPv6 at the network layer to support inter-subnet routing.

InfiniBand Applications

Key applications for InfiniBand include:

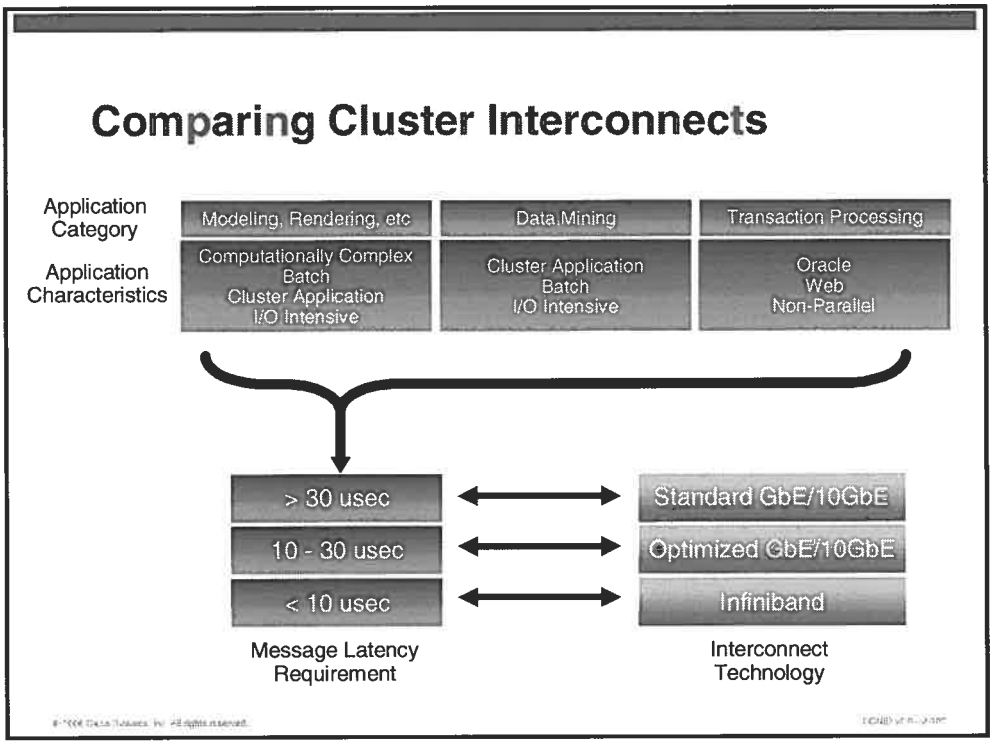
- High performance computing
- Grid and utility computing
- Database clustering

Drivers for InfiniBand adoption in these applications include:

- Processors increasing in performance
- I/O subsystem must not be a bottleneck
- High bandwidth, low latency applications

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cluster Requirements

Cluster interconnects impose stringent requirements on the network transport because the performance of the cluster is critically tied to the performance of the network. Factors include:

- Bandwidth
- Latency
- Non-blocking
- Buffering
- CPU impact

The cluster application is not necessarily related to the interconnect requirements, although parallel clusters tend to have more stringent requirements due to the greater amount of message-passing between nodes. Because of these requirements, cluster interconnects must operate in a Layer 2 domain. Gigabit Ethernet may be a viable solution for some types of clusters, but the latency of Gigabit Ethernet environments can place limitations on how tightly the cluster nodes can be integrated—that is, how closely they can share resources, how effectively they can balance workloads, and how statefully they can failover workloads from node to node.

Gigabit Ethernet

Standard Gigabit Ethernet is applicable in environments with higher latency tolerance (>30 microseconds), and some vendors now offer GigE NICs that support Remote Direct Memory Access (RDMA), which can reduce latency significantly.

© 2008 Cisco Systems, Inc.

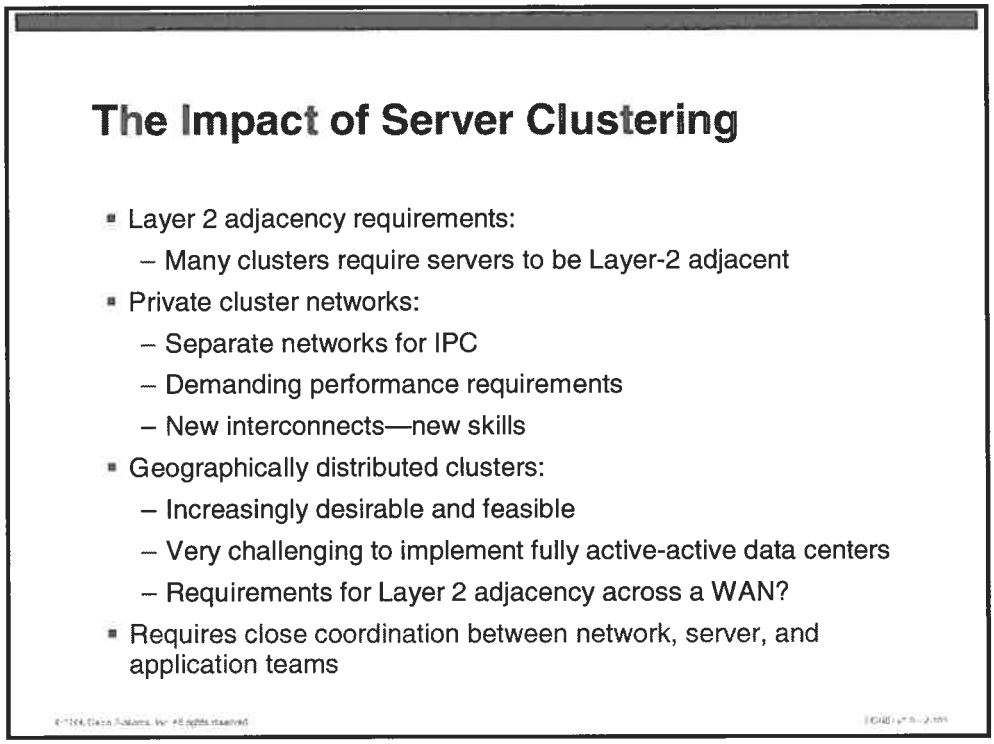
https://t.me/learningnets

InfiniBand

InfiniBand is a standards-based technology that offers very low latency at up to 30 Gbps, although most products today support 10 Gbps or 20 Gbps. It is an RDMA-based protocol that reduces the processing power needed for data transport.

Other cluster interconnects include Myrinet, a proprietary 2.5-Gbps serial interface, and Quadrics, a proprietary 340-Gbps parallel interface.

https://t.me/learningnets



The deployment of server clusters can impact the network design in several ways:

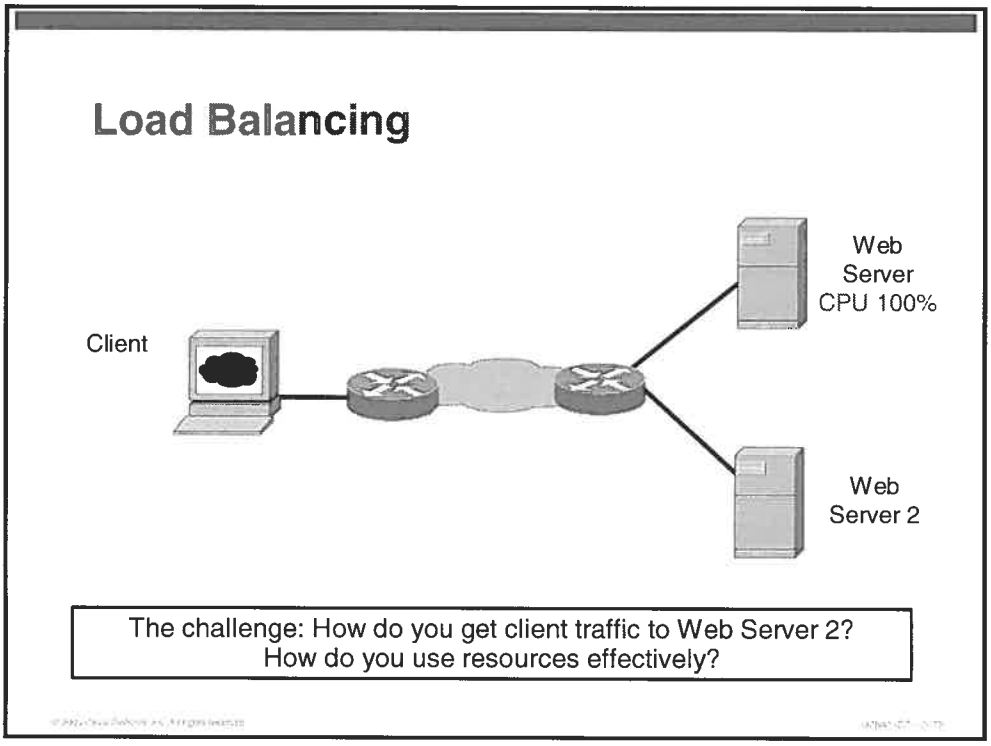
- Layer 2 adjacency requirements. Many clusters require servers to be Layer-2 adjacent.
- Private cluster networks require additional considerations. There might need to be separate networks for IPC, cluster control, and storage. Clusters have increasing network performance requirements, leading to the proliferation of new cluster interconnects, such as InfiniBand, leading in turn to the need for new engineering skills.
- Geographically distributed clusters are increasingly desirable and feasible today. However, it is still very challenging to implement fully active-active data centers. In some cases, there might be requirements for Layer 2 adjacency across a WAN, which obviously requires careful consideration.
- Deployment of server clusters requires close coordination between the network, server, and application teams.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Server Farms and Load Balancing

This topic discusses the concept of server farms and SLB functions within the data center, and their impact on the network design.



The figure illustrates a common situation. E-commerce and other forms of dynamic content for the website www.example.com have pushed server CPU utilization to 100%. A second server (web Server 2) has been added to distribute the load. However, how can traffic be transparently spread across both servers, and how can the distribution of traffic be easily managed and controlled?

Various methods of using multiple web servers have been available for several years. One popular method is to register multiple IP addresses in the DNS under the same system name.

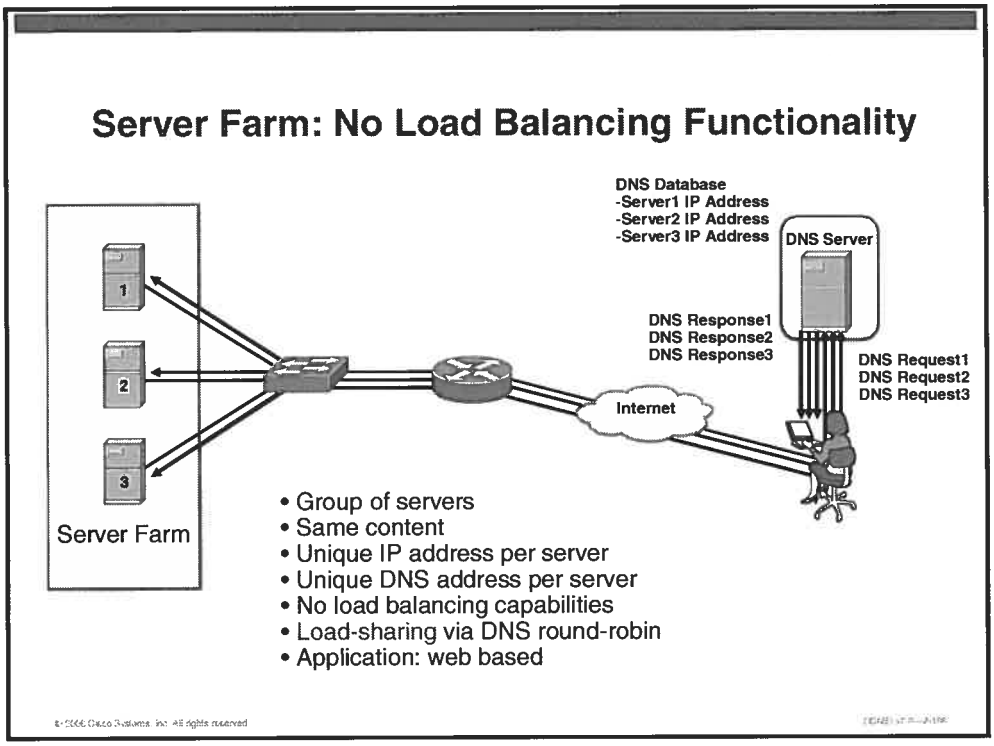
While this method provides some load-balancing capabilities, it has some limitations:

- Support is required in the client because not all web browsers properly handle multiple IP addresses.
- Round-robin name servers lack the information to make load-balancing decisions with any intelligence.

The end result is that use of the servers is not consistent, as shown in the diagrammed network. This inconsistency makes capacity planning and provisioning a difficult process. The lack of intelligent traffic management decisions leads to situations in which it is almost impossible to effectively use each web server.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

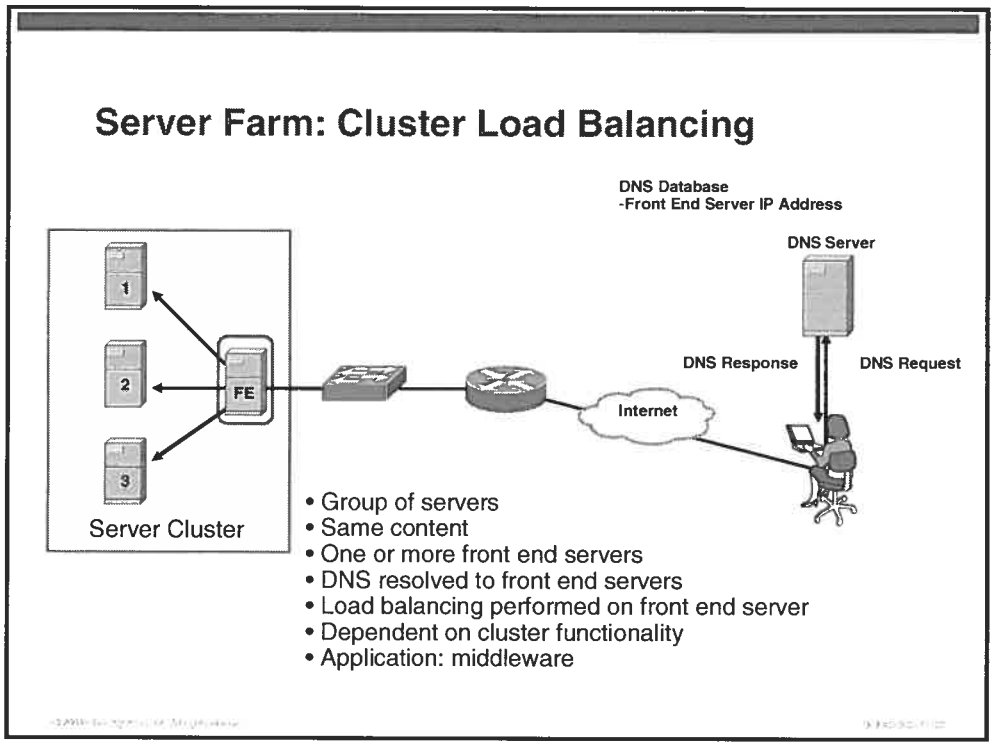


The figure shows a group of servers all serving the same content to clients. In this example there is no load balancing functionality, so each server must have a unique IP address and DNS name for external reachability. From a management perspective this creates overheads.

What would happen if the server to which the client was attached became overloaded and unable to accept any further connections, or alternatively, should go down?

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



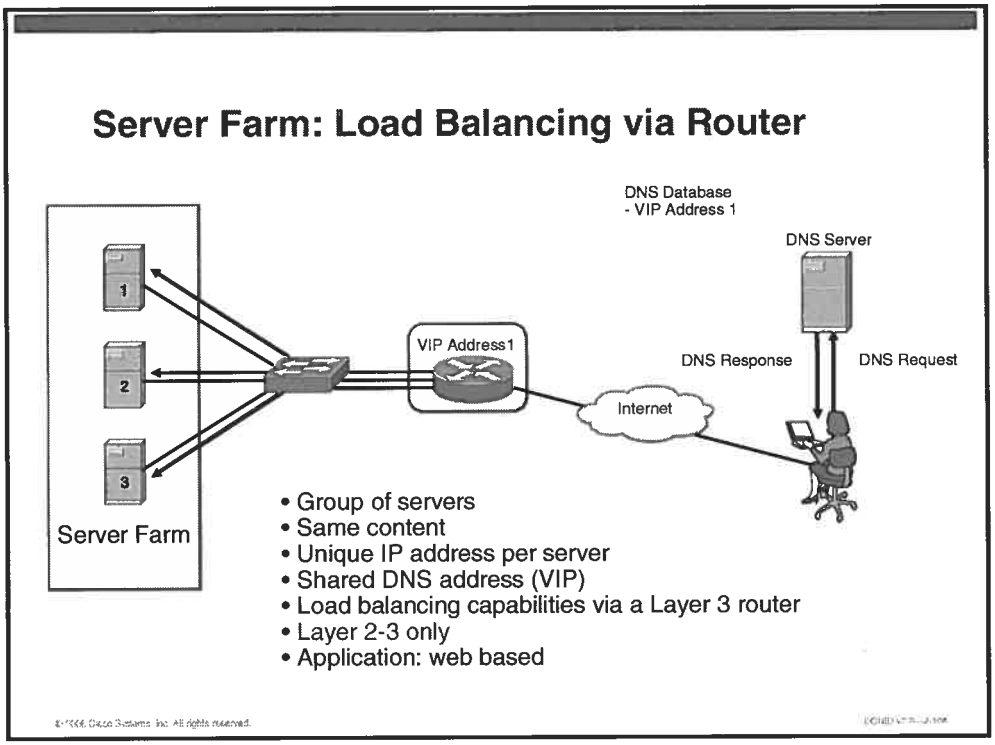
The figure shows a group of servers configured in a cluster with one or more front-end servers. All requests are sent through to the front-end servers who are responsible for making the load balancing decisions. The front-end servers require additional software to perform this function.

What happens if the front-end servers go down or become unreachable?

How do the front-end servers know which servers are available?

How do the front-end servers know which server is least loaded?

https://t.me/learningnets



The figure shows a group of servers, this time defined within a server farm. The Layer 3 router has a VIP address defined to which all content requests are sent The router then makes a load balancing decision as to which server within the server farm should process the request that was sent through.

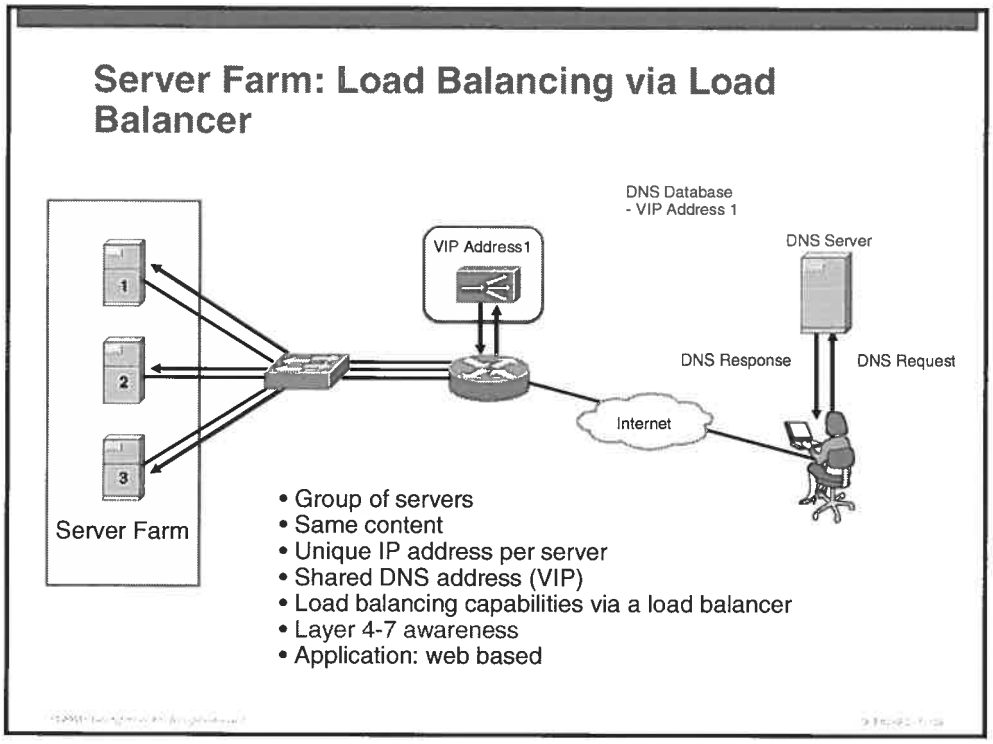
What happens though if a server should go down?

How does the router decide which server should process the request?

How does the router know which servers are available?

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure shows a group of servers, this time defined within a server farm. The load balancer has a VIP address defined to which all content requests are sent. The router forwards all requests through to the load balancer, who then makes a load balancing decision as to which server within the server farm should process the request that was sent through.

What happens though if a server should go down?

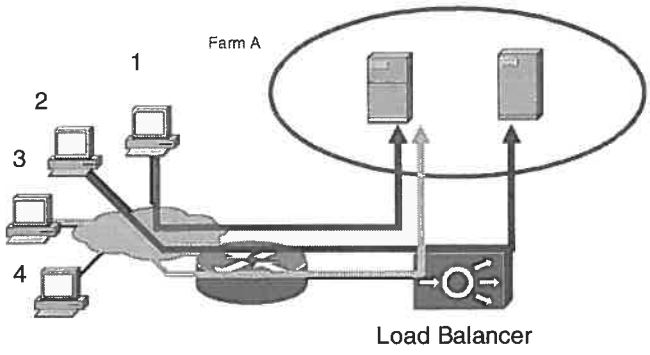
How does the load balancer decide which server should process the request?

How does the load balancer know which servers are available?

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Layer 4 Load Balancing

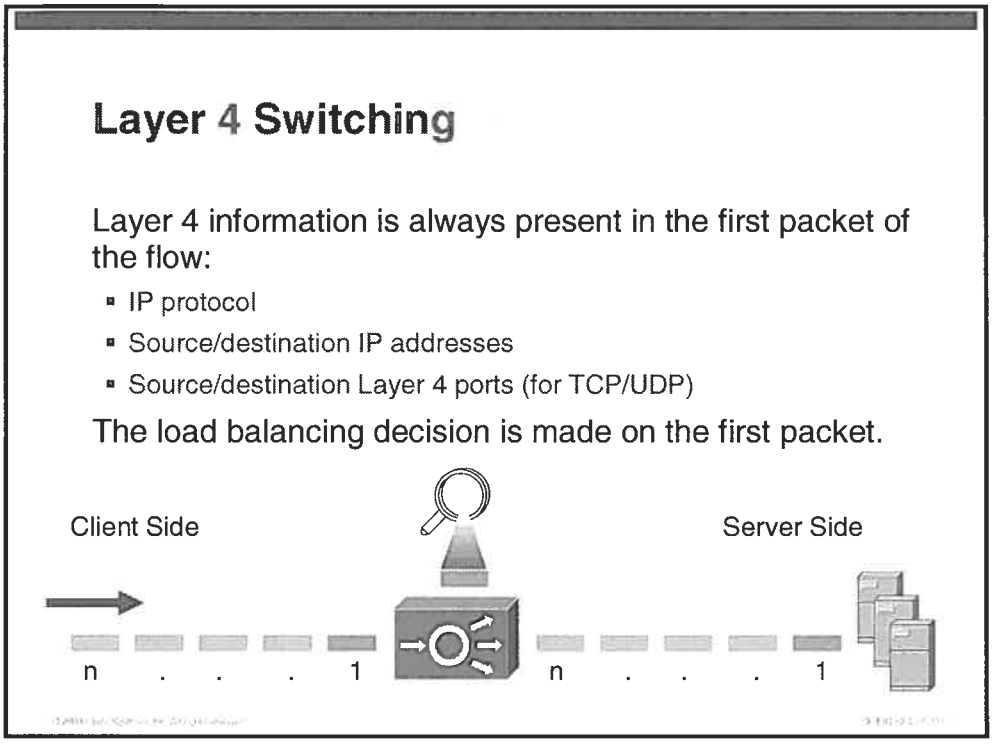


Cisco load balancers can make load balancing decisions that are based on Layer 4 information. These decisions are made on the first packet of a TCP connection (the SYN packet from the client) and on the first packet that is detected on a User Datagram Protocol (UDP) flow. Layer 4 load balancing decisions utilize the load balancing capabilities of a single server farm.

In our preceding sample network, we have a server farm with two real servers. Incoming client requests are being balanced between these servers at Layer 4.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



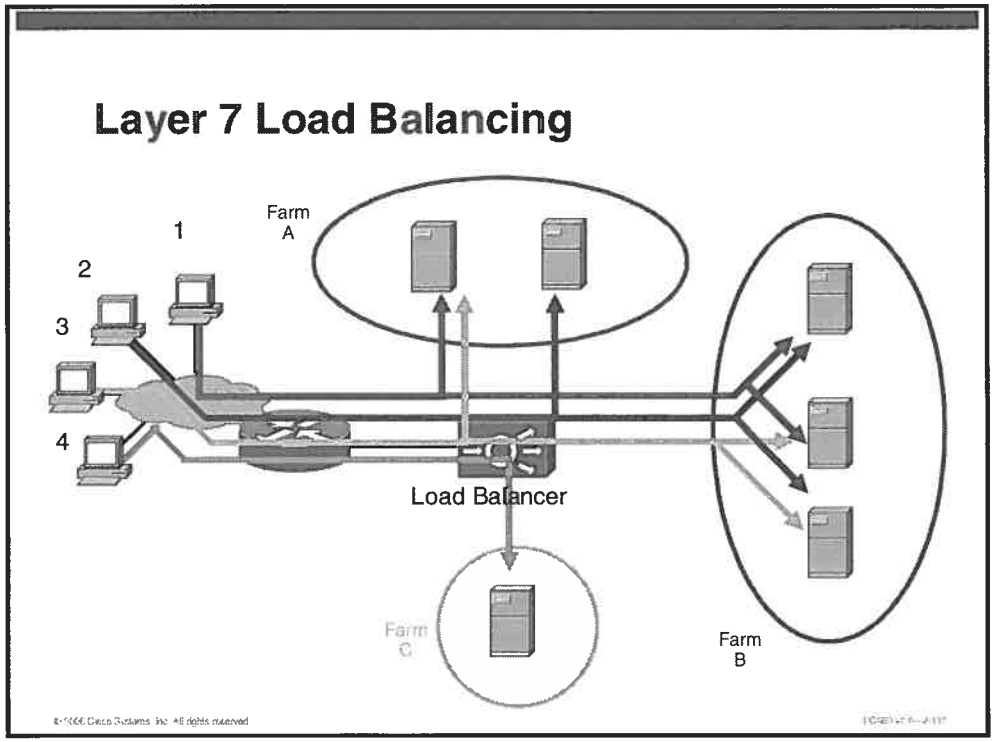
Layer 4 information in the packet includes the following fields:

- **IP:** This field is used to differentiate between the higher level protocols such as UDP and TCP that are supported by IP.
- **Source and destination IP addresses:** The IP address of the transmitting system and the intended recipient.
- **Source and destination port:** The port number that is being used on the transmitting system and the intended recipient. Note that port numbers are used to direct the IP traffic to a particular application process such as a web client or server. Well-known port numbers are defined for most IP based services; for example, port 80 is used for HTTP.

Layer 4 content switching decisions can be based on any of the Layer 4 fields listed above. With TCP connections the Layer 4 information is consistent for all packets in the connection. The Layer 4 information is often said to define a flow which is the communication path for a particular connection.

The preceding diagram shows a flow of packets coming from the client side of the network into a Cisco ACE appliance. The Cisco ACE examines the first packet in a new flow or connection, and a Layer 4 switching decision is made for the flow as a whole. The content switch makes this decision and then records the flow parameters and the switching decision. This table of switching decisions is used to switch every subsequent packet in the flow. Information is removed from the switching table when a connection is closed. In the case of Layer 4 switching of TCP packets, these decisions are normally made on the basis of synchronize and finish (FIN) packets and are done at TCP connection setup and termination. Reset (RST) packets are also analyzed as they are used to refuse a connection when it is requested or to abort an existing connection.

https://t.me/learningnets



Cisco load balancers are capable of policy driven load balancing, which extends the server selection process. Load balancing policies are used to select a server farm to process a client request. After a server farm has been selected, the server farm load balancing algorithm completes the server selection.

The diagram shows a hypothetical policy driven load balancing situation. We have three server farms. Farm A consists of two real servers, which are load-balancing using the round robin algorithm. Farm B consists of three real servers, which are also performing round robin load balancing. Farm C consists of only one server.

One virtual server IP is configured to which all clients target their requests. This virtual server implements the following policies:

1. Authorized users get HTML files from Farm A.
2. Authorized users get image files from Farm B.
3. Unauthorized users get an error page from Farm C.

Each client, in order, is trying to retrieve a home page from the virtual server. The home page contains two images that the clients also retrieve. Clients 1 through 3 are authorized; client 4 is not.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Client 1

1. Client requests the home page.
2. Policy assigns the request to Farm A.
3. Farm A load balancing assigns the request to the left server.
4. Client requests image 1.
5. Policy assigns the request to Farm B.
6. Farm B load balancing assigns the request to the top server.
7. Client requests image 2.
8. Policy assigns the request to Farm B.
9. Farm B load balancing assigns the request to the middle server.

Client 2

1. Client requests the home page.
2. Policy assigns the request to Farm A.
3. Farm A load balancing assigns the request to the right server.
4. Client requests image 1.
5. Policy assigns the request to Farm B.
6. Farm B load balancing assigns the request to the bottom server.
7. Client requests image 2.
8. Policy assigns the request to Farm B.
9. Farm B load balancing assigns the request to the top server.

Client 3

1. Client requests the home page.
2. Policy assigns the request to Farm A.
3. Farm A load balancing assigns the request to the left server.
4. Client requests image 1.
5. Policy assigns the request to Farm B.
6. Farm B load balancing assigns the request to the middle server.
7. Client requests image 2.
8. Policy assigns the request to Farm B.
9. Farm B load balancing assigns the request to the bottom server.

https://t.me/learningnets

Client 4

1. Client requests the home page
2. Policy assigns the request to Farm C.
3. There is only one server in Farm C and so the request gets assigned to it.

© 2008 Cisco Systems, Inc.

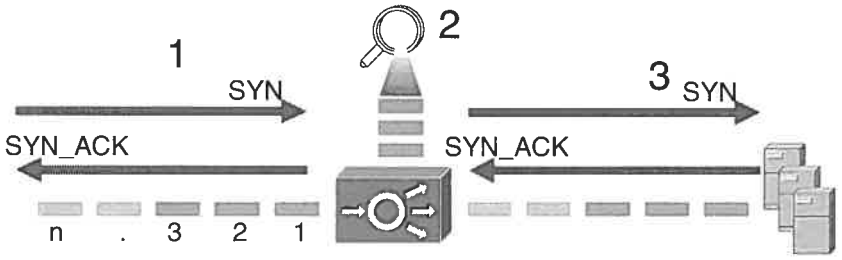
https://t.me/learningnets

Layer 7 Switching

L5 - L7 information is received only after the TCP setup and might span multiple packets:

- HTTP URLs, cookies, header fields
- SSL session ID
- FTP data channel port
- Generic application data

Requires TCP termination and buffering



Layer 7 information is available only after application data has been transmitted, but transmission requires that the TCP connection be fully functional, leading to a chicken-and-egg kind of problem. A server must respond to the client to fully start the TCP connection before the client sends the Layer 7 information that the content switch needs to select the server.

The content switch handles this chicken-and-egg dilemma by buffering client data and temporarily acting as a server. To do this, the content switch responds to the incoming SYN packet with its own SYN_ACK. The content switch then buffers packets until it has enough Layer 7 information to make a load balancing decision.

After a destination server has been selected, the content switch must make a connection to the server on behalf of the client. To establish the TCP connection to the server, a SYN packet is sent to the server, and then the load balancer waits for the SYN_ACK packet to be sent from the server. At this point, all buffered packets received from the client are sent to the server.

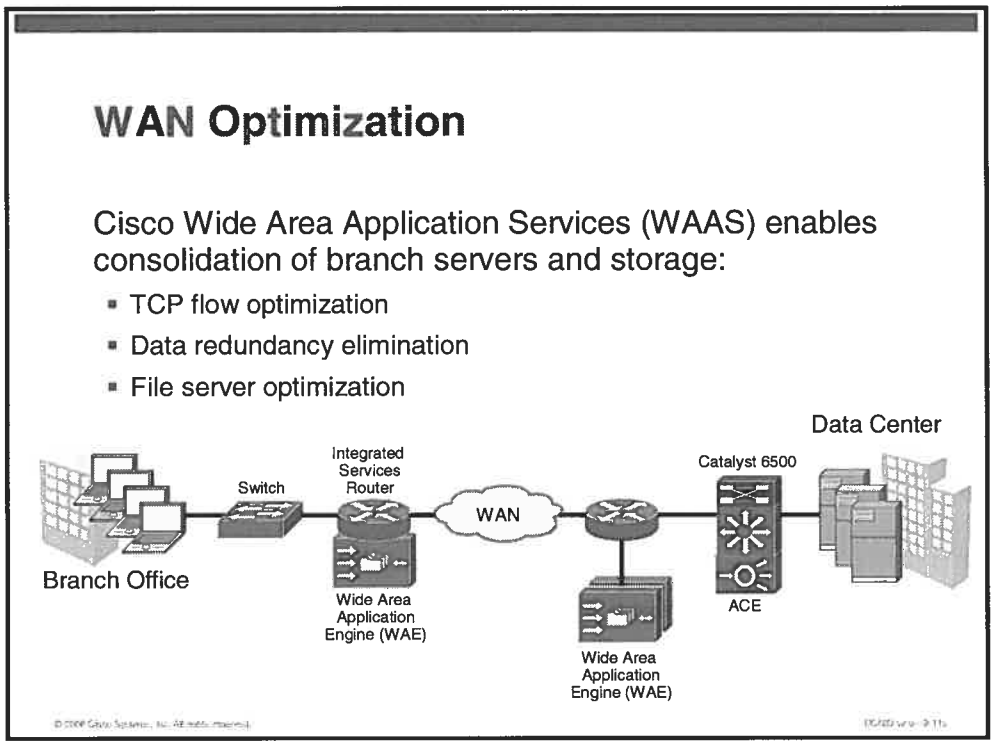
After the buffered packets have been sent, the two TCP connections can be spliced together by the content switch. This splicing is done by receiving packets from one connection and retransmitting them to the other.

Given that there are two different TCP connections from the content switch, one to the client and one to the server, there are likely two sets of sequence numbers in use, one on each connection. The content switch translates the sequence numbers from one connection to the other.

https://t.me/learningnets

WAN Optimization and Branch Consolidation

This topic discusses the impact of WAN optimization and branch consolidation on the overall network design.



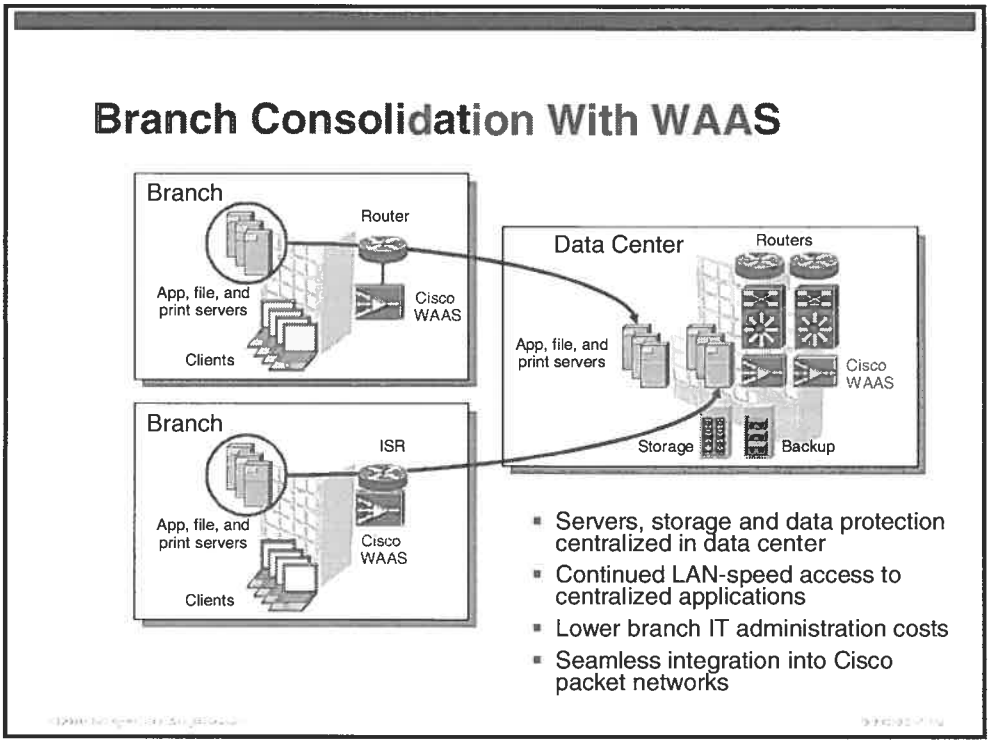
Cisco Wide Area Application Services (WAAS) is a powerful new solution that overcomes the challenges presented by the WAN. Cisco WAAS is a software package that runs on the Cisco WAN Application Engine (WAE) that transparently integrates with the network to optimize applications without client, server, or network feature changes.

A WAE is deployed in each remote office, regional office, and data center of the enterprise. With Cisco WAAS, flows to be optimized are transparently redirected to the Cisco WAE, which overcomes the restrictions presented by the WAN, including bandwidth disparity, packet loss, congestion, and latency. Cisco WAAS enables application flows to overcome restrictive WAN characteristics to enable the consolidation of distributed servers, save precious WAN bandwidth, and improve the performance of applications that are already centralized.

As shown in the preceding diagram, Cisco WAAS can be deployed using appliances that attach to the network as nodes, or it can be deployed within the Cisco Integrated Services Router (ISR). Alternatively, Cisco WAAS can be deployed on devices that are deployed physically in-path.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



According to research conduced by IDC, Gartner, and Cisco, companies spend 6 billion dollars per year on branch servers, storage, backup and management. According to a study conducted by Nemertes Research, the average branch office has four to six servers.

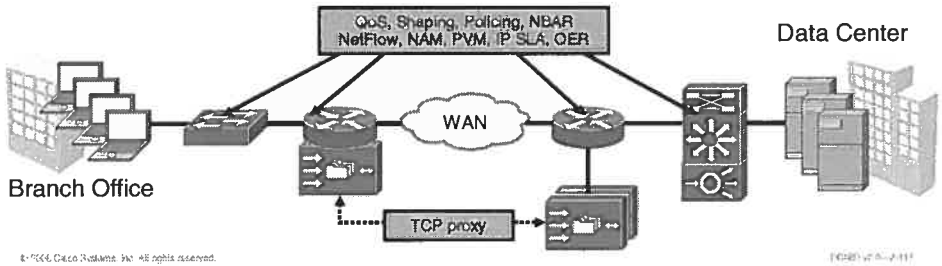
WAAS allows businesses to:

- Centralize file, print, application servers, storage, and data protection in the data center
- Provide continued LAN-speed access to centralized applications
- Reduce branch IT administration costs
- Maintain seamless integration of applications on Cisco packet networks

https://t.me/learningnets

Interoperability with Network Features

- Third-party WAN optimization products use GRE tunnels to transport optimized flows.
  - This breaks existing traffic prioritization and monitoring applications because original packet headers are not visible.
- WAAS uses a TCP proxy architecture rather than GRE.
  - WAAS integrates transparently with IOS features for network control, optimization, visibility, and monitoring.



Third-party WAN optimization products use GRE tunnels to transport optimized flows. GRE tunnels break existing traffic prioritization and monitoring applications because the original packet headers are not visible to the prioritization and monitoring applications.

Cisco WAAS uses a TCP proxy architecture for optimized connections, providing a transparent framework for accelerating applications. WAAS preserves information that is critical to network feature operation, such as packet header information. With its transparent optimization architecture, WAAS is uniquely positioned to offer feature compatibility with functions that are already used within the network, including:

- **Quality of Service (QoS)**: QoS policies provide traffic classification, prioritization, scheduling, queuing, policing, and shaping.
- **Network-Based Application Recognition (NBAR)**: NBAR provides protocol and application discovery.
- **Access Control Lists (ACLs)**: ACLs permit or deny traffic based on identification.
- **NetFlow**: NetFlow provides statistics on nodes communicating on the network, and the applications that are being used.
- **Firewall policies**: Firewall policies are used to prevent malicious traffic from entering or exiting a network.

By providing a transparent architecture, WAAS is able to provide interoperability with end-to-end performance analysis systems such as Cisco Network Application Performance Analysis (NAPA) and Performance Visibility Manager (PVM), as well as third-party products.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Module 5

Data Center Strategy

Overview

This module introduces the data center network strategy.

Module Objectives

Upon completing this module, you will be able to describe the Data Center network design strategies. This ability includes being able to meet these objectives:

- Describe and understand Data Center Drivers
- Describe and understand Data Center Network Architecture
- Describe and understand Cisco Data Center Virtual Switching Technology
- Describe and understand Unified Fabric and Unified I/O
- Describe and understand Data Center Ethernet (DCE)
- Describe and understand L2 Enhancements
- Describe and understand Data Center Sizing

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 1

Data Center Strategy

Overview

This lesson introduces concepts and basics of data center network strategy that designers are facing today.

Existing data center architectures are being revisited. Behind these efforts to craft a new architecture are a number of business drivers that collectively impose a new set of challenges. These challenges are addressed through innovative networking technology enhancements offered by Cisco data center switching products.

Objectives

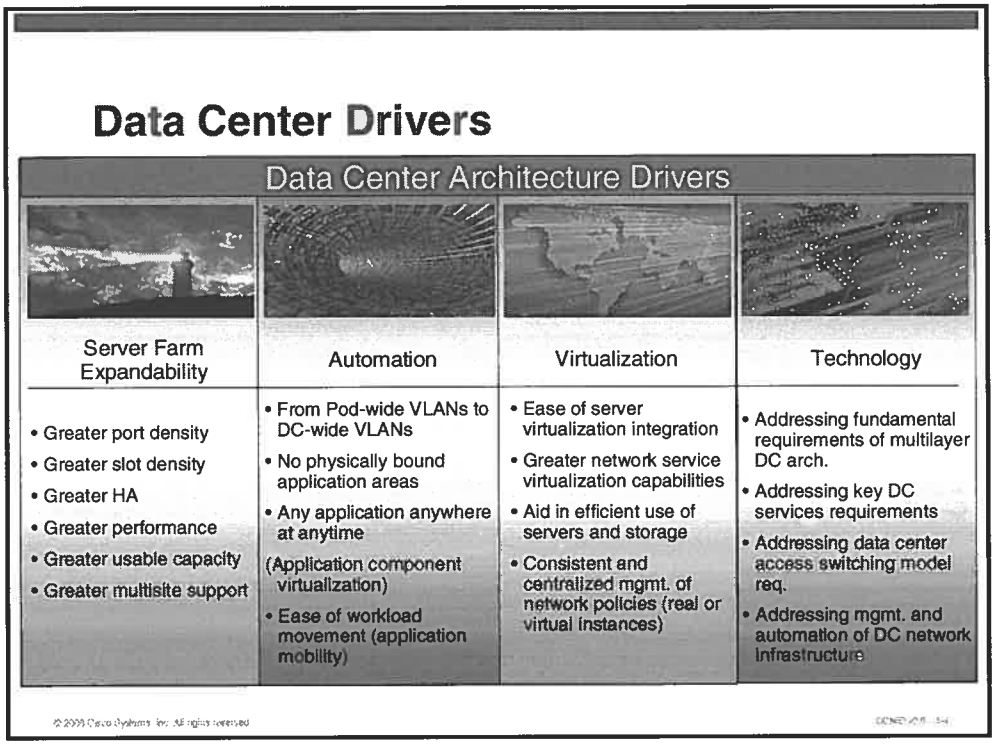
Upon completing this lesson, you will be able to meet these objectives:

- Describe and understand Data Center Drivers
- Describe and understand Data Center Network Architecture
- Describe and understand Cisco Data Center Virtual Switching Technology
- Describe and understand Unified Fabric and Unified I/O
- Describe and understand Data Center Ethernet (DCE)
- Describe and understand L2 Enhancements
- Describe and understand Data Center Sizing

https://t.me/learningnets

Data Center Driver

This topic describes the foundation and reasons for the data center Ethernet technology.



The most significant drivers and the force with regards to the data center architecture are the following:

Server Farm expandability or scalability and resource density: Server farms are getting larger for multiple reasons such as consolidation. Server farms are getting collapsed into a single physical security to have more servers per physical data center facility or footprint. The server growth is driven by the application growth. In addition, the growth of the server virtualization does not translate into a decrease in the growth for physical servers. Server virtualization simply slows down the curve of adoption of physical server technology. It is also important and interesting that the servers are more capable than they used to be, which is making the average price of the server increase dramatically. These are all the drivers for the greater port density and a for the higher slot density. Production of denser line cards was a response to the described data center server farm requirements. Server consolidation due to server virtualization is driving the need for more I/O bandwidth, such as 10Gigabit Ethernet channel availability to virtualized hosts. Digital media applications represent a small percentage of the data center

Greater high availability: Data center high availability has gained bigger importance and attention than before, since a single failure in a consolidated server farm can cause a wide-spread failure with a greater effect than a failure of a non-consolidated server farm. The new data center environment consists of Cisco IOS Software and Cisco NX-OS Software, and high availability must be provided by both OSs.

Greater performance: Data center design tends to lower the server oversubscription in order to improve the data center performance and eliminate the probability of triggering the congested data center network by highly oversubscribed server access layer ports and ISLs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Greater usable capacity: Greater usable capacity is related to the problems with the spanning tree blocking paths. Most Enterprise Data Center customers are using the Per VLAN Rapid Spanning Tree (PVRST) or some sort of Per VLAN Spanning Tree (PVST)-based protocol. There are several hundred service providers that are deploying Mono Spanning Tree in the data center infrastructure. A Mono Spanning Tree protocol is not a plug-and-play protocol like PVRST or PVST are. The Mono Spanning Tree protocol requires more operational maintainability and knowledgebase, which requires an interesting set of knowledgebase collisions. Many data center deployments have migrated from PVST to Mono Spanning Tree for better data center performance and stability reasons.

Demand for data center multi-site support: Multisite support is demanded mostly because of server high availability or because of applications such as VmWare Vmotion solutions. Vmotion enables workloads to be moved dynamically within a server farm.

Automation flexibility: The IT organization and the data center group have defined new standards about how the data center services are deployed. Data center services are the name for new servers, applications, and whatever comes along when the services are rolled out into a network. It includes the Layer 2 and Layer 3 configuration and the process that takes place to get that done. The standard includes the definition for the physical facility and the footprint that the new server farm takes. A rack that has a particular access switch or area that is all fed by a single access switch is also standard. When the new server is connected to the fabric, it is expected to be integrated within a particular VLAN on a particular subnet. If a different business department attempts to use the same subnet and VLAN, then only certain VLANs and certain subnets on a particular rack, or that particular set of access switches, will be available.

When the capacity of a particular area is reached, the growth of the business unit causes a new area for the business unit to be deployed. The growth of the business unit can therefore trigger a requirement for a new area for the business unit to be deployed. If the utilization of the next layer up—which is the aggregation layer—is already maximized, the designers are forced to either create another aggregation layer or attempt to use the one that exists from a different business unit. In either case, VLANs and subnets are the problem in this case. In the worst case scenario, the cables are strung from an access switch to an aggregation switch in order to assign the new servers to the same VLAN and subnet to which the other ones are assigned. Service providers have selected the mechanism where any VLAN could be available anywhere almost at anytime within the constraints of a 2-tier Layer 2 domain. The design guide is not broken by the service providers' VLAN availability but the second tier, which is the Layer 2 and Layer 3 boundary aggregation layer, is becoming a Layer 2 only. **The Layer 2 adjacency would be allowed across the entire data center up to the core layer.**

Applications anywhere at any time and Application Virtualization: The nature of rolling a new application on a non-virtual server environment has to do with how fast an application server can be deployed on the network to solve a particular problem. The problem originates from the physical ties between the network components and the logical definition of the network components on the physical servers. The new data centers must be flexibility in terms of how the new application can be rolled on the network. The time required must be cut from hours or days to minutes or seconds, if possible.

There is another aspect of application availability. Application virtualization would make it possible to start applications from any web browser at any location whenever possible. The web-based applications appear to play an important role due to application-wide availability. Application virtualization and availability are provided by the WmWare Vmotion middleware. An appliance will control which application is being moved where, when, and how; and the whole process is supposed to be accomplished dynamically. IBM and Opsware are providing application high availability, as discussed above, or at least some elements of it.

https://t.me/learningnets

The next step on the way to providing the best application availability is to control the application workload movement, which is the more efficient mechanism to tag the OS CPUs and memory, and another solution to a problem that could be a processing capacity anywhere on the network. This is probably the next stage of application automation that enables utilization of an idle CPU somewhere in the network for application processing purposes. When the code gets processed, the final output is sent back to the local application. There has to be a lot of technology that needs to exist which is not currently available.

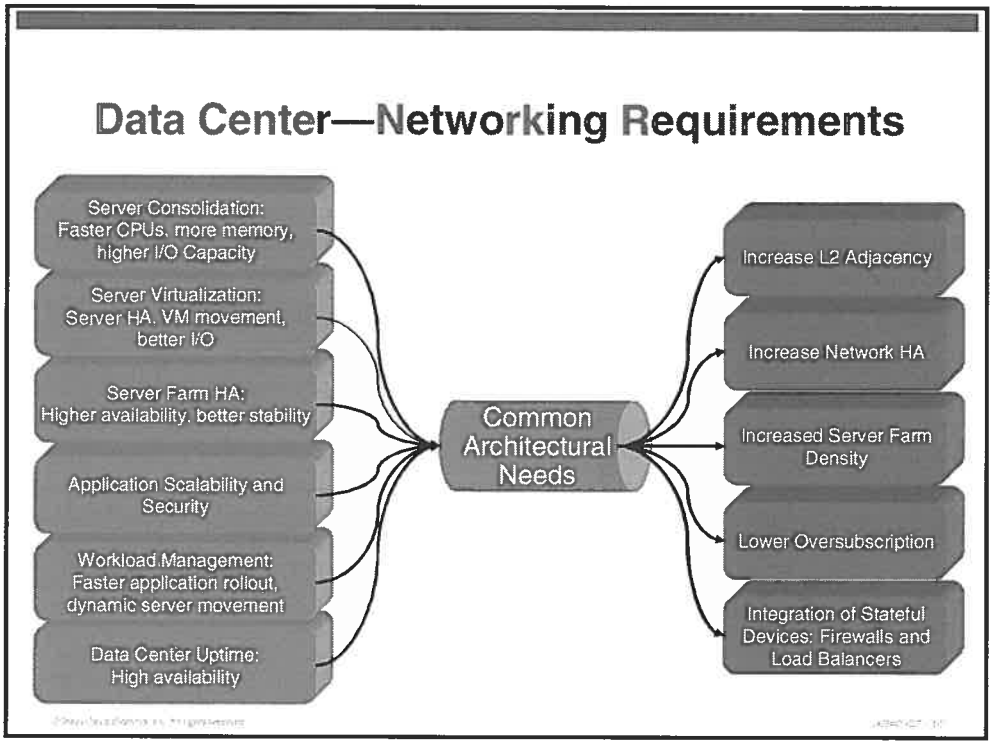
Virtualization: From the networking perspective, the VLANs and VSANs, and all the other types of virtualizations are useful to a certain extent. They make it possible for us to offer high-availability, centralized management and the ability to create network components that are grouping network entities regardless of their physical location. Virtualization, as already mentioned, is a necessary step toward the data center automation.

- **Servers Virtualization** : The capability of decoupling CPU, Mem and I/O functions from physical devices that provide them to increase their effective utilization, to enhance the flexibility in how they are utilized and to allow the dynamic management of logical instances
- **Storage Virtualization** : The capability of abstracting the physical location of data storage by presenting logical storage to the user for thus achieving location independence
- **Network-based Services**: The capability of manipulating service instances independent from the service device, thus providing flexibility in their usage
- **Network Infrastructure**: The capability of partitioning groups of networked resources providing logical isolation and common policy

Technology breadth: If there is a need for an access model that uses top-of-rack or blades, Cisco products are available meet the requirements of that design. For instance, if there is a need to load-balance firewall traffic, the Cisco FWSM can be sandwiched between ACS SM contexts. Cisco Data Center version 3 solutions ensure that future Data Center requirements are met as they arise.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The list of business drivers transforming the data center network environment is extensive. Main architecture drivers include the following:

- Server consolidation
- Server virtualization
- Application scalability and service availability
- Efficient workload management
- Data center uptime

Server Consolidation

Server Consolidation aims to control server hardware proliferation as well as standardize server hardware architecture. Server proliferation, also known as server sprawl, is the result of physical server growth. Server utilization, however, is not always efficient, yielding low average server utilization. Hardware standardization means establishing a set of guidelines for officially supported servers. Guidelines may dictate server sizes, numbers of CPUs or cores per server, memory, bus architecture, and overall I/O capacity. As a result, server growth continues but at a decreased pace (increased switch port density). Per server, capabilities continue to increase (more I/O capacity per server), which results in an increase in average load per server on the network.

https://t.me/learningnets

Server Virtualization

Server Virtualization is a form of server consolidation that helps control server sprawl and supports higher levels of server usage efficiency and application mobility. Higher server efficiency means higher average server utilization and therefore higher outbound load per server. Application mobility implies a virtual server infrastructure, where virtual servers can be moved across physical servers. There are implications such as a single physical server supporting different applications through multiple virtual servers, and multiple virtual servers requiring multiple IP addresses. Servers are also likely to need multiple I/O ports, VLAN trunking, and port channeling. Virtualized servers may require migration from Gigabit Ethernet (GbE) to 10 Gigabit Ethernet (10GbE) when multiple applications within a physical server demand more than 1 Gbps on a regular basis. Virtualized servers produce higher outbound traffic loads, which require a lower access layer over subscription target (ratio of server load to uplink capacity).

Application Scalability and Service Availability

Application scalability and service availability address both the need to control application response time jitter and the need to help ensure application uptime. Application response time predictability driven by end-user satisfaction means addressing application load congestion in the data center, whereas application uptime involves the need to protect application environments (security) and the need to help ensure network infrastructure uptime (network high availability).

Service availability is more focused on uninterrupted access to services expected by users or back-end systems. Besides the infrastructure components supporting services (network and servers), services are related to the applications and the associated service-level agreements (SLAs).

Efficient Workload Management

Workload Management is at the core of business automation. There are two main objectives: efficiency of resource utilization, and flexibility in achieving rapid on-demand service provisioning. Business needs call for applications to be placed in service in the shortest possible time frame. The rollout process efficiency depends primarily on the ability to add new physical or virtual servers to the network along with the required network policy changes (Layer 2, Layer 3, Layer 4 through Layer 7, and security). The implications are broad as resources are applied on demand based on a utility-like model. The physical and network environments must support network policy applied to any available portion of the physical environment (any rack with available power, cooling, and cabling capacity). Network policy is typically dependent on the most basic server identity (IP address), which implies that the server subnet (and its associated VLAN) should also be configurable in any physical rack when required. Layer 2 adjacency across the data center access layer becomes a key requirement.

Data Center Uptime

Data center uptime is critical. While this is not a new concept or requirement, data center downtime is more visible than ever because failures on denser server farms have wide repercussions. The resulting requirement has also wide ramifications: the network architecture and supporting switch architecture (software and hardware) must be highly scalable and highly available as uptime requirements arise.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

High Availability

High availability has and continues to be a major requirement. The expectations for high availability, however, are changing. A highly available network infrastructure requires switch-level and architecture-level high availability, so it is important to address these independently but cohesively. Switch-level high availability is addressed by hardware and software architectures and the capabilities they offer. Data center network architecture–level high availability centers on reliability, stability, and deterministic behavior, all of which need to be understood during a steady state and through normal operational procedures and failover conditions.

Data Center Networking Requirements

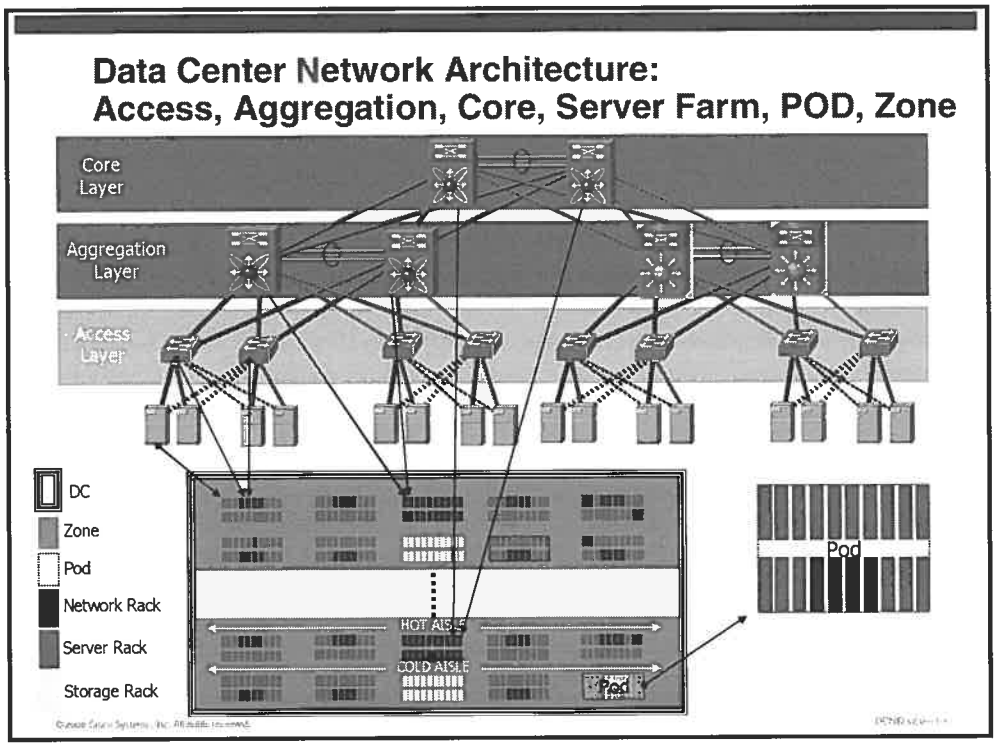
Data center network design and related topologies are influenced by requirements. The common set of networking requirements include higher density server farms, lower oversubscription, increased Layer 2 adjacency, increased high-availability needs, and integration of stateful services devices. Additional considerations include increased number of VLANs and subnets, increased number of MAC and IP address pairs, larger subnets, increased number of trunks and PortChannels, increased access control list (ACL) quantity, and so on, which are either interrelated or the result of larger networking requirements.

https://t.me/learningnets

Data Center Network Architecture

This topic describes the data center network physical and logical architecture.

Access, Aggregation, Core, Server Farm, POD, Zone



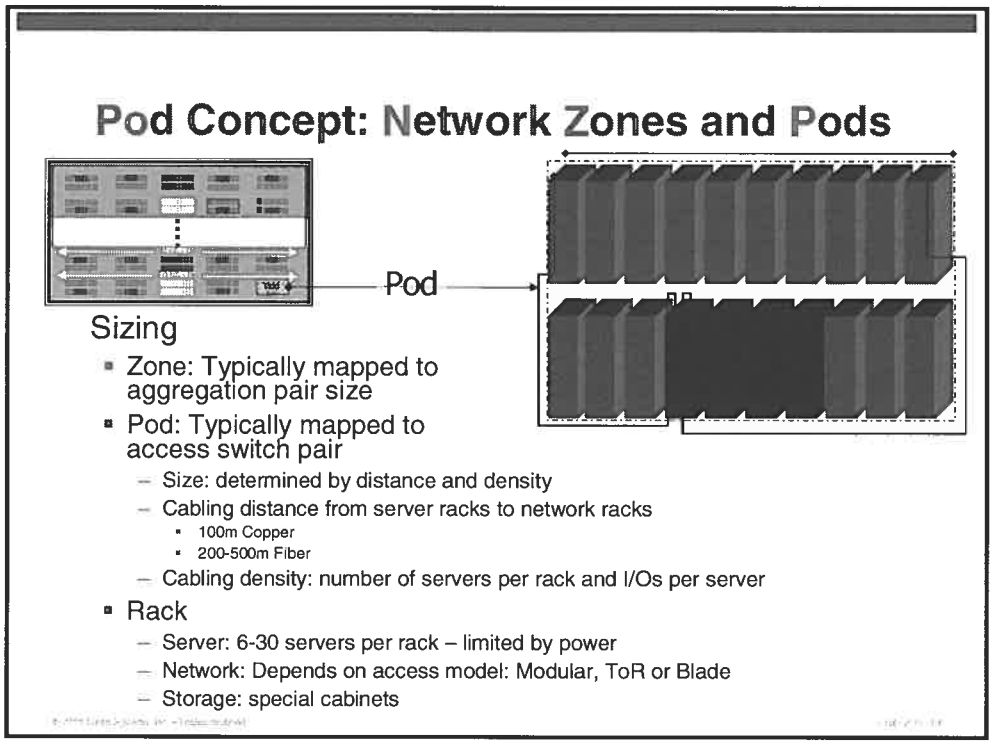
A server farm is a collection of servers maintained by an enterprise to meet transactional requirements far beyond the capacity of a single server. The POD is defined as a row of racks typically configured in an end-of-the-row design.

Most customers also define a construct called a Zone (sometimes called an area). The Zone (also referred to as an area or cluster) refers to a number of PODs within a single physical data center facility fed by common power, cooling, and cabling system. All of the components in a Zone meet at a common aggregation layer point. Power, cooling, cabling, and airflow usually present a greater constraint to Data Center design than network infrastructure components.

The product and model selections for the access layer are typically selected based on the density or based on cabling requirements. For example, data center designers will use the top-of-rack switches since the cabling dictates it.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The access layer switch selection in a particular Pod is determined by the density of that particular switch. The more advanced the selected switch is, the potentially larger the Pod can be, and potentially larger that Zone would be.

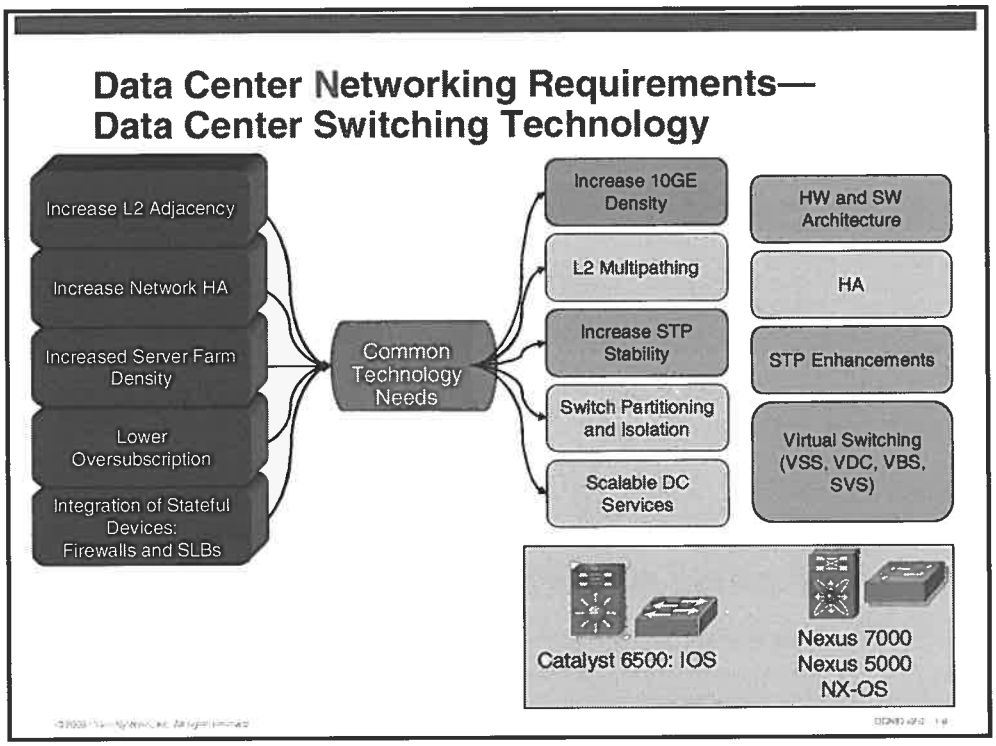
The limiting Pod factor is typically distance. So the size of the Pod is determined and limited by the cable distance. The worst-case distance scenario is where access switches are located at the farthest point apart from an aggregation switch. If the distance exceeds 100 meters, fiber optic connections are indicated. On the fiber connections side, which is usually the uplinks from the access switches to the aggregation layer, the size of the zone is a little bigger.

For each zone and pod there exists a typical airflow and the power cooling strategy. On average there are probably 6–12 servers per rack with approximately 22 ports per server, in the worst case. Every server can have one or two ports; of course, two usually stands for active-standby or active-active load balancing. Some servers may deploy four ports, two for active-standby and two for the backup applications.

Other design issues when deploying a POD or Zone include how many servers per rack can be deployed. There are two major constraints to server rack density, power and cooling. On the high-end where we have a lot of 1RU boxes, this translates to high power consumption and thermal load. In the enterprise space approximately 12–15 megawatts (MW) of power is required. In the SP space, between 30 and 40 MW of power are needed.

At the rack level, there is 3–5 kilowatts (kW) on an average basis. Cisco Nexus 7000 Series switches can, in the worst case scenario, consume 18 kW, which is required by an average cabinet at the peak time.

https://t.me/learningnets



The list on the left of the figure illustrate some of the top Data Center requirements: increased fabric capacity, L2 multipathing support, increased L2 stability, non-disruptive network service recovery mechanisms, and device VDC, VSS virtualization. These features increase data center high availability and data center design flexibility.

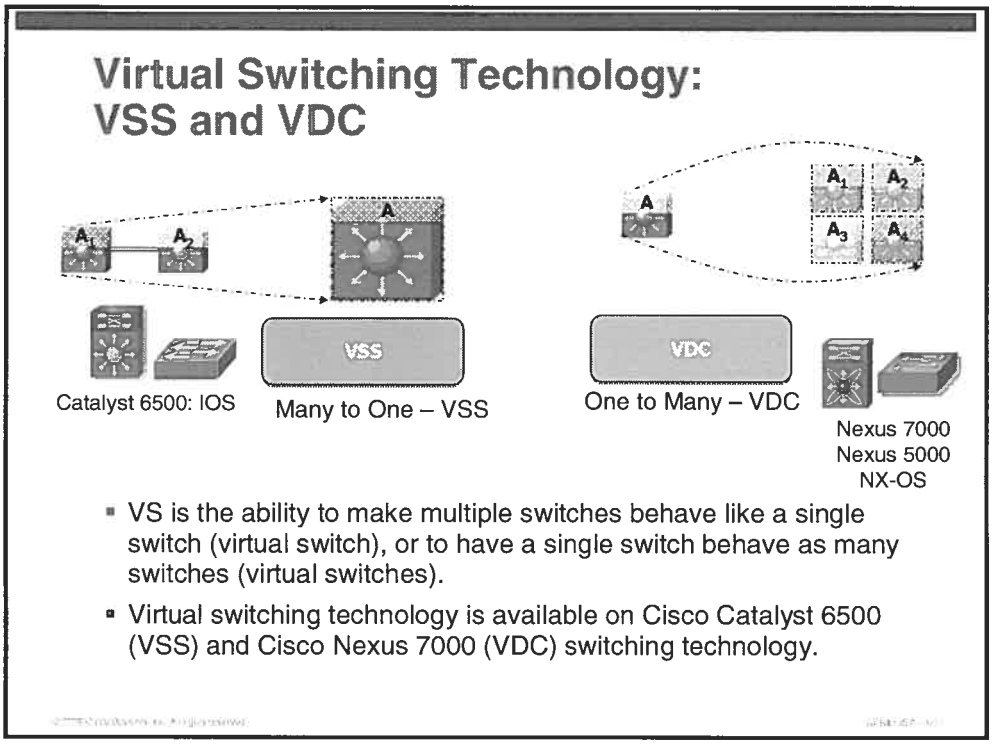
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Data Center Virtual Switching Technology

This topic describes the Virtual Switching Technology.

VSS and VDC



Virtual Switching (VS) is both revolutionary and evolutionary technology that supports a broad set of data center requirements. VS is revolutionary in that through innovative enhancements, it provides new architectural options to a data center network environment that are not currently available. VS is evolutionary in that it establishes a point of departure for next generation data center architectures from existing ones, and helps set the direction of technology in support of critical business drivers. VS technology is available on the Cisco Catalyst 6500 Series and the Nexus 7000.Series

https://t.me/learningnets

Virtual Blade Switching (VBS)

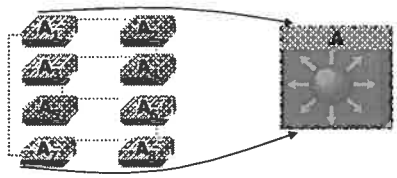
Virtual Switching Technology: Virtual Blade Switching (VBS)

Many switches look like one

- Two switches are physical
- One switch is virtual

Virtual Switch:

- All ports appear to be on the same physical switch
- Single point of management
- Single configuration
- Single IP/MAC
- Single control plane protocol instance
- Simplify infrastructure management
- One switch to manage

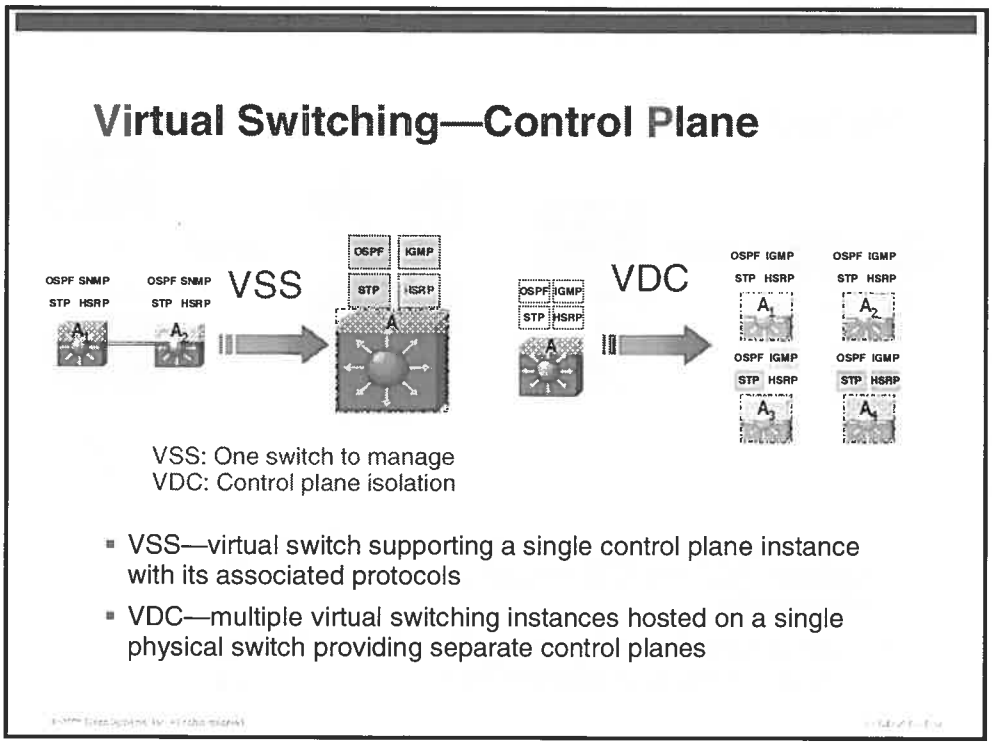


Many to one – VBS

© 2008 Cisco Systems, Inc. All rights reserved.

VBS provides new and simpler architectural option to a data center network. VBS technology is available on Cisco Catalyst Blade switches.

https://t.me/learningnets

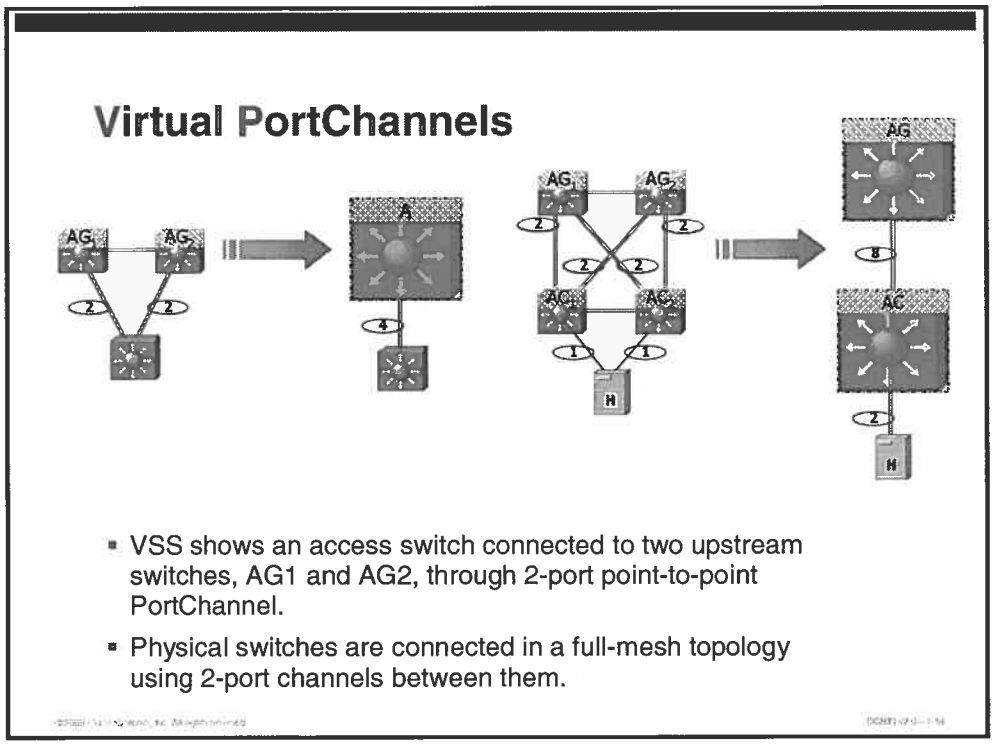


The preceding figure shows a view of how the control plane is handled on either kind of virtual switch. VSS shows a single virtual switch supporting a single control plane instance with its associated protocols. The single instance represents a single management and control plane point that carries a single configuration file and a single set of IP and MAC addresses.

VDC shows multiple virtual switching instances hosted on a single physical switch which provide separate control planes, thus separate control plane protocol instances. All instances are isolated from one another.

VDC shows a more complex topology with two sets of virtual switches, one at the aggregation layer and one at the access layer. All the physical switches are connected in a full-mesh topology using 2-port channels between them. Through virtual switching and virtual PortChannels, the equivalent environment contains two virtual switches, AG and AC, interconnected through an 8-port virtual PortChannel that contains the individual port channels between the physical switches.

https://t.me/learningnets



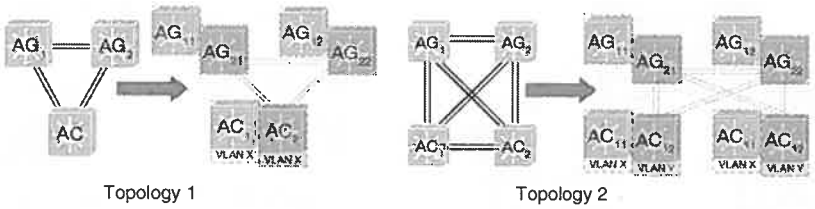
VSS and VDC Topology using MCEC

VSS shows an access switch connected to two upstream switches, AG1 and AG2, through 2-port point-to-point PortChannel. By utilizing virtual switching and virtual PortChannels, the equivalent topology becomes one in which the access switch is connected through a 4-port channel to the virtual switch. The virtual PortChannel is established between the access switch and AG1 and AG2, forming a multichassis EtherChannel.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Virtual Switching—Network Topologies



Topology 1 Topology 2

- Virtual switching and virtual PortChannel technology allows a Layer 2 network to use all available Layer 2 paths concurrently.
- Hosts and switches that are connecting the virtual switches use any available L2 path concurrently.
- Isolation at the topology level is done by assigning specific ports to particular virtual switching instances.

Topology 1 shows switches AG1, AG2 and AC which collectively support two virtual switching instances. Two instances of VLANx map to each of the virtual switching instances, thus creating two parallel topologies both supporting the same VLAN, but completely isolated.

Topology 2 shows a more complex topology that essentially supports two virtual switching instances on access and aggregation switches however each supporting a different VLAN.

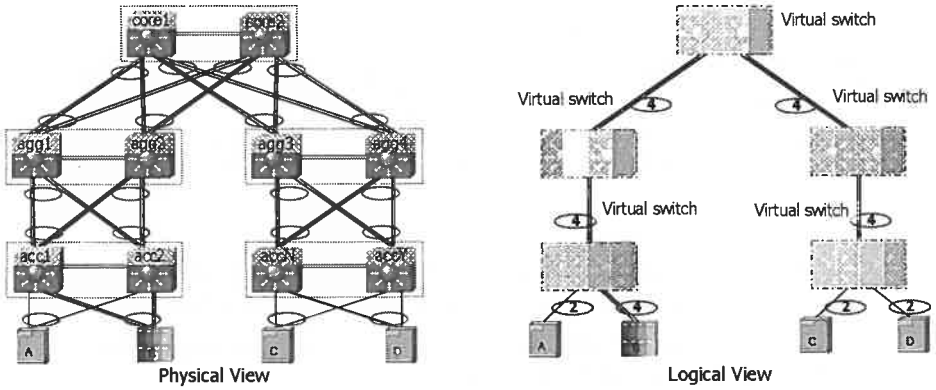
The physical separation is achieved by having ports dedicated to a particular switch instance.

There are additional benefits on Layer 2 topologies. In a typical Layer 2 topology, Spanning Tree Protocol (STP) provides loop avoidance and allows path resiliency. A byproduct of path resiliency is that only one path is active while all other redundant paths are blocked. This is also true for dual-homed servers that use network interface card (NIC) teaming, with one of the adaptors active and the second on standby.

Virtual switching and virtual PortChannel technology allows a Layer 2 network to use all available Layer 2 paths concurrently. Hosts and switches connecting the virtual switches use any available L2 path concurrently. The traffic distribution is granular to a flow (source and destination MAC addresses, source and destination IP addresses and Layer 4 port number). In addition to non-blocking paths, virtual PortChannels capabilities provide both switch and path redundancy.

https://t.me/learningnets

Topology—Virtual Switching and Virtual PortChannel



- Virtual switching and virtual PortChannel capabilities across multiple Layer 2 tiers: access, aggregation, and core
- Optimized L2 pathing through virtual PortChannels
- Less reliance on STP
- Isolation of L2 domains

The preceding figure shows the physical and logical views of a large Layer 2 network using VS technology. The diagrams show a generic multilayer switch providing virtual switching and virtual PortChannel capabilities across multiple Layer 2 tiers: access, aggregation, and core. The physical view shows a number of dual-homed servers that have established cross-chassis channel connections with a pair of access switches. For instance, server B, which is a collection of virtual servers, uses two groups of two I/O ports to connect to the access switches (two to each switch). These connections effectively form a 4-port channel when connecting to a set of access switches that are VS capable. Through this mechanism, server B is capable of using all links concurrently. Likewise, a non-virtualized server using a single IP and MAC address is also capable of using the two links concurrently. In either case, the servers are unaware of the fact that they are connecting to two distinct switches through a single PortChannel.

In a switch-to-switch environment, a switch pair cross-connection, as shown in the preceding figure, is established between all switches. In this case, acc1 and acc2 have two links to both agg1 and agg2. The equivalent logical topology shows a 4-port channel connection between two virtual switches: one logically formed by the access layer pair, and the other one by the aggregation switch pair. The logical view of the diagram shows a loop-free 4-port channel connection.

A Layer 2 topology built taking advantage of VS technology contains no loops, no blocking ports, and no blocked paths. The forwarding path is established by the VS mechanisms, and since STP is not used to help determine the forwarding topology, it is used strictly as a fail-safe backup protocol.

© 2008 Cisco Systems, Inc.

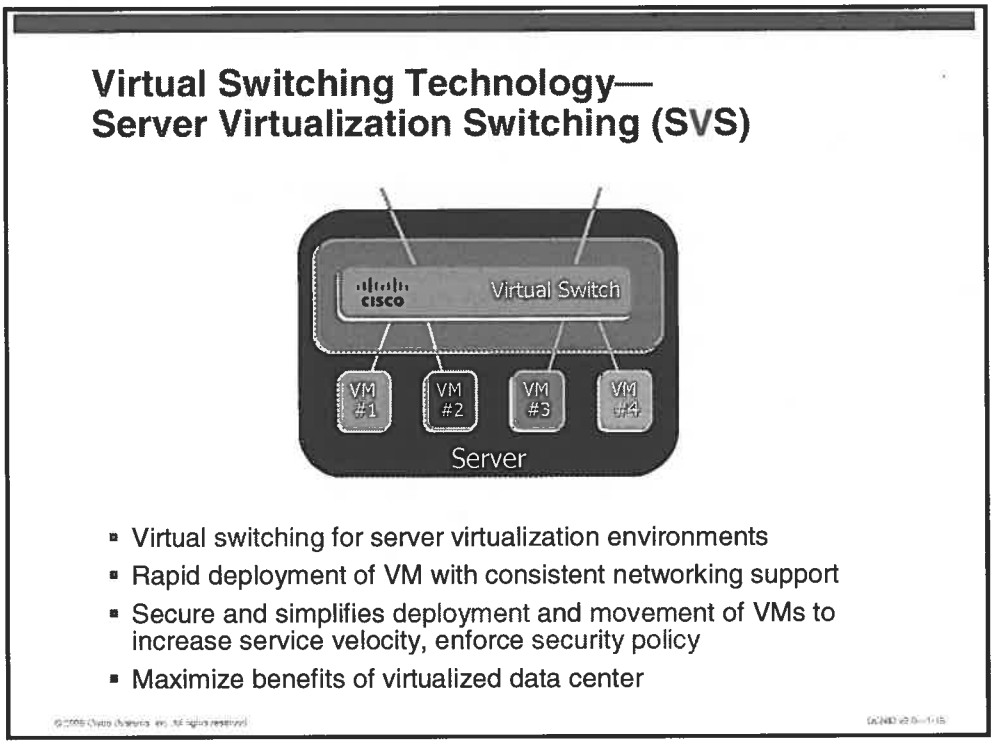
https://t.me/learningnets

Key characteristics of New Topology Using Virtual Switching:

- Optimized L2 pathing through virtual PortChannels
  - Increase bandwidth usage on: server to switch and switch to switch
  - Loop-free forwarding paths
- Less reliance on STP
  - Loopfree topology established through Virtual Switch mechanism
  - STP is strictly used as a fail-safe mechanism
- Isolation of L2 domains
  - Separate logical topologies
  - Distinct STP topologies

https://t.me/learningnets

Server Virtualization Switching (SVS)



SVS Characteristics

- Enables consistent network features to all servers (physical and virtual)
- Maintains familiar network/server operations and management models, simplifying troubleshooting and maintenance
- Safely allows Virtual Media (VM) from different functional units to mix on the same physical server
- Deploys VM just as easily and confidently as physical servers with the same segmentation and security policies

Benefits

- Rapid provisioning / deployment of VMs
- Preserving virtual server segmentation, and security policy
- Enhance diagnostic and monitoring capability
- Application driven migration of network policy

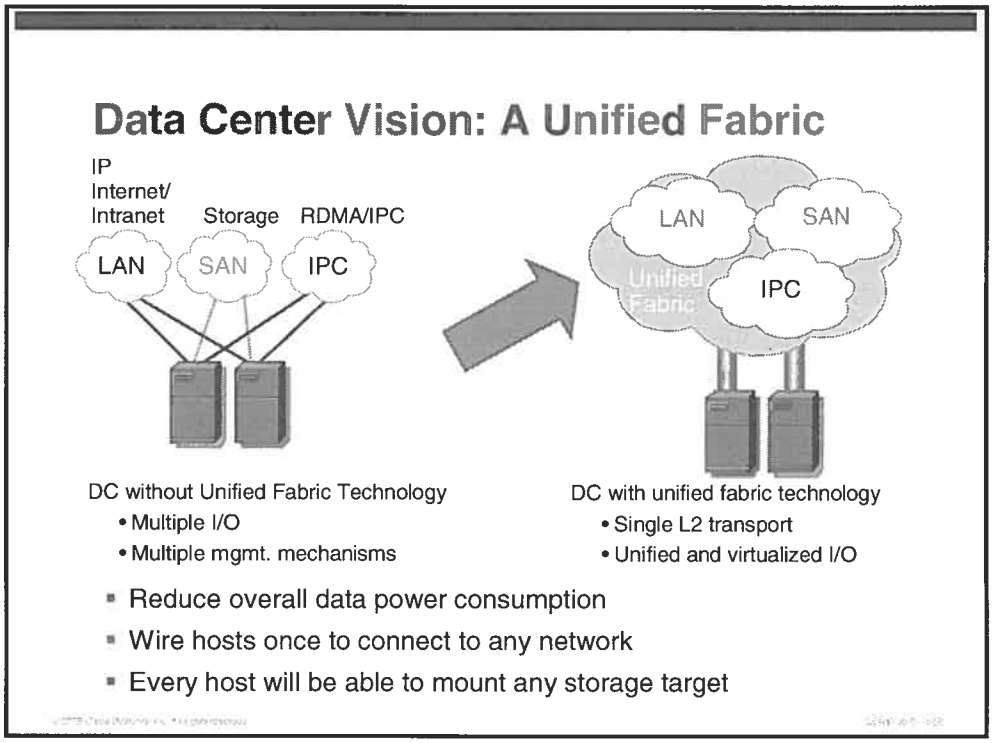
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Unified Fabric and Unified I/O

This topic describes the basics and benefits of unified fabrics and unified I/O in the data center.

Data Center Vision: A Unified Fabric

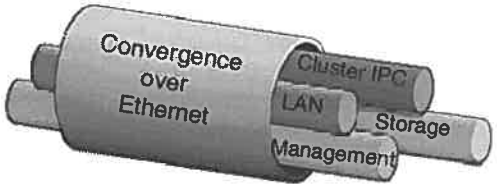


The benefits of deploying unified fabrics include:

- Reduce overall data power consumption.
- Extend the lifecycle of current data center.
- Wire hosts once to connect to any network—SAN and LAN. Faster rollout of new apps and services.
- Every host will be able to mount any storage target. Drive storage consolidation and improve utilization.
- Rack, Row, and X-Data Center VM portability become possible.

https://t.me/learningnets

Convergence over Ethernet



Converge functions:

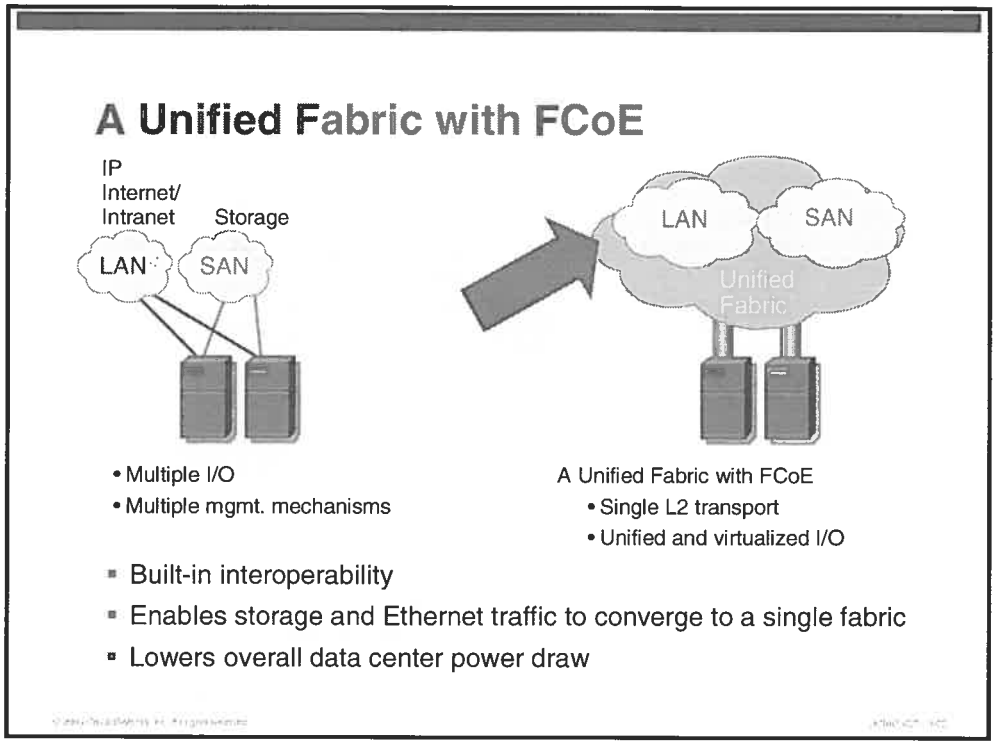
- Multiple functions (SAN, LAN, IPC, mgmt.) can be consolidated to a single fabric type
- Blade server storage connectivity (low cost)
- Consolidated ports
- Leverage Ethernet pervasiveness, knowledge, cost and volume
- Leverage existing standard Ethernet equipment
- Lower TCO – fast and low-cost scalability

Converged Ethernet should enable support of:

- Efficient transport of data, storage (FC), iSCSI, RDMA processes and management traffic and TOE
- High throughput/density, minimized transport delay
- Lossless fabric (no frame loss, storage traffic requirements)
- Different ports consolidated on the common converged Ethernet Port Adapter
- Multiple fabrics consolidated on the common converged fabric
- Enable low-cost and fast data center scalability

© 2008 Cisco Systems, Inc.

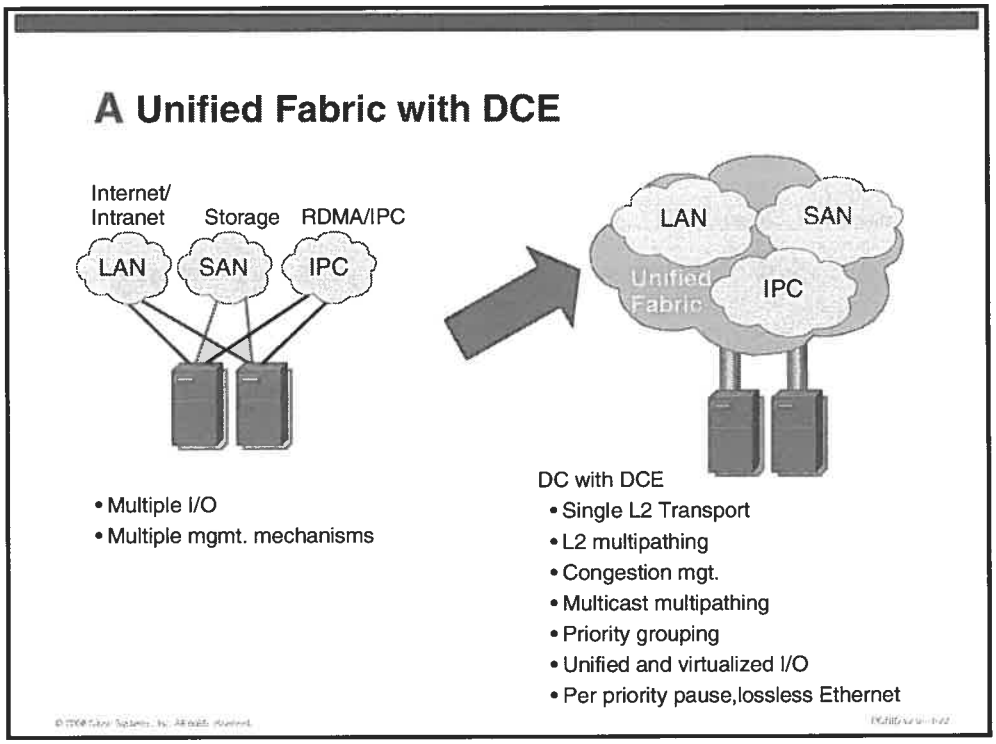
https://t.me/learningnets



Fibre Channel over Ethernet (FCoE) characteristics are:

- Reduce overall data power consumption. Extend the lifecycle of current data center.
- Wire hosts once to connect to any network—SAN and LAN. Faster rollout of new apps and services.
- Every host will be able to mount any storage target. Drive storage consolidation, and improve utilization.
- Rack, row, and cross–data center VM portability become possible.

https://t.me/learningnets



Data Center Ethernet (DCE) is an architectural collection of Ethernet extensions designed to meet new data center application requirements and data center converged networking .DCE is not an open standard but is rather a collection of technologies that are useful in the data center. A DCE enables unified converged fabrics that will run the SAN service, the LAN service, and the IPC type service over a the common physical infrastructure which is based on an improved Ethernet protocol.

DCE represents an L2 transport providing the same functionality such as a SAN, LAN, and an IPC network. DCE enables the following functions in addition to unified I/O:

- **Layer 2 multipathing**: Spanning Tree is replaced with IS-IS to enable multipathing of unicast traffic at Layer 2. IS-IS is better than Spanning Tree because it allows an edge switch to have active links to multiple core switches. In addition, it will always pick the shortest path through the network.
  - For environments where peer-to-peer or server-to-server traffic dominates, this optimal forwarding and maximum cross-sectional bandwidth are very important.
  - In addition, STP is considered to be unreliable and a limit to scaling of layer 2 domains. Eliminating it will allow customers to build larger Layer 2 networks.
- **Per Priority Pause for flow control**: Also called "Priority Flow Control," is the ability for the receiving end of a link to signal to the sending end to stop sending packets. This is an extension of the IEEE 802.3x Ethernet PAUSE standard that allows sending a separate PAUSE for each of eight different priority levels. Thus, the low priority traffic can be slowed down without affecting high priority traffic.
  - This flow control mechanism allows us to provide a class of service in the network that requires no frame drop or "no-drop" service by having the receiving port signal to the send port to stop sending traffic when the receivers' buffers are nearly full. A no-drop service allows us to carry Fibre Channel packets over the DCE network without TCP flow control or the FCoE standard.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **Congestion Management**: Cisco has built congestion management mechanisms into our DCE switches and DCE NICs. This congestion management system will detect congestion in switches and signal back to the source of the congestion at the edge of the network (causing it to reduce its transmissions). This pushes congestion to the edge of the network where it can be managed more effectively and will minimize the impact of one data stream on another (sometimes referred to as 'collateral damage').

  Congestion Management allows components to use smaller buffers and run them in a state that is near empty (as opposed to near full buffers as TCP does). This near empty state reduces latency and enables the switch to better absorb instantaneous bursts of traffic before the flow control mechanism (listed above) is engaged.

- **Multicast multipath**: Like multipathing at Layer 2 mentioned above which applies to unicast traffic, we are providing a similar capability for multicasting at Layer 2. The difficulty with multicast is that it ultimately must follow a tree structure. The trick that we are doing is to build multiple trees and when a multicast packet comes in, we will load-balance it across those trees. If the trees cross different links, the multipathing is achieved.

- **Priority Grouping**: Priority Grouping is a sophisticated scheduling algorithm for traffic across the eight priority levels. Although it is pretty general in its definition a specific example can illustrate what it allows the customer to do. If a normal IP traffic, IPC traffic, and Storage (FCoE) traffic is carried across a DCE link they can be grouped into three groups: IP, IPC, and Storage respectively. IP traffic might be assigned three priority levels, IPC might get three priority levels, and Storage might get two priority levels for a total of eight available priority levels.

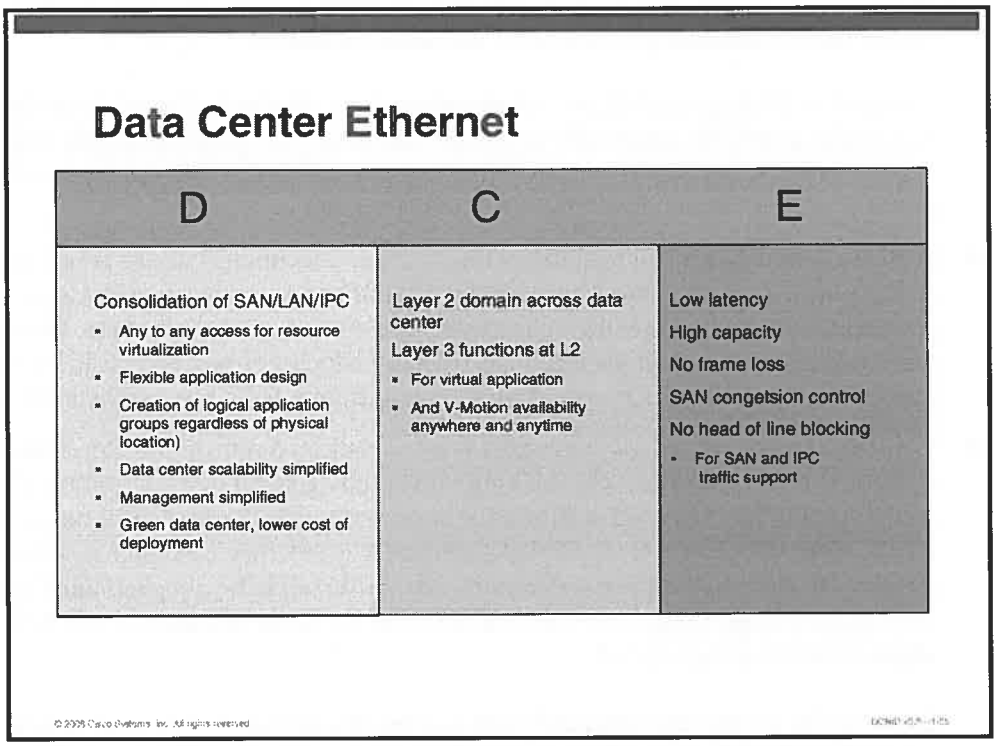
  Between the groups, the link bandwidth can be allocated and each group can be assigned one of the priorities to be high priority. The high priority of each group will be scheduled before the non-high priorities of each of the other groups as long as the total bandwidth for its group has not been exceeded. This feature is called priority grouping with priority promotion.

Note	FCoE as one of the components of DCE. FCoE in a very important technology in the data center that enables a unified fabric and I/O consolidation, but it will be running over a DCE infrastructure. FCoE can work over any Ethernet infrastructure that can guarantee no packet loss like DCE does.

https://t.me/learningnets

Data Center Ethernet

This topic describes the characteristics and features of the data center Ethernet fabrics.

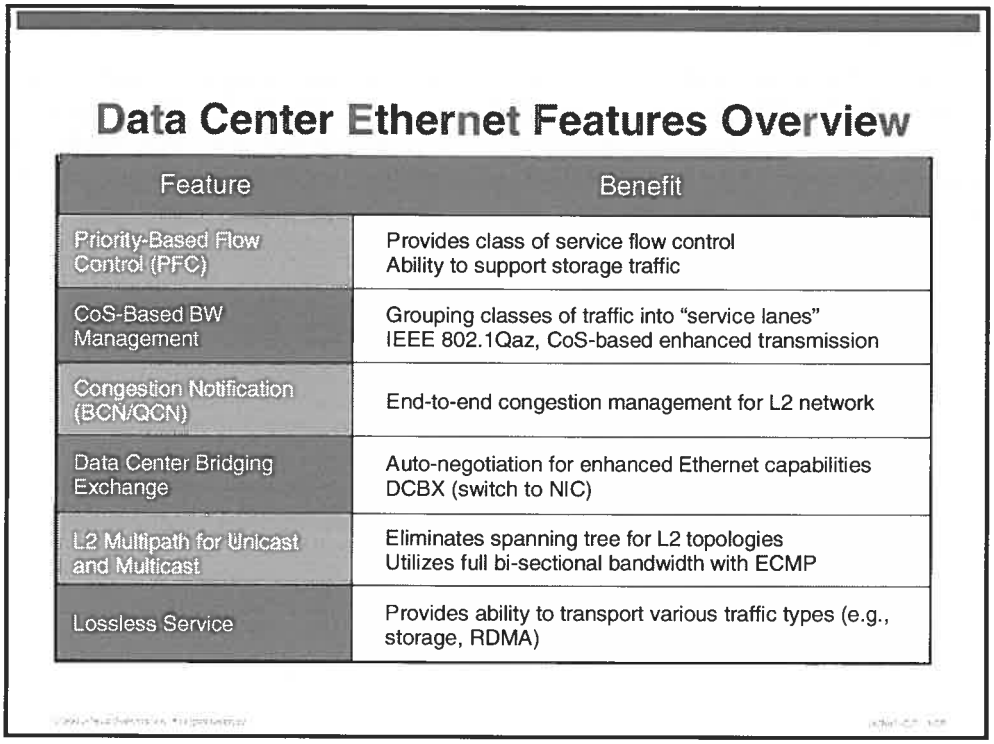


Application availability dictates new data center networking forms. Next generation data center must accommodate these new application requirements:

- Fast application deployment
- Lower overall data center power draw
- Fast Layer 2 domain must be extendable across the data center without boundaries
- Flat networks (Layer 2 adjacency) are important for Application availability such as: Hypervisor technology (VMware, Zen) and V-Motion that require L2 adjacency between these virtual machines in order for them to work effectively.
- Application performance, bandwidth, latency, and resource utilization requirements are always increasing: grid and clustered applications are being increasingly deployed in the data center
- Frame propagation/switching latency reduction to the micro-or-millisecond range for IPC or storage networking
- No head of line blocking to ensure consistent frame propagation latency
- No packet drop: Lossless Ethernet to enable FC SAN transport
- Consolidation of different data center traffic types—client-to-server, server-to-server, server-to-storage—support for short, steady or bursty messages types (seen in SAN) storage, Ethernet, IP, and HPC traffic can converge to a single network data center. Consolidation enables simple, cheaper data center scalability and design flexibility.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Priority-Based Flow Control (PFC)

Priority based flow control (PFC) is CoS-based PAUSE-ing based on CoS bits in the IEEE 802.1p standard. CoS-based PAUSE-ing is called priority-based flow control, which enables selective pausing of the traffic waiting in the buffer to be sent across DCE ISL.

CoS-Based Bandwidth Management

CoS-based Bandwidth Management or priority grouping is the ability to group classes of traffic into the virtual links to prioritize them and assign them link bandwidth. For example, virtual lane (VL) 1 is assigned 1 Gbps and represents LAN traffic, the IPC traffic is assigned VL 2 and 2 Gigabits of DCE ISL bandwidth, the VL 3 goes to SAN traffic and is assigned the rest of the DCE ISL bandwidth. There will be a traffic shaper at the end of the link enforcing the bandwidth allocation rations per priority group or per virtual links. The standard is called IEEE 802.1Qaz or Enhanced Transmission Selection protocol (ETS).

Congestion Notification or Congestion Management

Congestion notification or congestion management is based on the fiber channel congestion control. DCE cannot count on the TCP flow control for congestion management purposes. Data Center Ethernet (DCE) cannot depend on TCP flow control for congestion management purposes. Congestion management functionality is built into L2 hardware in modern data center environments. The FC SAN pushes the congestion impact to the edge of the network where the source of congestion exists and slows down the rate at which the servers are injecting the data into the network, thus causing the congestion. A similar function will be deployed by the DCE congestion notification and management.

Data Center Bridging Exchange (DCBX)

Data center bridging exchange (DCBX) is a negotiation mechanism on the DCE link between the host and switch, or switches and switches. Every VL is configured with some types of VL services, which are negotiated by the DCBX protocol. It is an automatic configuration protocol.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

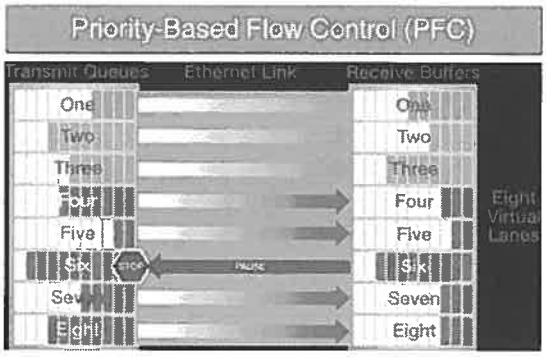
Multipathing at Layer 2

Layer 2 Ethernet networks use STP for frame forwarding. Using a link state algorithm such as FSPF or OSPF or IS-IS without using STP can simplify topologies and utilize the full bi-sectional bandwidth that is available in the core data center network. DCE will enable to use L2 multi-pathing based on these features in the future and lossless services that will be selectively applied to VL lanes.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Ethernet Features



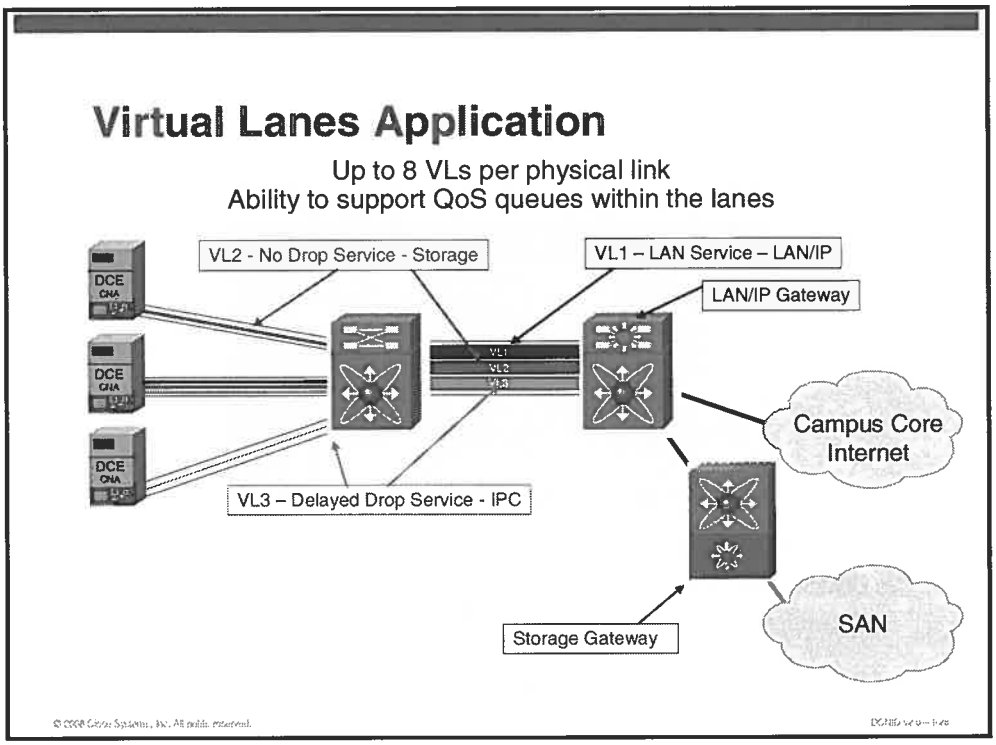
- Enables lossless fabrics for each class of service
- PAUSE sent per virtual lane when buffers limit exceeded
- Network resources are partitioned between VLs (e.g., input buffer and output queue)
- Switch behavior is negotiable per VL

CoS-based PAUSE-ing is based on CoS bits in the IEEE 802.1p standard. CoS-based PAUSE-ing is called priority-based flow control, which enables selectively pausing the traffic waiting in the buffer to be sent across DCE ISL. A physical link can be split into up to eight selective VLs, using the CoS bits. In this example, the traffic in VL 6 is storage traffic that is being PAUSED back. Priority flow control on the particular VL is enforced in order to achieve lossless behavior.

An NIC card or a converged network adaptor might be supporting per-priority based flow control. Note that when the interface buffer limit is exceeded, the PAUSE is sent for each virtual lane. Congestion management is enforced before the PAUSE process. Congestion management will first slow down the traffic injected into the fabric and then the PAUSE mechanism will be applied to ensure the lossless behavior.

© 2008 Cisco Systems, Inc.

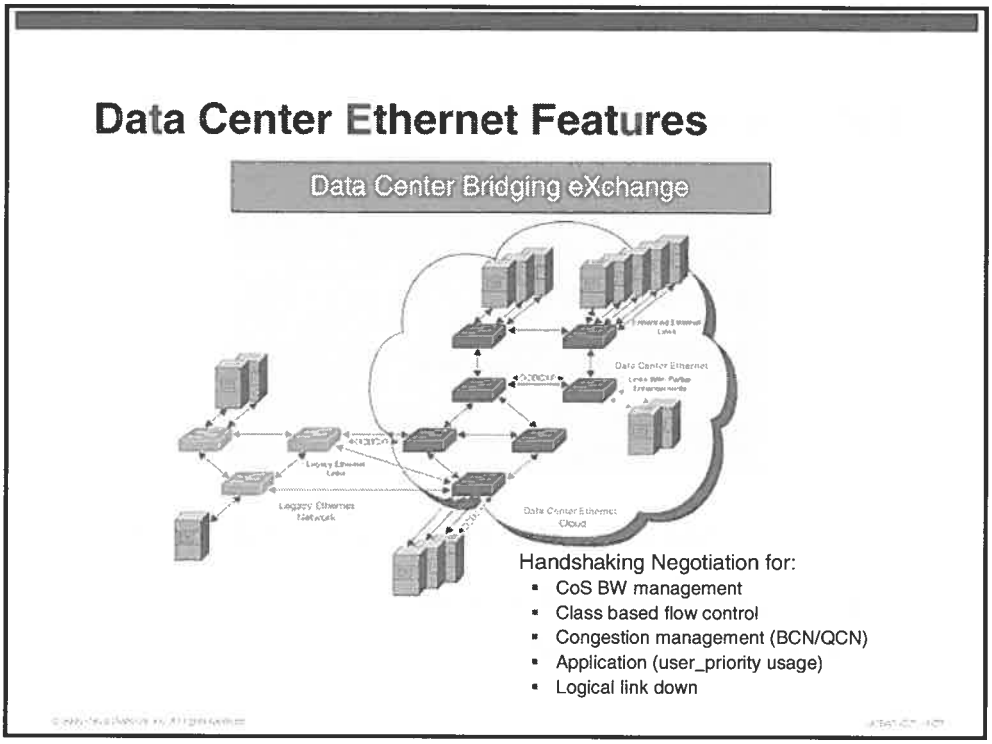
https://t.me/learningnets



There are three types of application traffic with different transporting requirements that can be met by assigning a separate Virtual Lane to each service. Some host adaptors need no drop Service since they are carrying Fibre Channel traffic. Another host adapter may need all of the services that are supported. The last host may need just the delay drop services. All three services can be deployed across the common DCE ISL link, for example across a 10G link.

© 2008 Cisco Systems, Inc.

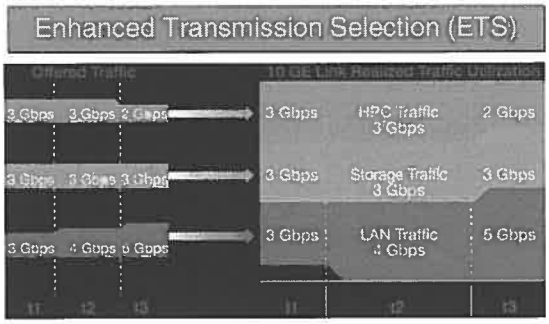
https://t.me/learningnets



This diagram is presenting an example of the data center bridging exchange protocol that can be enabled on the DCE host to switch and switch to switch connections. The Bridging Exchange is a protocol used to negotiate whether a link is up or down, class-based flow control policies, congestion management functionality that should be available on a virtual lane, CoS bandwidth management, FCoE support, and lossless lane support.

https://t.me/learningnets

Data Center Ethernet Features (Cont.)



- Enables intelligent sharing of bandwidth between traffic classes control of bandwidth
- Being standardized in IEEE 802.1Qaz
- Also known as priority grouping

© 2008 Cisco Systems, Inc. All rights reserved.

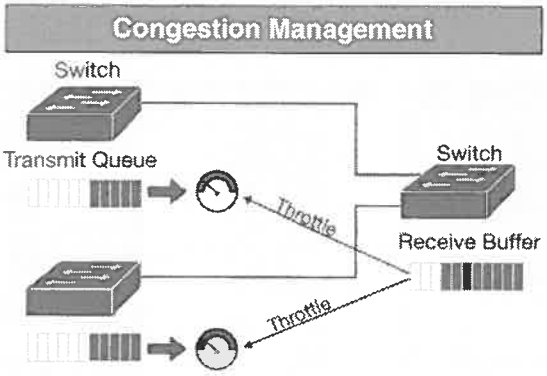
Enhanced transmission selection (ETS) or priority grouping is actually a version of the priority grouping initial proposal. IT provides the ability of deployment of bandwidth management groups for different traffic types: SAN, LAN and IPC. Each traffic group is assigned a specific bandwidth from the common channel. When 1Gbps bandwidth is still unassigned and free and storage traffic starts bursting, it can consume the available bandwidth. ETS is being standardized in the IEEE known as priority grouping, and provides the shaping or bandwidth management capabilities on a per VL basis.

Note When a given load in a traffic class does not fully utilize its allocated bandwidth, enhanced transmission selection allows other traffic classes to use the available bandwidth. This helps accommodate the bursty nature of some traffic classes while maintaining bandwidth guarantees.

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

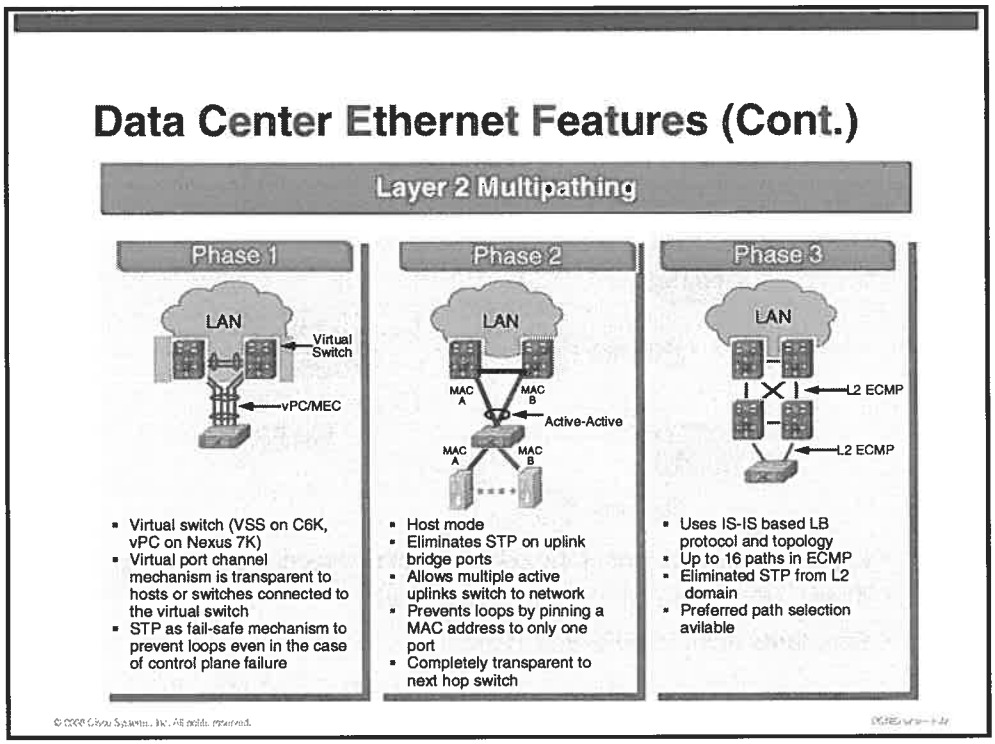
Data Center Ethernet Features (Cont.)



- Moves congestion out of the core to avoid congestion spreading
- Allows end-to-end congestion management
- Standards track in IEEE 802.1Qau

DCE congestion management is applied at Layer 2. As soon as congestion is detected on the switch port, the congestion notification is sent back to the edge (the last switch) of the DCE fabric where the source server of the congestion resides and slows down the server data injection into the fabric. The congestion detection and the reaction time are quicker in DCE fabric than in MDS fabric. It is faster than the PAUSE mechanism reaction to congestion. This is a standard based protocol: IEEE 802.1Qau.

https://t.me/learningnets



Layer 2 Multipathing Implementation Phase 1

Multipathing at Layer 2 is called VSS and is available in the Catalyst 6500 Series Switches. Multi-pathing capability can be deployed from the access switch to the distribution switch at Layer 2. The Nexus 7000 Series Switches will support Multi-pathing as well but it will be called Virtual PortChannel. The difference between the VSS and virtual port channel is that VSS does a lot of other things that virtual port channel doesn't support. One configuration file is managed as one unit, whereas virtual PortChannel is just that multipathing capability on the Nexus 7000 Series switch. MEC in the initial phase will use STP as a failsafe mechanism but it is used only as a failsafe mechanism on the Catalyst 6500 Series Switches.

Layer 2 Multipathing Implementation Phase 2

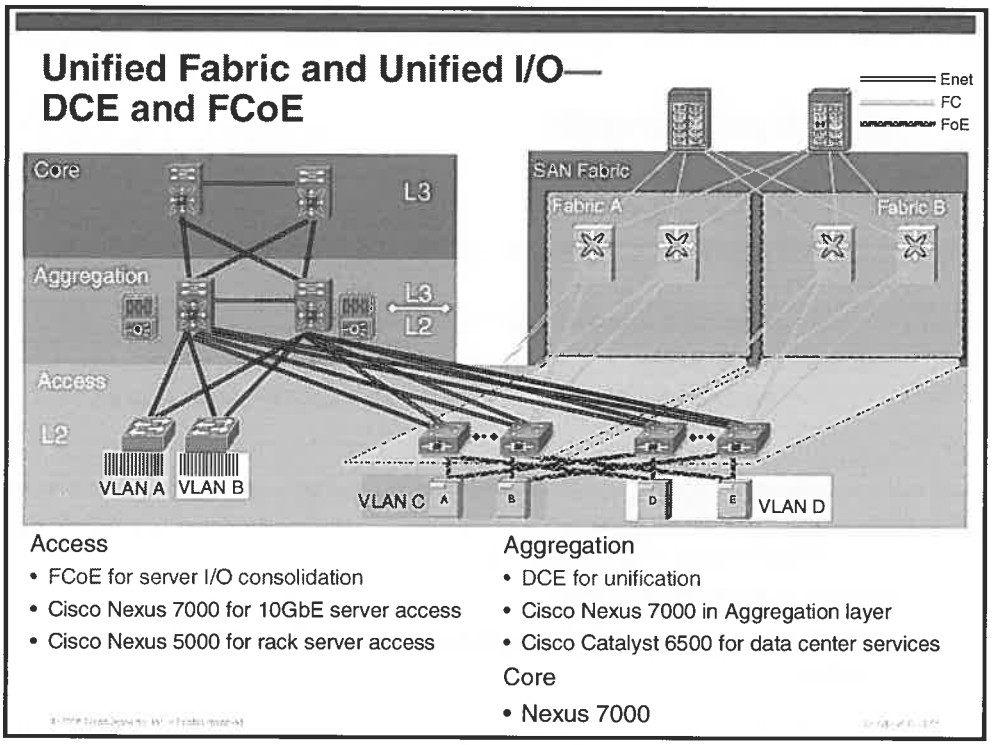
In the Phase 2, a real implementation of the multipathing with the uplink member active and with multipathing across both of the showed connections will be supported. MAC address is used for load-balancing and sticks to a particular session. Traffic from MAC A is always sent along the red link; and traffic from MAC B is always sent along the blue link. It is not important what the upstream switch is—the connection will still work. So it will be possible to have multipathing deployed from the rack switch or edge switch to the third-party switch in the core. Forwarding loops are prevented by pinning up the MAC addresses and will be completely transparent to the next-hop switch, and it now provides the active/active forwarding capability without relying on the spanning tree blocking process.

Layer 2 Multipathing Implementation Phase 3

The DCE variant of multipathing initial implementation will use ISIS. It will be IS-IS-based load-balancing across up to 16 paths in one ECMP channel. Spanning tree will be redundant and eliminated from the Layer 2 domain since IS-IS deploys loop-free topology. Preferred path selection will be possible to enforce certain traffic path selection using certain routes. This will be actually classic routing strategy at Layer 2.

© 2008 Cisco Systems, Inc.

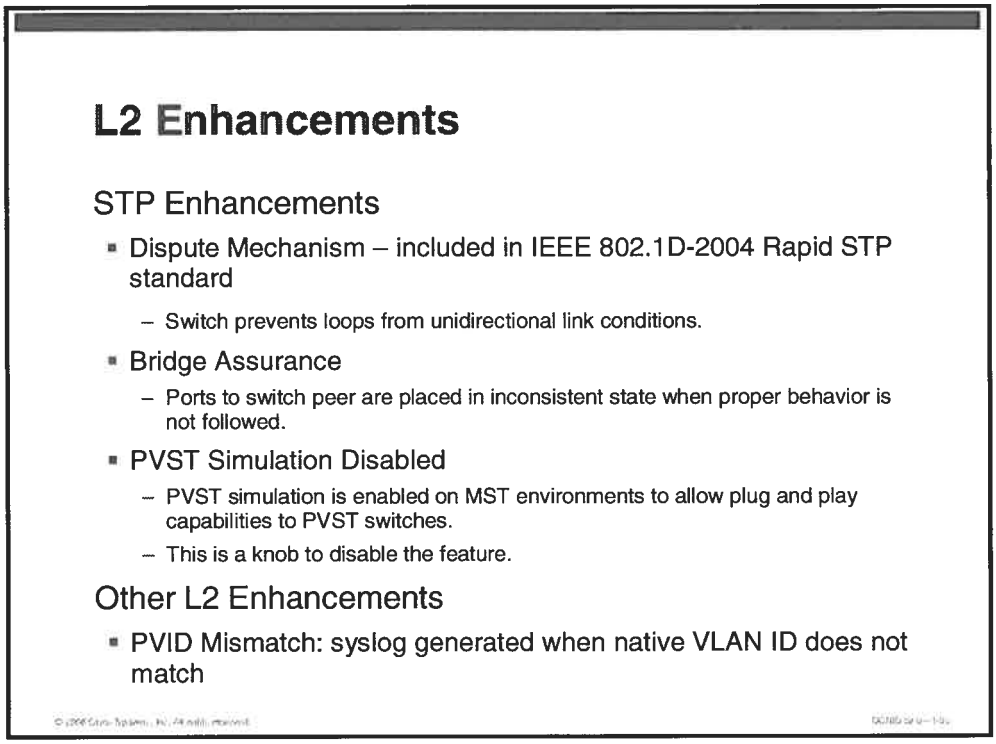
https://t.me/learningnets



A Unified fabric is the end state network where LAN, SAN, and IPC traffic are converged onto a single network infrastructure. FCoE is the enabling technology for delivering a unified fabric and I/O interfaces. DCE provides seamless integration of Ethernet-based data center fabrics with existing FC SAN environments. DCE is the foundation for delivering a Unified Fabric. DCE will also appeal to customers that are not planning for consolidated I/O due to its L2 multipathing capabilities and its lossless service behavior implementation of the Cisco Nexus 5020 Series switch—a single box connected to the Ethernet side on one side, and to a single fabric on the other side. The rules of how you build a SAN will not change because the Nexus 5020 Series switch is doing the unified I/O. The host itself will continue to be connected to two fabrics, just one 10 GigE pipe on one, one 10 GigE pipe on the other.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



STP stability is increased by addressing typical failures seen in operational environments. Many of these failures are related to the immediately available nature of the STP. Solutions to the various STP concerns therefore include dealing with some plug-n-play aspects of immediate availability through the use of advanced features that address specific conditions. In addition to the advanced features discussed here, the use of virtual switching technology to build loop-free deterministic topologies lessens the reliance on STP. Effective STP becomes a background fail-safe protocol that is not actively determining a loop-free topology.

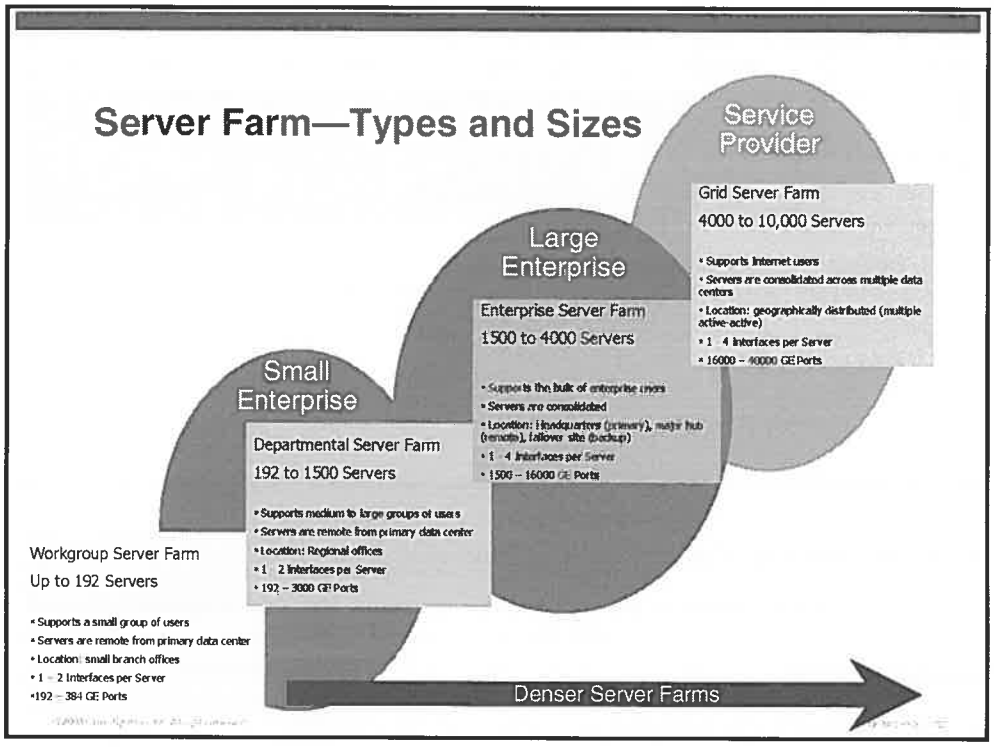
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Sizing

This topic discusses some of the data center sizing parameters.

Server Farm—Types and Sizes

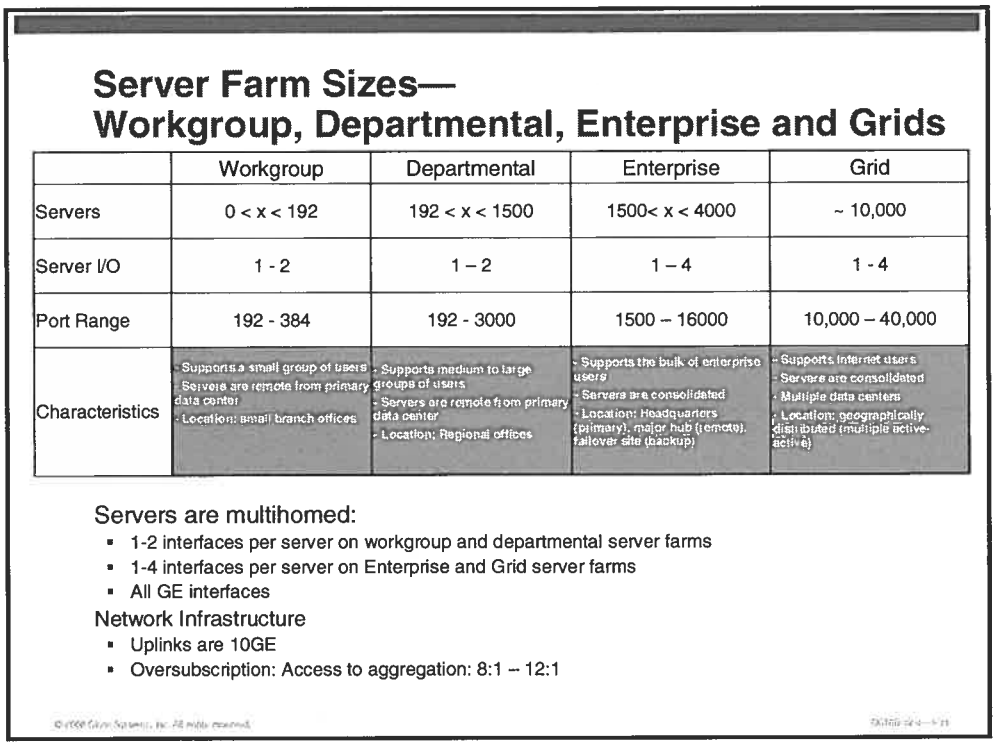


Typical server farm growth and data center consolidation produce denser server farms. Consolidated data centers exhibit faster growth than distributed ones. While the rate of growth is slowed down by server virtualization, growth is still experienced. Server growth requires a flexible network architecture to accommodate additional port density without affecting uptime or oversubscription. Server growth equates to increased switch port density at a rate determined by the number of physical interfaces per server. Virtual server growth equates to a higher number of IP addresses and their associated MAC addresses, a higher number of trunks and PortChannels from server to switch, and use of more I/O ports per server (4 x GbE) or higher I/O capacity (2 x 10GbE).

More IP addresses require more subnets and VLANs, which in turn drive additional service policy: ACLs (security and QoS), load balancing policy, firewall policies, and so on. More capable servers coupled with increased server quantities require higher network transport capacity, which must be preprovisioned to a reasonable point. Such a point is determined by the target oversubscription that is to be maintained as network growth is experienced. Increasing numbers of access switches demand higher port density at the aggregation layer, which in turn drives higher density at the core layer.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



There are a number of servers and a number of I/Os per server.

```
10,000 ports at four per port = 40,000 ports.
```

Example of Financial Business

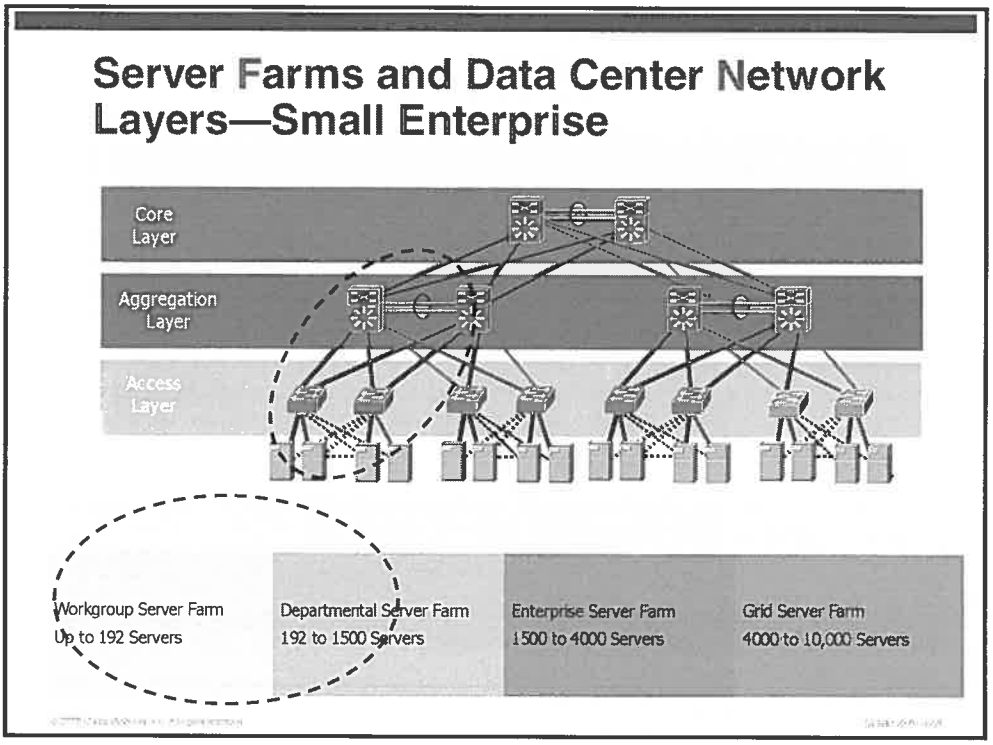
A customer had roughly 8,000 servers for a total of 46,000 ports on a single data center. Even though the customer has translated to 300 odd servers based on 80 physical servers, 80 servers have 22 physical interfaces each and 80 VMs per physical server. When discussing the virtual servers we are talking about a different level of oversubscription and a different level of high availability.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Server Farms and Data Center Network Layers

Small Enterprise

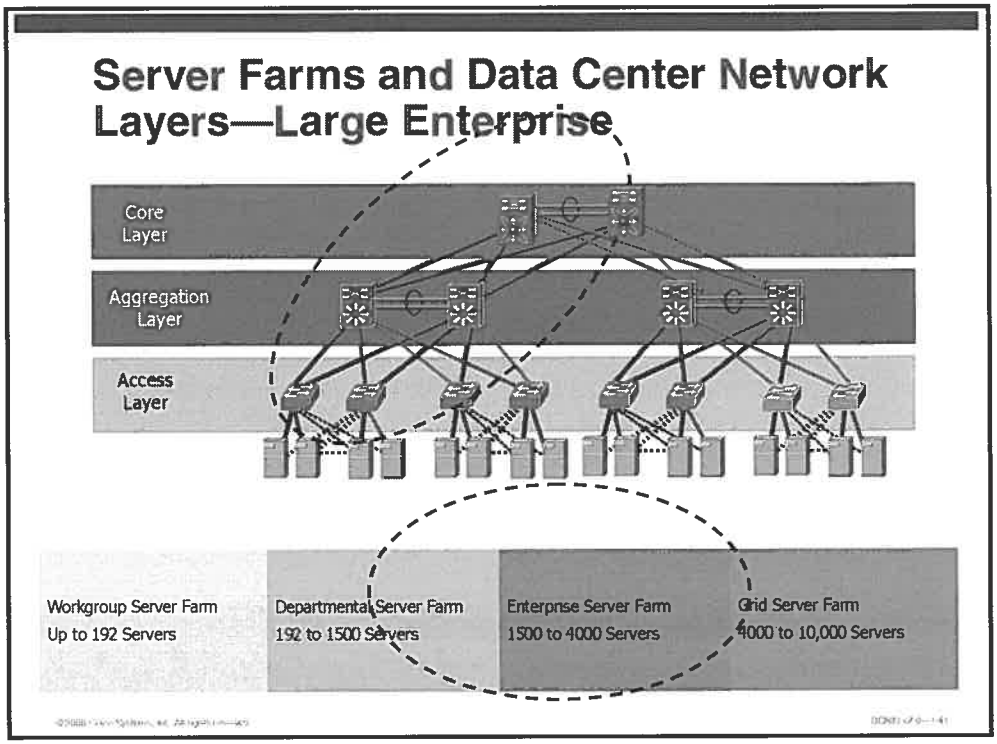


Server farms mapping to data center network layers:

- The small departmental typically maps down to a single aggregation switch pair (no core).

https://t.me/learningnets

Large Enterprise



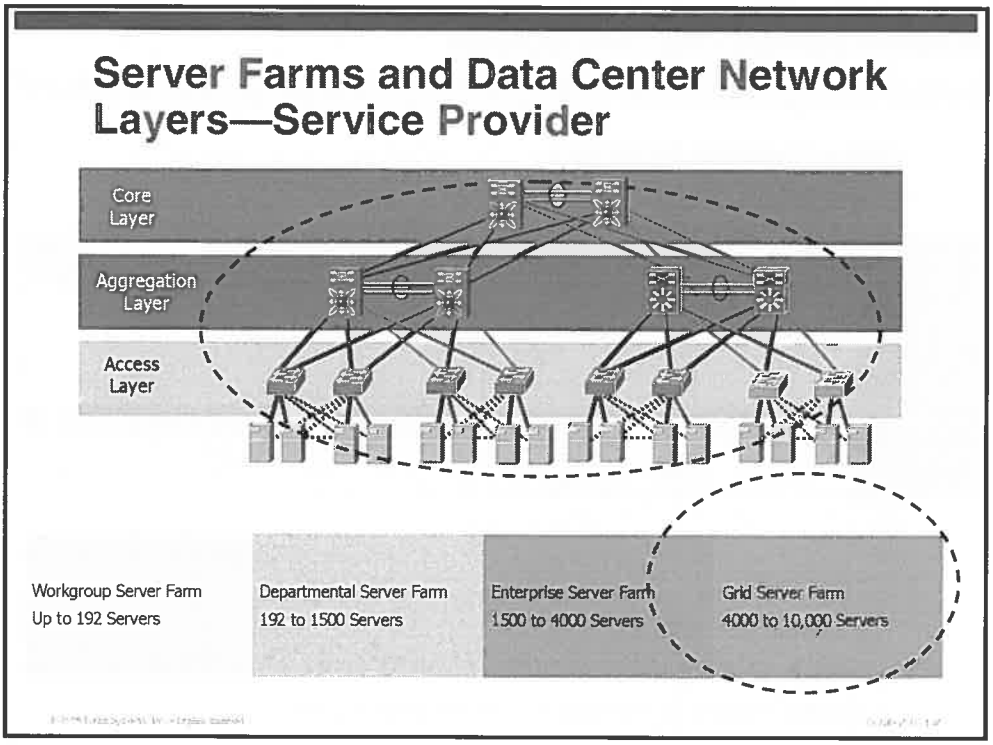
Server farms mapping to data center network layers:

- Large enterprises tend to have the core, aggregation, and the access tier.
- They may have more than one pair of aggregation switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Small Enterprise



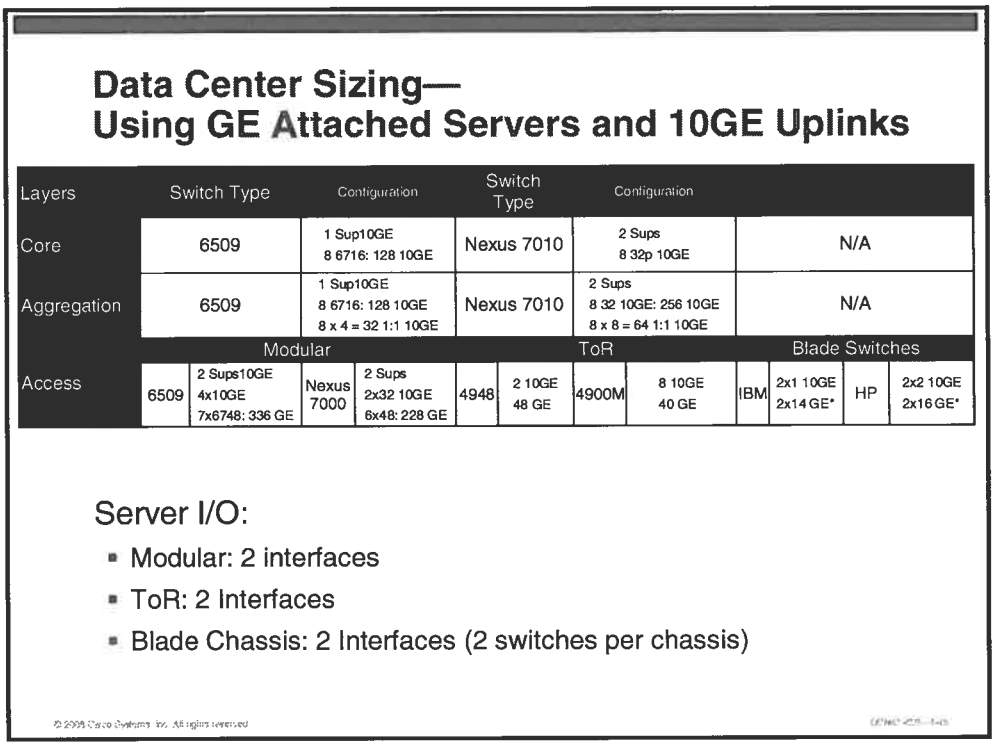
SPs have a core tier and multiple aggregation tiers. Large enterprises and SPs typically have a core tier, at least one aggregation layer, potentially more.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Sizing

Using GE attached Servers and 10GE Uplinks



The modular platforms are Catalyst 6500 Series switches. The aggregation and core layer will use Catalyst 6500 Series switches and Nexus 7000 Series switches.

Assumptions

On the modular access, there are two interfaces per server along with the top-of-rack and blade switch.

- The configuration of the switch is one supervisor module at the core and aggregation layer with eight 16-port line cards and additional four ports on the supervisor module.
- On the Nexus Series switch, similar topology or similar configuration exists. We can use two supervisor modules and then the rest of the cards can be 10 GE. GE line cards can connect appliances directly.

Standardization of server hardware implies use of the latest technology: faster processors, latest bus architecture, and increased and higher server I/O capacity. Considerations such as percentage of the I/O capacity used on average and whether the number changes when the server is multi-homed are often the subject of discussion between server and networking groups. Whatever the answer is, the network must be able to maintain the established target oversubscription while providing for server growth and associated network equipment growth.

Oversubscription for access layer switches ranges from 1:1 to 20:1, depending largely on the application environment. Applications that require high bandwidth benefit from very low oversubscription (1:1 through 4:1). Typical client applications are in the oversubscription range of 10:1 to 20:1 and are generally co-hosted with backup high-bandwidth applications for certain time slices requiring a lower oversubscription range of 4:1 to 12:1. Virtual servers are expected to have higher output than non-virtual servers. Hence, their oversubscription trends lower, in

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

the range of 5:1 to 10:1. Access layer oversubscription in modular access switches is expected to be in the 4:1 to 12:1 range, assuming that each port is capable of wire-rate performance. Once the server actual oversubscription is factored in, the overall oversubscription decreases.

Oversubscription

Oversubscription

Server Side:

- Single homed server:
  - 1GE, 500Mbps or 200 Mbps
- Multi-homed server:
  - 2GE I/O: 2Gbps, 1 Gbps, 500Mbps or 200Mbps

Network:

- In general: must factor I/O module oversubscription
- Access to aggregation: aggregate outbound server capacity to aggregate uplink capacity
- Aggregation to core: aggregate access uplink to aggregate uplink to core capacity
- Core to Network: aggregate aggregation uplink capacity to outbound link to network capacity

Oversubscription Ranges: 1:1 – 20:1

Factor in Server Oversubscription to obtain more accurate values

Modular Oversubscription:
6500: 336 servers to 4x 10GE uplinks ~ 8.4:1
DC3: 336 servers to 8x 10GE uplinks ~ 4.2:1

ToR Oversubscription:
4948: 48 servers to 2x 10GE uplinks ~ 2.4:1
4900M: 40 servers to 8x 10GE uplinks ~ 0.5:1

Blade Oversubscription:
IBM: 14 Servers to 1x 10GE uplinks ~ 1.4:1
HP: 16 Servers to 2 x 10GE uplinks ~ 0.8:1

Workgroup Modular Access: 6509

Data Center Sizing— Workgroup Modular Access: Cisco Catalyst 6509

	Small	
Servers	0 < x < 192	
Ports*	384	
Switch Types	Catalyst 6509	Nexus 7010
Core Switches	0	0
Aggregation Switches	2: 6509s LCs: 4 4x10GE Oversubscription: 8.4:1	2: NX7000s LCs: 2 32x10GE Oversubscription: 8.4:1
Access Switches	2: 6509	2:6509

Maximum Capacity

Access: 336 x 2 = 672 ports ~ 336 Servers

4 10GE uplinks per chassis = 8 10GE

Oversubscription: 336:40 ~ 8.4:1

Core: No switches Required

Aggregation:

6509: 8x10GE ÷ 4 ports per LC = 2 LCs

2 LC for cross connectivity

NX7010: 8x10GE ÷ 8 ports per LC = 1 LC

2 LC for cross connectivity

Access: 384 ÷ 336 = 1.4 ~ 2 chassis

4 10GE uplinks per chassis = 8 10GE

Oversubscription: 336:40 ~ 8.4:1

https://t.me/learningnets

Enterprise Modular Access: 6509

Data Center Sizing— Enterprise Modular Access: Cisco Catalyst 6509

	Enterprise	
Servers	1500 < x < 4000	
Ports*	8000	
Switch Types	Catalyst 6509	Nexus 7010
Core Switches	2: 6509 LCs: 6 4x10GE	0
Aggregation Switches	4: 6509s LCs: 26 4x10GE Outbound Capacity: 80GE Oversubscription: 96:16	2: NX7Ks LCs: 14 32x10GE Outbound Capacity: 40GE Oversubscription: 96:4
Access Switches	24: 6509 Oversubscription: 8.4:1	24: 6509 Oversubscription: 8.4:1

Maximum Capacity

Access: 336 x 24 = 8064 ports ~ 4032 Servers

- 4 10GE uplinks per chassis = 8 10GE
- Oversubscription: 336:40 ~ 8.4:1

Core:

- 6509: 16 10GE ÷ 4 ports per LC = 4 LCs
  - 2 LCs for cross connectivity
  - Uplinks: based on DC outbound requirements
- NX70103: none needed

Aggregation:

- 6509: 96x10GE ÷ 4 ports per LC = 24 LCs
  - 2 LCs for cross connectivity
  - 26 LCs ÷ 8 Slots = 3.25 ~ 4 chassis
  - 4 chassis x 4 uplinks = 16 10GE ports
- NX7010: 96x10GE ÷ 8 ports per LC = 12 LC
  - 2 LCs for cross connectivity
  - 14 LCs ÷ 8 Slots = 1.75 ~ 2 chassis

Access: 8000 ÷ 336 = 23.8 ~ 24 chassis

- 4 10GE uplinks per chassis = 96 10GE
- Oversubscription: 336:40 ~ 8.4:1

© 2008 Cisco Systems, Inc. All rights reserved.

Grid Modular Access: 6509

Data Center Sizing— Grid Modular Access: Cisco Catalyst 6509

	Grid	
Servers	4000 < x < 10000	
Ports*	20000	
Switch Types	Catalyst 6509	Nexus 7010
Core Switches	2: 6509 LCs: 10 4x10GE	0
Aggregation Switches	8: 6509s LCs: 62 4x10GE Outbound Capacity: 80GE Oversubscription: 240:32	2: NX7Ks LCs: 32 32x10GE Outbound Capacity: 40GE Oversubscription: 240:4
Access Switches	60: 6509 Oversubscription: 8.4:1	60: 6509 Oversubscription: 8.4:1

Maximum Capacity

Access: 336 x 60 = 20160 ~ 10080 Servers

- 4 10GE uplinks per chassis = 8 10GE
- Oversubscription: 336:40 ~ 8.4:1

Core:

- 6509: 32 10GE ÷ 4 ports per LC = 8 LCs
  - 2 LCs for cross connectivity
  - Uplinks: based on DC outbound requirements
- NX7010: none needed

Aggregation:

- 6509: 240x10GE ÷ 4 ports per LC = 60 LCs
  - 2 LCs for cross connectivity
  - 62 LCs ÷ 8 Slots = 7.75 ~ 8 chassis
  - 4 chassis x 4 uplinks = 16 10GE ports
- NX7010: 240x10GE ÷ 8 ports per LC = 30 LCs
  - 2 LCs for cross connectivity
  - 32 LCs ÷ 8 Slots = 2 chassis

Access: 20000 ÷ 336 = 59.5 ~ 60 Chassis

- 4 10GE uplinks per chassis = 240 10GE
- Oversubscription: 336:40 ~ 8.4:1

© 2008 Cisco Systems, Inc. All rights reserved.

https://t.me/learningnets

Data Center Sizing— Maximum Server Farm Density

Core: 2 Catalyst 6509
- 8 LCs x 4 10GE ports x 2 chassis: 64 10GE
- 4 10GE per agg switch = 16 chassis
  - 14 aggregation chassis & 80GE outbound

Aggregation: 2 Catalyst 6509
- 8 LCs x 4 10GE ports x 2 chassis: 64 10GE
- 4 10GE per acc switch = 16 chassis
- 2 Service Chassis and 14 Access Chassis
- Oversubscription=
  - 14 access x 4 : 2 core x 4 = 7:1
  - 12 access x 4 : 2 core x 8 = 3:1

Core: 2 Nexus 7010
- 8 LCs x 8 10GE ports x 2 chassis:128 10GE
- 4 10GE per agg switch = 32 chassis
  - 30 aggregation chassis & 80GE outbound

Aggregation: 2 Nexus 7010
- 8 LCs x 8 10GE ports x 2 chassis:128 10GE
- 4 10GE per acc switch = 32 chassis
- 2 Services Chassis and 14 Access Chassis
- Oversubscription=
  - 30 access x 4 : 2 core x 4 = 15:1
  - 28 access x 4 : 2 core x 8 = 7:1
  - 24 access x 4 : 2 core x 16 = 3:1

Switch Types		Catalyst 6509			Nexus 7010		
Access Switches		Modular: 336	ToR: 48	Blades: 28	Modular	ToR	Blades
GE ports per agg pair	7:1	14 X 336 = 4704	14 X 48 = 672	14 X 28 = 392	28 X 336 = 9408	28 X 48 = 1344	28 X 28 = 784
	3:1	12 X 336 = 4032	N/A	N/A	24 X 336 = 8064	N/A	N/A
GE ports per core pair	7:1	14 x 4704	14 X 672	14 x 392	30 X 9408	30 X 1433	30 X 784
	3:1	14 x 4032	N/A	N/A	30 X 8064	N/A	N/A

Data Center Sizing

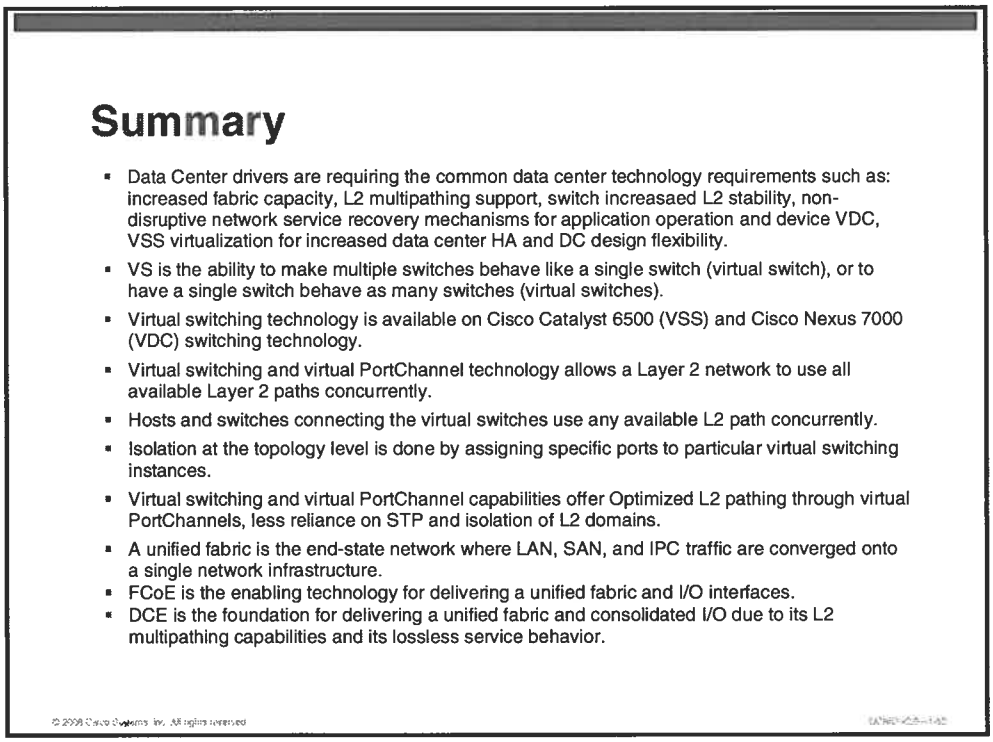
1. Migration to 10 uplinks in access has taken hold.
2. Access Environment:
   - Modular: oversubscription rate determines whether to use oversubscribed agg or core ports
   - ToR and Blade Access Environments: oversubscription at aggregation and core port is recommended
     - Provide higher density L2 domain
     - Oversubscription rates are kept in reasonable ranges
   - Oversubscription
     - Access:
       - 336 GE: 4 10GE = 8.4:1 (Sup Uplinks)
       - 240 GE: 8 10GE = 3:1 (2 4x10GE LCs)
     - Aggregation:
       - 6500: 64 10GE ports per aggregation pair
         - 3:1 = 48 ports to access + 16 ports to core
         - 7:1 = 56 ports to access + 8 ports to core
       - Nexus 7010: 128 10GE ports per aggregation pair
         - 3:1 = 96 port to access + 32 ports to core
         - 7:1 = 112 ports to access + 16 ports to core
   - Port Count per agg Module
     - Modular: ~5000 to ~13000
     - ToR: ~1000 to ~2500
     - Blade: ~600 to ~1500
   - Agg Modules per Core.
     - Range from 4 to 8 based on oversubscription

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

This topic summarizes the key points that were discussed in this lesson.



 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Module 6

Data Center Network Design

Overview

This module introduces different types of data center networking design.

Module Objectives

Upon completing this module, you will be able to create a redundant, scalable, and highly available Access Layer, Aggregation Layer and Core Layer designs. You will also be able to analyze topology alternatives using Cisco Blade switches, ToR switches, Catalyst 6500 switches, Nexus 7010 switches and the Nexus 5000 family of switches. This ability includes being able to meet these objectives:

- Explain generic reference topologies as they apply to multi-tier data center design following the hierarchical network design model: Core, Aggregation, and Access.
- Describe service deployment options
- Explain various DCdata center design architecture options including looped, delta, V, square, U, loop-free, minimal L2, and enhanced L2 topologies and the analysis the failure scenarios.
- Describe virtual switch-based network design deploying virtual switching system (VSS), virtual routing and forwarding (VRF), and VDC virtualization options.
- Explain routing design as applied to data center environment.
- Describe Layer 2 data center interconnect options

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 1

Data Center Network Design

Overview

This lesson discusses different data center design topologies and related data center high availability issues.

Objectives

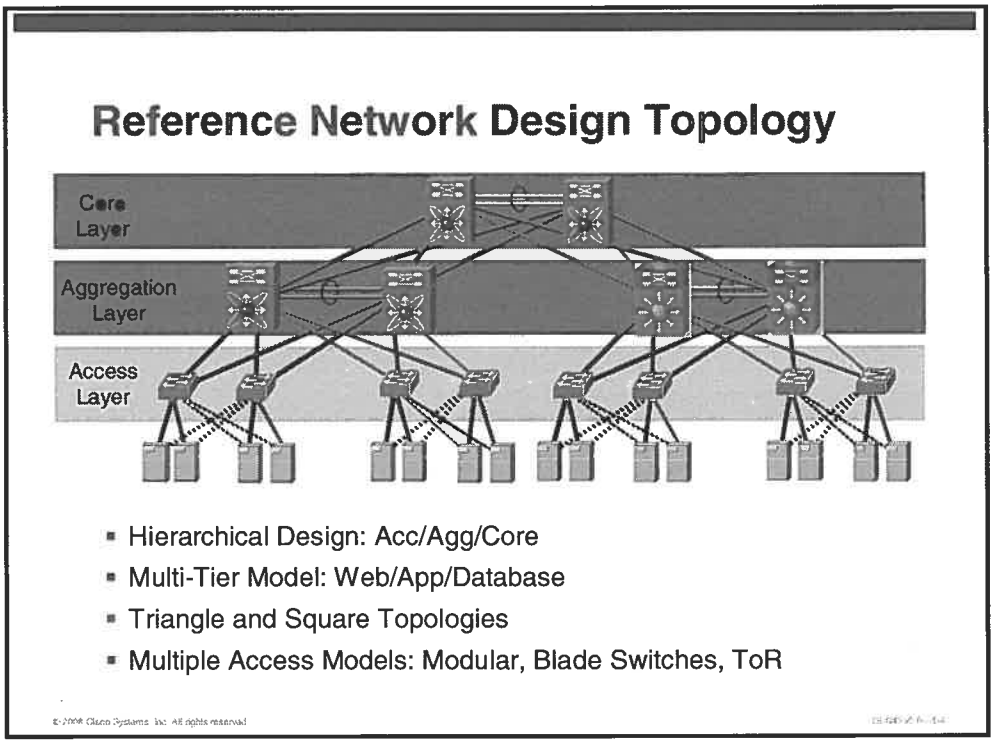
Upon completing this lesson, you will be able to meet these objectives:

- Explain generic reference topologies as they apply to multi-tier data center design following the hierarchical network design model: Core, Aggregation, and Access
- Describe service deployment options
- Explain various data center design architecture options including looped, delta, V, square, U, loop-free, minimal L2, and enhanced L2 topologies and the analysis the failure scenarios
- Describe virtual switch-based network design deploying virtual switching system (VSS), virtual routing and forwarding (VRF), and VDC virtualization options
- Explain routing design as applied to data center environment
- Describe Layer 2 data center interconnect options

https://t.me/learningnets

Data Center Network Design

This topic describes the reference data center design topology.

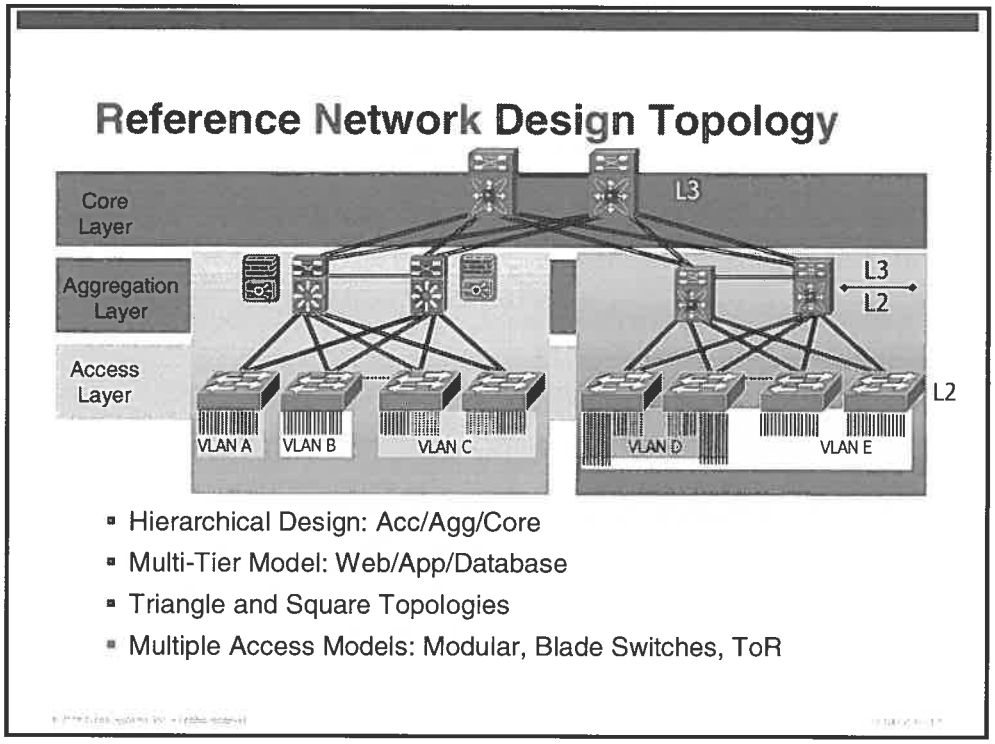


Throughout this module we will be reffering to various data center network design topologies such as:

- Hierarchical Design
- Multi-Tier Data Center Model
- Triangle and Square Topologies
- Multiple Access Models: Modular, Blade Switches, ToR

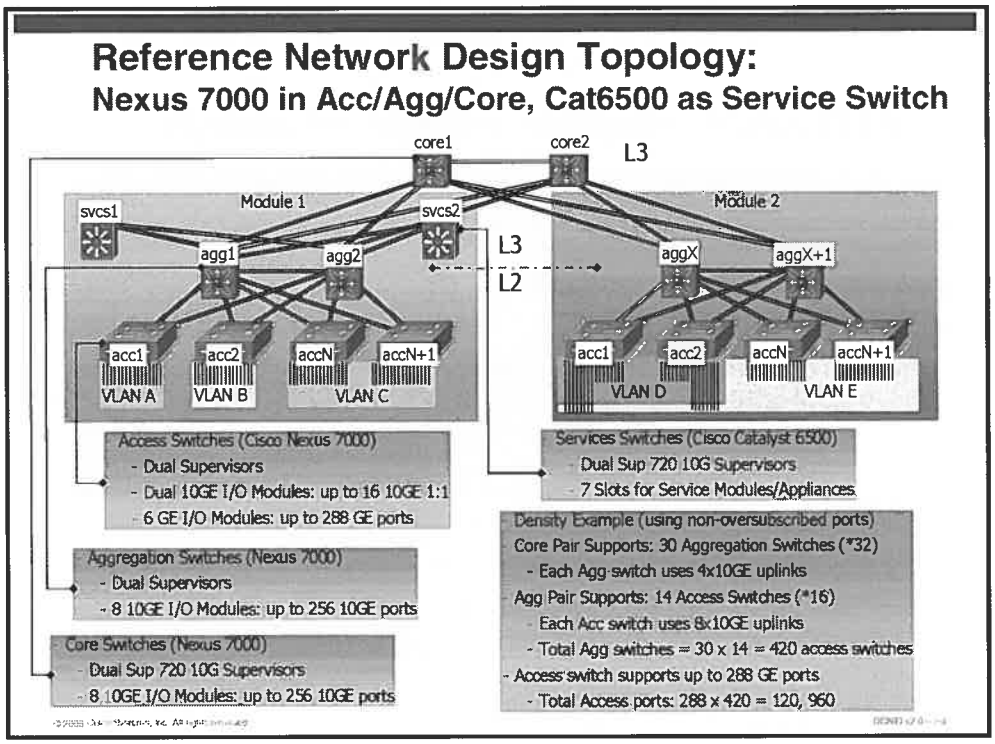
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Layer 2 and Layer 3 domain in the reference topology diagram map to a single or multiple switches and spread across the Access, Aggregation, or Core Layers.

https://t.me/learningnets

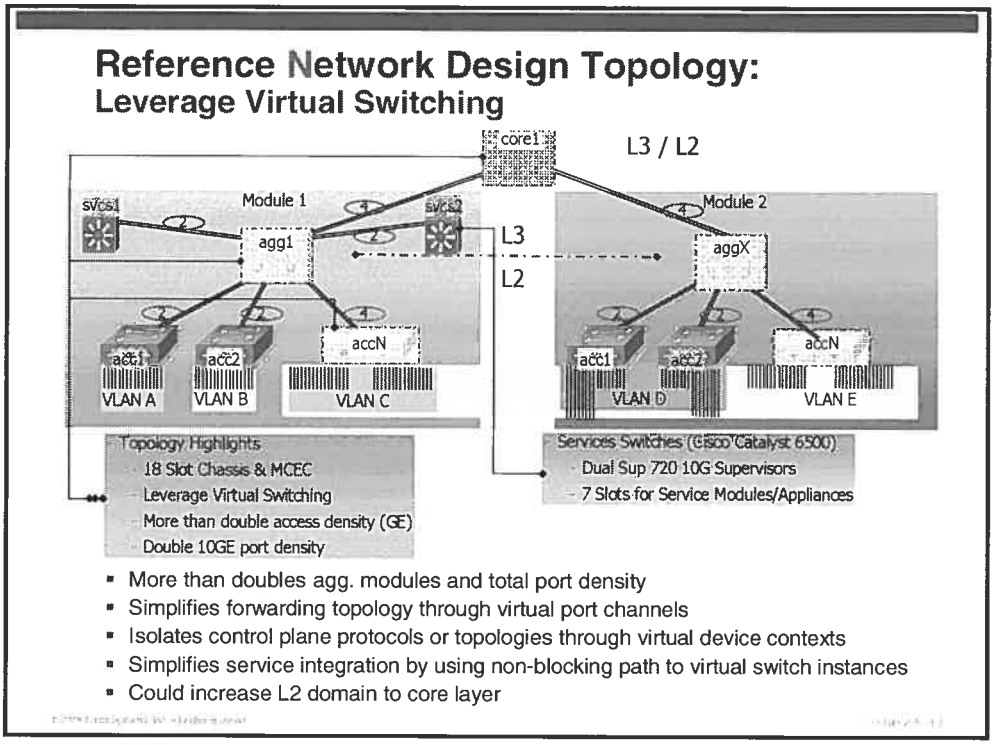


The Reference Data Center topology may use the following configuration:

- Access Switches (Nexus 7000 Series Switches)
  - Dual Supervisors
  - Dual 10GE I/O Modules: up to 16 10GE 1:1
  - 6 GE I/O Modules: up to 288 GE ports
- Aggregation Switches (Nexus 7000 Series Switches)
  - Dual Supervisors
  - 8 10GE I/O Modules: up to 256 10GE ports
- Core Switches (Nexus 7000 Series Switches)
  - Dual Sup 720 10G Supervisors
  - 8 10GE I/O Modules: up to 256 10GE ports
  - Services Switches (Catalyst 6500 Series Switches)
  - Dual Sup 720 10G Supervisors
  - 7 Slots for Service Modules/Appliances
  - Density Example (using non-oversubscribed ports)
- Core Pair Supports: 30 Aggregation Switches (*32)
  - Each Aggregation switch uses 4x10GE uplinks
- Aggregation Pair Supports: 14 Access Switches (*16)
  - Each Acc switch uses 8x10GE uplinks
  - Total Aggregation switches = 30 x 14 = 420 access switches
- Access switch supports up to 288 GE ports
  - Total Access ports: 288 x 420 = 120, 960

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

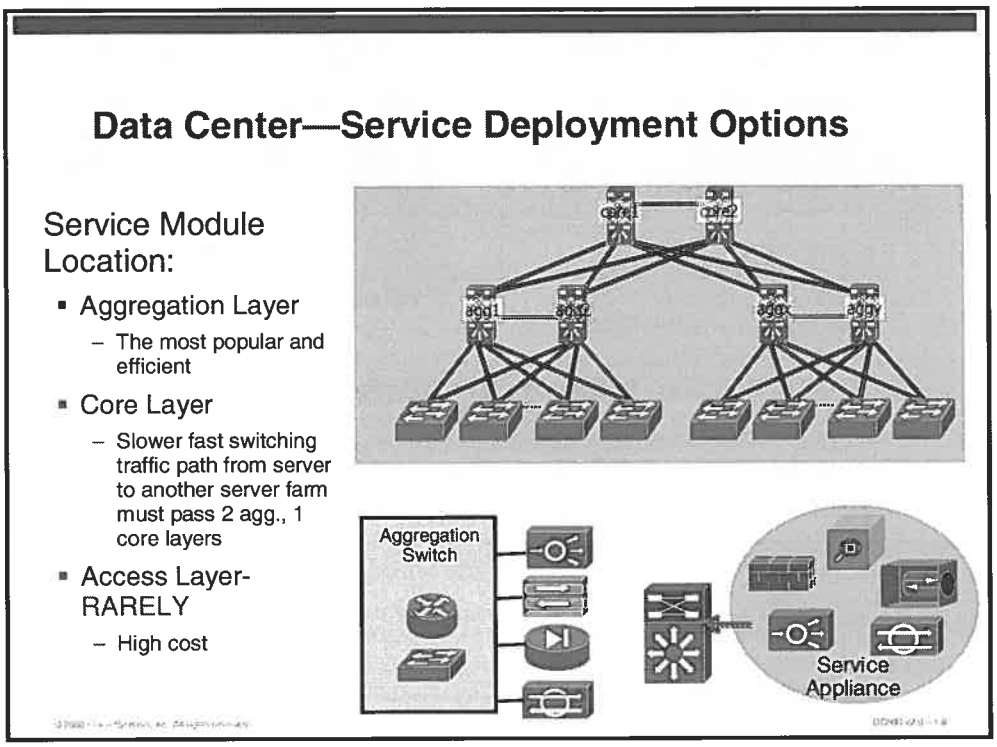


- Topology Highlights:
  - 18 Slot Chassis & MCEC
  - Leverage Virtual Switching
  - More than double access density (GE)
  - Double 10GE port density
- Services Switches (Catalyst 6500 Series Switches)
  - Dual Sup 720 10G Supervisors
  - 7 Slots for Service Modules/Appliances
- More than doubles agg. modules and total port density
- Simplifies forwarding topology through virtual port channels
- Isolates control plane protocols or topologies through virtual device contexts
- Simplifies service integration by using non-blocking path to virtual switch instances
- Could increase L2 domain to core layer

https://t.me/learningnets

Service Design

This topic describes the various options of service integration into the data center network topology.



Data center services can be integrated in two different ways: using service modules integrated into the Aggregation switch or using appliances connected to the Aggregation switch. Data center services are firewalls, Intrusion Detection/Prevention Systems (IDS/IPS), Load Balancers with Caches, and Secure Socket Layer (SSL) offloaders. The firewalls and load balancers support default gateway function on behalf of the server farms. Default Gateway has been provided by the router traditionally. Firewalls and load balancers can provide stateful failover which is supported by specific redundancy protocol. SSL offloader is typically integrated into the load balancer but SSL offloaders do not support default gateway services. IDS/IPS is transparent to the design and integrates well with any existing device. Caches are using Traffic redirection protocols such as PBR and WCCP running on Layer 2 or Layer 3 switches and load balancers to distribute the load among the cache cluster.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Service Module Location

Services such as firewalls can be provided using an appliance or a module located in a chassis such as the Catalyst 6500 Series Switches. The use of modules is common in the data center for reasons such as the reduction in cabling that is required and the high throughput they can provide due to direct connection to switch backplane. Some incorporate high bandwidth connection to the switch fabric.

Although it is possible to place service modules into the Access Layer switches this is uncommon. Placing a module such as the Firewall Service Module (FWSM) in every access switch would result in a high cost. Use of service modules in access switch by traffic destined to another results in unnecessary traffic flows within the Access Layer.

The Core Layer is primarily concerned with fast switching of packets. The placement of service modules in core switches reduces their ability to switch packets at the highest possible rate. Also, to reach service modules located in the Core Layer, traffic from one application tier to another would be required to pass through the Aggregation Layer twice and the Core Layer once. An example of this would be where a firewall is required between application tiers.

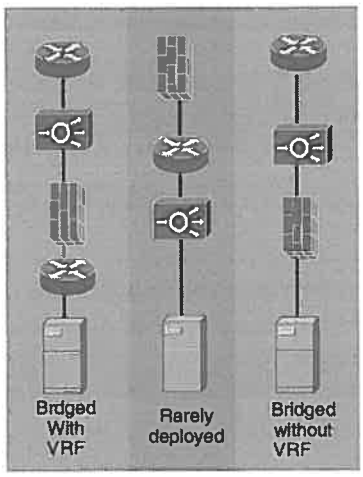
The use of a separate service layer is discussed later.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Service Module Design Choices

- Bridging or routing across access with VRF
- Routed or transparent mode
- Inline or one arm design
- What VLAN to extend between switches
- Which access design to use
- How to accomplish server-to-server LB
- Bridged design w/VRF:
  - Routing by MSFC or with VRF in Acc Layer
- VRF needed for routing between servers
  - Added firewall reduces throughput
- The FWSM context above MSFC:
  - But it is rarely used:
    STP into the core, link failure invisible to MSFC
  - Alternative: Add VRF



© 2008 Cisco Systems, Inc. All rights reserved.

ACE and Firewall Service Modules Design Choices in Catalyst 6500 Series Switches

When deploying these services, there are a number of choices to make, including:

- Bridging versus routing toward the access with VRFs
- Routed versus transparent mode
- Inline versus one-arm design
- Which VLANs to extend between Aggregation switches
- How traffic will flow
- Which access design to use
- How to accomplish server-to-server load balancing

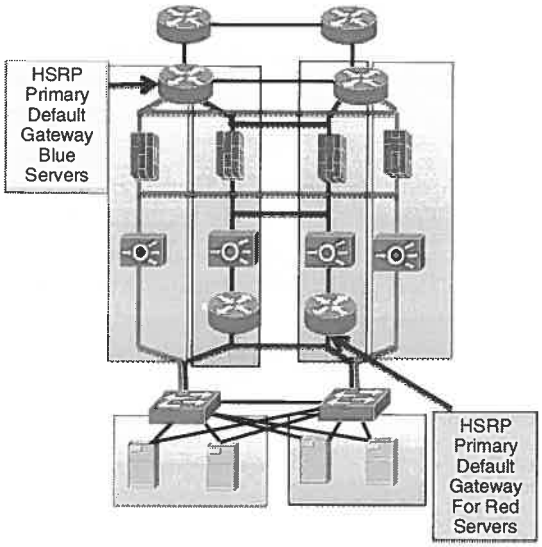
The bridged design leaves all of the routing to the global Multilayer Switch Feature Card (MSFC). The VRF routed design adds a VRF south of the service modules to do the access layer routing. This is useful when there is a requirement to route between subnets behind the firewall. For example, NAS storage resources might be located on a different subnet than the web server, but within the same security domain. Forcing these flows through the firewall reduces the overall capacity, without providing any additional security.

Another option is to place the firewall context above the global MSFC, between the Aggregation and Core tiers. This approach, however, is undesirable for a number of reasons. Spanning Tree Protocol (STP) processes are introduced into the core of the network, the MSFC loses direct visibility to link failures, and the regular changes to FWSM contexts are potentially disruptive to the entire network. Alternatively, when dedicated VRFs are used to provide routing functionality, the integrity of the core is maintained, while maximum flexibility is provided for access connections. VRFs can also provide a way to manage overlapping address space without the need for Network Address Translation (NAT).

https://t.me/learningnets

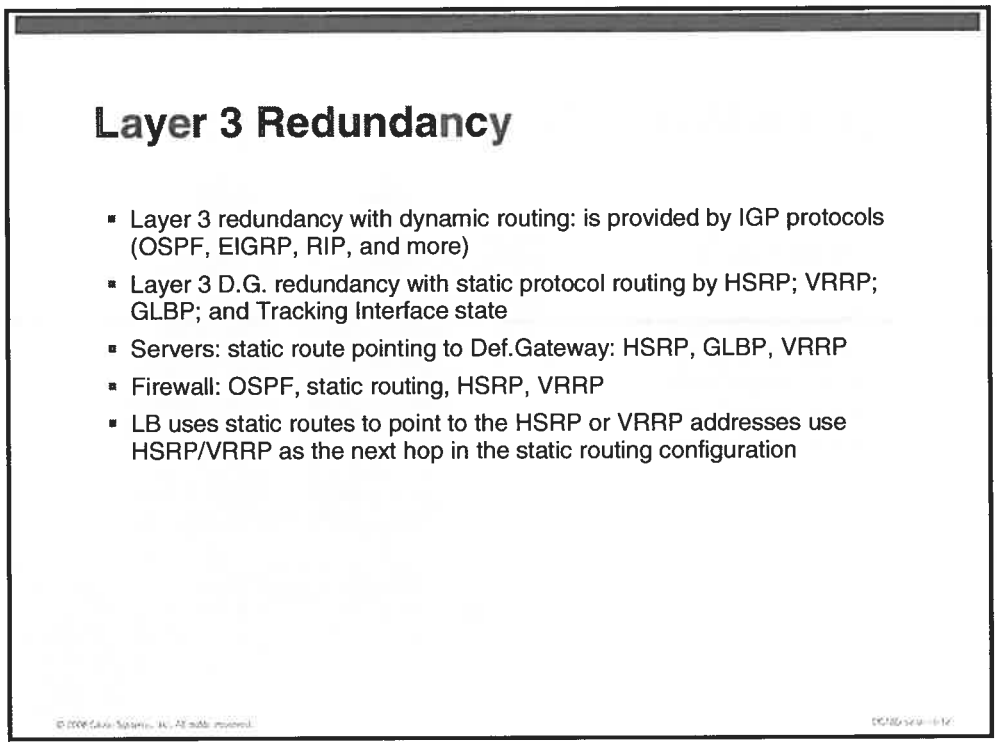
Service Module Design Choices (Cont.)

- The advantage of the red topology is simplicity.
- The blue topology is complex but used for environment with many subnets and heavy traffic flows.



The advantage of the blue service chain is simplicity. The red service chain is more complex, but is better suited to environments with multiple subnets and heavy traffic flows within the same security domain.

https://t.me/learningnets



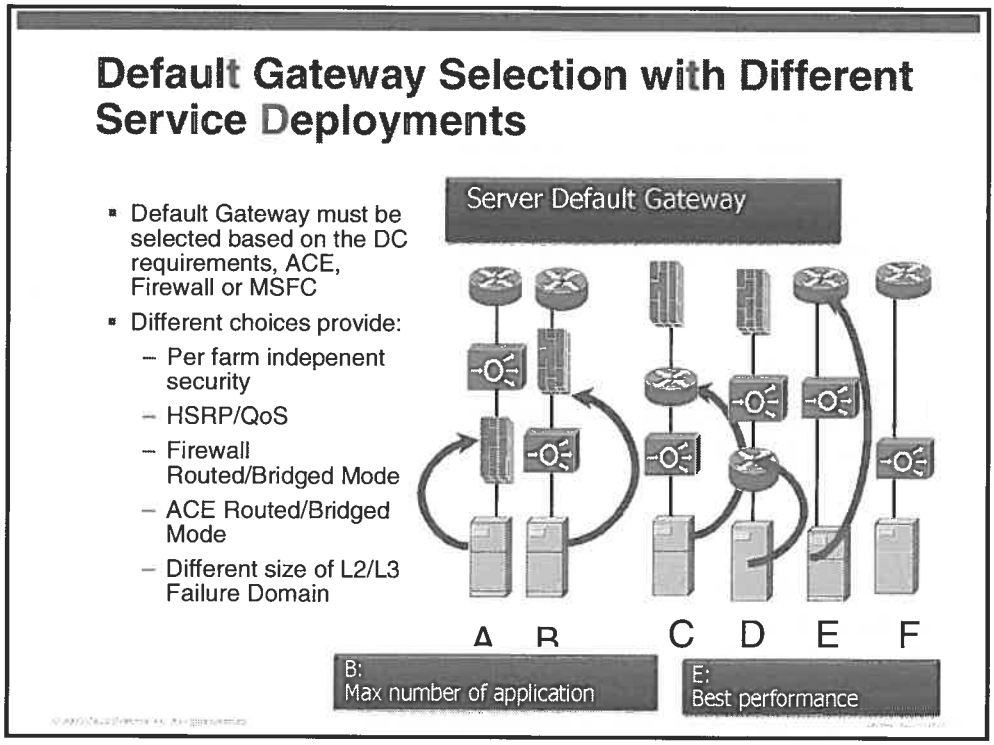
The convergence time-in a data center depends on the convergence time of all of the affected protocols running on all the hardware impacted by the particular failure. Most protocols can be configured to converge in a few seconds or even to have convergence times less than a second. A spanning tree with the IEEE 802.1w optimization converges in a less than 1 second. IGPs can be configured for similar convergence times. EIGRP is the fastest of the available IGPs, where as OSPF needs approximately 600 msec to reconverge after a topology change. If the default Hot Standby Router Protocol (HSRP) settings are maintained, recovery from a failure cannot be done in less than 10 seconds. If HSRP timers are configured to converge in less than 1 second, then HSRP flapping might be experienced upon Llayer 2 topology changes. The recommended HSRP timers are 1-3 seconds in most cases, and do not typically cause flapping at that speed (but this must be tested in every data center environment). When HSRP failure occurs, the newly elected peer in the HSRP group starts answering to ARP requests for the virtual IP address on the interface connected to the VLANs where the peers are redundant. The MAC address returned is the vMAC address.

When there is a change in the active-standby roles the newly active router interface sends out the gracious ARP that contains the VIP and the virtual MAC (vMAC) address and the source MAC of the packet is also the vMAC address. The gracious ARP will thus make sure that Layer 2 forwarding tables are refreshed.

Tracking enables an active HSRP interface to become a standby when the router loses an interface that is responsible to transport the traffic further (for example, from the HSRP redundant Aggregation Layer switches to the redundant core switches) used for the upstream traffic. When the tracked interface is down, the router will reduce the HSRP priority for the group 1 (10 by default) and will enable the standby HSRP interface to take over the active role.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Service Module Default Gateway Redundancy

Server Default gateway for a server in a data center can be a router, a firewall, or a load balancer. In the same data center, some sever farms use only Layer 2 and Layer 3 functions, some use only load balancing or firewalling, and some use both firewall and load balancing services. Firewalling and load balancing can be provided by placing these devices in front of the server farm either in transparent mode or bridge mode, or by making the firewall or the load balancer the default gateway for the server farm. Default Gateway must be selected based on the data center requirements. Different choices offer different disadvantages and advantages.

Example A

When load balancer is operating in router mode or bridge mode, the firewall configuration must enable server health-management traffic from the content switching module to the server farm. This adds management and configuration tasks to the design. In this design the firewall will provide the default gateway to the server farm.

Example B

Example B shows a server with the load balancer default gateway if it is deployed in router mode. If a load balancer were deployed in bridged mode, then the default gateway would be the firewall. This configuration facilitates the creation of multiple instances of the firewall and ACE combination for the segregation and for the load-balancing of each of the server farms independently. Placing the load balancer in bridge mode between the server farm and the firewall and configuring the firewall as the default gateway provides the maximum number of application and security services.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Example C

Firewall facing the Core IP network in this example must have routing capabilities to easy integration with the routed network. This makes the capability to secure each server farm independently more complicated since the packets can be routed from the server farm and back to the server farm by ACE and Firewall without passing the firewall. If the ACE is deployed in routed mode, then ACE or the router can be the Default gateway; when ACE is deployed in bridged mode, the default gateway will be the router.

Example D

The advantage of this design is that the router is the default gateway for the server farm; therefore, the servers can take advantage of HSRP tracking, QoS or DHCP relay functions, available only on routers.

Example E

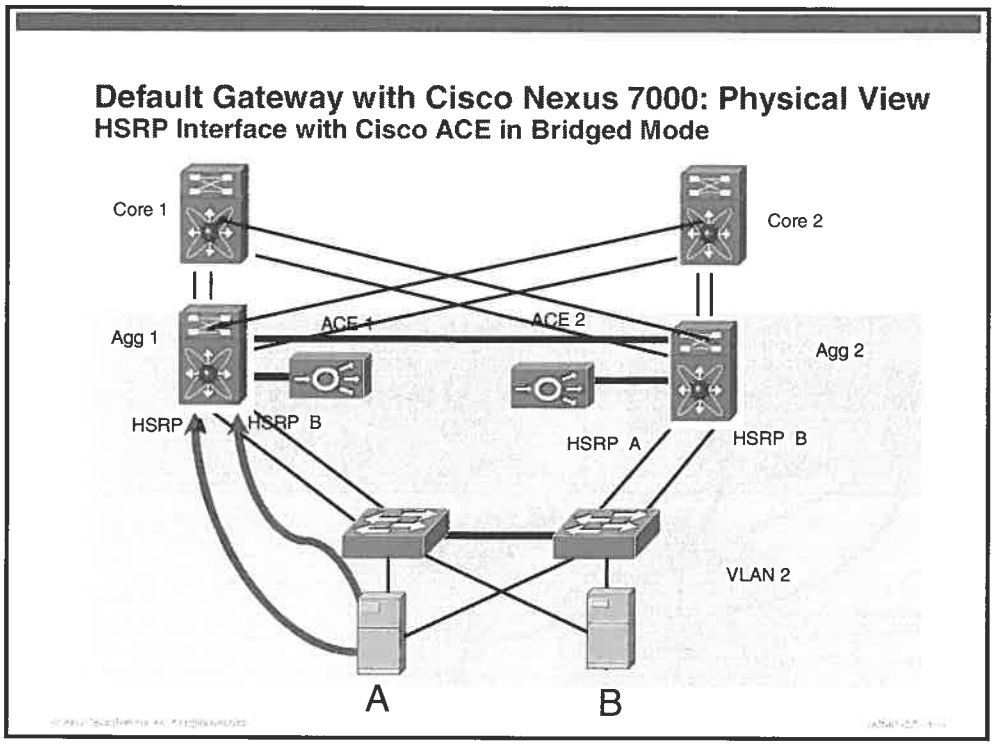
Servers send traffic directly to the router. The default gateway configured on the servers is the IP address of the router.

Note Configuring the router as the default gateway (A) provides the best performance.

Placing the FWSM in transparent mode, the traffic flowing from a server farm on one VLAN to a server on a different VLAN traverses the device twice because routing occurs on the MSFC.

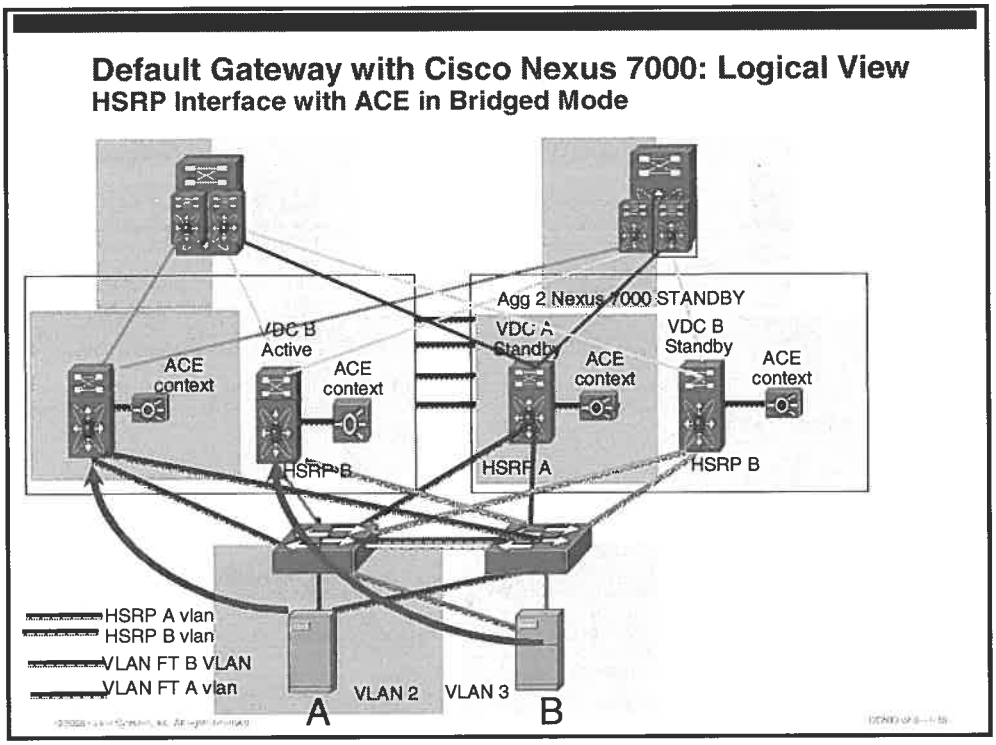
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



In this topology each of two Nexus 7000 Series switches are connected to a ACE appliance or ACE in the one arm mode-bridged mode. The Layer 2 server farms A and B are assigned to VLAN 2. Both Aggregation switches 1 and 2 deploy VDC A and VDC B for server farm A and server farm B on the same VLAN 2. Every server farm has a separate link to the core layer. Traffic from a different server farm uses a separate virtual switch (VDC); therefore, a separate physical port is needed. An external ACE appliance can be deployed in one arm mode or in transparent mode. An external firewall can be deployed in multi-context mode using context A and context B for each server farm. The Default gateway for server farms A and B will be an HSRP group A interface on the Aggregation switch 1 and an HSRP group B interface on the Aggregation switch 1, respectively.

https://t.me/learningnets



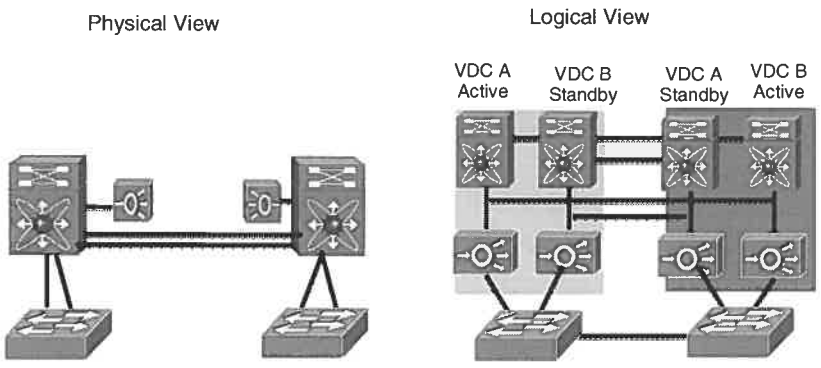
In this topology, two default HSRP gateway groups exist—HSRP A and HSRP B—and are assigned to server farm A and to server farm B, respectively. VDC A and VDC B are virtual switches that belong to server farm A and server farm B respectively. Server farm A and server farm B send traffic across VLAN 2 and VLAN 3, respectively. When HSRP group A fails over to the standby interface on Aggregation switch 2, the VDC A using HSRP tracking will failover to the standby VDC A on the Aggregation switch 2.

Note This solution is not scalable in the sense that every server farm consumes one dedicated VDC port on Nexus 7000 switches and another VDC port for ACE context connection. This topology represents a dedicated server farm network across the data center and server farm Layer 2 extended segmentation up to the Core layer, and it enables even further extension of Layer 2 domain across the WAN to the remote data center site server farm.

© 2008 Cisco Systems, Inc.

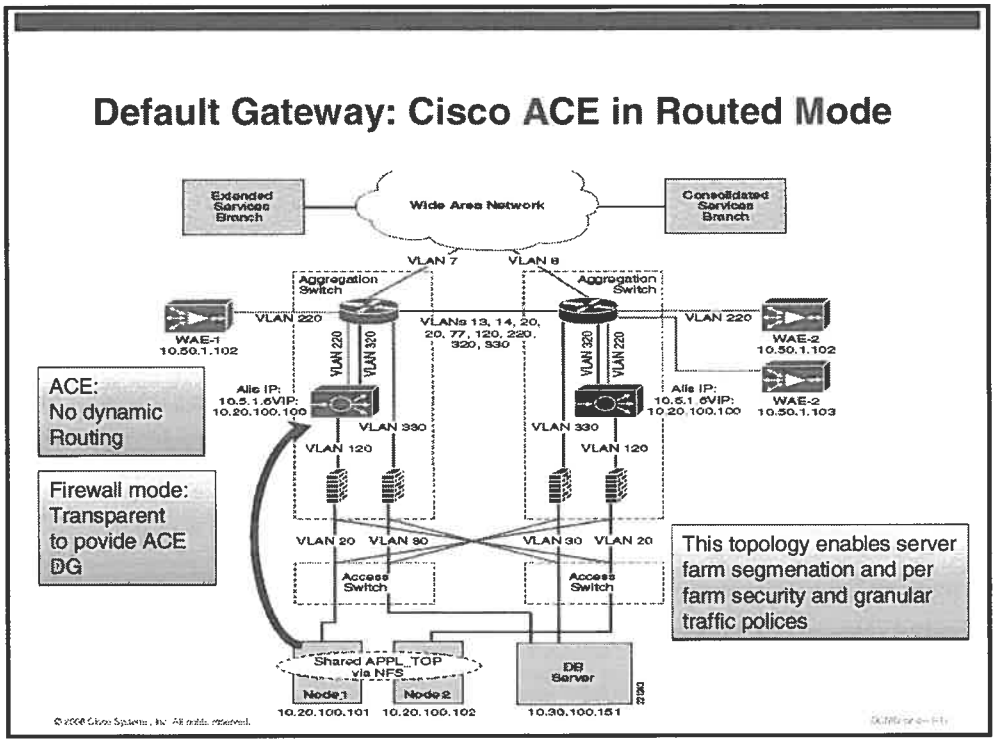
https://t.me/learningnets

Cisco ACE or Firewall in Bridged/Routed Mode Physical/Logical View



Two physical Fault tolerant links are provided between the two Nexus switches, one for each external ACE appliance in the respective Nexus VDC (one for each ACE context). Additional separate Layer 2 links can be deployed between the VDC A active and VDC A standby used to carry Layer 3 HSRP failover information (one for VDC A and the other for VDC B HSRP keepalive). Routing protocol might require another two Layer 3 links to be deployed between the physical Nexus switches, which are not shown on this figure. Similar figures apply to the external firewall attached to the Nexus 7000 Series switch in routed or bridged mode.

https://t.me/learningnets



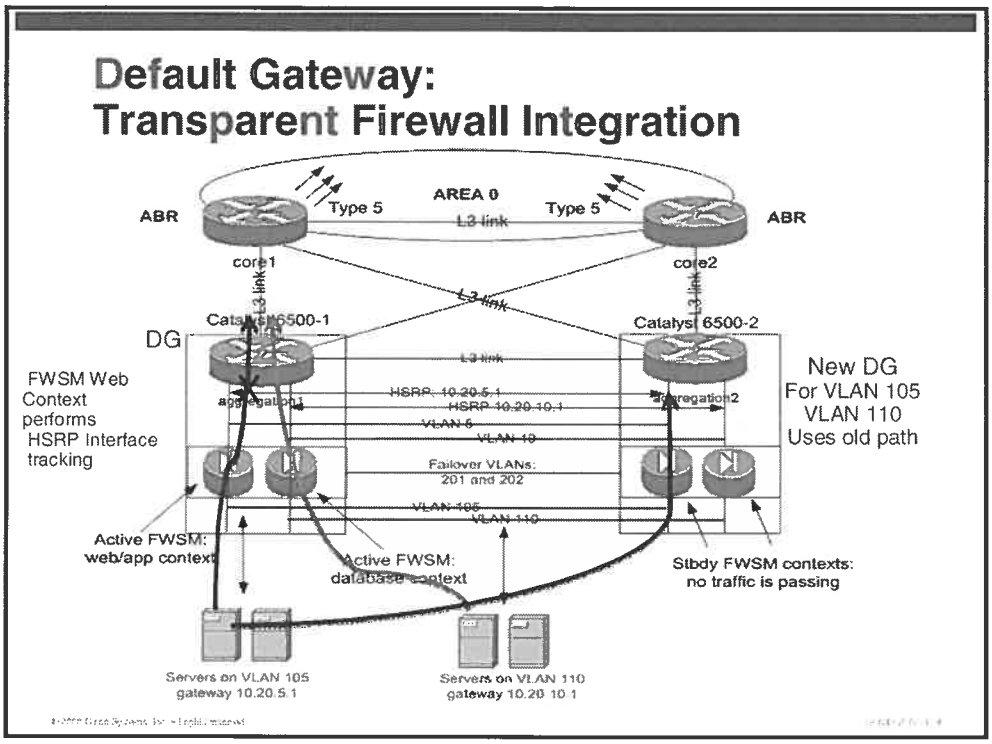
When the data center requires the ACE module or ACE appliance to be deployed in routed mode it implies that the ACE is the default gateway for the connected Server farms. The gateway is accessible via the services of a transparent firewall context. The ACE and firewall context are deployed in an active-standby scenario with stateful failover. An independent pair of transparent virtual firewalls secures the database environment as well.

Note The ability to segment and enforce granular traffic policies at each tier of the application topology is a powerful characteristic of this design.

Note Because ACE does not support dynamic routing but only static routes, this would break the possible routing neighborship between the VRF deployed below and the MSFC above the ACE in the SM chain.

© 2008 Cisco Systems, Inc.

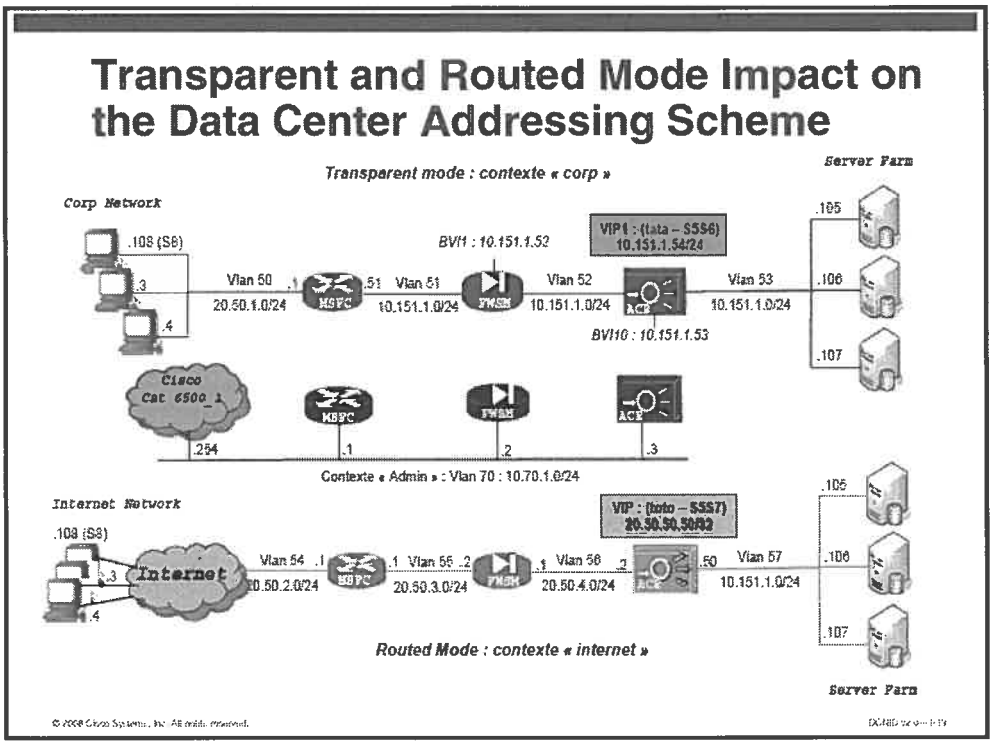
https://t.me/learningnets



There are two transparent firewall contexts deployed in redundant mode, each assigned a respective VLAN server farm traffic with a respective HSRP group. Each server group has a respective VLAN default gateway associated with its HSRP group interface.

If HSRP group 1 interface fails over to the standby interface on the Aggregation switch 2, the server's default gateway will be HSRP interface on the Catalyst 6500 Series Aggregation switch 2, and when both firewall contexts use HSRP interface tracking, the firewall context tracking the failed HSRP group will redirect VLAN 105 traffic to the standby firewall context in the Aggregation switch 2, while the VLAN 110 servers will keep the original path through the Aggregation switch 1.

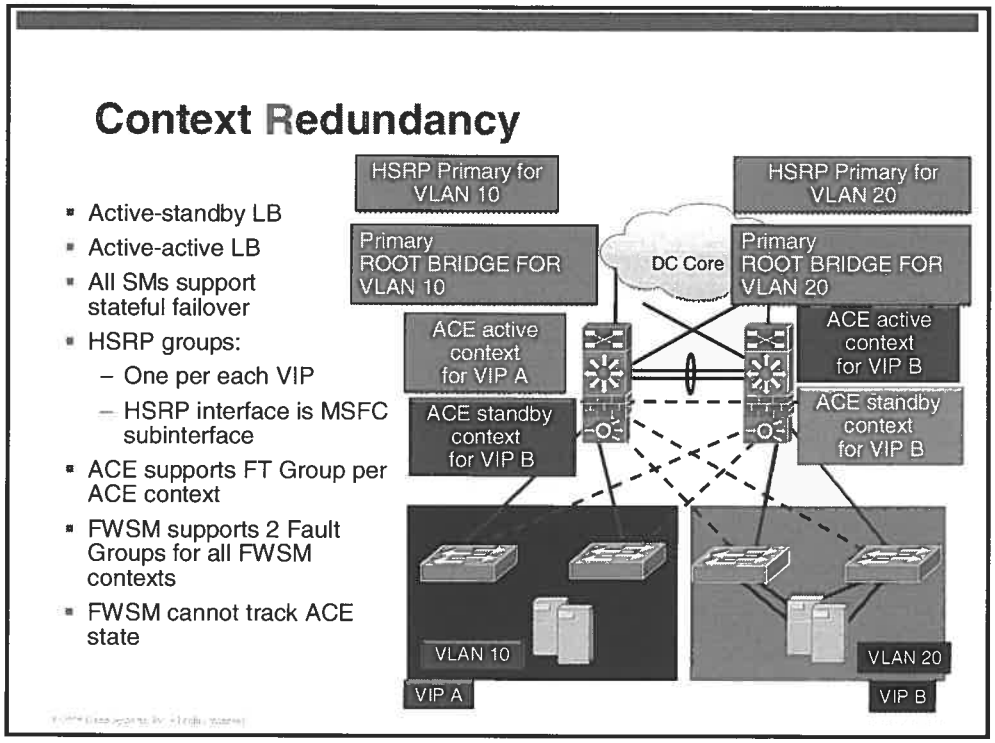
https://t.me/learningnets



This picture illustrates FWSM and ACE deployed in transparent and routing mode and the respective influence on the data center addressing scheme.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



A Layer 2 access layer permits services that are provided by service modules or appliances to be shared across the entire access layer. Examples of this are when using the FWSM, CSM. The active-standby modes of operation used by service modules require Layer 2 adjacency with the servers that use them.

Active-Standby Load-balancing

In active-standby environments, we have a pair of load balancers where only one of them is active. All the traffic from the core sent to the VIP is received by the primary load balancer and the connection or session state since there is a single forwarding path to the access layer switches. Management and troubleshooting is simpler. ACE load-balancer can be a server default-gateway.

Active-Active Load-balancing

In active-active environments, we have a pair of load balancers that are both active at the same time. Both load balancers are active for different VIPs. In active-active mode, each SLB is active for specific VIP and the other is in standby mode for that specific VIP.

Each physical SLB or firewall has context that is active for VIP A and context that is standby SLB context for the VIP B. In this case, it is recommended that the server default gateway points to the SLB that is primary for the server traffic load balancing. HSRP groups can be deployed on the Aggregation MSFC 1 and Aggregation MSFC 2 interface for different VIP default gateways redundancy. Both load balancers are active for the same VIP. Both SLB contexts on both Aggregation switches 1 and 2 can be active the same VIPs. The main challenge here is having the same MAC and IP address active in two different places. Traffic from the core into the access layer can be switched by the Aggregation MSFC on per packet or on a per-IP address basis, and so on.

https://t.me/learningnets

To make this work, the upstream device must ensure that the same load balancer sees all the packets for a single connection and all connections for the same session. This requires PBR on the router to direct certain traffic that matches certain parameters across certain paths. If one load balancer fails the other must be able to handle the combined load as a result of failure.

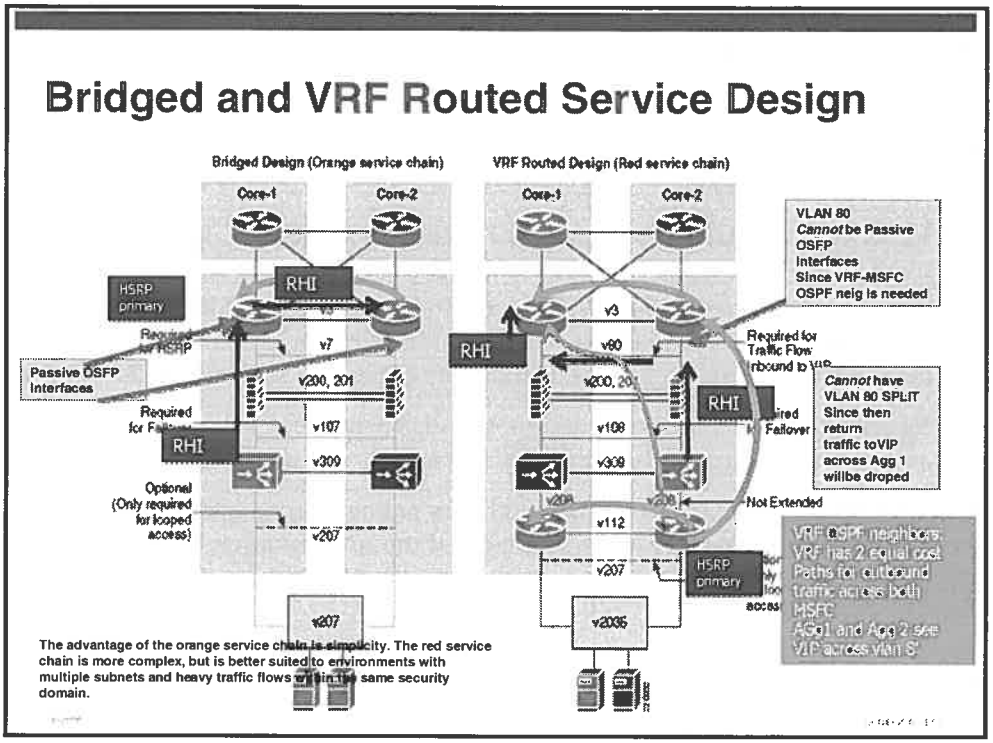
In the Layer 2 domain there might be several VLANs deployed that have been included to provide some scaling for Spanning-Tree and HSRP. Each of the seven devices in the Layer 2 domain participates in Spanning-Tree. The Aggregation layer switch 1 is configured as the primary STP root device for all VLANs in the Layer 2 domain, and the Aggregation switch 2 is configured as the secondary STP root. The STP that is used is PVST+ plus the rapid convergence of IEEE 802.1w (collectively referred to as Rapid PVST+ or rPVST+).

ACE and FWSM Context Redundancy

ACE and FWSM contexts are populated per each VLAN in each Aggregation switch. ACE can have an FT group defined per context, whereas FWSM can have only 2 FT groups for all FWSM contexts. Usually ACE failover will not trigger the FWSM context failover in the same design. The most popular design is deployment with ACE in inline bridged mode and FWSM in bridged mode, with an HSRP default gateway per each server farm.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Bridged Design (Orange service chain)

VLAN Extension for Bridged Design:

- **VLAN7:** is required for HSRP Failover because in this configuration the global MSFC is the default gateway for the servers. Each MSFC must see the HSRP keepalives across VLAN7 to know which should be active. v207 is optional (required only for looped access).
- **VLAN200, VLAN 201, VLAN 309:** are for fault tolerance.
- **VLAN 107:** is required for failover so that in the case of an FWSM or ACE context failure, where FWSM is active on one side and ACE is active on the other, the traffic is not blocked.
- **VLAN 3:** is for routing protocol exchange.
- **VLAN207:** is optional. For a looped access design, VLAN207 is extended and STP is relied on to prevent loops in the access. If loop-free access is required, VLAN207 is not extended and ACE VLAN tracking is used to recover from link failures.

https://t.me/learningnets

VRF Routed Design (Red service chain)

VLAN Extension for the VRF Routed Design:

- **In the VRF routed design:** VLAN80 and VLAN108 are extended, VLAN208 is split, and VLAN203 is optional.
- **VLAN80:** It is extended, but not for HSRP such as VLAN7. It is not the default gateway for the servers (although it would be useful as a default gateway for the management interfaces on the FWSM and ACE contexts). It would be advantageous not to extend VLAN80 to reduce load on the ISL, but in this design, this is not possible because of how the route health injection (RHI) static route to the VIP is propagated by OSPF.
- **VLAN108:** As in the orange design, the VLAN between FWSM and ACE must be extended to prevent traffic being blocked in failover situations.
- **VLAN208:** Not extended. VLAN208 does not need to be split, but splitting it provides an extra element of protection against loops if the ACE contexts become active-active.
- **VLAN203:** Extension is optional. The access VLAN must be extended only for looped access.
- **VLAN 3:** VLAN 112 for routing protocol exchange.
- **V207:** Optional (required only for looped access).

Note With both of these VLANs extended, there is the possibility of a loop if the FWSM or ACE goes into an active-active failover state. The most common scenario for active-active failover is when both modules are up but the failover VLAN is down. BPDU forwarding will prevent such loops and should always be enabled on ACE and FWSM, as shown in the following examples. Also note that whenever BPDUs are forwarded, two global commands must be disabled on the Catalyst IOS: spanning-tree portfast bpduguard default and spanning-tree loopguard default.

ACE BPDU forwarding configuration details:

```
access-list BPDU ethertype permit bpdu
interface VLAN1079 Service Module Design with ACE and FWSM EDCS-559859 VLAN Extension
bridge-group 1
access-group input BPDU
interface VLAN207
bridge-group 1
access-group input BPDU
FWSM BPDU forwarding:
access-list BPDU ethertype permit bpdu
access-group BPDU in interface inside
access-group BPDU in interface outside
```

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Access with FWSM

Typically, FWSM designs in the Aggregation tier have mandated a looped topology in the access layer, where an ISL link between the two Aggregation switches creates a loop, as shown in the preceding figure.

If there is a link failure on the primary link to the access layer, the traffic has a secondary path via AGG-2, but without the ISL link, it runs into a dormant firewall context that is unable to pass the traffic. An alternative to creating this loop in the access is for the FWSM to track the failure in the access and move the active context over to AGG-2 where the traffic has been redirected. This does not work very well with the FWSM because the FWSM is unable to track HSRP and VLAN autostate events such as the ACE module. It is able to detect when an FWSM interface becomes unreachable, but this process takes 50 seconds, which is much longer than the typical client timeout. In addition, each FWSM context is not in its own failover group as is the case with ACE. Thus, failover of one context moves over other contexts that might not need to be moved. The best way to get fast convergence for access link failures with an FWSM is looped access. With looped access and rapid PVST+, convergence is extremely fast—in the range of 0–1 seconds.

Note The failover times listed here and in the following section are measured by the number of millisecond pings lost during failover. These numbers tend to track pretty closely with Chariot (a popular testing tool).

Loop-Free Access with ACE

ACE, with its auto-state awareness and tracking mechanisms, is able to rapidly initiate context failovers on events other than a module failure. ACE can track VLANs, HSRP groups, and designated IP addresses. By associating the health of an access-link to the failover of the context, ACE provides loop-free access in the presence of service modules. The previous figure shows a "bridged design" service deployment with ACE deployed south of the FWSM. This figure also shows that the server VLAN of the ACE context has a single port. When the link fails, auto-state notifies ACE that the VLAN is down because there are no longer any active ports besides ACE itself. ACE can track this event to initiate a failover. As a result, when the primary link fails, ACE opens up the rest of the path so that traffic can pass back up to the primary HSRP router. This requires an ISL hop when there is an FWSM in the path.

The ACE tracking mechanism for VLANs is extremely fast. When the link fails, the ACE context fails over instantly and the outage is less than a second in duration. Currently, however, there is an issue (documented in CSCsf96889) associated with the preempt mechanism that makes the primary context active again when the link is reestablished, resulting in outages ranging from 5-10 seconds. This was tested with the U design with two access switches interconnected. The access switch with the primary uplink saw outages up to five seconds with preempt where the other access switch was consistently in the 8-10 second range.

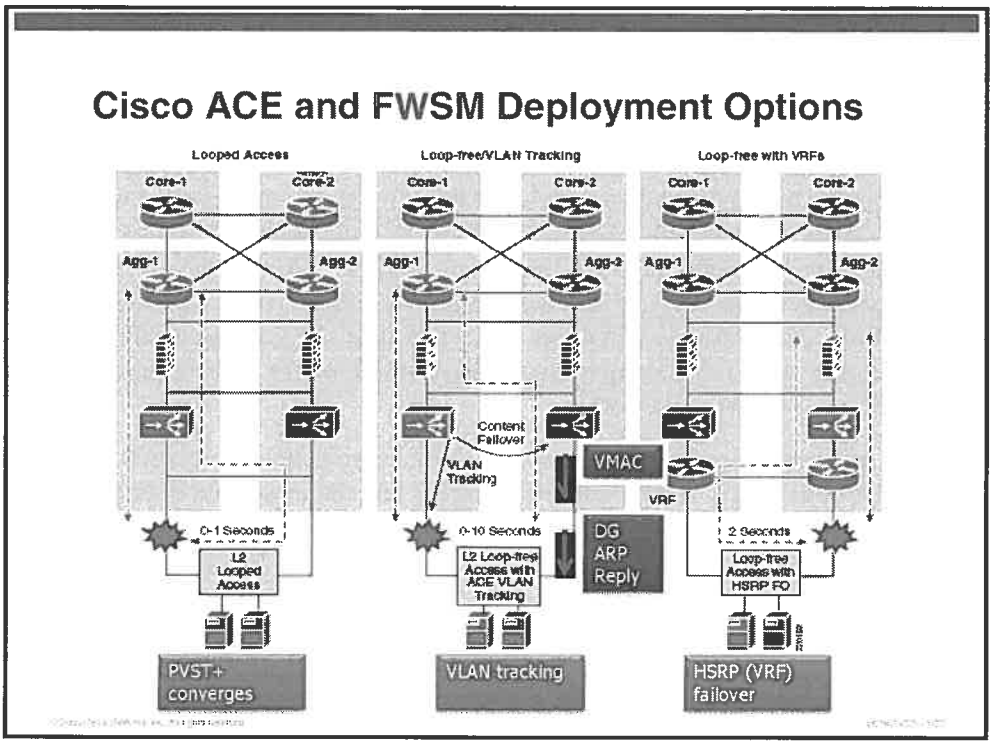
https://t.me/learningnets

VRF Routed Design and Loop-Free Access

Unlike service modules that use an active-standby failover mode, VRFs are always active and ready to forward traffic. As shown in the third scenario in the previous figure, when a Layer 3 path is available between the VRFs, a Layer 2 loop in the access can be avoided when HSRP converges. When loop-free access is configured this way, the gating factor is the time it takes HSRP to converge to the new active VRF serving as the default gateway. Using the default timers of HSRP, convergence takes place in 7–9 seconds. With the hello and hold timers set to 1/3, there is convergence within two seconds, which is not quite as fast as Rapid PVST+ and looped access, but close.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Looped Access with FWSM

- Typically, FWSM designs in the Aggregation tier have mandated a looped topology in the access layer, where an ISL link between the two Aggregation switches creates a loop, as shown in the preceding figure.
- If there is a link failure on the primary link to the access layer, the traffic has a secondary path via AGG-2, but without the ISL link, it runs into a dormant firewall context that is unable to pass the traffic. An alternative to creating this loop in the access is for the FWSM to track the failure in the access and move the active context over to AGG-2 where the traffic has been redirected. This does not work very well with the FWSM because the FWSM is unable to track HSRP and VLAN autostate events such as the ACE module. It is able to detect when an FWSM interface becomes unreachable, but this process takes 50 seconds, which is much longer than the typical client timeout. In addition, each FWSM context is not in its own failover group as is the case with ACE. Thus, failover of one context moves over other contexts that might not need to be moved. The best way to get fast convergence for access link failures with an FWSM is looped access. With looped access and rapid PVST+, convergence is extremely fast, in the range of 0–1 seconds.

Note The failover times listed here and in the following section are measured by the number of millisecond pings lost during failover.

https://t.me/learningnets

Loop-Free Access with ACE

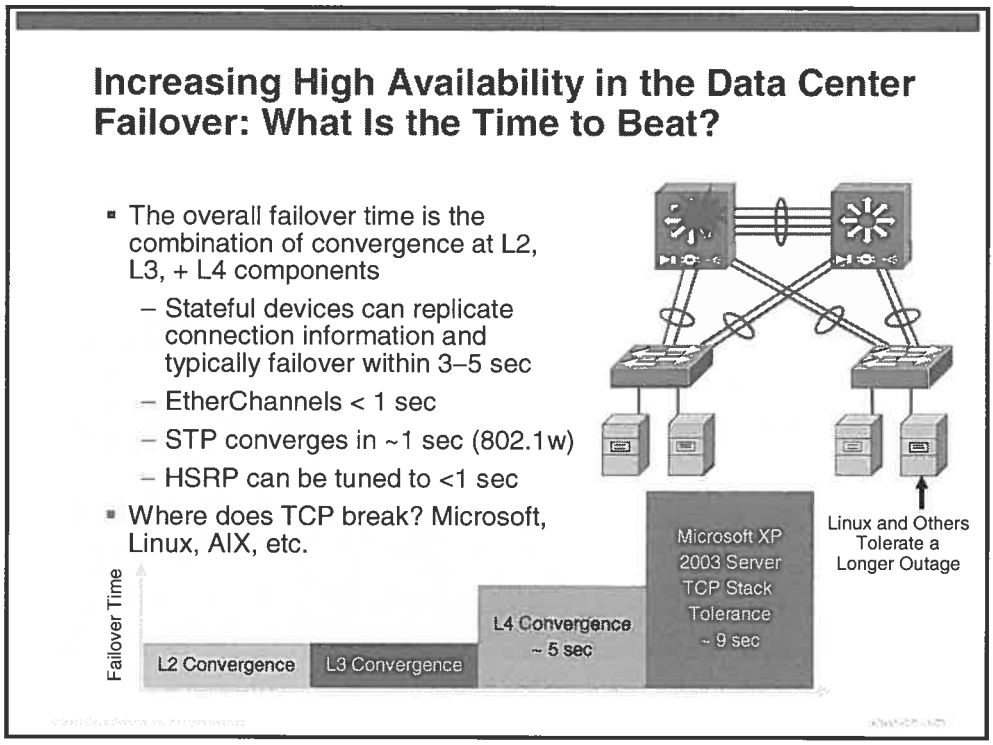
- ACE, with its auto-state awareness and tracking mechanisms, is able to rapidly initiate context failovers on events other than a module failure. ACE can track VLANs, HSRP groups, and designated IP addresses. By associating the health of an access-link to the failover of the context, ACE provides loop-free access in the presence of service modules. The preceding figure shows a "bridged design" service deployment with ACE deployed south of the FWSM. This figure also shows that the server VLAN of the ACE context has a single port. When the link fails, auto-state notifies ACE that the VLAN is down because there are no longer any active ports besides ACE itself. ACE can track this event to initiate a failover. As a result, when the primary link fails, ACE opens up the rest of the path so that traffic can pass back up to the primary HSRP router. This requires an ISL hop when there is an FWSM in the path.
- The ACE tracking mechanism for VLANs is extremely fast. When the link fails, the ACE context fails over instantly and the outage is less than a second in duration. Currently, however, there is an issue (documented in CSCsf96889) associated with the preempt mechanism that makes the primary context active again when the link is reestablished, resulting in outages ranging from 5–10 seconds. This was tested with the U design with two access switches interconnected. The access switch with the primary uplink saw outages up to five seconds with preempt where the other access switch was consistently in the 8–10 second range.

VRF Routed Design and Loop-Free Access

- Unlike service modules that use an active-standby failover mode, VRFs are always active and ready to forward traffic. As shown in the third scenario, when a Layer 3 path is available between the VRFs, a Layer 2 loop in the access can be avoided when HSRP converges. When loop-free access is configured this way, the gating factor is the time it takes HSRP to converge to the new active VRF serving as the default gateway. Using the default timers of HSRP, convergence takes place in 7–9 seconds. With the hello and hold timers set to 1/3, there is convergence within two seconds, which is not quite as fast as Rapid PVST+ and looped access, but close.

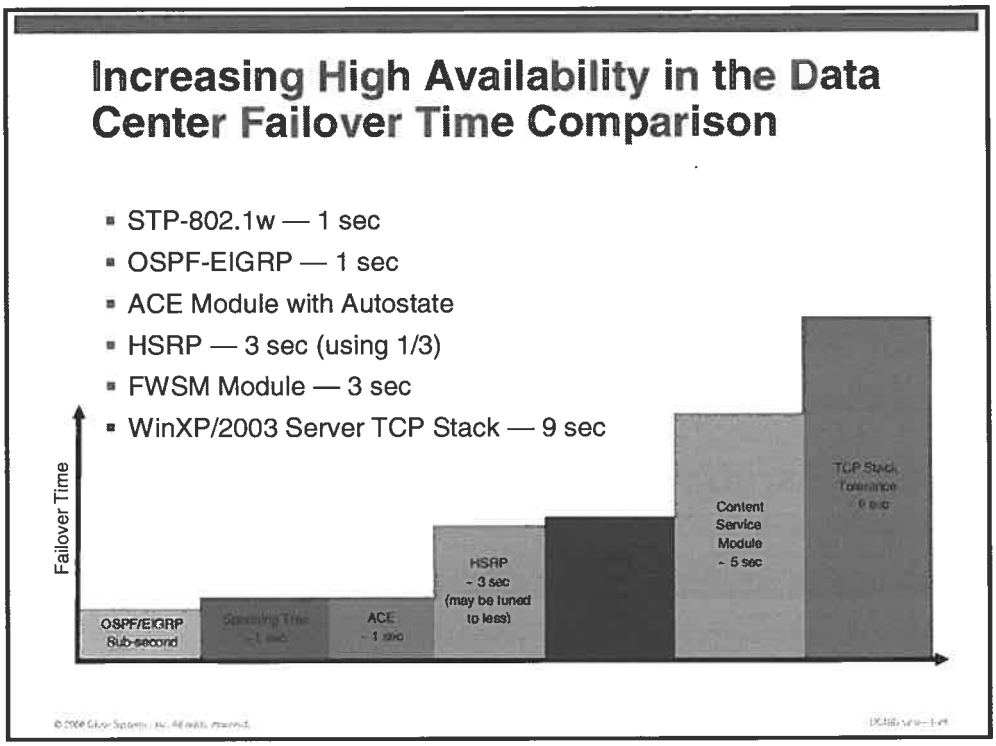
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure summarizes the characteristics of the Layer 2 and upper layer network convergence timers with respect to 9 seconds of the TCP stack tolerance.

https://t.me/learningnets



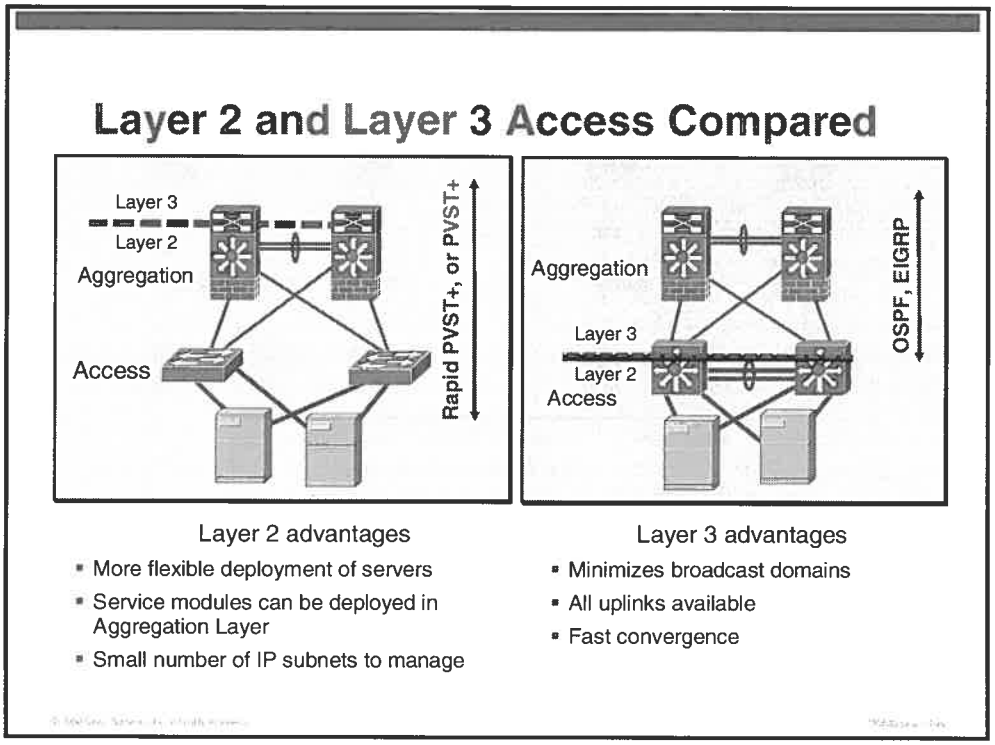
The figure summarizes the failover times for different protocols.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Layer 2 and Layer 3 Access Layer Design

This topic discusses the Layer 2 and Layer 3 access layer design options.



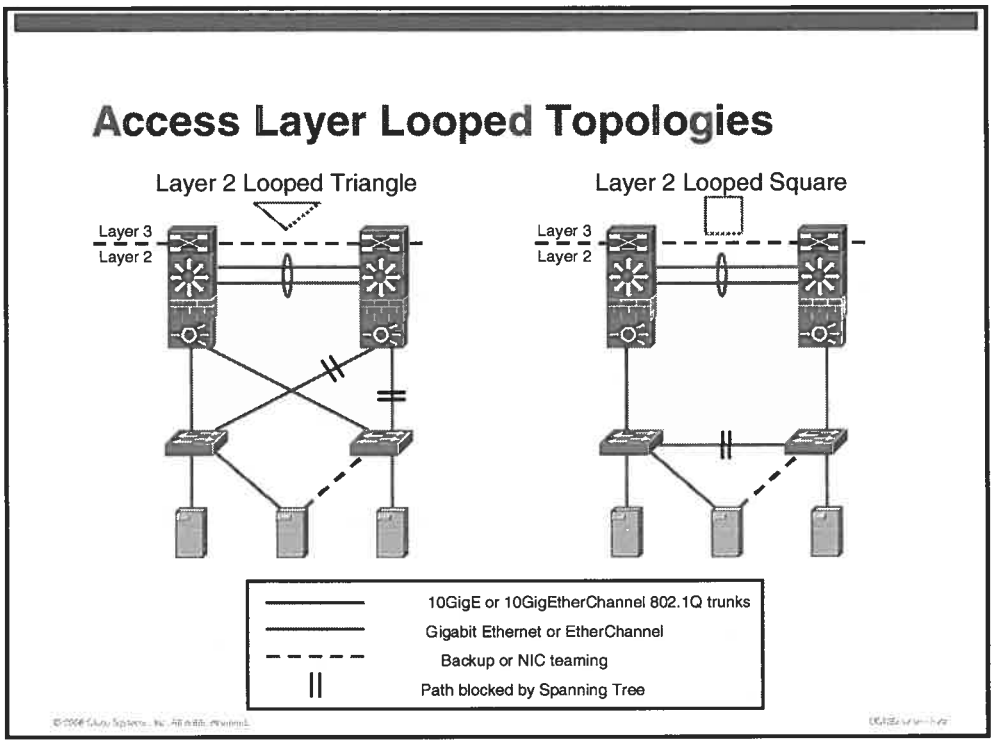
The key advantages of Layer 2 access are:

- More flexible deployment of servers
- Service modules can be deployed in Aggregation Layer
- Small number of IP subnets to manage

The key advantages of Layer 3 access are:

- Minimizes broadcast domains
- All uplinks available
- Faster convergence than STP, routing convergence is faster than spanning tree reconvergence

https://t.me/learningnets



In a Layer 2 looped access topology, a pair of access switches is connected to the Aggregation Layer using 802.1Q trunks. The two alternative topologies for looped access are based on triangle and square designs.

A VLAN is configured on each access switch on the corresponding 802.1Q uplink, and is also extended between Aggregation switches forming a looped topology for that VLAN. The left side of the figure shows an Access Layer with a triangle looped topology, and the right side shows a square looped topology. In the triangle looped topology the access switch is dual homed to each Aggregation switch. In the square looped topology, a pair of access switches are interconnected with each connected to a single Aggregation switch.

Because a loop is present, all links can not be in a forwarding state at all times. If all ports were forwarding then broadcasts/multicast packets and unknown unicast MAC address packets that are normally flooded could travel in an endless loop completely saturating the VLAN and adversely affecting network performance. An STP spanning tree protocol such as Rapid PVST+ or MST is required to automatically block a particular link and break this loop condition.

The dashed black lines on the Aggregation Layer switches represent the demarcation between Layer 2 and Layer 3 for the VLANs that are extended to the Access Layer switches. All packets processed in the VLAN beneath this line are in the same Layer 2 broadcast domain and are Layer 3 routed above the line. As denoted by the double solid lines, spanning tree automatically blocks one path to break the loop condition.

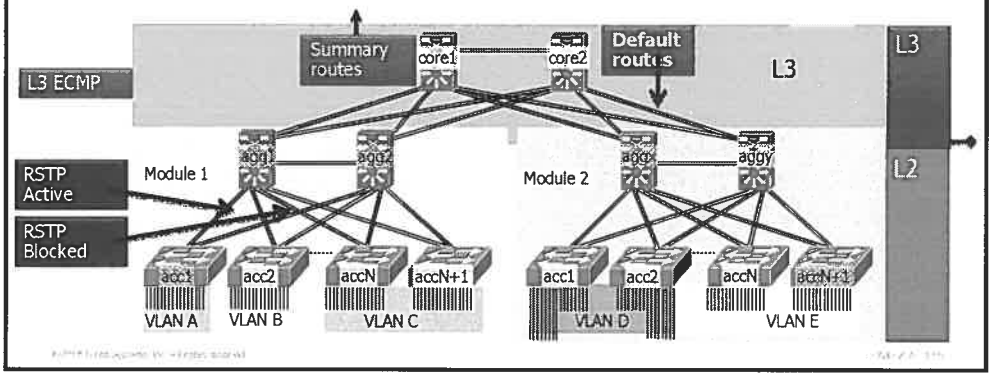
In both looped topologies, the service module fault-tolerant VLANs are extended between Aggregation switches over the 802.1Q inter-switch link. This permits active-standby hellos and session state communications to take place to support redundancy.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Classic Data Center Topology Using Δ & V Designs

- Probably most common topology
- Each module handles a number of VLANs
  - VLAN per access switch – typically loop free
  - VLAN per access switch pair – looped regardless of agg ISL and good for dual-homing
  - VLANs contained to agg module
- Failover is very predictable
  - Dual-homed environments maintain direct link to active service devices
  - Uplink capacity allows low oversubscription
- Access layer port oversubscription ranges: 6:1 – 12:1



This is probably the most widely used data center topology. This topology features a 2-tier Layer 2 domain boundary to the Aggregation Layer. PVST or RPVST are typically used, and there are a couple of ways that VLANs are used. Either a VLAN is dedicated to a switch, in which case a subnet is sized to match closely the capacity of that switch. Or a VLAN is dedicated to a pair of switches and network interface controller (NIC) teaming is used, in which case the subnet size also matches the size of that particular switch set. Sometimes a single VLAN is supported across multiple access switches, especially in Adapter Teaming active-active or active-passive scenario. The failure domain is spread out across the whole Aggregation module because that is where the spanning tree primary root and alternative root bridge are located.

The Layer 3 domain starts in the Aggregation Layer and spreads up to the core layer. Sometimes Layer 3 topology covers one tier, sometimes two tiers; this primarily depends upon whether MPLS will be in use. MPLS usually is done in the next upper tier, typically the DR is on the next tier. The route summarization can be deployed typically at the area level or at the AS level. Default routes are injected into to the core/backbone area or typically summary routes can be injected, sometimes even /32s are injected for hosts high availability into the routing tables of the core .

Aggregation to core layer connections are L3 equal cost multi-paths (ECMP) in and out the Aggregation Layer. The core layer summarizes the subnet routes toward the external/rest of the network and injects default routes for external networks toward or into the Aggregation Layer routing tables. In a triangle looped Access Layer design it is desirable to align the spanning tree root, HSRP default gateway, and active service modules on the same Aggregation switch, as shown in the figure. Aligning the Access Layer switch uplink that is in forwarding state directly to the same switch that is the primary default gateway and active service module/appliance optimizes the traffic flows. Otherwise, traffic flows can hop back and forth between Aggregation switches, creating undesirable conditions and difficulty in troubleshooting.

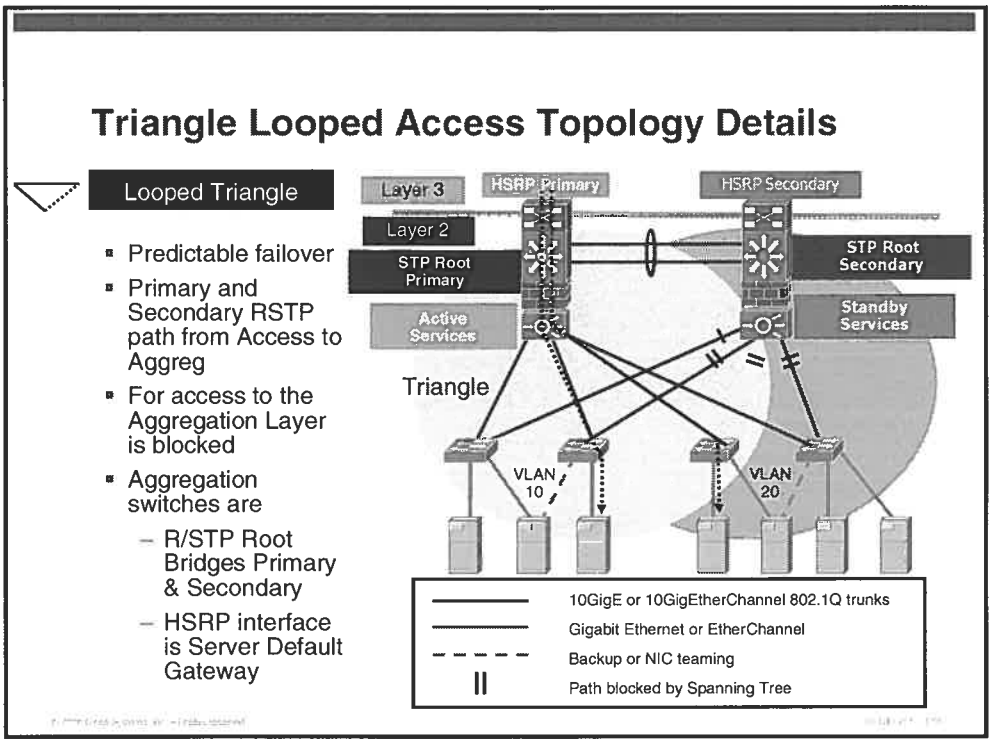
Access layer ports oversubscription ranges are from 6:1 to 12:1, depending on the type of Catalyst switches being used. Some enterprises currently use a 12:1 server oversubscription ratio but this trend is moving down to the 8:1 range to increase I/O capacity on a per server basis. With Dual-homed servers obviously you cannot assume that the same amount of link capacity as with a server that has a single NIC. The most advanced customers, when we say 5

https://t.me/learningnets

to 1, it means that a GE port puts out 200 megabits per sec (Mb/s). So then the bandwidth from five of those links could combine to a single GE uplink going out.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



In the triangular topology, the access switch connects to two Aggregation switches while the Aggregation switches interconnect carries the access layer switch VLAN(s), forming a looped topology. The triangle looped topology is currently the most widely implemented in the data center. This topology provides a deterministic design that makes it easy to troubleshoot while providing a high level of flexibility.

In the V-shaped topology, an access switch is also connected to two Aggregation switches; however, the Aggregation switches interconnect does not carry the VLAN used at the access switch, forming a loop-free topology.

The triangular and V-shaped topologies are widely used because of their predictability in steady and failover states where each access switch has a primary and redundant path to the Aggregation Layer (high availability).

Some of the characteristics of these topologies are:

- The uplinks are typically PortChannels, thus providing physical link redundancy.
- The Layer 2 and Layer 3 boundary is at the Aggregation Layer.
- Aggregation switches typically are the primary and secondary STP root bridges.
- FHRP active and standby instances match the primary and secondary root bridges, respectively.
- VLANs are contained to the Aggregation pair, and the failure domain is determined by how widely VLANs are spread to the access layer (one, many, or all access switches).
- Aggregation to core layer connections are Layer 3 equal cost multi-paths (ECMP) in and out the Aggregation Layer.
- The core layer summarizes routes toward the rest of the network and injects a default toward the Aggregation Layer.

https://t.me/learningnets

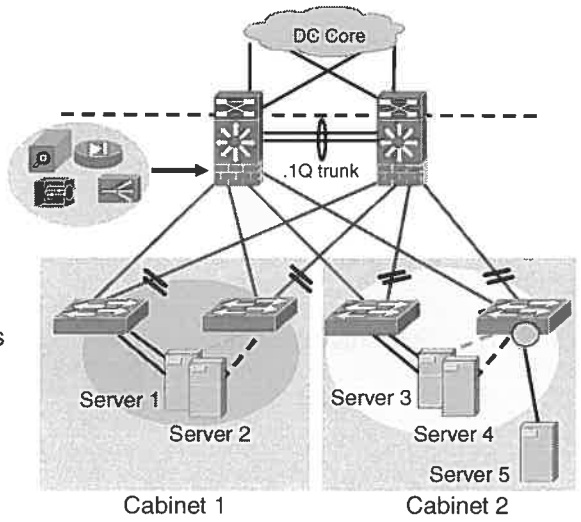
In a triangle looped Access Layer design it is desirable to align the spanning tree root, HSRP default gateway, and active service modules on the same Aggregation switch, as shown in the preceding figure. Aligning the Access Layer switch uplink that is in forwarding state directly to the same switch that is the primary default gateway and active service module/appliance optimizes the traffic flows. Otherwise, traffic flows can hop back and forth between Aggregation switches, creating undesirable conditions and difficulty in troubleshooting.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Characteristics of the Looped Topology

- Services modules can be shared
- VLANs are primarily contained between pairs of access switches
- NIC teaming and clustering can be supported across Access Layer modules

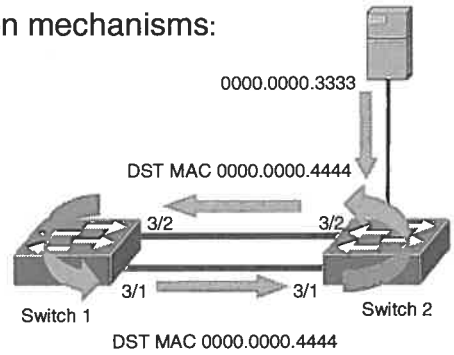


When using a looped topology, service modules such as the FWSM and ACE can be deployed at the Aggregation Layer and shared across multiple Access Layer switches. VLANs are primarily contained between pairs of access switches. A VLAN can be provisioned on a different access switch if administrative reasons require this. NIC teaming and clustering can be supported across Access Layer modules.

https://t.me/learningnets

Drawback of Layer 2 Looped Design

- If a loop occurs, the network may become unmanageable
- Employ loop prevention mechanisms:
  - Spanning Tree
  - BPDUguard
  - Rootguard
  - UDLD
  - LoopGuard



© 2008 Cisco Systems, Inc. All rights reserved.

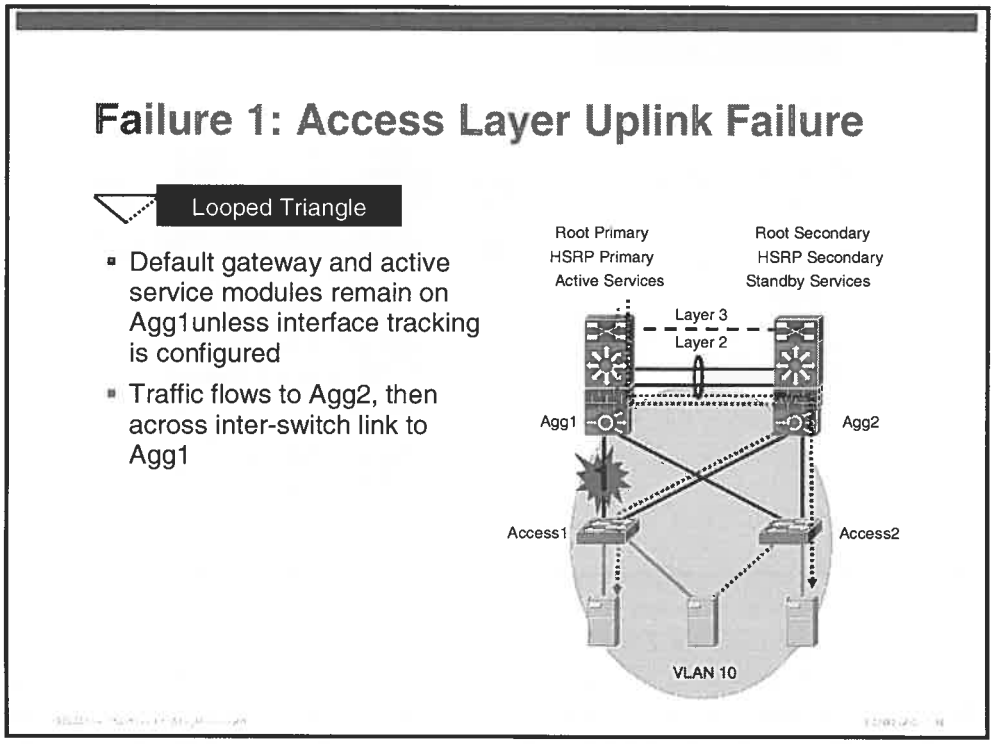
The main drawback of the looped topologies is that if a loop does occur, the network might become unmanageable due to the infinite replication of frames. Therefore, spanning tree should always be run. Conditions can occur that cause loops even when spanning tree is being run. In addition to spanning tree the following loop prevention mechanisms should be employed:

- **BPDUguard**: This is used to protect the switched network from the problems that can be caused by the receipt of BPDUs on ports that should not be receiving them. The receipt of unexpected BPDUs can be accidental or might be part of an unauthorized attempt to add a switch to the network.
- **Rootguard**: This protects against a switch outside the designated network attempting to become the root bridge by blocking its access until the receipt of its BPDUs ceases.

https://t.me/learningnets

Access Layer Looped Triangle Failure Analysis

This topic provides analysis for failures in the looped triangle data center access layer.



Upon Access Layer Link Failure the following occurs:

- Default gateway and active service modules remain on the switch Aggregation 1 unless tracking mechanisms are configured and triggered.
- Traffic flow goes through Aggregation 2 and uses the inter-switch link to Aggregation 1 to reach the active HRSP default gateway and active service module.

The convergence characteristics of this failure scenario depend on the spanning tree reconvergence. Test laboratory results show that with Rapid-PVST+ implementations, this value should be under ~1.5 seconds, but can vary based on the number of spanning tree logical and virtual ports per line card values used.

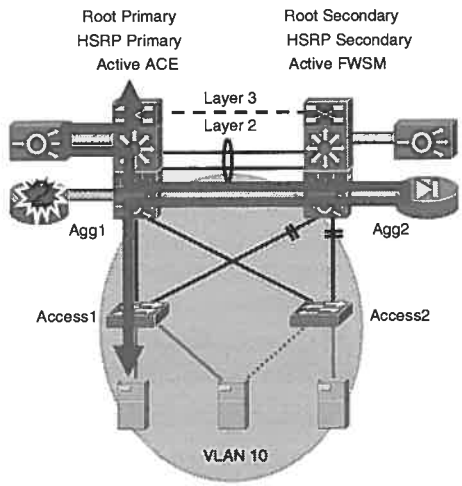
Blocked STP ports are unblocked immediately if they are part of the STP PortFast group and RSTP will reconverge almost in a second. The HSRP failover does not occur here. Some IP packets might be lost and TCP will retransmit them but the TCP sessions will not be restarted. The TCP session will not be dropped and restarted since the disconnection/disruption or L2 reconvergence does not persists for more than 8 seconds.

https://t.me/learningnets

Failure 2: Service Module Failure

Looped Triangle

- No Spanning Tree reconvergence
- FWSM on Aggregation 2 moves to active HSRP Gateway, and other service modules still active on aggregation 1
- Flow crosses the inter-Aggregation switch link twice



In this failure scenario, the FWSM is deployed in transparent mode and ACE appliance is deployed in one-arm mode. There is no spanning tree convergence up one FWSM failure and the primary default gateway remains active on the switch called Aggregation switch 1. The backup service module moves from the standby to the active state on the Aggregation 2 switch because it no longer receives hello packets from the active service module across fault tolerant link within the timeout interval. When the FWSM fails, traffic flow is redirected by the Aggregation 1 switch across the inter-switch link to the Aggregation 2 switch, through the now active FWSM module context and back across the inter-switch link to the active HSRP default gateway on the Aggregation switch 1 MSFC. Since other service modules are still active on Aggregation 1, return traffic flow is directed to them based on PBR configuration on the MSFC VLAN interface and on to the client via the core.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Firewall FWSM Failover Behavior

The failover feature lets you use a standby FWSM to take over the functionality of a failed FWSM. Failover is compatible with both routed and transparent firewall modes, and with single and multiple context modes. The FWSM supports two types of failover:

- **Regular failover**: When a failover occurs, all active connections are dropped and clients must reestablish connections when the new active unit takes over.
- **Stateful failover**: During normal operation, the active unit continually passes per-connection stateful information (for each context) to the standby unit. The interval between stateful information updates is 10 seconds, but if you set the unit polltime to be greater than 10 seconds, then that interval is used.

After a failover occurs, the same connection information is available at the new active unit. Supported end-user applications are not required to reconnect to keep the same communication session. The state information passed to the standby unit includes the following data:

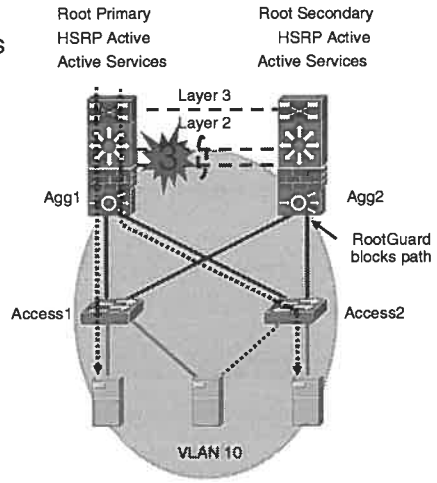
- NAT translation table
- TCP connection states
- UDP connection states (for connections lasting at least 15 seconds)
- HTTP connection states (Optional)
- H.323, SIP, and MGCP UDP media connections
- ARP table
- MAC address table (Transparent firewall mode only)

https://t.me/learningnets

Failure 3: Inter-Switch Link Failure

Looped Triangle

- Spanning tree unblocks Access Layer inter-switch link
- RootGuard blocks Access2 to Agg2 path
- HSRP hellos cannot be exchanged
- HSRP and service modules all go in the active-active state
- Black-holing or asymmetrical flows
- Use route tuning and RootGuard to avoid black-holing



This failure scenario has many side effects to consider. First, spanning tree unlocks the uplink to Aggregation 2 because no loop exists within a second. Rootguard on the Aggregation switch then automatically disables the link to access 2 because it sees root BPDUs on the now-unblocked path to Aggregation 1 via the Access Layer switch. With the inter-switch link down and RootGuard disabling the path to Aggregation 1 via Access 2, the HSRP multicast hello messages no longer have a path between Aggregation 1 and Aggregation 2, so HSRP goes into an active state on both switches for all VLANs. Because the service module failover VLANs are configured across the inter-switch link only, service modules in both Aggregation switches determine that the other has failed and become active (this is referred to as a split-brain effect).

(Wrong in the old version, of course: If the inbound traffic from the core flows into the Aggregation 2 switch during this failure scenario, it attempts to flow through the now active service modules and stops, because RootGuard has the path to the servers blocked. If for some reason RootGuard is not configured, this still results in asymmetrical flows and breaks connectivity.)

(Correction: The return traffic is load-balanced across both Aggregation switches, and service modules might disconnect the TCP connections due to traffic load-balancing.)

For these reasons Cisco recommends tuning the aggregation-core routing configuration such that the Aggregation 1 switch is the primary route advertised to the core for the primary service module-related VLANs. Route tuning plus RootGuard prevents asymmetrical connection and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions.

The infrastructure initially is without any "service," but you can build it in such a way that adding these Layers 4–7 functionalities and security does not force a redesign of the network. Service modules are stateful devices—that is, they keep track of TCP connections, NAT translate entries and higher level protocol states and synchronize them across fault tolerant link. For these reasons "service" devices must see both directions of the traffic—client-to-server and server-to-client—all the time.

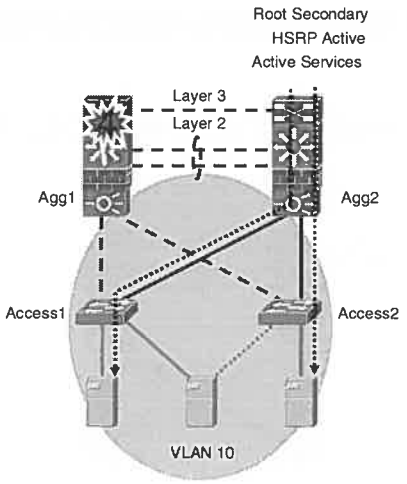
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Failure 4: Switch Power or Supervisor Failure (Non-Redundant Supervisor)

Looped Triangle

- HSRP and service modules active on Agg2
- Spanning tree root on Agg2
- Convergence time depends on service module timers

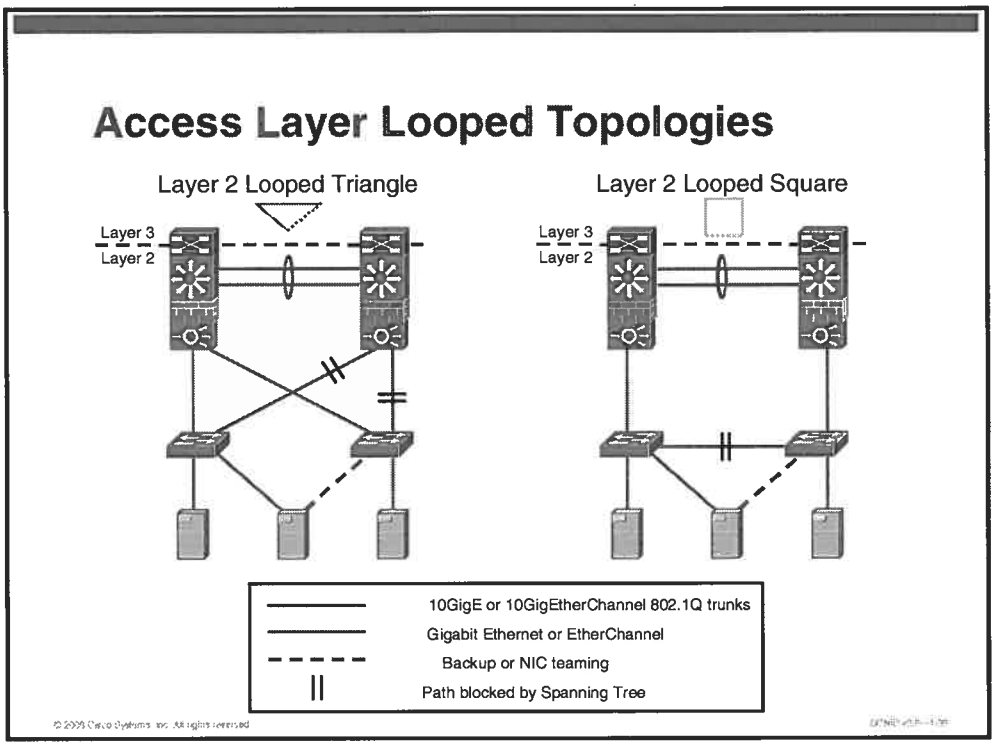


In this failure scenario, the spanning tree root, primary default gateway, and active service modules transition to the secondary Aggregation switch. The convergence characteristics of this failure scenario depend on spanning tree, HSRP, and service module failover times. Because the spanning tree and HSRP failover times are expected to be within a few seconds, they are under the other service modules failover timers. The actual convergence time depends on service module timer configurations.

https://t.me/learningnets

Access Layer Looped Topologies

This topic discusses the characteristics of the access layer looped topologies.



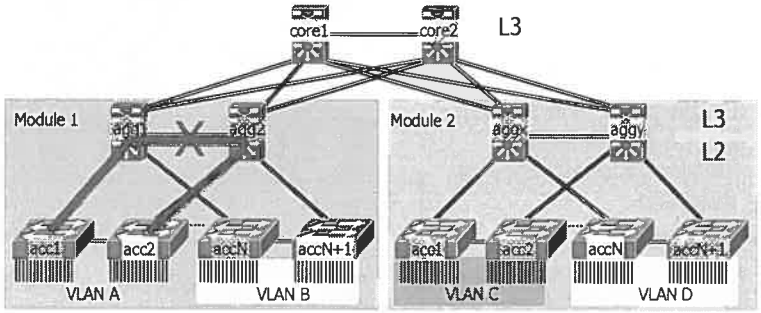
- In a Layer 2 looped access topology, a pair of access switches is connected to the Aggregation Layer using 802.1Q trunks. The two alternative topologies for looped access are based on triangle and square designs.
- A VLAN is configured on each access switch on the corresponding 802.1Q uplink, and is also extended between Aggregation switches, forming a looped topology for that VLAN. The left side of the figure shows an Access Layer with a triangle looped topology, and the right side shows a square looped topology. In the triangle looped topology, the access switch is dual homed to each Aggregation switch. In the square looped topology, a pair of access switches are interconnected with each connected to a single Aggregation switch.
- As we have see in earlier topology diagrams with a loop is present, there must be a blocking port present to prevent endless forwarding of broadcasts, multicast packets, and unknown unicast MAC address packets. A loop left in place would completely saturate the VLAN in seconds and could adversely affect network performance on other vlans connected to the same hardware. An STP spanning tree protocol such as Rapid PVST+ or MST is required to automatically block a particular link and break this loop condition.
- The dashed black lines on the Aggregation Layer switches represent the demarcation between Layer 2 and Layer 3 for the VLANs that are extended to the Access Layer switches. All packets processed in the VLAN beneath this line are in the same Layer 2 broadcast domain and are Layer 3 routed above the line. As denoted by the double solid lines, spanning tree automatically blocks one path to break the loop condition.
- In both looped topologies, the service module fault-tolerant VLANs are extended between Aggregation switches over the 802.1Q inter-switch link. This permits active-standby hellos and session state communications to take place to support redundancy.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Classic Data Center Topology using □ & U Designs

- Not as common but still utilized since it could introduce split subnet
- Each module handles a number of VLANs
  - VLAN per access switch pair – typically loop free
  - VLANs contained to Agg module
- Failover is suboptimal
  - No primary path to active service device
  - Uplink capacity limits lower oversubscription
- Oversubscription Ranges: 6:1 – 12:1

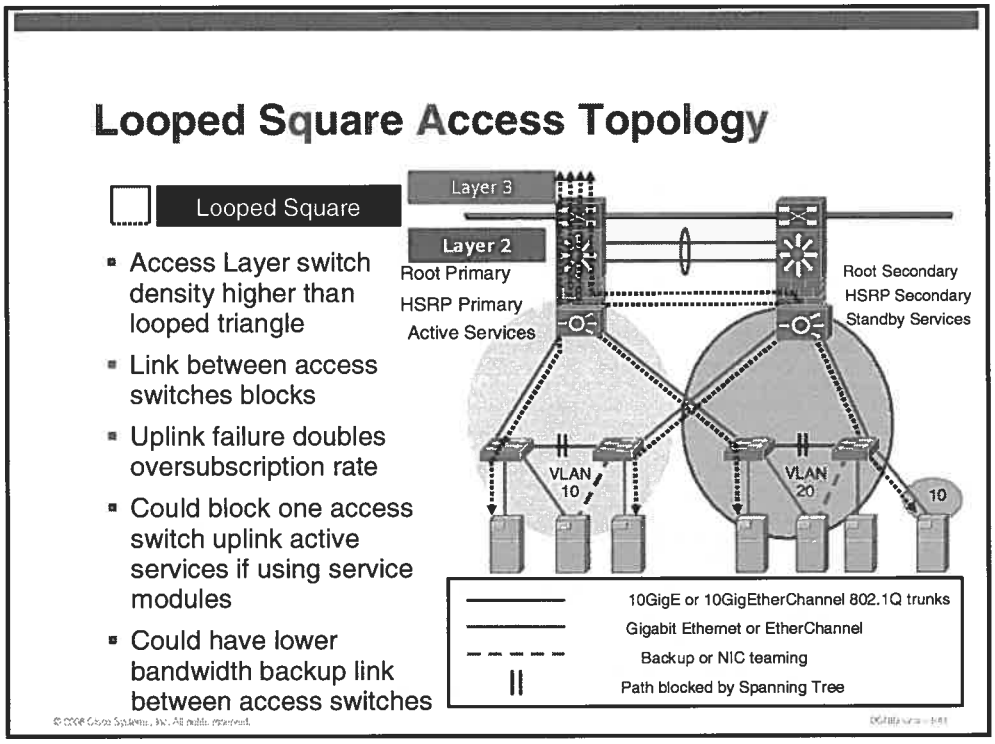


The square topology and U-shape topology are very similar. U-shape topology introduces only a pair of access switches that are interconnected. Typically a VLAN is shared between the two switches (or more than one VLAN for dual-homed servers). The problem with this model is it has a single point of failure at the access layer.

The opposite model is also possible, where the U-shaped VLAN connecting the Aggregation Layer switches is not extended to connect the access layer switches. Any network failure collision in the U-shape design can cause a major problem, especially when the redundant topology or a loop-free topology is not deployed. For instance, when the U-shape topology does not include the link that would carry a VLAN across the access layer switches and the ISL between the Aggregation Layer switches fails then there will be two isolated areas in the topology and the whole subnet will be split. However, some companies have deployed this model since they operationalized it to the extent that they know when a failure occurs, and they how to solve it. Whether this type of design makes sense depends on the situation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The square-based looped topology is not as common today in the enterprise data center as the triangle based looped topology but has recently gained more interest. The square looped topology increases the Access Layer switch density when compared to a triangle loop topology while retaining the same loop topology characteristics. This becomes particularly important when 10GE uplinks are used. This topology is very similar to the triangle loop topology, with differences in where spanning tree blocking occurs. Spanning tree blocks the link between the Access Layer switches, with the lowest cost path to root being via the uplinks to the Aggregation switches. This allows both uplinks to be active to the Aggregation switches while providing a backup path in the event of an uplink failure. The backup path can also be a lower bandwidth path because it is used only in a backup situation. This might also permit configurations such as 10GE uplinks with GEC backup.

The possible disadvantages of the square loop design relate to inter-switch link use, because 50 percent of Access Layer traffic might cross the inter-switch link to reach the default gateway/active service module. There can also be degradation in performance in the event of an uplink failure because, in this case, the oversubscription ratio doubles.

The figure shows the spanning tree blocking point on the link between the access switch pair. This is ideal if active services are deployed in each Aggregation switch because it permits the uplinks to be load balanced without traversing the Aggregation Layer inter-switch trunk.

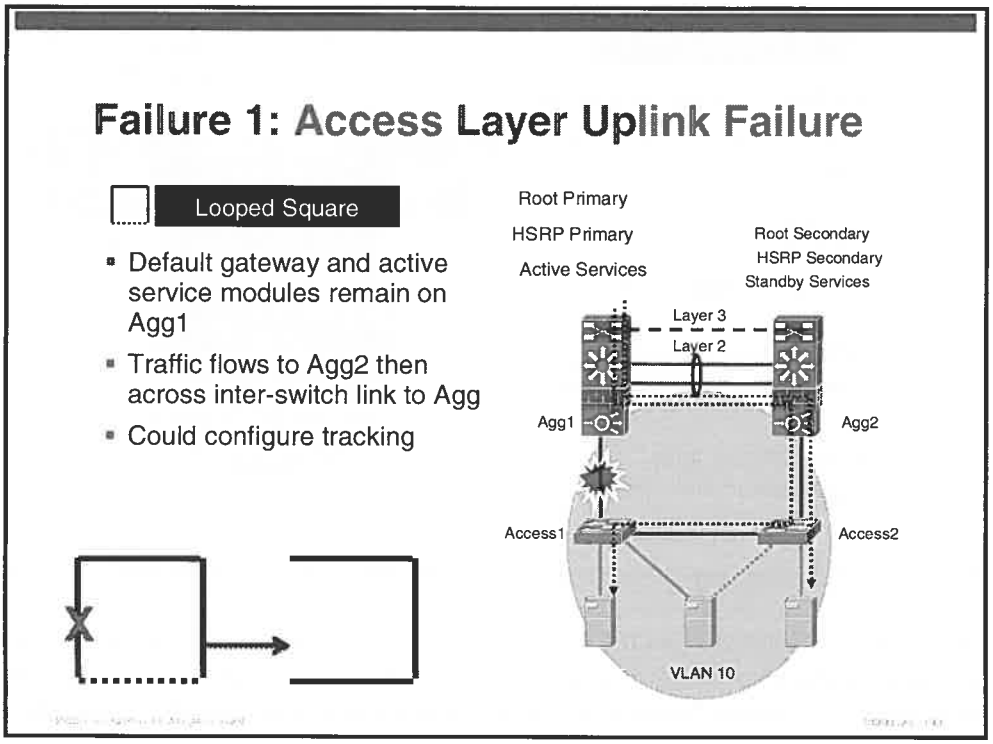
If active services are only on Aggregation 1, it might be desirable to adjust the STP cost such that the uplink to Aggregation 2 is blocked instead of the link between the access pair. This forces all traffic to the Aggregation 1 switch without having to traverse the Aggregation Layer inter-switch link. Similar to a triangle design, it is desirable in a square looped Access Layer design to align the spanning tree root, Hot Standby Router Protocol (HSRP) default gateway, and active service modules on the same Aggregation switch, as shown in the figure. By aligning the Access Layer switch uplink that is in the forwarding state directly to the same switch that is the primary default gateway and active service module/appliance, traffic flows are optimized. Otherwise, traffic flows can hop back and forth between Aggregation switches, creating undesirable conditions that are unpredictable and difficult to troubleshoot.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Failure Analysis in Looped Square Design

This topic discusses the failure in the Looped Square Design.



In this failure scenario, spanning tree unblocks the link between access switches because a loop no longer exists. The default gateway and active service modules remain on Aggregation 1 unless tracking mechanisms are configured and triggered. Traffic flows go through Aggregation 2 and use the inter-switch link to Aggregation 1 to reach the active HSRP default gateway and active service module.

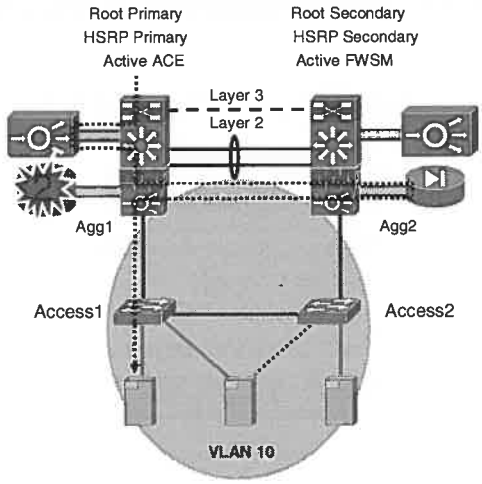
The convergence characteristics of this failure scenario depend on the spanning tree reconvergence. Test laboratory results show that with Rapid-PVST+ implementations this value should be under ~1.5 seconds, but can vary based on the number of spanning tree logical and virtual ports per line card values present.

https://t.me/learningnets

Failure 2: Service Module Failure

Looped Square

- FWSM in transparent mode
- ACE deployed in one-arm-mode
- No spanning tree convergence
- FWSM on Agg2 moves to active
- HSRP gateway and other service modules still active on Agg1
- Flow crosses inter-Aggregation link twice



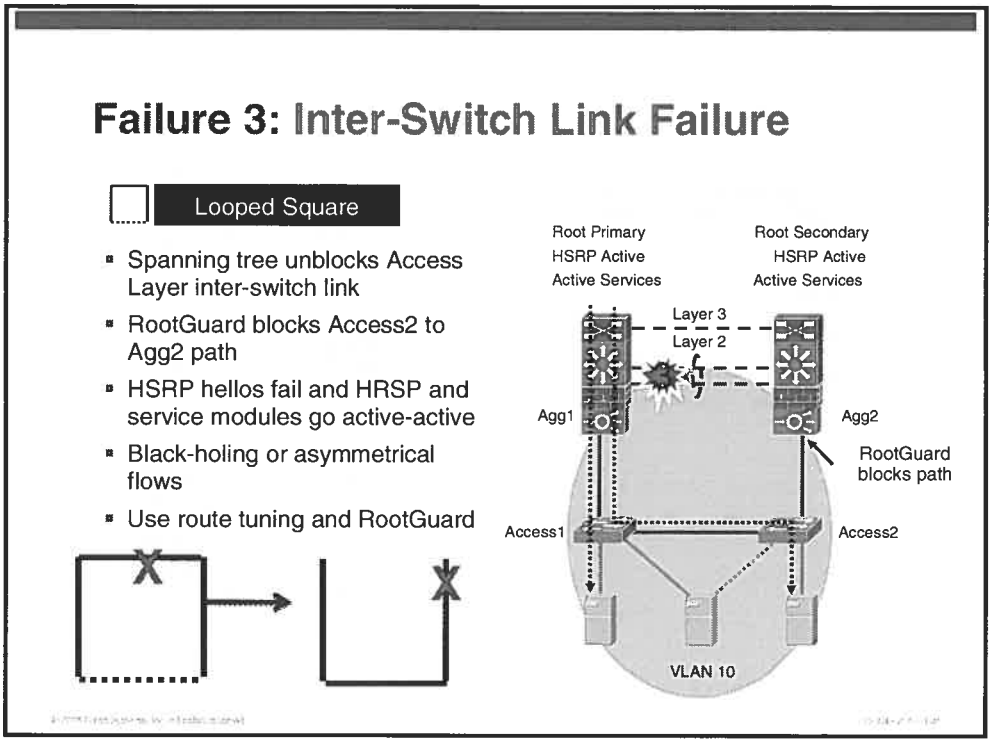
In this failure scenario when the FWSM is deployed is transparent mode and ACE appliance/module in one-arm mode. When the FWSM fails there will be no spanning tree reconvergence, and the primary default gateway remains active on the Aggregation 1 switch.

The backup service module moves to the active state on Aggregation switch 2 because it no longer receives hello packets from the active service module, and times out. FWSM failover is stateful; no TCP connection is lost or closed.

When the FWSM fails traffic flow goes through Aggregation 1 and across the inter-switch link to Aggregation 2, through the now-active FWSM module context, and back across the inter-switch link to the active HSRP default gateway on the Aggregation 1 MSFC. Since other service modules are still active on aggregation, return traffic flow is directed to them based on Policy-Based Routing (PBR) configuration on the MSFC VLAN interface and on to the client via the core.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



This failure scenario has many side effects to consider. First, spanning tree unblocks the Access Layer inter-switch link because the loop no longer exists. RootGuard on the Aggregation switch then automatically disables the link to access 2 because it receives superior BPDUs from the access layer switch.

With the inter-switch link down and RootGuard disabling the path to Aggregation1 via access 2, HSRP multicast hello messages no longer have a path between Aggregation 1 and 2, so HSRP goes into an active state on both switches for all VLANs. Because the service module failover VLANs are configured across the inter-switch link only, service modules in both Aggregation switches determine that the other has failed. This results in service modules in Aggregation 1 remaining in the active state, and service modules in Aggregation 2 moving from standby to the active state as well. This is commonly referred to as split-brain effect and is very undesirable.

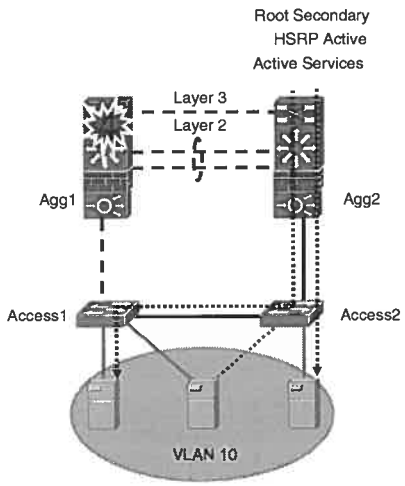
If inbound traffic from the core flows into the Aggregation 2 switch during this failure scenario, it attempts to flow through the now-active service modules and stops, because RootGuard has the path to the servers blocked. If for some reason RootGuard is not configured, this still results in asymmetrical flows that break connectivity. For these reasons Cisco recommends tuning the aggregation-core routing configuration such that the Aggregation 1 switch is the primary route advertised to the core for primary service module-related VLANs. Route tuning plus RootGuard prevents asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions.

https://t.me/learningnets

Failure 4: Switch Power or Supervisor Failure (Non-Redundant Supervisor)

Looped Square

- HSRP and service modules active on Agg2
- Spanning tree root on Agg2
- Convergence time depends on service module timers



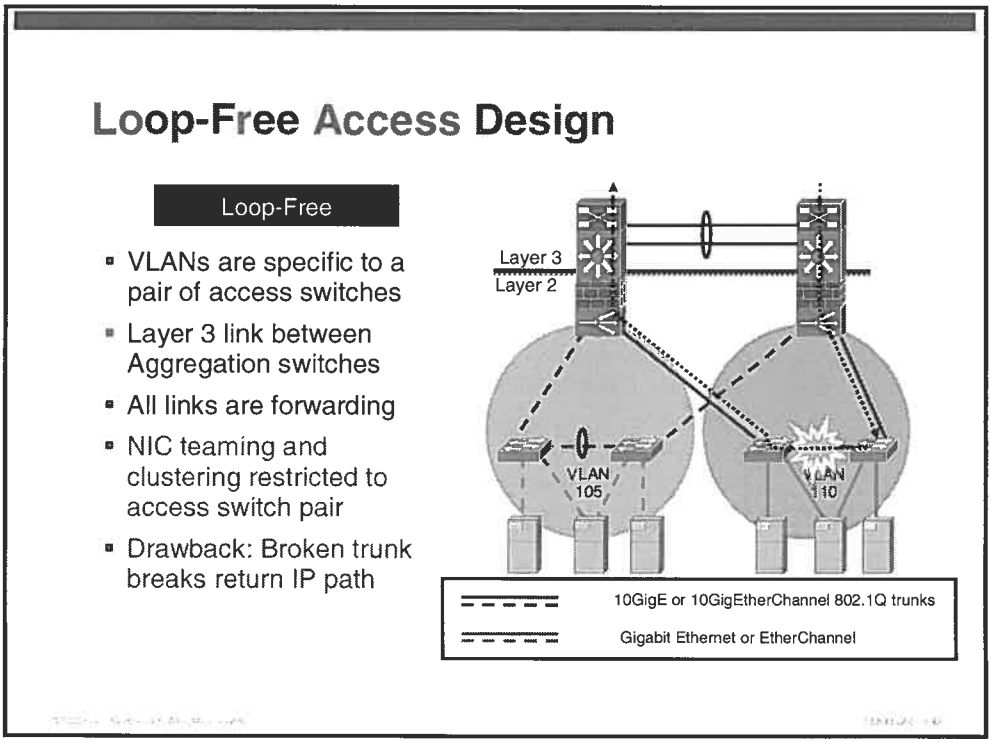
In this failure scenario where the supervisor failure occurs and the spanning tree root, primary default gateway, and active service modules transition to the Aggregation 2 switch. The convergence characteristics of this failure scenario depend on spanning tree, HSRP, and service module failover times. Because the spanning tree and HSRP failover times are expected to be under that of service modules, the actual convergence time depends on service module timer configurations. Fault tolerance heartbeats activate all standby services in the standby aggregation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Loop-Free Design

This topic describes discusses the loop-free data center design.



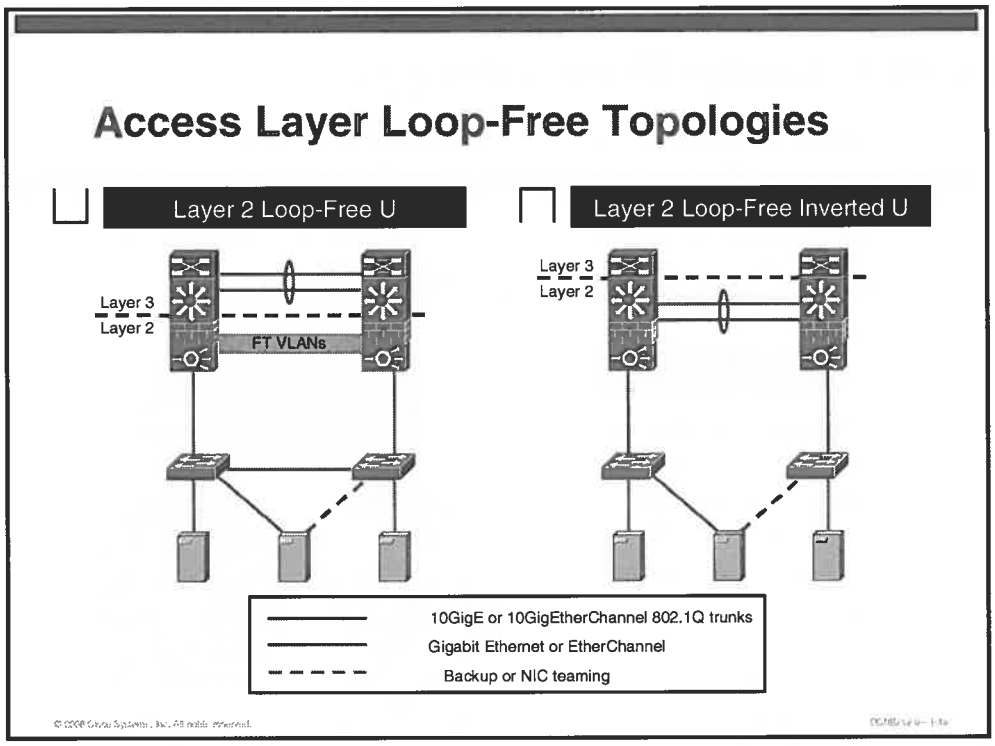
In the loop free topologies each pair of access switches are assigned a set of VLANs that are specific to that pair. No VLANs are trunked between Aggregation switches. Instead, the inter-Aggregation Layer switch link is a Layer 3 link. Spanning tree does not block any ports under normal operation but is still run as a protection against configuration and cabling errors. NIC teaming and clustering can only be supported within Access Layer modules.

- VLANs are not extended across inter-switch link trunk for loop-free access.
- In the loop free topologies each pair of access switches are assigned a set of VLANs that are specific to that pair. No VLANs are trunked between Aggregation switches. Instead, the inter-Aggregation Layer switch link is a Layer 3 link. Spanning tree does not block any ports under normal operation but is still run as a protection against configuration and cabling errors.
- NIC teaming and clustering can be supported only within Access Layer modules.

The use of a loop free topology has several drawbacks:

- If the trunk between the access switch pairs is broken, the return IP path might be broken.
- VLANs must be restricted to access switch pairs.
- If VLANs are extended between Access Layer modules, then spanning tree blocking can occur.

https://t.me/learningnets



The two alternative topologies for the Layer 2 loop-free model are loop-free U and loop-free inverted U. Note that the Layer 2–Layer 3 line of demarcation is different in each design. In a loop-free U, a VLAN is configured on each access switch, and on the 802.1Q inter-switch link between access switches and its corresponding 802.1Q uplink, but is not extended between Aggregation switches, thereby avoiding a looped topology. In a loop-free inverted U design, a VLAN is configured on each access switch and its corresponding 802.1Q uplink, and is also extended between Aggregation switches, but is not extended between access switches, avoiding a looped topology.

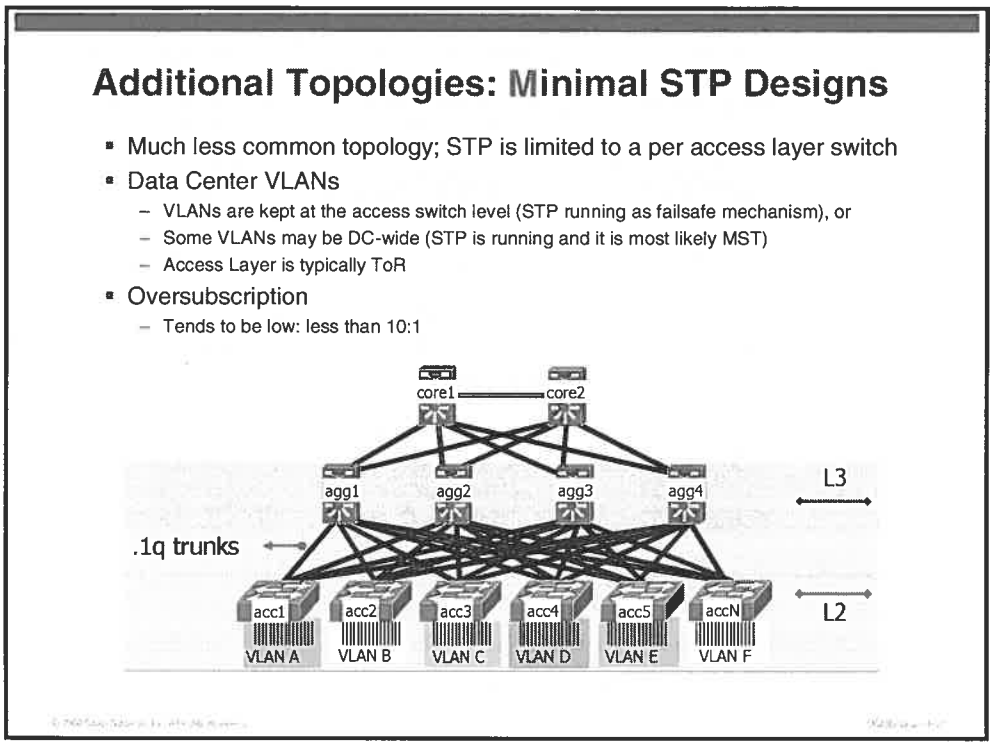
Although no loop is present in either loop-free design topology, it is still necessary to run STP as a loop prevention tool. In the event that a cabling or configuration error that creates a loop is encountered, STP prevents the loop from possibly bringing down the network. In the loop-free U design, you cannot use RootGuard on the Aggregation to Access Layer links because the Aggregation 2 switch would automatically disable these links because root BPDUs would be seen. In both loop-free topologies, the service module fault-tolerant VLANs are extended between Aggregation switches over the 802.1Q inter-switch link. This permits active-standby hellos and session state communications to take place to support redundancy.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

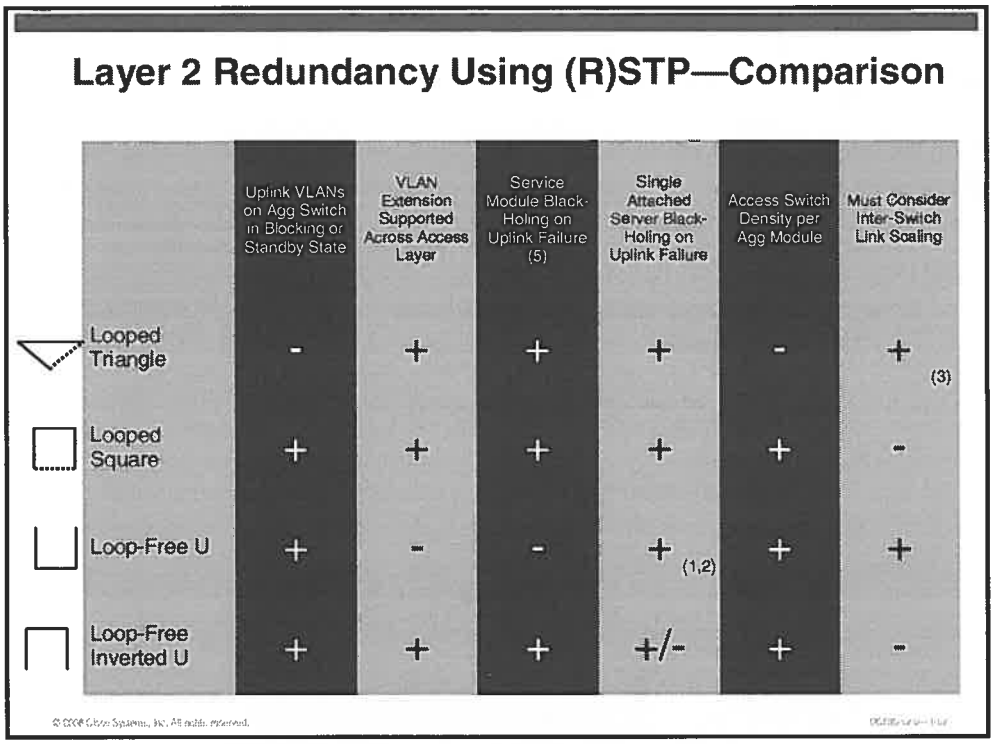
Minimal Layer 2 Design

This topic describes the minimal Layer 2 design.



Spanning tree, a layer 2 protocol, has slow convergence compared to some of the Layer 3 topology convergence mechanisms available and causes problems with blocked STP ports. In order to minimize the STP convergence impact, the STP tree has to be minimized. The Access Layer VLANs are not connected through Aggregation switches. Access Layer switches have 802.1q trunks up to Aggregation Layer switches, which is still a L2 connection, but none of Aggregation layer switches communicate with Access Layer Spanning trees. The Spanning tree is basically limited to a per access layer switch. (Oversubscription of this access layer can be as high as 10:1.)

https://t.me/learningnets



The table in the preceding figure summarizes the characteristics of the four most common spanning tree design models and their attributes:

- **Uplinks in blocking or standby state:** Some Access Layer designs can use both uplinks (active-active), while others have one link active and the other blocked on a per-VLAN basis by spanning tree, or completely unused in a backup mode only. A plus is given to those models that have both uplinks active.
- **VLAN extension across the Access Layer:** A plus is given to those access design models that permit a VLAN to be extended to all access switches that are connected to a common Aggregation module.
- **Service module black holing:** An uplink failure on the Access Layer switch could break connectivity between the servers and the service modules being used.
- **Single attached server black holing:** If an access switch has a single uplink, it could be a large failure exposure point. Uplinks that use Distributed EtherChannel can reduce the chances of black holing. Server load balancing to a VIP that includes servers physically connected across multiple access switches is another technique that can be used, as well as server NIC teaming.
- **Access switch density per Aggregation module:** When 10GE uplinks are used, port density at the Aggregation Layer can be a challenge. Some Access Layer designs permit a larger number of Access Layer switches per Aggregation Layer module than others.
- **Inter-switch link bandwidth scaling:** Some Access Layer designs send all traffic toward the primary root Aggregation switch, while other designs send traffic toward both Aggregation switches. When sending to both Aggregation switches, 50 percent of the traffic typically passes over the inter-switch link to reach the active HSRP default gateway and active service module pair. The amount of bandwidth used for the inter-switch links becomes very important in these designs and can create scaling challenges

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

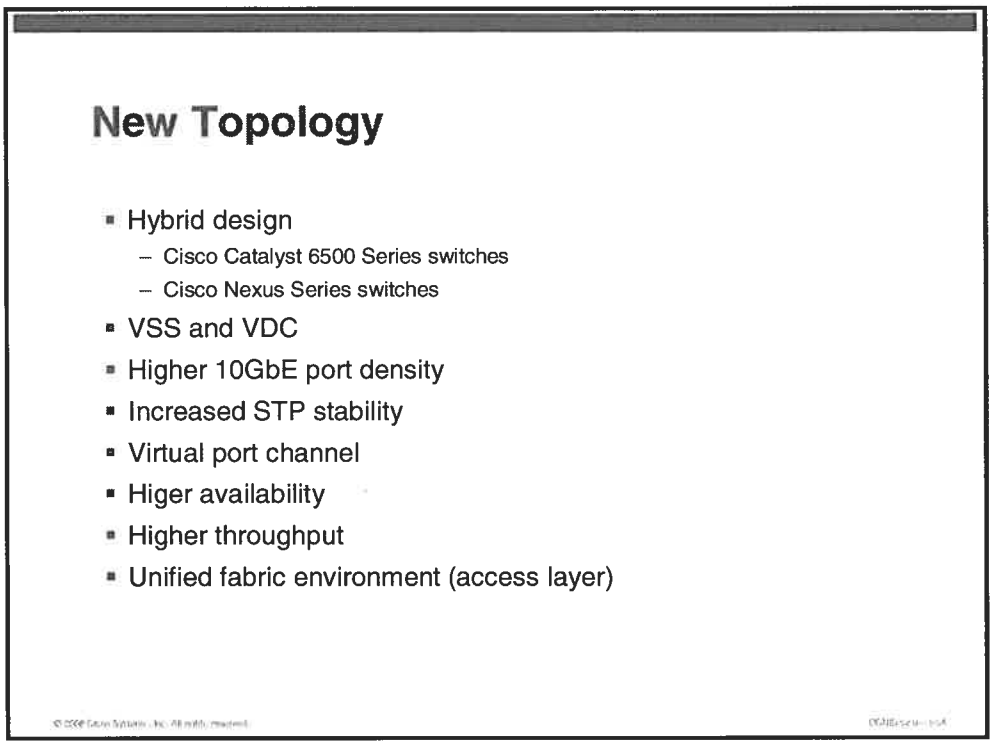
Note

- Use of Distributed EtherChannel greatly reduces chances of a black holing condition.
- NIC teaming can eliminate a black holing condition.
- When service modules are used and active service modules are aligned to aggregation, ACE module permits Layer 2 loop-free access with per-context switchover on uplink failure.

https://t.me/learningnets

Data Center New Topologies

This topic discusses integration of VSS, VDC and VRF features in the new data center network topology.



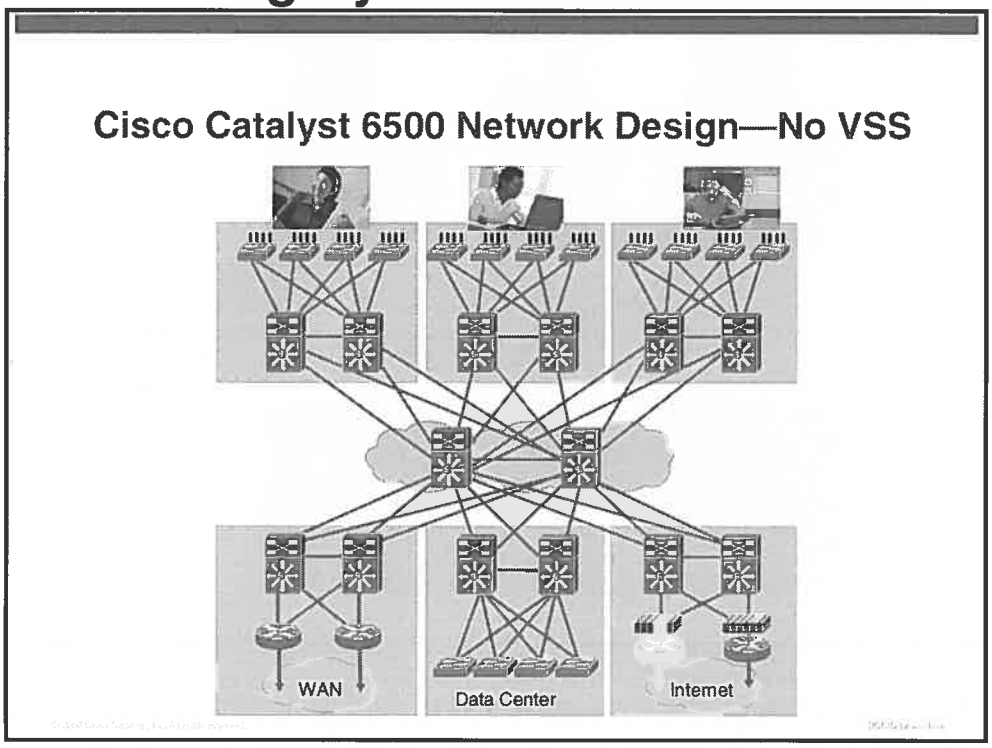
The new design topologies are taking advantage of the new technology enhancement available on the Catalyst 6500 Series and Nexus 7000 Series switching products, providing the following new functions:

- Higher 10GbE port density
- Increased STP stability
- Virtual Switching
- Virtaul PortChannel

© 2008 Cisco Systems, Inc.

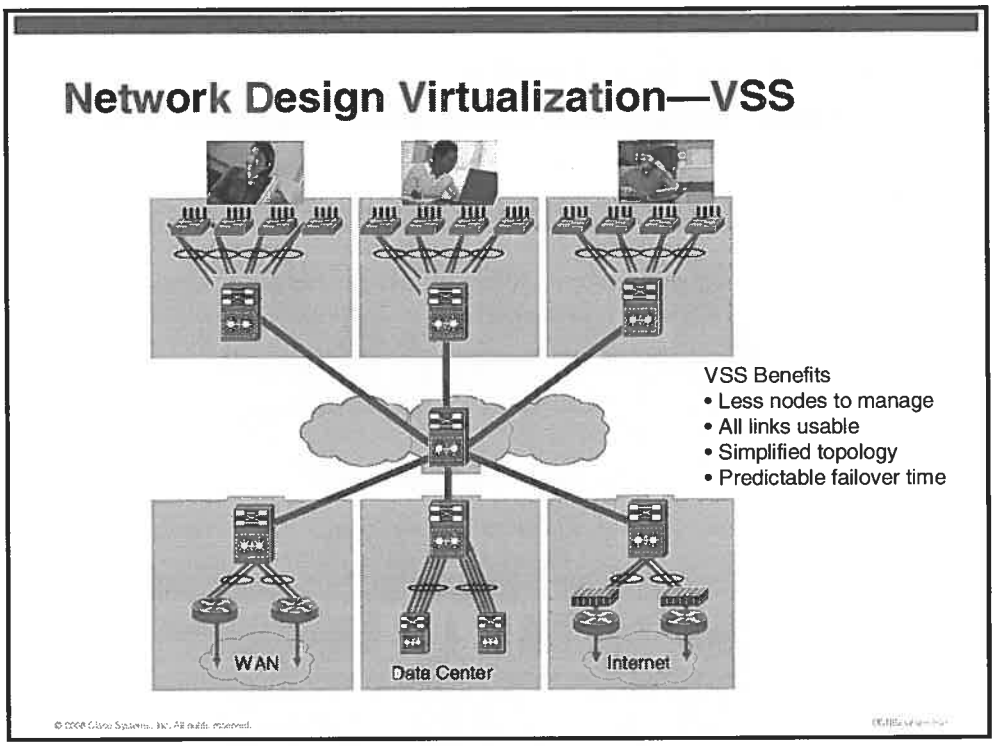
https://t.me/learningnets

Virtual Switching Systems



This is a traditional example of a data center network with a campus layer, data center layer, Internet DMZ zone, and WAN zone.

https://t.me/learningnets



With VSS, the architecture is simplified with less number of nodes and with all links actively propagating the traffic. No blocking link exists in this VSS respective design and the failover is predictable because the non-stop switching failover is predictable, which is sub-200 milliseconds. The Cisco Catalyst 6500 Series Virtual Switching System (VSS) 1440 increases operational efficiency by enabling multiple Catalyst 6500 Series switches to share a single point of management, single IP address, and single routing instance, as well as by eliminating the dependence on the STP. The system boosts non-stop communications by delivering deterministic, subsecond Layer 2 link recovery through inter-chassis stateful failovers and the predictable resilience of EtherChannel.

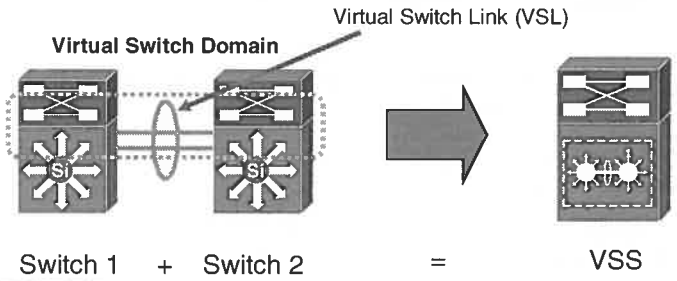
Cisco VSS 1440 scales system bandwidth capacity to 1.4 Tbps by activating all available bandwidth across redundant Catalyst 6500 Series switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

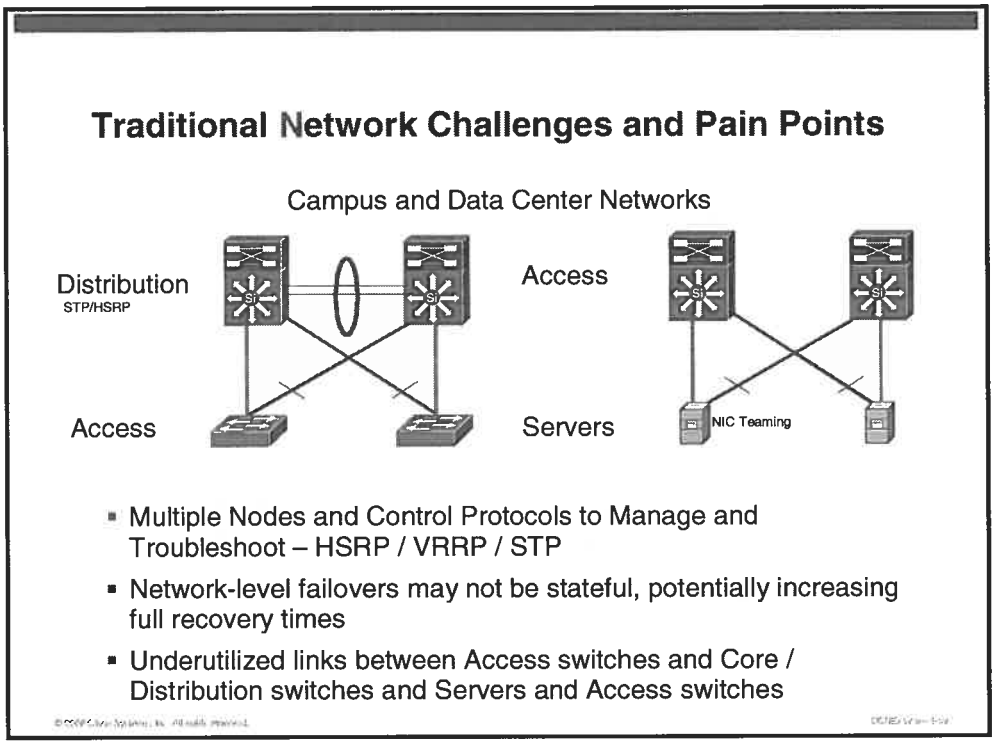
Virtual Switch: Catalyst 6500 Virtual Switching System (VSS)

- Virtual Switching System consists of two Catalyst 6500 Serial switches defined as members of the same virtual switch domain running a VSL (Virtual Switch Link) between them
- VSS Acts like a Single Control Plane with Dual Active Forwarding Planes
- Extends NSF/SSO infrastructure to Two Switches



VSS technology can be used in the data center Access Layer, the Aggregation Layer, and sometimes even in the Core Layer. The most typical deployment is the Aggregation layer. It is the central critical point of a data center network where two redundant distribution switches are to be provided. Two switches in the Aggregation Layer introduce a lot of complexity in terms of managing, monitoring, and troubleshooting the data center along with the complexity of defining which HSRP group is active for which of the server VLANs. VSS simplifies the problems of this nature by deploying two physical switches operating as a single logical switch image, which eliminates the complexity mentioned above. VSS is meant to provide increased availability of a single switch, faster convergence, and better traffic capacity using load balancing.

https://t.me/learningnets

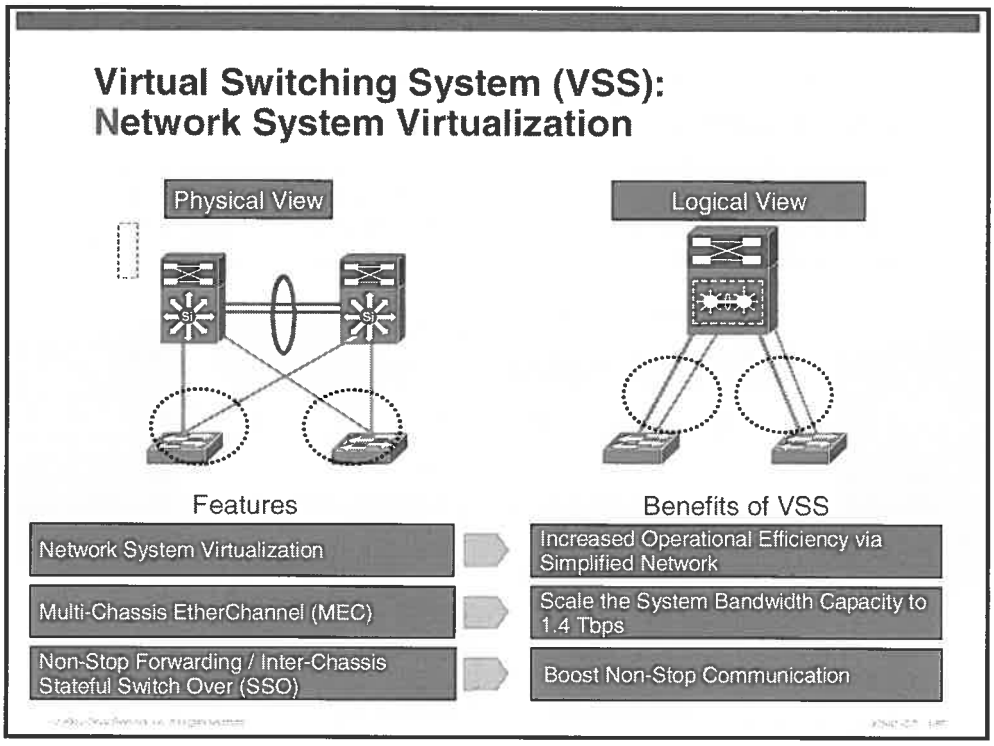


When there is a failover protocol such as HSRP, GLBP and others, the network needs a lot more time to converge as opposed to a network that does not rely on these failure protocols upon failure recovery. In the data center, servers are typically using NIC adapter teaming where one of the NICs is active and another is in standby mode and, in this case, the bandwidth is not fully utilized.

In the campus access space, access layer switches are also configured as dual homed to distribution layer switches and are also not fully utilizing the bandwidth because spanning tree is blocking in one of the links that prevents full utilization of the available bandwidth. VSS is supposed to resolve these types of data center challenges.

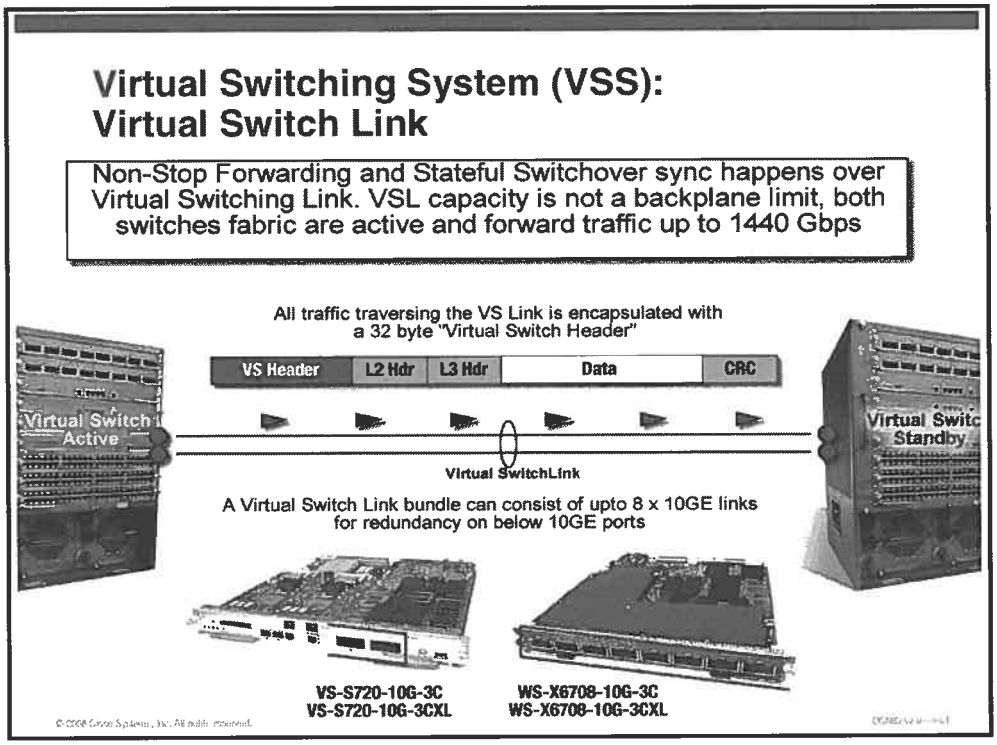
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The virtual switching system is essentially two Catalyst 6500 Series switches, acting like one single virtual switch. VSS allows scaling the capacity of two Catalyst 6500 Series switches to 1440 system capacity, providing the deterministic failover times. With the cross-chassis NSF, VSS can guarantee sub 200-millisecond failover times. The chassis does not have to be configured in the same way.

https://t.me/learningnets



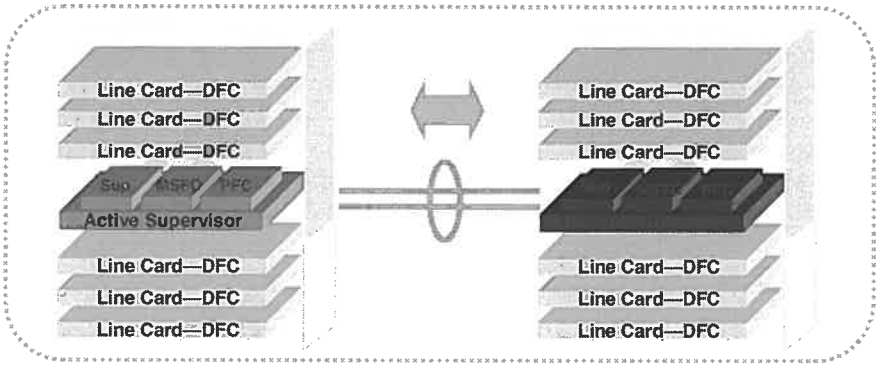
VSL can be any 10-Gb link, or up to 8 10G links bundled and is used to transfer control traffic between the two switches: control traffic meaning NSF-SSO features. The amount of bandwidth used for this control traffic is typically less than 5 percent of a single 10 Gbps link. For redundancy purposes it is recommended that at least two 10-Gb links are configured as part of the virtual switching link.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Virtual Switching System (VSS): Single Control Plane

- Dual chassis VSS leverages Non-Stop Forwarding (NSF) and Stateful Switchover (SSO) with the redundant supervisors split between two chassis.
- OIR events are processed by the active supervisor.
- Power management is distributed to active and standby supervisor.

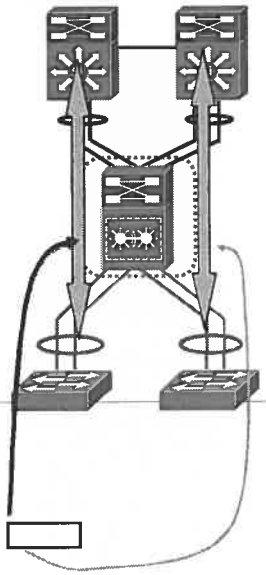


VSS fundamentally is a hot standby supervisor running in a second chassis, forming a single control plane with dual active forwarding planes. VSS is effectively takes control of the redundant supervisor model using the active supervisor module to run across two physical switch chassis, maintaining two active data planes. The active supervisor is actively managing the switch with all the hardware programmed. Additionally, the standby supervisor is running the same software and is actively learning CAM entries by monitoring packets. Both modules perform hardware synchronization functions.

https://t.me/learningnets

Virtual Switching System (VSS): Dual Active Forwarding Planes

- HSRP not needed.
- STP problem eliminated.
- Load Balancing on MEC per flow base.
- Both supervisors are actively forwarding data traffic (PFCs).
- All Distributed Forwarding Engines (DFCs) are actively forwarding.



VSS uses one active control plane, thus eliminating the need for HSRP and eliminating Spanning Tree loops. VSS uses dual active forwarding planes. (A single control plane is active with a high availability supervisor model with dual active control planes.) HSRP can be active on one side, while the other uplink is not utilized, sort of odd and even load balanced VLANs, which is considered to be the best practice.

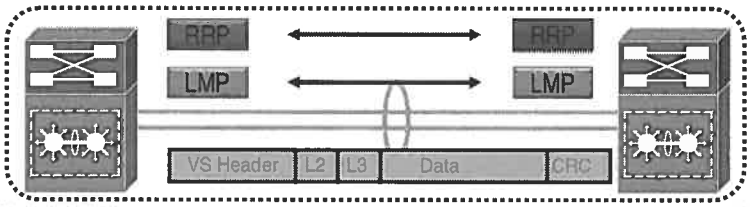
Both switches are actively forwarding traffic from all links, with the traffic load balanced on a per-flow basis, enabled by the multi-chassis EtherChannel (MEC).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

VSS Architecture: Virtual Switching Link (VSL)

- VSL provides two functions:
  - Control plane extension - Provides CPU to CPU communication
  - Data plane extension - Extends the switching fabric
- Configured over a 10GE EtherChannel link between both switches.
- QoS is predefined on the VSL link to protect control plane.
- Link Management Protocol (LMP) runs on each individual link that is part of the VSL, programs forwarding indices, as well as link maintenance checks.
- Role Resolution protocol (RRP) provides the mechanism to check HW/SW compatibility and determine Active/Standby roles.



VSL is a 10 Gbps EtherChannel enabled as a VSL link, which provides for the extension of the control plane bus and the control plane and switching fabric/data plane interconnection across the two Catalyst chassis. Two new protocols are running over the VSL—the Link Management Protocol (LMP), which is a link maintenance protocol that checks the link status and bundles all the links together into the VSL bundle and creates an EtherChannel out of this group of links.

The second protocol is the Role resolution protocol (RRP), which actually decides on the switch boot up—which (supervisor) side is active and which side is standby—and checks whether all the hardware and software is enabled at the required level. VSL is configured over a 10GE EtherChannel link between both switches. QoS is predefined on the VSL link to protect the control plane. LMP runs on each individual link that is part of the VSL for programs forwarding indices, as well as link maintenance checks. RRP provides the mechanism to check hardware and software compatibility and determine active-standby roles. The 32-byte header is encapsulating all frames, control planes, and data frames being sent across the link. The figure shows the 32-byte header that is passed across the bus to the remote PFC when the packets are forwarded within an individual chassis, thus extending the forwarding capability across the two chassis.

https://t.me/learningnets

VSS Architecture: DFC Requirements

- Distributed feature support requires that all PFC/DFC are operating in the same mode.
- Switch will revert to the lowest common denominator in order to ensure all features consistent across switch.
- VSS requires *both* switches be running in the same mode (either PFC3C or PFC3CXL).



The distributed feature support requires that all PFCs/DFCs are operating in the same mode. The switch will revert to the lowest common denominator in order to ensure that all features are consistent across the switch. VSS requires *both* switches to be running in the same mode (either PFC3C or PFC3CXL.) In addition, VSS is currently operating as a fabric-only supported architecture. In the current release only Cisco 6700 Series line cards are supported; all the other service modules will be support in the next release.

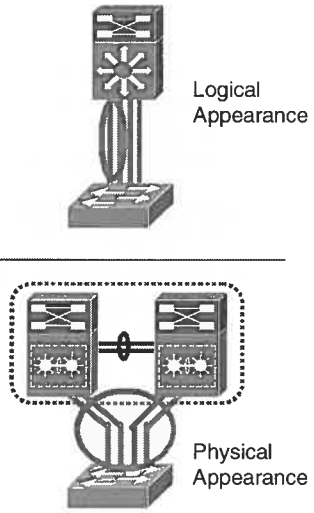
Note The only interfaces that support VSL with 32-byte header encapsulation are the Cisco 6708 Ethernet line cards and the 10 G interface ports on the new VSS Sup720 processor with a PFC3 based supervisor. The capability is needed in the new PFC to have the LTL logic to address ports on two physical switches. Because the old ASICS did not understand a 1/1/2 port on the second switch.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Virtual Switch System (VSS): Multi-Chassis EtherChannel (MEC)

- Multi-Chassis EtherChannel (MEC) replaces spanning tree as the means to provide link redundant loop-free topology.
- MEC allows the physical members of the EtherChannel bundle to be connected to two separate physical switches.
- MEC links on both switches are managed by PAgP or LACP running on the Master Switch via internal control messages.
- PAgP or LACP packets for all links in the MEC bundle are processed by the active supervisor.



Multi-chassis EtherChannel

EtherChannel is used for the fabric stability, high availability, load balancing and for an increased link capacity, and it traditionally runs between two switches—for example, between an access switch and a distribution switch, or two core switches, and so on.

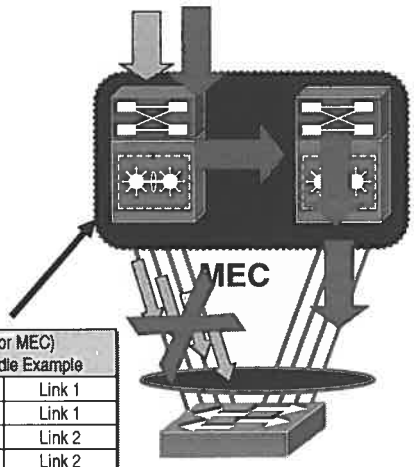
VSS is deploying the EtherChannel and allows it to run across the two chassis; therefore, it is called multi-chassis EtherChannel. From the access perspective, when there are redundant connections to two physical switches, they still represent an EtherChannel that acts like a single logical link connecting upstream to the distribution layer switch. From the control plane perspective, the Spanning Tree is not needed since EtherChannel is providing redundancy. VLAN load balancing on the EtherChannel uses hash load-balancing on a per flow basis on the traffic running upstream.

MEC replaces spanning tree as the means to provide link redundant loop-free Topology. MEC allows the physical members of the EtherChannel bundle to be connected to two separate physical Switches. MEC links on both switches are managed by PAgP or LACP running on the Master Switch via internal control messages. PAgP or LACP packets for all links in the MEC bundle are processed by the active supervisor module.

https://t.me/learningnets

Impact of VSS on the Campus Design: MEC Traffic Flows

- In a MEC configuration traffic is forward over local members of the Etherchannel bundle (all 8 buckets hash to a local link).
- Designed to prevent sending traffic across the VSL link unnecessarily.
- If all local links fail, the RBH is programmed to forward across the VSL link.



RBH (for MEC) 8 Link Bundle Example	
Bit 7	Link 1
Bit 6	Link 1
Bit 5	Link 2
Bit 4	Link 2
Bit 3	Link 3
Bit 2	Link 3
Bit 1	Link 4
Bit 0	Link 4

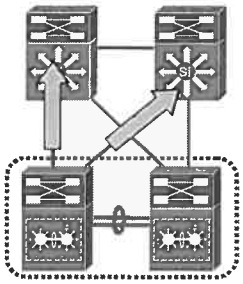
The VSL is a 20 Gbps Ethernet fabric extension. The VSS transports up to 720 Gbps across a chassis. Local skewing is used for any packet entering either one of the two switches and the VSS will try to never send it across the VSL link. Regardless of how many links are participating in the EtherChannel, traffic will be load balanced appropriately across all available links except the VSL. The RBH value used and hash lookup performed direct the packet across a particular link. Each switch will only return an RBH for the locally attached links. Only when there are no other links available will the RBH traffic be sent across the VSL bundle.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Impact of VSS on Campus Design: ECMP Traffic Flows

- Logically 4 ECMPs, but every VSS member in HW sees/uses only two paths for ECMP.
- ECMP follows a similar behavior, local links are preferred and all traffic is forwarded out of a locally attached link.
- Hardware FIB inserts entries for ECMP routes using locally attached links.
- If all local links fail, the FIB is programmed to forward across the VSL link.



```
cr2-6500-vss#sh ip route 10.121.0.0 255.255.128.0 longer-prefixes
D       10.121.0.0/17
          [90/3328] via 10.122.0.33, 2d10h, TenGigabitEthernet2/2/1
          [90/3328] via 10.122.0.27, 2d10h, TenGigabitEthernet1/2/1
          [90/3328] via 10.122.0.22, 2d10h, TenGigabitEthernet2/2/2
          [90/3328] via 10.122.0.20, 2d10h, TenGigabitEthernet1/2/2

cr2-6500-vss#sh mls cef 10.121.0.0 17 switch 1

Codes: decap - Decapsulation, + - Push Label
Index  Prefix              Adjacency
102400 10.121.0.0/17       Te1/2/2        , 0012.da67.7e40 (Hash: 0001)
                           Te1/2/1        , 0018.b966.e988 (Hash: 0002)
```

4 ECMP Entries

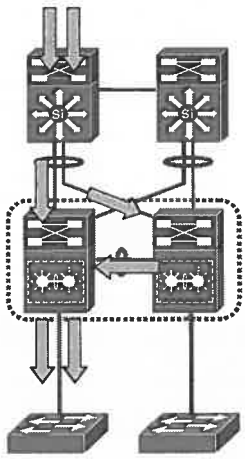
2 FIB Entries

When equal cost path links are present between the two routing points the VSS traffic path selection behavior will be the same as listed on the previous figure. There are four links from the routing protocols perspective from the VSS system to the core two switches. In hardware on every VSS switch member only the two local have been programmed which means that traffic is sent only over the links that are connected to this particular VSS switch member. The same approach is followed for the return traffic—the VSS will try to make sure that no traffic flows is sent across the VSL link.

https://t.me/learningnets

Impact of VSS on Campus Design: Single Homed Devices—Orphan Port

- Singly attached devices can result in sub-optimal traffic flows.
- When using MEC there is no method to communicate to the core which of the two VSS switches to forward traffic to (it looks like a single switch to the core).
- On average 50% of all traffic will pass over the VSL link.
- In practice, does not differ from a traditional design with single homed subnets using route summarization to the core.



It is possible to connect hosts or switches to a single VSS switch member but it is not recommended. VSS is not meant to be used for that purpose. These types of devices are called orphan ports or ports that are logically attached to only one switch. Orphan ports are not recommended because data flows will get suboptimal.

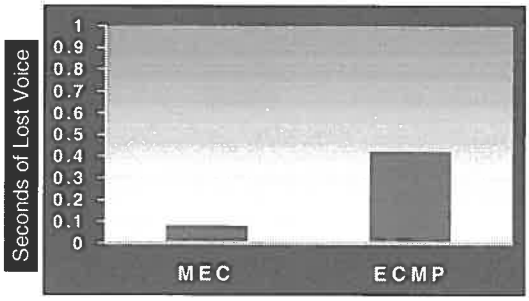
This is similar to building a network with an access switch with a single link. The network will still get traffic running across the distribution link; but from a VSS perspective it would create suboptimal flows. VSS works best when there are 2 connections for everything because of added redundancy, load balancing, fast convergence, and fast failover.

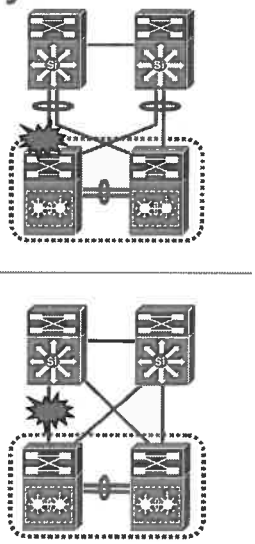
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

ECMP or MEC: Link Recovery

- MEC convergence is consistent, independent of the number of routes.
- ECMP convergence is dependent on the number of routes.



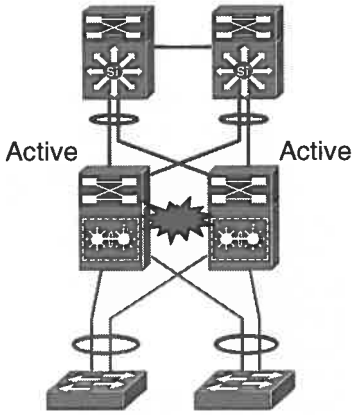


From a convergence perspective EtherChannel is very consistent, has a very deterministic and a very fast failover mechanism and should be, therefore, preferred over ECMP when fast convergence is the design priority.

https://t.me/learningnets

VSS VSL Failure: Dual Active VSS State Upon

- VSL provides for control plane extension between the two Supervisors with one Supervisor active and one in hot-standby mode.
- On catastrophic loss of all members of the VSL bundle, the standby supervisor will go ACTIVE.
- Potentially, two switches with same MAC address, IP address, … will exist on the network when this occurs.
- Two mechanisms to provide recovery from dual active state.
  - Enhanced PAgP
  - Dual Active Detection over IP-BFD

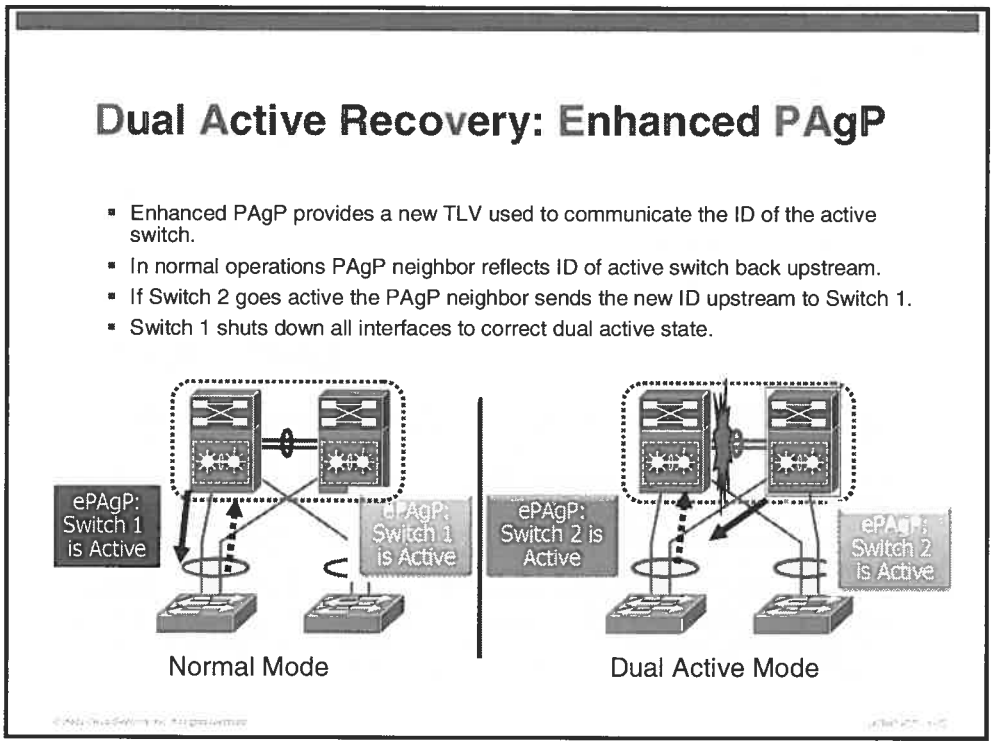


© 2008 Cisco Systems, Inc. All rights reserved.

The VSS is sensitive to the failure of the physical links between the two VSS switch members. When there are two switches and one is in the hot standby mode, if all possible VSL connectivity between the switches is lost, both switches will become active. The result will be two switches in the network, both with the same IP address and the same MAC addresses.

The dual active state should never occur, as there should always be a role resolution via an in-band or out-of-band mechanism to restore and recover returning the network to a stable state.

https://t.me/learningnets

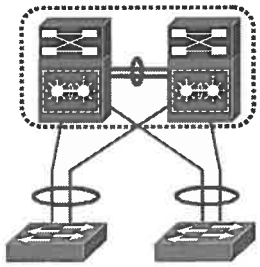


PAgP+ is an enhanced version of Cisco's PAgP protocol and is similar to LACP also known as 802.3ad. Cisco's PAgP+ has been enriched by adding another TLV frame in order to enable both of the physical switches in the VSS communicate downstream to all of their attached neighbors when they are configured with PAgP. This, in addition to being part of this EtherChannel, means that the switch has informed them that it is the active or standby switch in the respective EtherChannel bundle. When the client, the access switch, any neighbor switch, or the core switch are running the enhanced PAgP+ protocol, then he can receive the TLV frames, inspect them and discover the active and the standby VSS members and reflect this information back upstream. During normal operations that process continues. If there is a catastrophic fiber failure where all the fiber is cut between the two VSS switches, and both become active, an out-of-band arbitration responds. All of the access switches that were configured with this function will report to the VSS that they are receiving active switch message from both sides. In response to that, the switch that was originally active will be shut down and will allow the switch that has just recently become active to take control, remain active, and to maintain forwarding through the remaining chassis.

https://t.me/learningnets

Dual Active VSS Members Recovery: Enhanced PAgP

- PAgP dual active detect is globally enabled by default.
- Need to explicitly trust specific neighbors.



```
cr2-6500-vss(config)#switch virtual domain 10
cr2-6500-vss(config-vs-domain)#dual-active detection pagp trust channel-group 205

cr2-6500-vss#sh switch virtual dual-active pagp
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

Channel group 205 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
           Dual-Active     Partner              Partner   Partner
Port       Detect Capable  Name                 Port      Version
Gi1/8/19   Yes             cr7-6500-3           Gi5/1     1.1
Gi1/9/19   Yes             cr7-6500-3           Gi6/1     1.1
Gi2/8/19   Yes             cr7-6500-3           Gi5/2     1.1
Gi2/9/19   Yes             cr7-6500-3           Gi6/2     1.1
```

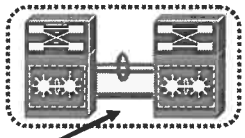
PAgP+ is enabled globally by default on the VSS switch however, at a minimum one trusted switch is needed in order to make the Dual active VSS recovery possible. Software supporting this feature is currently supported on for the Catalyst 6500, both Sup720 and Sup32. For the Catalyst Series 2900, 3x50, and 4500, support for PAgP+ will be available in the next software release.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Dual Active Recovery: IP-BFD

- IP-BFD utilizes a "heartbeat" mechanism to provide dual active detection.
- Utilizes a direct pt-pt fiber connected to an interface on each switch.
- Must have a unique IP address on each end of the link.



IP-BFD pt-pt link

```
cr2-6500-vss(config)#interface gigabitethernet 1/5/1
cr2-6500-vss(config-if)#no switchport
cr2-6500-vss(config-if)#ip address 200.230.230.231 255.255.255.0
cr2-6500-vss(config-if)#bfd interval 100 min_rx 100 multiplier 50
cr2-6500-vss(config-if)#interface gigabitethernet 2/5/1
cr2-6500-vss(config-if)#no switchport
cr2-6500-vss(config-if)#ip address 201.230.230.231 255.255.255.0
cr2-6500-vss(config-if)#bfd interval 100 min_rx 100 multiplier 50
cr2-6500-vss(config-if)#exit
cr2-6500-vss(config)#switch virtual domain 100
cr2-6500-vss(config-vs-domain)#dual-active detection bfd
cr2-6500-vss(config-vs-domain)#dual-active pair interface g 1/5/1 interface g 2/5/1 bfd

adding a static route 200.230.230.0 255.255.255.0 Gi2/5/1 for this dual-active pair
adding a static route 201.230.230.0 255.255.255.0 Gi1/5/1 for this dual-active pair
```

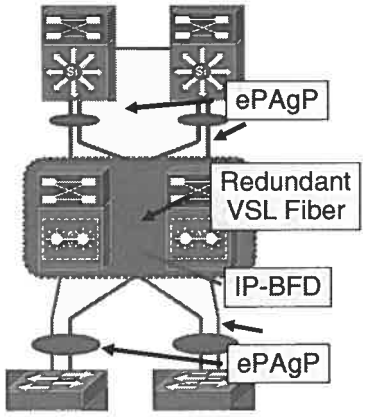
Requires unique IP subnets on the two interfaces

In cases where code upgrade is not possible because of older generation switch hardware or if there is non-Cisco switching third-party hardware attached to the VSS, an alternative mechanism is available called IP-BFD. A heartbeat cable is connected between the two VSS switches creating an out-of-band path available to accomplish the role resolution. When all the connectivity between the two distribution member VSS switches is lost then there is still a path available.

https://t.me/learningnets

Dual Active Recovery: Multiple Mechanisms

- Ensuring the availability of the VSL link is a high priority.
- Redundant fiber paths recommended to protect against physical fiber failures.
- IP-BFD provides a "heartbeat" like backup link which can follow another fiber path.
- PAgP only needs to be run on a single neighbor but …
- Leveraging enhanced PAgP on all interfaces will ensure that in the worst case if any switch is connected to both members of the same VSS pair, then a path will exist to correct the problem.



Dual Active Recovery mechanisms properties:

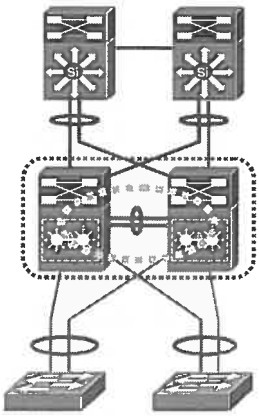
- Ensuring the availability of the VSL link is a high priority.
- Redundant fiber paths recommended to protect against physical fiber failures.
- IP-BFD provides a "heartbeat" like backup link that can follow another fiber path.
- PAgP needs to be run on only a single neighbor, but leveraging enhanced PAgP on all interfaces will ensure that, in the worst case, if any switch is connected to both members of the same VSS pair then a path will exist to correct the problem.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Impact of VSS on Campus Design: Capacity Planning for the Virtual Switch Link

- Capacity planning and link sizing for VSS is almost identical to traditional Multilayer Design.
- The only traffic that should flow across the VSL under normal conditions is control plane traffic.
- In an access switch uplink failure, half of the downstream traffic will be forwarded across the VSL link.
- Control Plane load is very small and sent with strict priority over the VSL link.
- Redundancy of the VSL is critical and should take priority over capacity planning.

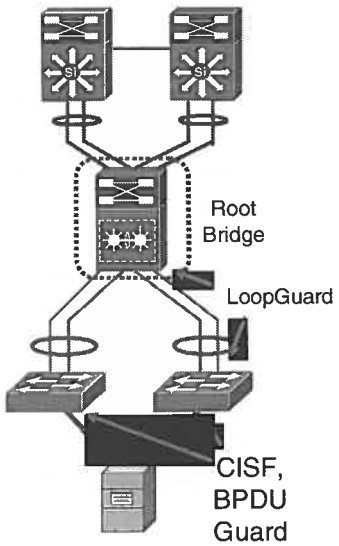


Traffic should never pass across the VSL link except in the case of an access to distribution uplink failure. The ideal configuration is to run two 10 Gbps connections between the VSS members, even if in production data throughput only 2Gbps of capacity is needed. This is needed for redundancy and for control plane connectivity.

https://t.me/learningnets

Impact of VSS on Campus Design: Control Plane Simplification

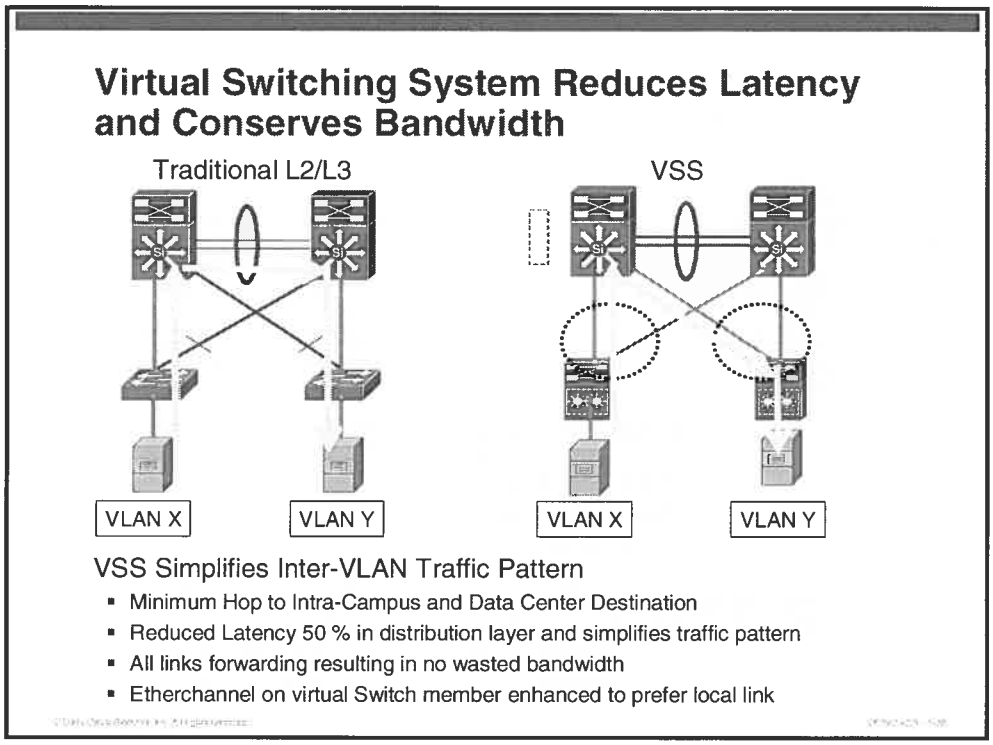
- Virtual Switch Design simplifies the topology.
- Redundant supervisors provide resiliency via SSO.
- No need for HSRP, GLBP, or VRRP.
- A single multicast router on the access subnets simplifies the multicast topology.
- No L2 loops in the topology, so need for spanning tree to provide for link redundancy.
- Do *not* disable spanning tree as it is still possible to create an external loop.



The value VSS brings to the network is simplified logical topology with a single distribution device, and eliminated need for the HSRP, GLBP, VRRP protocol and Layer 2 loops, thus eliminating the need for STP. The VSS network topology is loop-free, yet it provides all the same physical hardware and software redundancy as the legacy traditional multi-tier data center design. The Cisco enhanced value-add Spanning Tree features are still implemented, since physically it is still possible to create a topology that results in forwarding loops due to a design mistake or engineer error. CISF security features, BPDU-Guard, and Portfast features that are deployed on the edge of the network are still expected to be enabled.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



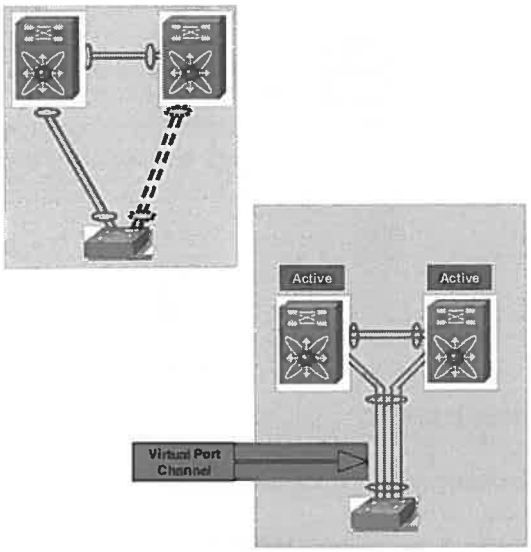
The figure on the left shows the traditional Layer 2, Layer 3 network topology where routing is deployed between the two different VLANs.

In the figure on the right, VSS is used so all links are forwarding. There is no need to send inter-VLAN traffic to the other switch. Instead traffic is sent directly - saving bandwidth and reducing the latency of traffic propagation due to a smaller number of hops.

Latency is an increasingly important factor in video and voice applications.

https://t.me/learningnets

NX-OS: Virtual PortChannel (vPC) (a.k.a. MCEC)



- Virtual Switch retains physical switches independent control and data plane –light version than VSS
- Active/Active data plane where local forwarding is preferred to avoid potential stress on inter-switch links
- Increase usable bandwidth, by eliminating STP blocked ports
- Virtual PortChannel mechanism is transparent to hosts or switches connected to the virtual switch
- STP as fail-safe mechanism to prevent loops, even in the case of control plane failure

One of the key technologies that Cisco will be introducing in the NX-OS 4.1 release is support for Virtual Port Channel. This has also been called Multi -Chassis EtherChannel. This has also been called Multi \Chassis EtherChannel. vPC or virtual PortChannel is a new name assigned to it. This is a lighter complexity implementation than VSS, and does not require sync up of the complete control plane. vPC addresses the issue with STP. STP will block one set of uplinks, and only part of the BW will be used. The Virtual Switch is the pair of upstream switches; they retain independent control plane and data planes. As a result we have active-active data planes where forwarding takes place on the local switch in the majority of cases, to avoid traffic stress on inter-switch links. The benefit of the vPC feature is that it keeps all the links active (STP loop avoidance mechanism will block some of the links), and hence increasing the cross sectional bandwidth to the servers. The new virtual port channel is transparent to the hosts or downstream switch, and does not require any special configuration. It is advised to continue running STP as a fail-safe, as it will continue to protect against the accidental loops or misconfiguration.

© 2008 Cisco Systems, Inc.

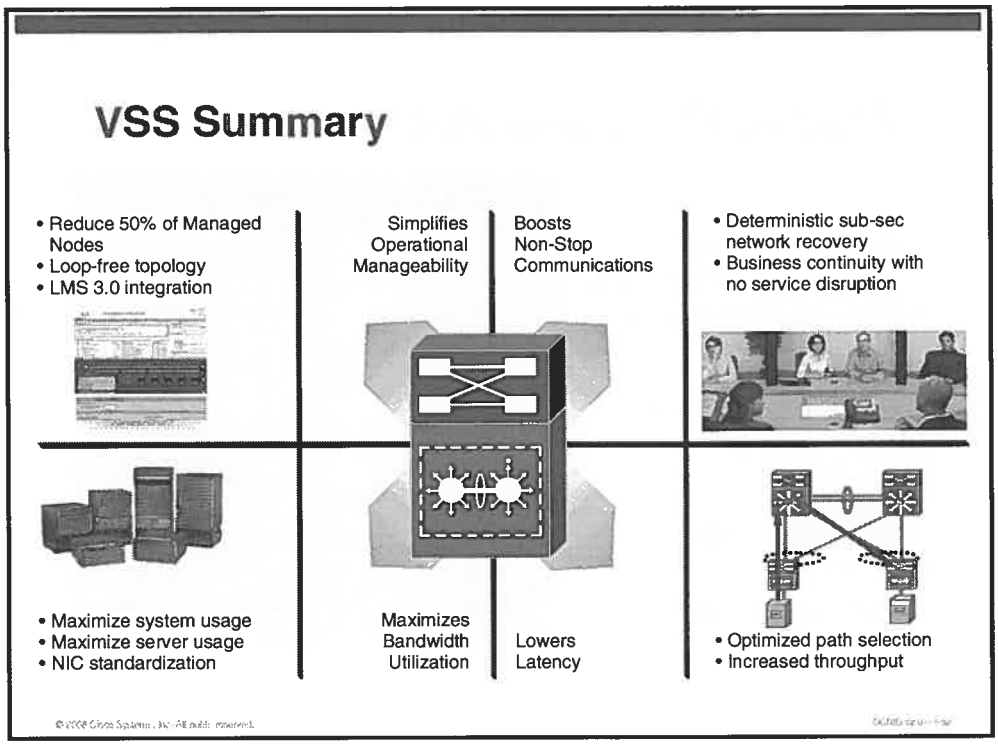
https://t.me/learningnets

vPC and VSS Comparison

	vPC (Virtual Port Channels)	VSS (Virtual Switching System)
Multi-Chassis PortChannel	Yes	Yes
Loop-Free Topology (no blocking ports)	Yes	Yes
STP as a "Fail-Safe" Protocol Only	Yes	Yes
Switch Control Plane	Two Independent Nodes, both active	Single Logical Node
Switch Redundancy (sup failover)	Intra-chassis	Inter-chassis
Control Plane Protocols	Instances per Node	Single instance
Switch Configuration	Separate Configs (w/ consistency checker)	Combined Configs
Maximum Physical Nodes	2	2
ISSU Support	Yes	12.2(33)SXI
Inter-Switch Link Hardware	32 Port 10GE Module	PFC3C mode, Sup 70 10G, 6708, 6716

Here is a comparison on the differences and the similarities between virtual PortChannel and virtual switching system, VSS. In terms of the switch control plane, the Nexus 7000 Series has two independent nodes, both of which are active. VSS is a single node look. A supervisor switchover will not cause an outage on the VS, and state is preserved. Redundancy, within the Nexus 7000 Series the failover of the supervisors is within the chassis, on the VSS the failover is between chassis, the result of which we end up with a separate config on each platform. The VSS is a single config. ISSU is supported straightaway with virtual port channel. We'll both have support for ISSU around the same time. Control plane runs in both devices Nexus 7000 Series switches, rather than one instance in VSS. Support for 2 supervisors in one chassis, in active-standby mode, and 1 virtual switch has an active-active model of data plane and control plane. 10GE module is required for bandwidth reasons.

https://t.me/learningnets



VSS increases operational efficiency by enabling multiple Catalyst 6500 Series switches to share a single point of management, single IP address, and single routing instance, as well as by eliminating the dependence on STP. It boosts non-stop communications by delivering deterministic, subsecond Layer 2 link recovery through inter-chassis stateful failovers and the predictable resilience of EtherChannel. VSS scales system bandwidth capacity to 1.4 Tbps by activating all available bandwidth across redundant Catalyst 6500 Series switches. None of the Cisco competitors have this feature.

VSS Key Benefits:

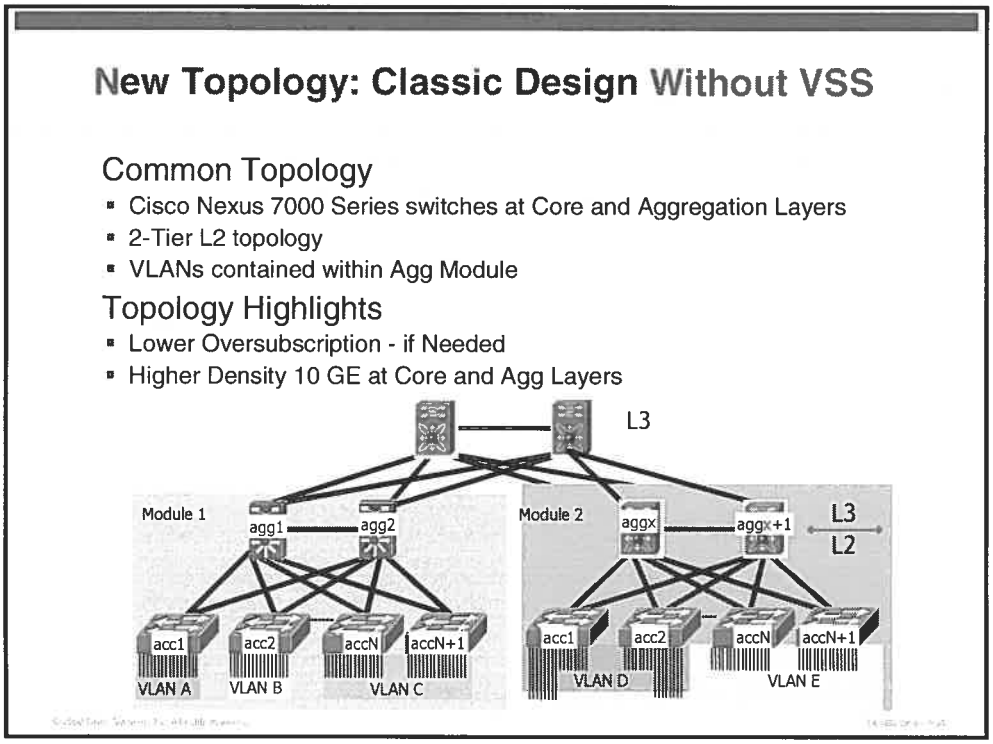
- Simplifies operational, because of the 50 percent reduction in configuration management and SNMP management
- Eliminates the usage of many protocols and simplifies the network design. You get non-stop communication with sub 200-millisecond failover times and the failover is stateful. Applications are not impacted when there is a failover from one distribution layer switch to another. The failover is predictable.
- When an uplink is lost connecting access layer to distribution layer only one link is lost. Spanning tree uplink fast is not used to recover this failure which gives a predictable failover and it also maximizes bandwidth utilization. Bandwidth utilization is improved between switches and from servers to switches. Optimal routing is acheived because all links are forwarding and you can route traffic through your network optimally

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Classic Design with VSS, VRF and VDC

This topic describes the classic data center design using VSS, VRF and VDC.



Catalyst 6500 Series switches (6509) in the Aggregation Layer are populated with 48-port linecards in all 7 switching slots, whereas the remaining two supervisor slots offer uplinks to the core layer. There are 336 server 1Gpps Ethernet ports available and 4 supervisor 10Gbps Ethernet uplinks. Nexus 7000 Series switch supervisor modules, on the other hand, do not offer uplink Ethernet ports; ports on the 10Gbps linecards should be provisioned instead.

The Nexus 7010 Series switch offers 8 payload slots. The 32-port 10Gbps Ethernet switching modules in 2 slots may be used for the redundant uplinks; each module connects to1 Core switch. On the server side, the Nexus 7010 Series switch is populated with the 48-port 1Gbps Ethernet ports, integrated into the other 6 modular slots of the Nexus 7010 chassis. Overall, the Nexus 7010 Series switch architecture will be less oversubscribed than the Catalyst 6500 Series.

The Nexus 7010 Series switch in this design has comparable high availability level, but much lower oversubscription than the Catalyst 6500 Series switch. Low oversubscription translates into more expensive solution, since more ports should be brought to connect the same servers to the fabric, but it is at the same time, a more reliable solution, with less potential or probability for triggering network congestion due to lower oversubscription. Nexus will welcome in the data center with 4:1 to 6:1 oversubscription ratio requirements of some low and mostly high end servers attached to the Nexus 1Gbps Ethernet ports. Due to it is huge performance and high availability with lower oversubscription the Nexus 7010 Series switch is a core and Aggregation Layer switch.

In contrast, the Nexus 7010 Series switch 32-port (with 1Gbps Ethernet ports) switching module has increased the access layer server ports throughput/density in comparison to the existing Catalyst 6500 Series 16-port switching modules.

https://t.me/learningnets

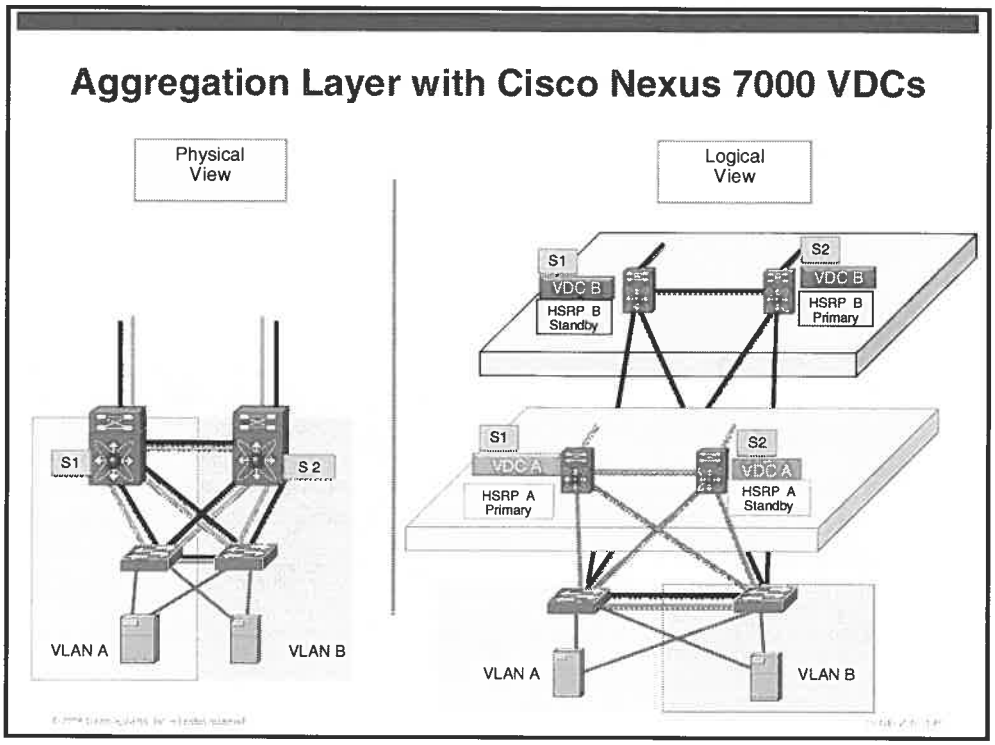
A 32-port Nexus 7010 Series line card offers the same port oversubscription ratio, 4:1, compared to the 16-port Catalyst 6500 Series line card, which enables more servers to connect to the switching module at the same 4:1 oversubscription ratio. This fact enables to double the Layer 2 domain size using Nexus 32 port linecard instead of Catalyst 6500 Series switch line card.

Note The Nexus 7010 Series switch has increased server throughput capacity on its switching modules, maintaining the same level of (switch port) oversubscription as before the respective 16-port line card in the Catalyst 6500 Series chassis. On the top-of-rack blade design, the number of servers is typically small, but the number can be doubled.

Note that a bigger Layer 2 domain means that a failure might impact a bigger number of servers in this domain.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



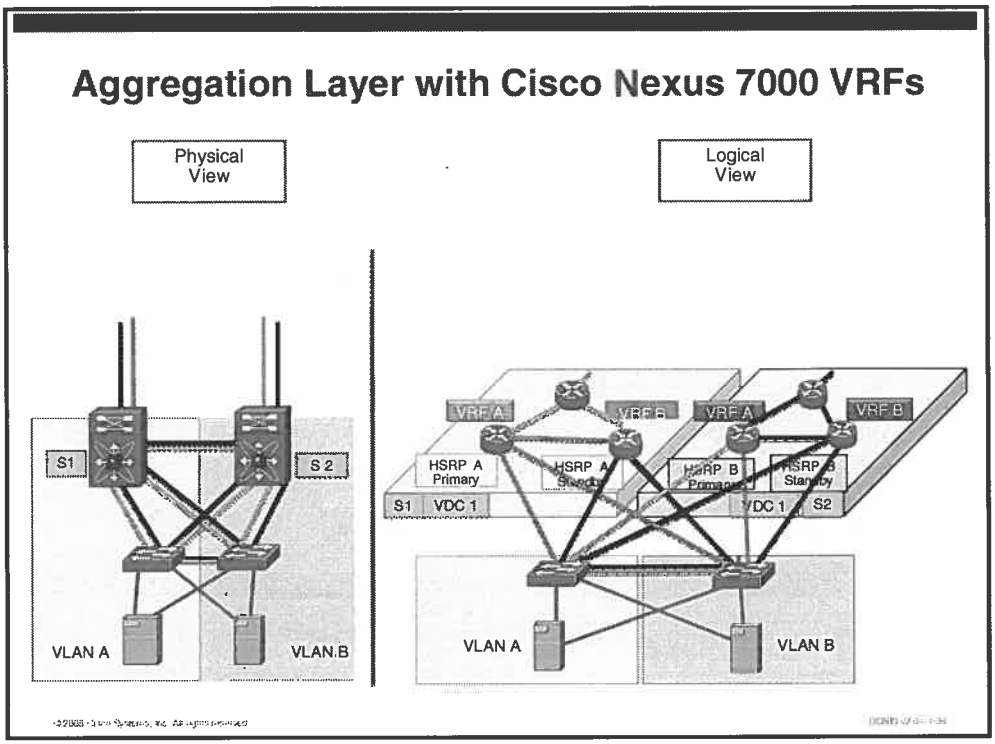
This deployment is appropriate even when communication between different server farm VLANs need to be deployed. VLANs can map to VDCs. Each VLAN has the default gateway on the respective VDC interface. One FHRP group per VLAN can be defined with primary interface on VDC A and secondary on remote VDC A.

An FHRP interconnection is needed for FHRP redundancy for each VDC pair (A and B). VDC A on the first physical switch can be active for VLAN A and standby for another VLAN C, and the VDC A on the remote switch can be active for VLAN C and standby for VLAN A.

Both pairs of VDC A and VDC B are connected with Layer 3 links in order to exchange routing information.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



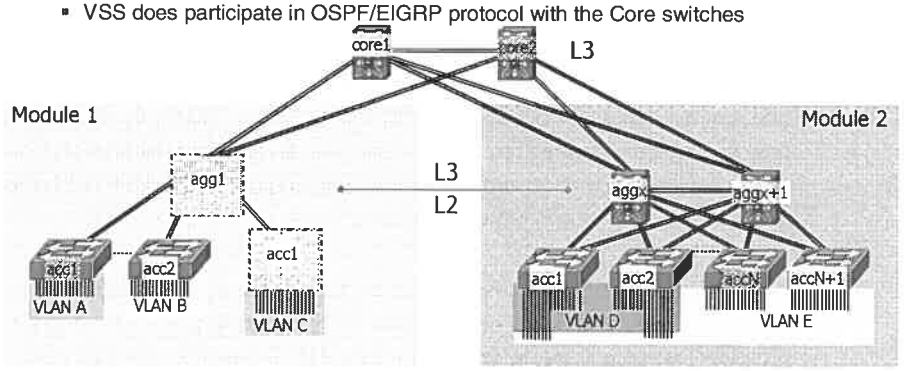
This deployment is appropriate even when communication between different server farm VLANs is needed, which is done by respective interconnecting VRF instances. Additional single (physical) VRF-interconnect between Nexus 7000 Series switches is needed for VRF A and for VRF B routing protocol exchange or FHRP group redundancy, if needed. Although VDC 1 on access switch one is called VDC 1 active and VDC 1 standby, the VRFs are both active and synchronizing routing database information.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Classic Design Using VSS in Access and Aggregation Layers

- Charactersitics:Virtual Switching and STP Stability
- Module 1: VSS in Access and Aggregation
- Module 2: triangular STP topology, VDCs for L2 failure domain segmentation
- VSS architecture has L2 interoperability (RSTP)with Cisco access switches
- Virtual PortChannel : MCE connects access VSS to Aggregation VSS
- Active-active forwarding for members of virtual PortChannel
- VSS does participate in OSPF/EIGRP protocol with the Core switches



Data center Catalyst 6500 VSS architecture supports Layer 2 interoperability with Cisco access switches, including the Catalyst 6500, Cisco Catalyst 4948 and Cisco Catalyst 4900-M Series switches using 802.1w Rapid Per VLAN Spanning Tree.

The preceding figure illustrates the topology that takes advantage of the Virtual Switching and STP stability features offered by both the Catalyst 6500 and Cisco Nexus 7000 Series Switches.

Module 1 shows virtual switches at the access and Aggregation Layers through the use of VSS technology.

Module 2 uses a classic triangular topology taking advantage of the advanced STP features, thus addressing STP stability in a high-density 10GbE Aggregation environment. Notice that module 2 could also reduce the size of the Layer 2 failure domain by using VS technology available through VDC, by creating virtual switches and assigning ports to each and allowing each logical topology to run isolated instances of control plane protocols.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

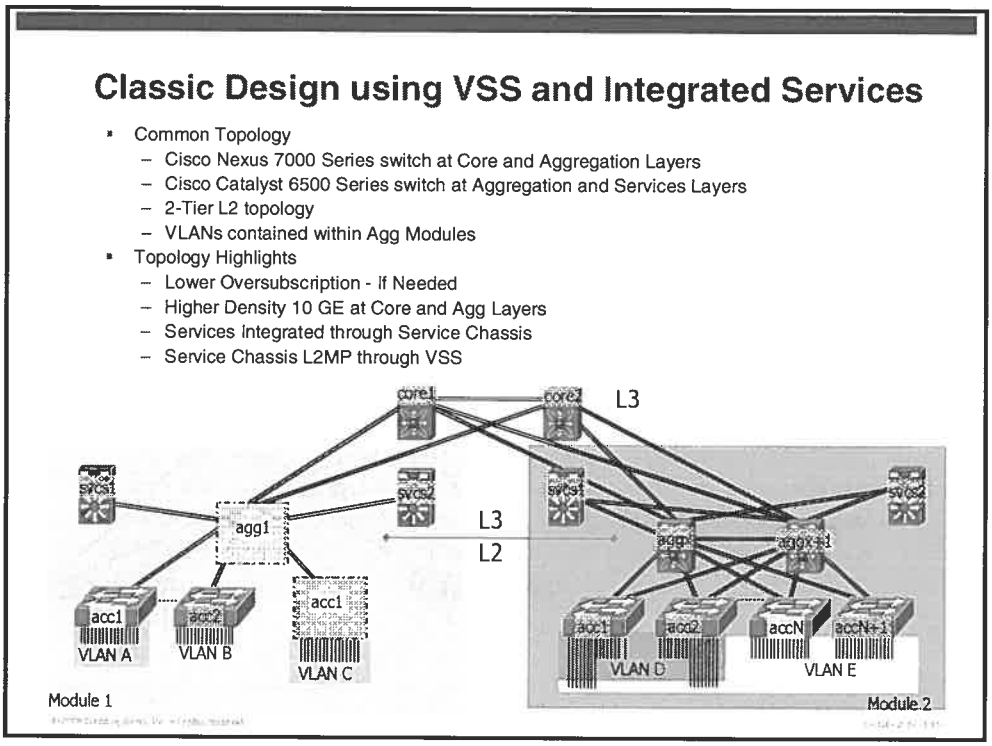
Note In a VSS environment, STP protocols are used strictly as fail-safe mechanisms as the loop-free forwarding topology is established through VSS. Using Catalyst VSS in the access and Aggregation Layer, the two physical switches act like one single switch with a single control plane. The virtual access switch 1 uses virtual PortChannel connection to the Aggregation switch 1. This design offers higher availability and survives a single switch failure without performance degradation when the access layer switch 1 is designed properly. The traffic that is imposed on the physical switch must not exceed the throughput or capacity of one single physical switch in the VSS system.

One of the advantage of the virtual switching and virtual PortChannel technology is that it allows a Layer 2 network to use all available Layer 2 paths actively and concurrently as opposed to STP blocking mechanism. Hosts and switches connecting to the virtual switches use any available L2 path concurrently and actively. The traffic distribution is granular to a flow (source and destination MAC addresses, source and destination IP addresses and Layer 4 port number). In addition to non-blocking paths, virtual PortChannels capabilities provide both switch and path redundancy.

The key feature that enables integration of VSS into the existing environment is the use of MCE for connection of the access layer to the Aggregation Layer, and the connection of the multi-NIC servers to the access layer. VSS offers Layer 3 interoperability with Cisco Catalyst 6500 core switches for both EIGRP and OSPF in a tiered data center design.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Nexus 7010 Series switches do not support services integration within the chassis; instead, service switches must be used as shown by the figure. The Service switch is another Catalyst 6500 Series switch populated with the service modules located on the side of a Nexus 7000 Series switch VSS system.

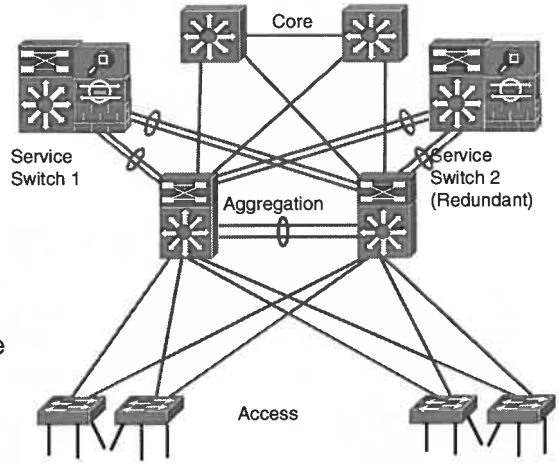
The second Module has Service switches connected to two aggregation Nexus 7000 Series switches. The L2/L3 boundary is kept the same as it is. The Nexus 7000 Series switches in this design deploy active-active VDCs per server farm and active-active or active-standby context firewall and load-balancing.

Module 1 offers a direct path between the active or the standby service module to the nearest Layer 3 router, thus eliminating STP reconvergence problems. This is only possible through a VSS system since the current implementation of Nexus 7000 Series does not support a VSS system. There are many application service modules that must be stable, and Layer 3 connectivity seems the best solution for service switch connection to the Aggregation switch.

https://t.me/learningnets

Service Layer Switches

- Move certain services out of Aggregation Layer
- Particularly works well for SLB modules
- Extend only necessary L2 VLANs to service switches via 802.1Q trunks
- Service layer can utilize EtherChannel uplinks
- Use QoS or separate GE links for fault-tolerance paths



The service layer switch provides a method of scaling up services using service modules without using slots in the Aggregation Layer switch. Moving certain service modules out of the Aggregation Layer switch increases the number of available slots and improves Aggregation Layer performance. For example, this is useful when a farm of ACE modules is required.

Typically, ACE modules are located in the Aggregation switch to provide services to servers across the access layer switches. By locating these service modules in a separate standalone switch that is connected using 802.1Q trunks, the Aggregation Layer can support a higher access layer uplink density and continue to operate in compact switching mode to optimize performance. This is particularly useful when 10 GigE port density requirements increase at the Aggregation Layer.

Mixing older line cards or service modules with the Cisco Supervisor Engine 720 integrated switch fabric architecture can limit the overall switching capacity and might not meet the performance requirements of the Aggregation Layer switches.

The ACE modules connect to the Catalyst 6500 Series bus using a 4 Gbps EtherChannel connection on the backplane. This 4 Gbps EtherChannel interface is used for all traffic entering and exiting the load balancer and uses hashing algorithms to distribute session load across it just as would an external physical EtherChannel connection. The CSM is also based on the classic bus architecture and depends on the Supervisor Engine 720 to switch packets in and out of its EtherChannel interface because it does not have a direct interface to the Sup720 integrated switch fabric.

If a single 4 Gbps EtherChannel does not provide enough bandwidth to meet design requirements, then multiple ACE modules can be used to scale, as required. Supporting multiple ACE modules does not create any particular problem but it does increase the number of slots used in the Aggregation switch, which might not be desirable. (?)Because the ACE is a bus-based, it must send truncated packet headers to the Sup720 PFC3 to determine packet destination on the backplane. When a single classic bus module exists in the switch, all non-DFC enabled line cards must perform truncated header lookup, which limits the overall system

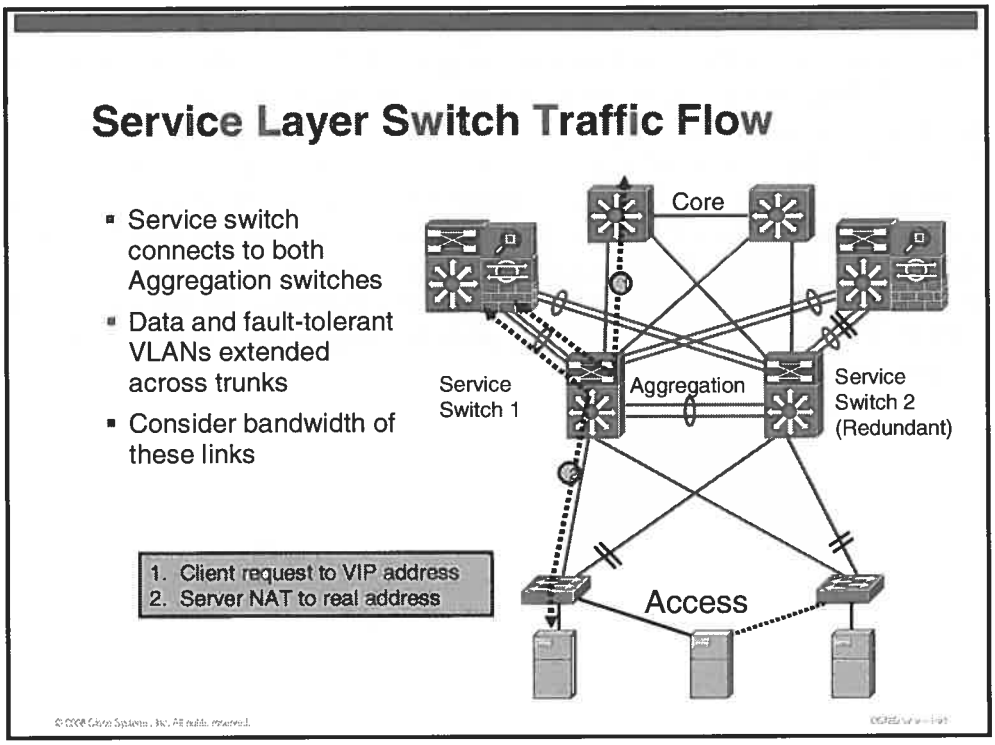
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

performance. The service switch can be any of the Catalyst 6500 Series platforms that use a Cisco Supervisor Engine 2 or Supervisor Engine 720 CPU module. The supervisor engine choice should consider sparing requirements, future migration to next generation modules, performance requirements, and uplink requirements to the Aggregation module. For example, if 10 GigE uplinks are planned, you must use the Supervisor Engine 720 to support the Cisco 6704 or Cisco 6708 GigE modules. A Supervisor Engine 32-10 GigE can also be used if only classic bus-enabled modules are being used. If an ACE in a service switch is configured for Route Health Injection (RHI), a Layer 3 configuration to the Aggregation Layer switch is necessary, because RHI knows how to insert a host route into only the routing table of the local MSFC. A Layer 3 link permits a routing protocol to redistribute the host route to the Aggregation Layer.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The service switch is connected to both Aggregation switches with Gigabit EtherChannel or 10 GigEtherChannel links configured as 802.1Q trunks. From a logical perspective, this can be viewed as extending the service module VLANs across an 802.1Q trunk.

The figure shows the flow of a session using an ACE module in one-arm mode on a service layer switch. The VLANs used in supporting the ACE configuration are extended across 802.1Q trunks from the Aggregation Layer switch to each service switch. The Fault Tolerant VLANs are also extended across these trunks. Spanning tree blocks only a single trunk from the secondary service switch to the Aggregation 2 switch. This configuration provides a forwarding path to the primary service switch from both Aggregation switches.

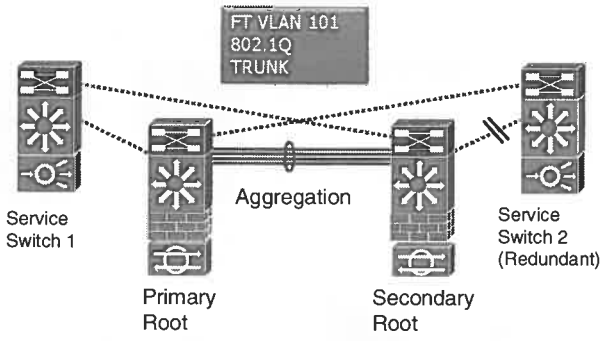
The bandwidth used to connect service switches should be carefully considered. The ACE is capable of supporting several Gbps of server and Fault Tolerant (replication and state) traffic on its bus interface. The switch uplink should be enough to avoid congestion on the Fault Tolerant path to ensure that ACEs do not go into an active-active state. A separate link containing the Fault Tolerant VLANs can be used to isolate the data and Fault Tolerant paths.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Redundancy with Service Layer Switch Design

- Deploy service switches in pairs
- Extend fault-tolerant VLAN across all link to aggregation switches
- FT for sesion state synchronization beetween the Service switches
- Do not use Service switches across more than one Aggregation module



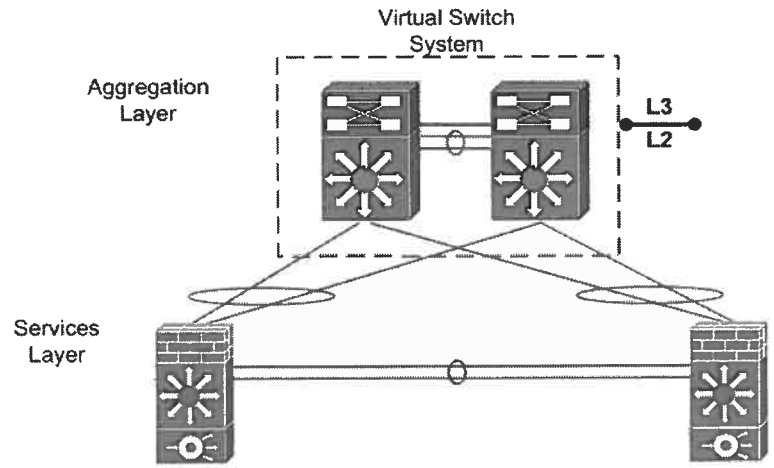
The service module is provisioned to participate in Spanning Tree and automatically elects the paths to the primary root on Aggregation 1 for the server VLAN. If Service Switch 1 fails, service module Fault Tolerant communication times out and Service switch 2 becomes primary.

The service switch layer should be deployed in pairs to support a fully redundant configuration. This permits primary-active service modules to be located in one chassis and backup-standby in a second chassis. Both service switches should be redundantly connected to each Aggregation switch to eliminate any single point of failure. In the figure, Service Switch 1 is configured with the active service modules while Service Switch 2 is used for standby service modules. If Service Switch 1 fails, Service Switch 2 becomes active and provides stateful failover for the existing connections. The Fault Tolerant VLAN is used to maintain session state between service modules. The FT VLAN must be extended between service switches across each 802.1Q trunk that connects to the Aggregation Layer. The Fault Tolerant VLAN is not passed across the link connected between the Aggregation switch pair. The service module is provisioned to participate in spanning tree and automatically elects the paths to the primary root on Aggregation 1 for the server VLAN. If Service Switch 1 fails, service module Fault Tolerant communication times out and Service Switch 2 becomes primary.

Note Service layer switches should not be used across more than one aggregation module. This is required to keep the spanning tree domains isolated between aggregation modules; otherwise, root switches in one aggregation module could become root for another, creating undesirable path selection and link blocking.

https://t.me/learningnets

Aggregation Layer Classic Design Using VSS with Service Modules (Physical View)



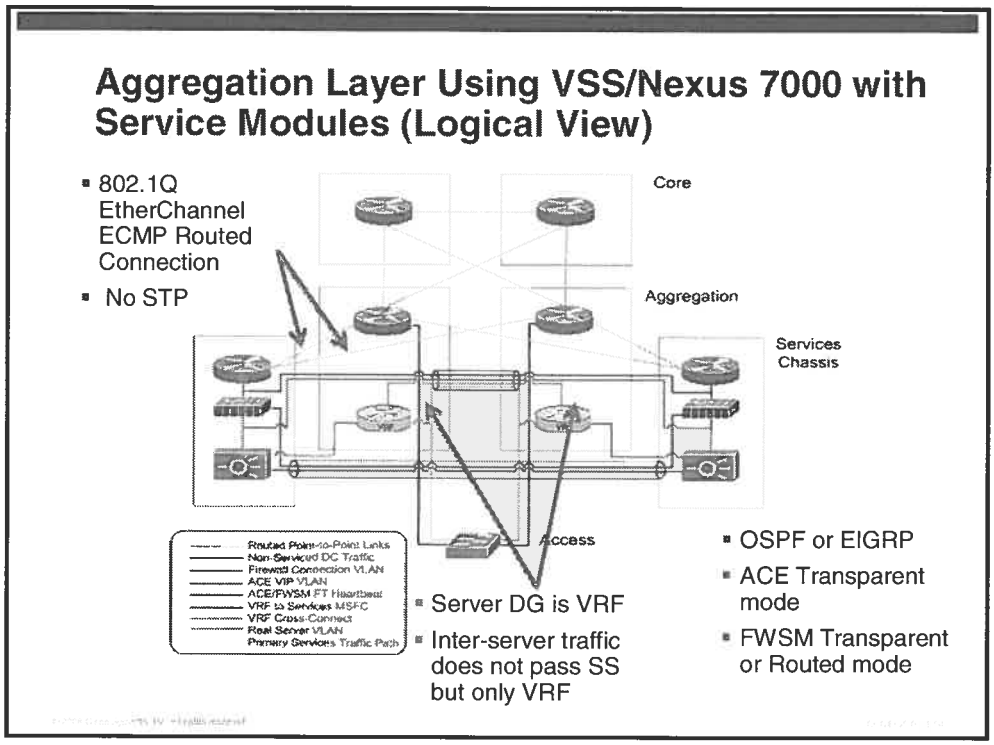
Catalyst 6500 Series switch with the Cisco Supervisor Engine 720-10GE module is needed to deploy the VSS system. Virtual Switch Link is connecting two physical switch members of the VSS system. A new PFC called the PFC3C is required for VSS switching. There are many possible variations of logical topologies for the integration of the Services chassis into the VSS-based Aggregation Layer. The Services Layer switches act like external service modules for the VSS Aggregation Layer system. The pros and cons of each design decision vary depending on customer traffic patterns and requirements. Some of the major design decisions include:

- Whether to enable the MSFC to perform IP Routing within the services chassis
- Whether to configure the service modules in transparent or routed mode, or a combination of the two modes for different modules in the service chain for those that support combination
- How to handle inter-server traffic within subnets and between subnets

The preceding figure shows one of the optional Aggregation Layer deployments with the VSS system with a flexible Services chassis model that takes advantage of the routing capabilities of the Cisco switching platforms to enable simple deployment of new services.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure shows the Services chassis logical topology that leverage routing in the chassis itself and a Virtual Routing and Forwarding (VRF) instance in the Aggregation Layer to allow direct routing of traffic between server subnets as needed. The FWSM can be deployed in both transparent and routed modes; the routed mode uses OSPF which is currently the only IGP supported by this module. The ACE is deployed in transparent mode only, since it does not support a routing protocol, and in routing mode would isolate the VRF from peering with the rest of the IGP in the data center.

Note EIGRP can also be used with both FWSM and ACE in transparent mode.

Attributes of this Services Chassis architecture include:

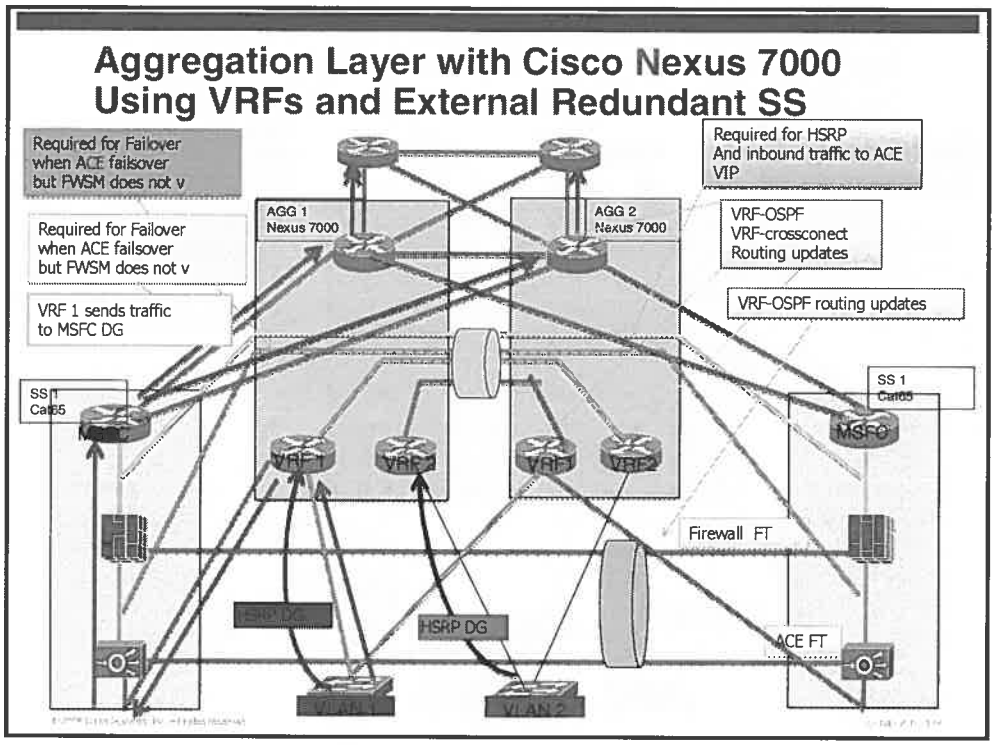
- It is applicable to traditional, VSS, and Nexus 7000 Series-based Aggregation Layers.
- Each Services Chassis is physically Dual-Homed into the Aggregation Layer with EtherChannel 802.1Q Trunks.
- The routed path into the Services Chassis eliminates STP from primary ingress traffic path, provides ECMP load balancing in and out of the Services Chassis, and permits RHI for ACE deployment. The Server default gateway resides on VRF instance on the Aggregation Layer.
- Services may be added into or removed from the chain while the IP routing destinations for inbound client traffic and server sourced traffic remains consistent since they are based on MSFC primary or VRF instances.
- Inter-server traffic may be routed by the VRF instance without passing through the Services Chassis.
- Traffic not requiring services may bypass the Services Chassis using separate VLANs for direct traffic.

https://t.me/learningnets

- Several variations on this architecture may be implemented but are not be included in the initial baseline configuration, including:
- If inter-server traffic must pass through a firewall, a separate transparent context may be added to the FWSM, or a separate physical firewall may be deployed as dictated by customer-specific bandwidth requirements and optimal traffic patterns.
- Multicast traffic may be serviced with all modules in transparent mode, by configuring PIM on the FWSM, or by using a direct non-serviced path.
- This design may be adapted to customer-specific requirements by removal of ACE, FWSM, or VRF instance from the traffic flow as needed. Proper addressing changes will be required and primary versus backup traffic paths should be considered.
- Longer term, this architecture also lends itself well to creating a VSS-based Services Chassis block, once the Whitney 2 code is available to support Catalyst 6500 Series services modules in a VSS-enabled switch.
- VLANs extended between services module layers could be migrated off of the Aggregation Layer onto the VSS EtherChannel between chassis.

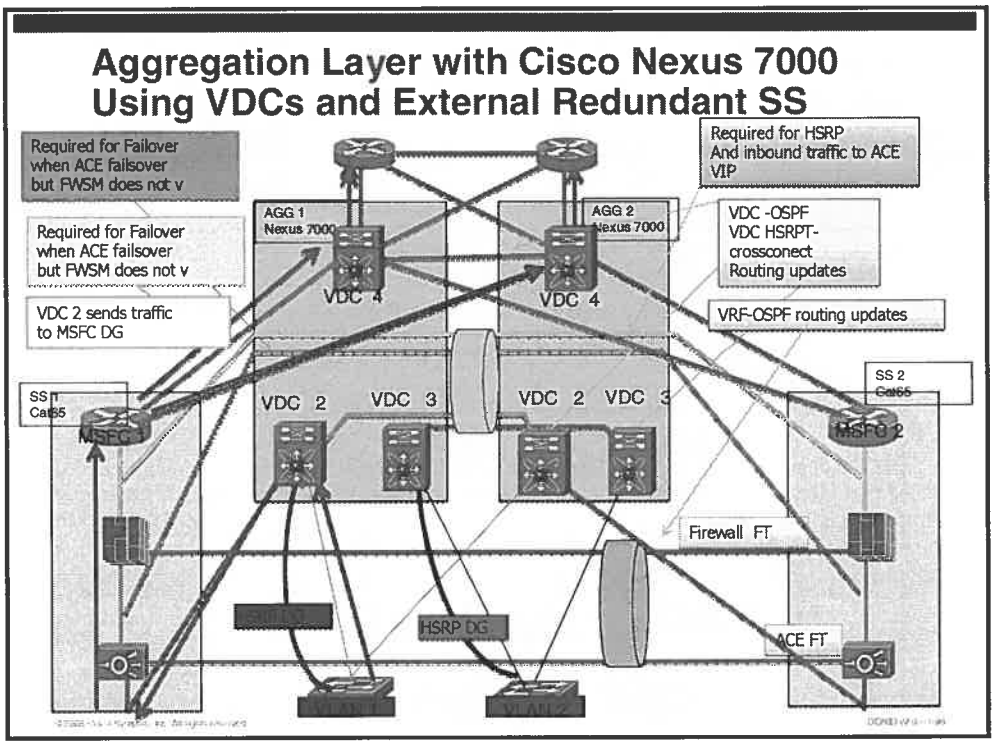
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Per VLAN traffic separation and network scalability can be achieved by deploying a dedicated VRF, a Service switch or a service line of the ACE and Firewall context, a dedicated HSRP group and a dedicated OSPF/EIGRP routing instance for each VLAN.

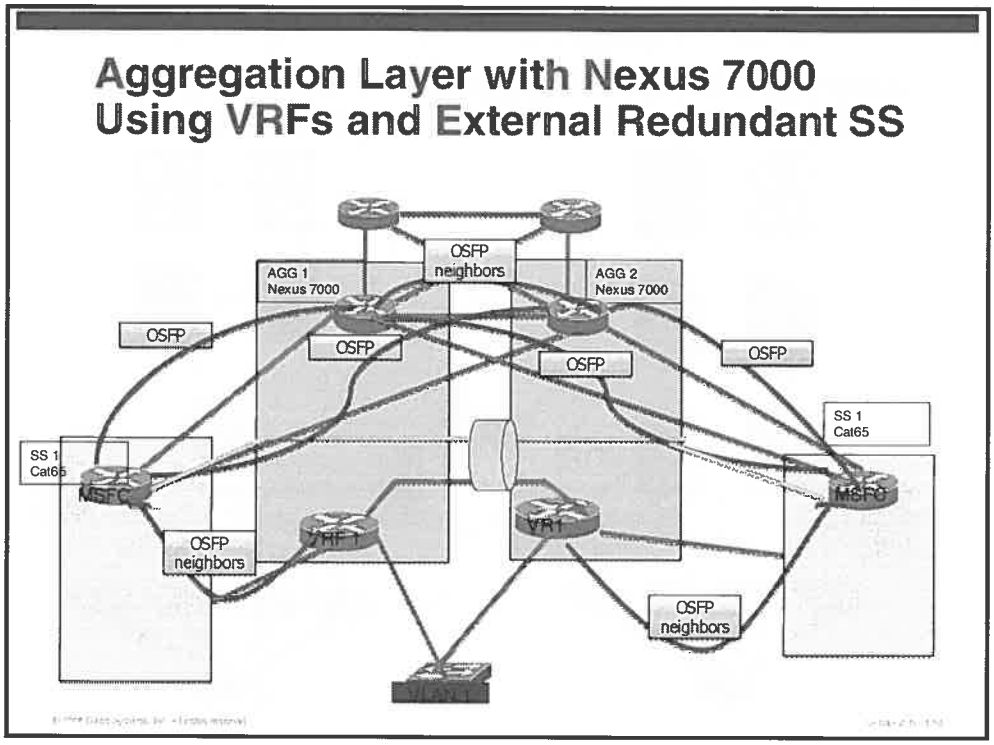
https://t.me/learningnets



There can be one VDC (2) for each VLAN (VLAN 1) in the Aggregation Layer and the single VDC 4 for traffic forwarded from all VLANs to the core. A separate VDC can be deployed only for traffic that is not sent to Service switches but passes directly through a single dedicated VDC to the core. Currently this design is not scalable since only 4 VDCs can be enabled on the Nexus 7010 Series switch.

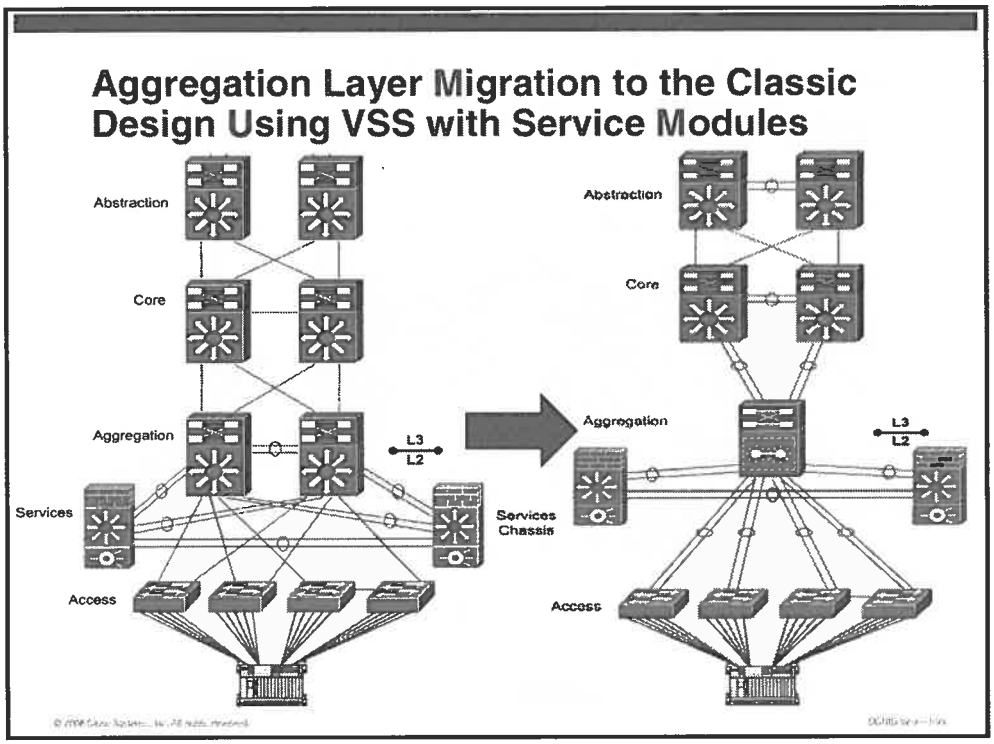
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure shows possible OSPF neighborships established among the routers in this topology. The VRF router might establish OSPF neighborship with one or both MSFCs in SS switches.

https://t.me/learningnets



Layer 2 Design Characteristics

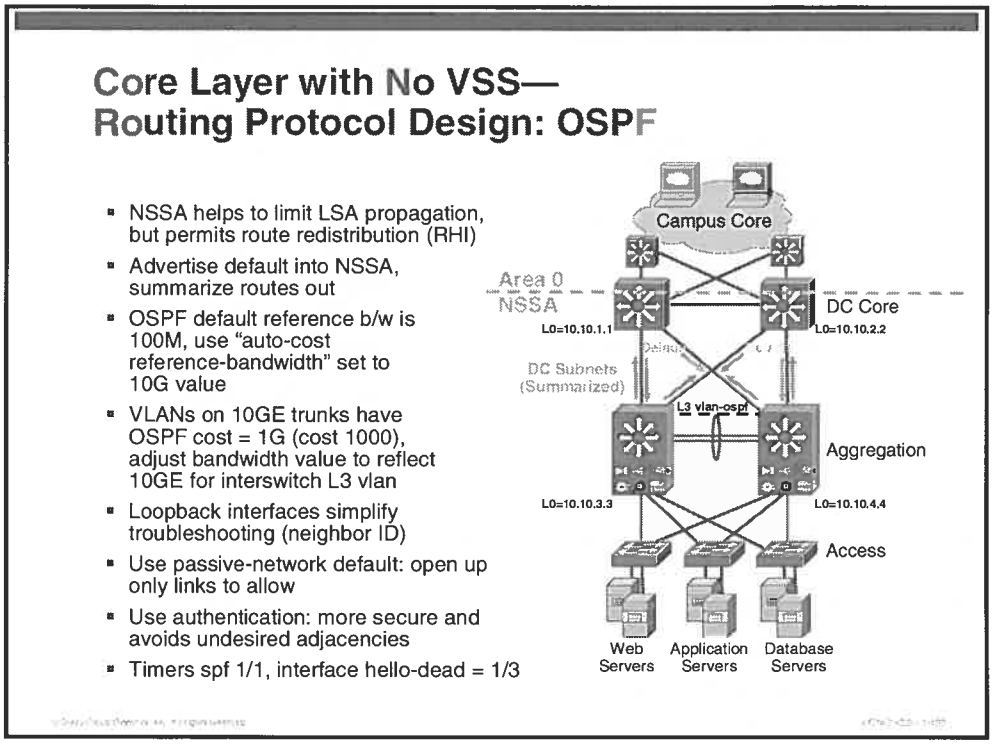
- Spanning Tree Root – map L3 path to L3 HSRP primary path using the`*spanning-tree vlan <vlan list> root [primary|secondary]*` configuration command for specified server VLANs.
- Enable Rapid Spanning Tree using `*spanning-tree mode rapid-pvst*` configuration command.
- All access layer nodes carry VLAN definitions for the purposes of STP.
- Pruning is enabled for the unused VLANs off of the trunks to ensure that the L2 logical VLAN paths are carefully controlled.
- Loop Guard enabled on the inter-switch links for Aggregation and Services switches.
- Root Guard enabled on ports facing the access layer.
- UDLD enabled globally.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Routing Design

This topic describes the routing model that can be applied to the data center design.



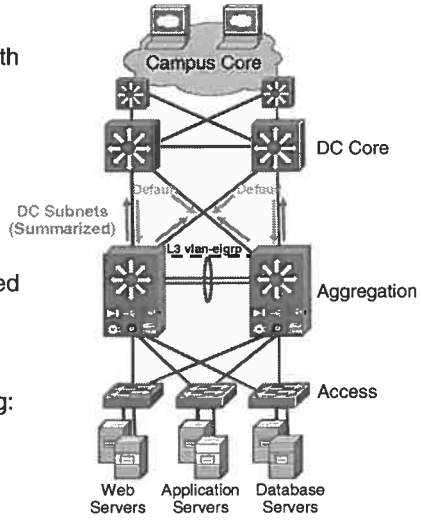
The non-VSS data center Aggregation and Core Layer can be applied OSFP routing protocol as shown by the figure:

- NSSA helps to limit LSA propagation, but permits route redistribution (RHI).
- Advertise default into NSSA, summarize routes out.
- OSPF default reference b/w is 100M, use "auto-cost reference-bandwidth" set to 10G value.
- VLANs on 10GE trunks have OSPF cost = 1G (cost 1000), adjust bandwidth value to reflect 10GE for interswitch L3 vlan.
- Loopback interfaces simplify troubleshooting (neighbor ID).
- Use passive-network default: open up only links to allow.
- Use authentication: more secure and avoids undesired adjacencies.
- Timers spf 1/1, interface hello-dead = 1/3sec.

https://t.me/learningnets

Core Layer with No VSS—Routing Protocol Design: EIGRP

- Advertise default into data center with interface command on core:
  - `ip summary-address eigrp 20 0.0.0.0 0.0.0.0 200`
  - Cost of 200 required to be preferred route over the NULL0 route installed by EIGRP
- If other default routes exist (from internet edge for example), may need to use distribute lists to filter out
- Use passive-network default
- Summarize data center subnets to core with interface command on agg:
  - `ip summary-address eigrp 20 10.20.0.0 255.255.0.0`



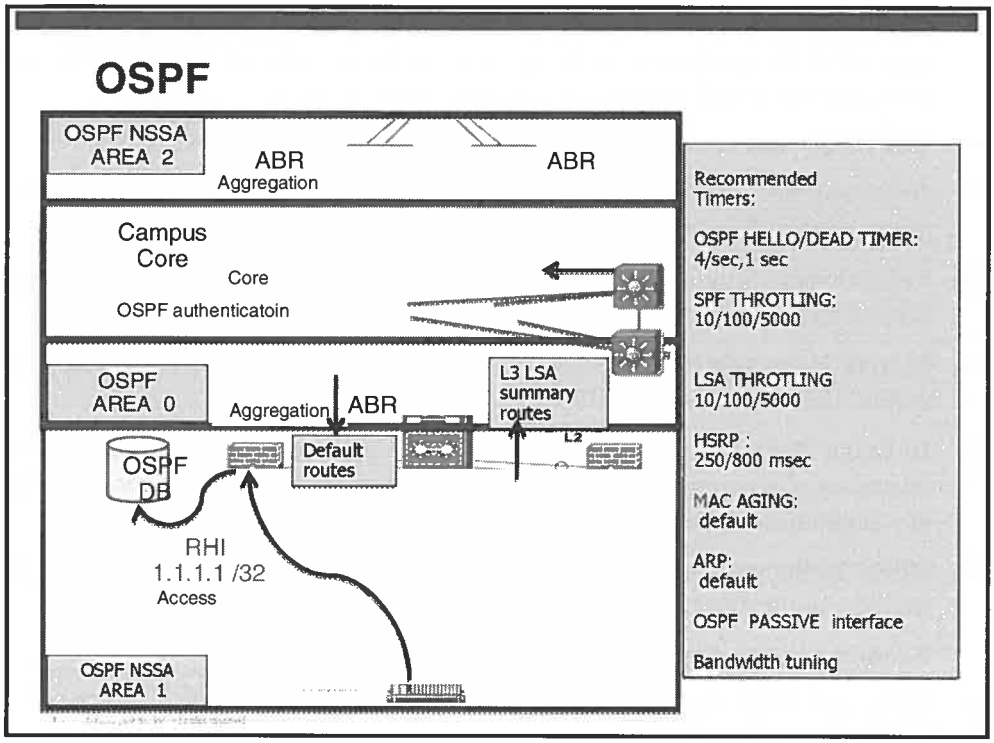
The non-VSS data center Aggregation and Core Layer can be applied EIGRP routing protocol as shown by the figure:

- Advertise default into data center with interface command on core:
  ip summary-address eigrp 20 0.0.0.0 0.0.0.0 200.

  Cost of 200 required to be preferred route over the NULL0 route installed by EIGRP.
- If other default routes exist (from internet edge for example), may need to use distribute lists to filter out.
- Use passive-network default.
- Summarize data center subnets to core with interface command on Aggregation switch:
  ip summary-address eigrp 20 10.20.0.0 255.255.0.0.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Layer 3 Design Characteristics

The preceding figure shows design details on using OSPF in the data center with an ABR router on the Aggregation Layer data center switch.

- **Layer 2/3:** Boundary can be maintained at the Aggregation Layer.
- **Routing:** Can be deployed in the Services chassis, but the physical link will still be a trunk to accommodate ingress and egress chassis traffic.
- **IP Addressing**: IP addresses enable summarization at the Aggregation Layer.
- **ECMP Load Balancing**: Catalyst 6500 Series switches configured globally for mls ip cef load-sharing.
- **Routing table population**: The IP routing tables can be injected with external routes from EIGRP/OSPF peering with the abstraction layer router.
- **Routing Protocols:** Such as EIGRP and OSPF variants can be deployed.
- For OSPF or EIGRP: Use passive interface default, and only un-passive interfaces that will run the protocol on.
- **OSPF Router ID:** Originated and defined by loopback interfaces.
- **OSPF Areas**: Links between the Aggregation and core, and between core switches are in OSPF Area 0.
- **Links within an Aggregation and Services Layer block:** Are in an OSPF area 1 corresponding to the Aggregation block. Aggregation block areas should be configured as NSSA to minimize LSA propagation, while permitting the use of RHI.
- **OSPF Authentication**: OSPF Authentication should be enabled to verify interoperability between VSS enabled IOS and IOS, with the respective authentication key.

https://t.me/learningnets

- **OSPF Hello/Dead Timers**: OSPF timers set for best practices optimizations. Fast Hellos are used at the interface level using the ip ospf dead-interval minimal hello-multiplier 4 command for a dead interval of 1 second, and 4 hellos per second.
- **OSPF SPF and LSA throttling:** Used according to best practices.
- **OSPF router:** spf and lsa throttling timers can be set to the following: spf 10 100 5000 and timers throttle lsa all 10 100 5000.
- **EIGRP Stub**: Stub routing should be set so that the core knows that the Aggregation block is not a transit path to other subnets to minimize queries on reconvergence.
- **EIGRP Hello and Hold Timers**: Set to 1 and 3 with ip hello-interval eigrp <router no> 1, ip hold-time eigrp <router no> 3.
- **IP Event Dampening**: Should be enabled using the dampening command on all routed interfaces. This prevents interface state changes from continuously propagating in the event of a flapping condition.
- **HSRP Hello and Dead Intervals**: Set to 250 and 800 for rapid transition of the default gateway in the event of losing the access layer link.

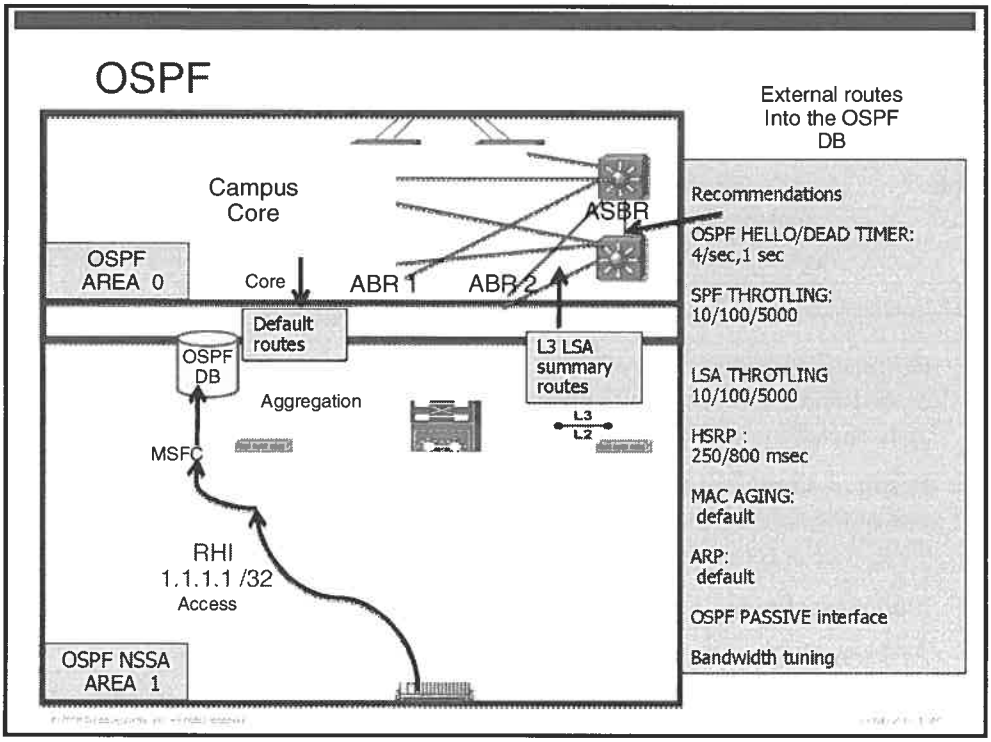
  Example: `standby 1 timers msec 250 msec 800`
- **MAC and ARP:** Aging timers used may be defaults.
- **Graceful restart and FHRP impact:** Should be minimized on data center traffic.

Layer 2 Design Characteristics

- **Spanning Tree Root**: map L3 path to L3 HSRP primary path using spanning-tree vlan <vlan list> root.
- **[primary|secondary]:** for specified server VLANs.
- **Enable Rapid Spanning Tree**: Spanning-tree mode rapid-pvst.
- **Added STP Control Plane Load**: Not shown in the physical and logical topology drawings is the intent to add STP control plan load to the system by including 16 additional Access Layer switches in the Aggregation Block under test. These switches will be low-end, 3548 or equivalent, and only 100 Mbps attached to the Aggregation Layer. This provides additional STP control plane load, without a great amount of additional space and equipment cost. All Access Layer nodes under test will carry VLAN definitions for purposes of STP.
- **Prune:** Unused VLANs off of trunks to ensure that the L2 logical VLAN paths are carefully controlled.
- **Enable Loop Guard:** On inter-switch links for Aggregation and Services switches.
- **Enable Root Guard:** On ports facing the Access Layer.
- **Enable UDLD:** Globally.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Layer 3 Design Characteristics

Design Details on using OSPF in the data center with ABR router on the core layer data center switch:

- The core switches belong to area 0 and they are also the ABRs for the Area 1.
  - The data center belongs to a separate area 1. This area can be totally stubby, or stub area, or NSSA. (Totally Stubby Area will not receive LSA type 3 but only default routes.)
- Connections to the Core Layer are Layer 3 links.
- Both ABR routers are injecting the summarized route information from area 1 into the OSPF area 0.
- When the load-balancer is deployed in bridged mode, it injects RHI server /32 host routes into the MSFC's OSFP routing table. The L3 LSA inject summarized routes from the NSSA/Stub area to Campus core switches. This minimizes network instability within Area 1 in the event of topology change or network failure.
- The data center Area 1 switches are receivers of default route topology link state information, injected by the ABR routers from the Core into the Aggregation area pointing to the ABR routers. The reason, again, is to keep the area 1 isolated from Core networks flapping, routing reconvergence and SPF recalculations as a result of Core topology instabilities.
- We can eliminate (R)STP on the links between the Aggregation and Core network and let routing protocols prevent forwarding loops.

Note By default, NSSA ABR will not originate a summary default as do stub ABRs and totally stubby ABRs. However, ABRs can be configured to inject a default route into the data center.

https://t.me/learningnets

- NSSA limits the scope of LSA flooding and thus SPF recalculations, and enables configuration of static Route redistribution. That is welcome when dealing with Firewalls, Load Balancers and with Internet Edge Routing.

Note LSA flooding triggers SPF calculations on every router and during the calculation the traffic is routed based on the old routing topology. The SPF uses a lot of switch CPU and lasts approximately 600 msec, depending on the complexity of the OSPF topology, number of OSPF routing nodes and number of interconnecting networks.

- Origination and injection of summary routes from one area into the other area by the ABR or ASBR routers limits the impact of SPF instability or OSPF reconvergence within the area caused by topology changes such and prevents route flapping.
- Result of summarization is that when a network topology failure occurs that affects only one of the subnets that the ABR summarizes into the next area, the ABR will not generate a Type-3 LSA for such a failure.
- Stub areas, for example, are isolated and not affected by other areas' SPF calculations.
- ABR does not originate an LSA Type 3 into the totally stubby area.

Passive Interface

- OSPF passive interface will not establish OSPF neighborship with other OSPF routers on the passive interface network. This feature is welcome in the data center design.
- When ABR routers share a common network on many VLAN interface in the OSPF area 1 they do not have to establish a neighbor relationship per every VLAN. Instead, passive OSPF subinterfaces can be configured on the Trunk connection between the ABR routes. An SVI interface can be used instead of a physical or subinterface for the ABR neighbor.
- Both ABRs send LSA information into the core network 10.10.0.0/16 that summarizes all VLAN subnet networks (10.10.x.0/24). When ABR 1 looses connection to the VLAN 1, then ABR will still receive the traffic destined to VLAN 1 subnet but must redirect it to ABR 2. If ABR 1 and ABR 2 are not neighbors, this is not possible. An additional Layer 3 link should connect them (for this purpose only). Then, ABR 2 will announce subnet to ABR 1 and ABR 1 will redirect traffic to ABR 1.

Note In order to for the OSPF SPF to distinguish the 1Gbps and 10Gbps links form the Fast Ethernet links, the referenced bandwidth must be set to 100. By default, the auto-cost reference-bandwidth is set to 10.

© 2008 Cisco Systems, Inc.
https://t.me/learningnets

OSPF Convergence

OSPF convergence time with default timers is 6 seconds on average:

- The convergence time for the link or router failure is given by the time required to do the following:
  - Detect link failure
  - LSA information propagation
  - Waiting for SPF calculation to be finished
    - Router calculates the SPF in 5 sec delay after receiving LSA
    - Configurable with timers spf delay holdtime
  - Running SPF alghorithm a few hundred miliseconds
  - Update the routing table

OSPF Convergence time can be brought down to a few seconds using the timers spf 1 5

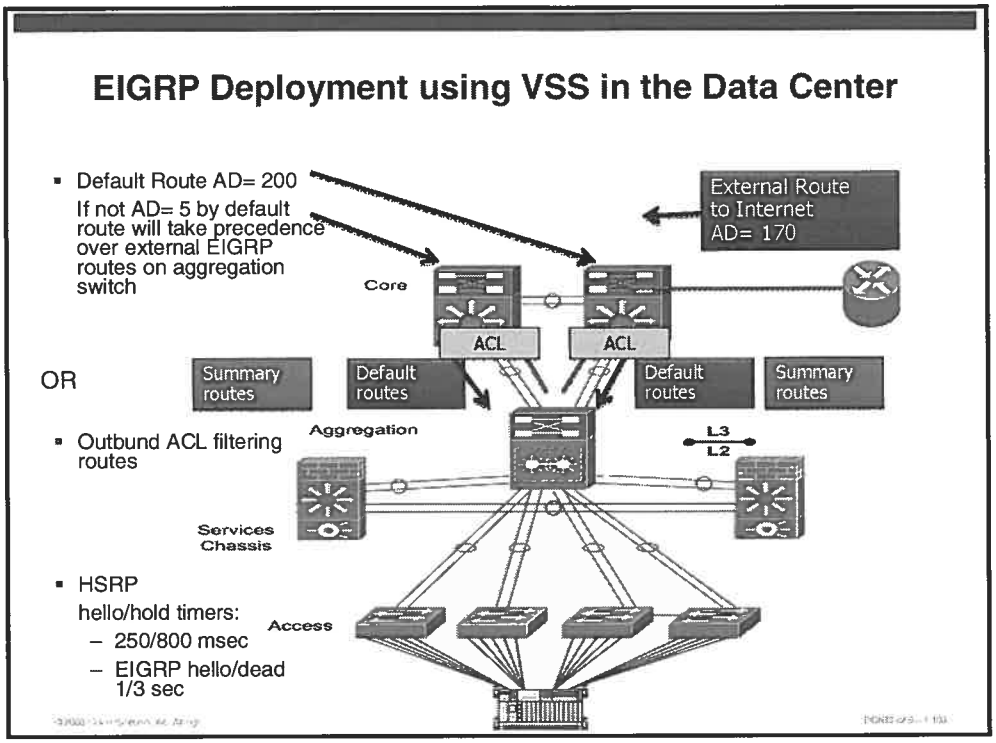
- Reducing the spf delay/holdtime may cause permanent SPF recalculation upon route flapping. Holdtime between 2 consequtive SPF calculation is 10 sec by default.

OSPF convergence time with default timers is 6 seconds for an average topology. These steps can aid in reducing OSPF convergence time:

- Reduce the spf delay holdtime
- Enable incremental SPF to avoid recomputing the entire SPF
- Run SPF algorithm a few hundred milliseconds

Note The use of SPF timers is recommended only in an environment using a well-structured area, route summarization design, and link flap damping features.

https://t.me/learningnets



EIGRP has faster convergence than other distance-vector protocols, due to the hybrid nature of its architecture. The goal of using EIGRP in the data center is to minimize the convergence time upon the topology change in a stable data center network. The following design describes one of the EIGRP deployment options in a data center network:

- Connections between the Aggregation routers and the Core are Layer 3 links.
- The core routers advertise default routes to the data center using distribute lists.

Note Default routes can be injected using `***ip summary-address eigrp***` or `***distribute list interface***` configuration command.

If summary-address is used, the AD cost must be set to 200 or the default routes advertised by the core to the data center. There are external EIGRP default routes with external administrative cost 170 which must be prioritized over the one injected by the core into the data center. Route summarization is used per interface from the data center to the core switches (using ip summary-address eigrp) Inside the data center the network statement must include all the subnets.

Passive interface can be used to reduce the number of neighbors that each Aggregation router needs to maintain and rather keep a separate Layer 3 link or Layer 3 SVI interface for the neighborship purpose.

EIGRP Route summarization from the Aggregation to the Core Layer will limit the scope of EIGRP queries upon topology change. If topology changes occur, the query reaches the core, by asking for an existing path to certain destination network. The fact that core routers have no information about particular subnets but keep only more general summarized information will stop the query process from being propagated further through the network. The Core will report the infinite destination to the network, thus accelerating the completion of the DUAL computation.

© 2008 Cisco Systems, Inc.

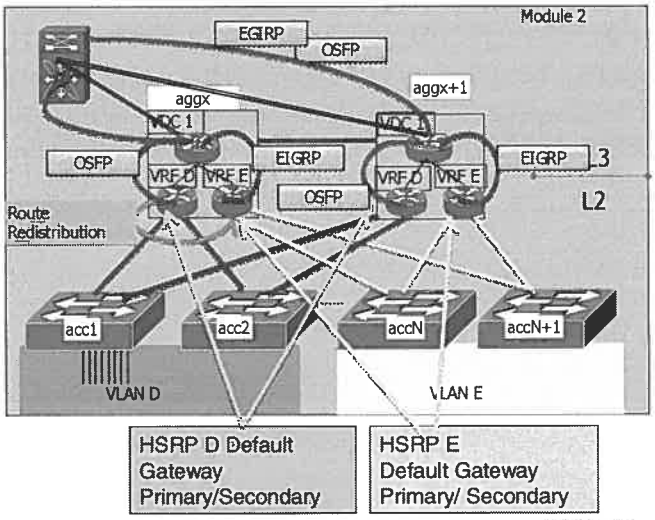
https://t.me/learningnets

When summary routes are configured, EIGRP automatically installs a NULL route into the routing table. So if a summary 0.0.0.0 0.0.0.0 is configured without administrative cost, EIGRP will install a NULL route for the 0.0.0.0 0.0.0.0 route cost 5. The Null route would then take precedence over the external EIGRP route. As a result, any traffic that reaches the core and does not match any route would be blackholed instead of getting pushed to the edge. An alternative is to filter the routes from the core to the data center in such a way that only default routes (injected by the edge routers) will be advertised to the data center. (Use ACL and distribute list.) Both options are valid but different upon the router being powered up. In the first case, the powered up router would immediately advertise the default route without having complete routing table information. In the second case, the router needs the routing information to reach the edge routers before advertising the default route.

https://t.me/learningnets

High Availability with Cisco Nexus 7000 in the Aggregation Layer Using VRFs

- Server DG is HSRP group per VRF pair
- HSRP group per VRF
- Inter-VRF route redistribution using BGP & VRF route import/export
- VRF to VLAN mapping for the ports to access layer
- Incoming traffic is assigned to VRFs based on destination subnet or VLAN/VRF subinterface
- Each VRF runs one OSPF routing process or EIGRP routing process



This variation of the Nexus 7000 Series switch deployment in the Aggregation Layer is therefore less prone to failures, less isolated, less virtualized and less stable than the previous topology, but more scalable since VLANs are mapped to VRFs within the same VDC.

The top router in VDC 1 allows route redistribution between the two VRF instances, but it is a design option, as is also the use of an external router to do so.

The links between the two Aggregation switches are used for HSRP group D and group E keepalives and for VRF D and VRF E interconnection.

VRF D/E is the default gateway for the server farm VLAN D/E in this scenario.

Each VRF is running a separate routing protocol to isolate routing instabilities per server farm.

Inter-VRF route redistribution is done with BGP, and VRF route import/export feature.

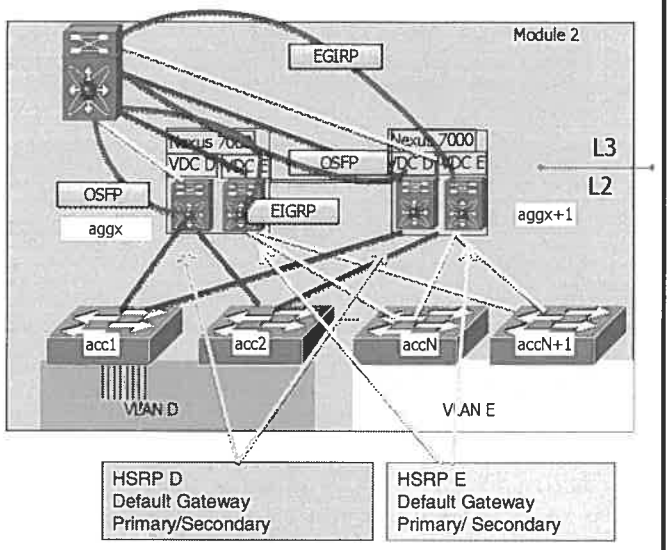
The alternative is to dedicate a separate interface per each VRF from the Aggregation to the Core Layer, escaping the common router to both VRFs shown on the top and another routing hop in the Aggregation Layer switch.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

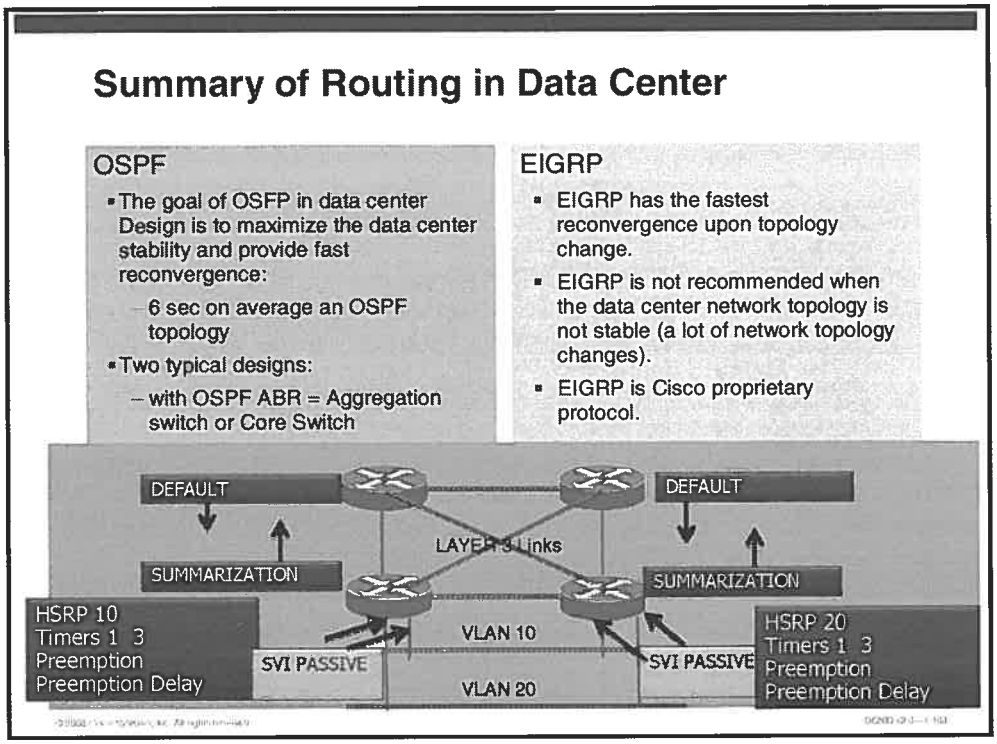
Cisco Nexus 7000 in the Aggregation Layer Using VDCs

- HSRP group per VDC pair.
- Each physical interface on agg switch is assigned to respective VDC.
- Incoming traffic is assigned to VRFs based on destination subnet or VLAN/VRF subinterface.
- VRF represents one OSPF router that establishes neighborship with VDC default routing instance.
- Inter-VDC route redistribution using external physical link with ends assigned to different VDC members' ports on the same physical Nexus switch.



The advantage of this deployment is high availability of this topology and isolation of routing and switching instabilities per Layer 2 domain that is achieved through the creation of a logical data center network using VLANs mapped VDCs. This type of deployment of Module 2 assigns a separate active-standby pair of VDCs to each access layer VLAN or group of VLANs. The maximum number of VDCs currently deployed per Nexus 7000 Series switch is four. The mapping of VLAN to VDC creates a segmented virtual network infrastructure that is highly available and stable. The routing protocol can be deployed per each VDC and can be OSFP for VDC D and EIGRP for VDC E, for example. VDCs can be used in active-standby of active-active mode using respective FHRP protocol and the fault tolerant connection between the two Aggregation switches. VDC D and VDC E will be actively exchanging routing information across VDC interconnection. Server Default Gateway will be the primary interface on the active VDC in Aggregation switch.

https://t.me/learningnets



Summary of OSPF and EIGRP Deployment in the Data Center Network

The core routers are connected with the Aggregation routers with Layer 3 links. The key design goal with the routing protocols is to advertise a default from the core routers to the Aggregation routers and summarized routes from the data center to the core network. The default gateway function is provided by HSRP, VRRP or GLBP. You assign the priorities to match the Layer 2 topology for the optimized traffic forwarding. Use preemption and the preemption delay feature.

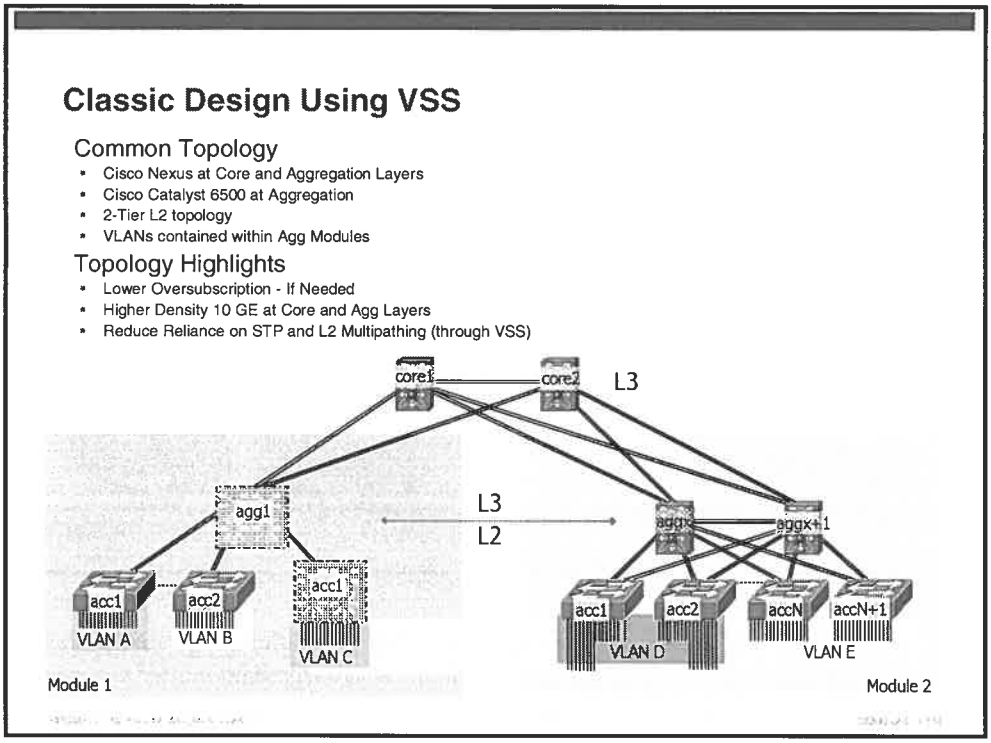
HSRP timers must be configured to detect a failure in a few seconds. OSPF or EIGRP are typically deployed. The goal of OSPF design is to maximize the data center stability since EIGRP does not prove to be stable in the environment with many topology changes. EIGRP in turn proves to be a better choice when the routing convergence must be immediate and where the network topology is more or less stable. The ABR routers can be Aggregation routers or Core routers. When Core routers are ABR routers, the routing processing on the Aggregation switches in the data center is minimized. The core routers have whole topology view of the data center; therefore, the route flapping in the data center is perceived in the core. Making the Aggregation routers the ABRs prevents this problem. The router configuration has passive SVI interfaces for the data center VLANs, and a Layer 3 link between the Aggregation routers ensures that Aggregation routers establish neighborship. Layer 3 configurations with summarization and default routes limit the umber of topology recalculations at Layer 3. Internal redundancy is achieved using the SSO/NSF.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enhanced L2 Topology

This topic describes the enhanced data center L2 design.



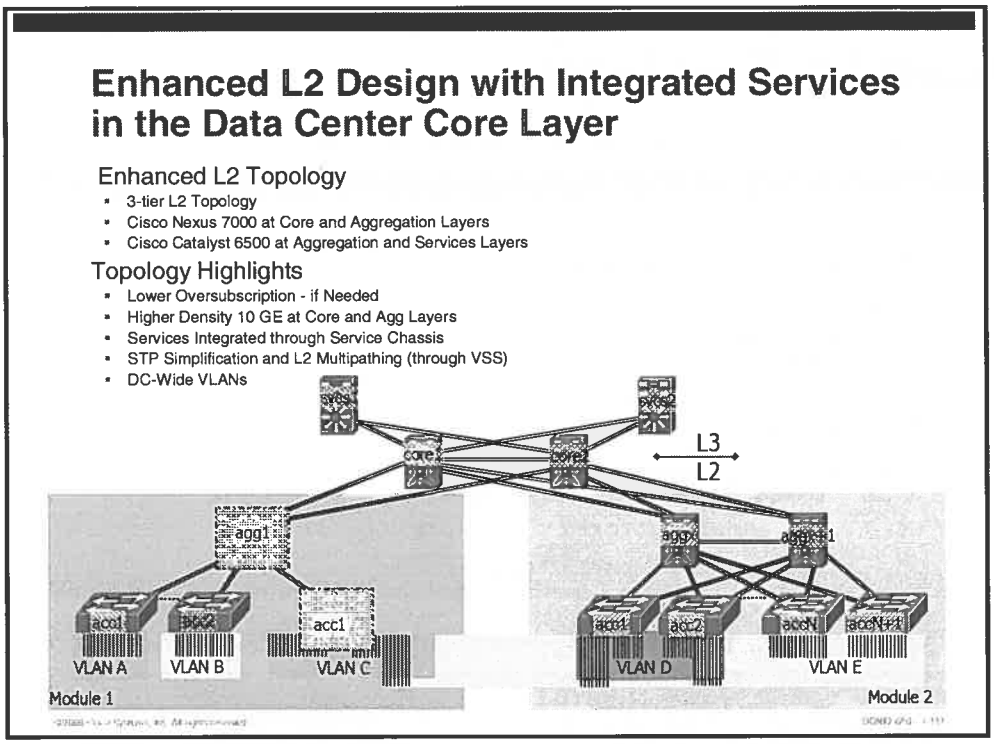
VSS uses Layer 2 multipathing on its virtual PortChannels toward the access or Aggregation switches.

In the previous topology without VSS, the access switch was a dual-homed switch connected to two Aggregation switches. In this topology the access switch still uses two physical connections, but they are used to build up a single virtual link, virtual PortChannel. The PortChannel selection is granular to a flow and represents a stretched definition of Layer 2 multipathing on the top of an EtherChannel. This EtherChannel is built across two different physical switches, but it still enables a host to send the data out using the same source IP address across two different Layer 2 links at the same time.

At any one point different flows can take different Layer 2 paths. This is analogous to applying Layer 3 CoS forwarding across multiple paths applied at Layer 2.

Host connected to an access switch providing VSS is better redundancy than NIC teaming since the host in the access layer used to be connected to the access layer switch using NIC teaming in active-active mode with the same source IP and MAC address which caused problems.

https://t.me/learningnets



This figure shows the optional topology with Service modules integrated at the core layer into the Service switches. This design is applicable to very large data center environments where the Layer 2 domain must span throughout the geography of the data center. Since Nexus 7010 Series switches currently cannot support routing at the size of the Internet routes, the Nexus 7010 switches right now do not fit into the Internet facing data center.

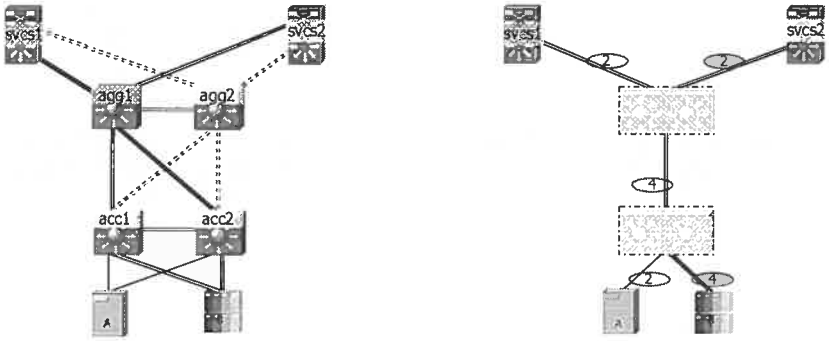
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Design Using L2 Multipathing with Integrated Services and Multipathing

Key characteristics

- Without L2 Multipathing
  - Always a blocking path to secondary root switch
  - Suboptimal traffic paths when:
    - Primary FHRP peer changes
    - Primary link to root switch fails
- With L2 Multipathing
  - Always a non-blocking path from access to agg multi-chassis switch
  - Always a non-blocking path from service to agg multi-chassis switch
  - Simpler support for active-active service environments



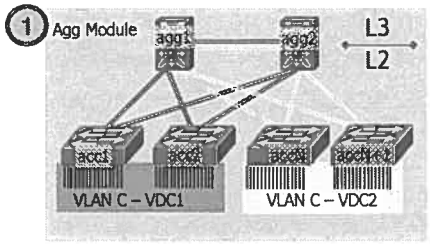
In the following two topologies the physical and the logical clouds are represented. The topology represents an attempt to simplify the looping problems so that a service device that is dual-homed does not participate in a forwarding Layer 2 Loop because a forwarding path exists to where the services are located. From the Aggregation to the Access Layer this topology is a non-looping topology on the right side, and from the access to the host layer there are multiple equal cost Layer 2 paths deployed to the Access Layer VSS system. The left section of the figure represents the limitations of a redundant L2 topology without L2 multipathing. If a trunk goes down in the left network, Spanning Tree convergence must complete before an alternate path can be used. STP blocks the paths from the access to the Aggregation Layer switches and the traffic path is suboptimal upon the primary FHRP peer change and primary link to the root switch failure. The topology on the right side of the figure illustrates that with L2 multipathing, there is always a non-blocking path between Aggregation switches.

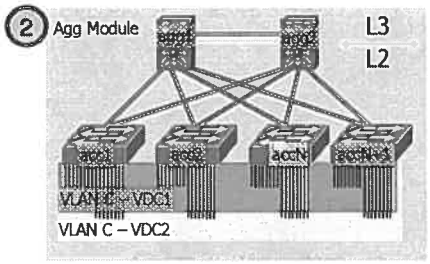
https://t.me/learningnets

Isolating Collapsed L2 Domains Virtual Device Contexts at Aggregation Layer

- Pods are isolated at aggregation layer. Each Pod runs its own STP instance (instance per VDC).
- Multiple pods could exist in a single VDC.
- VLANs contained within Agg Module per VDC.

- Pods are isolated at access layer.
- Each Pod belongs to multiple VDCs.
- Each VDC topology requires dedicated Ports.
- VLANs contained within Agg Module per VDC.





- Higher 10GE Port Density Allows multiple Agg Pairs to be collapsed.
- Collapsed Agg Pair could still be L2 isolated (different STP instances).
- VLAN IDs could be replicated on different VDC – shared infrastructure.

The left Model 1 has 2 VDCs deployed, one for each VLAN C. A VDC trunk port does not currently exist. One mechanism that enables the separation of the failure domain is by dropping a VDC. Each VDC is running its own spanning tree process; therefore, different VDCs in the same chassis will have spanning tree topologies logically, and physically will be separated. Module 1 Aggregation switch has a VDC that has dedicated ports connected to the data center topology of eight pairs of the Aggregation switches that have two VDCs. Access layer switches are not aware of VDCs at all; physical ports are connected to a particular physical port assigned to a particular VDC. This topology isolates VLAN 2 failure domain for VLAN C server farm.

The Module 2 figure on the right has a slightly different arrangement but the same physical topology. Each access switch is connected to two different VDCs, which is not recommended in every case. There must be dedicated physical ports that connect to either VDC to realize this topology.

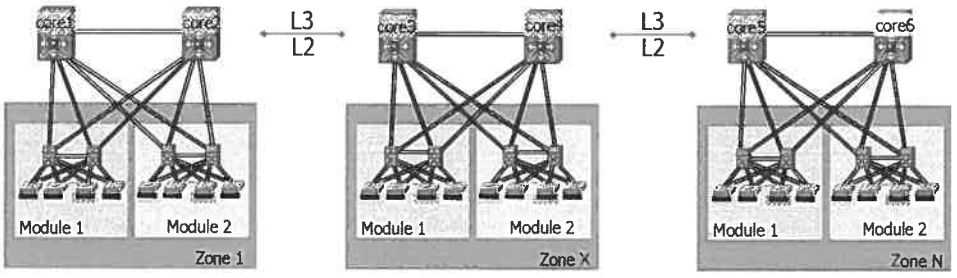
These two topologies provide an example of a separation of spanning tree domain and thus separation of failure domains at the VDC level.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enhanced L2 Collapsed Core Virtual Device Contexts at Core Layer

- Enhanced L2 Topology with Collapsed Core
  - Benefits of 3-tier L2 Topology
  - Zones are still isolated (an STP instance per each zone)
  - Core switches are managed independently by VDC

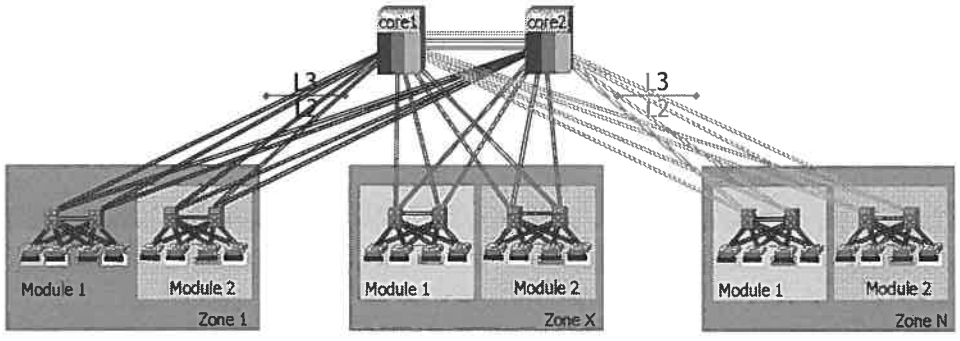


The 3-tier Layer 2 design is becoming interesting due to some recent initiations of extending Layer 2 domain across the data center to enable application mobility and high availability. A 3-tier Layer 2 design requires having a VLAN; for example, yellow VLAN to be extended and deployed across all zones.

https://t.me/learningnets

Enhanced L2 Collapsed Core
Virtual Device Contexts at Core Layer

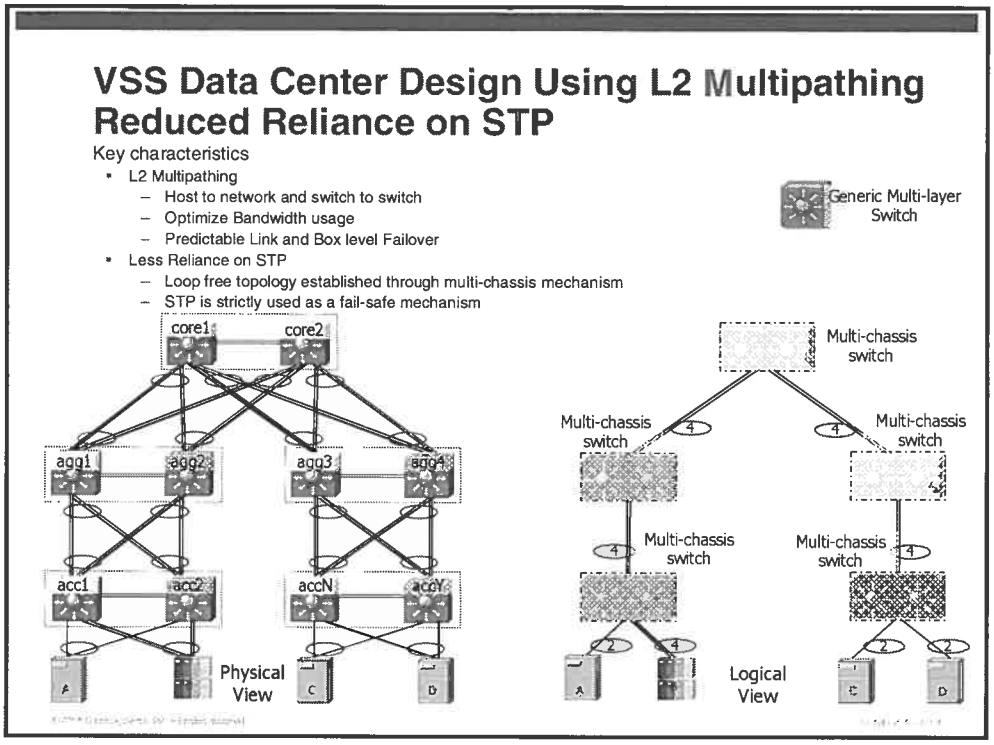
- A virtual Core switch for each Zone
- Inter-Module within the Zone communication
- Zones are still isolated (an STP instance per zone)
- Core switches deploy independent VDCs that enable



In this design each zone has a dedicated VDC in the Core layer that is enabled in active-active or active-standby mode. The Layer 2/Layer 3 boundary is pushed out from the data center Aggregation to the Core Layer in order to meet application requirements. From the data center deployment perspective, zones established in this design enable zones to be bigger than zones in the past. VDC is isolating and terminating the Layer 2 failure domain. VDCs have no way to communicate (based on the configuration) through those boxes. There must be a physical wire that interconnects the VDCs within the physical switch.

© 2008 Cisco Systems, Inc.

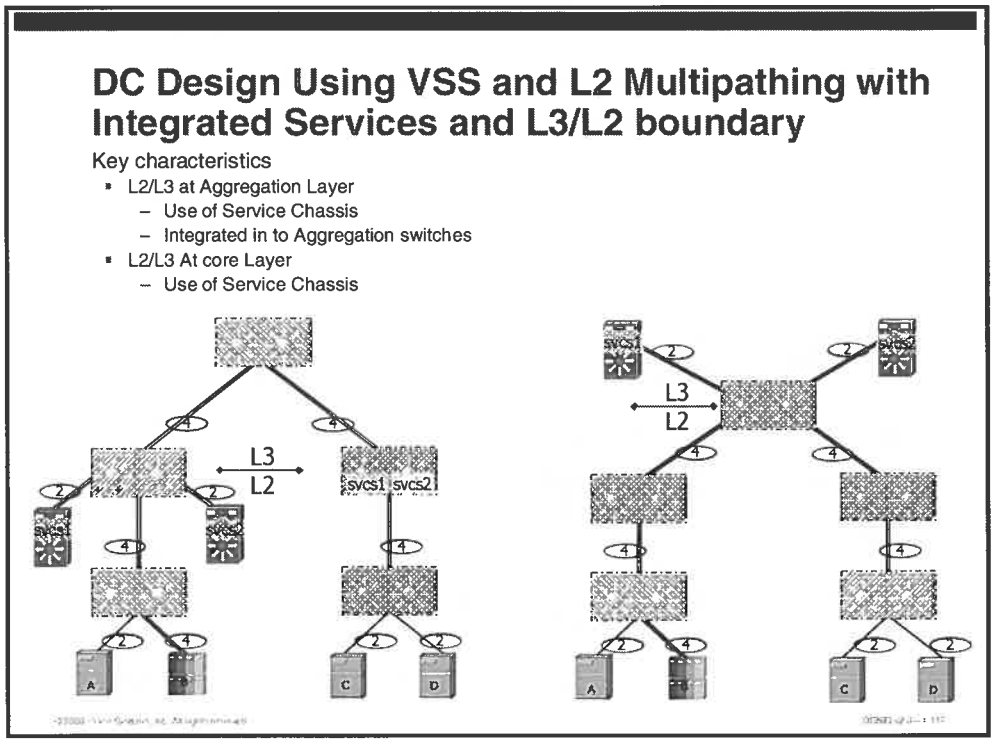
https://t.me/learningnets



Data center topology can be deployed in the most reliable way using VSS system for high availability at every single layer, as shown by the figure. The left figure shows the physical and the right figure the respective logical data center topology.

Layer 2 multipathing is interconnecting the VSS systems at different layers all the way from the host through the network to the core layer. The STP is still part of this design but there is no physical loop in the topology; therefore, Spanning Tree operation is not needed. This topology simplifies the traffic pathing all throughout the infrastructure and brings higher data center availability and reliability.

https://t.me/learningnets



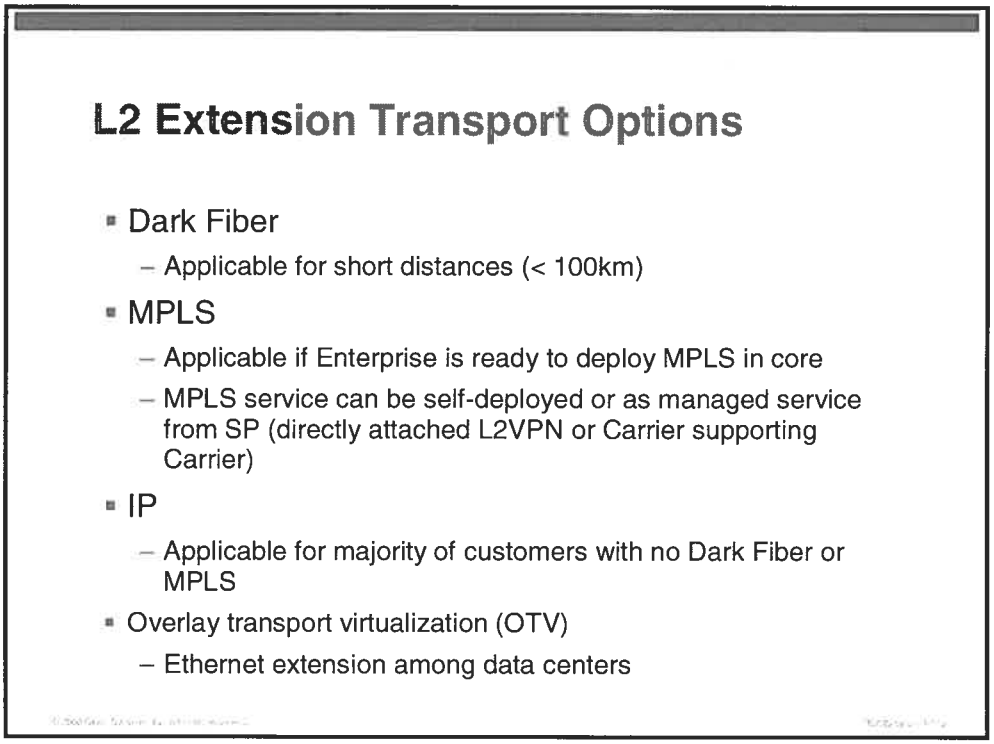
The figure represents an optional 3-tier data center high availability design with Layer 2/Layer 3 boundaries at the core or at the Aggregation Layer. Service integration is accomplished in the Aggregation Layer or in the Core Layer, respectively. The Catalyst switch represents here and now a service switch in both cases due to both high port density and very advantageous architecture supporting the latest data center services.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Layer 2 Extensions for Data Center Interconnect

This topic describes the L2 Extension Transport Options.



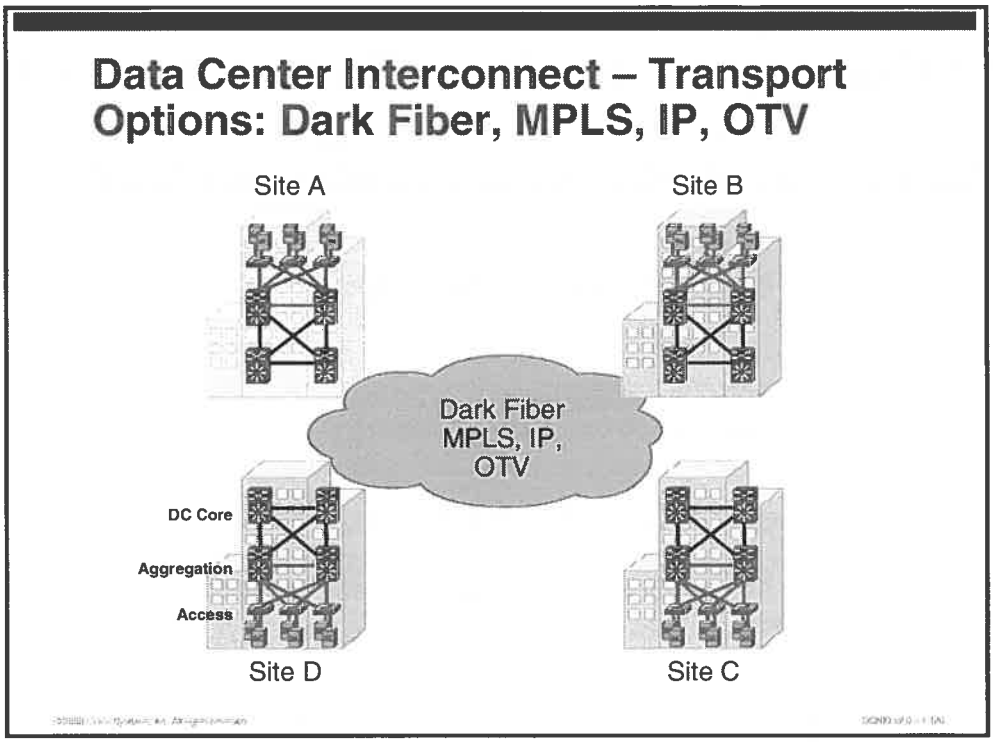
Data center L2 extension transport options are:

- Dark Fiber
- MPLS
- IP
- OTV

This document will give hardware forwarding-engine designers an idea of what is required to support OTV. OTV is a new way of providing L2 and L3 VPN support over an L2 or L3 infrastructure. OTV provides an "over-the-top" method of doing virtualization versus traditional "in-the-network" style mechanisms where multiple routing and forwarding tables are maintained in every device from a source to a destination. OTV can be incrementally deployed and reside in very few number of boxes at an edge site. We call these boxes "Edge Boxes," which perform typical Layer 2 learning and forwarding functions to provide the ability to connect L2 networks, L3 networks, or hybrid networks together.

OTV is the Nexus 7000 Series switches' alternative to the MPLS.

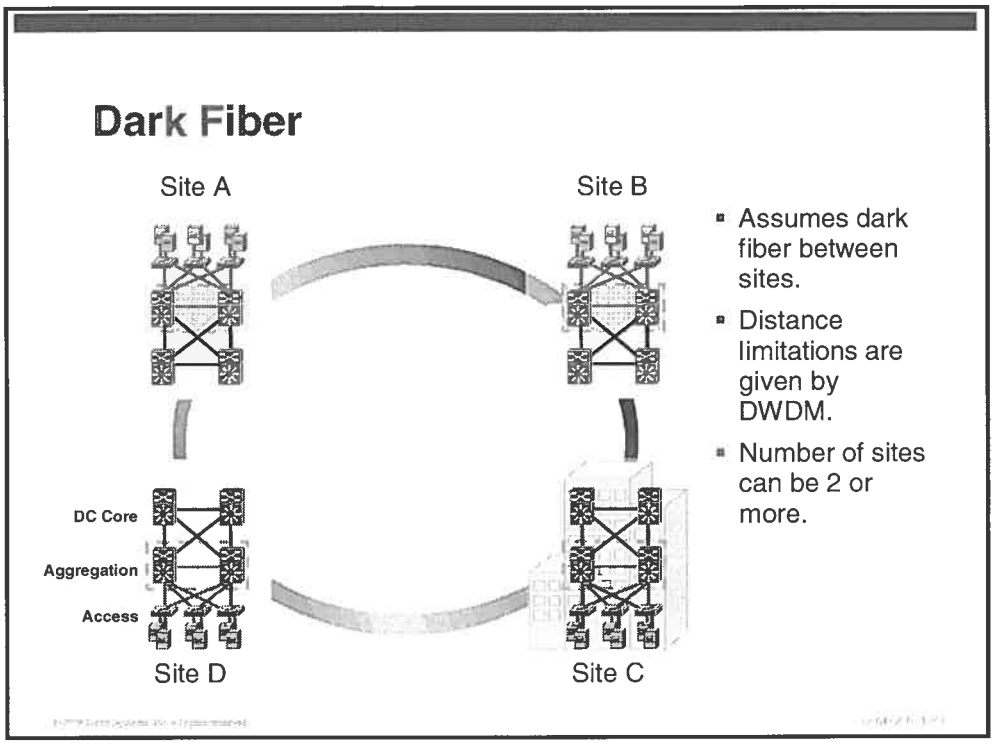
https://t.me/learningnets



The figure shows a simple data center topology with disaster recovery site or with multiple locations connected using Dark Fiber, MPLS, IP and OTV.

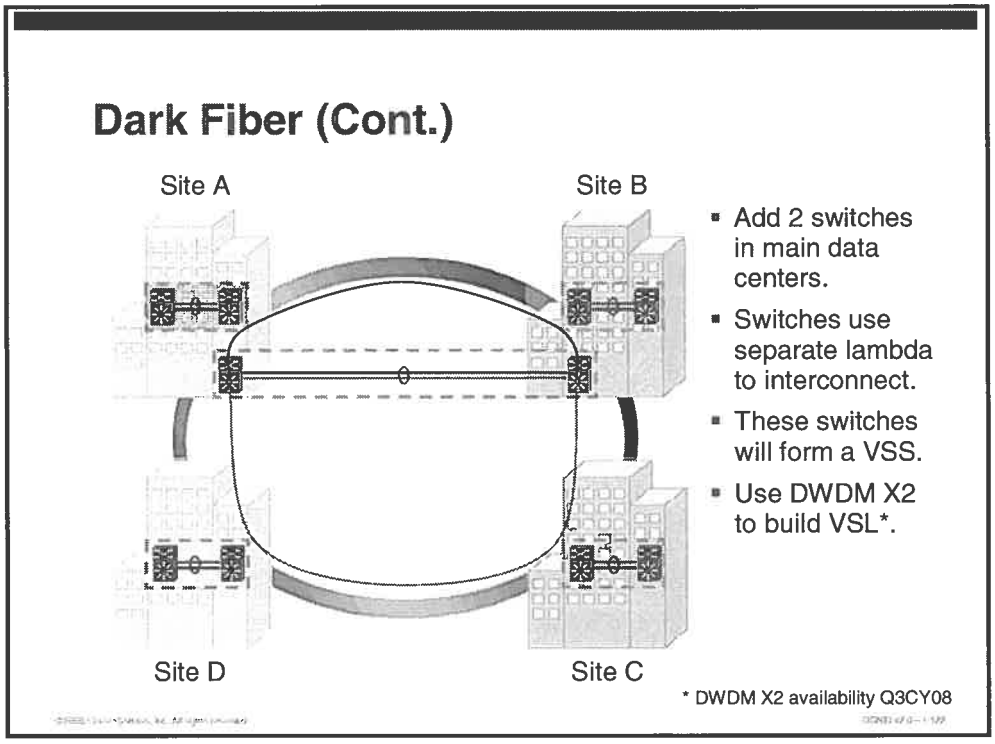
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Data center sites can be connected using dark fiber (DWDM), which provides the highest bandwidth and fast (sub-second) failover.

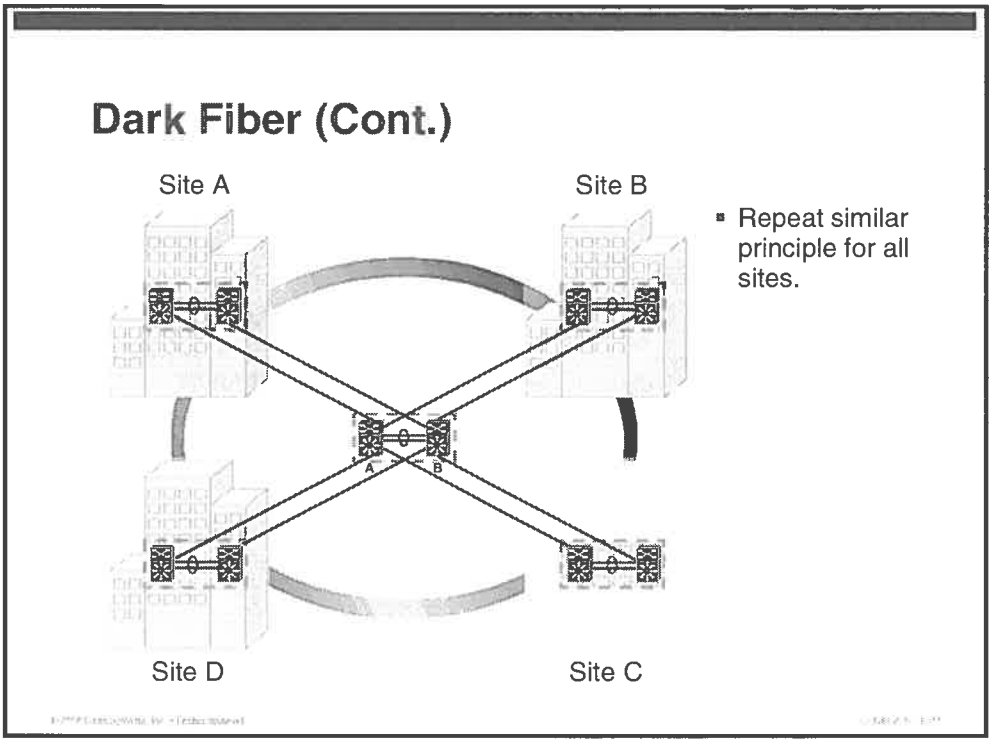
https://t.me/learningnets



Data center sites can be interconnected by VSS VSL that is based on DWDM interconnection. The two members of the VSS reside in two different sites: Site A and Site B on the figure.

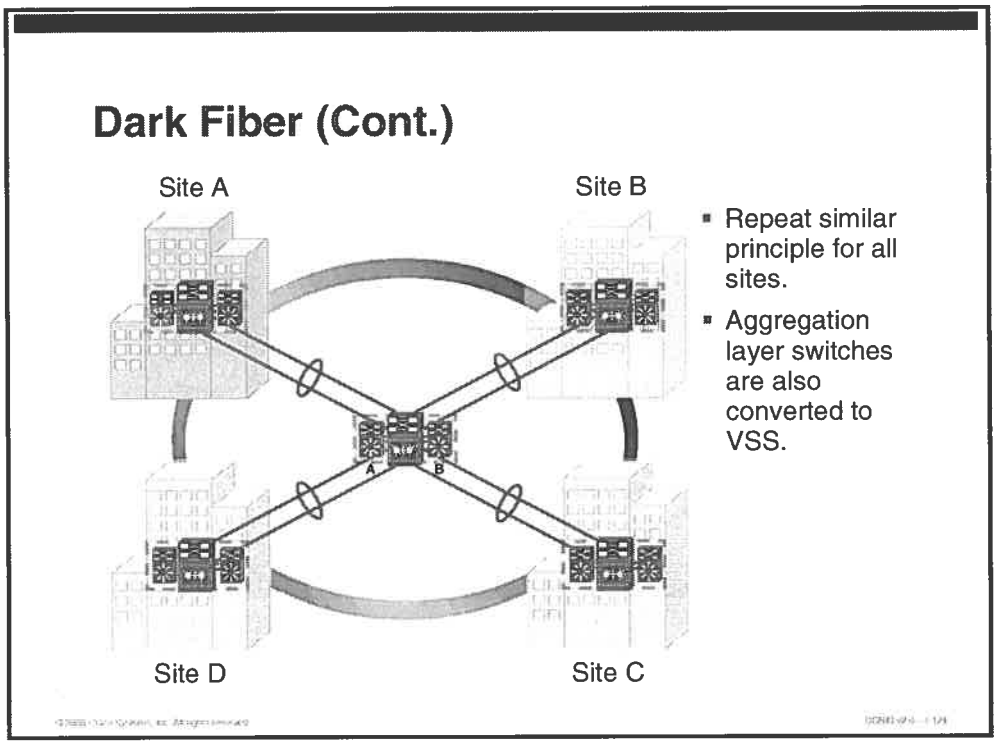
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure shows a core VSS system interconnecting data center VSSs from different sites using vPC.

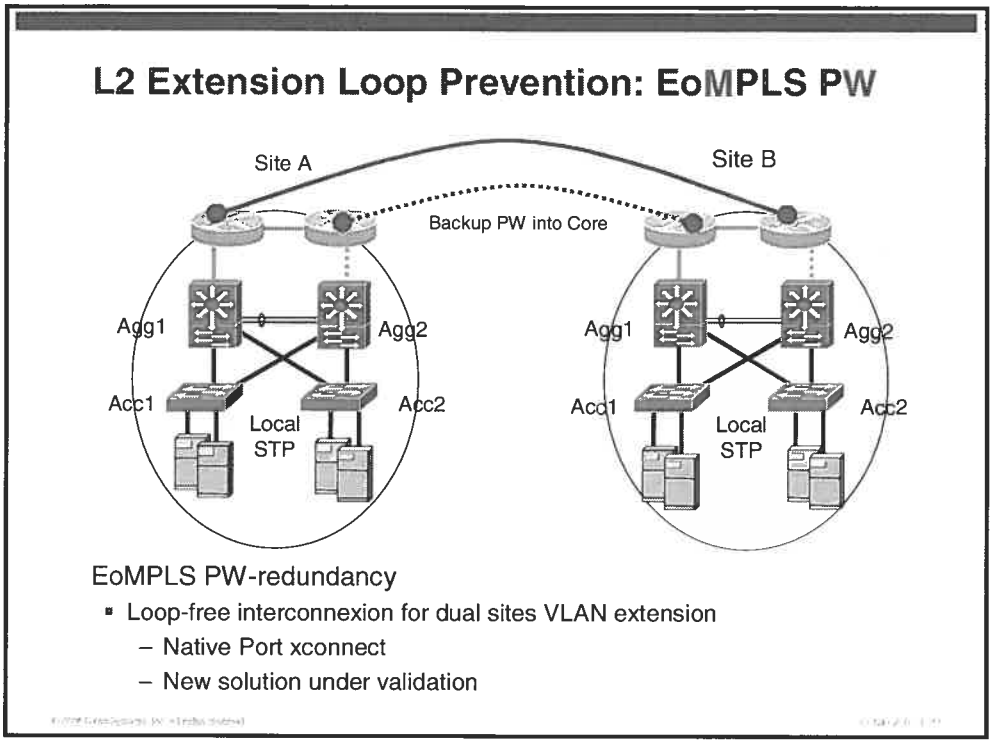
https://t.me/learningnets



The figure shows a core VSS system interconnecting data center VSSs from different sites using vPC along with Catalyst 6500 Service switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



VPLS technology began in mid-2001, when various service providers that were evaluating the possibility of deploying Metro Ethernet services identified concerns about the deployment of basic Ethernet switching architectures. They correctly noted that Ethernet switching technologies that were developed up to that point were deficient in providing network services that could support the SLA characteristics comparable to existing traditional WAN services, such as Frame Relay, T-1/E-1 and ATM.

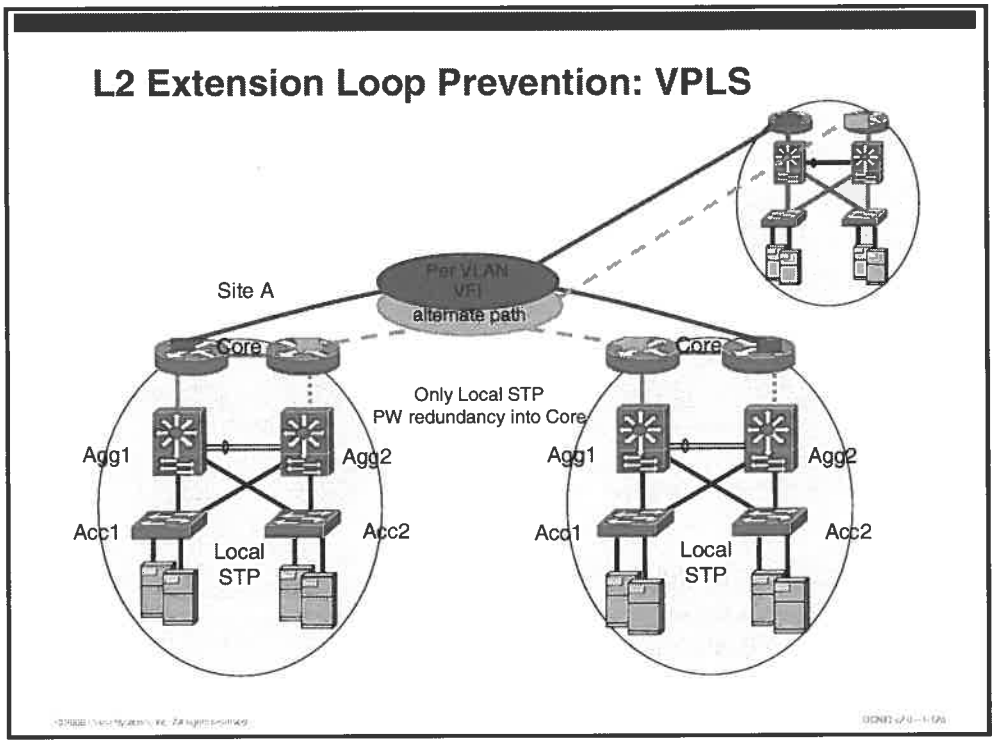
These service providers sought a technology that could support multipoint data services (transparent LAN services) over an infrastructure that provided traffic engineering, high availability, and OAM characteristics similar to their existing services. The obvious technology direction to investigate was MPLS, and specifically EoMPLS.

EoMPLS had been previously defined in the IETF specification originally referred to as the Martini Draft, and was posted to the IETF draft site toward the end of 2000. This document explained the format for point-to-point L2 VPN services using Ethernet frames transported over an MPLS pseudowire, with LDP specified as the signaling and OAM mechanism for this point-to-point service.

This EoMPLS draft specification supported most of the service characteristics that many of the service providers were seeking, except that it did not support a multipoint service architecture, only a point-to-point or hub-and-spoke service architecture. However, the basic encapsulation and signaling mechanisms of EoMPLS would prove to be a reasonable basis for defining a multipoint service architecture.

Note Use EEM to enable the backup PW.

https://t.me/learningnets



VPLS Architecture Characteristics

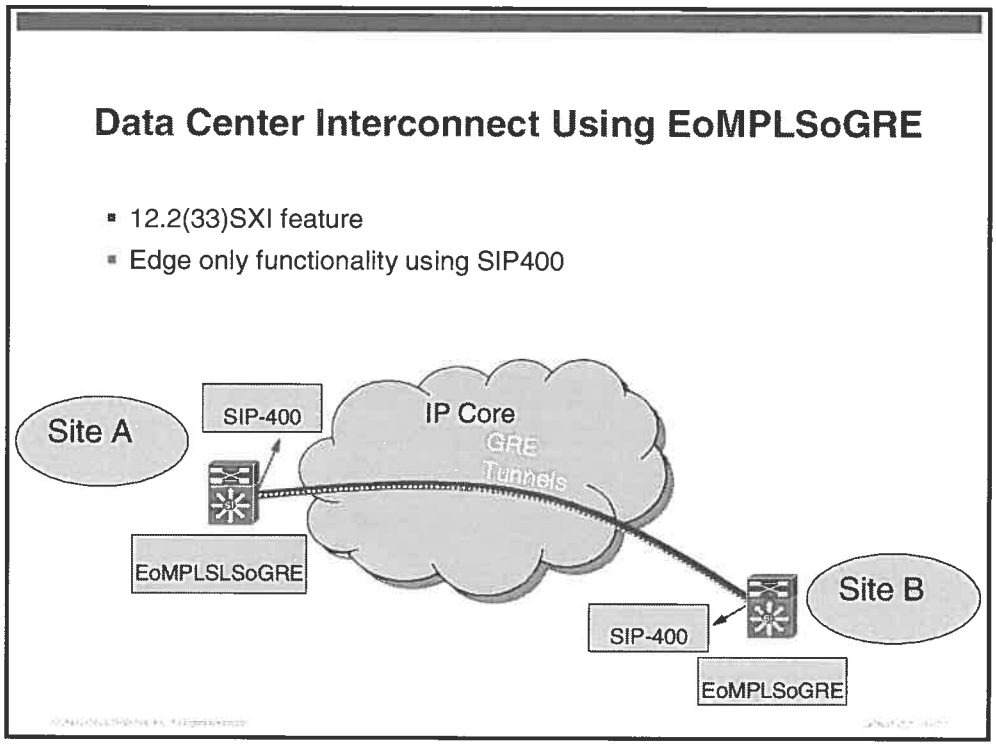
- Suitable for small customer implementations
- Simple provisioning
- Full mesh of directed LDP sessions required between participating PEs
- VLAN and Port level support (no QinQ)
- VPLS OFFERS Single Flat Hierarchy
- MPLS to the Edge

Drawbacks

- No hierarchical scalability
- Scaling issues:
  - PE packet replication, Full mesh causes classic - N*(N-1) / 2 concerns

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



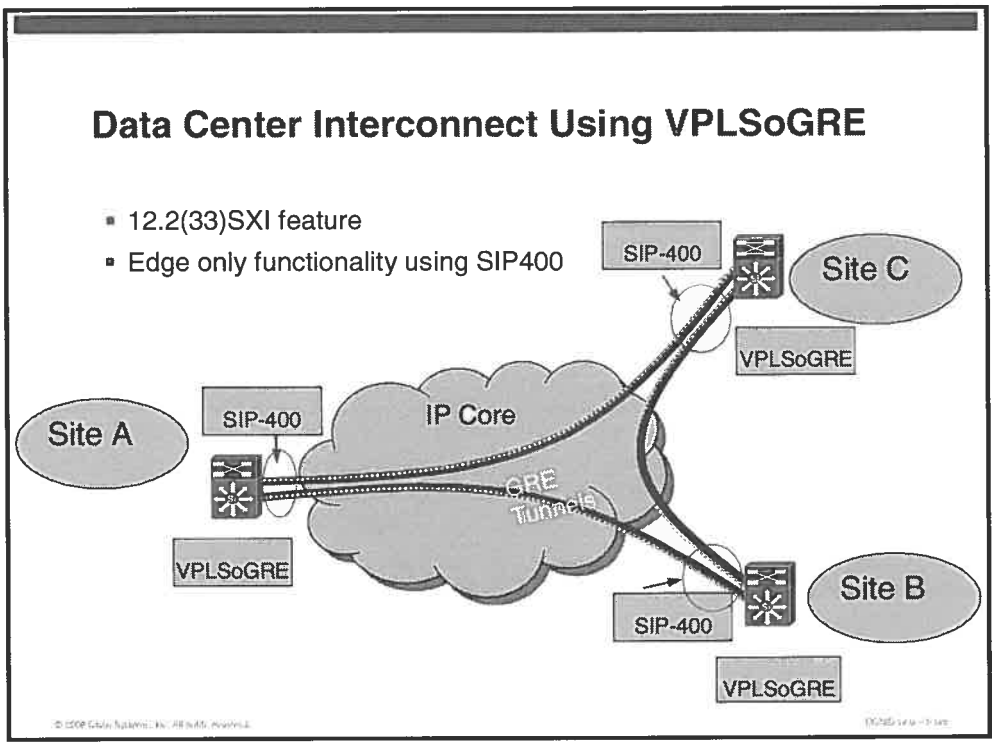
The figure shows a data center interconnection based on EoMPLSoGRE.

The following steps are needed:

- Establish GRE tunnels between routers.
- Establish MPLS LSP tunnels on these GRE tunnels, also called MPLSoGRE tunnels.
- Establish AToM sessions over these MPLSoGRE tunnels. MPLSoGRE tunnel IP Core.
- R2 AToM VCs AToM VCs IP GRE Tunnel.
- R1: AToM over GRE Tunnel.

By establishing MPLS tunnels over GRE tunnels (MPLSoGRE tunnels), Catalyst 6500 Series switches can be configured to provide AToM services on an IP-only network (AToMoGRE).

https://t.me/learningnets



The figure shows a data center interconnection based on VPLSoGRE.

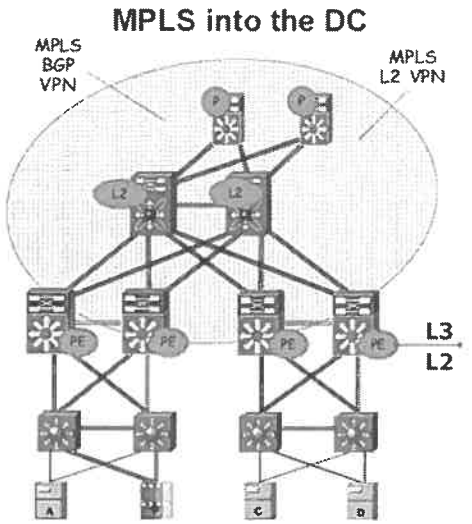
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Nexus 7000 Series Switches

Cisco Nexus 7000 Series Switches in the Segmented Data Center Today

- N7k Bridged Core
- Catalyst 6500 Aggregation
- MPLS traffic is bridged transparently in the core
- Challenges:
  - –Spanning Tree in the core
  - –Larger L3 peering domain



Over-the-Top Virtualization (OTV)

- OTV is something that addresses the requirements for an MPLS-like feature for the enterprise.
- OTV enables multipoint connectivity, loop preventions.
- OTV controls the Layer 2 broadcast and the Layer 2 domains.
- The only thing it needs is an IP core network or pop maximum benefits and IP multicast core.
- OTV addresses a problem for a large portion of the data center market.
- This is especially true overseas and below the "fortunate 50" watermark.
- OTV is a fundamental block in the OPEN Ethernet architecture, which has been agreed upon by multiple instances across the company.
- OTV provides the IP L2VPN complement to our enterprise direction of IP based L3VPNs (vNETs).
- OTV enables DCE to scale to a very large footprint.
- OTV allows the suppression of broadcast traffic needed in traditional data plane learning. By connecting DCE clouds together with OTV, larger DCE clouds are possible.
- Actually, OTV allows L2 in general to safely scale beyond its current limitations.
- OTV is primarily positioned for Ethernet extension among data centers.

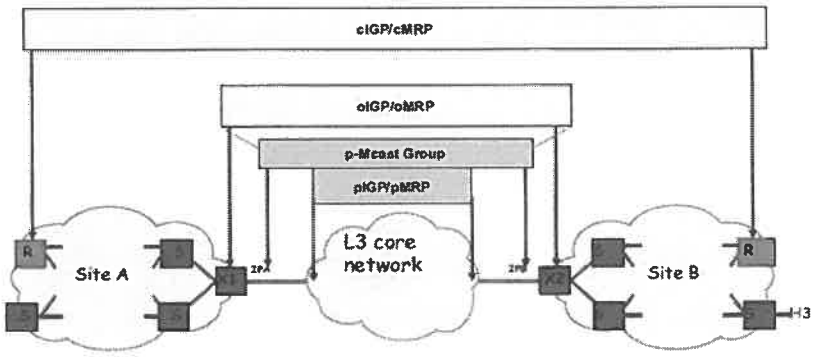
Other use cases:

- Direct L2 Interconnection of Branch Routers to avoid PE-CE routing operations.
- May be leveraged (with OPEN) by SPs to provide multi-point L2VPN services.
- OTV does not create a lossless Ethernet extension.

https://t.me/learningnets

Data Center Interconnect – Transport Option: OTV

OTV Control Planes

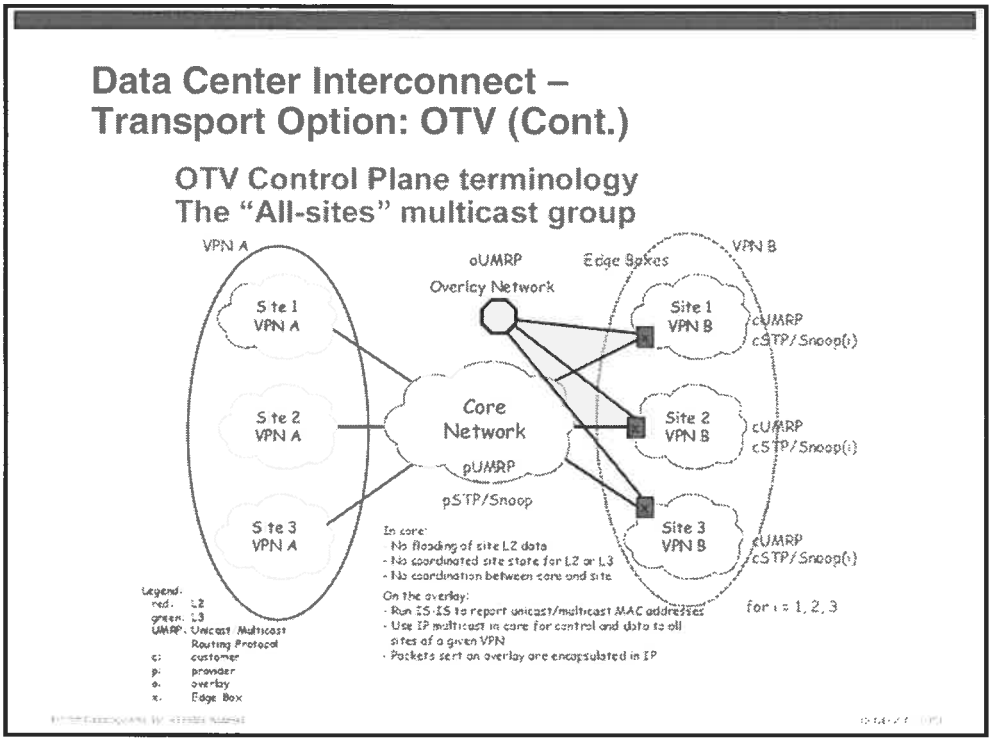


Over-the-Top Virtualization (OTV) Entities

- **Terminology Edge Box**: This is a modified L2 switch that performs OTV functions. It will typically run as an L2 device but can be co-located in a device that performs L3 routing on other L3-enabled ports.
- **Internal Links**: These are L2 enabled links that connect to other traditional L2 switches on the site-side of a network.
- **Overlay Link**: This is a logical multi-access interface. The interface can replicate broadcast and multicast packets efficiently. The interface takes L2 frames and encapsulates them in IP unicast or multicast headers.
- **Broadcast Forwarder (BF)**: If there is more than one Edge Box at a site, one of them is elected as BF to send and receive broadcast frames to and from the Overlay link.
- **MAC Table**: This is a forwarding table of 48-bit MAC addresses. The table can contain unicast or multicast MAC addresses. The table is populated by two means, one being traditional data-plane learning and the other by the IS-IS protocol at the control-plane.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Forwarding Rules

Most of the interesting forwarding cases are when a packet comes from the Overlay Link to be forwarded to an Internal Link, or the reverse. But for completeness, we will describe how forwarding between Internal Links will occur. There can be multiple Overlay Links, one per VPN, but typically when a packet is forwarded it is not forwarded between or across Overlay Links.

- **Forwarding between Internal Links:** When an Edge Box has Internal Links, it operates like a traditional L2 switch. That is, it will send unicast packets on a port where the MAC was learned, it will send multicast packets on the ports it has IGMP-snooped, and it will send broadcast packets out all ports. This is done on a per VLAN basis.
- **Forwarding from an Internal Link to an Overlay Links:** An Edge Box will decide to forward a unicast, multicast, or broadcast packet over the Overlay Link when IS-IS has put the logical port of the Overlay Link in the MAC table for the corresponding unicast or multicast MAC address. For broadcast packets, there is BF bit per Overlay Link on the Edge Box. If the bit is not set, the packet is dropped, otherwise it sent over the overlay network.

  When a packet is sent over the overlay link, it is first prepended with an IP header. The packet as received from the Internal Link is not touched other than to remove the Preamble and FCS from the frame. The IP header, outer MAC header, and physical port the packet is to go out is all cached as a rewrite string. This is so all the information to physically forward the packet is all together to easily prepend and send.

  The encapsulation format will use RFC 3378. That is, the outer MAC header will use ethertype 0x800, the IP header will use protocol number 97, and the original MAC frame sent by the host comes next. The IP addresses and the outer MAC addresses are all provided and stored for the hardware by the control-plane software.

https://t.me/learningnets

- **Forwarding from an Overlay Link to an Internal Link:** When a packet is received on the Overlay Link, it will need to be IP decapsulated to reveal the inner MAC header for forwarding. The inner MAC header SA and DA addresses will used for the MAC table lookup. When a unicast or multicast packet is received on the Overlay Link, an Edge Box must determine it should forward it. It must do an RPF check to decide if the packet should be dropped or to continue processing. The procedure for the RPF check is:
  - Do a hash on the inner SA and DA.
  - Do a MAC table lookup on the SA. If the header rewrite string IP DA equals the IP SA from the received packet, the RPF succeeds and processing continues. Otherwise, the packet is dropped.

  This RPF check is necessary when the source MAC site has flooded the packet to all Edge Boxes and each one has determine they have a route to the destination MAC in this site. We want to avoid duplicates from entering this site, so the RPF check eliminates the all by one copy.

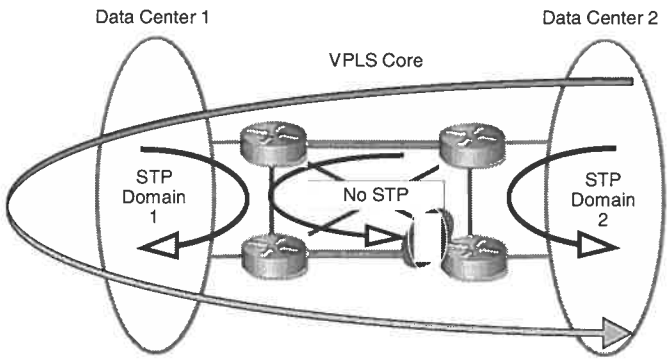
  For broadcast packets, a BF bit per Overlay Link is tested. If the bit is clear, the packet is dropped. Otherwise, packet forwarding continues. Once the packet is decapsulated, the inner MAC DA is looked up in the MAC table and L2 forwarded out Internal Link port. If the packet is a broadcast packet, it is flooded on all Internal Links in the VLAN. In the 3 packet type cases, after a packet is decapsulated (the IP header is stripped), the Edge Box sets bit 41 (the L/G bit) in the inner MAC header SA. This tells other Edge Boxes who may get this packet, which the SA is still sending from a remote site. This step is necessary for fast MAC moves between sites.

Learning Rules

- **Learning between Internal Links:** When a packet is received on an Internal Link and the SA has not been cached in the MAC table, the SA is cached with the port the packet was received on. If a packet is received from the same SA on a different Internal Link port, the MAC entry port is updated to the new port. This is traditional L2 MAC learning.
- **Learning from an Internal Link to an Overlay Link:** If a MAC table entry exists and points to an Internal Link and an IS-IS advertisement indicates the MAC is reachable over the Overlay Link, then the MAC entry is updated. This is known as the "IS-IS preference rule". This case occurs when the MAC has moved out of the local site and has been IS-IS advertised by the site it now resides in. Subsequent packets coming from the moved MAC will have bit-41 set. In this case, the same MAC will have two entries in the MAC table, the MAC with bit-41 cleared pointing to the Overlay Link and the MAC with bit-41 set pointing to the Internal Link (the one from where the last packet was received). The latter will never be used to forward packets on since hosts do not send with bit-41 set.
- **Learning from an Overlay Link to an Internal Link:** If a MAC has moved into the local site and an Edge Box has the MAC entry pointing to the Overlay Link, it must move the entry to a local port as soon as possible. That means as soon as the MAC starts sending in the local site, the Edge Box must either forward packets locally or remotely to the destination the MAC is sending to with no packet loss. So we have an exception to the "IS-IS preference rule". The exception states if the MAC table entry already existed for the SA when the packet is received and the entry has the Overlay Link cached, then it is *not* updated with the Internal Link the packet was received on if bit-41 (the L/G bit) of the SA is set. This is because the MAC is still remote and has not moved. The packet just came around to an Edge box (and was decapsulated by the entry Edge Box). So, when bit-41 is cleared, we do want to change the existing MAC table entry's cached port from the Overlay Link to an Internal Link so we can react to the MAC-move between sites.

© 2008 Cisco Systems, Inc.
https://t.me/learningnets

VPLS with Split-Horizon Loop-free Interconnection with STP Isolation

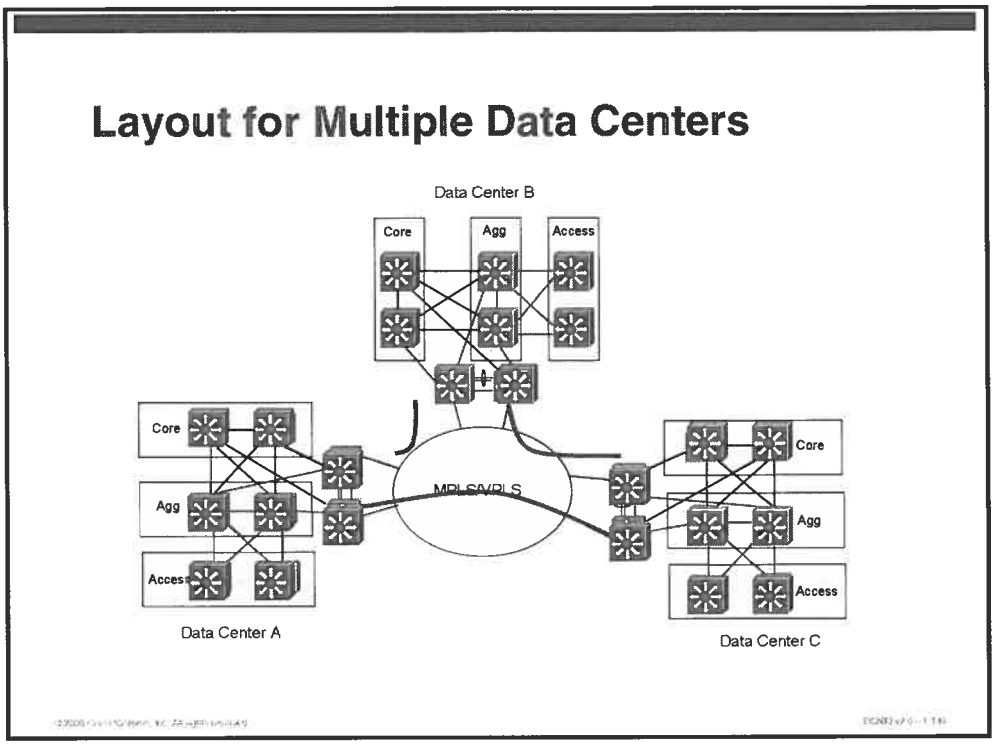


- Split Horizon prevents loops in the DC - not needed in core
- An end-to-end loop could possibly form
- Loops mean risks of permanent broadcast storms

The Spanning Tree Protocol (STP) may be running in Data Centers to prevent loops. A VPLS core interconnecting two DCs does not run STP. Consequently, each Data Center will likely be running its own instance of STP which is not connected to instances of STP at other DC sites.

In this situation, much like in MPLS networks, end-to-end loops are possible and extra care must be taken when designing these networks to avoid such end-to-end loop scenarios.

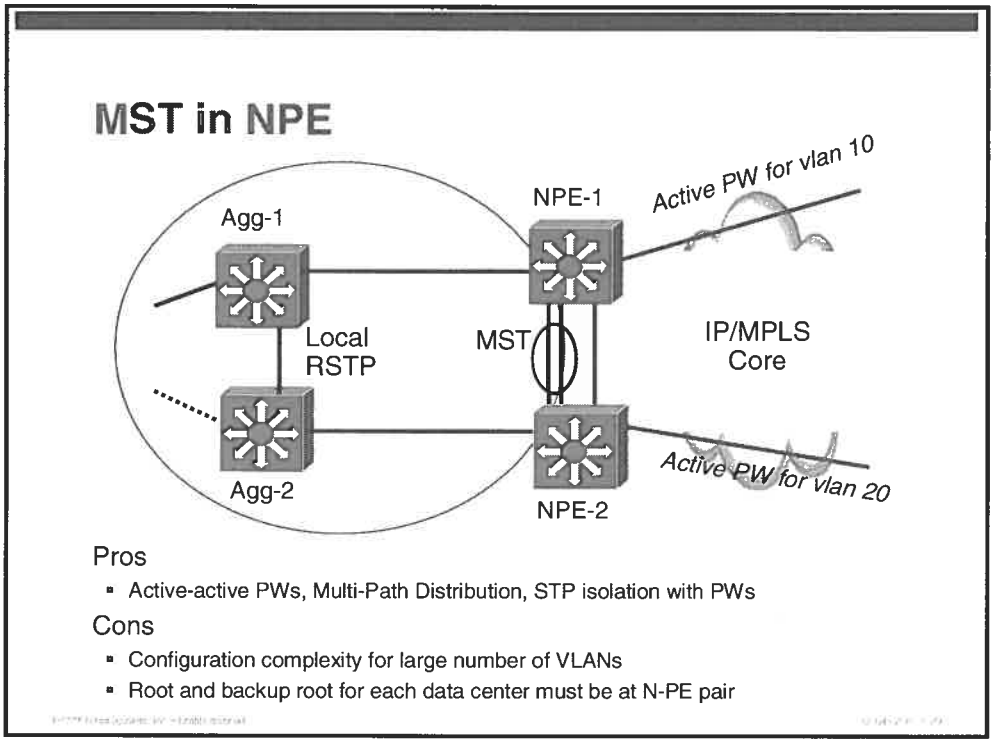
https://t.me/learningnets



Using MPLS/VPLS the topology would be a logical full mesh with one logical connection from each site to each other site. At layer one there would still be physical redundancy with multiple links to the service provider.

© 2008 Cisco Systems, Inc.

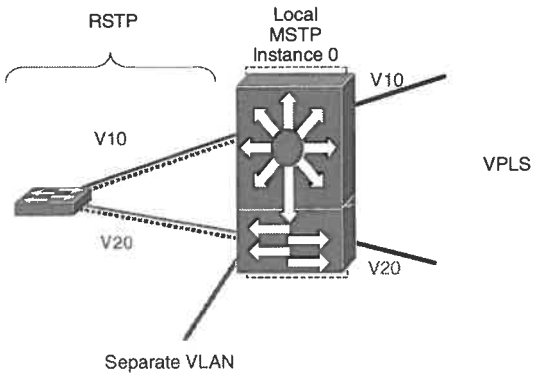
https://t.me/learningnets



Multiple Spanning Tree (MST) (IEEE standard 802.1s) can be used with NPE (Network Provider Edge) multilayer switches. MST allows a separate instance of STP to be run for each VLAN. Each instance runs independently of the others STP instances inside of the MST region. MST maintains an Instance 0 as the Internal Spanning-Tree instance. This is reserved for interacting with other Spanning-Tree Protocols (STPs) and other MST regions. In order to maintain a loop-free topology, the spanning-tree states, such as forwarding and blocking for all boundary ports on the edge of the MST region, match the spanning-tree state from the IST.

https://t.me/learningnets

MST with per VLAN Path-Splitting

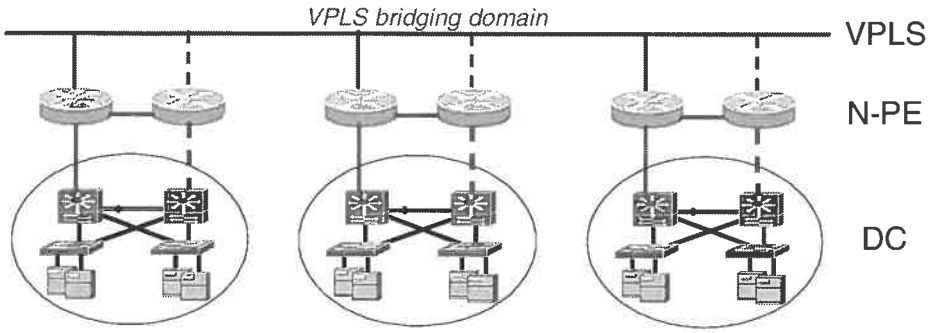


- From outside of MST, MST region is seen as a single switch
- Control the forwarding path per VLAN using port cost on the Aggregate Layer switch

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

EEM in N-PE over VPLS



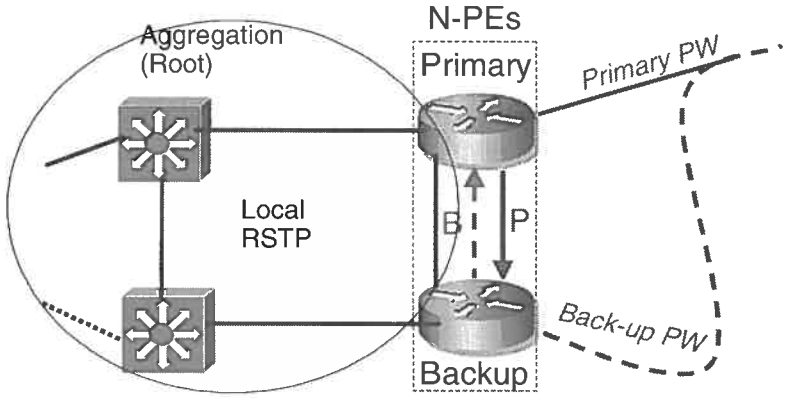
Pros:

- Scalable to any number of data centers
- Total DC STP isolation
- Does not require changing Root Bridge
- Scalability to any number of VLANs (QinQ)

Weakness:

- Scripted technology, not native, but is scripting really a Con ???

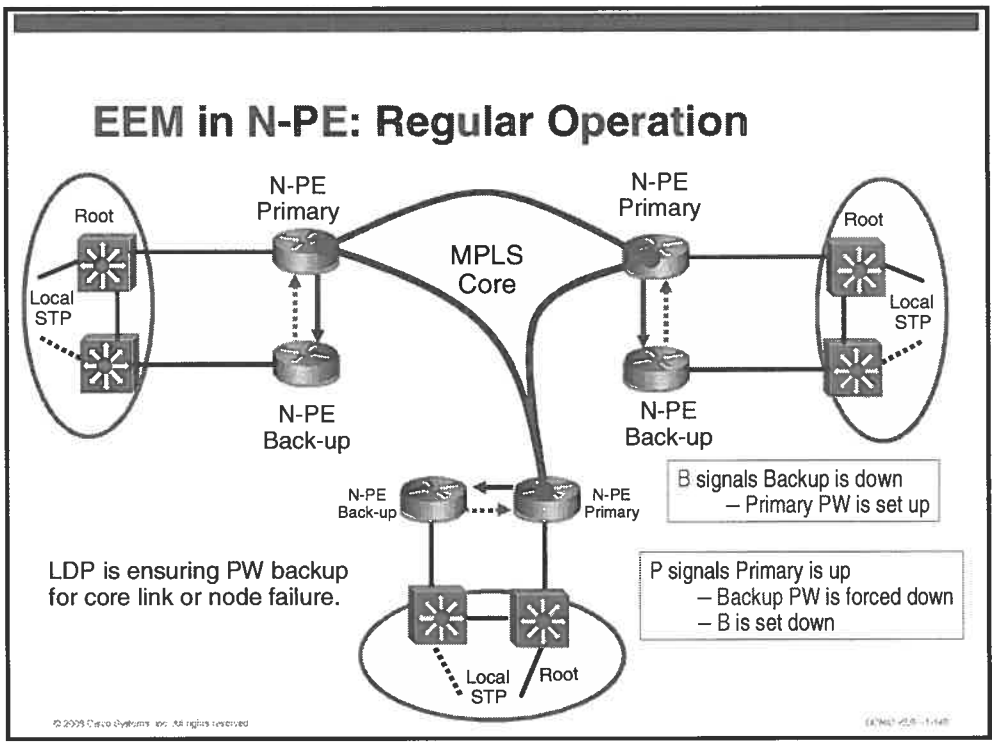
EEM in N-PE = "VPLS Node Redundancy Using EEM" Active/Back-Up PW Signaling with EEM



- Primary and Backup PW per VFI
- Loop Prevention at VPLS core (only one PW active per vlan-vfi)
- Handshaking using EEM loopback tracking (P and B loopbacks)
- Very similar topology as MST in N-PE
- Easy evolution to QinQ solution (just edge port add-on)

https://t.me/learningnets

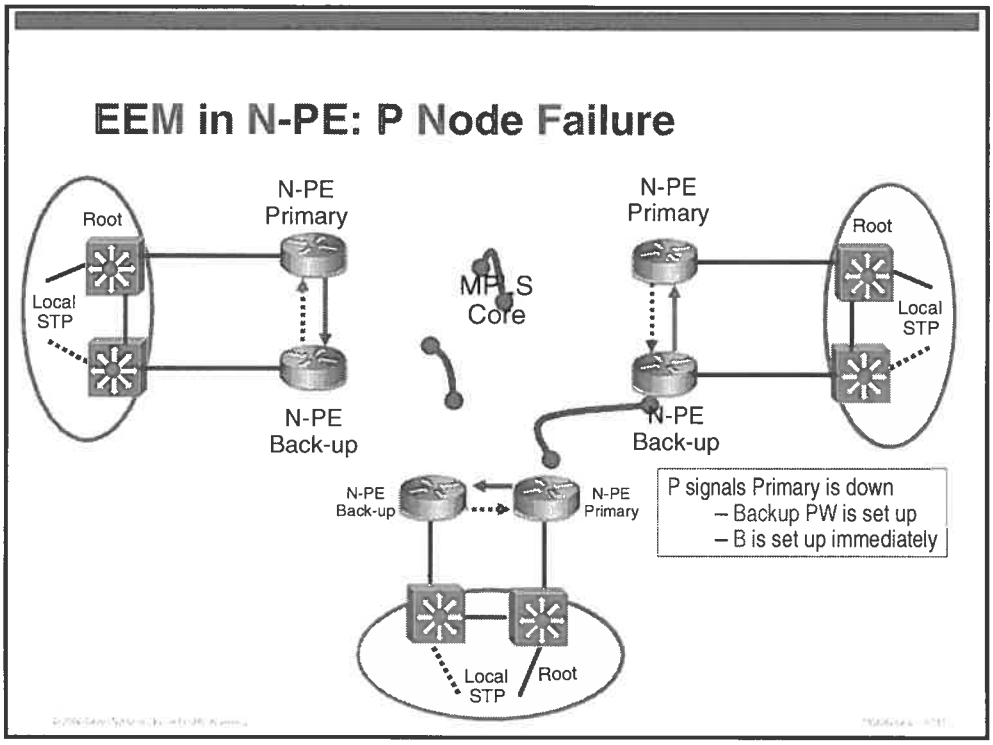
EEM in N-PE: Regular Operation



© 2008 Cisco Systems, Inc.

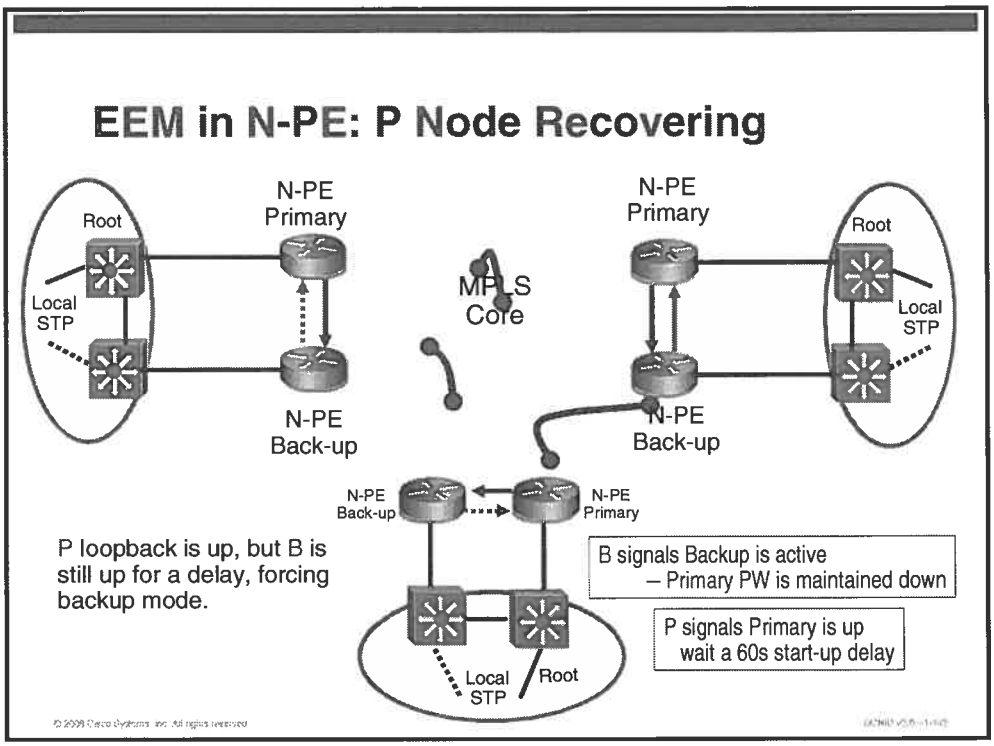
https://t.me/learningnets

EEM in N-PE: P Node Failure



https://t.me/learningnets

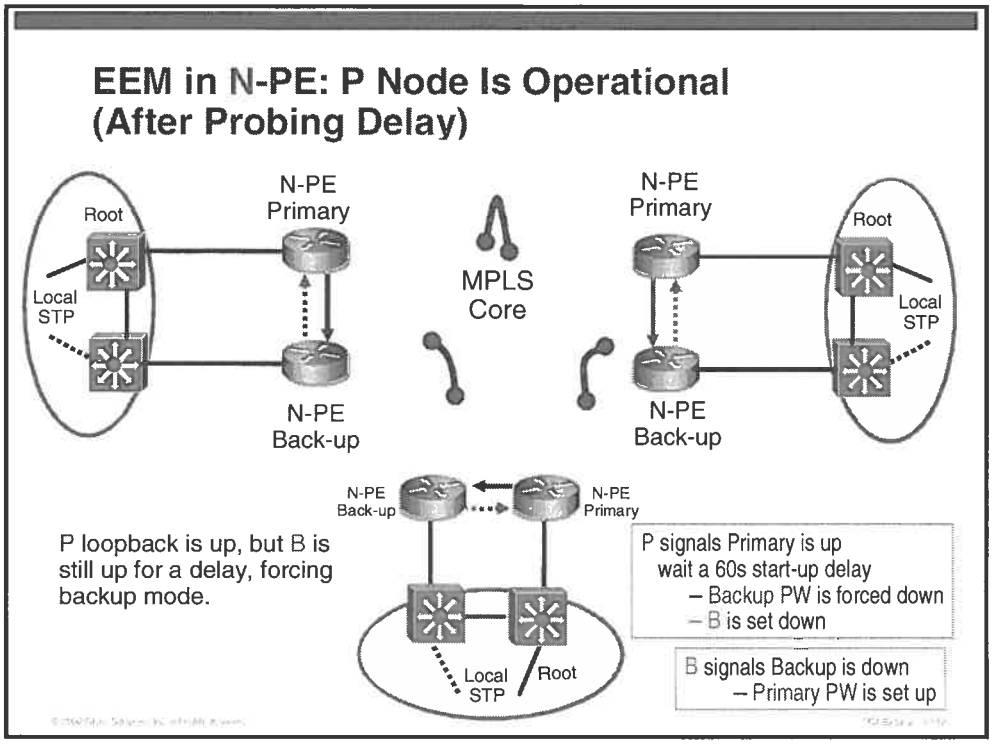
EEM in N-PE: P Node Recovering



© 2008 Cisco Systems, Inc.

https://t.me/learningnets

EEM in N-PE: P Node Is Operational (After Probing Delay)



https://t.me/learningnets

Reference Design Summary

Common Benefits:

- STP isolation across sites
- Redundancy (dual N-PE)
- L2 loop protection
- Following transport options are supported:
  - MPLS using VPLS
  - IP using VPLSoGRE

EEM-Based Benefits:

- No need to change STP design
- Scalable (using QinQ)

MST-Based Benefits:

- Active load balancing

Reference Design Summary

L2 extension tailored to the enterprise needs

Need	VPLS	EoMPLS	IP PW (L2TPv3)	OTV
Core Transparency	MPLS	MPLS	IP unicast	IP unicast & multicast
Multi-point Capable	Full mesh	No	No	Stateless
Loop prevention	Large STP	Large STP	Large STP	Built-in
Optimal traffic handling	Head-end replication Flooding	Head-end replication Flooding	Head-end replication Flooding	Native multicast No flood
Multi-homing	Complex	Complex	Complex	Built-in
Scalability	Poor	Poor	Poor	Good
Manageability	Many protocols Complex add/deletes	Many protocols Complex add/deletes	Many protocols Complex add/deletes	Single protocol Simple add/delete

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

This topic summarizes the key points that were discussed in this lesson.

Summary

- RSTP and MST are suitable for a Layer 2 Access Layer topology.
- A knowledge of traffic flows, oversubscription rates and redundancy requirements is important when selecting link types.
- Deployment of service modules influences the design of the Access Layer due to Layer 2 adjacency requirements.
- When deploying Layer 2 looped topologies, links must block to prevent traffic looping at Layer 2.
- Looped square topologies allow a single uplink to be deployed from each access switch while still providing redundant paths to the Aggregation Layer.
- Looped triangle is the most widely deployed and proven topology.

Summary (Cont.)

- In loop-free topologies, all uplinks may be forwarding but VLANs are restricted to access switch pairs.
- The loop-free U topology is not recommended if service modules are present.
- In a loop-free inverted U topology, there is only one uplink to the Aggregation Layer from an access switch. Failure of an aggregation switch leads to black-holing.
- In a Layer 3 Access Layer topology, it is possible to equal cost load balance over up to eight uplinks to the Aggregation Layer; however, access VLANs are restricted to pairs of access switches.
- Scalability considerations for the Access Layer include oversubscription rates and the types of switches to be deployed.

https://t.me/learningnets

Summary (Cont.)

- L2 extension options between multiple data centers are redundant, scalable and secure.
- The use of top-of-rack, end-of-row, and blade switches should be considered.
- Multiple models can be appropriate in the same data center.
- The selection of appropriate platforms and modules for use in the Access Layer is based on the choice of cabling models and on the bandwidth requirements of the design.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets