DCNID

Data Center Network Infrastructure Design

Volume 1

Version 2.0

Student Guide

Text Part Number: 97-2679-01

https://t.me/learningnets

Table of Contents

Volume 1

Course Introduction ***1***

Overview 1
Learner Skills and Knowledge 1
Course Goal and Objectives 2
Course Flow 3
Additional References 4
Cisco Glossary of Terms 5
Introductions 6

Data Center Design Models ***1-1***

Overview 1-1
Module Objectives 1-1

Data Center Business Objectives **1-3**

Overview 1-3
Objectives 1-3
Business Objectives 1-4
Data Center Definition, Solutions and Goals 1-4
Data Center Business Objectives and Design Criteria 1-6
Data Center High Availability Metrics for Service Availability 1-8
Scope of Outage Impact and Recovery Mechanisms 1-11
Data Center Design Trends 1-12
Consolidation in Data Centers 1-16
Server I/O Consolidation 1-17
Server and Server Farm Consolidation Drivers 1-18
Virtualizing Network Services 1-20
Business Solutions: Virtualization 1-20
Virtualized Services: Logical and Physical Data Center View 1-22
Data Center Automation 1-25
Summary 1-27

Data Center Networking Platforms and Modules **1-29**

Overview 1-29
Objectives 1-29
Cisco Catalyst 6500 Series Switches 1-30
Cisco Nexus Switching Portfolio for the Data Center 1-32
Cisco Nexus 7000 Series Data Center Class Switches 1-32
Cisco Nexus 5000 Series Switches 1-40
Top-of-Rack Consolidated I/O 1-45
Cisco Catalyst 6500 Series Switching Portfolio 1-47
Cisco Catalyst 6500 Series Modular Architecture 1-51
Cisco Catalyst 6500 Virtual Switching System 1-53
Virtual Switch Link (VSL) 1-53
Network System Virtualization 1-55
Security Services Modules: FWSM and IDSM-2 1-58
Firewall Service Module 1-58
Integrated Network Analysis 1-60
Cisco Network Analysis Modules (NAMs) 1-60
Application Control Engine 1-64
A New Generation of Application Delivery Solutions 1-64
Cisco Catalyst 4900M Series Switches 1-69
Data Center Switching Architecture 1-69
Cisco Blade Switches 1-73
Cisco Server Fabric Switches 1-85
Cisco InfiniBand Gateways 1-87
Fibre Channel Gateway 1-87
Ethernet Gateway 1-88

https://t.me/learningnets

Storage Networking 1-89
The MDS 9000 Platform 1-90
Optical Transport 1-95
Data Center Network Platforms: Optical 1-95
Summary 1-97
Data Center Environmental Requirements 1-99
Overview 1-99
Objectives 1-99
Environmental Requirements 1-100
Cisco Data Center Network Architecture Physical Facility 1-101
Green Data Center Initiatives 1-131
Summary 1-138
Cisco Nexus 7010 Switch 2-1
Overview 2-1
Module Objectives 2-1
Cisco Nexus 7010 Hardware Architecture 2-3
Overview 2-3
Objectives 2-3
Cisco Nexus 7010 Switch Supervisor Engine Architecture 2-4
Nexus 7010 Switch Supervisor Engine Architecture 2-4
Cisco Nexus 7010 Switch I/O Modules Architecture 2-14
32-Port 10 Gigabit Ethernet I/O Module Features 2-14
Cisco Nexus 7010 Switch Fabric Modules Architecture 2-23
Cisco Nexus 7010 Switch Forwarding Engine Architecture 2-24
Cisco Nexus 7010 Switch Fabric Scalable and Performance and Fabric Capacity 2-25
Cisco Nexus 7010 Switch Power Supply, Cooling System, and Cable Management 2-28
Summary 2-41
Cisco Nexus 7010 Software Architecture 2-43
Overview 2-43
Objectives 2-43
Cisco Nexus 7010 NX-OS Software Architecture Design 2-44
Cisco Nexus 7010 NX-OS Software Layer 2, Layer 3 and Routing Protocols 2-52
Layer 2 Packet Flow 2-60
Cisco Nexus 7010 NX-OS Software Licensing Model 2-108
Summary 2-111
Cisco Nexus 7010 Continual Availability 2-113
Overview 2-113
Objectives 2-113
Cisco Nexus 7010 Switch Continual Availability 2-114
Summary 2-152
Cisco Nexus 7010 Switch Network Management 2-153
Overview 2-153
Objectives 2-153
Cisco Nexus 7010 Switch Configuration Methods 2-154
Cisco Nexus 7010 Switch Management using DCNM 2-159
Cisco Nexus 7010 Switch Operating System Serviceability 2-170
Cisco Nexus 7010 Switch Operating System Manageability 2-193
Summary 2-202

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Security 2-203
Overview 2-203
Objectives 2-203
Cisco Nexus 7010 Switch Security Overview 2-204
Cisco Secure Data Center 2-204
Cisco Nexus 7010 Switch Integrated Security Features 2-206
Control Plane Policing (CoPP) 2-206
Summary 2-233
Cisco Nexus 7010 QoS Implementation 2-235
Overview 2-235
Objectives 2-235
Cisco Nexus 7010 Packet Forwarding and VOQ 2-236
Cisco Nexus 7010 QoS Components and Features 2-242
QoS Goals 2-242
Data Center Network QoS 2-271
Disaster Recovery Center (DRC) Topology 2-271
Summary 2-280
Cisco Nexus 7010 Switch Positioning in the Data Center 2-281
Overview 2-281
Objectives 2-281
Cisco Nexus 7010 Switch Data Center Positioning 2-282
Characteristics of the Data Center infrastructure 2-282
Summary 2-290

https://t.me/learningnets

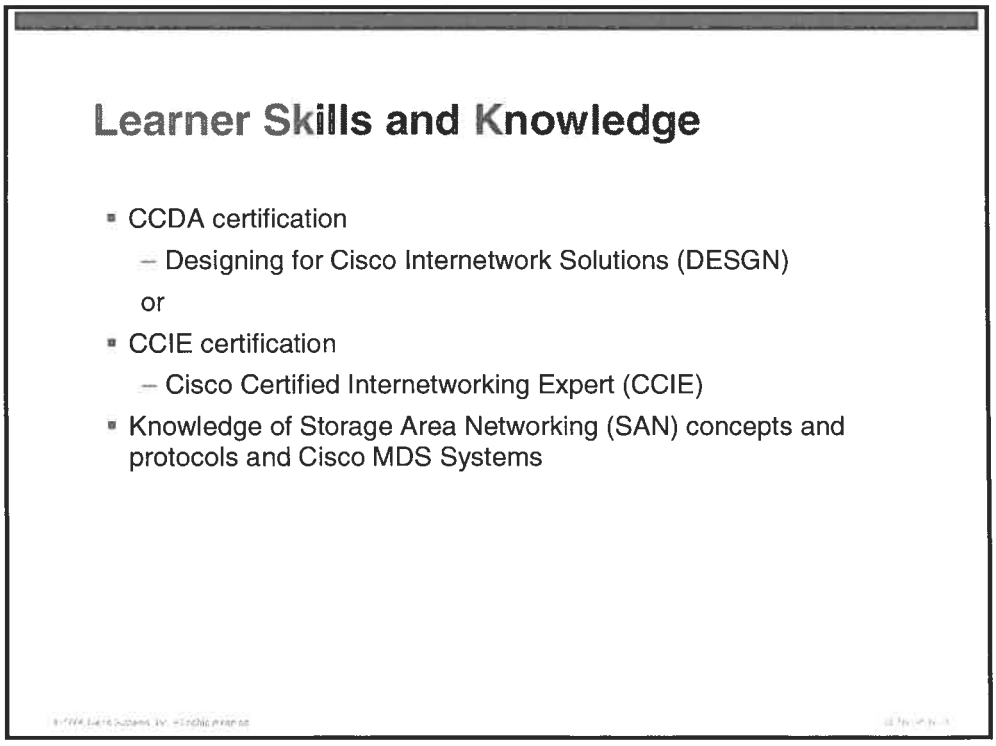
Course Introduction

Overview

After completing this course, you will be able to create a Data Center network design that optimizes availability, scalability, performance, and security, using the Nexus 7000 , Nexus 5000, Catalyst 6500, Catalyst 4900M, Catalyst Blade Switches, Firewall Services Module (FWSM), Intrusion Detection/Protection Services Module(IDSM/IPSM), Network Analysis Module (NAM),Application Control Engineer Module (ACEM) and ACE Appliance.

Learner Skills and Knowledge

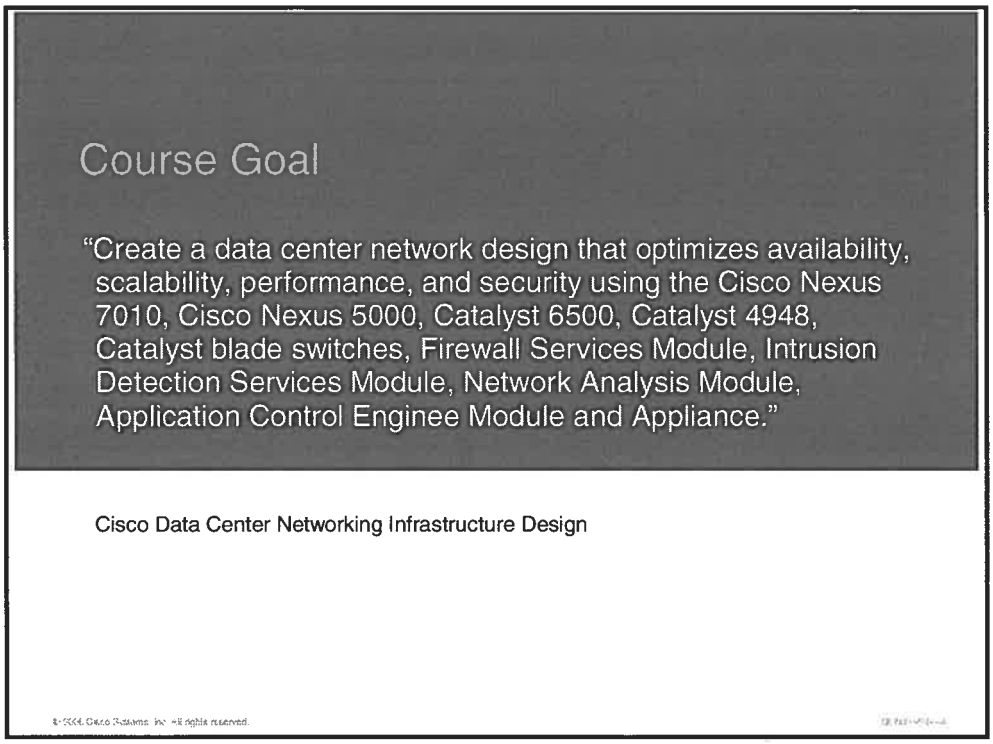
This subtopic lists the skills and knowledge that learners must possess to benefit fully from the course. The subtopic also includes recommended Cisco learning offerings that learners should first complete to benefit fully from this course.



https://t.me/learningnets

Course Goal and Objectives

This topic describes the course goal and objectives.



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

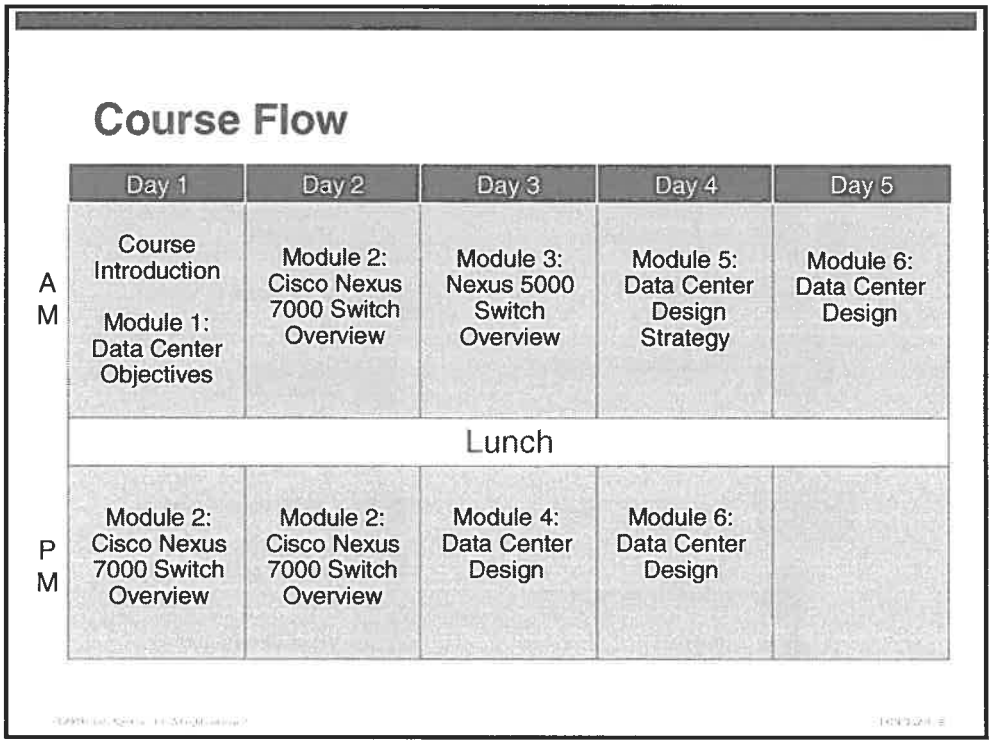
- Create a Data Center network design that optimizes availability, scalability, and performance for that environment
- Create an access layer design that optimizes data flows
- Create an aggregation layer design that optimizes data flows
- Create a core layer design that optimizes data flows
- Optimize a Data Center design to achieve high availability
- Design a security implementation using Nexus 7010, Nexus 5000 and Catalyst 6500 service modules
- Implement the Catalyst 6500 NAM to provide centralized network monitoring services
- Implement Data Center Server Load Balancing using Application Control Appliance or Catalyst 6500 Module.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Course Flow

This topic presents the suggested flow of the course materials.

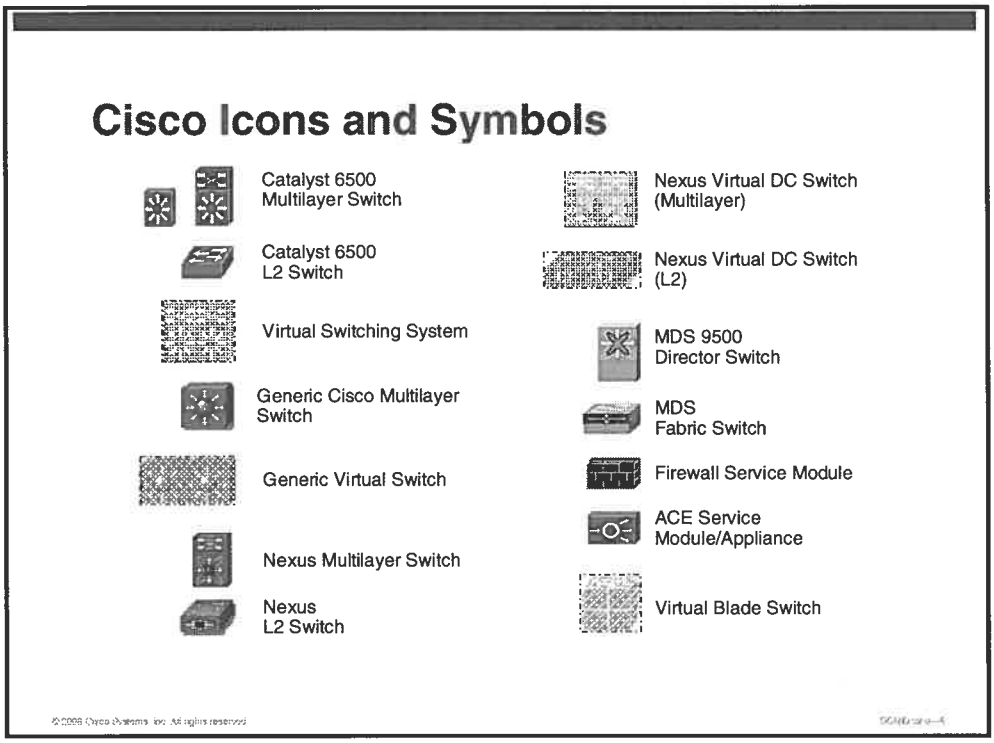


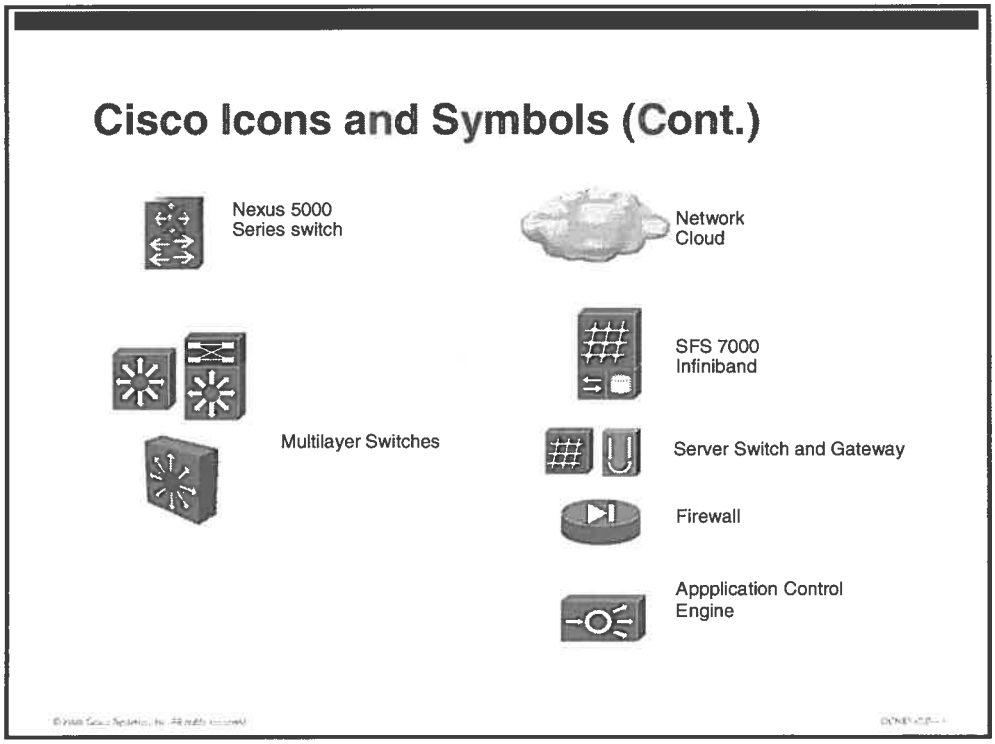
The schedule reflects the recommended structure for this course. This structure allows enough time for the instructor to present the course information and for you to work through the lab activities. The exact timing of the subject materials and labs depends on the pace of your specific class.

https://t.me/learningnets

Additional References

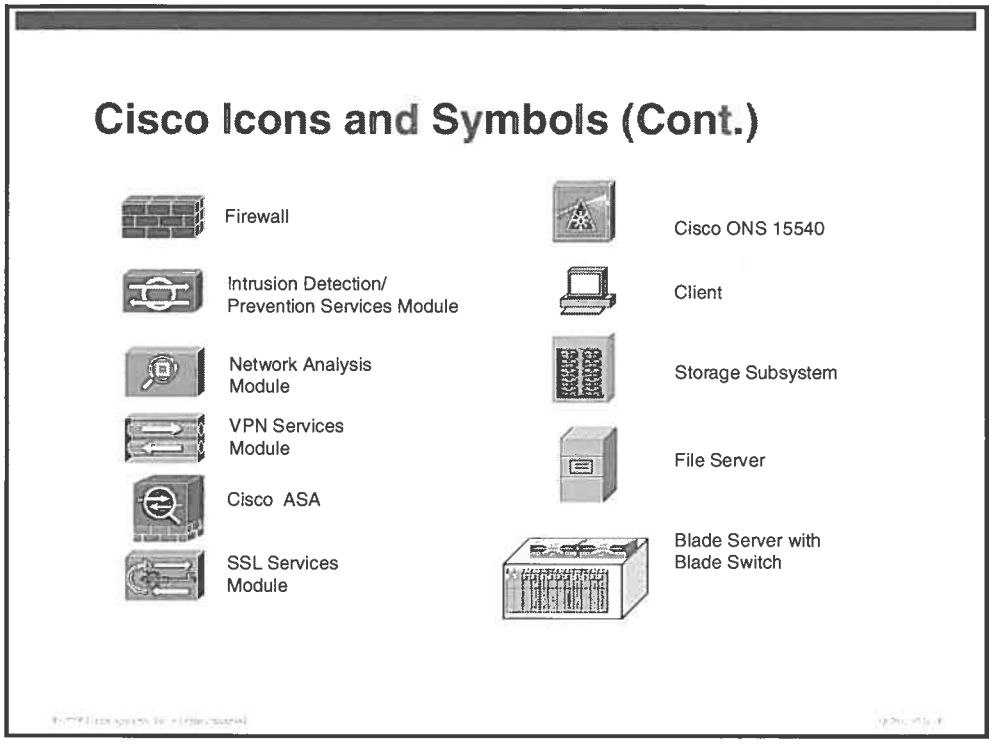
This topic presents the Cisco icons and symbols that are used in this course, as well as information on where to find additional technical references.





 © 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Glossary of Terms

For additional information on Cisco terminology, refer to the Cisco Internetworking Terms and Acronyms glossary at http://www.cisco.com/univercd/cc/td/doc/cisintwk/ita/index.htm.

https://t.me/learningnets

Introductions

Participant Responsibilities

Learner responsibilities:

- Complete prerequisites
- Participate in Case Studies
- Ask questions
- Provide feedback



© 2008 Cisco Systems, Inc. All rights reserved.

Participant Introductions

- Your name
- Your company
- Prerequisite skills
- Brief history
- Objective



https://t.me/learningnets

Module 1

Data Center Design Models

Overview

This module introduces concepts and basics of designing Data Center (DC) network architecture, and the challenges that designers are facing today.

Module Objectives

Upon completing this module, you will be able to describe the Data Center (DC) network architecture design basics and challenges. This ability includes being able to meet these objectives:

- Describe the key Data Center concerns and objectives
- Describe the Data Center networking components
- Describe the Data Center environmental requirement

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 1

Data Center Business Objectives

Overview

This lesson introduces concepts and basic characteristics of the components of the data center. With this knowledge, you can map these concepts and characteristics to the network design model so that the data center components can be accommodated within the data center network design.

Objectives

Upon completing this lesson, you will be able to discuss the key objectives for the data center architecture. This ability includes being able to meet these objectives:

- Define a data center and the goals and solutions provided by data centers
- Discuss the key drivers and objectives that influence data center design
- Discuss high-availiability metrics
- Relate environmental characteristics, such as power and cooling, to scaling a data center
- Describe the drivers for virtualization, consolidation, service integration and automation that are applied to data center components
- Describe the characteristics of a green data center

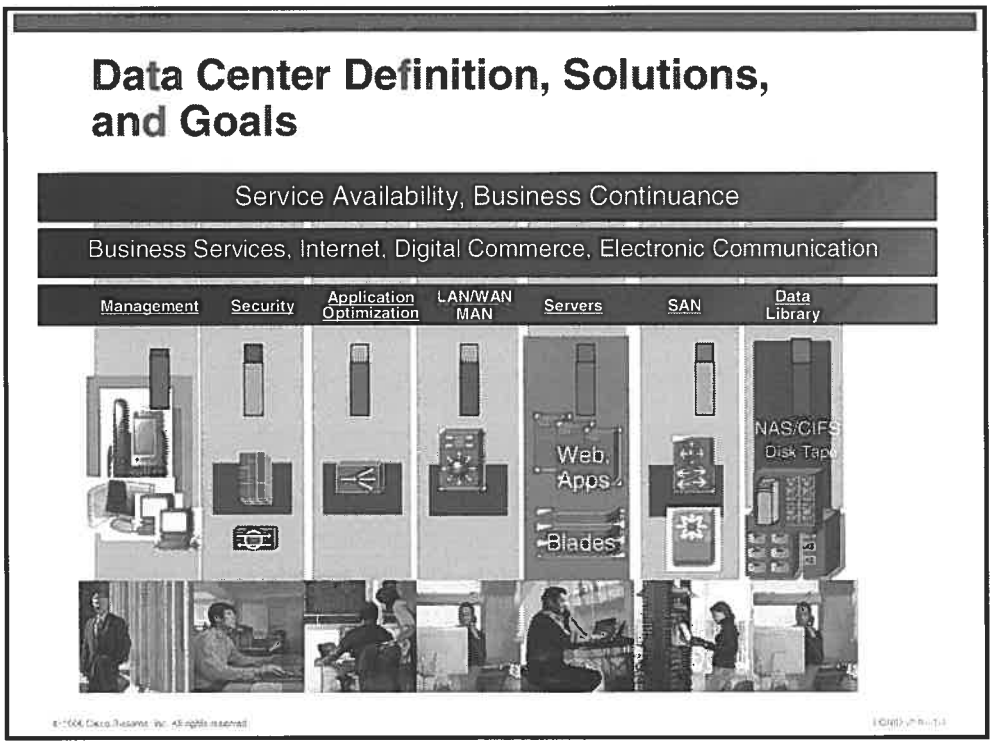
https://t.me/learningnets

Business Objectives

This topic discusses the evolving data center and the challenges that are influencing the future of the data center network design.

Data Center Definition, Solutions and Goals

This topic discusses the evolving data center and the challenges that are influencing the future of the data center network design.



Data Center Definition

A data center is a centralized or geographically distributed group of departments housing the computing systems and their related storage equipment or data libraries. A data center has controlled centralized management that enables an Enterprise to operate according to business needs.

Computing resources include web applications and servers; file servers, print servers, messaging servers, software application servers and the operating systems running on them; storage subsystems and network infrastructure with data networks and Storage Area Networks (SAN).

Applications include financial and human resources, e-commerce, and business-to-business applications. Network operation applications today include application optimization protocols, Network Time Protocol (NTP), Dynamic Host Configuration Protocol (DHCP), Simple Network Management Protocol (SNMP), TFTP, File sharing protocols, Network File System and Common Internet File System Protocols (NFS/CIFS), IP telephony, IP video conferencing, routing and switching protocols, security protocols, network monitoring and others.

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Solutions

A data center infrastructure is an essential component that supports Internet services, digital commerce, electronic communications and other business services and solutions.

Data Center Goals

A data center goal is to sustain the business functions and operations, along with flexibility for future data center changes. A data center network must be flexible and must support non-disruptive scalability of applications and computing resources to adopt the infrastructure for future business needs.

Business Continuance Definition

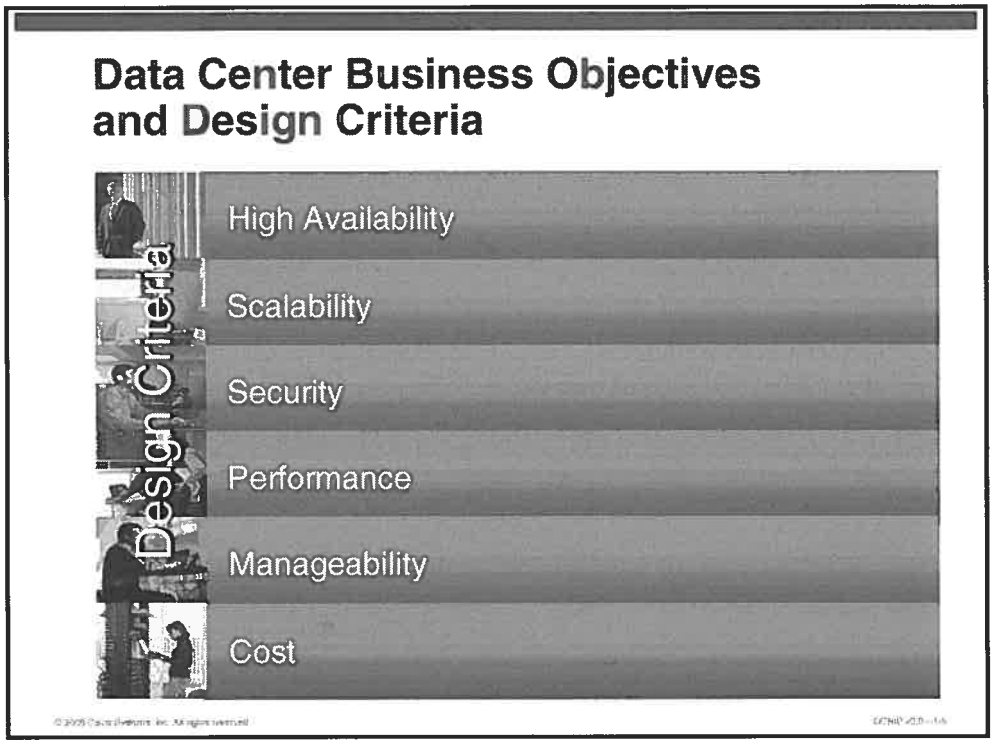
Business continuity is the ability to adapt and respond to risks as well as opportunities in order to maintain continuous business operations. There are four primary aspects of business continuity:

- High availability (disaster tolerance)
- Continuous operations
- Disaster recovery
- Disaster tolerance

https://t.me/learningnets

Data Center Business Objectives and Design Criteria

This topic discusses the data center business objectives and design criteria.



The benefits provided by a data center include business goals that generate the following data center business objectives. data center network design and architecture are dictated by business continuance and application requirements such as:

High Availability

Data center networking architecture must follow compliances and regulations needed to address regulatory issues and enable business resilience. Data centers need to implement an infrastructure that can protect and recover applications, communications, and information, and also provide uninterrupted access. Data center architecture must be redundant and highly available to support delivery of the desired recovery point objective (RPO) and recovery time objective (RTO). RTO determines how long it takes for a certain application to recover from a failure, and RPO determines to which point (in backup and data) the application can recover. These objectives also outline the requirements for disaster recovery and business continuity. If these needs are not met in a deterministic fashion, an enterprise carries significant risk in terms of its ability to deliver on the desired service level agreements (SLAs).

Scalability

Data center network infrastructure must be highly scalable in almost every sense, with minimized impact on data center operation. Effective scalability of data center servers, of data center storage capacity, of network capacity and of capacity of other computing resources can be achieved using data center consolidation, virtualization and automation mechanisms. In addition, data center network architecture must support seamless incorporation of new, advanced technologies into the existing data center architecture.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Security

Data center design must provide physical and logical security. All critical components of server and data security are within the data center itself. All entry points to the data center must be controlled and monitored. Data center servers, storage devices and fabric operation can be attacked over the network. Almost all of an organization's valuable information is stored online. Therefore, it is mandatory to protect data and servers from people with malicious intent using different restricted access policies, authentication and authorization mechanisms.

Performance

Data center performance must meet application requirements and enable service availability or prevent service degradation, loss of data and data center downtime. Performance parameters such as fabric throughput, utilization, latency, link capacities, and efficiency must be taken into consideration when designing data center networks.

Manageability

Data center networking infrastructure should have a centralized-management infrastructure and department, independent of an application networking infrastructure as much as possible. It should be able to manage, monitor, and access the data center networking infrastructure, in order to predict and prevent potential failures and in order to react to failures in data center operation.

Cost (Total Cost of Ownership, Management and Deployment)

The data center infrastructure, deployment and operations model must be economically sustainable. Operational expenses (OpEx) and Capital expenditures (CapEx) must be controllable and minimized. Cost of management is a significant part of the overall expenses and should be automated to minimize management expenses.

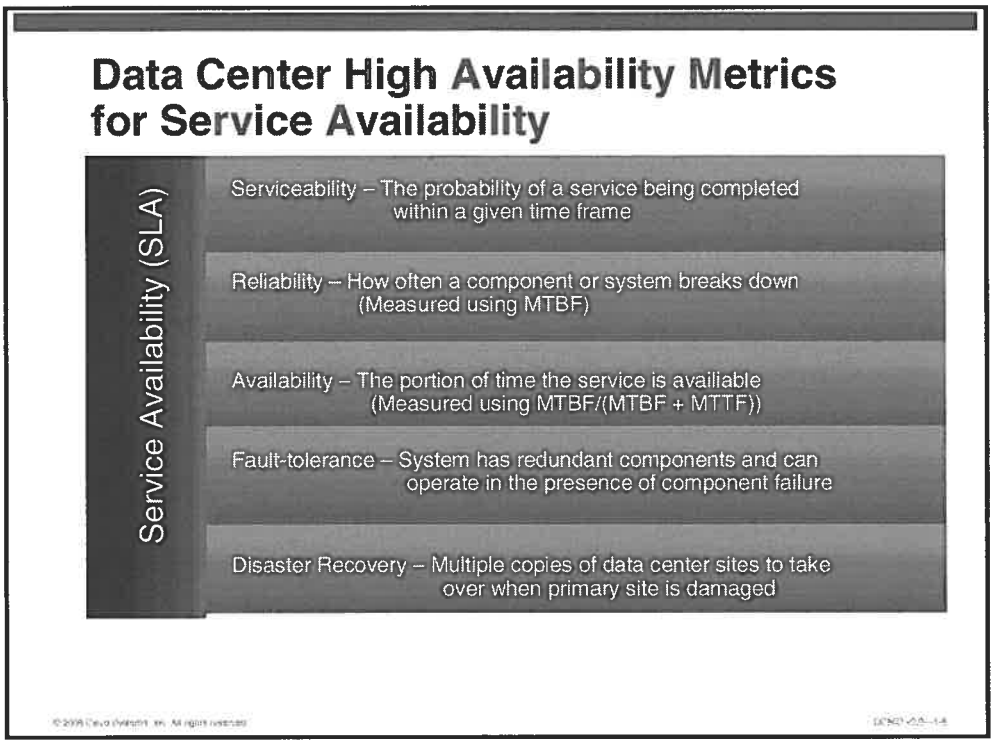
The cost of the recovery mechanism is increasing with the increasing cost of disruption, with the increased amount of lost data and with the increased disruption in the data center operation upon recovery.

A data center design that doesn't consider these risks might not be able to sustain the expected service level agreements (SLAs). Data center downtime, service degradation or inability to roll new services implies that SLAs are not met, which leads to a loss of access to critical resources and an impact on normal business operation. The impact can be increased response time or loss of data.

https://t.me/learningnets

Data Center High Availability Metrics for Service Availability

This topic describes the data center high availability metrics.



Protecting against failure is expensive, as is downtime. It is important to identify the most serious causes of service failure and to build cost-effective safeguards against them. A high degree of component reliability and data protection with redundant disks, adapters, servers, clusters and disaster recovery decreases the odds of service outages. Data center architecture providing service availability is a combination of several different levels of data center high availability features, and depends on the following:

Serviceability

Serviceability is the probability of a service being completed within a given time window. For example, if a system has serviceability of 0.98 for 3 hours, then there is a 98% probability that the service will be completed within 3 hours. In an ideal situation, a system can be serviced without any interruption of user support.

Reliability

Reliability represents the probability of a component or system not encountering any failure over a time span. The focus of reliability is to make the data center system components unbreakable. Reliability is a component of high availability that measures how rarely a component or system breaks down, and is expressed as the mean time between failures (MTBF). For example, a battery may have a useful life of 10 hours, but its MTBF is 50,000 hours. In a population of 50,000 batteries, this translates into one battery failure every hour during its 10-hour life span. Mean time to repair (MTTR) is the average time required to complete a repair action. A server with 99% reliability will be down for 3.65 days every year.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

If we have a system with 10 components where each component can fail independently but have 99% reliability, then the reliability of the entire system is not 99%. The entire system reliability is 90.44% (0.99 to the tenth power). This translates to 34.9 days of downtime in one year. Hardware reliability problems cause only 10% of the downtime. Software, human, or process-related failures comprise the other 90%.

Availability

Availability is the portion of time that an application or service is available for productive work. A more resilient system results in higher availability. An important decision to consider when building a system is the required availability level, which is a compromise between the downtime cost and the cost of the high availability configuration. Availability measures the ability of a system or group of systems to keep the application or service up and running. Designing for availability assumes that the system will fail and that the system is configured to mask and recover from component-to-server failures with minimum application outage. Availability could be calculated as follows:

Availability = MTBF/(MTBF + MTTR) or Availability = Uptime/(Uptime + Downtime)

To achieve the property of availability requires either building very reliable components (high MTBF) or designing components and systems that can rapidly recover from failure (low MTTR). As downtime approaches zero, availability approaches 100%.

Fault-tolerance

Fault-tolerant systems are systems that have redundant hardware components and can operate in the presence of individual component failure. Several components of the system have built-in component redundancy. Clusters are also examples of a fault-tolerant system. Clusters can provide uninterrupted service despite node failure. If a node running on one or more applications fails, one or more nodes in the cluster takes over the applications from the failed server. A fault-tolerant server has a fully replicated hardware design that allows uninterrupted service in the event of component failure. The recovery time or performance loss caused by a component failure is close to zero and information and disk content are preserved. The problem with fault-tolerant systems is that the system itself is a single point of failure.

Disaster Recovery (DR)

Disaster recovery is the ability to recover a data center at a different site if a disaster destroys the primary site or otherwise makes the primary site inoperable. Disaster, in the context of online applications, is an extended period of outage of mission-critical service or data caused by events such as fire or attacks that damage the entire facility. A disaster recovery solution requires a remote, mirrored (Backup/Secondary data center) site where business and mission-critical applications can be started within a reasonable period of time after the destruction of the primary site.

Setting up a new, off-site facility with duplicate hardware, software, and real-time data synchronization enables organizations to quickly recover from a disaster at the primary site. The data center infrastructure must deliver the desired Recovery Point Objective (RPO) and Recovery Time Objective (RTO). RTO determines how long it takes for a certain application to recover, and RPO determines to which point (in backup and data) the application can recover. These objectives also outline the requirements for disaster recovery and business continuity. If these requirements are not met in a deterministic fashion, an enterprise carries significant risk in terms of its ability to deliver on the desired service level agreements (SLAs). SLAs are

https://t.me/learningnets

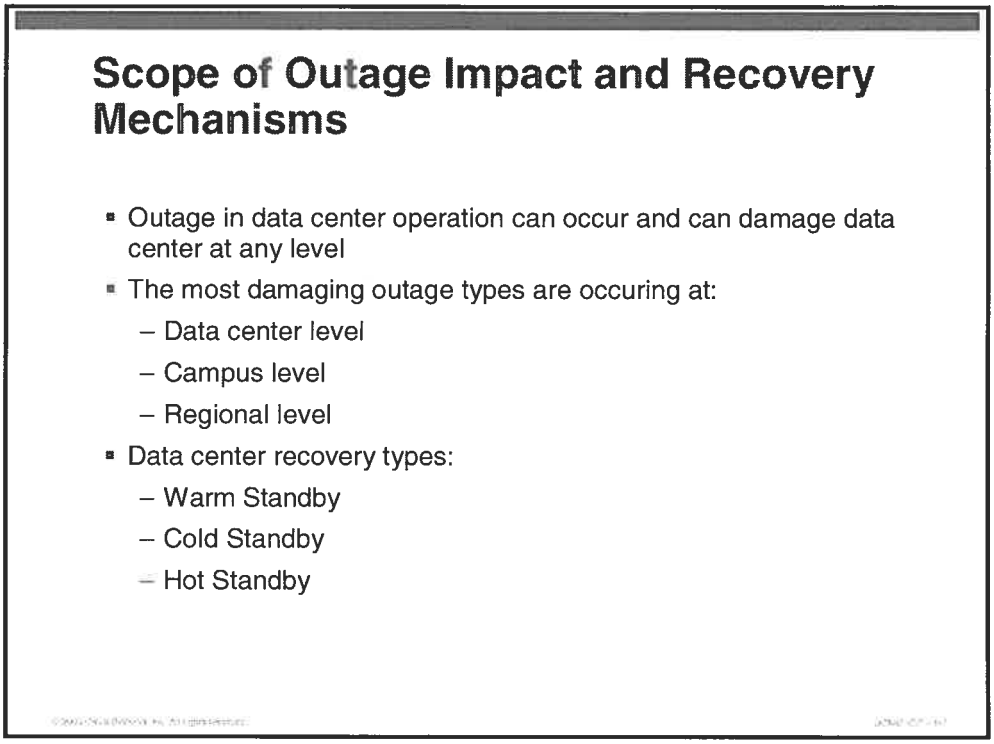
fundamental to business continuity. Ultimately, SLAs define your minimum levels of data center availability and often determine what actions will be taken in the event of a serious disruption. SLAs record and prescribe the levels of service availability, serviceability, performance support, and other attributes of the service, such as billing and even penalties, in the case of violation of the SLAs. For example, SLAs can prescribe different expectations in terms of guaranteed application response time, such as 1 second, 0.5 second or 0.1 second; guaranteed application resource allocation time, such as 1 hour or automatic; and guaranteed data center availability, such as 99.999%, 99.99%, or 99.9%. Higher levels of guaranteed availability imply higher SLA charges.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Scope of Outage Impact and Recovery Mechanisms

This topic discusses the types and scopes of outages in data center networks and recovery mechanisms.



There are different types of outages that might be expected and might impact the data center functions and operation. Typically, they are classified based on the scope of the outage impact:

Outage with an Impact at the Data Center Level

An outage of this type is an outage of a system or a component such as hardware or software. These types of outages can be recovered using reliable, resilient and redundant data center components, using fast routing and switching reconvergence, and stateful module/process failovers.

Outage with an Impact at the Campus Level

This type of outage impacts a building or an entire campus. Fire or loss of electricity can cause damage at the Campus level and can be recovered using redundant components such as power supplies and fans, or by using the secondary data center site or Power over Ethernet (PoE).

Outage with an Impact at the Regional Level

Outages that impact a region, such as earthquakes, widespread power outage, flooding or tornados. Such outages can be recovered using geographically-dispersed, standby data centers that use global site selection and redirection protocols to seamlessly redirect user requests to the secondary site.

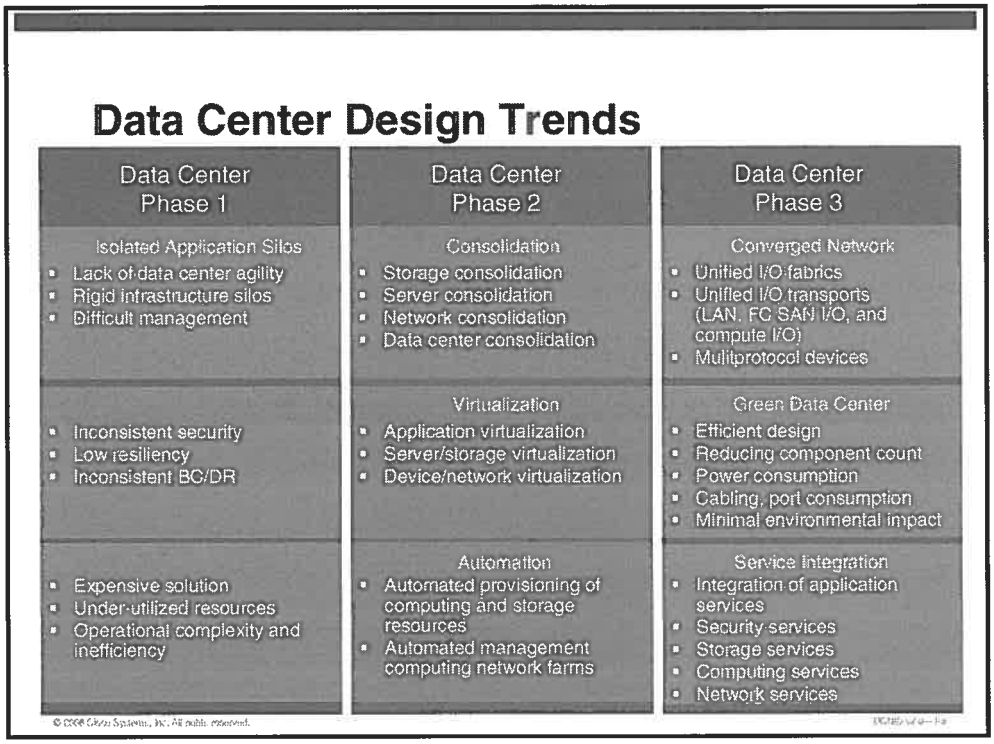
Data Center Recovery Types

Different types of data center recovery provide different levels of service and data protection, such as Cold Standby, Warm Standby, Hot Standby, Immediate Recovery, Continuous Availability, Continuous Operation, Gradual Recovery, and Back-Out Plan.

https://t.me/learningnets

Data Center Design Trends

This topic discusses the main data center design trends impacting the data center network architecture.



Data center trends that have impacted the data center architecture and design can be summarized using the following most significant phases and stages:

Isolated Application Silos

Data centers are about servers and applications. The first data centers were mostly mainframe, glass house, raised-floor structures that housed the computer resources, as well as the intellectual capital (programmers and support staff) of the enterprise. Over the past decade, most data centers have evolved on an ad-hoc basis. The goal was to provide the most appropriate server, storage, and networking infrastructure that supported specific applications. This strategy led to data centers with stove-pipe architectures or technology islands that were difficult to manage or adapt to changing environments.

There are many server platforms in present-day data centers, all designed to deploy a series of applications. For example:

- IBM mainframe applications
- E-mail applications on Microsoft Windows servers
- Business applications on IBM AS/400 servers
- Enterprise resource planning (ERP) applications on UNIX servers
- R&D applications on Linux servers

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

In addition, a broad collection of storage silos exists to support these disparate server environments. These storage silos can be in the form of integrated, Direct-Attached Storage (DAS), Network-Attached Storage (NAS), or small Storage Area Network (SAN) islands.

This silo approach has led to underutilization of resources, difficulty in managing these disparate complex environments, and difficulty in applying uniform services such as security and application optimization. It is also difficult to implement strong consistent disaster-recovery procedures and business continuance functions.

Consolidation

Consolidation of storage, servers, and networks has enabled centralization of data center components, any-to-any access, simplified management, and technology convergence, and consolidation has reduced the cost of data center deployment.

Virtualization

Virtualization is creation of another abstraction layer that separates physical from logical characteristics and enables further automation of data center services. Almost any component of a data center can now be virtualized: storage, servers, networks, computing resources, file systems, file blocks, tape devices, and so on. Virtualization in a data center enables creation of huge resource pools, increased, efficient, and more flexible resource utilization, and automated resource provisioning, allocation, and assignment to applications. Virtualization represents the foundation for further automation of data center services.

Automation

Automation of data center services has been made possible by consolidating and virtualizing data center components. The advantages of data center automation are automated dynamic resource provisioning, automated Information Life Cycle Management and Data Center management. Computing and networking resources can be automatically provisioned whenever needed. Other data center services can also be automated. For example, automated data migration, automated mirroring, and volume management functions. Monitoring data center resource utilization is a necessary condition for an automated data center environment.

Converged Network

Converged networks promise the unification of various networks and single all-purpose communication application. Converged networks potentially lead to reduced IT cost and increased user productivity. Unified data center fabric is based on a unified I/O transport protocol, which could potentially transport SAN, LAN/WAN, and clustering I/Os. Most protocols today tend to be transported across a common unified I/O channel and common hardware and software components of data center architecture.

Green Data Center

A green data center is a data center in which the mechanical, lighting, electrical and computer systems are designed for maximum energy efficiency and minimum environmental impact. The construction and operation of a green data center includes advanced technologies and strategies, such as minimizing the footprints of the buildings, using low-emission building materials, sustainable landscaping, and waste recycling, and installing catalytic converters on backup generators. The use of alternative energy technologies such as photovoltaics, heat pumps, and evaporative cooling is also being considered.

https://t.me/learningnets

Building and certifying a green data center or other facility can be expensive initially, but long-term cost savings can be realized on operations and maintenance. Another advantage is the fact that green facilities offer employees a healthy, comfortable work environment. There is growing pressure from environmentalists and, increasingly, the general public, for governments to offer green incentives. That is, monetary support for the creation and maintenance of ecologically responsible technologies.

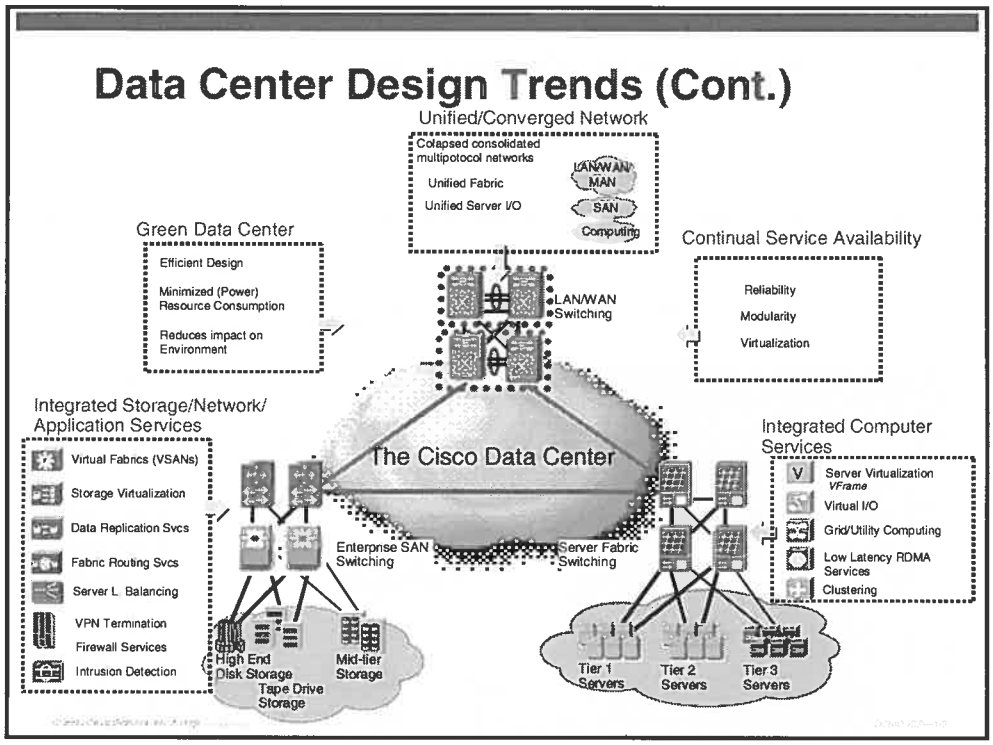
Today, green data centers provide an 85% power reduction using virtualized, integrated modules. Rack space is being saved with virtualized, integrated modules and additional savings are being derived from reduced cabling, port consumption and support cost.

Service Integration

The data center network infrastructure is integrating network intelligence and is becoming application-agnostic, security-agnostic, computing-agnostic, and storage-agnostic by integrating application services, security services, computing services, and storage services, among others..

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The next wave of data center deployments are focused on intelligent, automated, converged data center network architecture that provides continual service availability and integrates application, network, storage and computing services at maximum energy efficiency and minimum environmental impact.

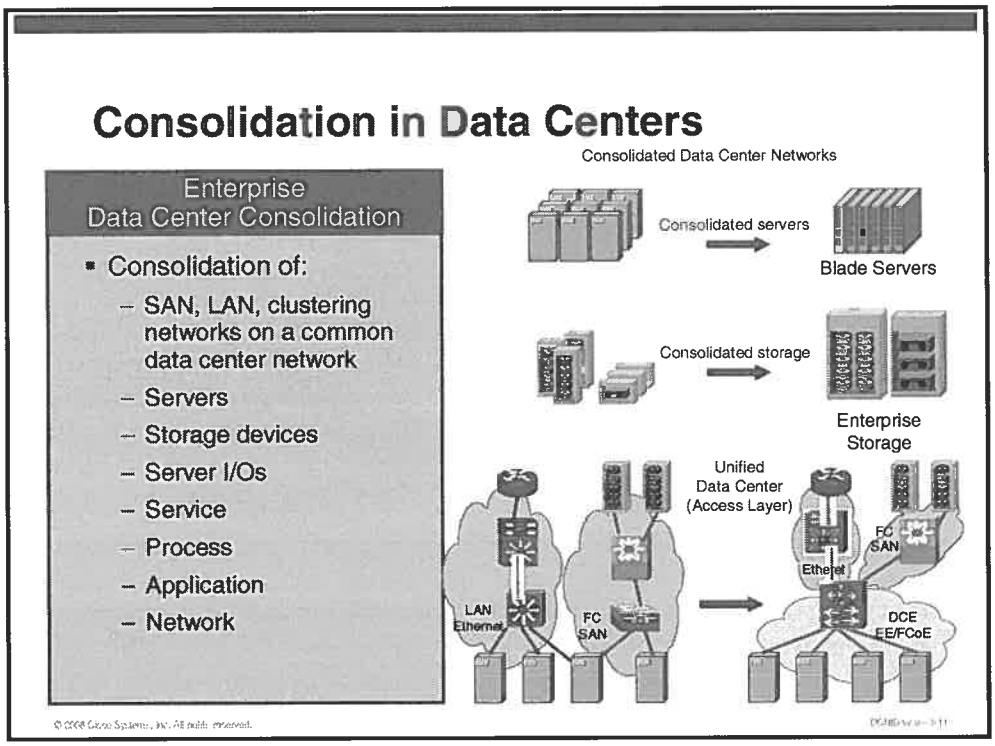
A data center converged network is an intelligent, automated, massive, parallel fabric that supports huge performance, deterministic latency for multi-service transport. A data center converged network is also very reliable and flexible because it utilizes modularity and virtualization of hardware and software components.

The new wave of data center architecture is based on consolidated hardware and software modularity, virtualized at almost every layer with integrated network services to provide continual service availability.

https://t.me/learningnets

Consolidation in Data Centers

This topic discusses some of the reasons behind consolidation and why more businesses are looking at these tools within the data center to improve resource utilization.



Consolidation is defined as the process of bringing together disconnected parts to make a single and complete whole. In the data center, it means replacing several small devices with a few highly capable pieces of equipment to provide simplicity.

The key reason for consolidation is the sprawl of equipment and processes required to manage the equipment. It is crucial to understand the functions of each piece of equipment before consolidating it. The key reasons for server, storage, server I/O, network, application and process consolidation are:

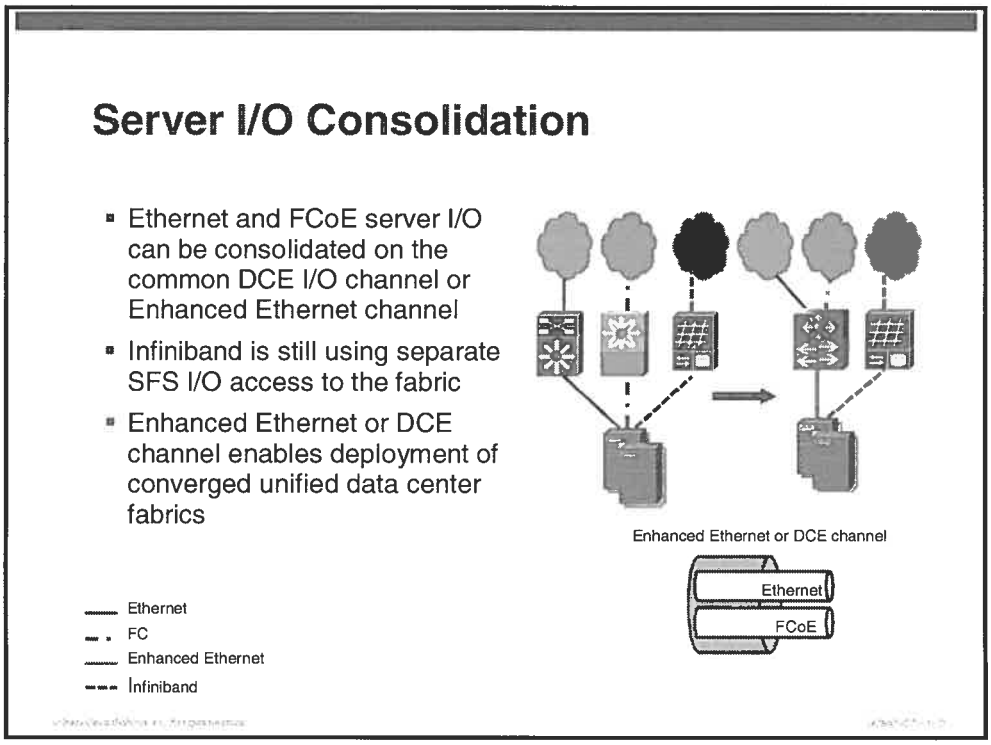
- Reduced number of servers, storage devices, networks, cables, and so on
- Increased usage of resources using resource pools (of storage and computing resources)
- Reduced centralized management
- Reduced expenses due to smaller number of equipment pieces needed
- Increased service reliability

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Server I/O Consolidation

This topic discusses the server I/O consolidation as a step towards converged data center networking infrastructure.



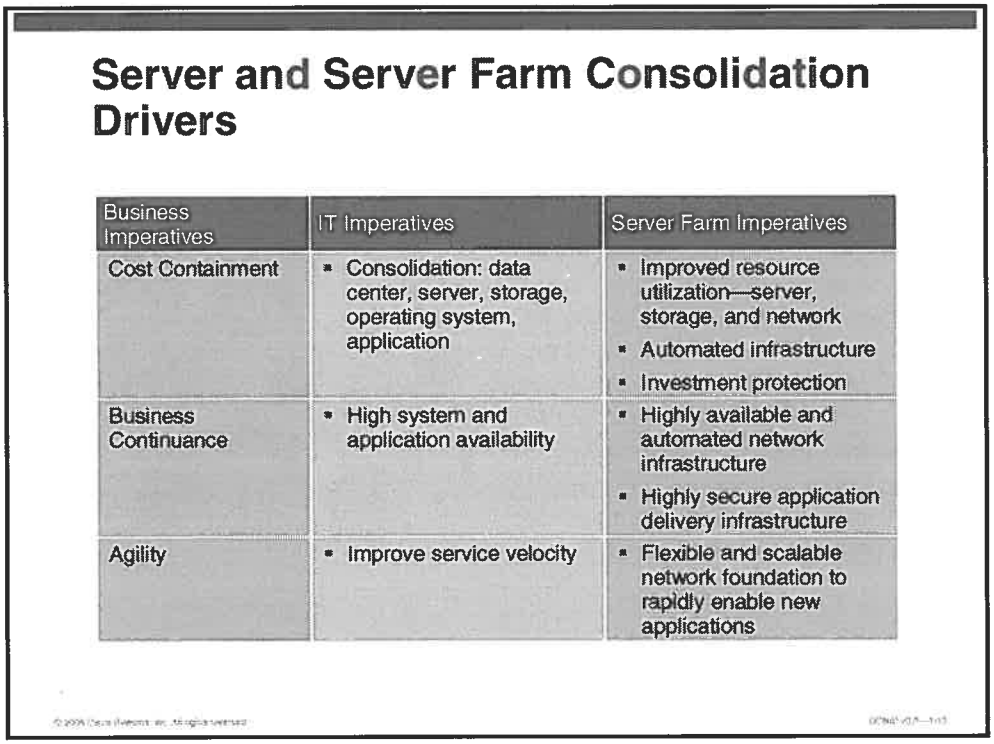
Server I/O consolidation has been attempted several times in the past by the introduction of Fibre Channel, InfiniBand and iSCSI protocols in order to carry storage I/Os, data I/Os and clustering I/Os across the same channel. All initial attempts have been unsuccessful.

Enhanced Ethernet is a new converged network protocol that is designed to transport unified data and storage I/Os. Key enabling technologies are PCI-Express and 10 Gigabit Ethernet. A growing demand for network storage is driving the network bandwidth demands. Server virtualization allows the consolidation of multiple applications on a server, thus driving the server bandwidth requirement of 10 Gb/s. 10 Gb/s Enhanced Ethernet uses Copper and Twinax cables with short distances (10m), but with lower cost, lower latency and lower power requirements than 10BaseT. FCoE and classical Ethernet can be multiplexed across the common physical Enhanced Ethernet connection. Enhanced Ethernet is sometimes referred to as Data Center Ethernet (DCE).

https://t.me/learningnets

Server and Server Farm Consolidation Drivers

This topic discusses the driving factors for server farm consolidation.



Data center server farms have very diverse requirements from an integration, performance, and services perspective. The more complex the IT infrastructure becomes, the more issues are raised:

- High operational costs; proliferation of disparate computer platforms across multiple data centers and branches; mainframe, UNIX, Windows
- Low average CPU utilization; Windows 5% to 15%, UNIX and Linux 10% to 30%
- Significant investment in mainframe
- Need for lower-cost high-performance data analysis and storage resources
- Need for faster server and application deployment; new applications and services, development environments, surges in demand
- High-density servers cause problems with cooling, server I/O, network connectivity

Many organizations are looking toward server consolidation that standardizes on blade centers or industry-standard 1U to 4U servers that can process information much faster and drive significant traffic volumes at line rate across Gigabit Ethernet and at significant fractions of 10 Gigabit Ethernet. Organizations are also adopting, or looking to adopt, virtual server techniques such as VMware that further increase server densities, albeit at a virtual level. Additionally, these same organizations must maintain heterogeneous environments that have different applications and server platforms (blade servers, midrange servers, mainframes, and so on) that need to be factored into the data center design.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Another important factor to consider is that the volume and value of data is increasing dramatically, and the network and associated systems must be able to scale to support these new loads. In the financial sector, techniques and technologies pioneered in high-performance computing (HPC), such as InfiniBand, are being deployed to improve systems performance. In many cases, organizations are deploying greater numbers of servers as they seek to gain competitive advantage and reduce costs. HPC clusters, once the domain of research and academic institutions, are now widely deployed in enterprise businesses for diverse reasons. In some cases, workloads traditionally performed on mainframes are now performed on clusters of industry-standard servers at a fraction of the cost of a mainframe.

As organizations look toward utility computing environments to provide business and operational agility, the requirements for network, as well as server, virtualization become critical. The ability to assign computer nodes to an application pool with the correct security, availability, and performance requires that the network is flexible enough to accommodate ad-hoc adds, moves, and changes, and also contains fault domains to mitigate against systemic failures. This requires that the data center network is designed to be partitioned, but also requires robust networking protocol implementations.

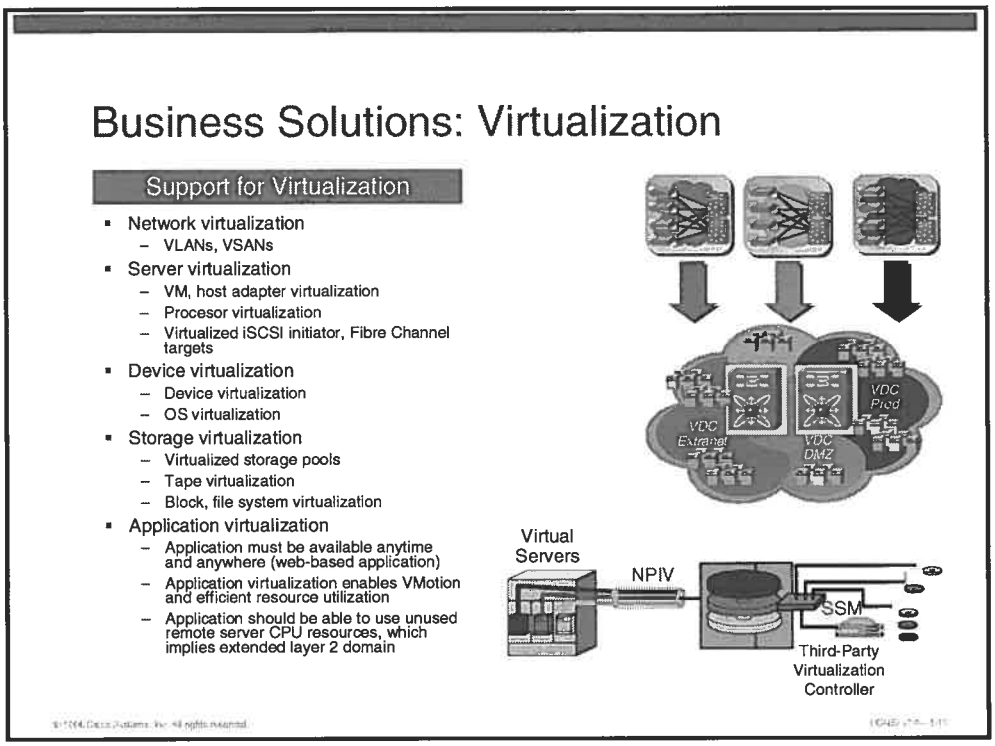
https://t.me/learningnets

Virtualizing Network Services

This topic describes how virtualization can be applied to the data center at various layers and the effect virtualization has on network and service availability.

Business Solutions: Virtualization

This topic explains the reasons for and benefits of implementing virtualization in data center networking infrastructure.



Virtualization delivers tremendous flexibility to build and architect data center solutions. Considering the diverse networking needs of different enterprises might require separation of a single user group or a separation of data center resources from the rest of the network for certain reasons. Separation tasks become complex when it is not possible to confine specific users or resources to specific areas in the network. When separation occurs, the physical positioning will no longer address the problem.

Network Virtualization

Network virtualization can address the problem of separation. Network virtualization also provides other types of benefits such as increasing network availability, better security, consolidation of multiple networks, segmentation of networks and increased network availability. Examples of network virtualization are VLANs and VSANs in Fibre Channel SANs. VLAN virtualizes Layer 2 segments, making them independent of the physical topology. This virtualization presents the ability to connect two servers to the same physical switch though they participate in different logical broadcast domains (VLANs). A similar concept represents VSAN in Fibre Channel SANs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Server Virtualization

Server virtualization enables physical consolidation of servers on the common physical infrastructure. Deployment of another virtual server is easy because there is no need to buy a new adapter and no need to buy a new server. For a virtual server to be enabled, only software is needed to be activated and configured properly. Server virtualization, thus, simplifies server deployment, reduces the cost of management and increases server utilization. WMW are and Microsoft are examples of companies that support server virtualization technologies.

Device Virtualization

Cisco Nexus 7010 and Catalyst 6500 switch support device virtualization or NX-OS virtualization. Virtual Device Context (VDC) represents the ability of the switch to enable multiple virtual and independent switches on the common physical switch to participate in data center networks. This feature provides various benefits to the application services, such as higher service availability, fault isolation, separation of logical networking infrastructure based on traffic service types, and flexible and scalable data center design.

Storage Virtualization

Storage virtualization is the ability to pool storage on diverse and independent devices into a single view. Features such as copy services, data migration and multiprotocol and multivendor integration can benefit from storage virtualization.

Application Virtualization

The web-based application must be available anytime and anywhere and it should be able to use unused remote server CPU resources, which implies an extended layer 2 domain. Application virtualization enables Vmotion and efficient resource utilization.

Common Goals

Common goals across virtualization techniques include:

- Driving utilization and reducing over-provisioning
- Isolation:

  Security must be effective enough to prevent any undesired access across the virtual entities that share a common physical infrastructure. Performance (QoS, SLA) must be provided at the desired level independently for each virtual entity. Faults must be contained.

- Management:

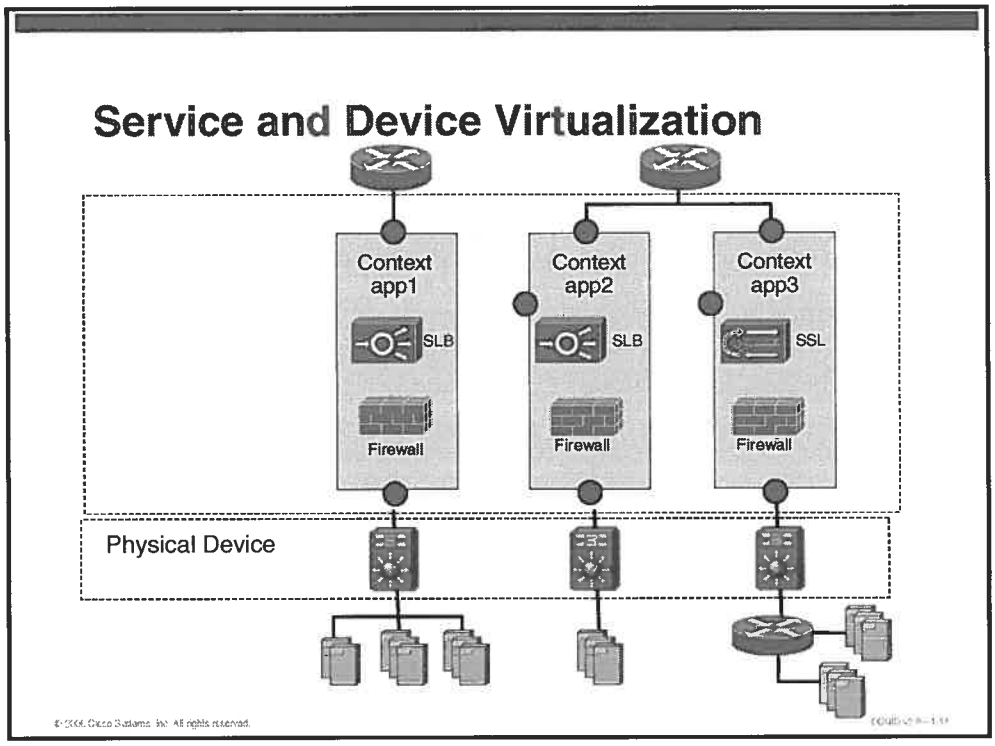
  Flexibly managing a virtual resource requires no hardware change in many cases. Individual administration for each virtual entity can be deployed using role-based access control (RBAC).

https://t.me/learningnets

Virtualized Services: Logical and Physical Data Center View

Virtualizing data center network services has changed the logical and physical data center network topology view.

The figure shows how virtual services can be created from the physical infrastructure, using features such as Virtual Device Context (VDC), VLANs and virtual SANs (VSANs). Virtual network services include virtual firewalls with the Firewall Services Module (FWSM), and virtual server load-balancing contexts with the Cisco Application Control Engine (ACE) and Virtual Intrusion Detection System (IDS).



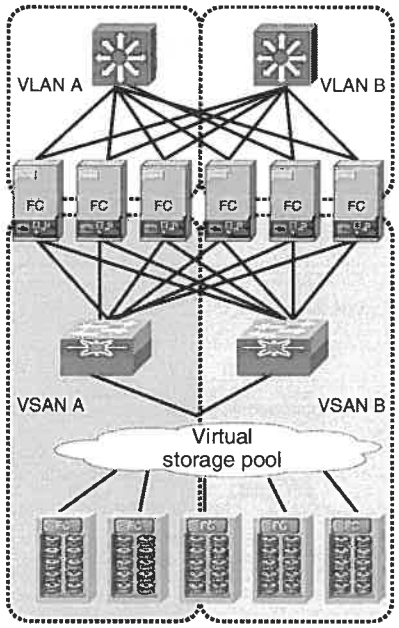
This slide depicts one physical service module that has been logically partitioned into several virtual service modules, and a physical switch that has been logically partitioned into several virtual device contexts. This partitioning reduces the number of physical devices that must be deployed and managed, yet still provides the same functionality that each device could provide.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Storage Virtualization

- Storage networks allow storage sharing and consolidation across multiple servers:
  - Leads to SAN proliferation (SAN islands)
  - Poor utilization
- Virtual SANs (VSANs) allow further consolidation of SAN islands:
  - Increased utilization
  - Easier to manage
- Storage device virtualization coupled with VSANs enables dynamic storage allocation

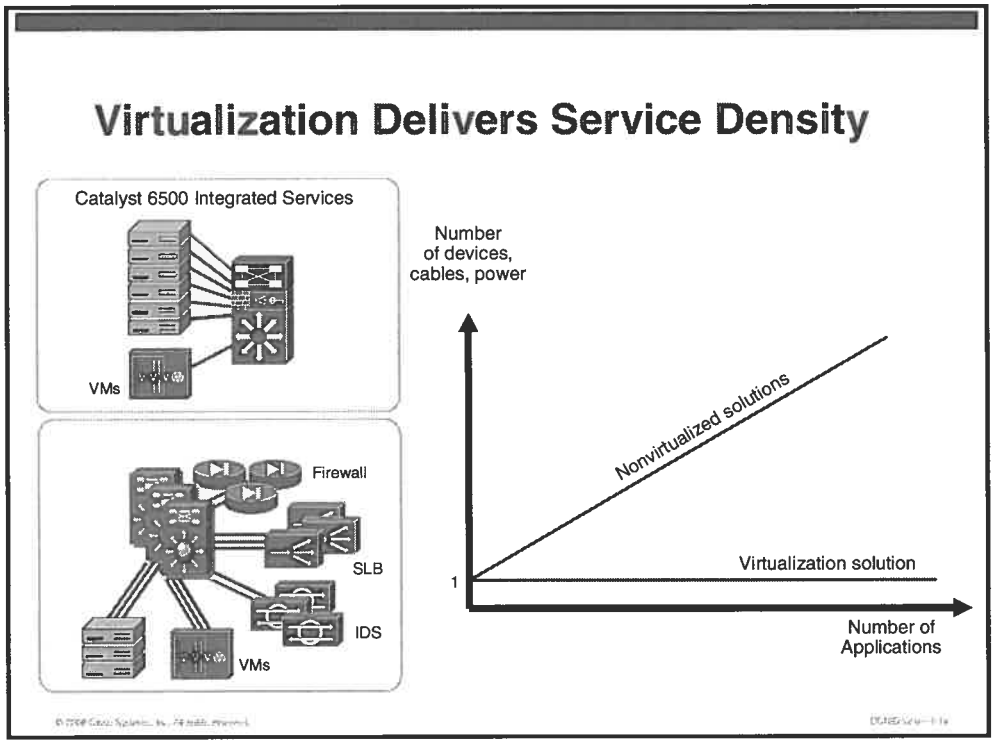


Data center storage virtualization starts with Cisco VSAN technology. Traditionally, SAN islands have been used within the data center to separate traffic on different physical infrastructures, providing security and separation from both a management and traffic perspective. To provide virtualization facilities, VSANs are used within the data center SAN environment to consolidate SAN islands onto one physical infrastructure while maintaining the separation from both a management and traffic perspective.

Storage virtualization also involves virtualizing the storage devices themselves. Coupled with VSANs, storage device virtualization enables dynamic allocation of storage. Taking a similar approach to the integration of network services directly into data center switching platforms, the Cisco MDS 9000 platform supports third-party storage virtualization applications on a MDS 9000 services module, reducing operational costs by consolidating management processes.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Network services virtualization enables higher service density by eliminating the need to deploy separate appliances for each application. The benefits of higher service density include:

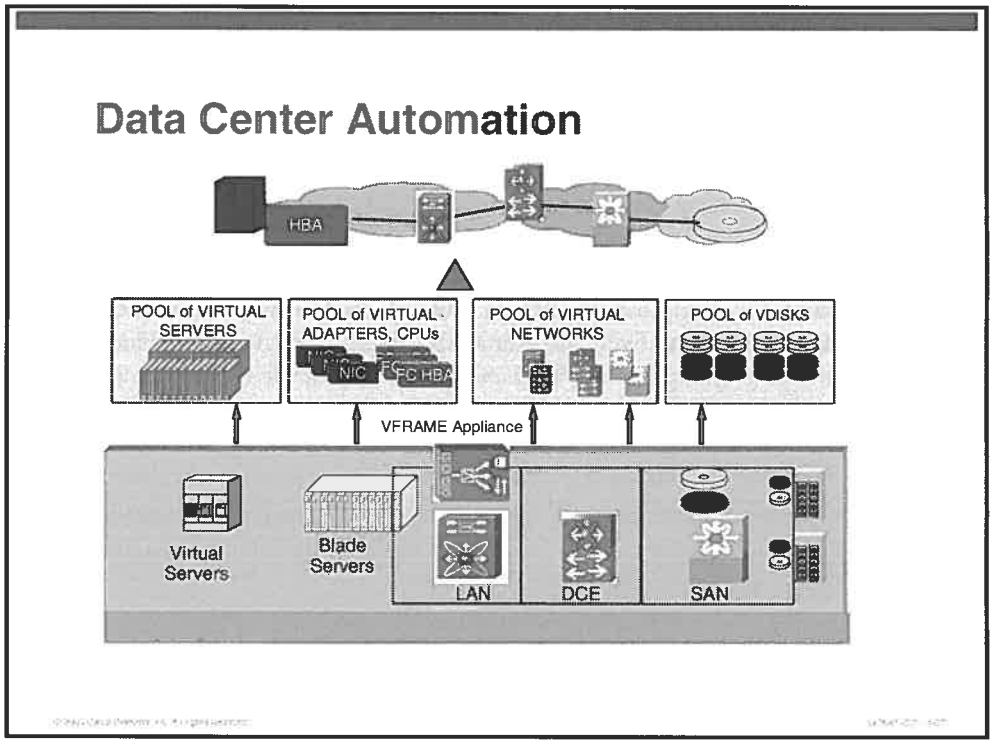
- Less power consumption
- Less rack space
- Reduced ports and cabling
- Simplified operational management
- Lower maintenance costs

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Automation

This topic discusses the reasoning behind the current trend of deploying automated data center networking architecture, and the impact of automation on data center networking design.



By pooling and virtualizing computing, storage and networking resources, several advantages can be obtained:

- Simplyfied management and troubleshooting
- Increased resource utilization resulting in reduced cost
- Data center service automation (makes deployment simplier and quicker)

Data center management tasks and operations can be automated based on consolidated pools of virtualized storage, computing and networking resources. Virtualization and consolidation enable creation of virtual server pools, virtual pools of adapters, pools of virtual processing units, virtual network pools and pools of virtual disks.

An appliance attached to an existing data center that monitors application processing, computing, storage and networking resource utilization can thus detect the missing processing power or the lack of application storage resources and can automatically react to it. An appliance can configure a virtual server, activate a virtual adapter, configure a server I/O channel and connect the channel across a virtual network to the dynamically allocated virtual disk, and then start the application on the newly allocated infrastructure. Cisco Data Center VFRAME appliance operation is an example of an appliance that is utilizing virtualized pools of resources in a very similar way.

https://t.me/learningnets

Besides policy-based resource provisioning—the ability to automatically increase the capacity of an existing application—data center service automation also provides the ability to roll out new applications that are critical to the success of many enterprises:

- E-commerce
- Customer relationship management (CRM)
- Supply chain management (SCM)
- Human resources (HR)

Policy-based provisioning is about providing an end-to-end IT service that is dynamically linked to business policies, allowing the ability to adapt to changing business conditions. For example, if a customer's order entry application suddenly experiences a surge in load, just allocating more CPU might not be enough; the application might also need additional storage, more network capacity, and even additional servers and new users to handle the increased activity. All of these changes must be orchestrated so that the dynamic allocation of multiple resource elements occurs seamlessly.

For servers, there are virtual server solutions like EMC VMware or modular blade server solutions. For storage, there are a variety of virtualization solutions; nearly every enterprise-class storage product supports some form of virtualization. Fabric virtualization is based on VLANs, VSANs, and on Virtual Device Context.

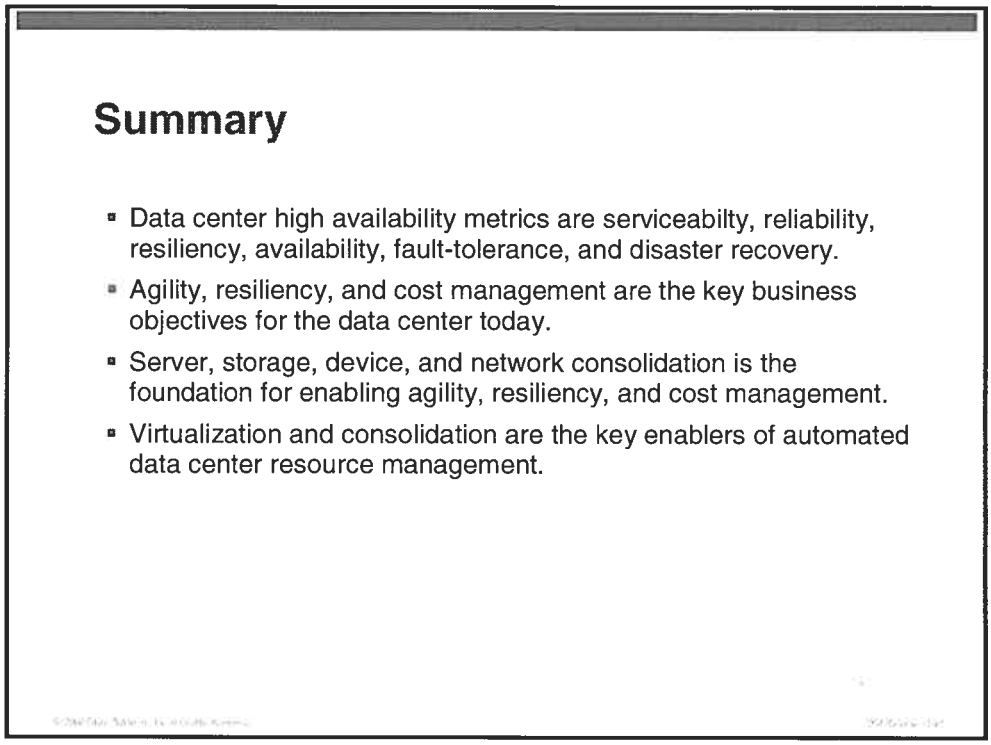
By reducing complexity, consolidation reduces management overhead and operational expenses (OpEx). At the same time, the ability to use resources is increased because resources are no longer locked up in silos, resulting in lower capital expenditures (CapEx).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

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



https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 2

Data Center Networking Platforms and Modules

Overview

This lesson introduces the various platforms and modules that are relevant to the data center network. It provides a brief overview of each component to provide a basis for selecting the correct hardware for the various layers of the data center architecture.

Objectives

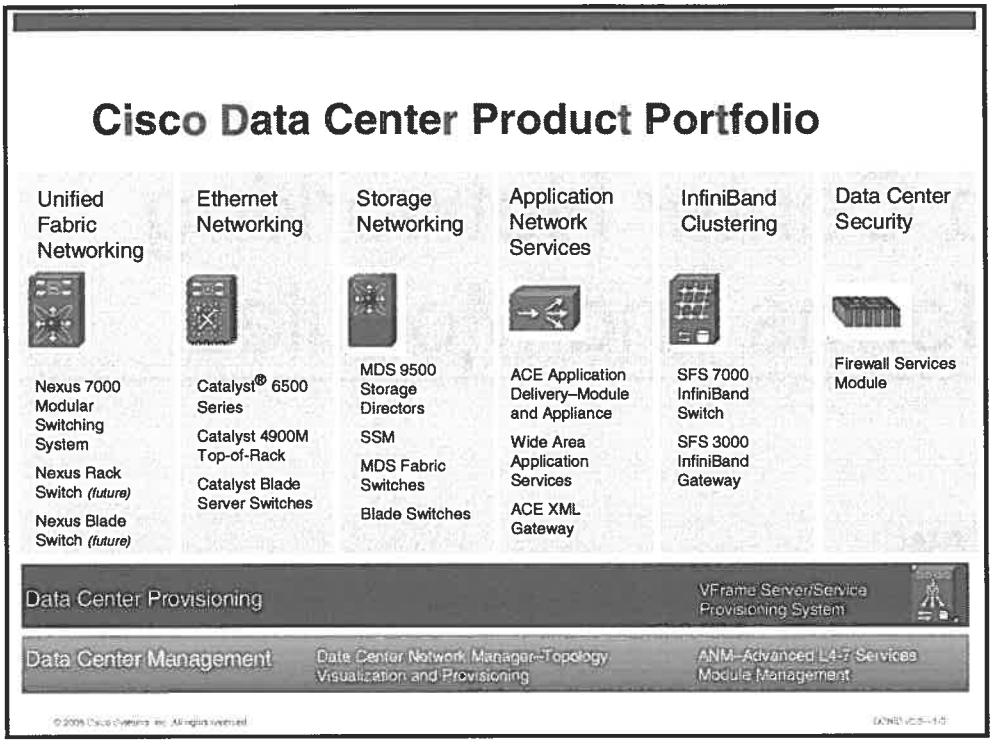
Upon completing this lesson, you will be able to identify the platforms and models that are most relevant to the data center network. This ability includes being able to meet these objectives:

- Describe the basic characteristics of the Cisco Nexus Family of switches portfolio for the data center:
  - Cisco Nexus 7000 Series Switches, Virtual Device Contexts (VDC) and Cisco Nexus 5000 Series Switches
- Describe the basic characteristics of the Cisco Catalyst Family of Switches portfolio for the data center:
  - Cisco Catalyst 6500 Series Switches, Virtual Switching System (VSS), Security service modules:
    - FWSM, IDSM-2 and Cisco Catalyst 4900 Series Switches
- Describe the basic characteristics of Cisco Ethernet blade switches
- Describe the basic characteristics of the Network Analysis Module (NAM)
- Describe the basic characteristics of the Cisco ACE module/appliance
- Describe the basic characteristics of the Cisco Multilayer Director Switch (MDS) platform
- Describe the basic characteristics of the Cisco ONS 15454

https://t.me/learningnets

Cisco Catalyst 6500 Series Switches

This topic introduces the Cisco Catalyst Series Switches and some of the Catalyst 6500 services modules, and describes the basic characteristics of the various models available and their recommended placement within the data center.

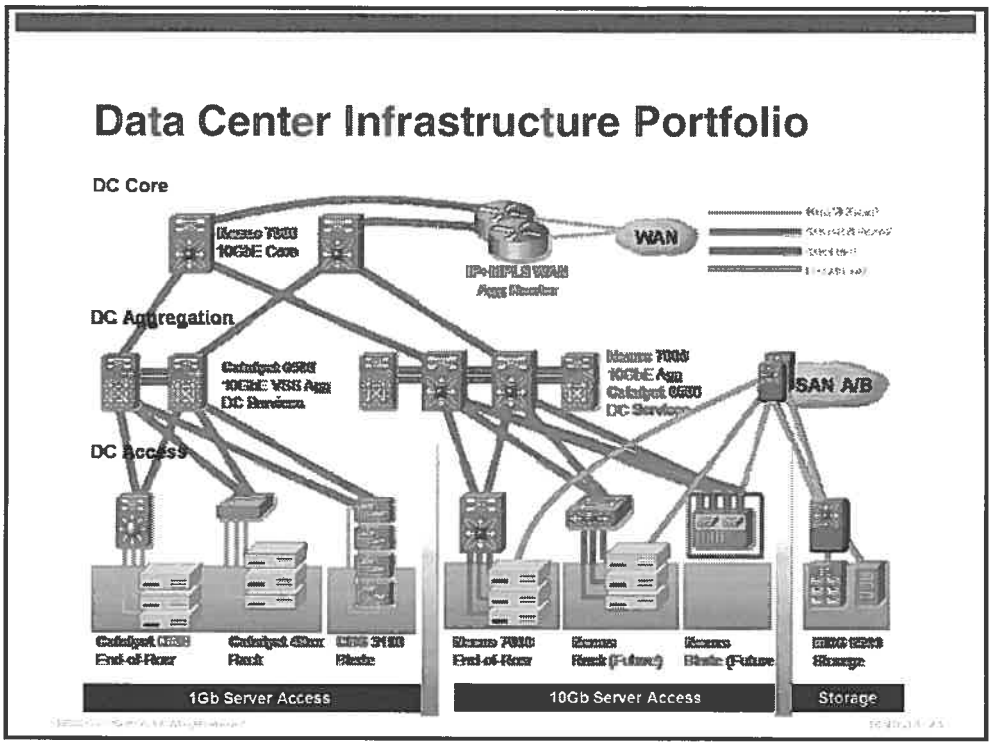


The Cisco Data Center products listed on this slide will be discussed in the following lessons:

- Cisco Nexus 7010 Series Switch
- Cisco Nexus 5000 Series Switch
- Cisco Catalyst 6500 Series Switches
- Cisco Catalyst 4900 Series top-of-rack switch
- Cisco Catalyst Blade Server switches
- Cisco MDS 9500 Series storage switches
- Cisco MDS 9100 Fabric switches
- Firewall Services Modules
- Intrusion Detection and Intrusion Protection Systems
- Cisco ACE Application Control Services Module
- Cisco Adaptive Security Appliances (ASA)
- Network Analysis Module
- Cisco SFS 7000 High-Density InfiniBand Computer Fabric Switch
- Cisco SFS 3000 InfiniBand Gateway
- ACE Appliance

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



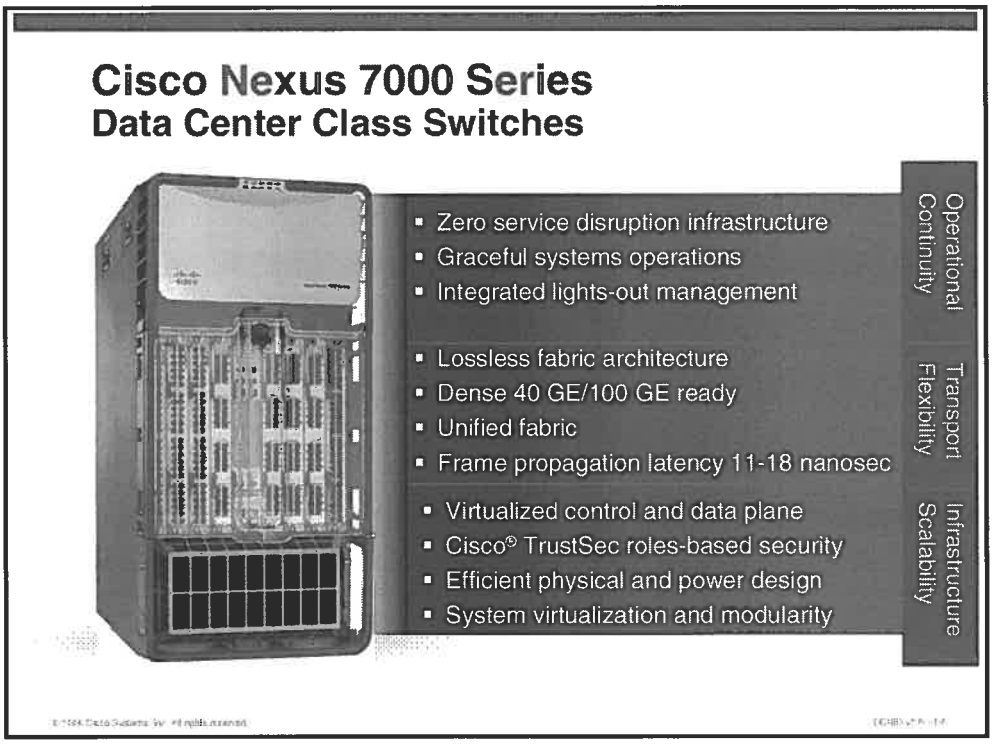
The figure represents the positioning of Cisco platforms in the data channel network infrastructure with respect to data channel core, aggregation, and access layer, and data channel interconnect and internet edge.

https://t.me/learningnets

Cisco Nexus Switching Portfolio for the Data Center

This topic introduces the Cisco Nexus 7000 and 5000 Series Switches and some of the Catalyst 6500 services modules, and describes the basic characteristics of the various models available and their recommended placement within the data center.

Cisco Nexus 7000 Series Data Center Class Switches



The data center class of switches represents a new architecture built specifically for data centers. General purpose systems are unable to meet the demands of data center environments. These data center switches are built on a new operational architecture supporting zero service disruption for both hardware and software upgrades. The complexity of support and customer OpEx is reduced by creating linkages between the hardware and the software for graceful operations and embedded lights out management support on the supervisor.. The transport architecture has a lossless architecture that enables the unified fabric. The scalability of the architecture can support dense 40 GB and 100 GB. These are the key items that the customer should be focused on. We look at the infrastructure architecture, embedding virtualization of the control plane and the data plane, support for virtual device contexts that allow the customer to segregate out operational control, and data plane traffic into different VDCs. This is the first Cisco platform to support Cisco TrustSec to scale security into the data center using source group ACLs and LinkSec encryption at wire-rate on all ports. Addressing power and cooling, the platform design is very efficient. It includes power supplies that are more than 90 percent efficient and fan trays that step down to a very low setting of 300 watts and then scale back up should the ambient temperature increase. The adjustable settings allow you to be more effective in a hot aisle/cold aisle deployment.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

The Nexus 7000 Series offers the following innovations in the data center:

- **A new operations architecture**:
  Zero service disruption through hardware or software upgrades. Graceful system operations, linking hardware changes to software functionality, delivering an intelligent approach to service insertion. Nexus 7000 Series switches support lights out management to reduce costs of support and operations.
- **A new switching fabric architecture**:
  Lossless fabric supports the unified fabric requirements for a lossless transport, 40 Gigabit Ethernet and 100 Gigabit Ethernet technologies with day one support for 230Gb/s per slot, and a total system architecture that scales to over 15 Tbps.
- **A new infrastructure architecture**:
  Virtualizing the control plane and data plane to allow for VDCs, operational separation and traffic separation. Securing the infrastructure through role-based security, scaling security beyond IP ACL complexity, and scaling the power and cooling infrastructure through efficient system designs.

https://t.me/learningnets

Cisco Nexus 7000 Series Switches
Unified Fabric and 40/100 Gigabit Ethernet

High Availability Design
- Redundancy of common equipment
- Rear access for all common equipment

Physical Characteristics
- 512 ports, 10 Gigabit Ethernet per rack
- Optimized for data center
- Efficient physical and power design

Operational Manageability & Serviceability
- Cable management
- Beacon LEDs
- Integrated Lights Out
- Integrated Wireshark analyzer

Highest Density for Scalable Designs
- 230 Gb/s per slot
- Flexible port modes:
  - Dedicated and shared
- Cisco TrustSec role-based security



Cisco Nexus 7000 Series Switches are built to support unified converged data lossless Ethernet fabrics with high availability and non-blocking internal switching architecture for non-disruptive data center operation. The Cisco Nexus 7010 Series Switch has high density—up to 512 ports per chassis.

These unified platforms can support the emerging technologies for not only unified I/O, such as FCoE and DCE, but also for 40 Gigabit Ethernet and 100 Gigabit Ethernet.

All hardware components are redundant: dual supervisors, redundant power supplies, fan systems and fabric modules.

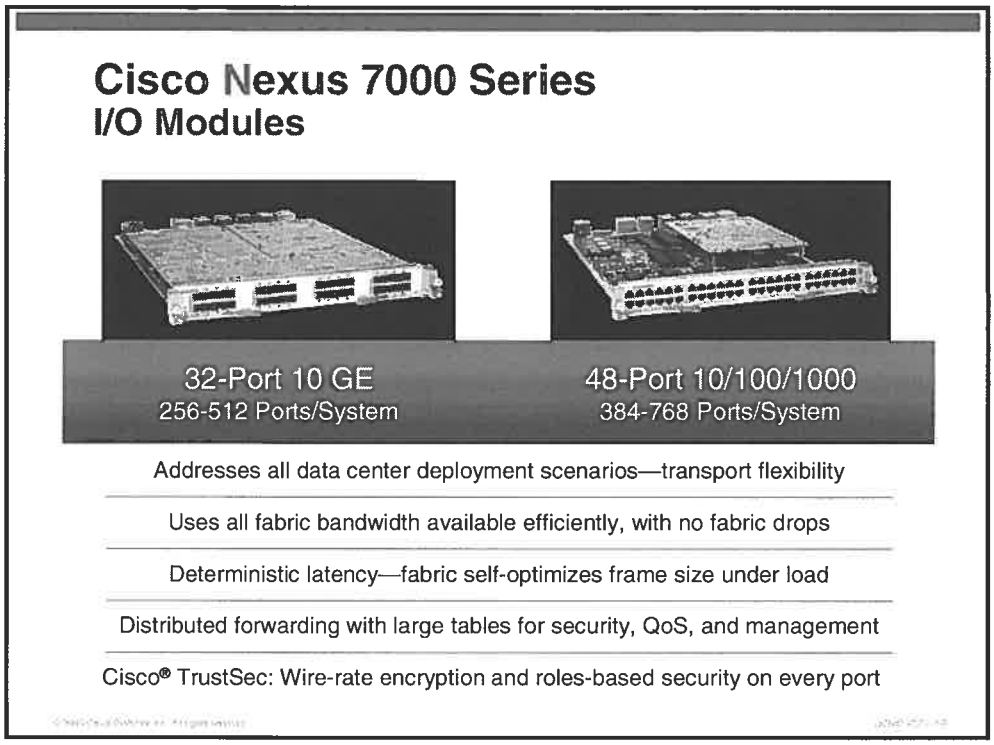
These components have been designed around the customer requirements for data center equipment, with efficient cooling and power usage. In well controlled environments it consumes significantly less power than at higher ambient temperatures. The power supplies are industry leading—over 90% efficient at full load, and 89% efficient at 50% load.

Cisco Nexus 7000 Series Switches support Cisco TrustSec.

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

I/O Modules



The Cisco Nexus 7010 Series Switch supports 32-port 10 Gigabit Ethernet and 48-port 10/100/1000 Ethernet line cards.

NX-OS is designed as an operating system for the data center and other Ethernet environments. Virtual Device Context (VDC) provides flexible separation and distribution of software components and hardware resources.

NX-OS is Linux-based

- Finer-grain memory protection
- Stateful process restart
- Easier debugging environments help bug fix cycle time

https://t.me/learningnets

Cisco NX-OS: Purpose Built for Data Center Networking

- Integrates multiple technologies and must be:
  - DCE/unified fabric ready, which translates to the following characteristics:
    - SAN traffic has the following character/requirements:
      - Sometimes ?dedicated? paths/bandwidth from server to storage must be ensured
      - High switching throughput
      - Nanosecond range of frame switching latency (11-18 ns of Nexus 7000 latency)
      - SAN fabric no more than end-to-end delay
      - NO FRAME LOSS (Frame loss triggers I/0 exchange to be restarted)
      - In-order delivery of frames within an exchange/sequence
      - Non-blocking switching architecture

© 2008 Cisco Systems, Inc. All rights reserved.

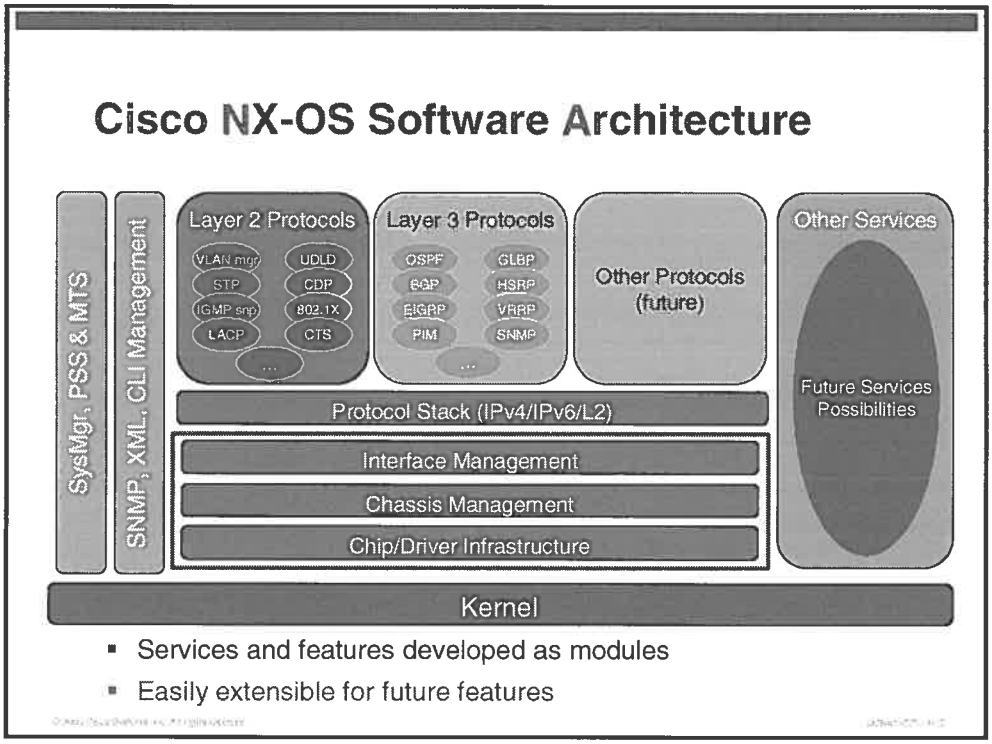
SAN fabrics have the following characteristics:

- "Dedicated" paths/bandwidth between the servers and storage devices
- High switching throughput and non-blocking fabric switching architecture
- Nanosecond range of frame switching latency
- No more than 4 seconds of SAN fabric end-to-end delay
- No frame loss (frame loss triggers I/0 exchange to be restarted)
- In-order delivery of frames and exchanges/sequences or upper layer I/O units
- Non-blocking switching architecture

NX-OS is designed to meet the operational needs of the data center:

- Built on SAN-OS
- Multi-protocol data channel-class operating system
- Comprehensive L2 and L3 feature set
- Cisco IOS-Like CLI
- Zero service disruption
- Perform maintenance anytime
- Virtual Device Contexts (VDC)
- Stateful process restart
- Stateful restart = NO reconvergence
- Unified management infrastructure
- XML and Web services

https://t.me/learningnets



NX-OS software architecture is modular and virtualized.

VDC enables:

- Flexible separation and distribution of software components and hardware resources
- Securely delineated administrative contexts

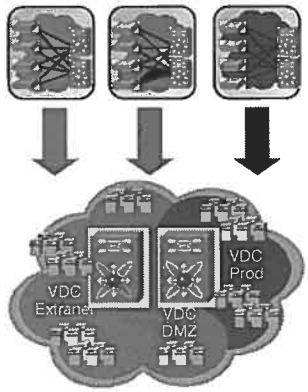
VDC are not:

- Capable of running different OS levels on the same box at the same time
- Based on a *hypervisor* model; a single infrastructure layer handles hardware programming.

https://t.me/learningnets

VDC—Application

- Many network designs used a silo approach for reasons of security, network administration, outsourcing/organizational boundaries, and fault isolation:
  - Production, development, test
  - Intranet, Internet, DMZ, extranet
  - ApplicationA, ApplicationB, ApplicationC
  - CustomerA, CustomerB, CustomerC
  - ClusterA, ClusterB, ClusterC
  - Storage Replication, Secure Transaction Database
- Using VDC enables collapsing physical infrastructure into logical infrastructure while maintaining the same security, network administration, outsourcing/organizational boundaries, and strong fault isolation.

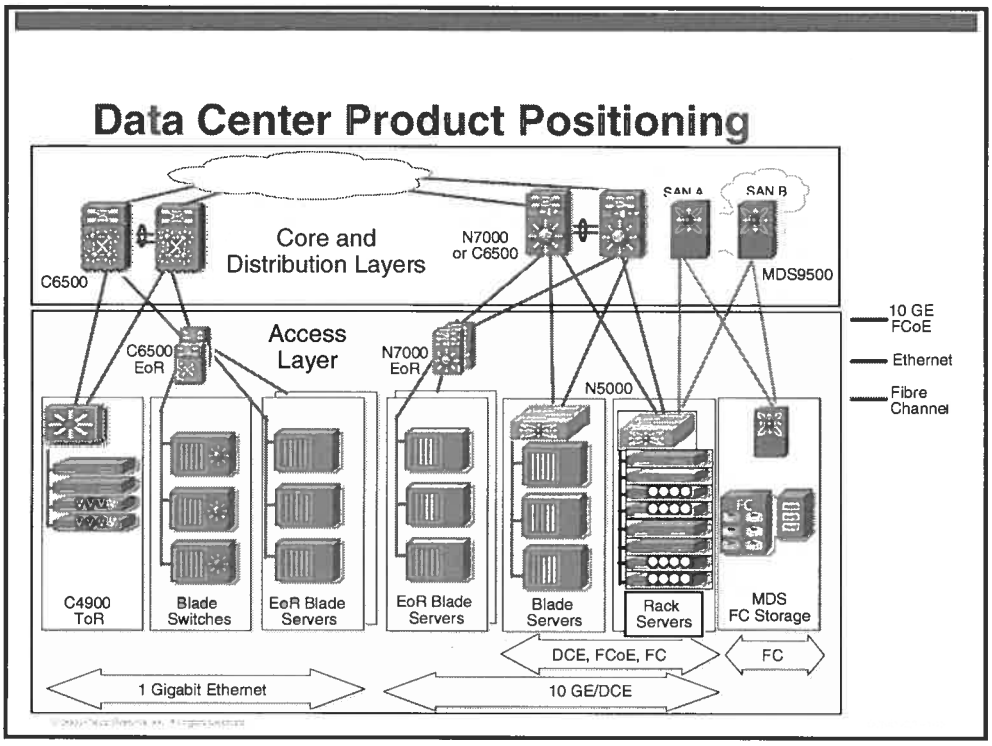


Physical network islands are virtualized onto common data center networking infrastructure

Physical data center network islands are virtualized onto common data center networking infrastructure using VDCs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



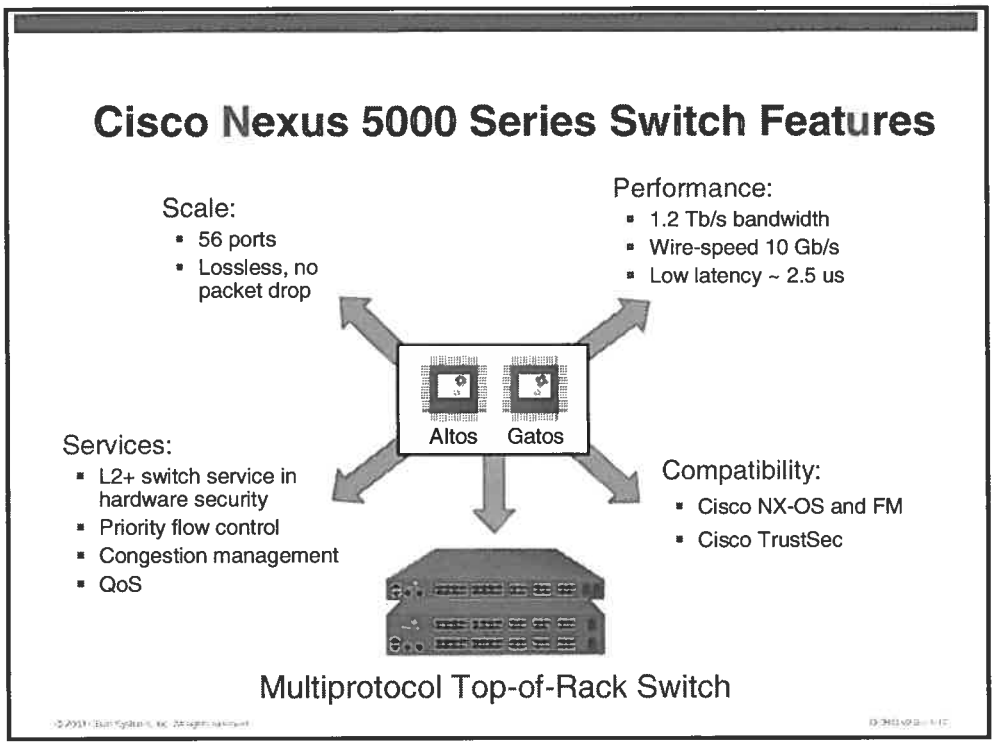
Cisco Nexus 5000 Series Switches are:

- Top-of-rack switches for 10 Gb/s unified I/O access layer
- End-of-row 10 Gb/s aggregation layer switches
- Top-of-rack unified I/O aggregation layer switches

Nexus 5000 Series Switches support a mix of Fibre Channel, Ethernet and FCoE ports on the 28-port and 56-port switches.

https://t.me/learningnets

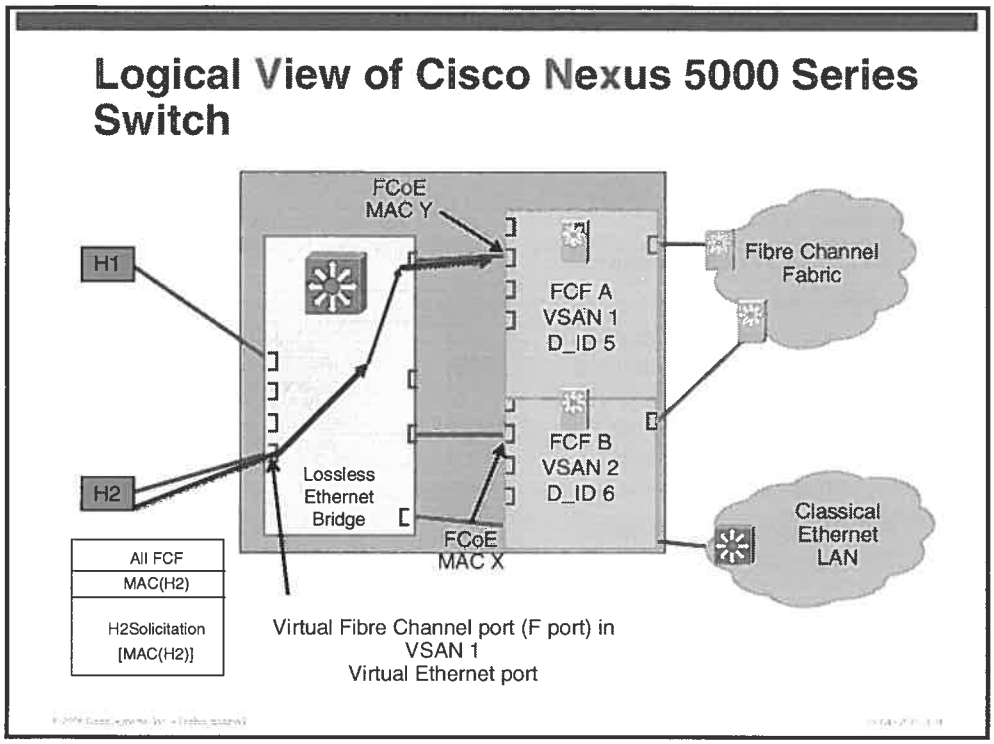
Cisco Nexus 5000 Series Switches



The Nexus 5000 Series Switch integrates SAN-OS and Layer 2 Catalyst switch features.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



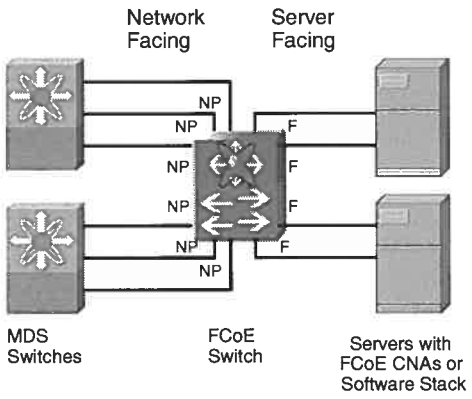
This figure represents the logical view of Nexus 5000 Series Switches. These switches consist of Layer 2 Catalyst 6500 Series Switches and Fibre Channel switches that are connected across an FCoE mapper, where Fibre Channel frames get encapsulated into Ethernet frames or decapsulated from Ethernet frames and then sent across native Fibre Channel fabric.

https://t.me/learningnets

Cisco Nexus 5000 Series Switch: NPV Mode Network Topology

N Port Virtualizer:

- No Fibre Channel domain required on switch
  - MDS switches see Nexus 5000 switch as Fibre Channel end host with many ports and many Fibre Channel IDs assigned.
  - Network-facing ports function as NP ports.
  - Server-facing ports function as F ports



Nexus 5000 Switch in N Port Virtualization (NPV) mode:

- Relays FLOGI/FDISC to true Fibre Channel switch upstream
- Has no E ports and no Fibre Channel services
- Needs NPIV support on F port in upstream Fibre Channel switches

Current deployment:

- Same feature is released in MDS 3.2(1) for blade and fabric switches
- Similar to HP Virtual Connect

NPV mode advantages:

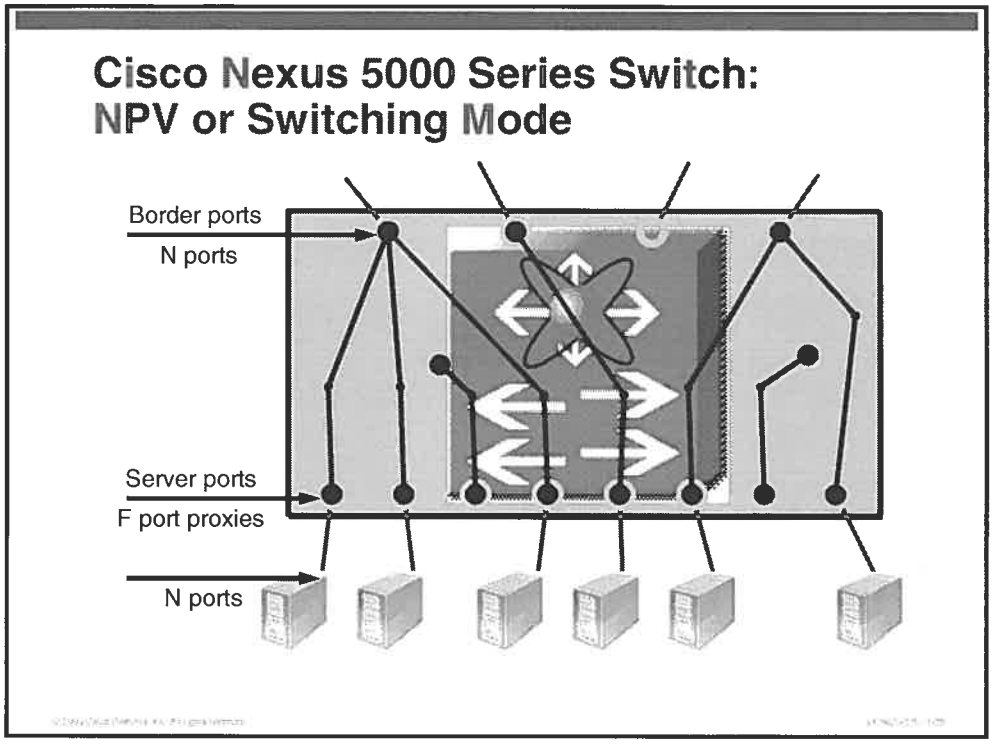
- Increases scalability by reducing the number of Fibre Channel domain IDs used
- Seamless interoperability between Fibre Channel blade switches and multi-vendor core SAN switches

Disadvantages:

- No local switching; all policy enforced by true Fibre Channel switch upstream.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

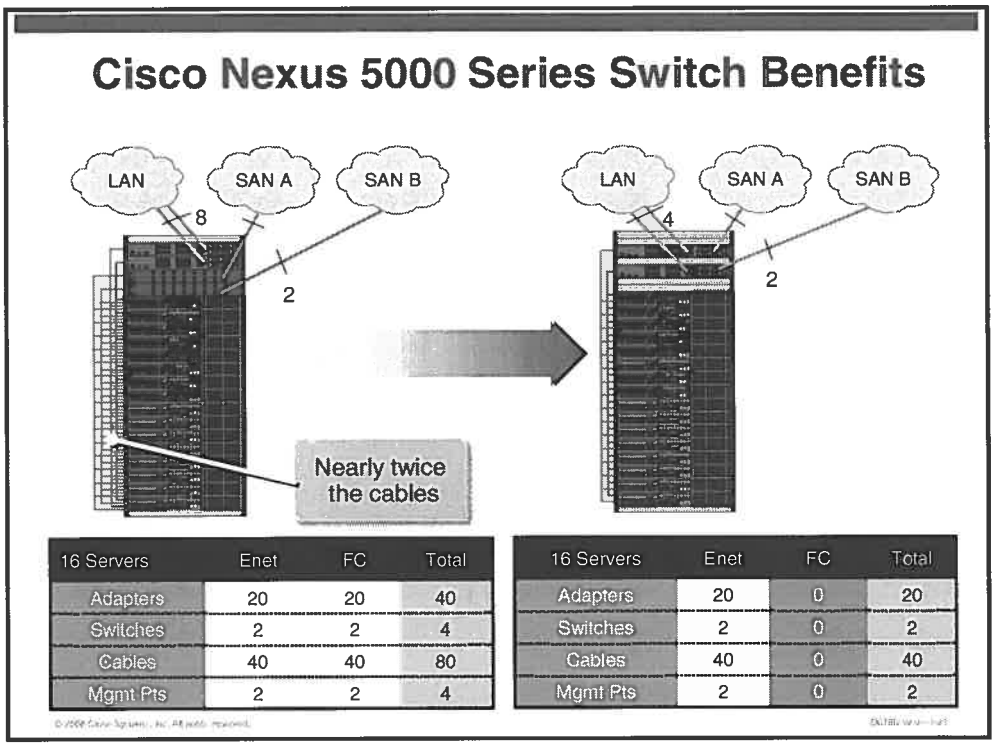


The Nexus 5000 Series Switch functions as either a Fibre Channel Switch or an NPV, but not both. Changing between modes requires reboot and erase of configuration.

NPV supports the following small subset of Fibre Channel F port features:

- **VSAN membership**
- **Pinning and load balancing**:
  Server-facing interfaces pinned to uplink interfaces based on lowest number of server interfaces. Server interfaces are pinned to uplink interfaces (not FLOGI basis)
- **Data packet on server interface**:
  All packets are sent to pinned external interface without any forwarding lookup. Binding check is done to make sure Security Identifier (SID) is coming on right server interface to prevent address spoofing.
- **Data packet on external interface**:
  Forwarding lookup is done based on destination ID (D_ID) of frame that will point to server Interface. The packet is discarded on miss.

https://t.me/learningnets



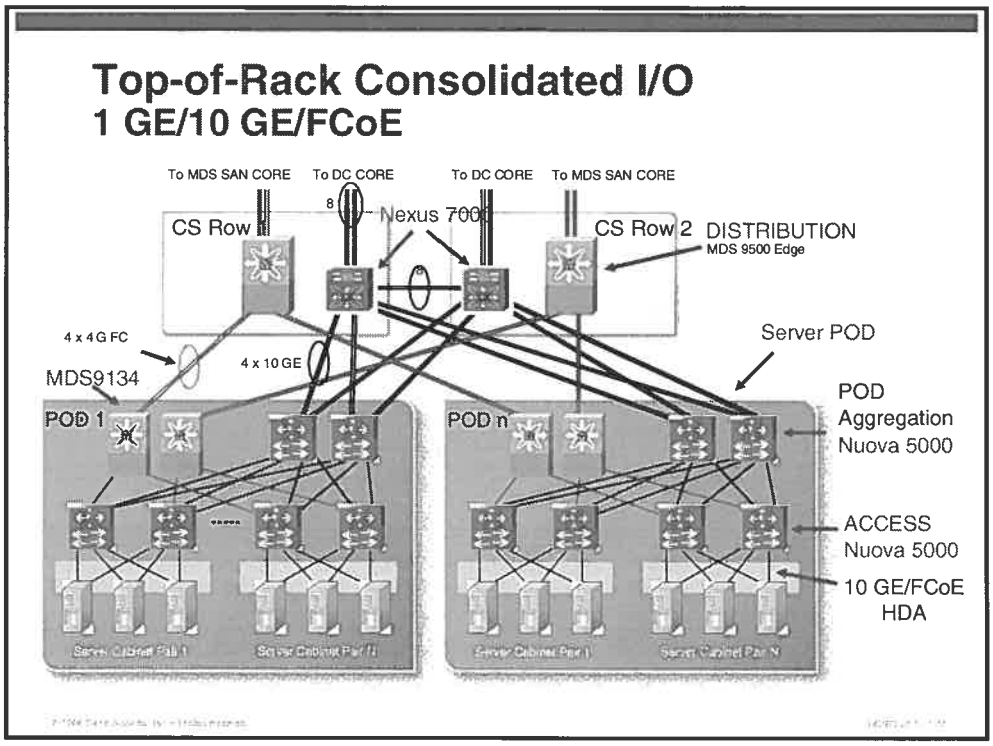
The slide shows the reduction in cost and networking components resulting from using Nexus 5000 Series Switches in the access layer.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

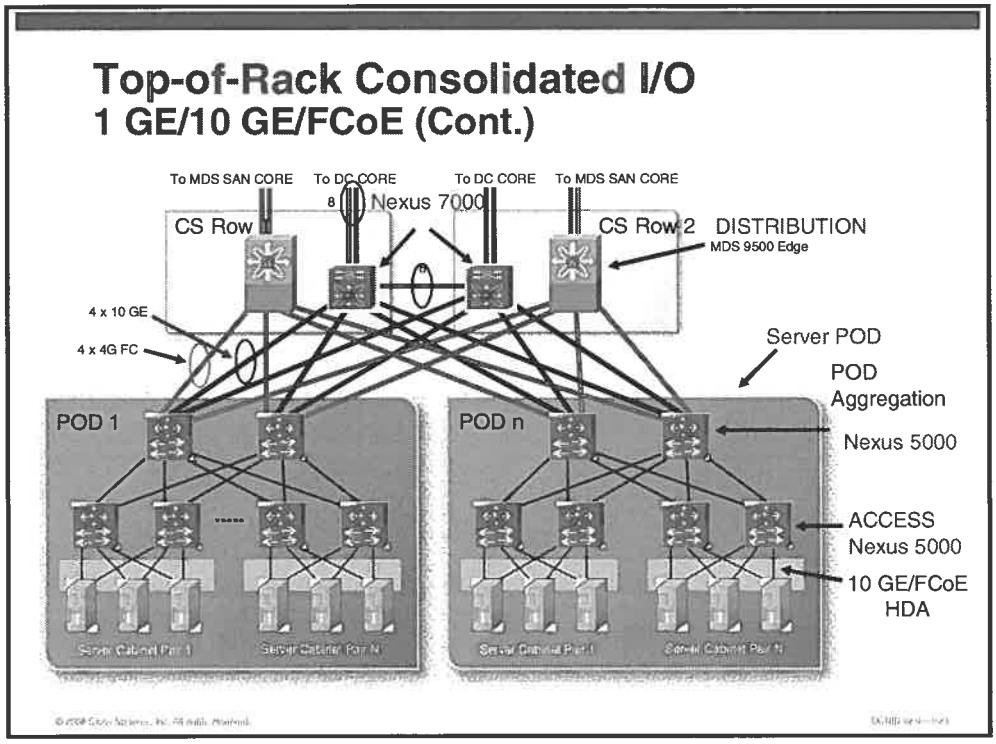
Top-of-Rack Consolidated I/O

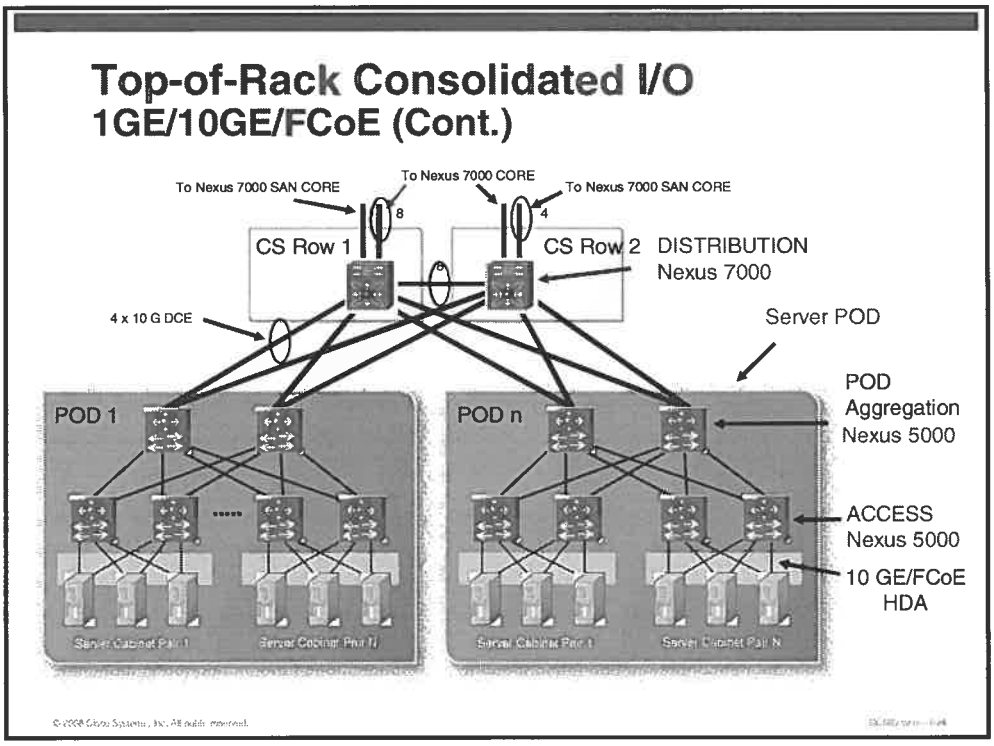
The following three topologies position the Nexus 5000 Series Switches in the data channel network.The figures show Nexus 5000 Series Switches as top-of-rack switches in the data center.



This slide shows 1 Gigabit Ethernet and 10 Gigabit Ethernet access layer and distribution layer switches.

https://t.me/learningnets



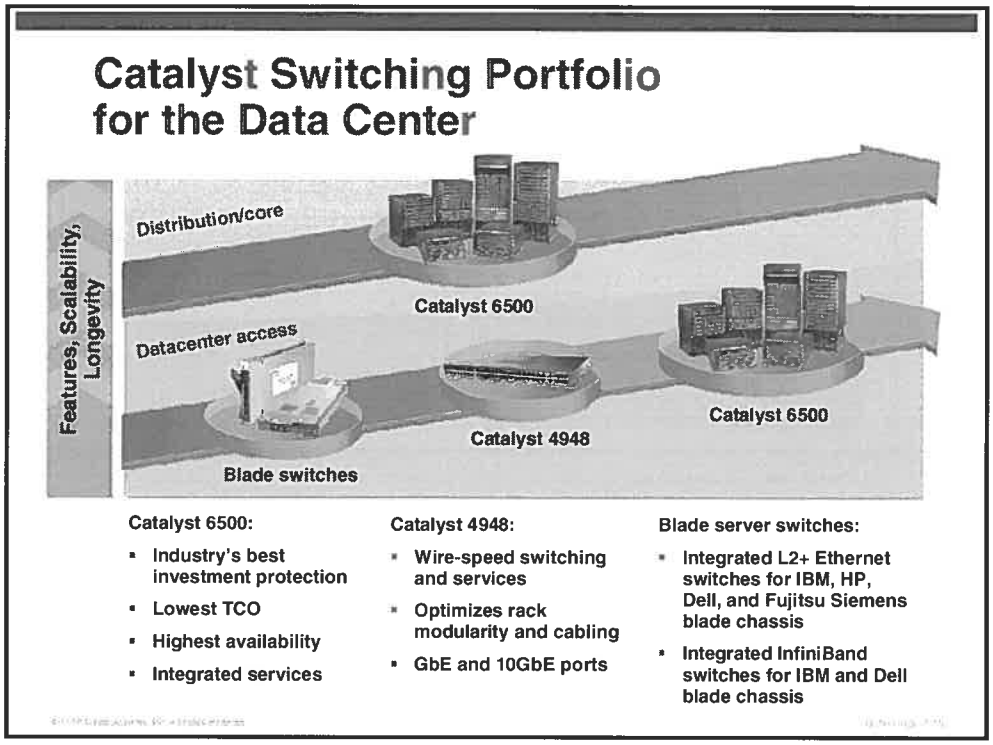


The figures above show 1 Gigabit Ethernet and 10 Gigabit Ethernet consolidated access layer and distribution layer switches that eliminate the need for MDS switches in the aggregation and distribution layers.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Catalyst 6500 Series Switching Portfolio



The Catalyst 6500 Series Switching portfolio for the data center includes the following models:

- **Catalyst 6500**

  This Cisco switch is the industry's leading switch, designed for all layers of the data center architecture.

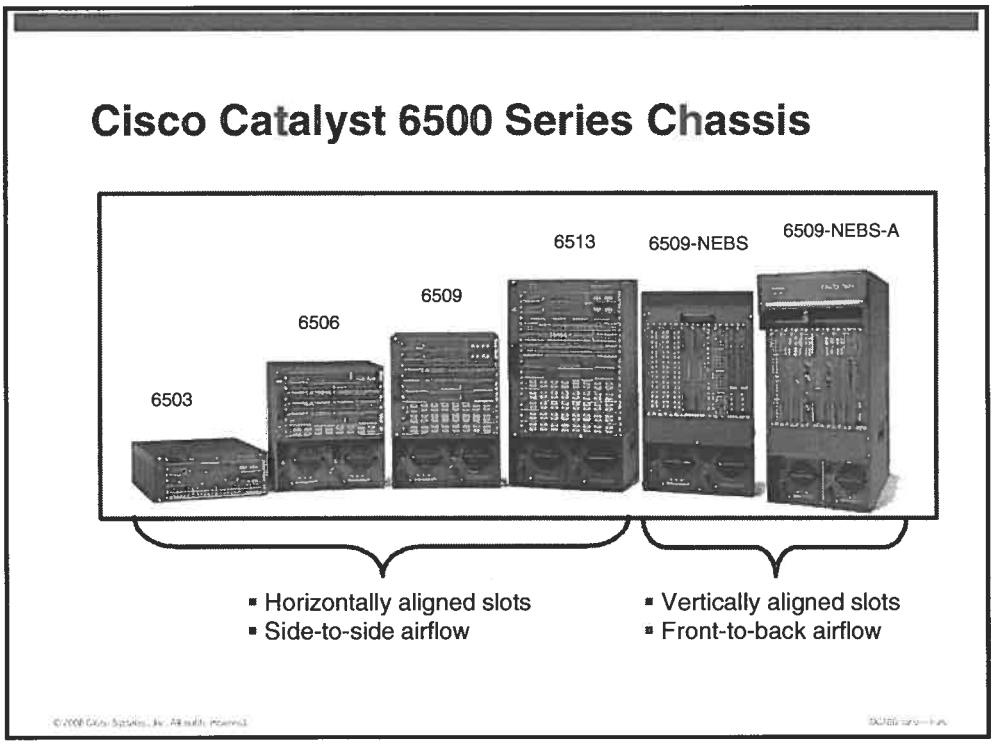
- **Catalyst 4948**

  This switch provides wire-speed switching and services within the data center access layer.

- **Blade Server Switches**

  These switches provide integrated Layer 2 services within the data center access layer.

https://t.me/learningnets



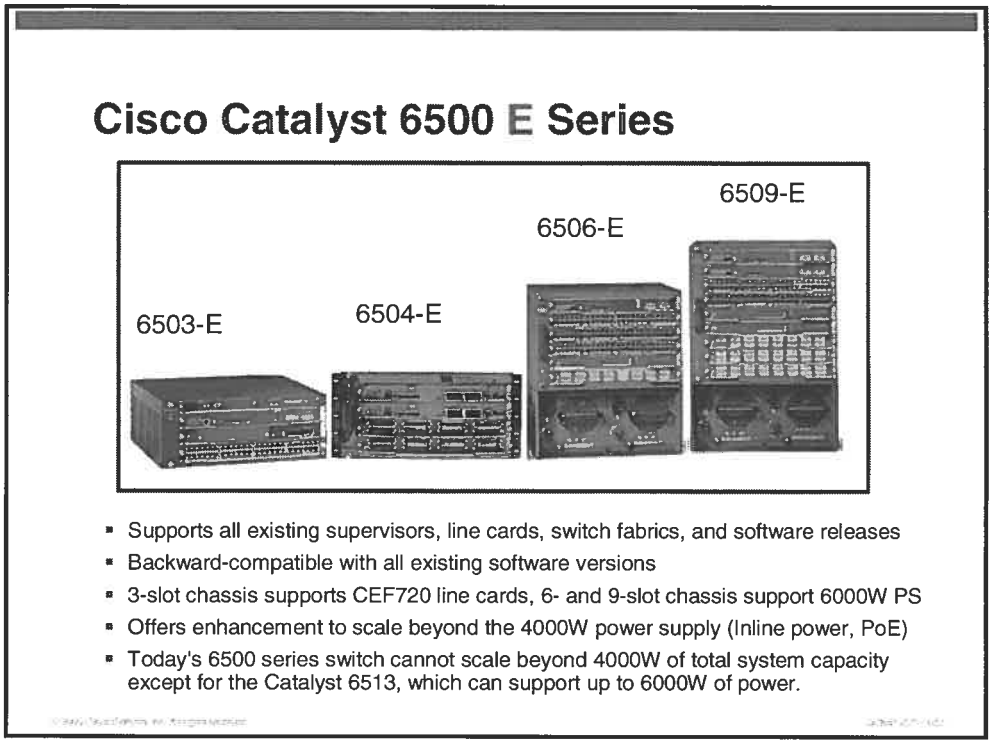
The Cisco Catalyst 6500 Series of switches is offered with either horizontally or vertically aligned slots. The following types of Catalyst 6500 chassis are offered:

- The Cisco Catalyst 6500 Series switches with horizontally aligned slots:
  - Cisco Catalyst 6503 Switch with a 3-slot chassis
  - Cisco Catalyst 6506 Switch with a 6-slot chassis
  - Cisco Catalyst 6509 Switch with a 9-slot chassis
  - Cisco Catalyst 6513 Switch with a 13-slot chassis
- The Cisco Catalyst 6500 Series switches with vertically aligned slots:
  - Cisco Catalyst 6509-NEBS Switch with a 9-slot chassis
  - Cisco Catalyst 6509-NEBS-A Switch with a 9-slot chassis

Note Network Equipment Building Systems (NEBS) criteria are a set of requirements and objectives for personnel safety, property protection, and operational continuity. NEBS documents describe both physical and electrical requirements. NEBS compliance is a critical issue to telephone companies when these companies evaluate the suitability of products for use in their networks. All Cisco Catalyst 6500 chassis are NEBS-compliant. The Catalyst 6500 chassis that are identified as NEBS have front-to-back airflow specifically designed for service provider installations. The non-NEBS Catalyst 6500 chassis have side-to-side airflow.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Catalyst E Series Chassis

The Cisco Catalyst 6500 E Series adds enhanced 3-, 4-, 6-, and 9-slot chassis to the Cisco Catalyst 6500 Series switches. The new 3-slot chassis (part number 6503-E) offers enhancements to support the Cisco Catalyst 6500 Series line cards. The new 6- and 9-slot chassis (part numbers 6506-E and 6509-E) offer enhancements that increase the overall system power capacity for industry leading Power over Ethernet (PoE) port-density scalability. All existing supervisors, line cards, switch fabrics, and software releases are supported in the new Catalyst 6500 E Series chassis. The following section summarizes the features of the Catalyst 6500 E series chassis:

Catalyst 6503-E

The 3-slot Cisco Catalyst 6503-E chassis offers a compact 4-rack unit (RU) height that is ideally suited for multi-Gb/s secure data centers, remote access, e-commerce, and converged network solutions. The Catalyst 6503-E chassis supports up to ninety six 10/100/1000 Ethernet ports and is well suited for small wiring closets. The Catalyst 6503-E chassis is also compatible with all Catalyst 6500 supervisor engines and can be configured with redundant power supplies.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Catalyst 6504-E

The 4-slot Cisco Catalyst 6504-E chassis delivers performance in a compact 5-RU form factor. This chassis can be configured in two ways: with a single supervisor engine and up to three line cards, or with dual supervisor engines and up to two line cards. The Cisco Catalyst 6504-E also supports redundant AC or DC power supplies. The interface density and breadth of the Cisco Catalyst 6504-E make it ideal for deployment in high performance applications such as the following:

- Enterprise access layer
- Small or medium enterprise core and distribution layers
- Metro Ethernet edge aggregation
- Enterprise WAN edge

6506-E Chassis

This chassis is ideal for many wiring-closet and core-network deployments. The 6-slot Cisco Catalyst 6506-E chassis provides intermediate port densities and supports a range of power supply options, including redundant power supplies ranging from 1000W AC or DC to 2500W AC or DC. When equipped with a single Catalyst 6500 Series supervisor engine, five payload slots are available to support a wide range of interface or services modules. The Catalyst 6506-E also features field-upgradeable fan trays for easy service.

6509-E Chassis

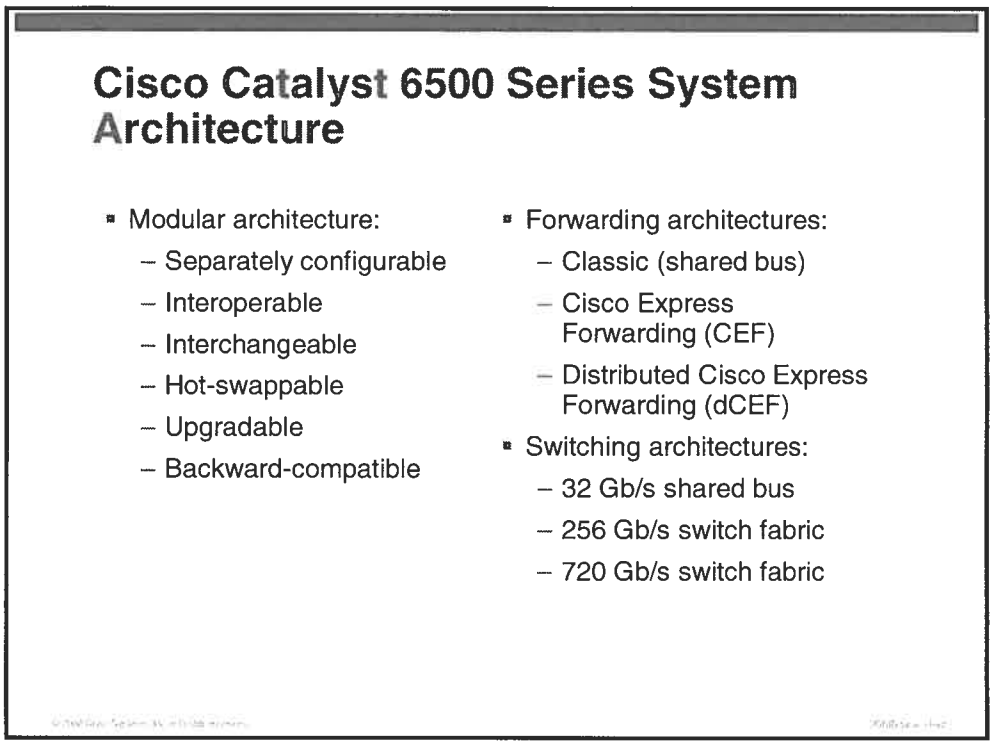
The 9-slot Cisco Catalyst 6509-E chassis offers scalable port densities, which is ideal for many wiring-closet, core, and data center deployments. When equipped with a single Catalyst 6500 series supervisor engine, eight payload slots are available to support a wide range of interface or services modules. The Catalyst 6509-E offers a range of power supply options, including redundant power supplies ranging from 1000W AC or DC to 4000W AC. The Catalyst 6509-E also features field-upgradeable fan trays for easy service.

Note The Cisco Catalyst 6513 Switch is not sold in an E variation at this time.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Catalyst 6500 Series Modular Architecture



The Cisco Catalyst 6500 Series is a modular system that can grow according to customer requirements by expanding to technological evolution. The ability for growth allows Cisco customers to upgrade and reconfigure systems by adding new modules, replacing existing modules, and adding and redeploying systems. Throughout the Cisco Catalyst 6500 Series product line, modules are:

- **Separately configurable:** Simplifies the addition of new services
- **Interoperable in the same chassis:** Provides flexible design options
- **Interchangeable among Cisco Catalyst 6500 Series systems:** Simplifies sparing and network expansion
- **Hot-swappable without requiring a chassis to be powered off:** Provides fast upgrade and repair
- **Upgradeable as newer modules are released:** Provides investment protection

https://t.me/learningnets

Forwarding Architectures

Cisco Catalyst 6500 Series modules use one of three forwarding technologies: classic, Cisco Express Forwarding (CEF), and distributed Cisco Express Forwarding (dCEF). Each technology has a unique architecture with the following characteristics and capabilities:

- Classic architecture connects modules to the 32 Gb/s shared bus.
- CEF is a switched fabric architecture. CEF is scaleable to 30 Mpps. This technology uses a central CEF engine located on the supervisor engine Policy Feature Card (PFC); it makes forwarding decisions for all line cards.
- dCEF is suited for the most demanding environments. This technology uses the dCEF engine located on the interface module Distributed Forwarding Card (DFC) daughter card and the dCEF table. The dCEF table is a local copy of the supervisor engine central CEF table located on the interface module DFC. The interface module makes all forwarding decisions locally. The interface module provides maximum performance and scalability of the architectures offered.

Note You need a Multilayer Switched Feature Card (MSFC) for CEF; the MSFC manages the CEF tables.

Switching Architectures

To allow platforms to scale in any deployment, Cisco has developed the following switching architectures for Cisco Catalyst 6500 Series modules:

- 32 Gb/s bus: Allowing access to a central shared bus
- 256 Gb/s switch fabric: Located on the Switch Fabric Module (SFM)
- 720 Gb/s switch fabric: Located on the Cisco Catalyst 6500 Series Supervisor Engine 720

Note More information on the Catalyst 6500 Series can be found at http://www.cisco.com/go/catalyst6500.

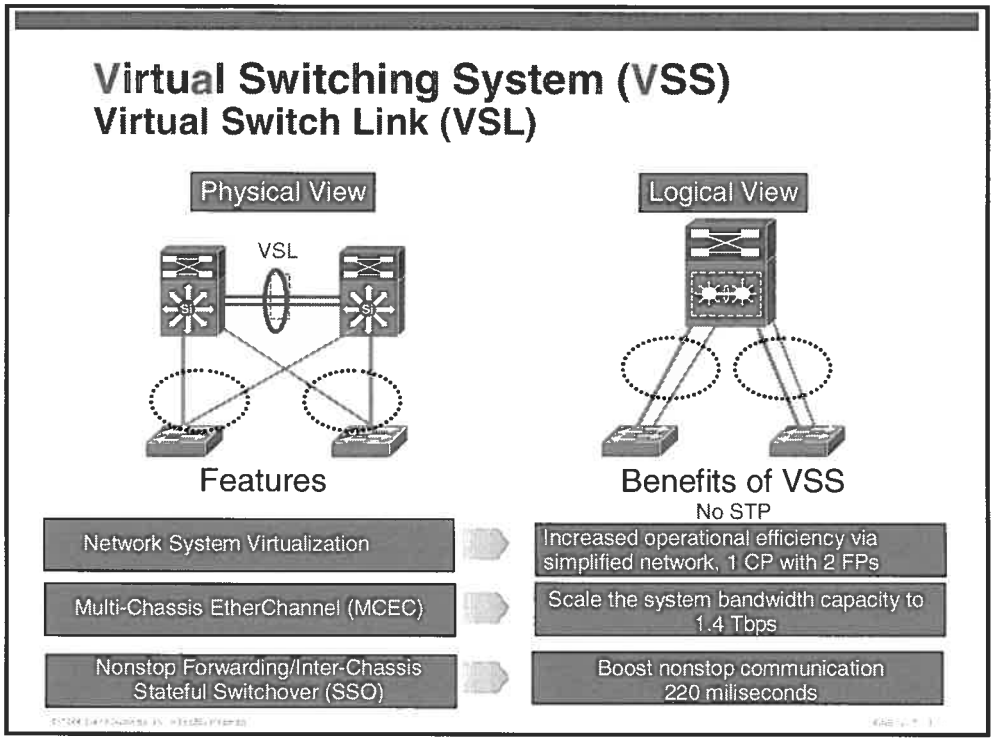
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Catalyst 6500 Virtual Switching System

This topic introduces the basic characteristics of the Catalyst 6500 Switches Virtual Switching System (VSS).

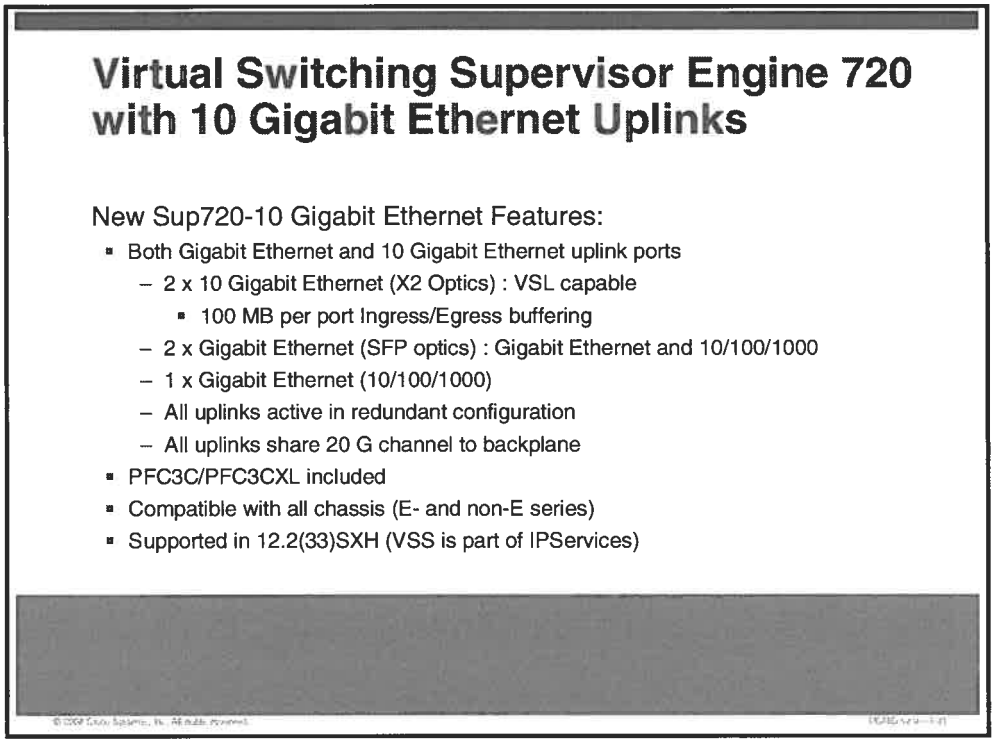
Virtual Switch Link (VSL)



The virtual switching system is essentially based on two Catalyst 6500 switches acting as a single virtual switch. VSS enables management of the virtual switch as a single network entity and reduces the spanning tree and Hot Standby Router Protocol (HSRP) concerns, since two switches are acting as a single logical switch and two or more physical links are acting as a single logical link Multi-Chassis EtherChannel (MCEC). The VSS creates a highly-available virtual switch based on two physical switches with fast network throughput. VSS provides deterministic failover times because it eliminates the use of HSRP and the Spanning Tree Protocol (STP). Using cross-chassis Cisco Nonstop Forwarding (NSF), sub 200 millisecond failover times are guaranteed by VSS.. The two chassis do not have to be configured in the same way. For example, a 3-slot and a 16-slot chassis can be configured together or there could be two modules in one chassis and another 10 modules in the other chassis. The virtual switch link is not the backplane. This is a 10 Gigabit Ethernet link-based EtherChannel. The supervisor uplinks are the 8-port 10 Gigabit Ethernet cards that can be interconnected to transfer control traffic between the two switches along with the NSF-Stateful Switchover (SSO) feature functions.

The amount of bandwidth typically used for the control traffic is less than 5% of a single 10 Gigabit Ethernet link. At least two 10 Gigabit Ethernet links are recommended but up to eight 10 Gigabit Ethernet links can be configured members of a PortChannel or a so-called Virtual Switch Link (VSL).

https://t.me/learningnets

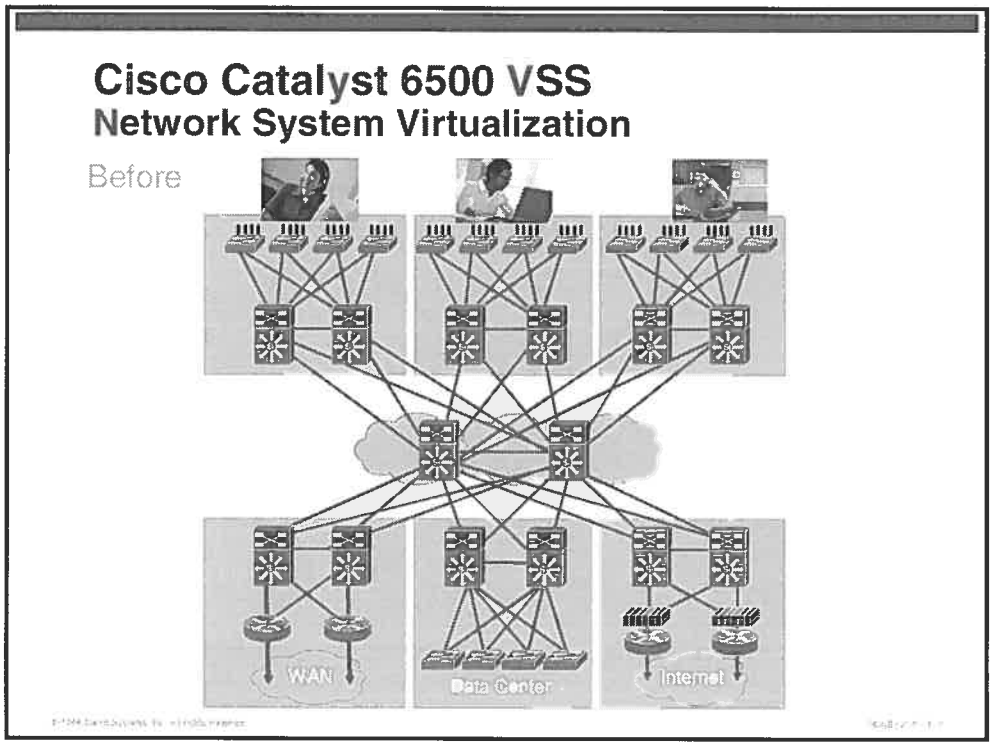


The VSS feature requires the VS720 supervisor with the software that supports the virtual switching system. From a control plane perspective, there is one active control plane but there are two chassis and a supervisor in each of the chassis. Both chassis fabrics are active so that 720 Gb/s throughput is provided by the first switch and 720 Gb/s throughput is provided by the next switch; together the switches provide 1440 Gb/s throughput. When the administrator logs in, he logs in to a single control plane and configures a single configuration file.

© 2008 Cisco Systems, Inc.

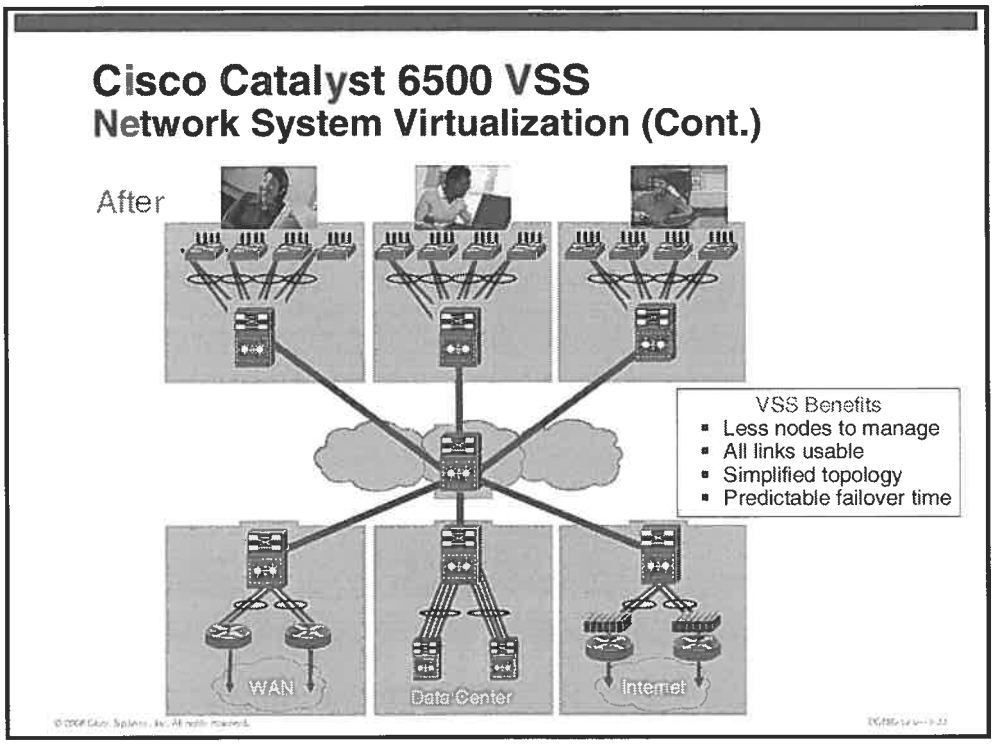
https://t.me/learningnets

Network System Virtualization



The VSS network simplifies the campus and other data center layers, the Internet, the DMZ zone and the WAN zone topology. The VSS network reduces the number of nodes to maintain and more effectively utilizes the data channel link capacity. This network also provides more predictable failover based on the nonstop switching failover (sub 200 milliseconds).

https://t.me/learningnets



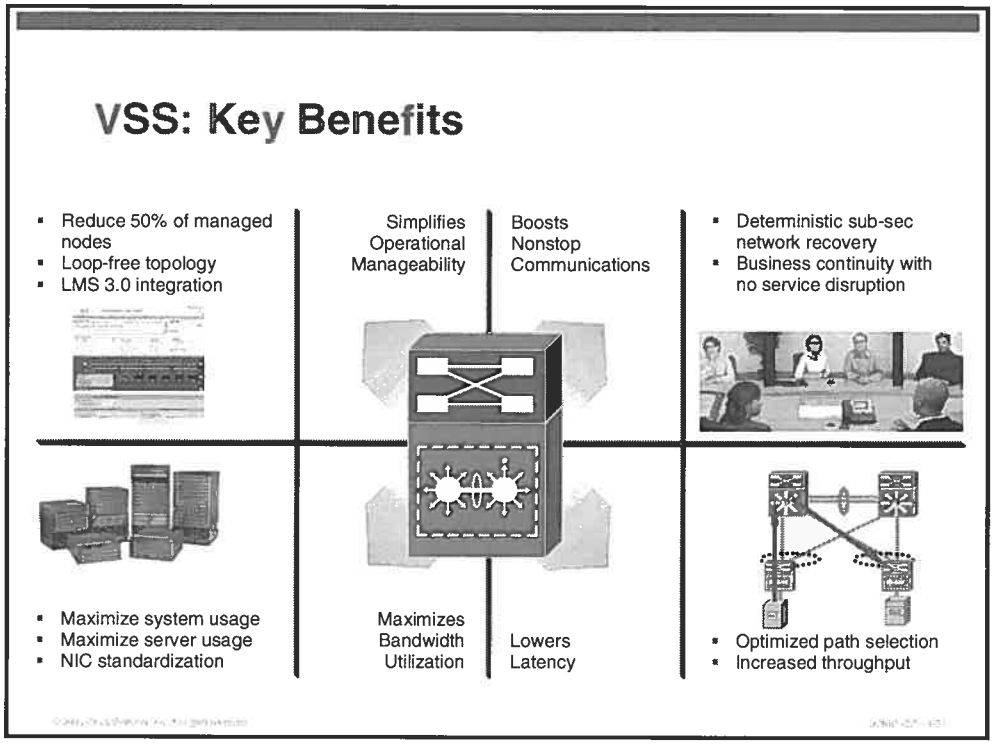
VSS boosts nonstop communication by delivering deterministic, sub-second Layer 2 link recovery through inter-chassis stateful failovers and the predictable resilience of the EtherChannel.

VSS 1440 scales system bandwidth capacity to 1.4 Tbps by activating all available bandwidth across redundant Catalyst 6500 switches.

VSS increases operational efficiency by enabling multiple Catalyst 6500 switches to share a single point of management, single IP address, and single routing instance, as well as by eliminating the dependence on STP.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The following list summarizes the benefits of using VSS in the data center network:

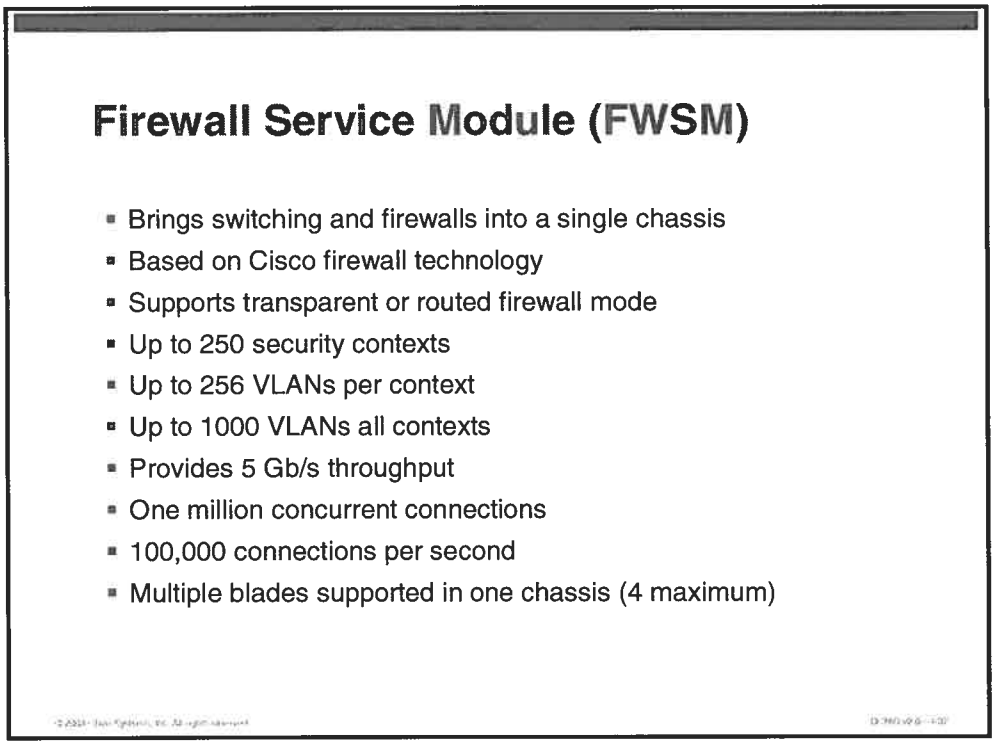
- Simplified operations resulting from the 50% reduction in management
- Eliminated STP protocol issues
- Simplified network topology
- Nonstop communication with sub 200 millisecond failover times
- Stateful failover
- Maximized link bandwidth utilization due to active members of EtherChannel
- Applications operation will not be impacted when there is a failover from one distribution to another distribution layer, and failover times are more predictable than with HSRP or other FHRP upper layer redundancy protocols.

https://t.me/learningnets

Security Services Modules: FWSM and IDSM-2

This topic introduces the Firewall Services Module (FWSM) and Intrusion Detection Services Module 2 (IDSM-2) which can be integrated within the modular Catalyst 6500 Series Switch.

Firewall Service Module



The key features of the Firewall Services Module (FWSM) are briefly summarized as follows:

- **Supports transparent or routed firewall mode**

  When configured to run in routed mode, the FWSM is considered to be a router hop in the network and performs Network Address Translation (NAT) between connected networks. When configured in transparent mode, the FWSM acts like a bump in the wire and is not considered a router hop. The inside and outside interfaces are the same networks but different VLANs, with the FWSM providing the connectivity.

- **Supports up to 100 security contexts**

  The FWSM can be in single or multiple context modes. In multiple context mode, up to 100 separate security contexts can be configured, depending on the software license being used. Multiple contexts are similar to having multiple standalone firewalls, conveniently contained within a single module.

- **Supports up to 256 VLANs per context**

  Up to 256 VLANs can be configured per context.

- **Supports up to 1000 VLANs all contexts**

  Across all contexts, a maximum of 1000 VLANs can be configured.

- **Supports 5 Gb/s throughput**

  The FWSM provides up to 5 Gb/s throughput.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **Supports one million concurrent connections**

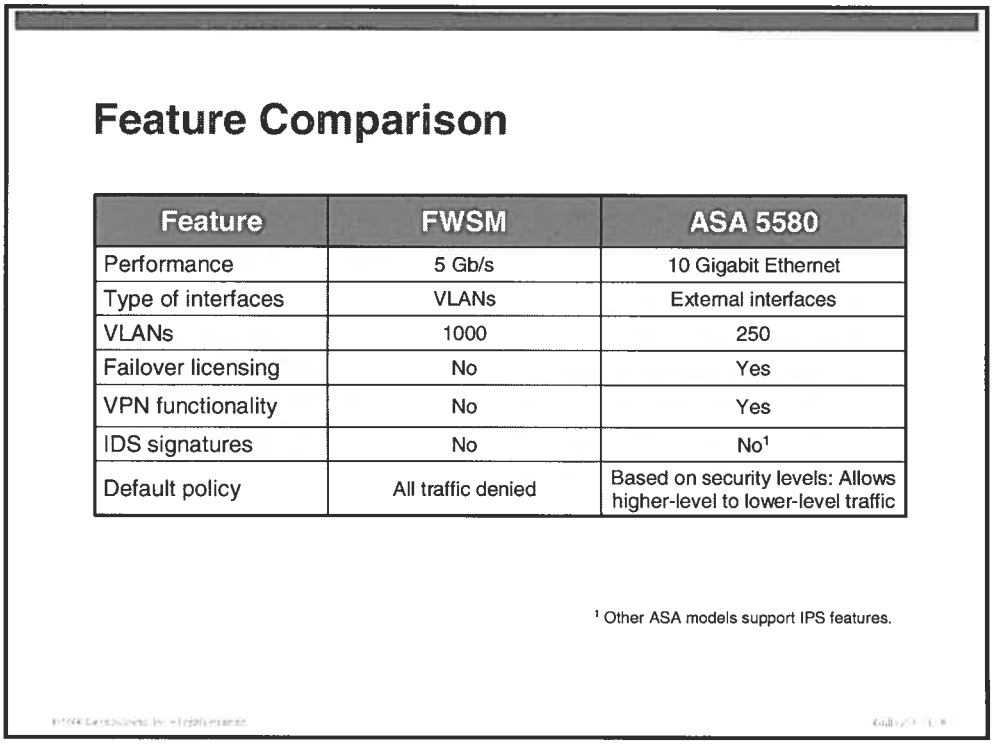
  The FWSM supports up to one million concurrent connections at any given time.

- **Supports 100,000 connections per second**

  Up to 100,000 connections can be established per second.

- **Multiple blades are supported in one chassis**

  In a single Catalyst 6500 series chassis, up to 4 FWSM modules can be supported.



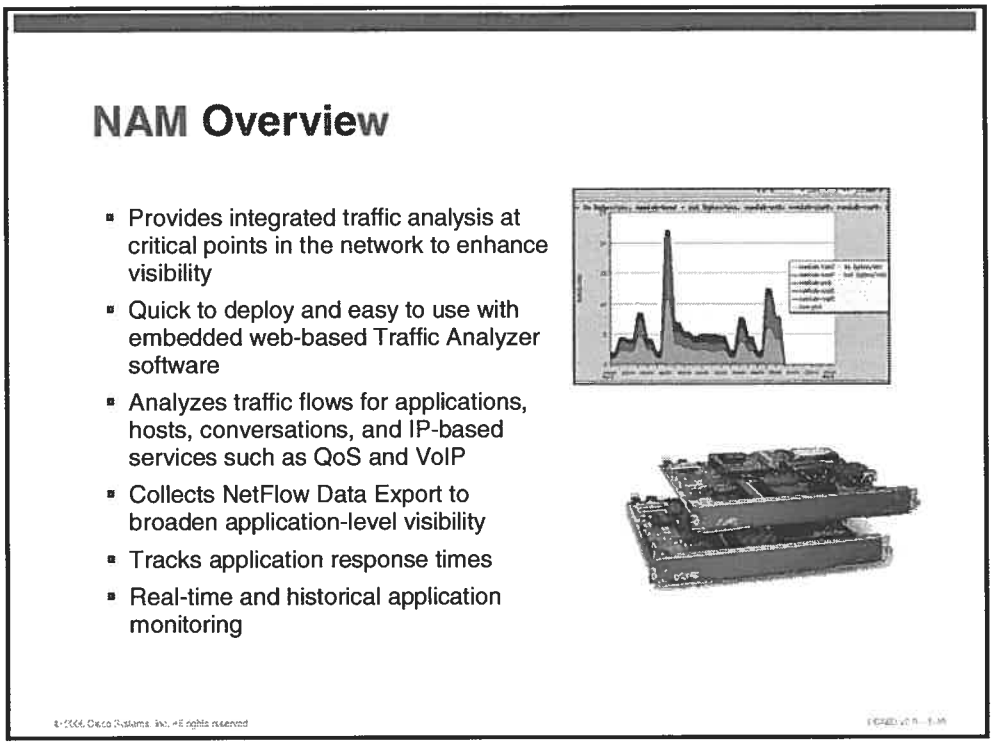
This figure lists the key differences between the FWSM and ASA. The FWSM is able to process more traffic than the ASA, but termination of virtual private network (VPN) connections for traffic flowing through the FWSM is not supported on a FWSM. The Catalyst 6500 provides intrusion protection services with the Cisco Intrusion Detection System Module (IDSM), and virtual private networking via VPN service modules (VPNSM).

https://t.me/learningnets

Integrated Network Analysis

This topic describes the basic characteristics of the Cisco NAMs.

Cisco Network Analysis Modules (NAMs)



Cisco has developed a second generation of Network Analysis Modules (NAMs) for Cisco Catalyst 6500 Series switches and Cisco 7600 Series routers. The second-generation Cisco NAMs are integrated and powerful traffic monitoring services modules that occupy a single slot in the chassis and provide application level visibility into network traffic to improve performance, reduce failures, and maximize returns on network investments.

Easy to Deploy and Use

The Cisco NAMs include an embedded, web-based Traffic Analyzer that delivers extensive monitoring and troubleshooting capabilities. Because the monitoring functions are integrated directly into the switch and have complete data collection and data analysis capabilities onboard, they are easy to deploy, and you can conveniently access data from anywhere using a web browser. For security purposes, you can give users role-based access and can secure web browser access with up to 168-bit encryption.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Many Data Sources Available for NAM Analysis

The Cisco NAMs use several data sources from local and remote switches and routers to provide combined visibility into LAN and WAN environments. The NAMs collect data from physical ports, virtual LANs (VLANs), or Cisco EtherChannel connections using the SPAN feature. For selective monitoring of large amounts of traffic or for traffic from WAN interfaces, VLAN access control list (VACL)-based captures can be used to filter traffic before it is sent to the NAM. In addition, the NAMs collect and analyze NetFlow data from local and remote devices to provide broad application level visibility, including remote WAN segments. The NAMs also collect data from remote switches and routers using the Remote Switched Port Analyzer (RSPAN) feature of the devices.

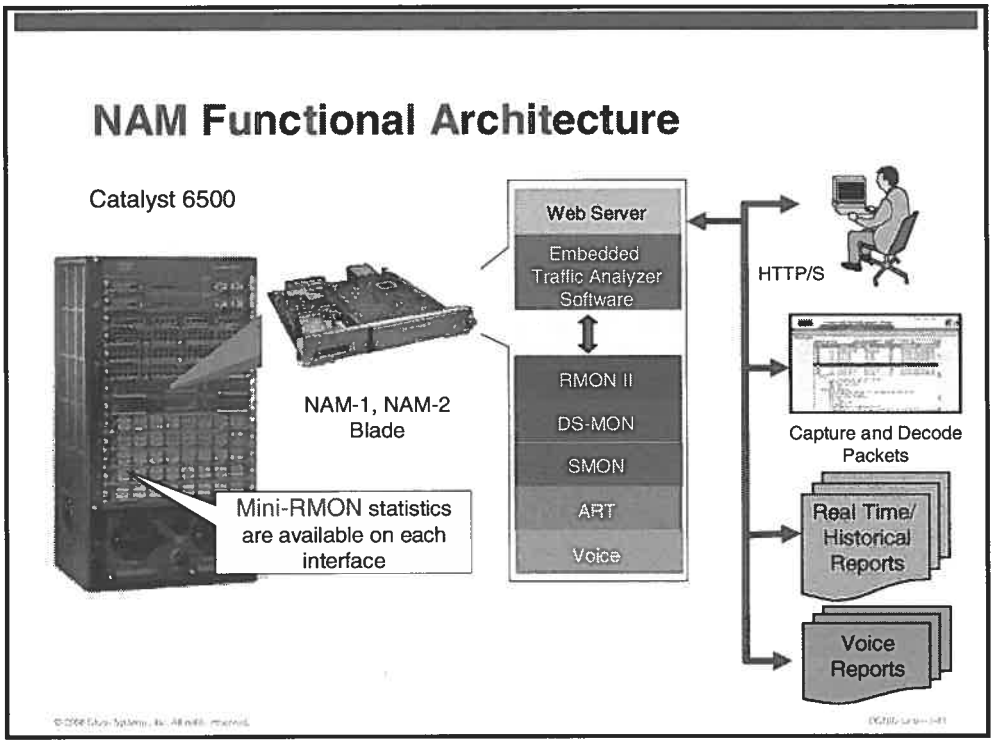
Application Response Time Analysis

The Cisco NAMs provide valuable information about delays in server responses to client requests. Using the Application Response Time (ART) Management Information Base (MIB), the NAMs can identify problems with applications or servers in critical environments such as e-commerce and IP telephony. With the URL monitoring feature, key statistics such as identifying frequently visited websites and the hosts visiting them, provide more granular information on the usage and activity of critical enterprise resources.

Real-Time and Historical Application Monitoring

The Cisco NAMs detect the applications on the network and, using Remote Monitoring (RMON), RMON2, and several extended RMON MIBs, provide detailed real-time and historical information on how these applications utilize the bandwidth, which hosts access the applications, and which client and server pairs generate the most traffic.

https://t.me/learningnets



The Cisco NAMs are available in two hardware versions, NAM-1 and NAM-2. Each version offers high performance monitoring and crossbar (fabric) connectivity to meet diverse network analysis needs in scalable switching and routing environments running at gigabit speeds. Included with the Cisco NAMs is an embedded, web-based Traffic Analyzer that provides full scale remote monitoring and troubleshooting accessible through a web browser.

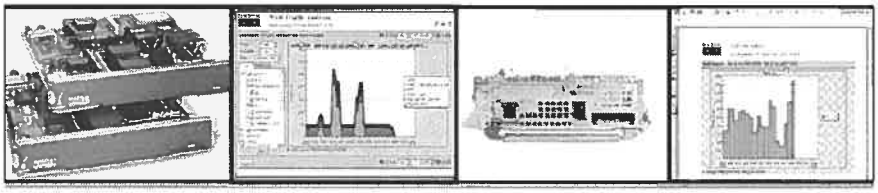
The Cisco NAMs provide visibility into all layers of network traffic by using RMON2 and other advanced MIBs. The NAMs access the built-in remote monitoring (mini-RMON) features of the Cisco Catalyst 6500 Series switches to provide port-level traffic statistics at the MAC or data-link layer. The NAMs also deliver the intelligence required to analyze traffic flows for applications, hosts, conversations, and network-based services, such as quality of service (QoS) and Voice over IP (VoIP).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NAM Benefits

- Using the NAM provides visibility into the data center from within the data center
- Allows administrators to be proactive and make the right decisions:
  - Make accurate business decisions about your IT resources
  - Identify traffic with greatest impacts to performance
  - Pinpoint latencies and isolate problems
- Size the network correctly to reduce network spending



Improperly managed networks lead to downtime and loss of access to important data, making users painfully aware that their data depends on a network.

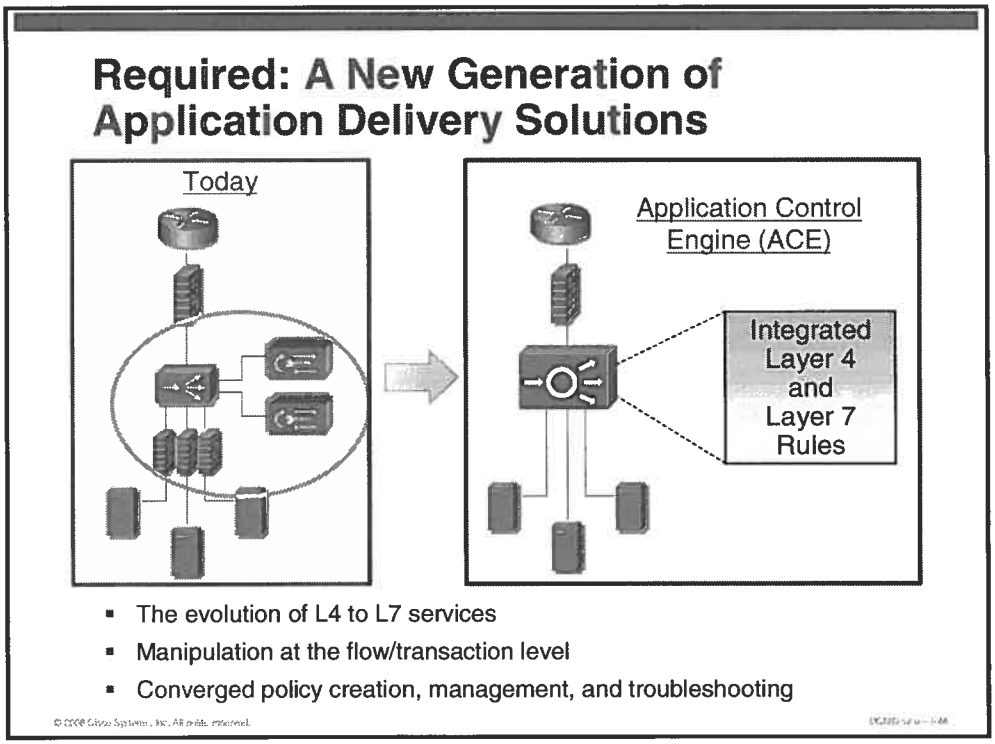
With performance monitoring, the key to decision making is the visibility within the network. A well thought out and implemented network management strategy provides users with a consistent high level of network services, increasing productivity. The collected management data also can be used to maximize return on investment (ROI), to verify third-party service agreements, and to quantify change and growth. Using the NAM leads to an overall increase in network reliability and effectiveness, in addition to saving money.

https://t.me/learningnets

Application Control Engine

This topic introduces the concept of server load balancing within the data center and describes the basic characteristics of the Cisco Application Control Engine Module (ACEM) and external appliance, which can be used to provide server load balancing and content switching functionality and ensure continual access to server farms and high availability.

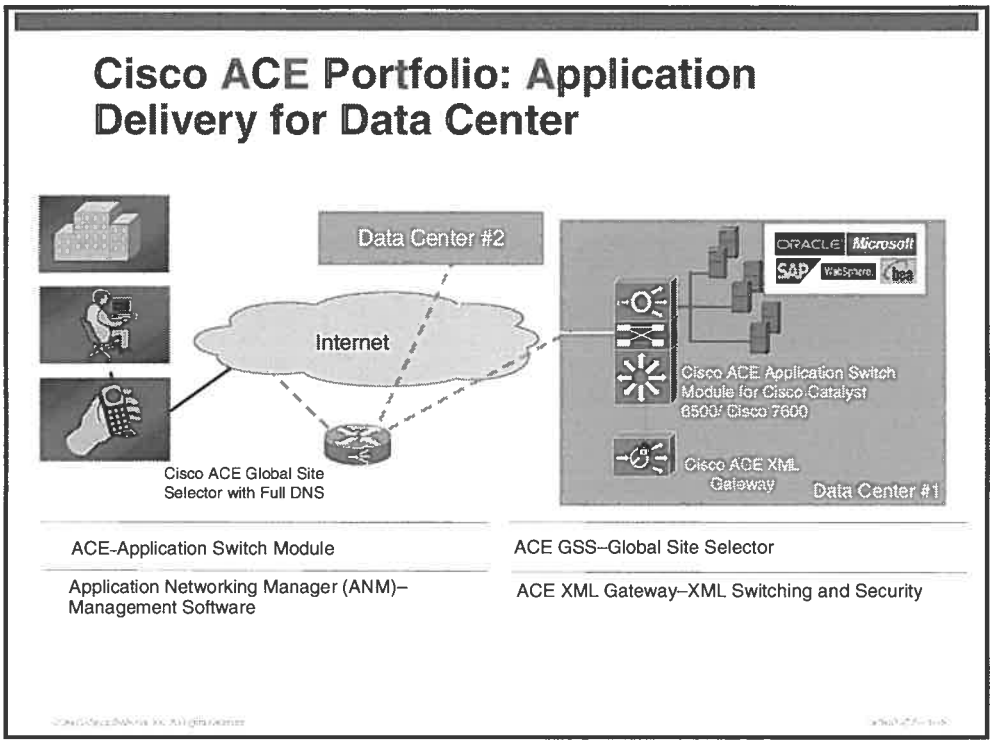
A New Generation of Application Delivery Solutions



A new generation of application delivery services is now consolidating application services on the common ACE software and hardware platform, integrating Layer 4 and Layer 7 load-balancing and security features along with SSL support.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

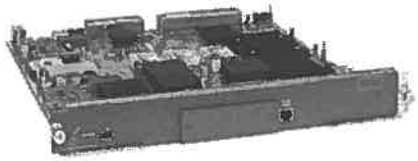


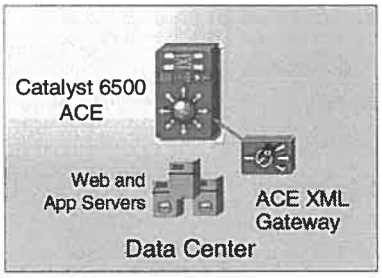
This figure describes the Cisco ACE portfolio:

- ACE appliance
- ACE GSS
- ACE XML gateway
- ACE application switch module.

https://t.me/learningnets

Cisco ACE Module: Application Switch





- Application availability
- Application acceleration
- Application security
- Virtualized architecture (ACE context)
- Server load balancing and content switching
- Support for Catalyst 6500 Switch and Cisco 7600 Router
- CSM ? ACE Modules

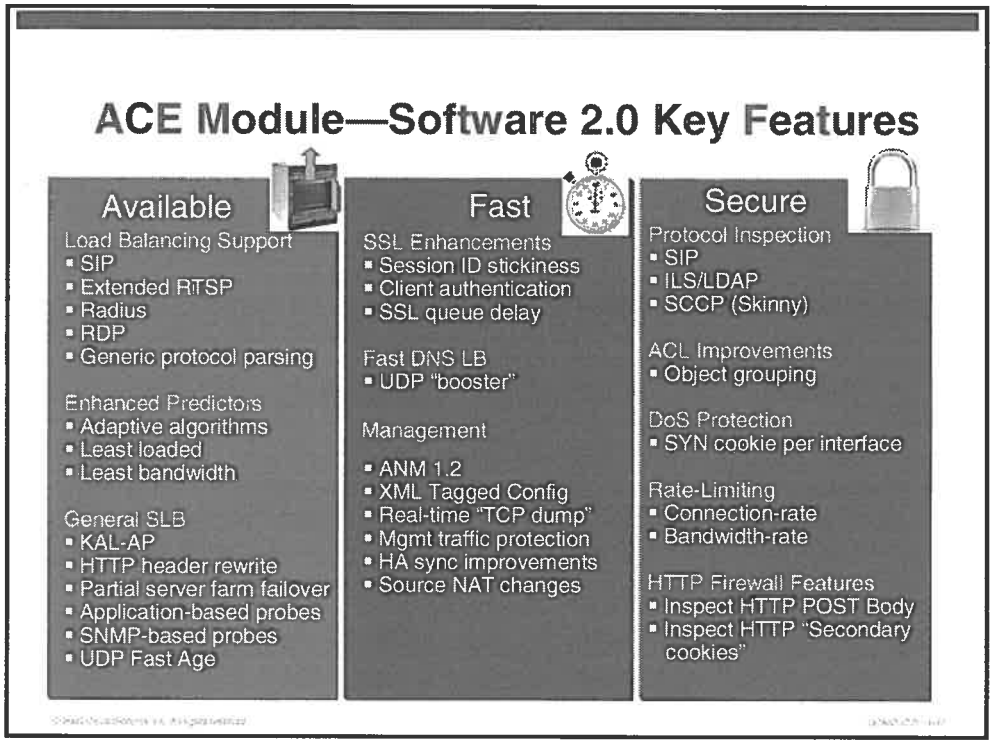
The Cisco ACEM provides the following advantages:

- Application availability
- Minimizes impact of application, device or site failure
- Maximizes availability through highly scalable (up to 64 Gb/s) load balancing and content switching
- Application acceleration
- Improves end user productivity with up to 300% faster applications
- Improves server performance by offloading SSL, TCP, XML processing
- Application security
- Comprehensive security for network, applications, XML and Web services
- Provides last line of defense for servers
- Protects against day zero attacks
- Virtualized architecture
- Provides faster application deployments and build-outs
- Increases IT productivity by deploying virtual devices, not physical devices
- Virtual devices accelerate application roll-out

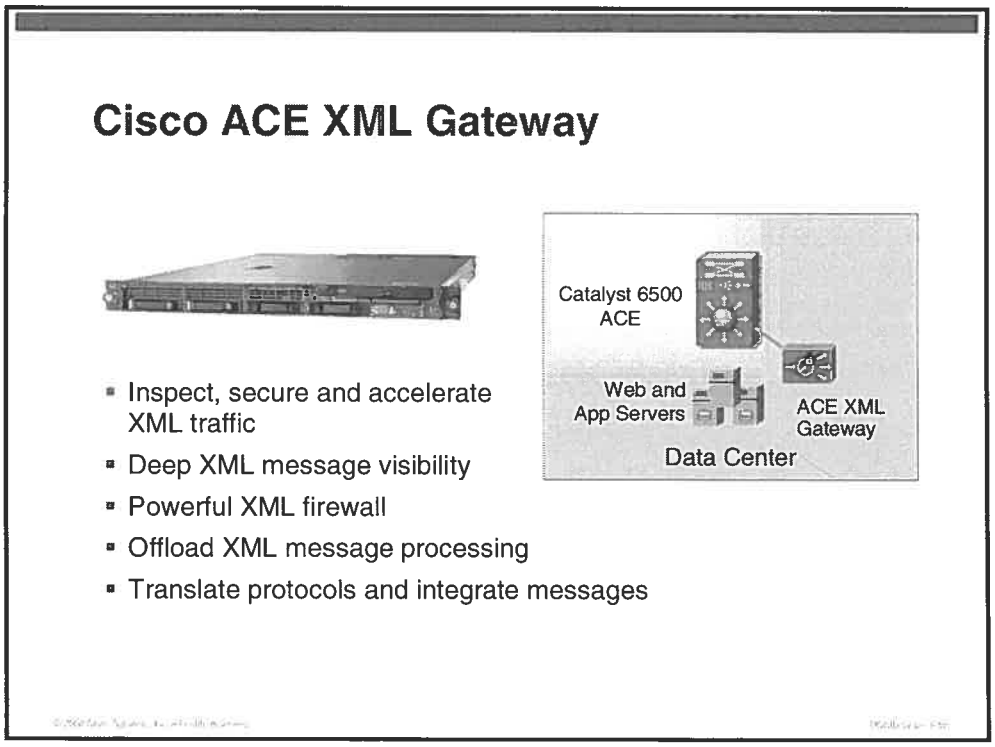
The Cisco ACEM is a CEF720 fabric-enabled line card connecting to the switch fabric with a single 20 Gb/s connection. ACE is supported only in Cisco IOS native mode and requires Cisco IOS 12.2(18) SXF4 or above. Four ACE modules can be deployed in a single Catalyst 6500 chassis. ACE configuration and management activities are handled by resources on the ACE module and require minimal support and interaction with the Catalyst supervisor and its configuration.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The ACE 2.0 software runs on an ACE appliance or an ACE module and supports many features listed in the figure above.



The Cisco ACE XML Gateway is used in the data center to inspect, secure and accelerate XML traffic. The Cisco ACE XML Gateway supports deep XML message visibility and a powerful XML firewall, and offloads XML message processing from servers.

https://t.me/learningnets

Cisco ACE Global Site Selector (GSS): Ensuring Data Center to Data Center Continuity





Cisco ACE Global Site Selector with Full DNS

s of user requests across distributed data centers within 1 second

- Intelligently selects best data center to ensure application SLAs
- Fully integrated DNS server
- Reacts to changes in local and global networks
- Scalable across multiple data centers, servers, and switches

Cisco ACE Global Site Selector (GSS) provides failover for thousands of user requests across the distributed data center within 1 second. Cisco ACE GSS intelligently selects the best data center to ensure application SLAs. It has a fully integrated DNS server and reacts to server availability changes in local and global networks with integrated server load balancers.

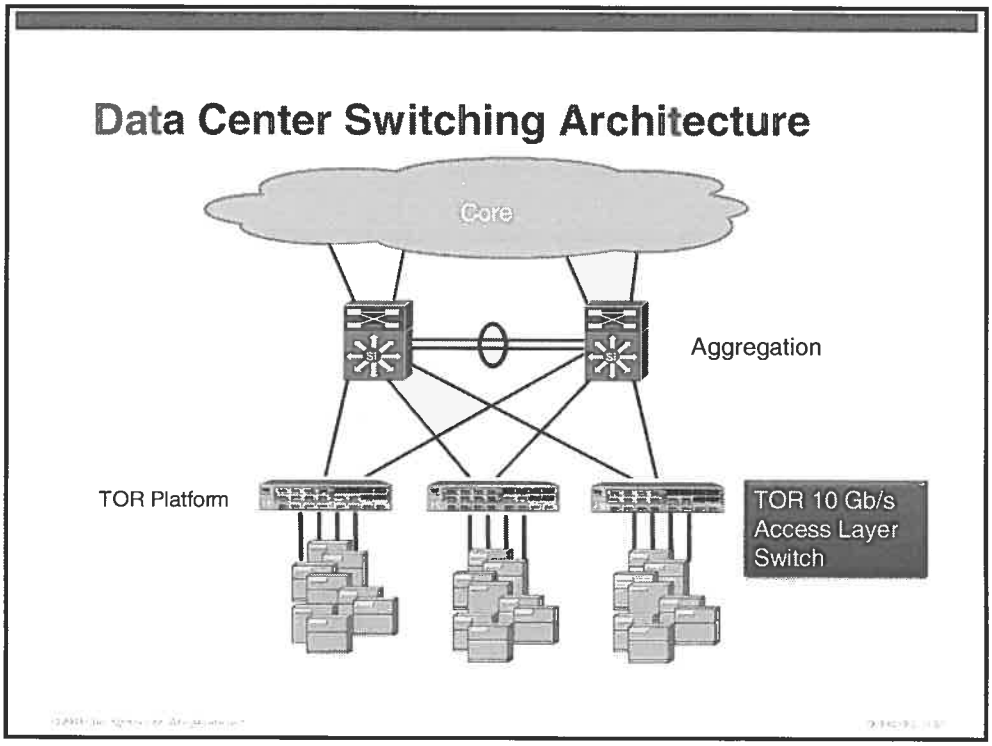
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Catalyst 4900M Series Switches

This topic introduces the characteristics of the Cisco Catalyst 4900M switches.

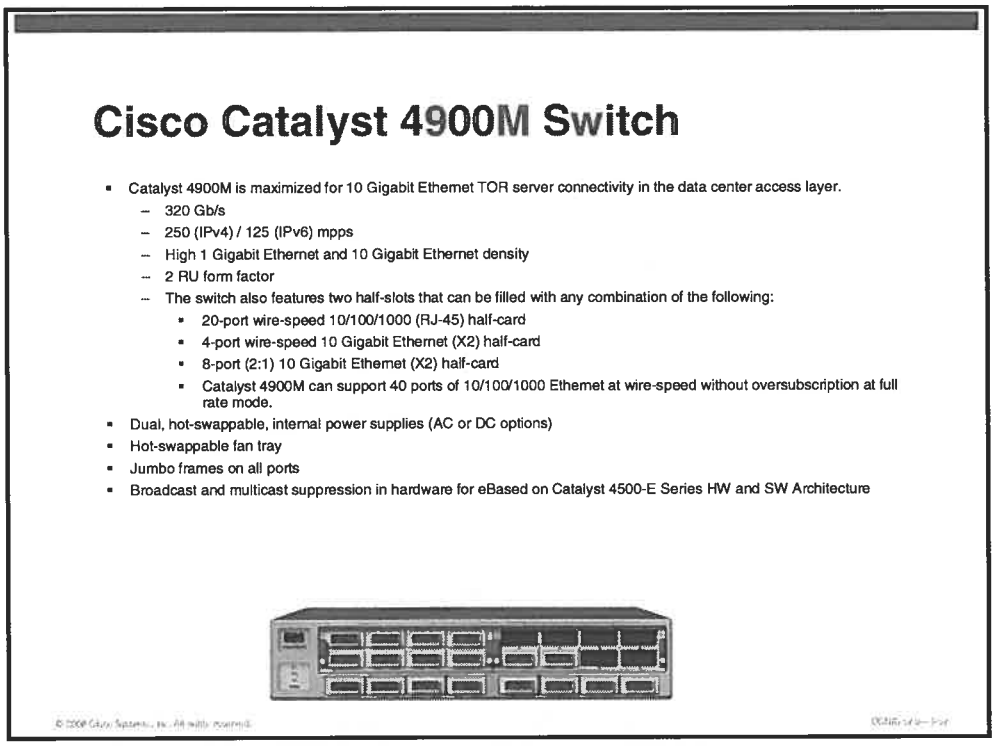
Data Center Switching Architecture



This topic introduces the Catalyst 4900M Series Switch, the next generation top-of-rack switch. The Catalyst 4900 M is a semi-fixed switch in 2RU, and its purpose is to facilitate the transition between 1 Gigabit Ethernet and 10 Gigabit Ethernet server access layer in the data center.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Cisco Catalyst 4900M Series is part of the Cisco Data Center switching portfolio. The Cisco Catalyst 4900 Series switches are designed as a top-of-rack access layer switch for use with rack-optimized servers. The Cisco Catalyst 4900M is purpose built as a top-of-rack switch and delivers the buffering, latency, and performance characteristics necessary to participate in the data center access layer.

The Cisco Catalyst 4900M is a two-rack-unit (2RU) top-of-server rack Ethernet switch optimized for 10/100/1000 and 10 GE access devices. It is unique in the top-of-rack space because it is semi-fixed. The design of the Cisco Catalyst 4900M addresses several challenges that data centers face today:

- Offers the benefit of hot-swappable modules
- Provides high availability
- Optional half cards can be changed as new servers require different media and port speeds, which helps to maintain the investment of the Cisco Catalyst 4900M base unit.
- 16 MB of shared buffers help prevent performance degradation under very heavy traffic loads.

The switch has 8 fixed wire-speed 10 GE ports that are typically used to uplink to an aggregation switch.

The Catalyst 4900M switch also features two half-slots that can be filled with any combination of the following:

- 20-port wire-speed 10/100/1000 (RJ-45) half-card
- 4-port wire-speed 10 GE (X2) half-card
- 8-port (2:1) 10 GE (X2) half-card (Cisco TwinGig Converter Module compatible)

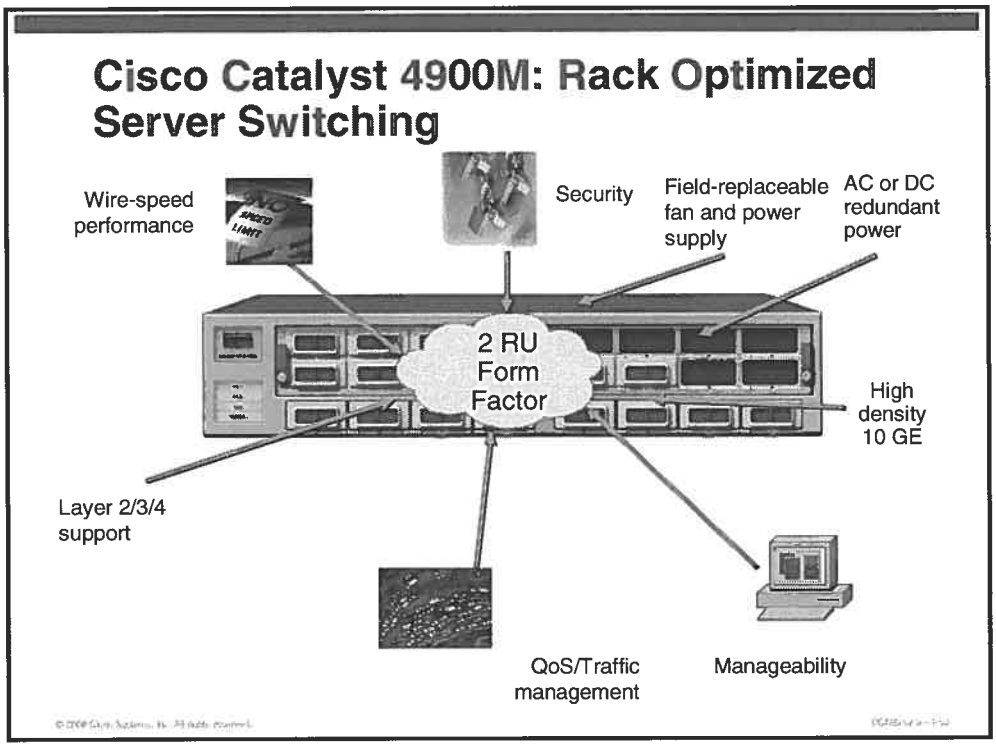
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

The Catalyst 4900M is maximized for 10 GE server connectivity in the data channel access layer.

The Catalyst 4900M can support 40 ports of 10/100/1000 Ethernet at wire-speed without oversubscription at full rate mode. The Catalyst 4900M switch allows customers to transition from 1 Gigabit Ethernet to 10 Gigabit Ethernet in data channel racks and transition to 10 BaseT in the future.

https://t.me/learningnets



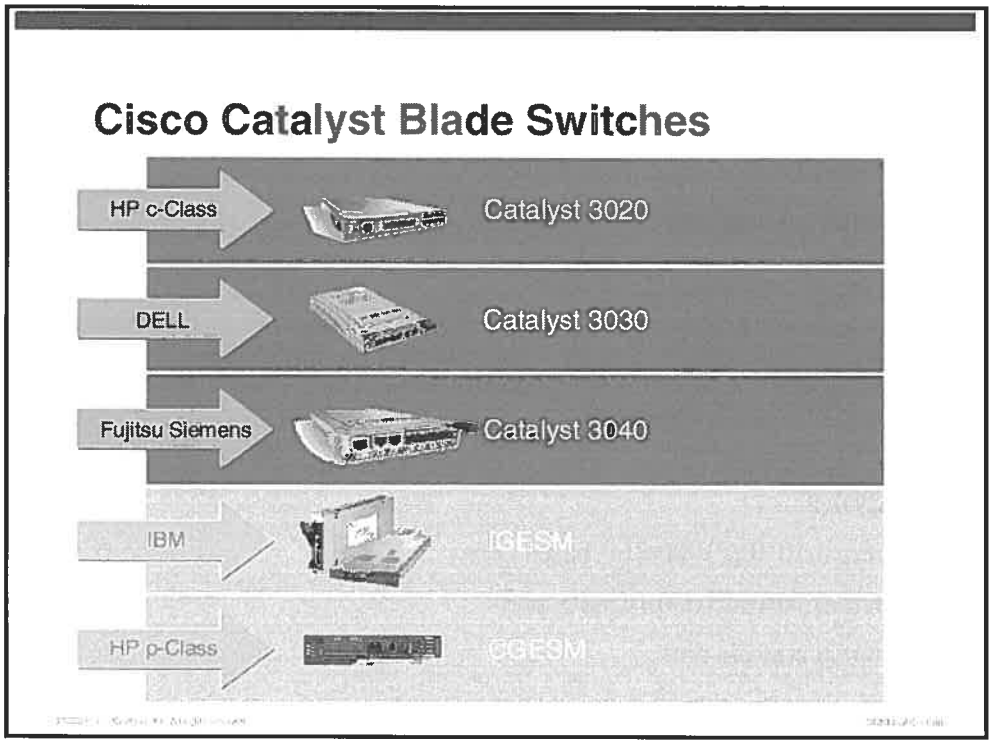
The Cisco Catalyst 4900M switch is a 2 RU optimized server switching platform that supports wire-speed performance, high density 10 Gigabit Ethernet ports, security features, field-replaceable fan and power supply, traffic management and QoS features, and Layer 2, Layer 3, and Layer 4 switching functions.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Blade Switches

This topic describes the basic characteristics of Cisco Ethernet blade switches



Cisco offers three Catalyst-based blade switches. All Catalyst Blade Switches offer a complete set of intelligent services to deliver security, QoS, and availability in the server farm access environment. A Catalyst Blade Switch extends Cisco infrastructure services to the server edge and uses existing network investments to help reduce operational expenses.

Cisco Catalyst Blade Switch 3020 for HP

The Cisco Catalyst Blade Switch 3020 for HP provides HP c-Class Blade System customers with an integrated switching solution. The Cisco Catalyst Blade Switch 3020 for HP provides the following hardware configuration:

- Sixteen internal 1000BASE ports connected to servers through the c-Class Blade System backplane.
- Up to eight external GE uplink ports:
  - Four external dual-media Ethernet interfaces; interfaces can be either 1000BASE-SX SFP or 10/100/1000BASE-T ports.
  - Four additional external 10/100/1000BASE-T ports; two of these ports can be configured to provide an internal crossover connection to an associated additional Cisco Catalyst Blade Switch 3020.
- One Fast Ethernet connection to the HP Enclosure Onboard Administrator.
- One external console port.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Catalyst Blade Switch 3030 for Dell

Designed for the Dell Power Edge Blade Server Enclosure, the Cisco Catalyst Blade Switch 3030 provides the following hardware configuration:

- 10 internal 1000BASE ports connected to servers through the PowerEdge enclosure backplane
- 4 external 10/100/1000 SFP-based ports; supports copper and fiber (SX) modules from Cisco Systems
- 2 external 10/100/1000BASE-T ports
- 1 external console port

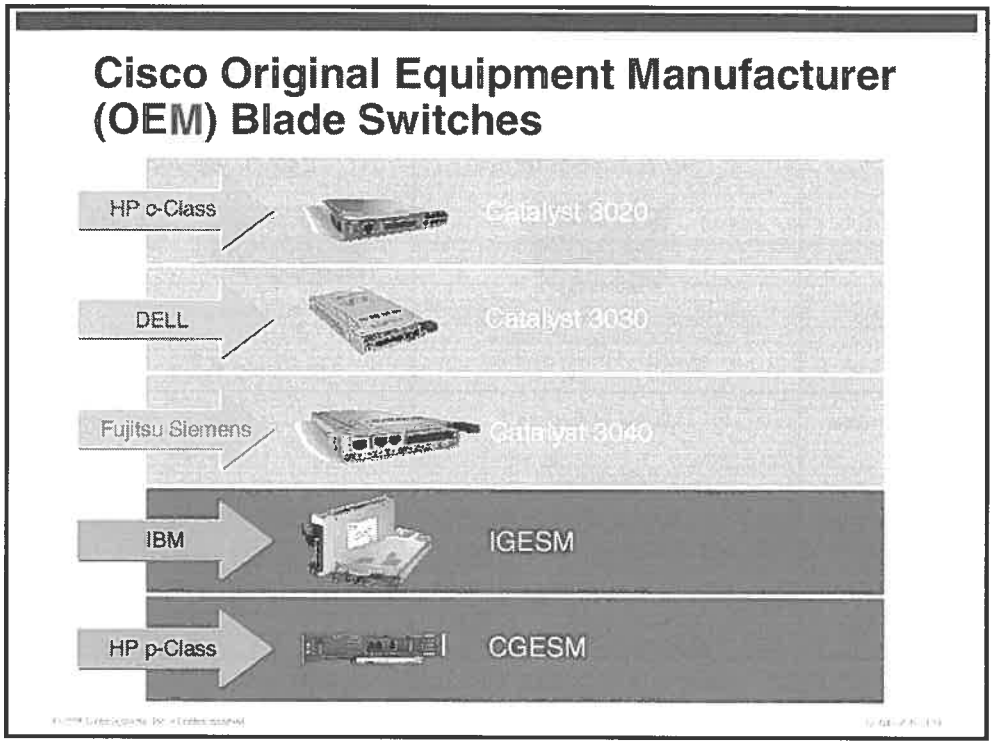
Catalyst Blade Switch 3040 for FSC

The Catalyst Blade Switch 3040 for FSC is an integrated switch for Fujitsu Siemens Computers (FSC) PRIMERGY Blade Server customers. The Catalyst Blade Switch 3040 provides the following hardware configuration:

- 10 internal 1000BASE ports connected to servers through the PRIMERGY chassis backplane
- 4 external 10/100/1000 SFP-based ports
- 2 external 10/100/1000BASE-T ports
- 1 external console port

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco also provides two OEM blade switches:

- The Cisco Intelligent Gigabit Ethernet Switch Module (IGESM) for IBM Blade Center
- The Cisco Gigabit Ethernet Switch Module (CGESM) for the HP p-Class Blade System

These custom switches offer many of the same features, benefits, and value as Catalyst switch blades, but are designed specifically for IBM and HP blade products.

https://t.me/learningnets

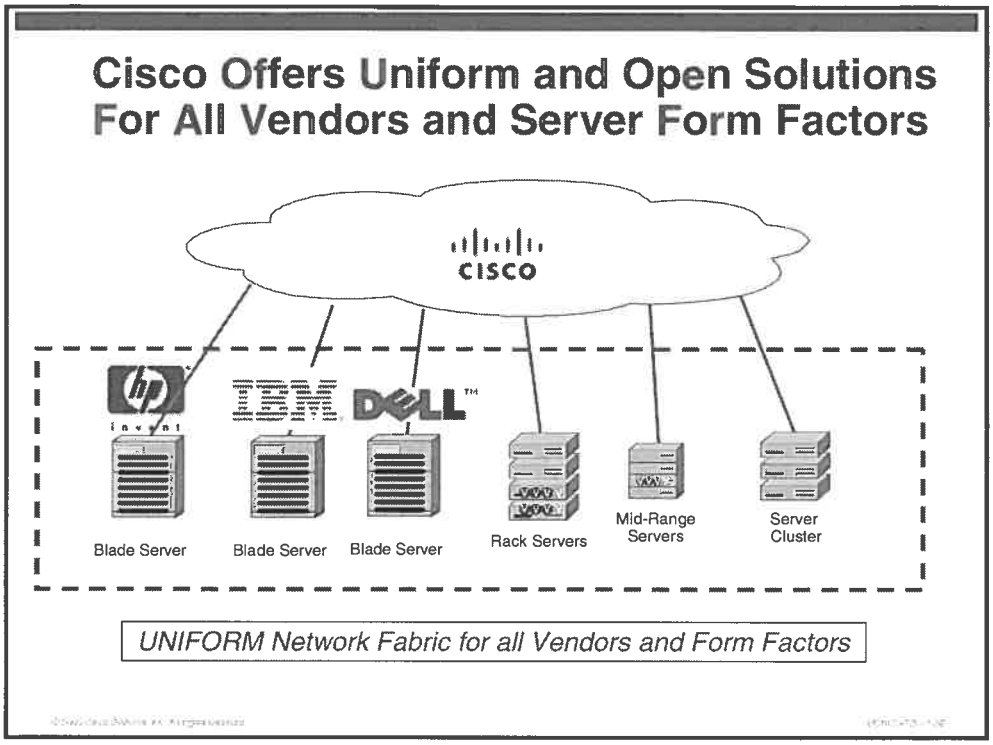
Blade Server Offerings from Server Vendors
HP and IBM Overview

	HP BladeSystem	*IBM BladeCenter*
		 
Chassis	*C-Class Enclosures (10 G capable)* *▪ C7000–16 server chassis* *▪ C3000–8 server chassis*	*BladeCenter Chassis* *▪ BC E–14 server chassis* *▪ BC H–14 server chassis (10 G capable)* *▪ BCT & BCHT–Telco chassis*
Switch Slots	*Up to 8 switch slots*	*▪ BC E – 4 standard switch slots* *▪ BC H – 4 standard switch slots; 4 High-speed slots*
Servers	*half-height or full-height servers*	*full-height servers*
Processor	*Intel, AMD and Itanium*	*Intel, AMD, Power and Cell*
Memory	*Up to 64 G*	*Up to 32 G*
Storage	*▪ Each server supports up to 2 disks* *▪ Also supports Storage blades*	*▪ Each server can support up to 2 disks* *▪ Also supports Storage blades*

Integrated blade switches present a third cabling design option in addition to end-of-row and top-of-rack. From a logical network perspective, a cabling design that uses integrated blade switches is most similar to a top-of-rack design, although the blade switch design can introduce complications for the spanning tree design because there are more access layer switches per rack.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Virtual Blade Switch (VBS) and Top-of-Rack Blade switches enable a unified fabric access layer connection to Cisco Data Center distribution layer switches.

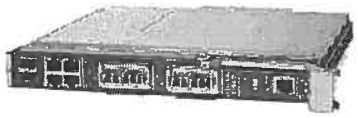
https://t.me/learningnets

Cisco Catalyst Blade Switches for Data Center Access Layer

BS 3030 for Dell PowerEdge BS

- 10 internal 1000Base-T Ethernet ports
- 4 external 1 Gigabit Ethernet ports

BS 3032 for Dell M1000E BS Enclosure



- 16 internal 1000Base Ethernet ports
- 8 external 1 Gigabit Ethernet uplink ports
- 48 Gb/s switching fabric (36Mpps)

BS 3130 for Dell 1000E BS Enclosure

- 3130X : 4 * 1 Gigabit Ethernet uplink + 2 * 10 Gigabit Ethernet ports
- 3130G : 8 * 1 Gigabit Ethernet uplink + 2 * 10 Gigabit Ethernet ports
- 64 Gb/s inter-switch links
- 160 Gb/s EtherChannel (Active-Active uplink forwarding)

The figure shows various Cisco Catalyst Blade switches for the data channel access layer.

The Cisco Catalyst Blade switch 3030 for the Dell PowerEdge Blade Server is an integrated switch that extends Cisco infrastructure services to the data center server edge farm and uses existing network investments to help reduce operational expenses.

The Cisco Catalyst Blade Switch 3030 for the Dell Power Edge Blade Server enclosures is the optimum choice for Small and Medium data center environment as well as for the Enterprise customers who need to deploy a secure, resilient and highly manageable Blade Server solution.

Cisco Catalyst 3030 Blade switch supports 10 internal 1000BASE-T Ethernet ports connected to servers through the Dell PowerEdge Blade Server enclosure midplane and 4 external 10/100/1000 Small Form-Factor Pluggable (SFP)-based ports. The blade switch supports copper and fiber 1000BASE-SX SFP Cisco Systems modules. Cisco Catalyst 3030 Blade switch has 2 external 10/100/1000BASE-T Ethernet ports and 1 external Console Management ports. Cisco Catalyst switch 3030 simplifies the data channel cable complexity in the rack and the server to fabric interconnections, releases the space constraints in high-density data channels, simplifies data channel manageability to the server edge, provides Cisco security IOS features and supports GUI device manager and Cisco Works for management purposes. Cisco Catalyst Blade Switch 3030 decreases cable complexity by up to 90 percent and completely eliminates the required cabling between servers and the integrated switch. This switch also reduces space requirements by providing an integrated switch in a chassis that can be more densely populated than the traditional server rack. Cisco Catalyst Blade Switch 3030 helps ensure consistency across the network for advanced services such as security and resiliency. Cisco supports network security features supported by the Cisco IOS Software LAN base image such as Extended Access Control Lists (ACLs) and Quality of Service (QoS) traffic policies to help mitigate flooding and attacks. The switch is also Network Admission Control (NAC) compliant. NAC allows for effective admission control, posture validation, and virus detection at the network level, before serious damage can reach servers or end stations. The Catalyst blade Switch 3030 also incorporates many resiliency features such as FlexLink and Trunk Failover, to create a highly available integrated switch. FlexLink can co-exist with

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

spanning tree and allows convergence time of less than 50 milliseconds on switch uplink ports providing incredible spanning-tree recovery intervals. Trunk Failover allows rapid failover to the redundant switch in the PowerEdge Blade Server Enclosure if all uplinks from primary switch fail, allowing uninterrupted network performance.

The Cisco Catalyst Blade Switch 3032 for Dell M1000e is an integrated switch for Dell M1000e Blade Server Enclosure. The Cisco Catalyst Blade Switch 3032 for Dell M1000e provides the following hardware configuration:

- 16 internal 1000BASE ports connected to servers through the Dell M1000e backplane
- Up to 8 GE uplink ports: 4 10/100/1000BASE-T ports and 4 SFP GE ports
- One external Console port

The Cisco Catalyst Blade Switch 3032 uses Cisco IOS Software with the IP Base image and offers a complete set of intelligent services to deliver security, QoS, basic IP routing, and high availability in the server farm access environment. The Cisco Catalyst Blade Switch 3032 offers a wide range of security features such as:

- Denial-of-Service (DoS) guard and other attack guards
- Access Control Lists (ACLs) based on source and destination MAC addresses and IP addresses, and TCP and User Datagram Protocol (UDP) ports
- ACL lookups are performed in hardware and forwarding performance is not compromised
- Port security
- Secure Shell (SSH) Protocol
- Kerberos Protocol
- Simple Network Management Protocol Version 3 (SNMPv3)
- TACACS+ and RADIUS authentication for centralized access control
- Web-based management interfaces provide the capability to give different levels of configuration capabilities to different administrators
- The MAC address notification feature can be used to monitor the network and track servers by sending an alert to a management station so that network administrators know when and where servers are plugged into or removed from a blade enclosure.

The Dynamic Host Configuration Protocol (DHCP) Interface Tracker (Option 82) feature can provide location-based IP address assignment by providing both the switch and the port ID to a DHCP server. An Option 82-aware DHCP server such as the Cisco Network Registrar can use this information to assign the specific IP address to the requesting server. The Private VLAN and Private VLAN Edge features isolate ports on a switch, helping ensure that traffic travels directly from the entry point to the aggregation device through a virtual path and cannot be directed to another port, thus helping to isolate a server from other servers in the same blade enclosure. The Cisco Catalyst Blade Switch 3032 offers several high-availability features to minimize network downtime, maintain mission-critical servers and applications, and reduce total cost of ownership (TCO). Enhancements to the standard Spanning Tree Protocol, such as Per-VLAN Spanning Tree Plus (PVST+), UplinkFast, and PortFast, maximize network uptime. UplinkFast and PortFast help reduce the standard 30- to 60-second Spanning Tree Protocol convergence time. Loop Guard and Bridge Protocol Data Unit (BPDU) Guard provide Spanning Tree Protocol loop avoidance. The Cisco Catalyst Blade Switch 3032 uses redundant power and cooling capabilities of the Dell PowerEdge M1000e infrastructure to provide maximum availability to customers. The Cisco Catalyst Blade Switch 3032 offers granular QoS

https://t.me/learningnets

features used to classify, police, mark, queue, and schedule incoming packets and can queue and schedule packets at egress. Packet classification allows the network elements to discriminate between traffic flows and enforce policies based on Layer 2 and Layer 3 QoS fields. To implement QoS, the Cisco Catalyst Blade Switch 3032 first identifies traffic flows or packet groups and classifies or reclassifies these groups using the differentiated services code point (DSCP) field or the IEEE 802.1p Class-of-Service (CoS) field. Classification can be based on criteria as specific as the source or destination IP address, source or destination MAC address, or Layer 4 TCP or UDP port. At ingress, the Cisco Catalyst Blade Switch 3032 will also police to determine whether a packet is in or out of profile; mark to change the classification label, pass-through, or dropout of profile packets; queue packets based on classification; and queue services based on configured weights. Control plane and data plane ACLs are supported on all ports to help ensure proper treatment on a per-packet basis. The Cisco Catalyst Blade Switch 3032 supports four egress queues per port, which allows the network administrator to be discriminating and specific in assigning priorities for the applications in the server farm. At egress, the switch performs scheduling and congestion control. The Cisco Catalyst Blade Switch 3032 supports Shaped Round Robin (SRR) and strict priority queuing. The SRR queuing algorithm helps ensure differential prioritization. The Catalyst Blade Switch 3032 comes with an embedded GUI device using SNMP network management platforms such as CiscoWorks. As part of CiscoWorks, the Web-based CiscoWorks Resource Manager Essentials (RME) offers automated inventory collection, software deployment, easy tracking of network changes, views into device availability, and quick isolation of error conditions. The Catalyst Blade Switch 3032 offers customers high-performance, basic IP routing. It uses Cisco Express Forwarding (CEF/dCEF) hardware routing architecture to deliver basic IP unicast routing protocols that include static routing, Routing Information Protocol, and Cisco Enhanced Integrated Gateway Routing Protocol (EIGRP) Stub. This blade switch does not support Open Shortest Path First (OSPF) Protocol and Border Gateway Protocol (BGP).

Cisco Catalyst 3130 Series Switch Blade Servers are playing an important role in networking solutions because they have access to enterprise-class networking, including availability, performance, and end-to-end network services. The Cisco Catalyst Blade Switch 3100 Series represents the next generation of networking solutions for blade server environments. Based on the Cisco hardware and IOS software, the Cisco Catalyst Blade Switch 3100 Series is engineered with innovative technologies specifically designed to meet the rigors of blade server based application infrastructure. Specifically, the switch is designed to support blade servers in their new role delivering scalable, high-performance, highly resilient connectivity while supporting ongoing initiatives to reduce server infrastructure complexity and total cost of operation (TCO).Cisco has coordinated development with Dell to bring the Cisco Blade Switch 3100 series to the Dell M1000E Blade Server enclosure with the release of the Cisco Blade Switch 3130G and 3130X. While offering identical functionality, the CBS 3130G supports up to 8 Gigabit Ethernet uplinks while the CBS 3130X supports 4 Gigabit Ethernet uplinks and 2 10 GE ports.

© 2008 Cisco Systems, Inc.

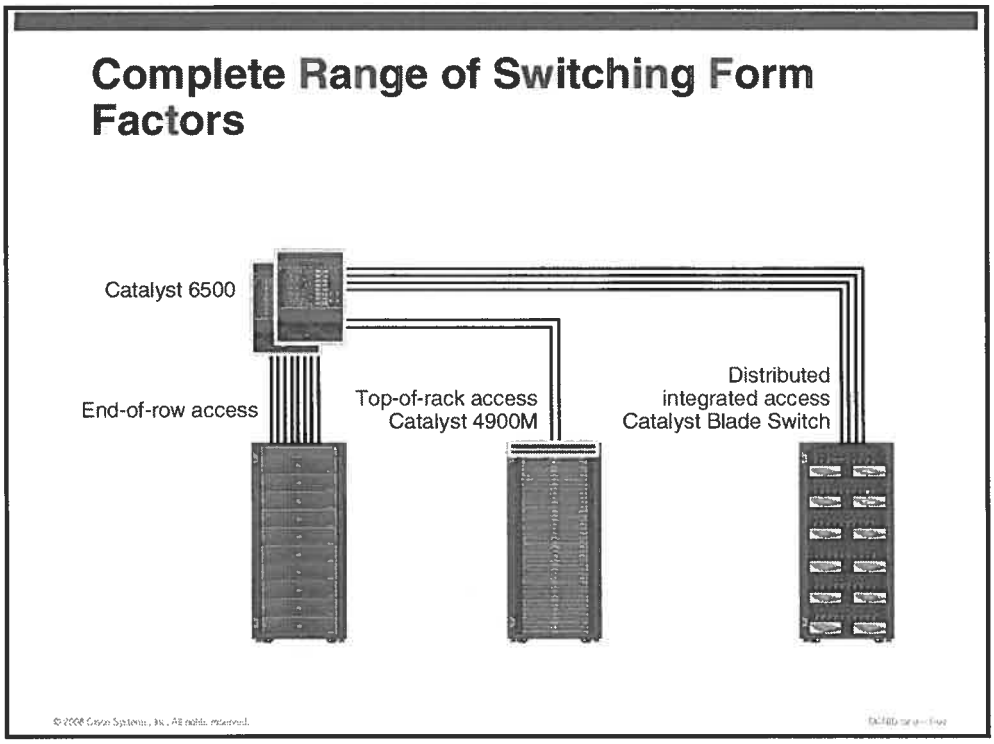
https://t.me/learningnets

The Virtual Blade Switch

With the Cisco Catalyst Blade Switch CBS 3130, Cisco introduces a unique technology called the virtual blade switch (VBS). Much like server virtualization technology, this switch virtualization technology treats the individual physical switches within a rack as one logical switch, an innovation that allows the switches to deliver better utilization, increased performance, and improved resilience while simplifying operations and management. The VBS appears as a single network device, so the network topology is greatly simplified, enabling quicker deployment, a more robust network, and faster troubleshooting and problem resolution. The VBS delivers a number of advantages to the supported servers: by significantly simplifying data center design and operation, the VBS reduces infrastructure complexity of management and improves the network resiliency. By providing up to 160 Gb/s upstream, the VBS offers exceptional performance. The VBS can double the bandwidth available to a server. The VBS makes extensive use of link virtualization for both upstream connectivity and server connectivity, providing the benefits of increased available bandwidth and increased redundancy. The VBS can recover from a physical link failure without having to reconverge the network, thus avoiding session drops. The VBS uses the same Cisco IOS Software interface, MIBs, and management tools as the rest of the Cisco Catalyst Series, which simplifies operations and management by the network operations team and helps ensure that attached servers can fully participate in end-to-end network services such as QoS and security policies, thus helping maintain regulatory compliance.

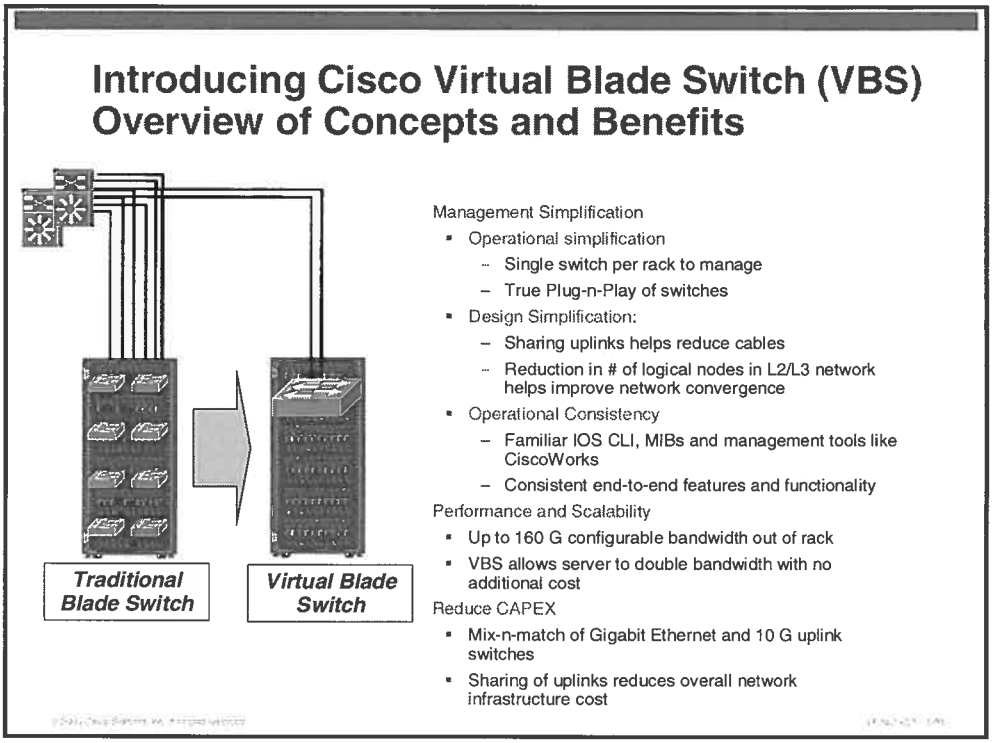
The Cisco Catalyst Blade Switch 3130 can mix and match 1 GE and 10 GE switches, providing customers with a cost-effective migration path. Switch capacity can be incrementally upgraded as server needs demand. Simplified infrastructure through VBS technology makes a rack full of switches appear as a single node at Layers 2 and 3. It simplifies management and operations through integration with Cisco VFrame immediately activated provisioning and Cisco IOS Embedded Event Manager (EEM), and Cisco Generic Online Diagnostics (GOLD) reduces power, cooling, and cable management requirements. High performance and scalability through Cisco EtherChannel provides superior performance and flexibility, provides 160 Gb/s aggregate upstream bandwidth, and doubles server bandwidth through active-active configuration. The Cisco Catalyst Blade Switch 3130 provides 64 Gb/s inter-switch VBS fabric. Continuous operation supports active-active server links through Cisco EtherChannel, provides link resiliency and aggregation layer diversity through VBS and avoids Layer 2 or 3 reconvergence after link or switch failure. Simple Layer 2 and 3 topologies also increase network stability. Switch modules can be added or removed without service interruption in remaining modules. Software and configuration is automatically remediated. Manageability offers integration with original equipment manufacturer (OEM) element manager and can be managed like any other switch running Cisco IOS Software, using the same tools (such as Cisco EEM and Cisco GOLD) and procedures through its consistent Cisco IOS Software feature implementation. Manageability also supports end-to-end features such as QoS.

https://t.me/learningnets



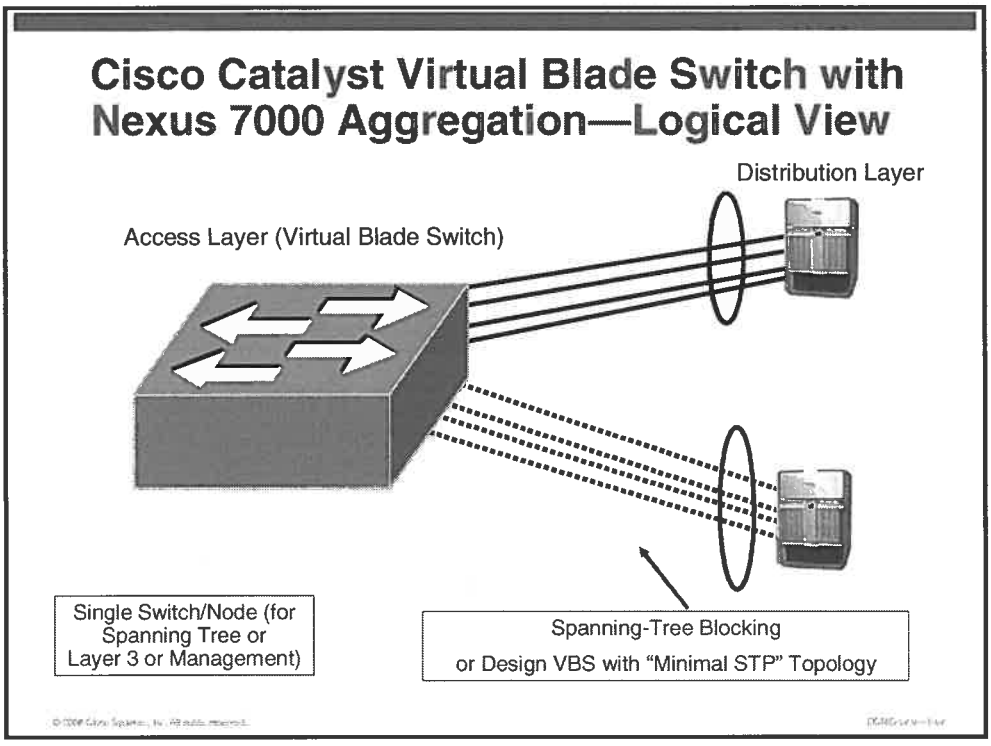
Integrated blade switches present a third cabling design option, in addition to end-of-row and top-of-rack. From a logical network perspective, a cabling design that uses integrated blade switches is most similar to a top-of-rack design, although the blade switch design can introduce complications for the spanning tree design because there are more access layer switches per rack.

https://t.me/learningnets

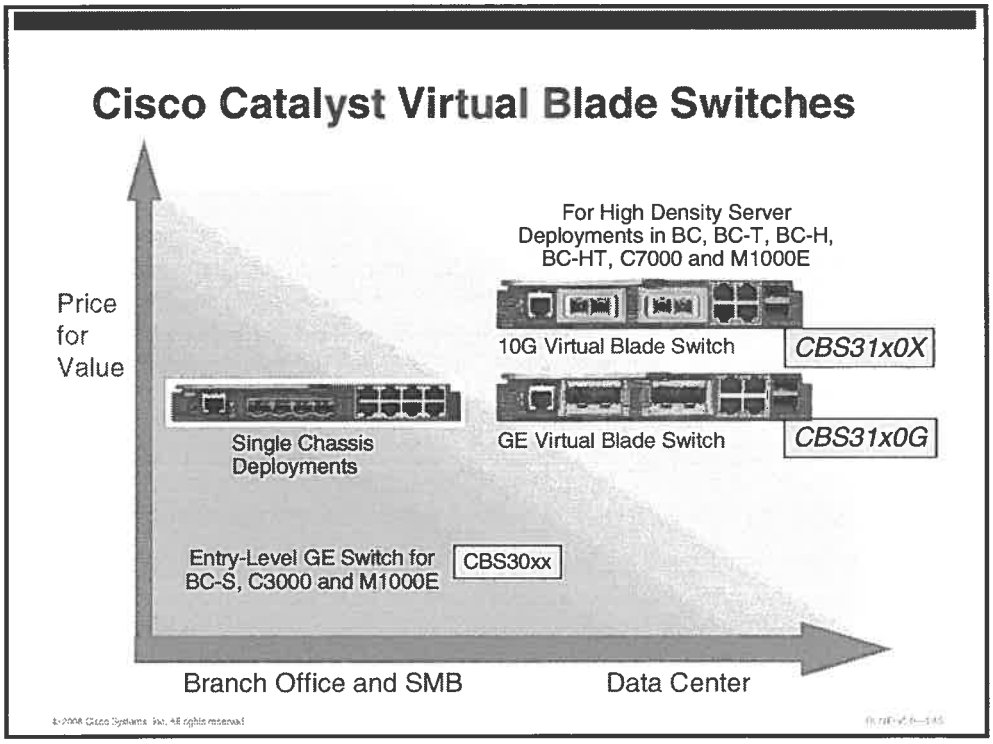


The Cisco VBS significantly simplifies management of the data channel blade switching access layer.

https://t.me/learningnets



The figure represents the logical topology view of the VBS deployed in the data channel access layer.



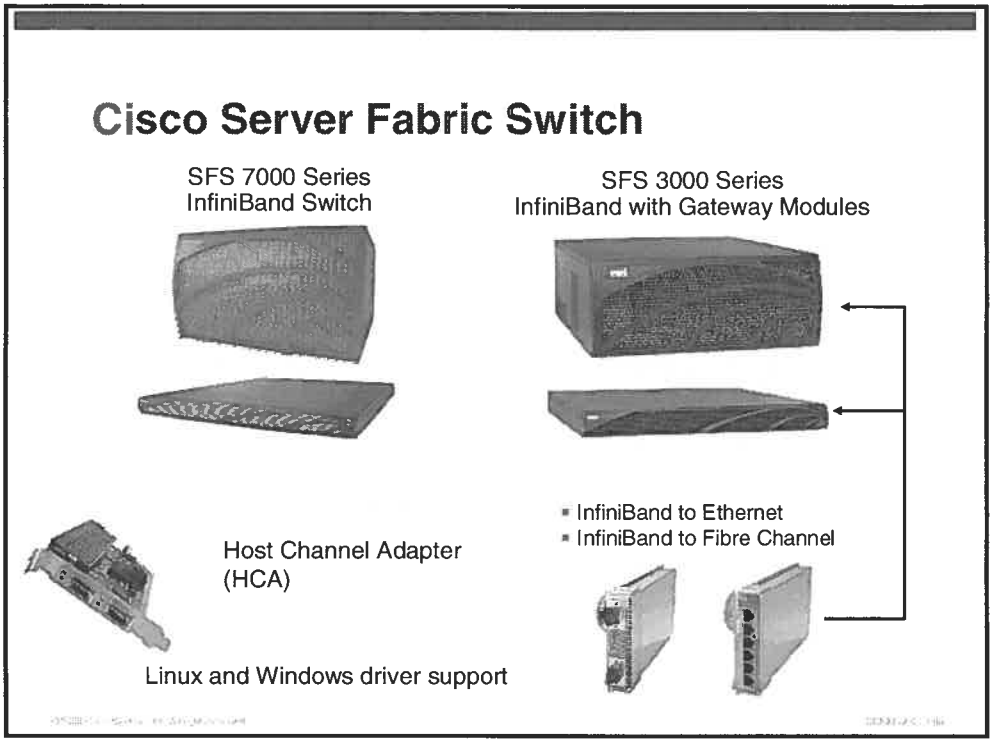
This figure positions the Cisco Virtual Blade switches in the data center networks and compares their price.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Server Fabric Switches

This topic describes the Cisco Server Fabric Switch (SFS) platform.

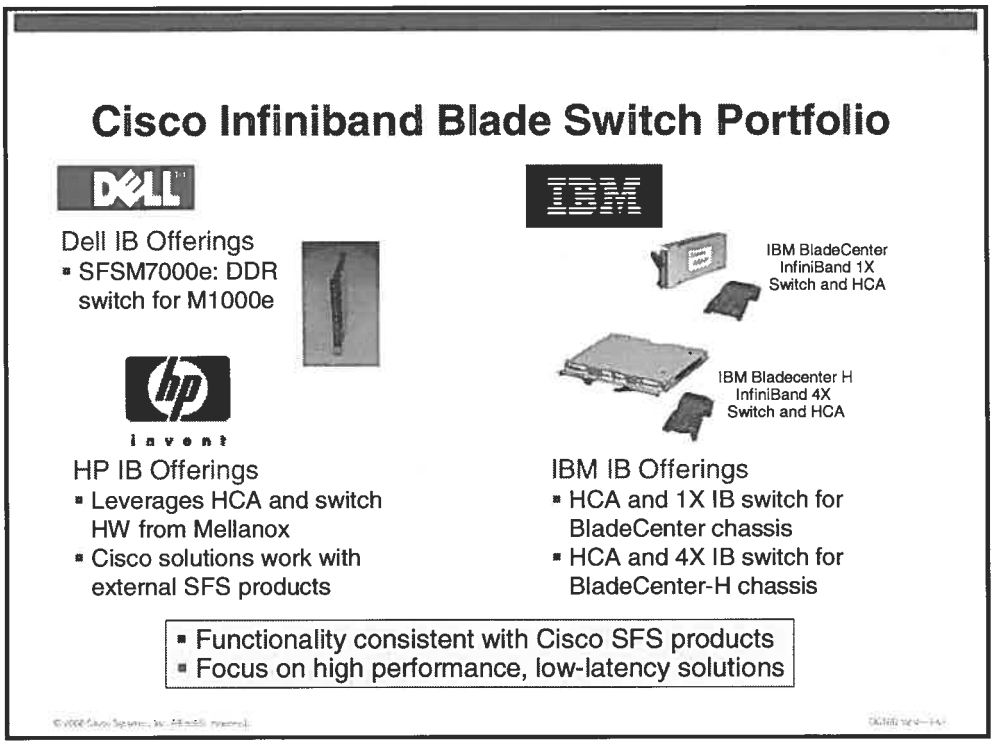


The Cisco SFS product line consists of the following products:

- **SFS 7000 Series InfiniBand Server Switches:** These are InfiniBand-only switches. The SFS 7000 is available with 24 ports of 10 Gb/s connectivity or 8 ports of 30 Gb/s connectivity in a single, compact 1U chassis. The SFS 7008 is a director-class 8-slot modular switch that supports ninety six 4X or thirty two 12X IB ports.
- **SFS 3000 Series Multifabric Server Switches:** These switches combine a high-performance InfiniBand server interconnect with Ethernet and Fibre Channel gateways, allowing an entire fabric of servers to share virtualized pools of I/O and storage resources connected through Ethernet or Fibre Channel switches such as Cisco Catalyst or Cisco MDS switches. The SFS 3001 provides up to twelve 4X InfiniBand ports with an expansion slot that supports either six GE ports or two 2 Gb/s Fibre Channel ports. The SFS 3012 is a director-class switch that provides 12 slots for InfiniBand, Ethernet, and Fibre Channel modules.

Cisco InfiniBand Host Channel Adapters (HCAs) bring high-performance 10 Gb/s InfiniBand connectivity to PCI-X and PCI-Express-based servers, with driver support for Linux, Windows, and Solaris x86 servers.

https://t.me/learningnets



The Cisco InfiniBand Host Channel Adapter Expansion Card for IBM and Dell Blade Centers (referred to as the HCA Expansion Card) provides InfiniBand I/O capability to processor blades in the Blade Center units. The host channel adapter adds InfiniBand ports to the CPU blade cards to create an InfiniBand-capable high-density cluster. HCA Expansion Cards communicate with one another through the Cisco InfiniBand Switch Module.

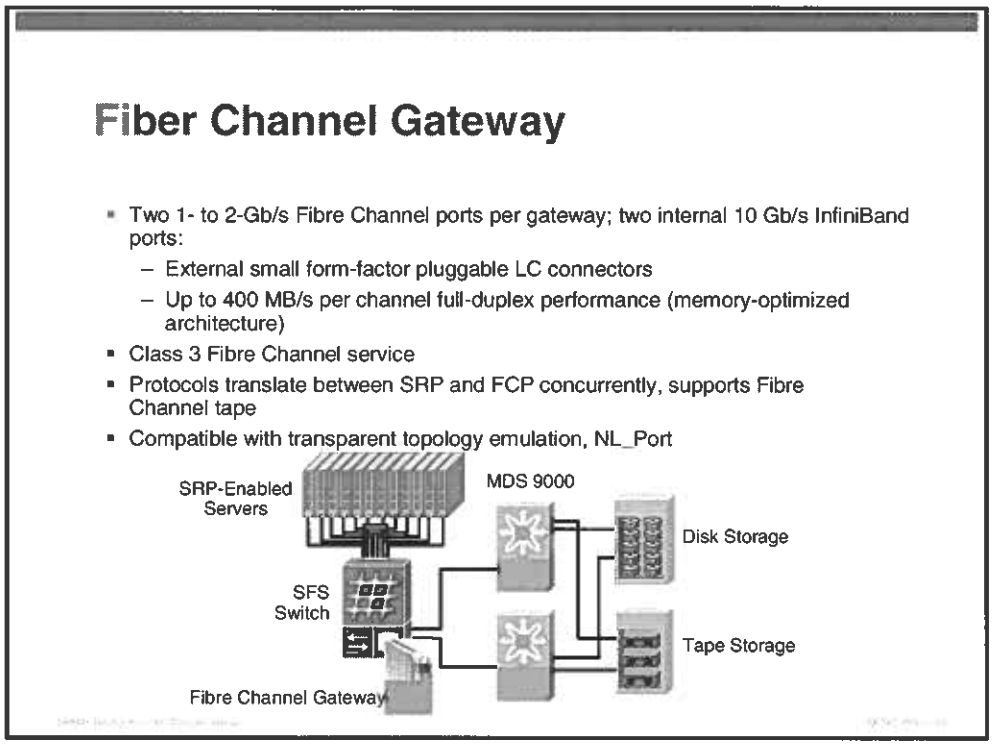
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco InfiniBand Gateways

This topic introduces the Cisco InfiniBand gateways.

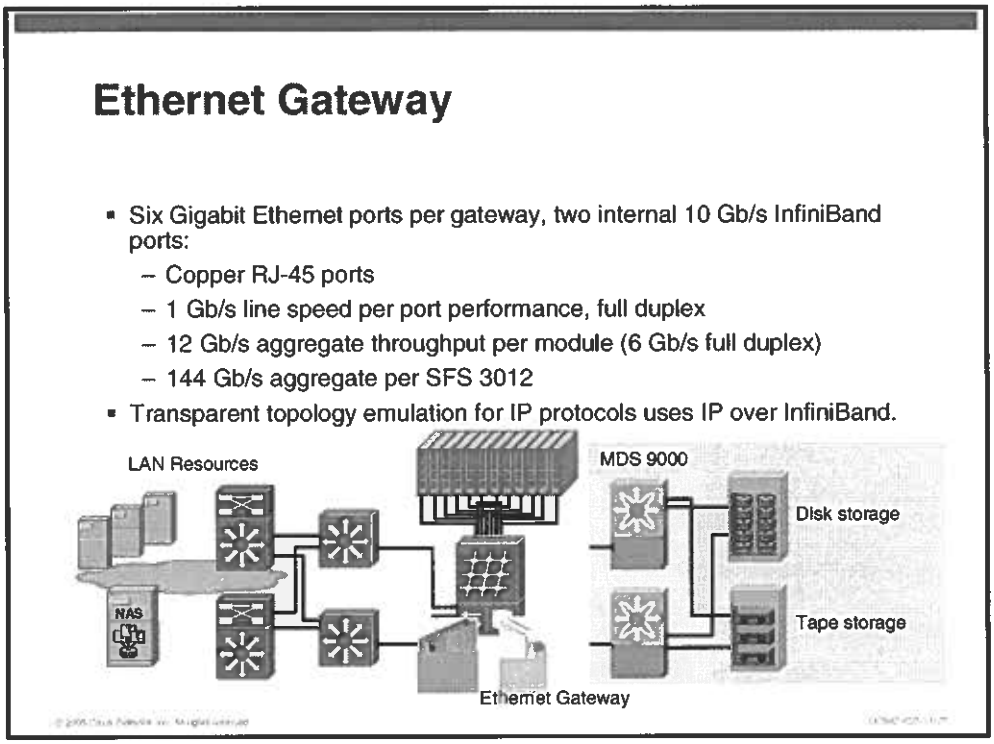
Fibre Channel Gateway



The Cisco SFS switch supports Fibre Channel gateway functions on 1-to 2-Gb/s Fibre Channel ports per gateway with two internal 10 Gb/s InfiniBand ports.

https://t.me/learningnets

Ethernet Gateway



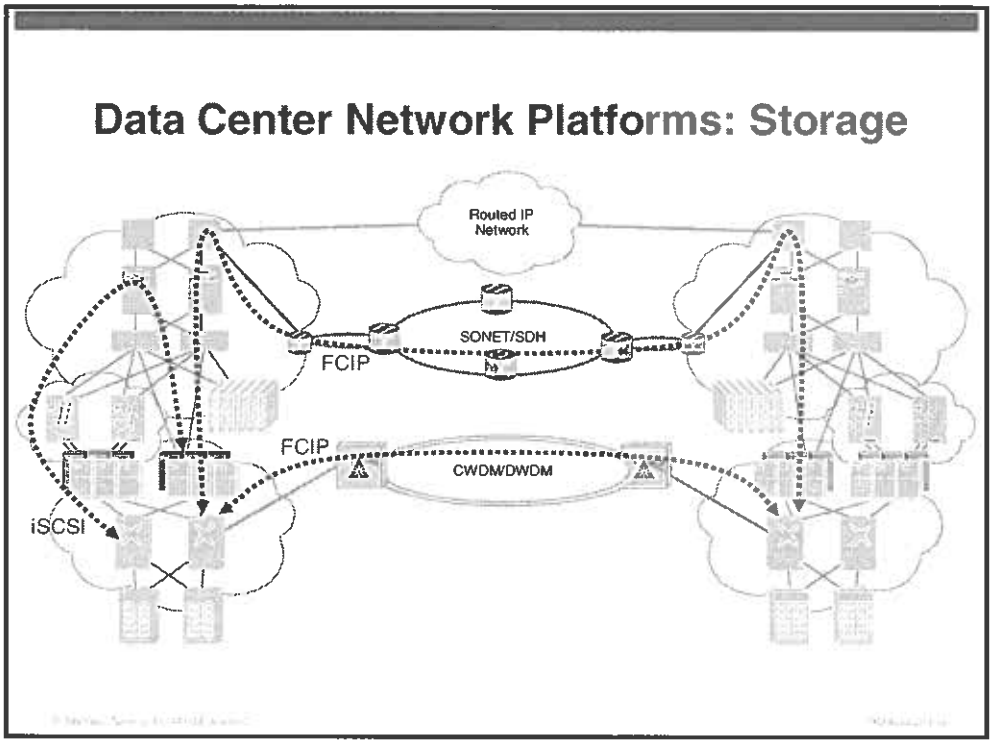
The Cisco SFS switch supports Ethernet gateway functions on 6 Gigabit Ethernet ports per gateway with two internal 10 Gb/s InfiniBand ports.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Storage Networking

This topic describes the Data Center Storage Networking.

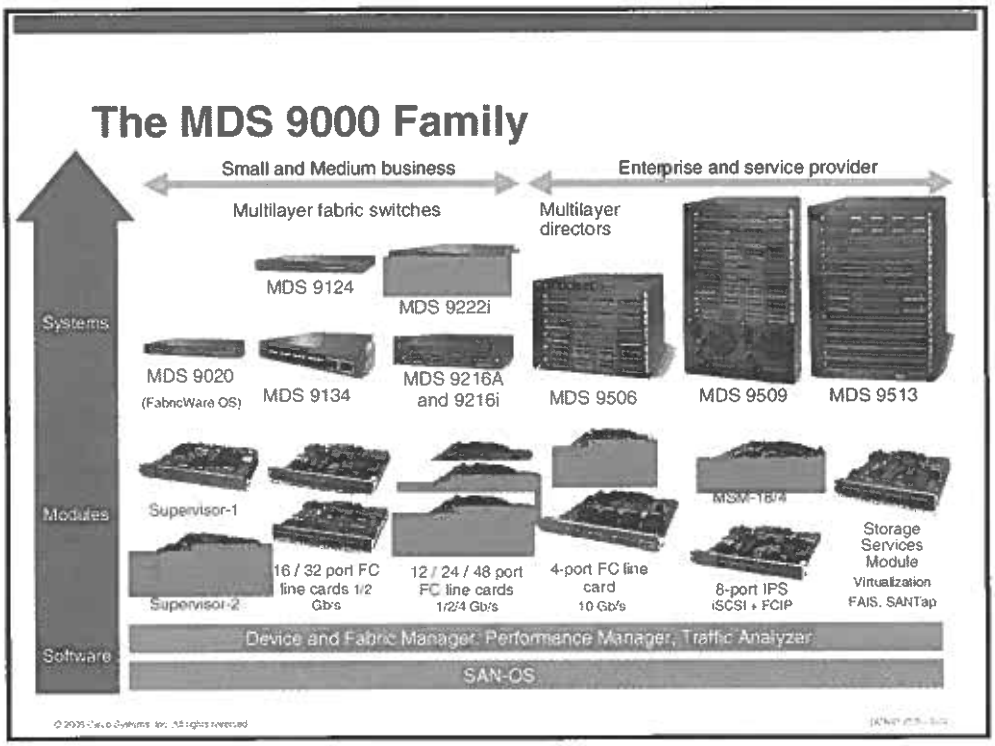


Until recently, the back-end SAN was a completely separate entity from the IP network, with different protocols, architecture, and management. Increasingly, however, the storage network and the IP network are converging. A modern data center storage network is designed along many of the same principles, such as the multitier model, as a modern IP network.

Of even greater relevance to the IP network engineer is the fact that storage traffic can now cross over into the IP network. The widespread adoption of Fibre Channel over IP (FCIP) as a means to transport storage data between sites, and the increasing adoption of Internet Small Computer Systems Interface (iSCSI) as an alternative to Fibre Channel for midrange servers, must be factored into the design of the IP network. FCIP and iSCSI have unique traffic patterns and requirements that must also be understood.

https://t.me/learningnets

The MDS 9000 Platform



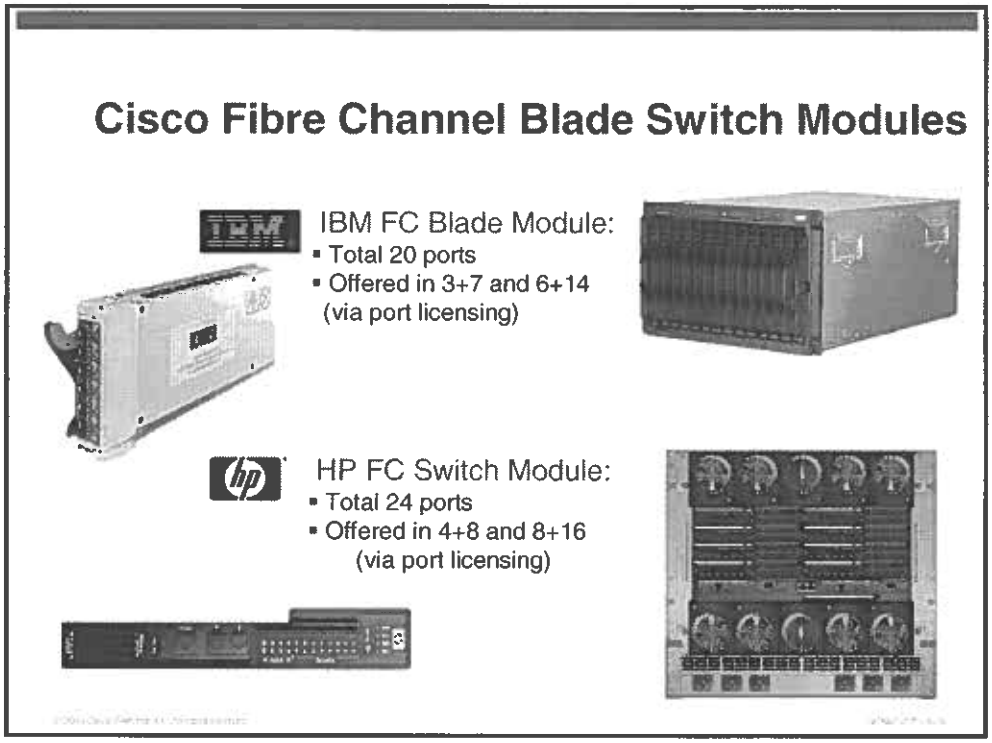
Like the Catalyst 6500 platform, the Cisco MDS 9000 is a modular platform that is designed to scale while preserving existing investments. Like the Catalyst 6500, the MDS 9000 consists of director-class, highly available chassis, supervisor modules, line card modules, and services modules. There are also fixed configuration switches for edge and workgroup applications.

The MDS 9000 is a multiprotocol switch. In addition to supporting the Fibre Channel interfaces that are the standard today for most SANs, the MDS 9000 supports storage-over-IP protocols, including FCIP and iSCSI. FCIP and iSCSI are supported on three MDS products:

- The Multiprotocol Services (MPS) module is a line card for MDS 9200 and 9500 chassis. The Multiservice Module (MSM) 18/4 is an enhanced version of MPS 14/2. The MPS module has 14 Fibre Channel interfaces and 2 GE interfaces. The MSM module has 18 Fibre Channel interfaces and 4 GE interfaces. The Intrusion Prevention System (IPS) module is also a line card for MDS 9200 and 9500 chassis. The IPS module has eight GE interfaces.
- The MDS 9216i contains a built-in supervisor and line card. The line card has the same features as the MPS module. The MDS 9216i also has an expansion slot that can be used for any other line card or services module.
- The MDS 9222i is new and enhanced version of MDS 9216i with an MSM 18/4 module built-in.

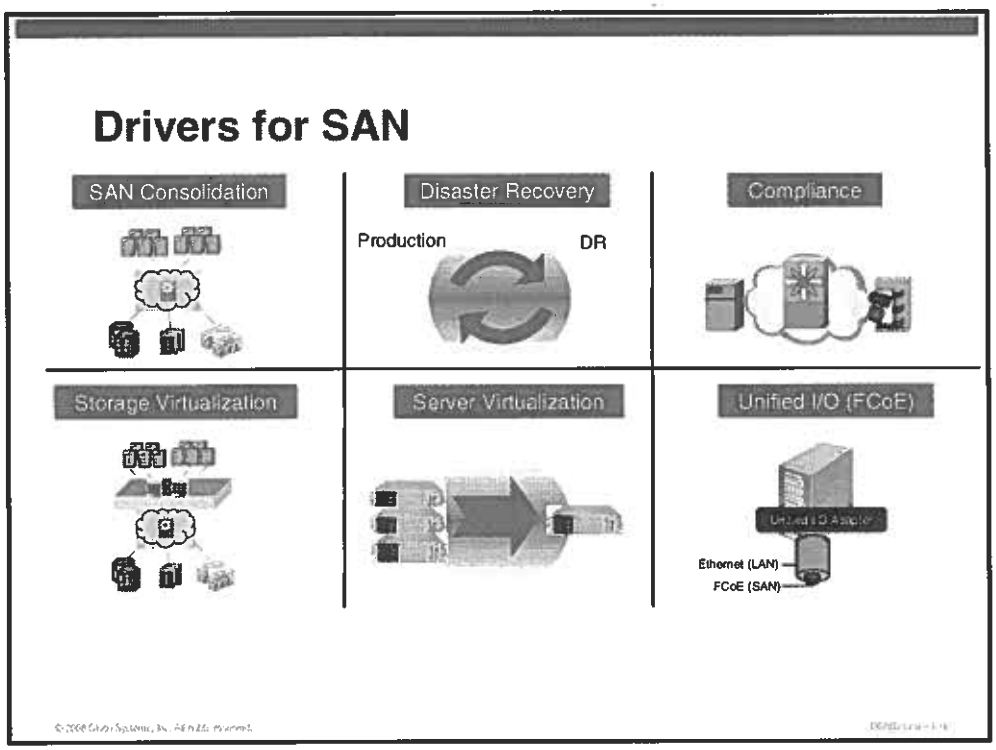
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco also offers Fibre Channel blade switches for the IBM Blade Center and the HP c-Class Blade System.

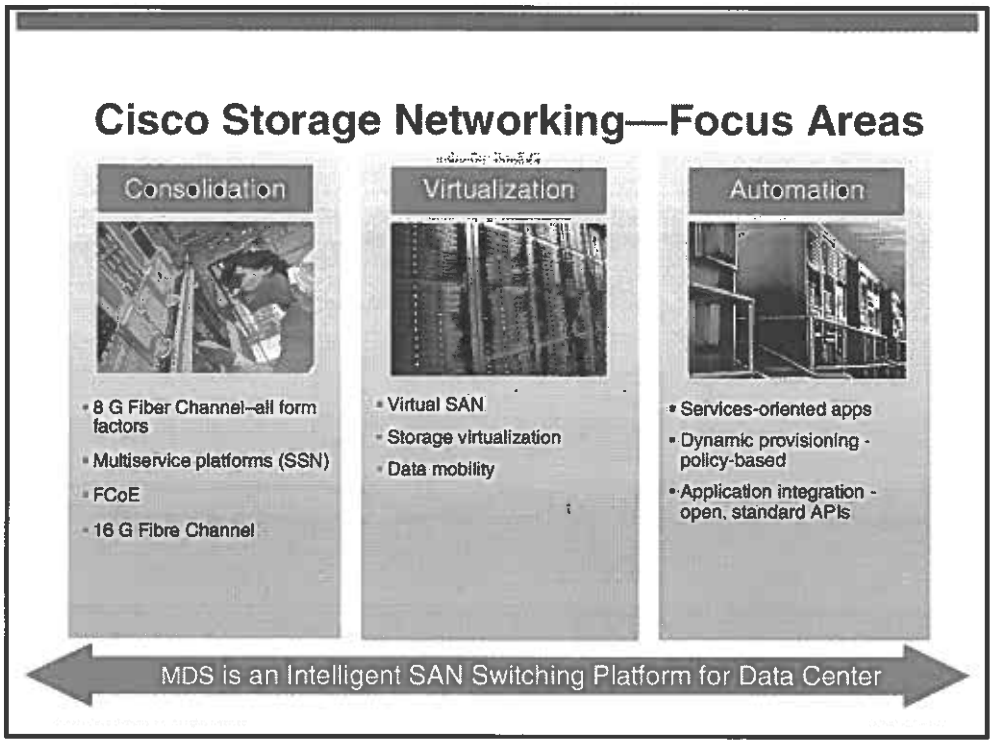
https://t.me/learningnets



SAN consolidation, storage virtualization and SAN extension solutions are effectively used for disaster recovery purposes. Fibre Channel protocol advantages and solutions will not go away, and as a result the SAN market will continue growing at about 6% compound annual growth rate (CAGR). The Cisco MDS 9124e Fabric Switches are used for HP c-Class Blade Systems, delivering a high performance 4 Gb/s embedded Fibre Channel switch that is especially suitable in a server virtualization environment. The integrated switch design frees up rack space, enables a reduction of components such as cables and SFPs, and utilizes shared power and cooling within the Blade System enclosure. This design results in space and cost savings as well as higher levels of availability through fewer points of failure. A hot-pluggable design supports non-disruptive software upgrades, and the unique ability to automatically and statefully restart failed processes. The integrated switch design further enhances the availability features of the overall solution. The Cisco MDS 9124e Fabric Switch supports VSAN, an industry standard for fabric virtualization that enables more efficient network use by creating hardware-based isolated environments within a single physical SAN fabric or switch. In addition, advanced traffic management features such as PortChannels, virtual output queuing, and Fabric Shortest Path First (FSPF) simplify deployment while optimizing the fabric. Management is simplified with a broad range of integrated tools that range from a consistent Cisco MDS 9000 Family Command Line Interface (CLI) to Cisco Fabric Manager to an industry-standard API that includes support for SNMP and SMI-s for integration with third-party and user-developed tools. Switch management is integrated with HP Blade System management tools including HP Onboard Administrator for Blade System, HP Systems Insight Manager, and HP Storage Essentials. These management tools help create a unique and tightly integrated experience for server and storage administrators.

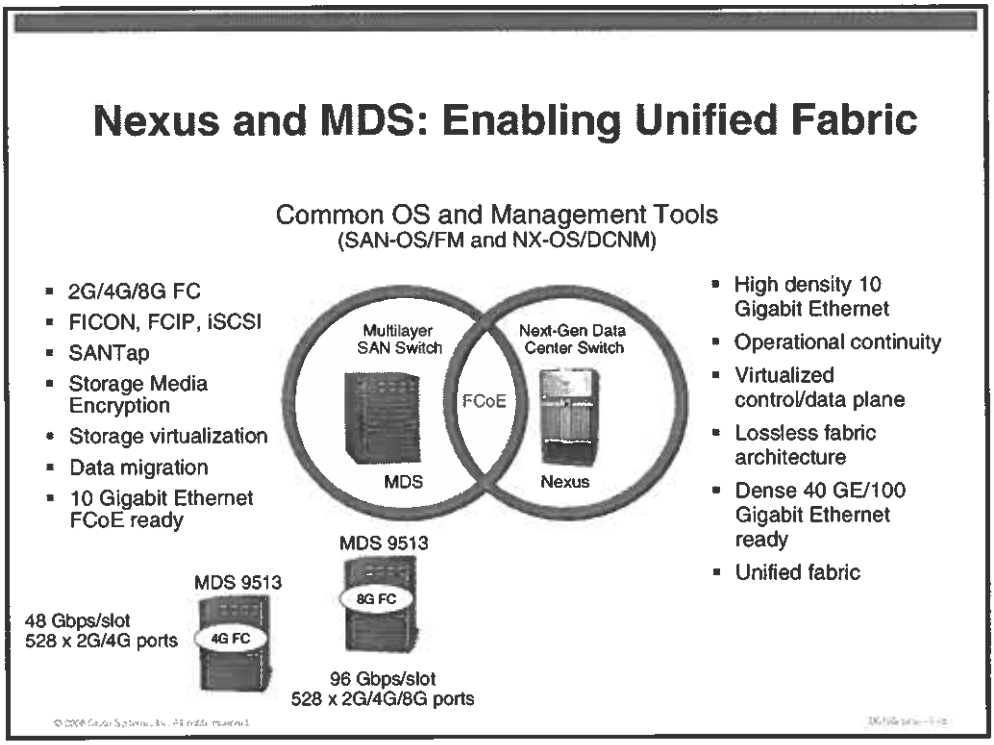
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



SAN consolidation enables any-to-any server and storage access that further provides the base for storage and server virtualization and also for the fabric automation. SAN fabric virtualization enables the fabric convergence. Fabric automation simplifies fabric management and reduces the fabric downtime. Fabric automation together with virtualization enables automated resource utilization and provisioning, which leads to self-managed data center SAN fabric. With virtualization processes such as hierarchical storage management and ILM, resource provisioning can be fully automated. The Cisco MDS platform will support 8 and 16 Gb/s Fibre Channel ports and FCoE in the future, as well.

https://t.me/learningnets



The Nexus 5000 or Nexus 7000 Series Switches and MDS Switches both support Fibre Channel protocol. In addition, Cisco MDS switching architecture is already 10 Gigabit Ethernet FCoE ready and represents lossless non-blocking and highly-available Fibre Channel architecture that has been deployed on Cisco Nexus 7000 Series Switches.

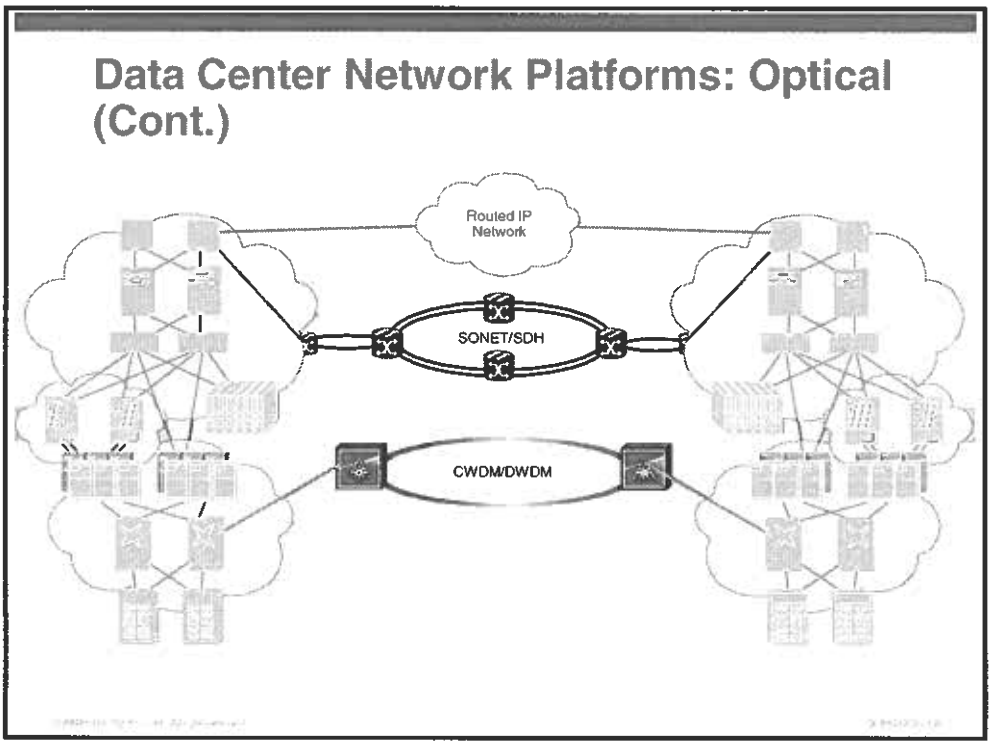
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Optical Transport

This topic describes the basic characteristics of the Cisco ONS 15454.

Data Center Network Platforms: Optical



Every enterprise-class data center has a disaster recovery plan. From a technical perspective, redundant sites are the key to effective data protection and business continuity. Applications like storage data replication often have unique requirements for connectivity between sites, leading to the adoption of high-bandwidth, low-latency, optical transports like Coarse Wave Division Multiplexing (CWDM) and Dense Wave Division Multiplexing (DWDM). In addition, the redundancy provided by the use of traditional time division multiplexing (TDM) networks like Synchronous Optical Network (SONET) and Synchronous Digital Hierarchy (SDH) are also factored into disaster recovery plans.

Optical transport engineering is a specialized field that is beyond the scope of this course. However, it is important to realize that many enterprises are currently dealing with multiple inter-site networks, including WDM, SONET, or SDH, and routed IP networks. The Cisco Optical Networking System (ONS) 15454 is the industry's first optical networking platform that enables the consolidation of all disparate WAN transports.

https://t.me/learningnets

Cisco ONS 15454 Multiservice Transport Platform (MSTP)

Carrier-class DWDM transport:

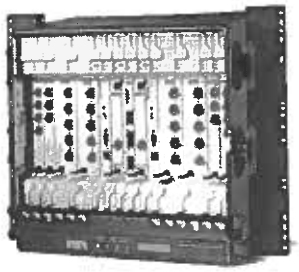
- TDM, Ethernet, and SAN integrated on one DWDM platform
- Integrated planning, install, and turn-up
- End-to-end management of Cisco IP DWDM networking

Reconfigurable networking:

- Reconfigurable Optical Add-Drop Mux (ROADM)
- Tunable 10 G lasers

Industry-leading advancements:

- 80 channel-capable solution
- Qualified for 40 G transport



15454 MSTP

The Cisco ONS 15454 is a carrier-class DWDM platform capable of using the same chassis for both DWDM and SONET or SDH. The MSTP provides a best-in-class integrated network design and planning software package, Metro Planner, which offers more than optical link budget calculations. Metro Planner provides everything needed to set up an optical network element for DWDM, including uploadable configuration files for each node, installation drawings, and bills-of-materials. After the configuration file is uploaded to a node, the Automatic Power Control (APC) function takes over and automatically keeps the network equalized. APC compensates for fiber degradations over time and equalizes channel powers at each node.

The MSTP Reconfigurable Optical Add/Drop Multiplexer (ROADM) is a popular feature that reduces initial costs and lowers ongoing operating expenses. ROADMs take the place of static OADMs, which might need to be upgraded to support increases in service demand. ROADMs can live in networks along with OADMs and all other optical components. The MSTP originally offered 32 channels in the C-band of the optical spectrum. This range might increase in a future release to 104 channels, utilizing the L-band in addition to the C-band.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

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

Summary

- The Cisco data center switching portfolio includes the Nexus 7000 series, Nexus 5000 series, Catalyst 6500 Series, the Catalyst 4900M, and Catalyst Blade Server Switches.
- The Nexus 7000 Series offers innovation in the data center: Virtualizing the control plane and data plane to allow for VDCs, operational separation, and traffic separation; lossless fabric to support the unified fabric requirements for a lossless transport.
- The Catalyst 6500 FWSM brings switching and firewalls into a single chassis, and supports a virtualization mode that allows a single FWSM to replace many firewall appliances.
- The Catalyst 6500 IDSM supports inline IPS and off-path IDS configurations integrated into the switch chassis, with a variety of traffic capture options.
- Cisco ACE, available as a Catalyst 6500 module or an appliance, provides application infrastructure control services, including Layer 2-7 server load-balancing.

Summary (Cont.)

- Blade servers are increasingly used in the data center, and can pose challenges for power, cooling, and cabling.
- Distributed cabling designs, including top-of-rack and end-of-row, improve the scalability and manageability of the data center network.
- The physical cabling design has implications for Layer 2 logical design and overall scalability and reliability.
- The Cisco Catalyst 4900M is a high-performance, rack-optimized switch that is designed for top-of-rack deployments.
- Cisco integrated blade switches present a third cabling design option.
- The Cisco SFS 3000 and 7000 are standards-based InfiniBand switches that provide an alternative to expensive, proprietary interconnects for high-performance computing fabrics.
- The Cisco ONS 15454 is the industry's first optical networking platform that enables the consolidation of multiple inter-site networks for business continuity.

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 3

Data Center Environmental Requirements

Overview

This lesson discusses concepts and basic characteristics of the components of the data center. With this knowledge, you can map these concepts and characteristics to the network design model so that the data center components can be accommodated within the data center network design.

Objectives

Upon completing this lesson, you will be able to discuss the business objectives for the data center architecture. This ability includes being able to meet these objectives:

- Discuss the environmental business objectives that are influencing data center design
- Discuss environmental characteristics such as power and cooling as they relate to scaling the data center
- Describe the data center thermal control model
- Locate devices in the data center hierarchical topology within the data center facility
- Describe the environmental benefits of deploying the data center blade servers and server racks
- Describe how resources are efficiently utilized in green data centers

https://t.me/learningnets

Environmental Requirements

This topic relates environmental characteristics such as power and cooling to scaling the data center.

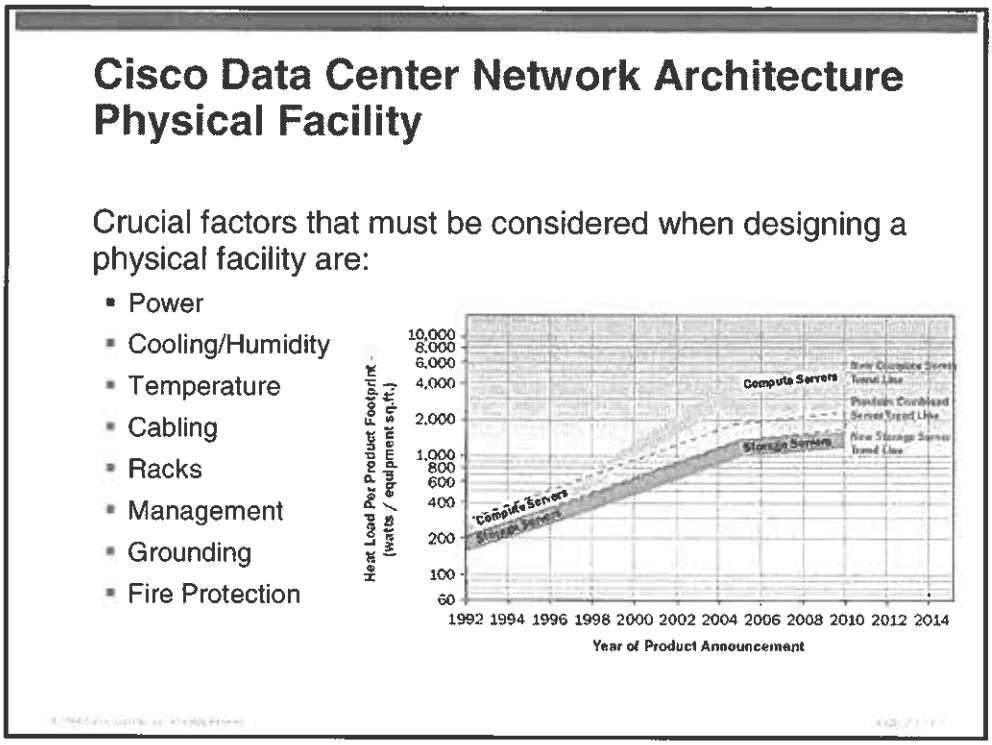
Consolidating the data center infrastructure requires careful consideration of environmental factors, such as power and cooling. This topic discusses these issues and how they relate to the overall data center design.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Data Center Network Architecture Physical Facility

A crucial part of the Cisco Enterprise Data Center Network Architecture is the physical facility—the power, cooling, physical housing, cabling, physical security, and fire protection—that allow Information Technology to function.



Power

Elements that make up the power infrastructure of the data center include the electrical service entrance of the building, the main distribution center, generator(s), uninterruptible power supply (UPS) systems and batteries, surge protection, transformers, distribution panels, and circuit breakers.

Power issues in data center facility include:

- Circuit breaker overload
  - Typically, staff does not account for fluctuations in power and might inadvertently create this issue.
- Heat variations
  - Power is dissipated as heat; more power = more heat. Local hot spots can occur.
- Loss of redundancy
  - Dual-homed server power supplies might cause a possible overload condition if more than half of the load is carried by the main power supply when a power failure occurs.

https://t.me/learningnets

Overheating is an equipment issue with high-density computing (blade servers):

- More heat overall
- Hot spots
- High heat and humidity threaten equipment life spans
- Computing power and memory requirements demand more power and generate more heat
- Data center demand for space-saving servers: density = heat; 3 kilowatts (kW) per chassis is not a problem for one chassis, but a 5 or 6 chassis rack = 20 kW
- Humidity levels affect static electricity and condensation; 40% to 55% relative humidity is recommended

Cooling

The cooling systems required to successfully remove heat from the data center include Computer Room Air Conditioners (CRACs) and their associated subsystems—chillers, cooling towers, condensers, ductwork, pump packages, piping—and any rack- or row-level cooling or air distribution devices. Sizing a precision cooling system requires an understanding of the amount of heat produced by the IT equipment, along with the heat produced by other heat sources typically encountered in a data center. Common measurements include BTUs per hour, tons per day, and watts. The mixed use of these different units causes a great deal of unnecessary confusion for users and specifiers. Fortunately, there is a worldwide trend among standards organizations to move toward a common cooling standard measurement—the watt.

Since the power transmitted by IT equipment through data lines is negligible, the power consumed from AC service mains is essentially all converted to heat. This fact allows for the straightforward representation of IT thermal output as watts, equal to its power consumption in watts. To further simplify, the total heat output of a system—therefore, the total cooling requirement—is the sum of the heat output of the components, which includes the IT equipment plus other items such as UPS, power distribution, air conditioning units, lighting, and people. Fortunately, the heat output rates of these items can be easily determined using simple and standardized rules.

The heat output of the UPS and power distribution systems consists of a fixed loss plus a loss proportional to operating power. Conveniently, these losses are sufficiently consistent across equipment brands and models to be approximated without significant error. Lighting and people can also be readily estimated using standard values. The only user-specific parameters needed to determine the total cooling load are a few readily available values, such as the floor area in square feet and the rated electrical system power. While air conditioning units create a significant amount of heat from fans and compressors, this heat is exhausted to the outside and does not create a thermal load inside the data center.

High and Low Data Center Humidity

High humidity can result in tape and surface deterioration, head crashes, condensation, corrosion, paper handling problems, and gold and silver migration leading to component and board failure. Low humidity greatly increases the possibility of static electric discharges. Such static discharges can corrupt data and damage hardware.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

High or Low Data Center Temperature

A high or low ambient temperature, or rapid temperature swings, can corrupt data processing and shut down an entire system. Temperature variations can alter the electrical and physical characteristics of electronic chips and other board components, causing faulty operation or failure. These problems may be transient or may last for days. Even transient problems can be very difficult to diagnose and repair.

Cabling

Cabling systems span a large group of different media. Data cables utilize different materials and connectors to optimize performance and flexibility, while the management of the systems maintains this optimization for the long haul. Power cables must also be addressed.

Racks and Physical Structure

There are many open rack and cabinet solutions that could be considered part of the physical structure of a data center. The most critical elements are the rack structures housing IT equipment, physical room elements such as dropped ceilings and raised floors, and pathways to manage cabling considerations.

Management

Management spans all facilities elements. To have reliable facilities, it is important to have visibility of all of the components of the physical infrastructure. Management includes systems, such as building management systems (BMS), network management systems (NMS), element managers, and other monitoring hardware and software.

Grounding

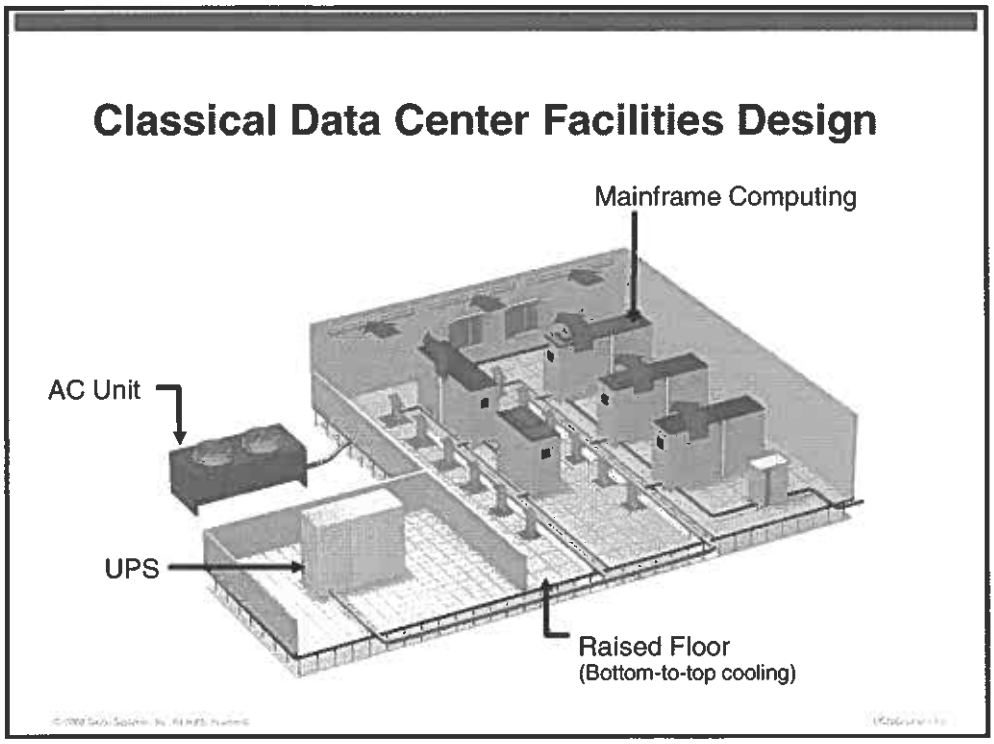
Grounding considerations cover the common bonding network and grounding gear that will protect your data center investment from electrostatic discharge.

Security and Fire Protection

Subsystems included here are physical security devices at the room and rack level and fire detection/suppression systems.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



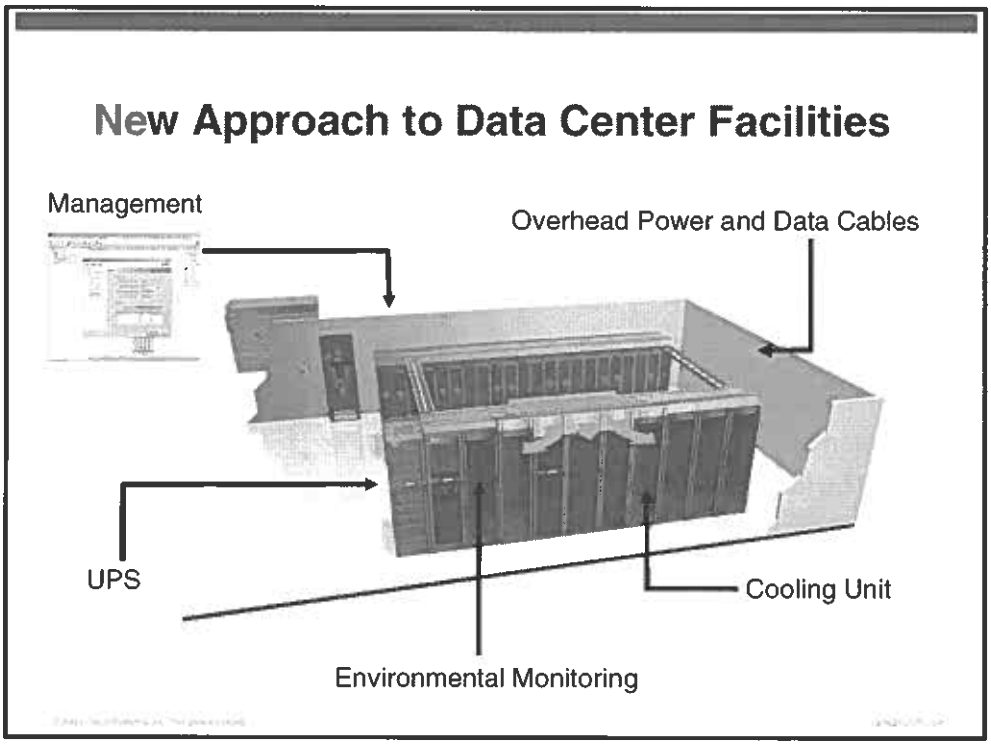
Traditionally, the data center evolved primarily to support mainframe computing. A feature of this environment is that change is the exception, rather than the rule, and this led to a Layer 0 infrastructure in which:

- Power and cooling were over-provisioned from onset.
- Floor was raised for bottom to top cooling.
- Racks and cabling did not need to be reconfigured.
- Floor loading was predetermined.
- Raised floors were the norm.

Today's server-based environment is one where change is a constant, yet the environmental infrastructure support is often relatively inflexible.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



In the figure, the UPS and air-conditioning are no longer segregated from the computer equipment. There is no longer a raised floor, the power and data cabling is being run over the tops of the racks, and the power distribution panels are now within the racks.

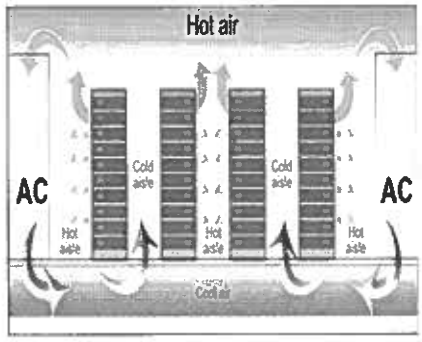
Incorporating the physical layer into an on-demand framework achieves significant cost savings, and the framework's modular nature results in much higher levels of availability versus investment. Additional benefits include:

- Significantly lower TCO
- Ability to add power and cooling as needed
- Shorter deployment cycles
- Ability to provision facilities the same as IT infrastructure
- Increased uptime

https://t.me/learningnets

Data Center Thermal Control Model

- Cold air is pumped from the AC units through the floor into the cold aisles between server racks
- AC air is pulled from the cold aisle through the racks and exits the back of the servers
- The heat from the server racks exhausts into the hot aisles where it is returned to the AC units to be chilled



A traditional data center thermal control model is based on controlled air flows. Before trying to solve the heat problem, it is important to understand what is happening in your data center environment that may be adding heat. To help dissipate heat from electrically active IT hardware, Air Conditioning (AC) treated air is pumped into the floor space below the rows of server racks. The chilled air enters the data center through vented tiles in the floor and fills the cold aisle in front of the rows of racks. For non-raised floor data centers, the chilled air enters through diffusers above the cold aisle. Although construction of data centers differs, the heat problem they face is similar, and the technology described here applies to both. The built-in fans of the stored hardware pull the cool air through each rack, chilling the warm components (processors, power supplies, etc.) of the hardware. Heat is exchanged and the air becomes warm. The heated air then exits out the back of the rack into the hot aisle. The heated air is pulled back into the AC units and chilled. The cycle repeats, and the data center environment should be kept at a safe, cool temperature, as long as the BTU capacity of the AC units is sufficient to cool all the equipment installed.

Summarization of the air flow in a data center facility:

- Cold air is pumped from the AC units through the raised floor of the data center and into the cold aisles between facing server racks.
- Air-conditioned air is pulled from the cold aisle through the racks and exits the back of the servers.
- The heat from the server racks exhausts into the hot aisles where it is returned to the AC units to be chilled.

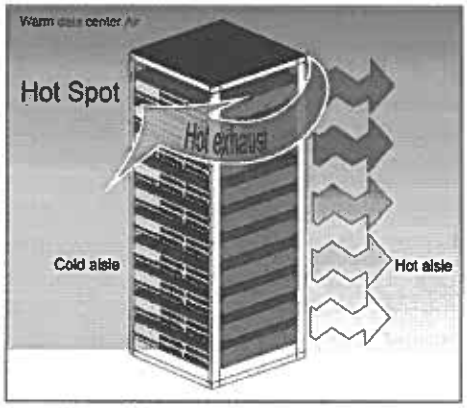
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Power and Hot Spots

High-density equipment produces a lot of heat:

- Temperature and humidity need to be controlled
- Direct cooling:
  - Cooling applied to DC zones, rack or individual devices
- Distribute the heat
- People generate heat—500 BTUs per person



Equipment Layout Common Cold and Hot Aisles with Hot Spot

The consequences of high-density equipment include:

- Hot spots
- Unpredictable cooling performance
- Loss of redundancy
- Loss of static pressure
- Excessive dehumidification

Many customers simply do not know how to plan for this kind of equipment. Consider some of the requirements:

- High-density equipment produces a lot of heat. An average rack uses 1.5 kW of power and produces 1.5 kW of heat. A populated HD rack can produce 15 kW.
- Review the recommended environment specifications. For example, the recommended inlet temperature might be 68°F to 77°F (20°C - 25°C), while the allowable range is 59°F to 90°F (15°C to 32°C). The recommended humidity range might be 40% to 50%, while the allowable range is 20% to 80%.
- Evaluate equipment air flow possibilities: front to rear, front to top, front to top and rear.

Some rooms have false ceilings; these can be incorporated into the system design. The ceiling void can be utilized as a warm air return plenum. This best practice should be combined with the best practice for rack layout. According to best practice, the rack layout is now fixed and a return air duct can be used in the hot aisle if no drop ceiling is present.

The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) is an international membership-based organization that offers guidelines (www.ashrae.org).

https://t.me/learningnets

Helpful Conversions

```
One watt = 3.413 BTUs.
```

This is a generally used value for converting electrical values to BTUs and vice versa. Many manufacturers publish KW, KVA, and BTU in their equipment specifications. Sometimes dividing the BTU value by 3.413 does not equal the published wattage. Where the information is provided by the manufacturer, use it. Where the information is not provided, the above formula can be helpful.

Hot Spots

Inadequate AC may allow hot spots to develop in the data center. Hot air migrates to the cold aisle and collects around the highest racked servers, fully populated racks and blade servers. Unless cooled, these hot spots may cause components to overheat and fail.

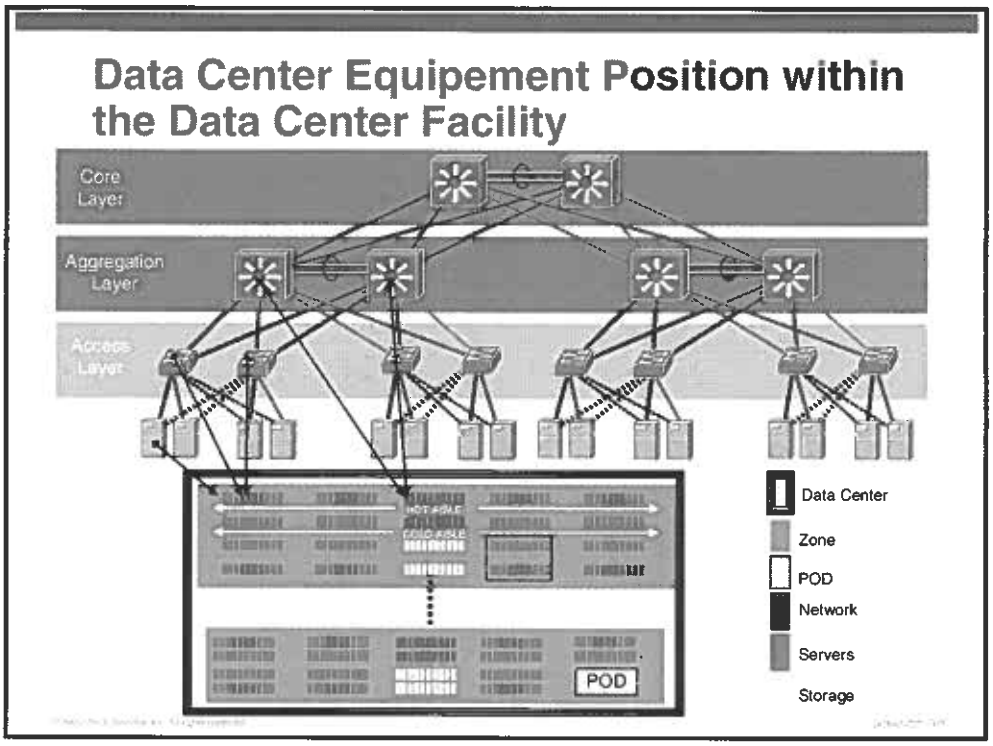
In an attempt to avoid damaging hot spots, two steps are usually tried:

- IT facility managers try to depopulate racks in an attempt to lower the heat load.
- The space between racks is increased so that adequate chilled air flow can be supplied to the racks.

The higher the server is located in the cabinet, the more difficult the cooling problem, as cool air is typically supplied from the floor of the data center. For example, imagine a server rack that has 600cfm of cool air available and is populated with 10 servers, each of which draws 100 cfm at its air intake. The lowest server draws 100cfm of the rising air, leaving 500cfm to cool the remaining nine servers. The next server up in the enclosure draws another 100cfm, leaving 400cfm to cool the remaining eight servers, and so on. Obviously, in this example, there will be no cool air left for the seventh through tenth servers at the top of the rack. Typically, the remaining servers simply draw air from whatever source possible, often times the hot exhaust from and around the ceiling.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



A data center facility with server racks, blade servers and switches can be organized as shown in the figure. The network equipment is located between the parallel hot and cold aisles. Providing the floor space and AC capacity to help meet the demands of a growing infrastructure is a major environmental data center challenge.

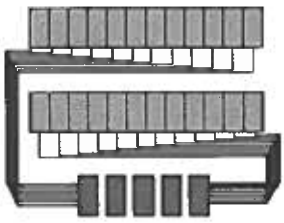
Cold aisles may be too narrow; air flow below the raised floor may become restricted by cables and piping. Smaller, more powerful and hotter processors may need to work faster and the AC units are often underpowered for the increased heat load.

Heat buildup may threaten the data center when it is growing or if we are trying to incorporate more processing power in the same footprint year after year.

https://t.me/learningnets

Cabling Design: Direct-Connect

- One main networking row
- Cabling routed directly to server cabinet locations
- Excellent for the logical element of a network





- Bad for the physical element of a network
- Scales poorly
- Prone to cable overlap

Smaller networks often use a direct-connect cabling design in which there is one main networking row. Cabling is routed directly from the networking row to server cabinet locations.

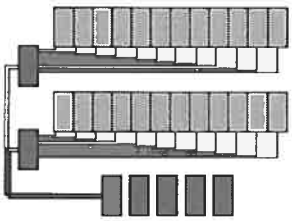
A direct-connect design is excellent for the logical element of a network, in terms of Layer 2 configuration, but a direct-connect design is bad for the physical element of a network. A direct-connect design scales poorly and is prone to cable overlap.

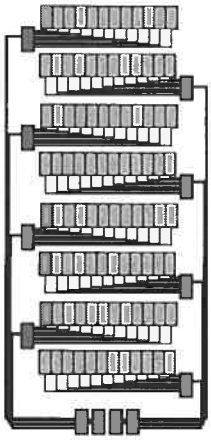
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Distributed Cabling Design

- Network substations
- Cabling to all server cabinets
- Subset of cables to the network row

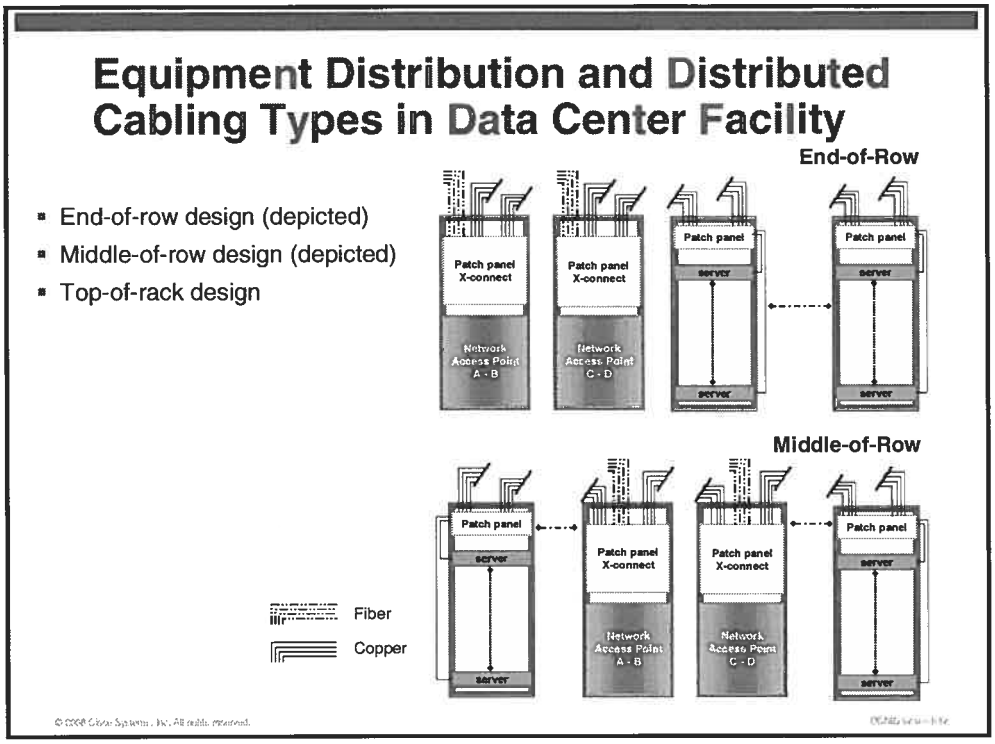




- Good for the physical element of a network
- Scales well
- No cable overlap
- Cable runs are shorter and better organized:
  - Easier to manage
  - Less expensive
  - Less restrictive for air flow

A distributed cabling design uses network substations within each server row. Server cabinets are cabled to the local substation, and from there to the main network row. This design is superior for the physical element of a network. This design scales well, and makes it possible to avoid cable overlap. Cable runs are shorter and better organized, which makes the physical network easier to manage, less expensive, and less restrictive for air flow.

https://t.me/learningnets



There are different types of network equipment distribution of the data center racks. The most typically used distribution type is end-of-row distribution, with servers and network access layer switches connected using copper cables. The middle-of-row and end-of-row design distributions are typically used for modular access, and cabling is done at data center build-out. These two types of rack distributions enable lower cabling distances (and lower cost) and allow dense access layer. On average these are 6-12 multi-RU servers per rack requiring 4-6 kW per server rack and 10 kW-20 kW per network rack. Every switch is supporting one or many medium or large subnets and VLANs.

End-of-row design characteristics

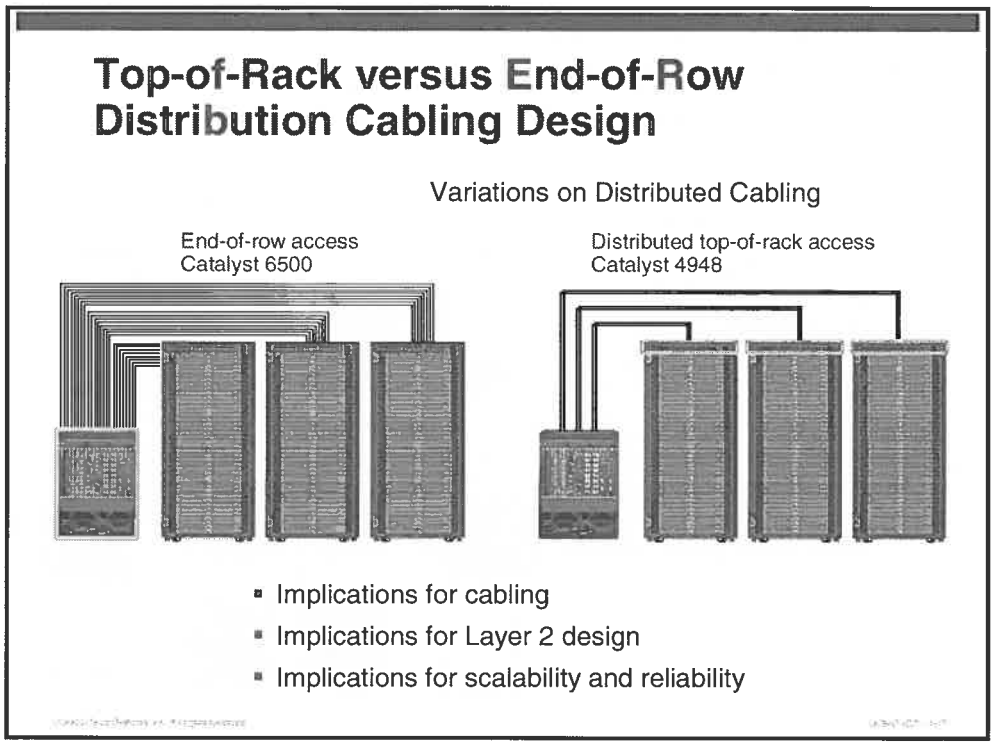
- Copper from server to access switches
- Poses challenges on highly dense server farms
  - Distance from farthest rack to access point
  - Row length may not lend itself well to switch port density

Middle-of-row design characteristics

- Use is starting to increase, given EoR challenges
- Copper cabling from servers to access switches
- Fiber cabling may be used to aggregate ToR servers
- Addresses aggregation requirements for ToR access environments

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



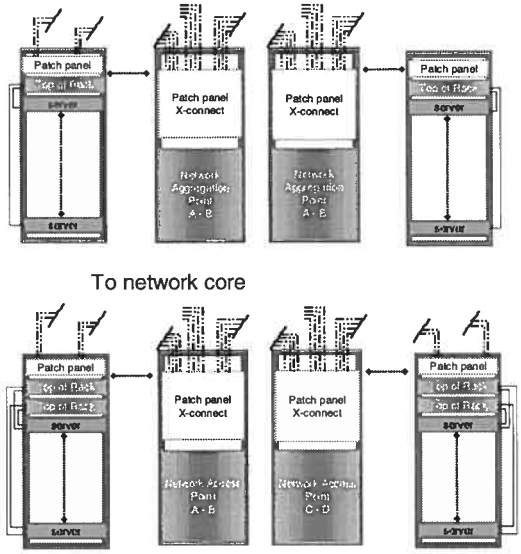
There is no single answer to the cabling question. Many factors need to be considered, including exiting cable plants, cabling maintenance costs, and logical network design. Both top-of-rack and end-of-rack designs are viable; the choice of which design to use is often dependent on the specific characteristics of each data center environment.

https://t.me/learningnets

Equipment Distribution and Distributed Cabling Types in Data Center Facility—ToR

ToR design model:

- Used in conjunction with dense access racks (1U servers)
- Typically one access switch per rack
- Typically ~10-30 servers per rack
- One or many small subnets per access switch



Top-of-rack (ToR) design model is appropriate for many 1RU servers, making it a lot easier for cabling compared to end-of-rack (EoR) design because the access switch is located on the top of every server rack and cabling happens within the rack. Outside the server access rack, the copper or fiber uplinks are used to the aggregation layer middle-of-row (MoR) switching rack. Every access layer switch is configured with one or more small subnets.

Top-of-rack design characteristics:

- Used in conjunction with dense access racks (1U servers)
- Typically one access switch per rack
  - Some customers are considering two + cluster
- Typically:
  - ~10-15 servers per rack (enterprises)
  - ~15-30 servers per rack (SP)
- Use of either side of rack is becoming popular
- Cabling:
  - Within rack:
    - Copper for server to access switch
  - Outside rack (uplink):
    - Copper (Gigabit Ethernet): Needs a MoR model for fiber aggregation
    - Fiber (Gigabit Ethernet or 10 Gigabit Ethernet): Is more flexible and also requires aggregation model (MoR)
- Subnets and VLANs:
  - One or many subnets per access switch

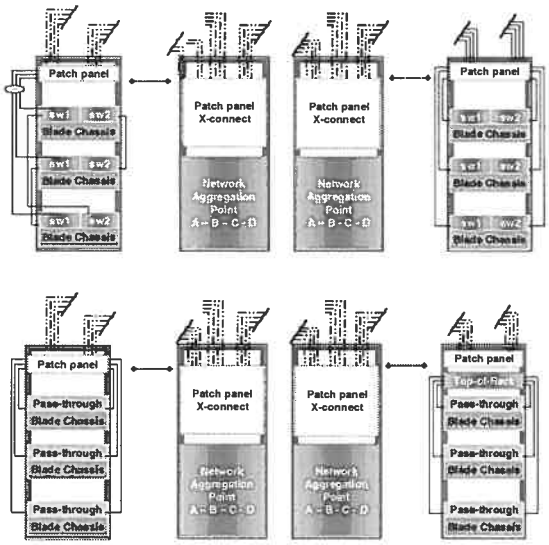
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Equipment Distribution and Cabling Types in Data Center Facility—Blade Chassis

Blade chassis rack design models:

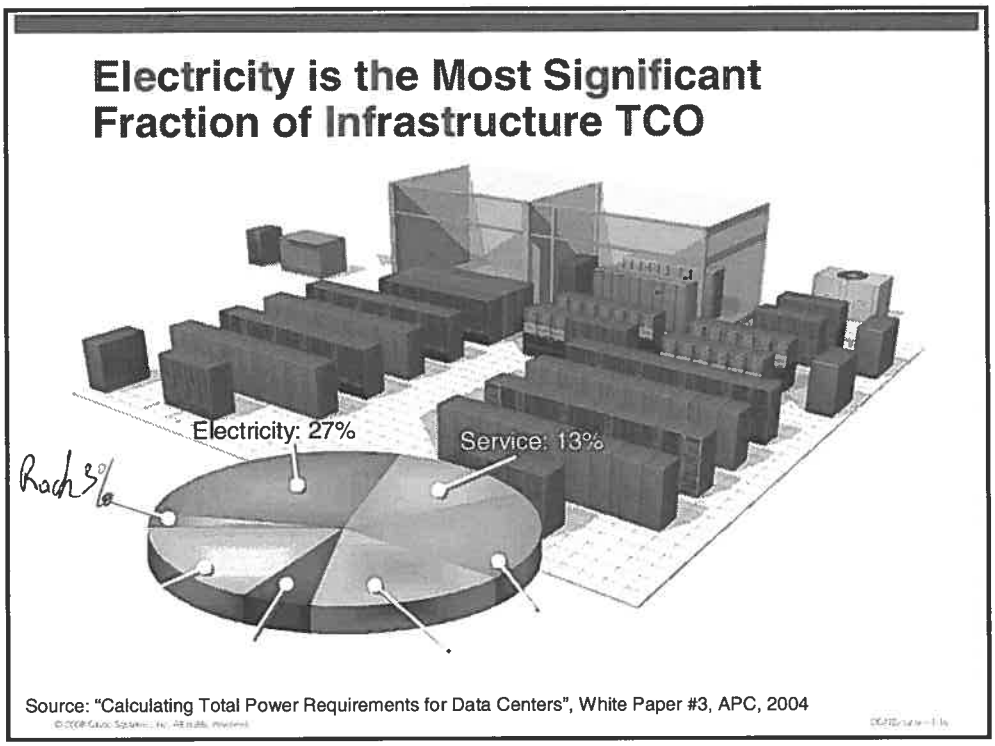
- End-of-row (switch to switch)
- Middle-of-row (pass-through)
- ToR blade server design



For blade chassis the following three design models can be used:

- End-of-row design (switch to switch) characteristics:
  - Scales well for blade server racks (~3 blade chassis per rack)
  - Most current uplinks are copper but the new switches will offer fiber
- Middle-of-row (pass-through) design characteristics:
  - Scales well for pass-through blade racks
  - Copper from servers to access switches
- ToR blade server design characteristics:
  - Have not seen it used in conjunction with blade switches
  - May be a viable option on pass-through environments if the access port count is right

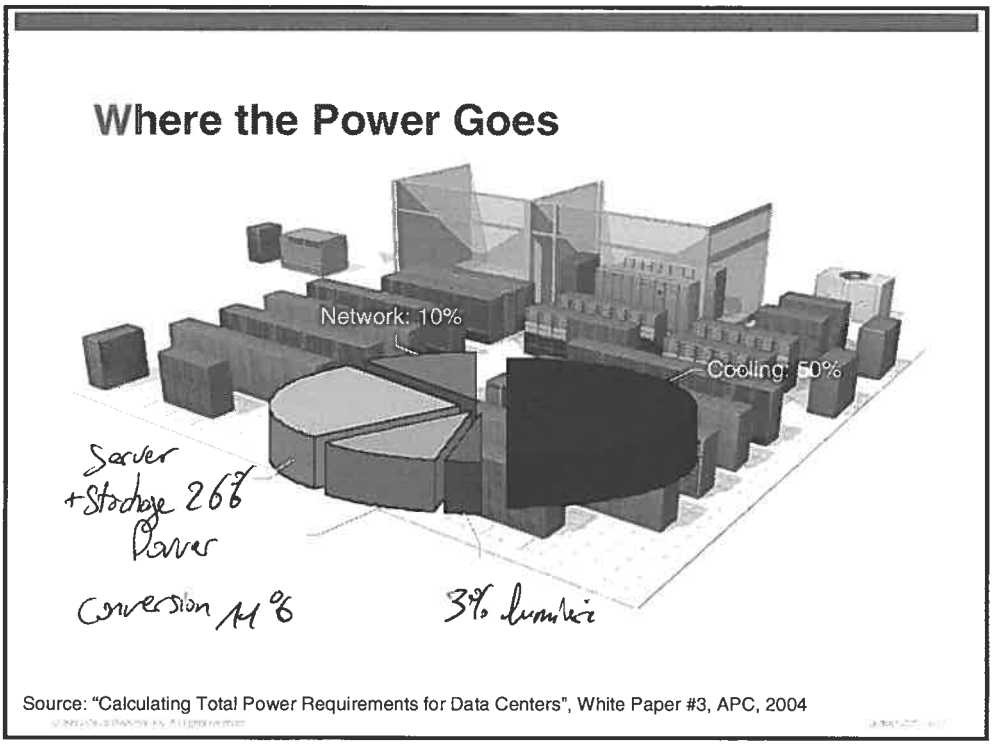
https://t.me/learningnets



The figure shows that electricity accounts for a significant fraction of the Total Cost of Ownership (TCO) of the data center infrastructure.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The figure shows a typical breakdown of power consumption in a data center. The network portion consumed can be from 10% to 20%, depending on the specific data center architecture.

https://t.me/learningnets

Calculating Efficiency and Operating Costs

Typical problem: What is the operating cost of a 1 MW UPS system that has a 90% efficiency at $0.10 per kWh, for one year at 50% load level?

Operating cost = kW x (1/efficiency at load level)) x Hours in a Year x Cost of Electricity x Load Level

= 1000 kW x (1 / 0.90) x 8760 x $0.10 x Load Level

= 1000 kW x (1.111) x 8760 x $0.10 x 0.50

= $486,618



A starting point for understanding the impact that a network design can have on power requirements is to understand the current efficiency and operating costs associated with running devices within the data center.

Helpful Conversions

```
One watt = 3.413 BTUs
```

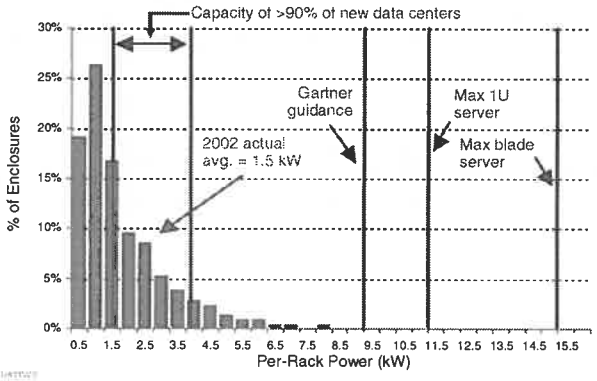
This is a generally used value for converting electrical values to BTUs and vice versa. Many manufacturers publish kW, kVA, and BTU in their equipment specifications. Sometimes, however, dividing the BTU value by 3.413 does not equal the published wattage. Where the information is provided by the manufacturer, use it. Where the information is not provided, the above formula can be helpful.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Powering the High-Density Data Center

- 90% of data centers can only accommodate 4 kW or less per rack:
  - High-density racks need 8-15 kW ...or more
- Server power consumption fluctuates:
  - Multiprocessor and diskless servers have the most dynamic power variation

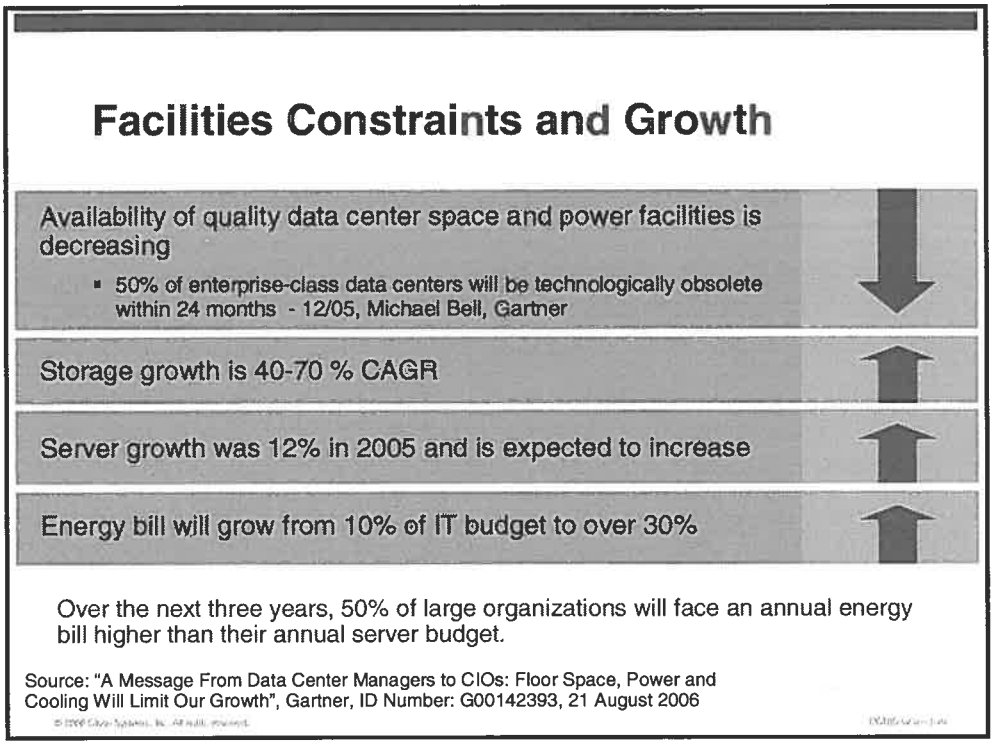


High-density rack power requirements have far exceeded existing power capacity. Data centers designed only a few years ago need many times the power they currently have to implement blade servers and other high-density technologies. The bar graph shows how power per rack is not keeping pace with the deployment of new and denser equipment.

When determining power requirements for equipment in a data center, one variable that is often hard to predict is the level of usage. In a server environment, the harder the computing power of the server must work, the greater the power draw from the AC supply and the heat output that needs to be dissipated.

The figure shows how levels vary with usage. Disk systems are less variable because the disks continue to spin regardless of whether they are being accessed or not.

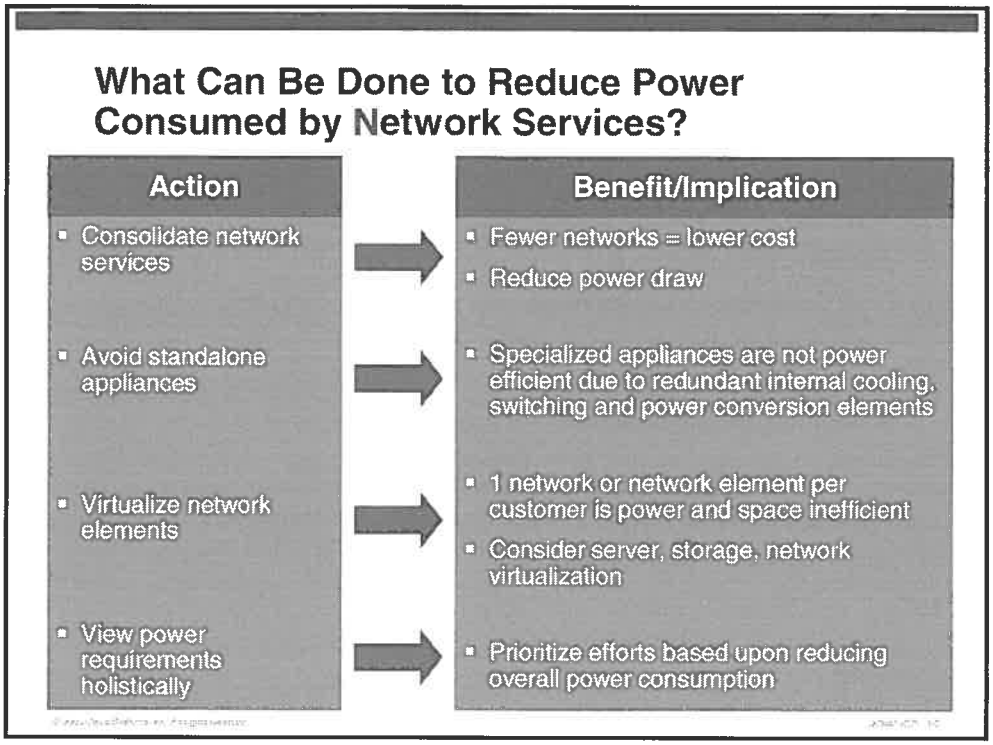
https://t.me/learningnets



Google is an example of one company where power costs already exceed server budget. One of the key recommendations from Gartner is to make better use of existing systems.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The data center design can dramatically affect how the power is consumed by network services. It is possible to reduce the power being consumed by:

- Consolidating networks and services
- Avoiding standalone appliances by integrating and consolidating network services
- Virtualizing network elements
- Viewing power requirements holistically and focusing on power consumption as an integral part of the network design

https://t.me/learningnets

Rack Weight and Floor Loading

- Locate racks where floor can support the weight
- Raised floor is used for:
  - Cabling and connectivity
  - Power to run the racks
  - Cooling
- Weight of a full blade rack = 970 kg in 60x60 cm
- Most floors are designed to 500 kg per sq. meter

Trends in data center technology are dramatically changing the requirements of the physical resources in the data center:

- High-density blade servers and increasingly dense storage arrays are greatly increasing the weight-to-floor area ratios, pushing the specifications beyond design capacities.
- As denser computing and storage power is deployed, the heat output from equipment and the power requirements per rack are increasing. This means that power and cooling requirements are also radically increasing.
- Placement of current and future equipment must be very carefully considered so that data center physical infrastructure and support is deployed optimally.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

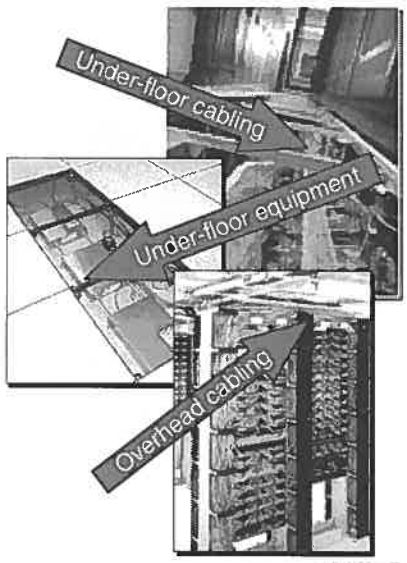
Cable Management

Challenge:

- Airflow
- Cable management

Solution:

- Integrated cable management system
- Above the rack instead of under the floor
- Cables located in the front or rear of the rack for easy access



Cable management is a major issue in any data center. Cable types (fiber optic types, copper types), lengths (minimum and maximum), runs, locations, bend radii, and so on, are all factors that need to be taken into account. How cable runs are actually implemented in a data center can affect other aspects of data center operation (no one wants a data center where the cables run across the floor, creating a health and safety hazard). Achieving appropriate cable runs raises some issues.

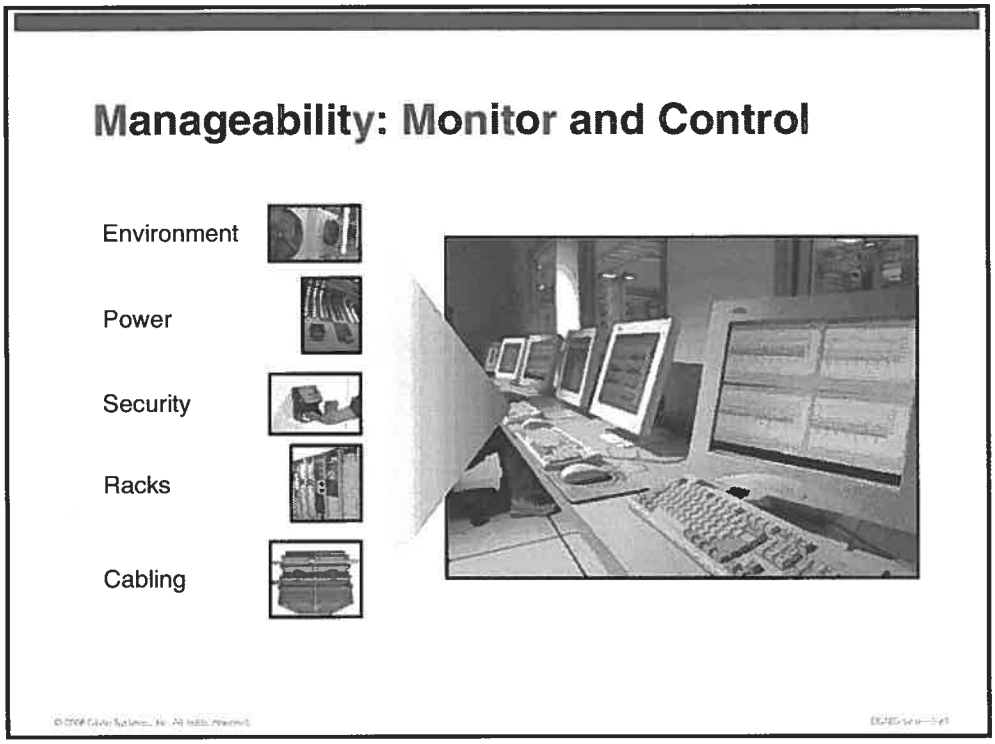
For example, with under-floor cabling, airflow is restricted by power and data cables. Raised flooring is a difficult environment in which to manage cables because cable changes mean lifting floor panels and potentially having to move equipment racks.

The solution is a cable management system consisting of integrated channels located above the rack for connectivity. Cables should be located in the front or rear of the rack for easy access. Typically, cabling is located in the front of the rack in service provider environments.

Cable management is a major topic in its own right. Work with the facilities team that installs and maintains the data center cabling to understand the implications of any new or reconfigured environment in the data center.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Techniques and technologies in Layer 0 are evolving and include such items as:

- Hot aisle containment
- Modular power
- Overhead cable management
- Environmental monitoring and intelligent power distribution
- Centralized management over IP network

Manageability

The challenge for the future is to be able to monitor and control the Layer 0 issues in the data center. Forms of centralized management of IT resources via a GUI interface are evolving to facilitate access to environmental and power variables as well as the security status of the rack.

Environment

Environmental considerations include:

- Temperature fluctuations and data center hot spots
- Relative humidity
- Smoke detection

Power

Power considerations include:

- Power to each rack
- Power draw per phase of 3 phase supplies

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Security and Access Control

Security considerations include:

- Pass cards
- Biometrics
- Intruder detection

Racks

Considerations for server and network racks include:

- Loading and weight
- Heat
- Power

Cabling

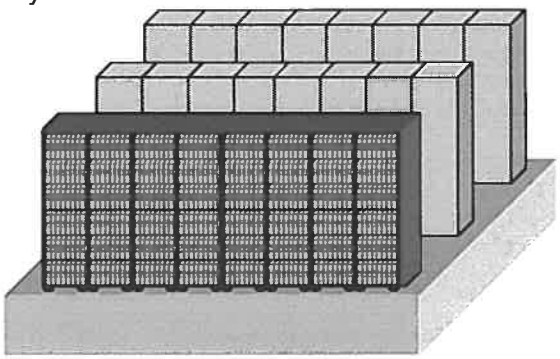
Considerations for cabling include:

- Cable labeling
- Patch panels in rack
- Cable management systems

https://t.me/learningnets

Blade Server Integration Considerations

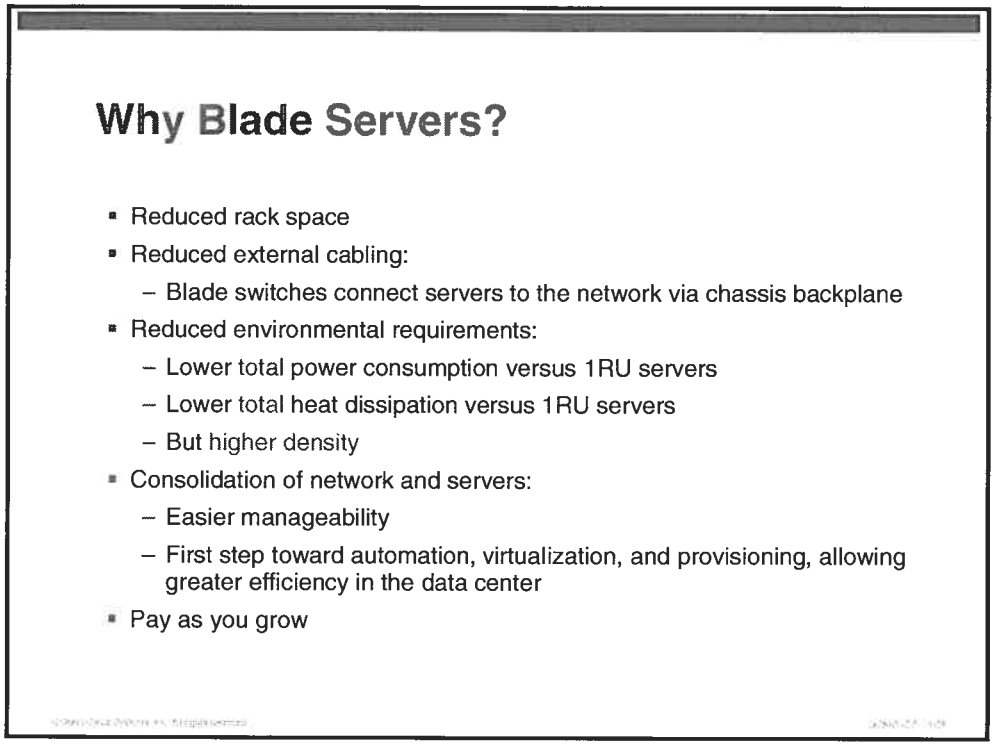
- Highly available and scalable physical architecture to support blade servers
- Primary services of facility:
  - Space
  - Power
  - Cooling
  - Cabling
- These services are challenged by the introduction of blade servers



© 2008 Cisco Systems, Inc. All rights reserved.

Blade servers are an integration alternative where space, power, cooling, and cabling are of primary consideration within the data center facility. Introducing blade servers addresses these concerns because blade servers provide tight integration. In order to support blade server architecture, the scalability and high availability of the physical architecture must be taken into consideration and built in to the data center design.

https://t.me/learningnets



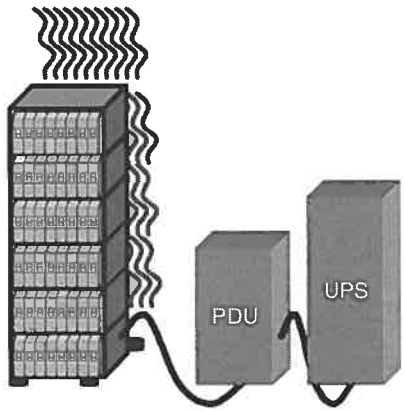
Blade servers are used within the data center for many reasons, including the following:

- Easier manageability
- Reduced rack space
- Reduced external cabling because blade switches connect servers to the network via chassis backplane
- Lower power consumption and lower heat dissipation
- Consolidation of network and servers, providing a first step toward automation, virtualization, and provisioning, all of which provide greater efficiency

https://t.me/learningnets

Physical Design: Increasing Heat

- Computing power and memory requirements:
  - Power = Heat
- IT staff demanding space saving servers:
  - Density = Heat
- Heat shortens equipment life cycle within the data center
- Blades and mainframes are the culprits:
  - Create hot spots



© 2008 Cisco Systems, Inc. All rights reserved.

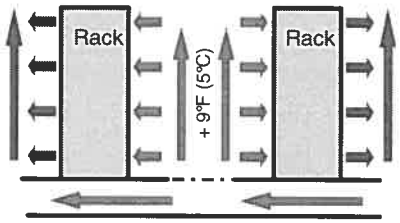
Increasing heat requirements must be considered when designing blade servers within the data center, because the higher density resulting from space-saving servers increases heat. The design must consider what cooling is required for the current sizing of the data center servers, plus future growth expectations:

- 3.5 BTU/kW
- 138 cfm/kW

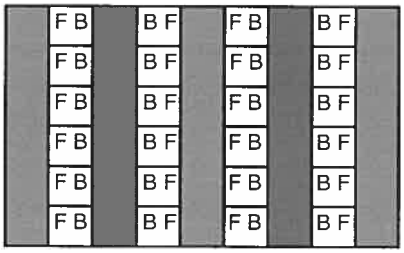
https://t.me/learningnets

Physical Design: Cooling Challenges

- Supply and remove cool and hot air to and from the rack/row
- ASHRAE guidelines:
  - Inlet temperature 68°F–77°F (20°C–25°C)
  - Humidity range of 40%–50%
- Cooling solutions:
  - Increase space
  - Increase number of HVAC units
  - Increase airflow
  - New approaches and technology



Rack and Row Design



The figure provides some guidelines for designing cooling for the data center blade servers.

© 2008 Cisco Systems, Inc.

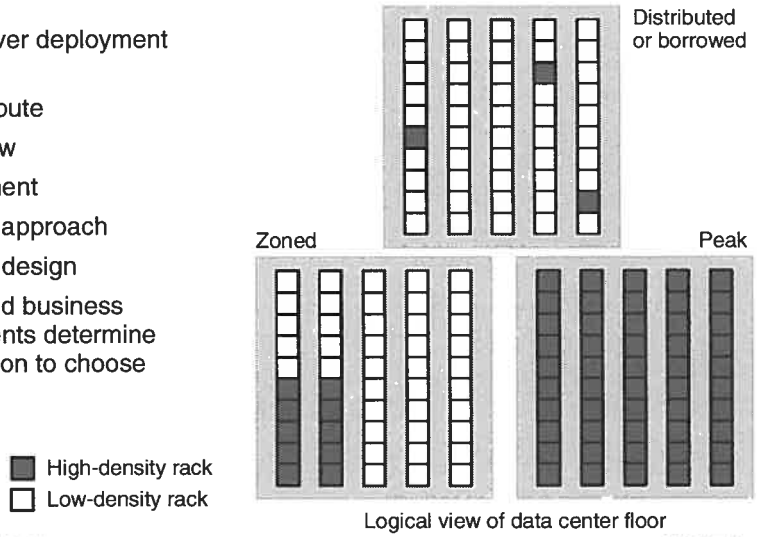
https://t.me/learningnets

Physical Design: Solutions

Blade server deployment options:

- Distribute
- Borrow
- Augment
- Zone approach
- Peak design

Facility and business requirements determine which option to choose



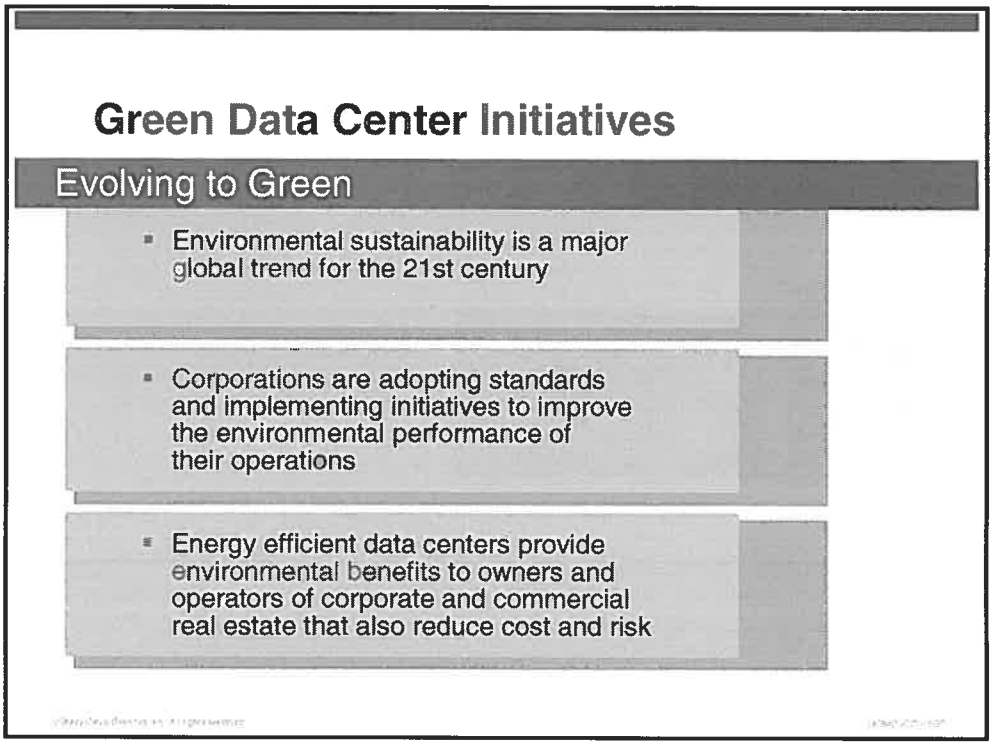
Planning a data center and operating a data center require different skill sets. It is important to work with the facilities team to plan data center physical design.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Green Data Center Initiatives

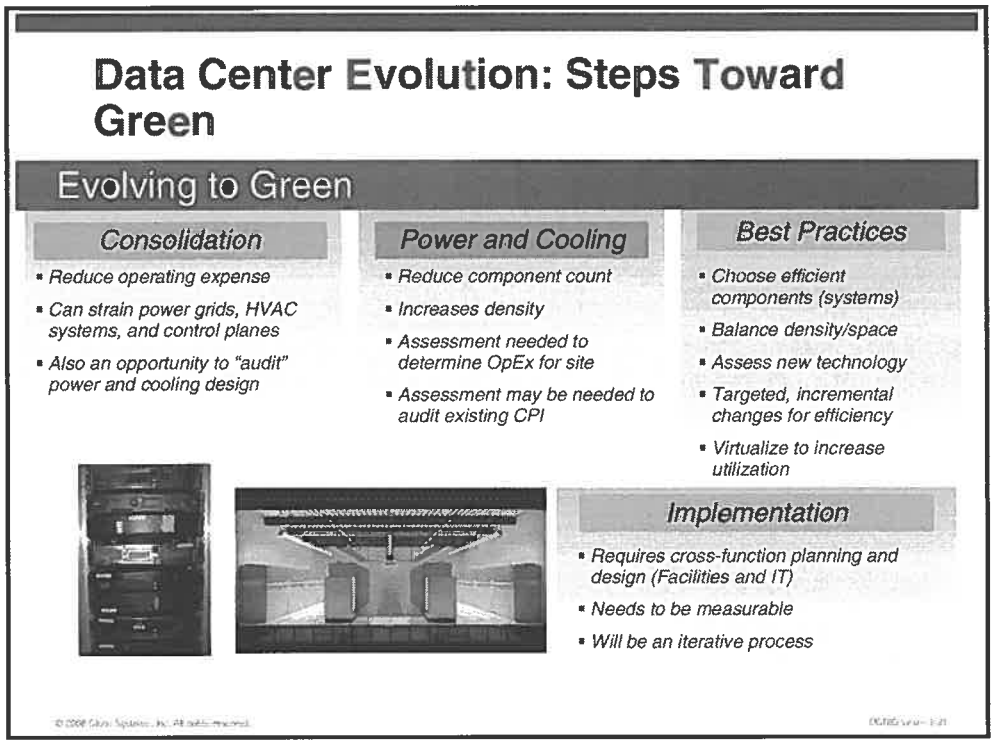
This topic identifies green data center initiatives and the corresponding Cisco activities and solutions.



Half of the data centers currently in use will have insufficient power and cooling capacity to meet the demands of high-density equipment in the near future. Through 2009, energy costs will emerge as the second-highest operating cost (behind labor) in 70% of data center facilities. Power demand for high-density equipment will level off or decline. In-rack and in-row cooling will be the predominant cooling strategy for high-density equipment

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

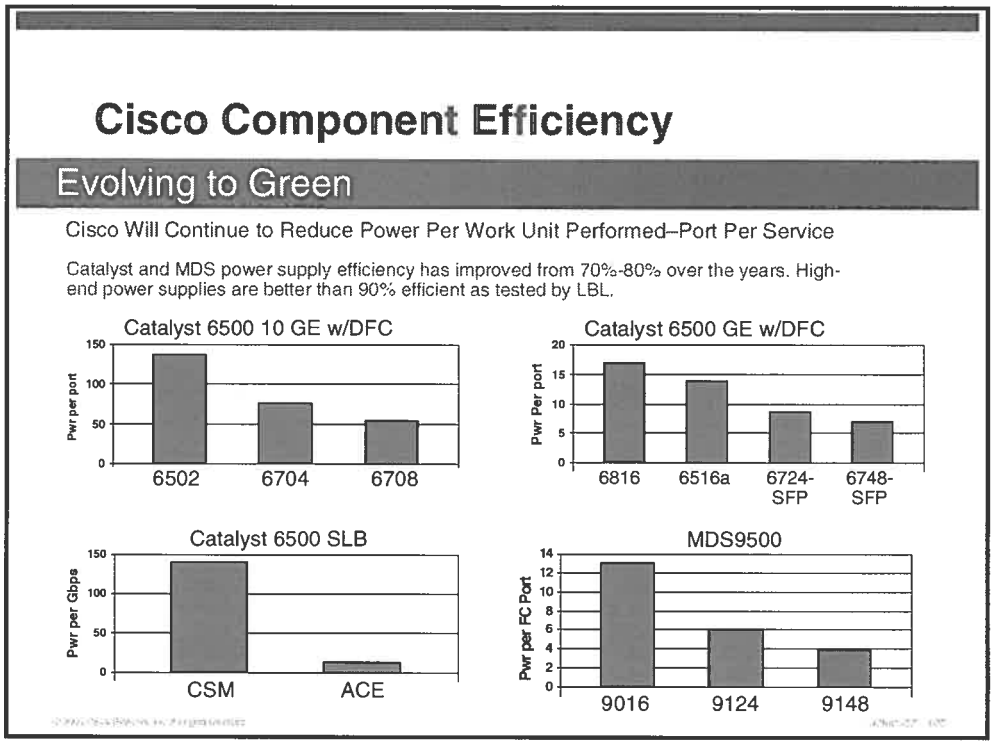


The following processes enable the green data center environment:

- Consolidation in data center reduces the data center operating expenses.
- Reduced data center power and cooling requirements are achieved by increasing the density of data center components.
- Virtualization of data center components increases resource utilization and reduces the data center footprint and management expenses.
- Cisco will continue to reduce power per work unit performed–Port Per Service

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

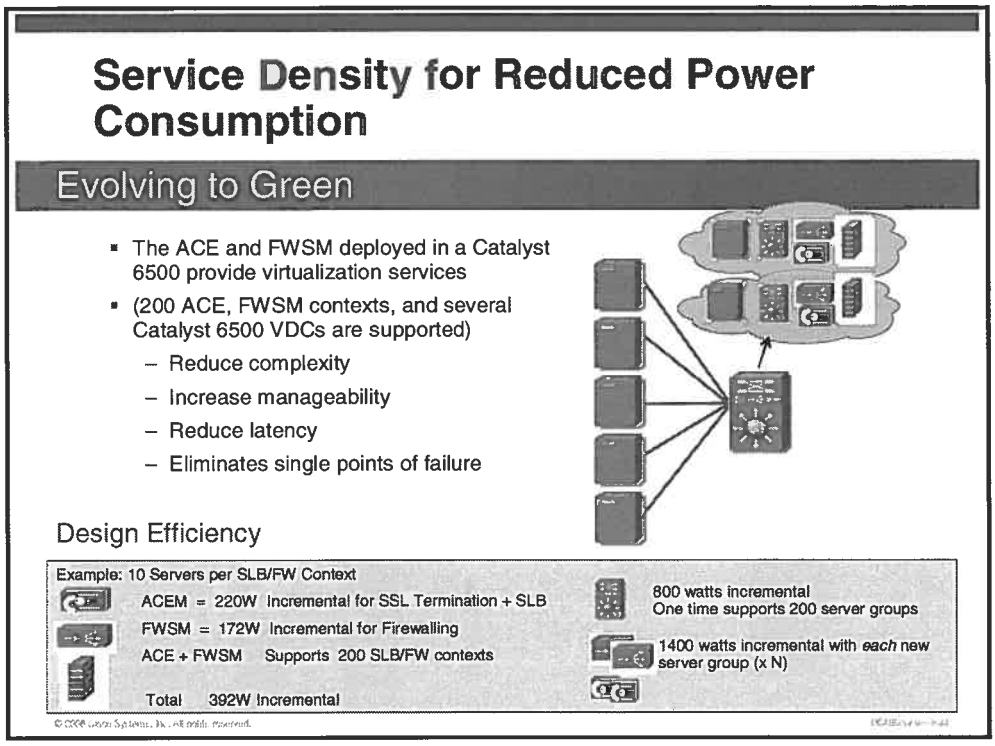


Cisco is also continuing to move the industry on Power Supply Efficiency. Cisco next generation power supplies are expected to be 90% efficient, resulting in less wasted power and less heat generated.

Cisco Catalyst and Cisco MDS Power Supply Efficiency has improved from 70% efficiency to 80% efficiency over the years. High-end power supplies are better than 90% efficient as tested by Lawrence Berkeley National Labs LBL.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The ACE and FWSM can be virtualized to support over 200 servers, each with unique policies.

Using virtualization, you can save large amounts of power versus using external appliances for tasks such as firewalling and server load balancing. For example, an appliance-based solution that includes a firewall and SLB would consume 1400W per server group in redundant mode.

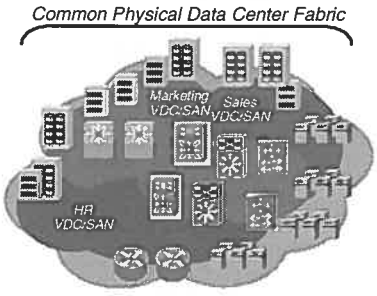
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Virtualization and Consolidation Solutions for Design Efficiency

Evolving to Green

- Build a network solution using virtual fabrics, virtual servers and storage devices
- Consolidate data center entities on the common physical infrastructure
- "Green" benefits:
  - Increased resource utilization (disk:~70%)
  - Greater power savings
  - Reduced footprint



Taking a tape subsystem off-line can save $3,800 in power and cooling per year

Virtualization Energy Benefits

Cisco virtualization solutions slow the growth of power demand through increased utilization and reduce component count. Cisco next-generation products and solutions will further reduce power consumption at the systems level. Cisco is focused on environmental concerns from executive direction to individual product design, and participates in the Green Grid. The Green Grid is a non-profit consortium dedicated to advancing energy efficiency in data centers and business computing ecosystems. Applying virtualization will save on real energy consumption in the data center. With the Cisco solution, we can use a single consolidated data center fabric, even down to one data center Catalyst or Nexus Series switch or MDS director, and still we can insure safe isolation of traffic using VDC, VLANs and VSANs. This solution saves money on acquisition costs of the network equipment, and lowers overall operating costs by saving power.

Consolidation Energy Benefits

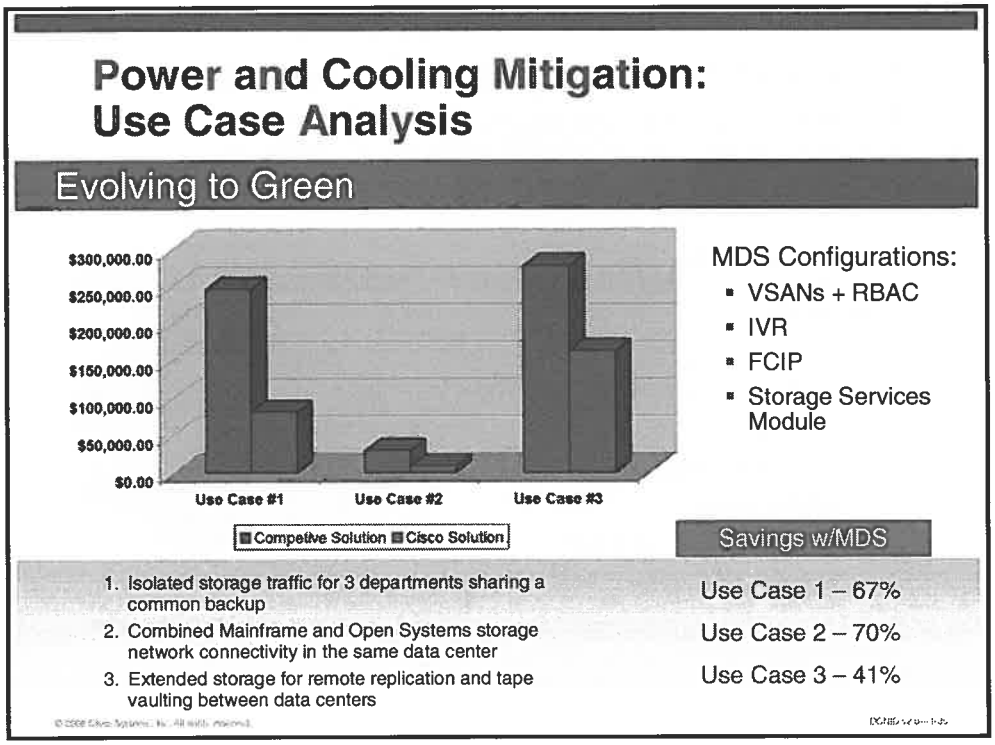
Computing, storage and network consolidation can save substantial energy. Purchasing fewer and larger blade servers and storage arrays provides these benefits:

- Consolidate and optimize computing and storage workloads
- Maximize resource utilization

Consolidated platforms enable shared resources and higher utilization. Tiered storage enables different storage for different workloads. "Pay as you grow" configuration flexibility enables more efficient utilization of capacity, performance and connectivity.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



This slide illustrates an example of a typical enterprise customer and what that customer expects to do over the next couple of years with respect to power and cooling. . Power consumptions for competitive gear were taken as 80% of kVA from data sheets compared to 100% of kVA on Cisco services modules. Power of the switched infrastructure was ignored because it is a common cost in either deployment model.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco MDS is More Power Efficient than the Closest Competitor

Evolving to Green

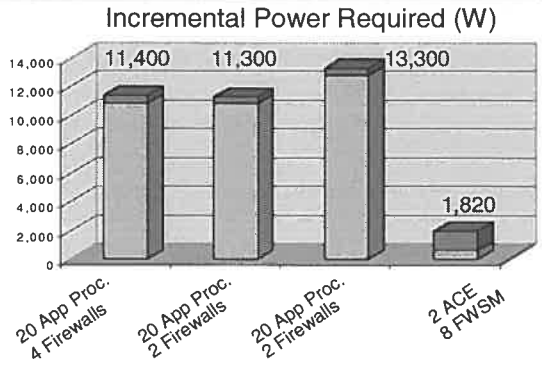
	Competitor's Solution	Cisco MDS Solution
Devices Required	5	1
Stranded (*wasted*) Ports and Slots	Yes - 90 ports plus 9 slots	None
Max. Number of Isolated SANs Available	3	256
Routing Bandwidth between SANs	Up to 5x4G ISLs: 20 G	Backplane BW: 48 G / slot
Power per Host Port	11.98 watts	3.23 watts
Power Breakdown	2 routers (600 W each) 3 Storage Directors: 2 CPs @ 100 W each 3 32-port modules @ 90 W each 2 16-port modules @ 90 W each 3 Chassis with fans @ 100 W each	1 Cisco MDS 9513: 2 48-port modules @ 130 W each 2 12-port modules @ 92 W each 2 Sup-2 @ 88 W each 2 Crossbars @ 44 W each 1 Chassis with fans @ 223 w
Total Power Required	3450 Watts	931 Watts

85% Reduction in Power Requirements with Networking Services Virtualization

Evolving to Green

Performance Requirement

- 10 Gb/s load balancing
- 20 Gb/s firewall
- 10 VDCs
- High availability



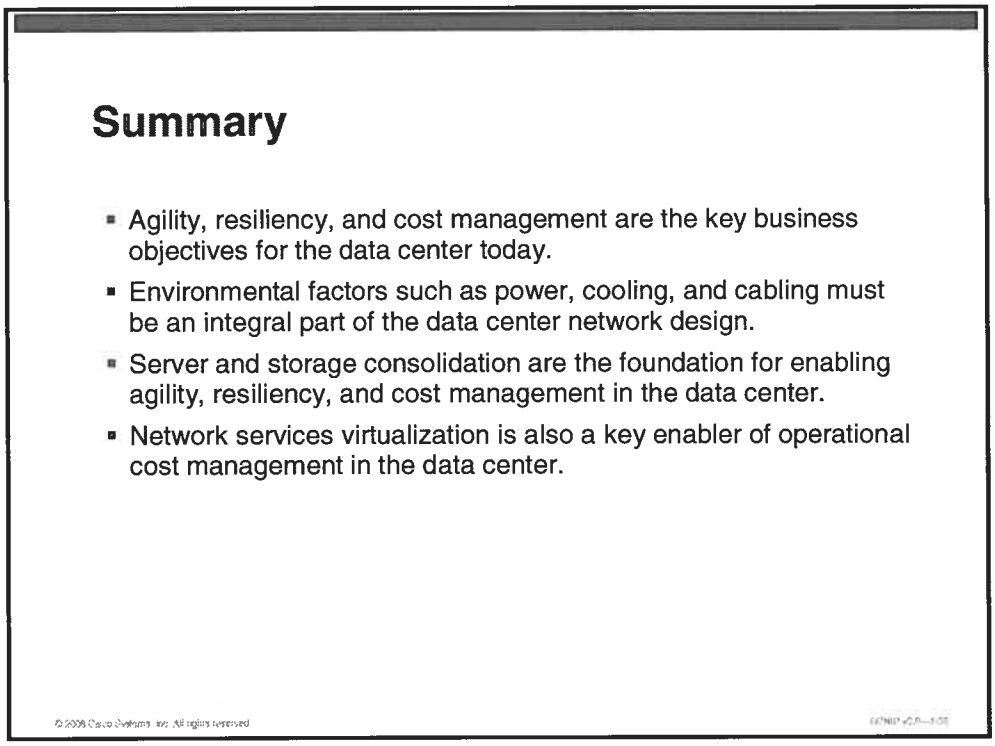
- 85% power reduction with virtualized, integrated modules ~11 kW
- Rack space saved with virtualized, integrated modules ~30RU
- Additional savings from reduced cabling, port consumption, and support costs

The figure shows the potential impact of consolidating network services appliances. Power consumptions for competitive gear were taken as 80% of kVA from data sheets compared to 100% of kVA on Cisco services modules. Power consumed by the switched infrastructure was ignored because it is a fixed cost in either deployment model. With 10 logical contexts on each ACE and FWSM, consolidating network appliances onto service modules saves 30RU and 11 kW.

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 2

Cisco Nexus 7010 Switch

Overview

This module introduces concepts and basic characteristics of the hardware and software architecture of Cisco Nexus 7010 switch, describes the modularity of hardware and software architecture, and describes the service availability for data center continual availability with zero loss.

Module Objectives

Upon completing this module, you will be able to discuss the hardware and software architecture of Cisco Nexus 7010 Switch. This ability includes being able to meet these objectives:

- Describe the Nexus 7010 Switch hardware architecture
- Describe the Nexus 7010 Switch software architecture
- Describe the Nexus 7010 Switch network management features
- Describe the Nexus 7010 NX-OS integrated security features
- Describe the QoS implementation on the Nexus 7010 Switch
- Position the Nexus 7010 Switch in the data center network

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 1

Cisco Nexus 7010 Hardware Architecture

Overview

This lesson describes the modularity of the Cisco Nexus 7010 Switch hardware architecture for data center continual service availability.

Objectives

Upon completing this lesson, you will be able to describe the components and operation of Cisco Nexus 7010 Switch hardware architecture. This ability includes being able to meet these objectives:

- Describe the Nexus 7010 Switch Supervisor engine hardware architecture
- Describe the Nexus 7010 Switch I/O module hardware architecture and the features relevant to data center network design
- Describe the Nexus 7010 fabric modules hardware architecture and system capacity
- Describe the Nexus 7010 Switch forwarding engine architecture
- Examine performance scalability factors, throughput capacity and bottlenecks of Nexus 7010 Switch hardware components.
- Describe the Nexus 7010 Switch power supply, system cooling and cable management hardware components

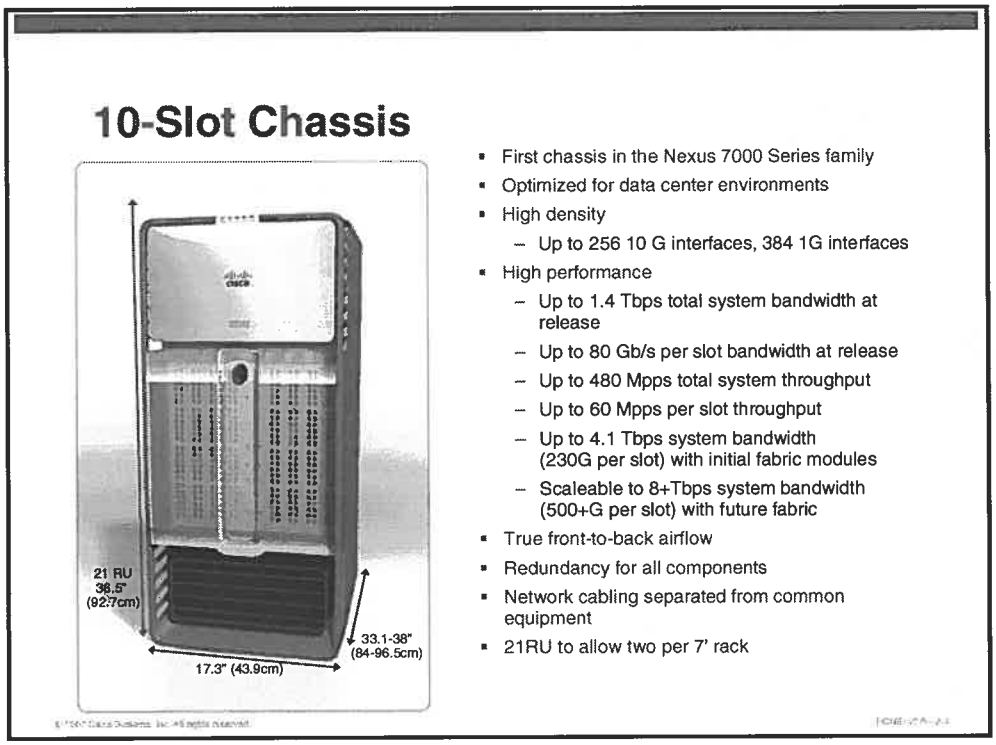
https://t.me/learningnets

Cisco Nexus 7010 Switch Supervisor Engine Architecture

This topic describes the characteristics of the Nexus 7010 Switch Supervisor Engine architecture.

Nexus 7010 Switch Supervisor Engine Architecture

This topic describes the Nexus 7010 Switch Supervisor Engine hardware architecture and management access provided by Connectivity Manager Processors (CMP).



The figure presents the front-face of the Nexus 7010 Series Switch hardware architecture based on a 10-slot chassis, 21RU high that supports up to 256 ports of 10 Gigabit Ethernet (GE) and 1.4 Tbps aggregated system throughput with up to 80 Gb/s per slot bandwidth for each I/O module. Two 10-slot Nexus 7010 Series chassis can be accommodated into a single rack. Cisco Nexus 7010 Series Switch supports front-to-back data center airflow. The fabric spines at the back, the airflow comes in the front and goes out the back of the chassis.

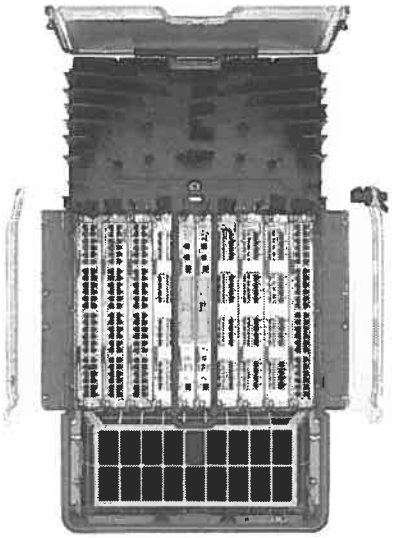
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 10-Slot Chassis

Nexus 7010 10-slot fabric

- Has dual redundant dedicated Supervisor Modules
- Supervisor Engine has the following characteristics and functions:
  - Control plane functions
  - Management functions
  - Provide continuous operation and high availability with no service disruption during HW/SW upgrade
  - Designed for mission critical data center networks



Cisco Nexus 7010 Series Switches are designed and focused specifically to meet data center requirements scalable to 15 terabits per second (Tbps). Nexus 7010 Switch has been positioned to provide unified I/O capabilities in the future using the current chassis, supervisor, and fabric modules. Nexus 7010 Series Switches comprise a modular data center-class product line designed to provide highly scalable 10 GE networks with future support for 40 and 100 GE interfaces. Purpose-built to meet the requirements of the most mission-critical data centers, Nexus 7010 Switch delivers continuous system operation and virtualized, pervasive services.

The Cisco Nexus 7000 Series is based on the Cisco NX-OS operating system, with enhanced features that deliver real-time system upgrades with exceptional manageability and serviceability. Its innovative design is built to support end-to-end data center connectivity, consolidating IP, storage, and interprocess communication (IPC) networks on a single Ethernet fabric.

The Cisco Nexus 7010 Supervisor module delivers scalable control plane and management functions for the chassis. The fully virtualized and distributed forwarding architecture of the Cisco Nexus 70000 Series performs all data plane processing on the I/O modules, freeing the supervisor to focus exclusively on scaling the control plane.

The 7010 Switch Supervisor Engine acts as the control module for the new architecture built around the fabric and crossbar modules that interconnect all the line cards with point-to-point, full-duplex serial connections. Supervisor Engine is a key component of the distributed forwarding architecture.

Dual dedicated supervisor modules provide exceptional high-availability features with no service interruptions even during hardware and software upgrades.

https://t.me/learningnets

The Nexus 7010 Series systems can be configured with one or two supervisors. If the system has two supervisor modules, you can replace one of the supervisors while the other one manages system operations. If the system has only one supervisor, you must bring the system down to replace the supervisor module because the supervisor module is required for managing operations.

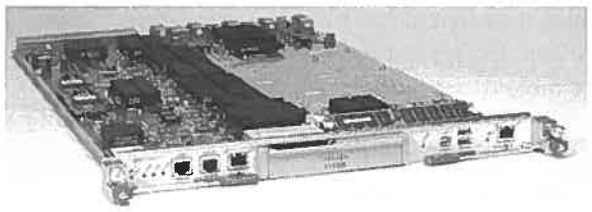
The Nexus 7010 Series Switch redundant supervisor modules can perform seamless supervisor switchover within a millisecond's time interval upon Nexus kernel failure and impact on data center operation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Supervisor Engine 1

- High-performance dual-core 1.6G Hz Intel Xeon processor with 4 GB DRAM
- Connectivity Management Processor (CMP) for lights-out management via dedicated 10/100/1000 Ethernet
- 2 MB NVRAM, 2 GB internal boot disk, 2 external compact flash slots
- 10/100/1000 management port with 802.1ae LinkSec (Not enabled in R4.0)
- Console and auxiliary serial ports
- One USB device port, 2 USB host ports
- Blue ID LED for easy location
- Resilient switched Gigabit Ethernet out-of-band channels to I/O modules
- Central arbiter providing I/O modules access to switch fabric



We are using a 1.6 GHz Intel Xeon with 4 GB of DRAM which ships standard on the Supervisor. There is a CMP for lights-out management, the connectivity management processes which we will go through in great detail in the next presentation. There is 2 MB of NVRAM, but we actually installed a config in flash so it is actually different there.

There are two external compact flash slots which are underneath the cover and there are two USB thumb drive slots which you can use to copy an image to the system.

There is another GE port, which is for the RP for out-of-band management and the Central Arbiter sits on the Supervisor very much like the MDS for the central arbitration for the fabric.

The Cisco Nexus 7010 10-Slot fabric module delivers the fault-tolerant fabric that provides parallel fabric channels to each I/O and supervisor module slot. Up to five simultaneously active fabric modules can work together delivering 230 Gb/s per slot of bandwidth.

The '32 Gb/s' Data Bus (and Results Bus) that runs up and down the Catalyst 6500 backplane is not implemented on the Nexus 7010 System switch. As such, the Nexus does not have such switching modes as 'truncated mode', 'compact mode', and 'flow through mode'.

The Nexus 7010 Series switch uses a mid-plane architecture that directly connects the fabric modules to the I/O modules for direct data transfer. All intermodule switching is performed via the multi-stage crossbar fabric.

For traffic between ports that are connected to the same bridge/replication ASIC (also known as Metropolis), the data does not need to hit the fabric and will stay locally within the I/O module. In any cases other than that, the traffic will need to go through the 3-stage fabrics even though the ingress/egress ports are located on the same I/O module.

https://t.me/learningnets

Unicast Traffic Load Balancing

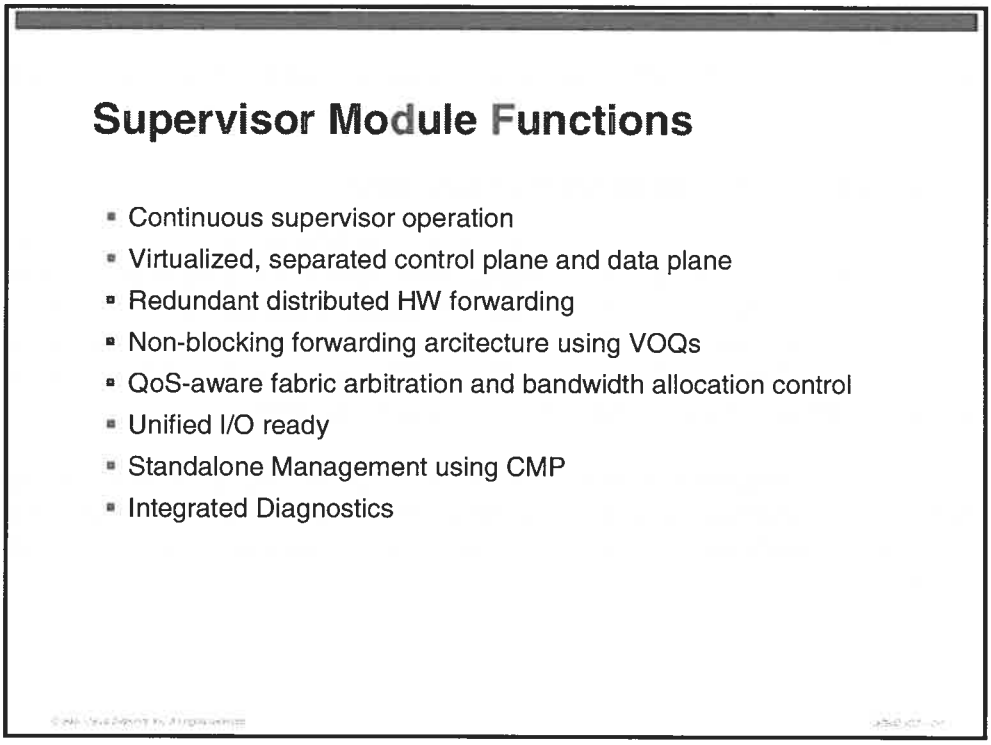
Nexus 7010 Series Switch utilizes a centrally arbitrated fabric. When packets are ready to be sent into the fabric, a request for a credit to send for a particular QoS level is sent to the central arbiter. The arbiter will send a grant back to the line card when there is buffer space available on the egress line card for that QoS level. A grant gives the line card the ability to send up to 9216 bytes. For frames larger than 2560 bytes, the ingress fabric device fragments the frame into 2560 byte fragments that are sent across all the available paths between the source and destination using a pseudo-random round-robin algorithm. The load-balancing is on per packet or per packet-fragment basis. The egress fabric interface is responsible for re-ordering and reassembling the packets into the original frame before sending onto the next stage.

Supervisor Bandwidth Capacity

Each supervisor slot has a single 'channel' of connectivity to each fabric card, for a total of 5 crossbar channels. Note that at least in first-generation Nexus 7000 supervisors there are neither front-panel ports nor any forwarding complex on the supervisors, so this crossbar capacity is used entirely for inband communication only and never for forwarding traffic. However, this does not preclude future supervisors from having a forwarding complex and/or front-panel ports. Five channels, all running at 23.07 Gb/s, provide 115.3846 Gb/s of 'usable' full-duplex system capacity per supervisor slot.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Continuous Supervisor Operation

The Cisco Nexus 7000 Series Supervisor module delivers a comprehensive set of features to address the needs of the most demanding data center deployments. This module is based on the high-performance dual-core Intel Xeon processor. The multithreaded operating system runs on a dedicated dual-core processor. The flexible dual-core design harnesses the maximum processing power for control plane processing with enhanced reliability. Supervisor module has resilient dedicated connections to each fabric module and I/O module slots. Dual redundant Ethernet out-of-band channels (EOBCs) to each module and fabric provide resilience for communication between the control and line card processors. Two supervisor modules operate in active and standby modes with stateful supervisor failover to enhance total system availability.

Virtualized Separated Data Plane and Control Plane with Distributed HW Forwarding

Fully decoupled control plane and data plane with no hardware forwarding on the module enable scalable control plane processing. Control plane can be instantiated to represent virtualized switch operations. The fully virtualized and distributed forwarding architecture of the Cisco Nexus 7010 Switch performs all data plane processing on the I/O modules for the supervisor to focus exclusively on scaling the Control plane.

Non-Blocking Forwarding Using VOQs and QoS-Aware Arbiter Fabric Arbitration for Bandwidth Control

Centralized arbiter is based on supervisor module and has active-active control of the flow of traffic through the switch fabric to ensure transparent switchover with no data losses. Arbiter is quality-of-service (QoS) aware for lossless fabric switching and uses huge number of VOQs to avoid the problems associated with head-of-line blocking. Arbiter controls and allocates available fabric bandwidth among I/O modules. Using VOQs arbiter enables non-blocking switching architecture for non-blocking forwarding among all fabric ports.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Unified I/O Ready

Hardware architecture has all the necessary features to enable future multiprotocol support for Fibre Channel over Ethernet (FCoE)or clustering protocols.

Standalone Management and Integrated Diagnostics

The Cisco Nexus 7000 Series Supervisor Module incorporates highly advanced analysis and debugging capabilities and proactive management tools to verify connectivity and mechanisms for capturing and analyzing traffic. It supports redundant standalone out-of-band Ethernet management, separated from other OS, which eliminates the need for a terminal server. The Power-On Self-Test (POST) and Cisco Generic Online Diagnostics (GOLD) provide proactive health monitoring both at startup and during system operation.

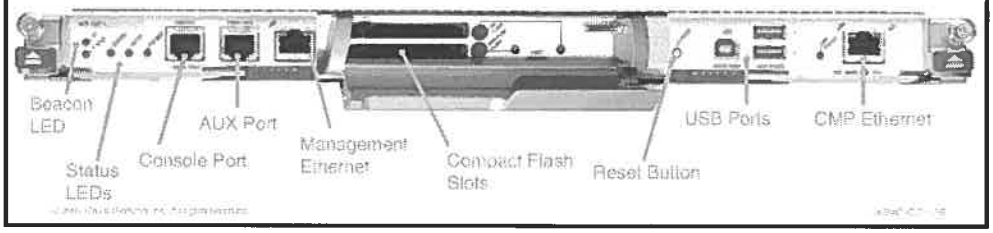
The supervisor module uniquely provides a built-in packet capture with an embedded control plane protocol analyzer—a protocol decoding tool that allows analysis of control plane traffic to improve network planning, provide faster operation response times to events, and reduce operating costs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Supervisor Engine Components

- Dual-core 1.66G Hz Intel Xeon processor with 4 GB DRAM
- Connectivity Management Processor (CMP) for lights-out management
- 2 MB NVRAM, 2 GB internal boot disk, 2 external compact flash slots
- 10/100/1000 management port with 802.1AE LinkSec
- Console and auxiliary serial ports
- One USB device port and two USB host ports for file transfer
- Blue Beacon LED for easy identification
- Resilient switched Gigabit Ethernet out-of-band channels to I/O modules
- Central arbiter providing I/O modules access to switch fabric



The Cisco Nexus 7010 Series Switch Supervisor Engine includes the following new components:

- Dual-core 1.66 GHz Intel Xeon processor with 4 GB DRAM
- Connectivity Management Processor (CMP) for lights-out management
- 2 MB NVRAM, 2 GB internal boot disk, 2 external compact flash slots
- 10/100/1000 management port with 802.1AE LinkSec
- Console and Auxiliary serial ports
- USB ports for file transfer
- Blue Beacon LED for easy identification

Three USB ports are also on the front panel to enable access to USB flash memory devices for software image loading and recovery. Two are designated for host use and one for device use. Host USB ports can be used to plug in devices such as a PC, while device ports can be used for attaching devices such as a Flash memory key.

LEDs provide visible indication of the supervisor ID, active or standby status, power management, and activity on the compact flash drive.

Blue Beacon LEDs allow for easy FRU identification for servicing. Using the Beacon feature, the administrator can clearly identify the chassis and supervisor module.

Log and expansion flash memory can be used for removable devices for logging and software images.

The compact flash memory slot cover protects against accidental damage or removal.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Connectivity Management Processor (CMP) enables management connection without using a terminal server. It uses Ethernet management interface exclusively for the system management. There is dedicated "management" virtual routing and forwarding (VRF) with the following characteristics:

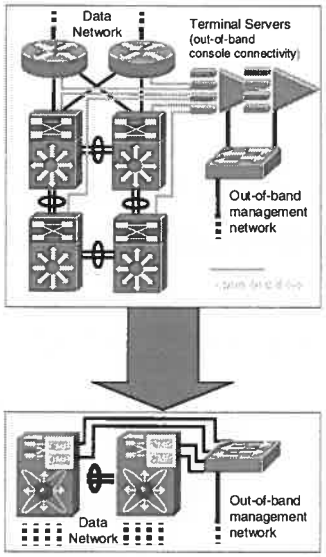
- Prevents data plane traffic from entering and exiting from management 0 interface
- Cannot move management interface to another VRF
- Cannot assign other system ports to management VRF
- Is capable of IEEE 802.3ae LinkSec encryption (not enabled in 4.0 release)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Connectivity Management Processor (CMP)

- Standalone, always-on microprocessor on supervisor engine
- Provides lights-out remote management and disaster recovery via 10/100/1000 interface
  - Removes need for terminal servers
- Monitors supervisor and modules, access log files, power cycle supervisor, etc.
  - Runs lightweight Linux kernel and network stack
  - Completely independent of DC-OS on main CPU



The CMP provides a complete out-of-band management and monitoring capability independent from the primary operating system. The CMP is a standalone, always-on microprocessor on the supervisor engine that runs lightweight Linux kernel and network stack completely independent of the NX-OS on the main CPU.

The CMP provides remote management and disaster recovery via 10/100/1000 Ethernet interface. The CMP enables remote monitoring and management of the supervisor module, all other modules and the Cisco Nexus 7010 Series Switch without the need for separate terminal servers with the associated additional complexity and cost.

The CMP delivers the remote control through its own dedicated processor, memory, and boot flash memory and a separate Ethernet management port. The CMP can reset all system components, including power supplies if necessary. It can also reset the host supervisor module to which it is attached, allowing a complete system restart.

The CMP enables system reset while retaining out-of-band Ethernet connectivity. The need for onsite support during system maintenance is reduced. Administrator has the capability to view boot-time messages remotely and has complete visibility during the entire boot process. Local operator intervention is not needed to reset power for devices or for system power shutdown and restart.

CMP supports login authentication for secure access to the out-of-band management environment. The CMP has access to critical supervisor log information to enable rapid detection and prevention of potential system problems. It is possible to take full console control of the supervisor and complete control of the operating environment. The CMP has dedicated front-panel LEDs that clearly identify the CMP status separate from that of the supervisor.

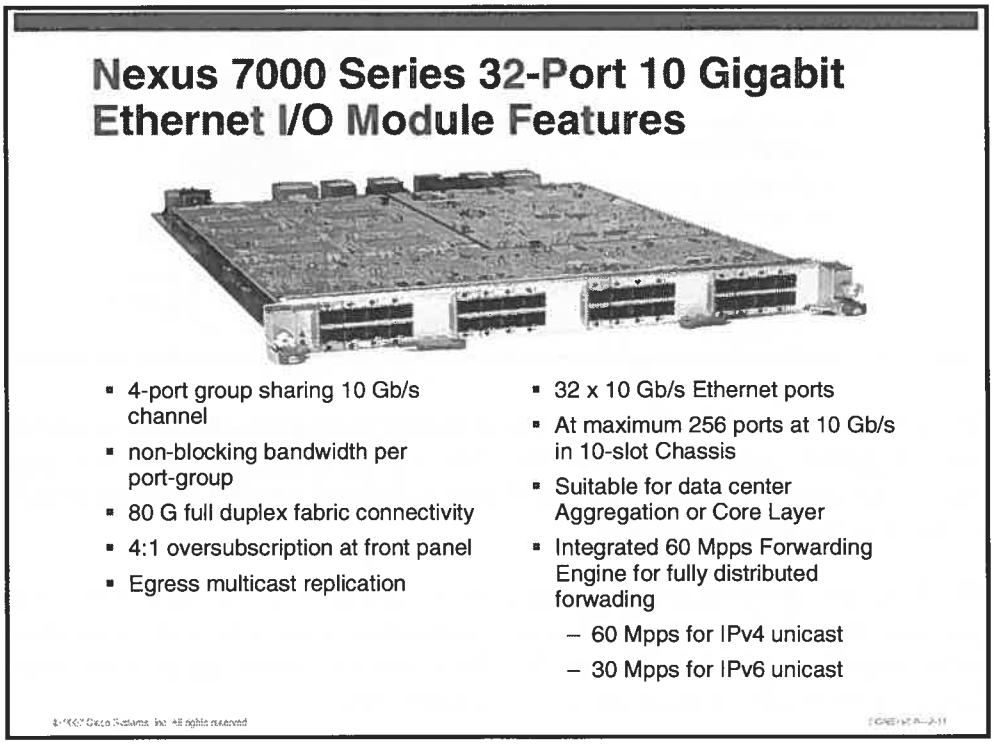
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch I/O Modules Architecture

This topic describes the characteristics of the Nexus 7010 Switch I/O module hardware architecture. The Nexus 7010 Series Switch 32-port and 48-port I/O module hardware architecture and functions will be introduced and discussed.

32-Port 10 Gigabit Ethernet I/O Module Features



The Cisco Nexus 7000 Series 32-Port 10 Gigabit Ethernet Module with 80 Gb/s of bandwidth to the fabric is a high performance, high-density 10 Gigabit Ethernet Module designed for mission-critical Ethernet networks. This module supports online insertion and removal (OIR).

Populating the 10-slot chassis with this module delivers up to 256 ports of 10 Gigabit Ethernet in a single chassis, or up to 512 ports of 10 Gigabit Ethernet in a single rack. This configuration is ideal for the core layer or the aggregation layer of a data center network, where density, performance, and continuous system operation are critical.

This module can also be used as a 10 Gigabit Ethernet uplink when employing the Cisco Nexus 7000 Series 48-port 10/100/1000 Module for the access layer.

All Cisco Nexus 7000 Series I/O modules contain an integrated forwarding engine. This architecture scales the forwarding performance of the chassis linearly by the number of I/O modules employed. The forwarding engine on the Cisco Nexus 7000 32-Port 10 GE module is part of the Cisco Nexus 7000 M Series of forwarding engines. Because this is the first generation of the M Series, the engine is referred to as the M1 forwarding engine. The M1 forwarding engine delivers 60 million packets per second (Mpps) of Layer 2 and 3 IPv4 unicast forwarding or 30 Mpps of IPv6 unicast forwarding across all ports on a single I/O module. A 10-slot chassis with eight M1 I/O modules delivers up to 480 Mpps of IPv4 unicast forwarding.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Multicast forwarding is built into the I/O module performing egress replication. The M1 forwarding engine also delivers access control list (ACL) filtering, marking, rate limiting, and NetFlow with no effect on performance.

The fabric interface on the Cisco Nexus 7000 32-Port 10 GE Module delivers 80 Gb/s of bandwidth in each direction. Traffic destined for a different module is distributed across up to five fabric modules. At least two fabric modules are required in the chassis to realize the full 80 Gb/s of bandwidth. Although this configuration offers fault tolerance, if one of the fabric modules fails, loss of bandwidth will occur. With three fabric modules installed in the chassis, one of the fabric modules can fail with no loss of bandwidth.

I/O Module Bandwidth Capacity

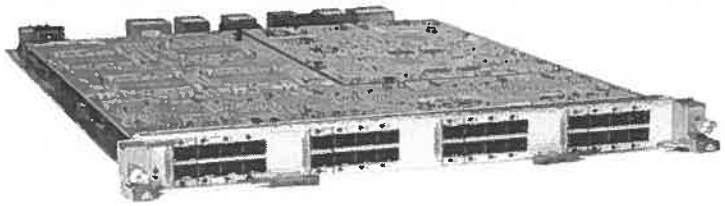
Initially shipping I/O module bandwidth: 80 Gb/s per slot, assuming there are 8 * 10 G ports in dedicated mode per I/O module in the Nexus 7010 chassis:

```
(80 Gb/s/slot) * (8 payload slots) = 640 Gb/s
(640 Gb/s) * (2 for full duplex operation) = 1280 Gb/s = 1.2
Tbps system bandwidth
```

Note The I/O modules shipping with the first software release (4.0) won't be able to leverage the full bandwidth of the 10 available crossbar channels. They are bandwidth-limited by other ASICs used on the modules. So while the system can provide 230 Gb/s of usable bandwidth per slot from day one, whether this bandwidth is fully utilized depends on the module design. Future modules already in the roadmap will leverage all the available slot bandwidth.

https://t.me/learningnets

Nexus 7000 Series 32-Port 10 Gigabit Ethernet I/O Module Features (Cont.)



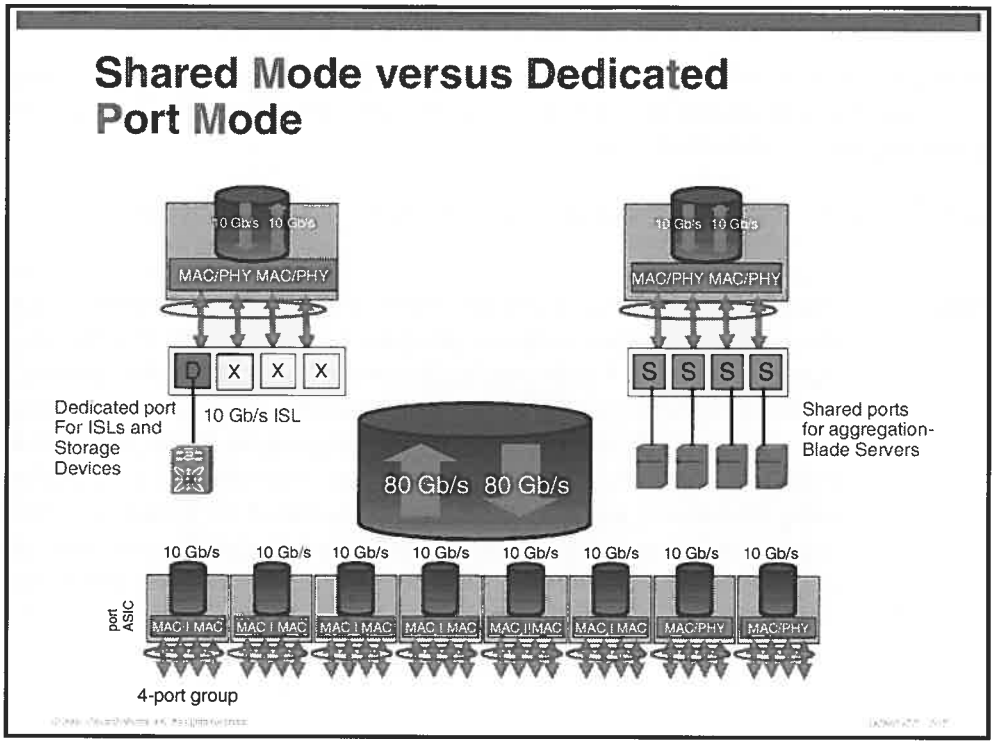
- Cisco TrustSec–for scalable security policy enforcement based on Security Group Tag (SGT)
- Hop-by-hop ISL encryption 802.1AE LinkSec on every port
  - Based on AES chip
  - 128-bit key
- ACL filtering, marking, rate limiting
- Netflow supported without impact on performance
- HW-based security enforcement :
  - L2-L4 ACL
  - Secure Group ACLs (SGACL)
- Virtual Output Queue (VOQ) for non-blocking forwarding architecture
- Quality of Service (QoS)
  - QoS queues: 8q2t ingress, 1p7q4t egress
- Buffers: 7.5 MB ingress, 6.2 MB egress

Powerful ACL processing supports up to 64,000 entries per module, where entries can address Layer 2, 3, and 4 fields in addition to new Cisco metadata fields that employ security group tags (SGTs).

The Cisco Nexus 7000 32-Port 10 GE Module offers exceptional security with integrated hardware support for Cisco TrustSec. This security includes line rate data confidentiality, data integrity and ACL processing for Security Group Tags. Data confidentiality and integrity conform to the IEEE MAC security standard (IEEE 802.1AE). All 32 ports on the module support the Advanced Encryption Standard (AES) cipher using a 128-bit key. New security ACLs are enhanced through hardware support for Cisco metadata headers capable of carrying SGTs. Security group ACLs (SGACLs) use SGT information to provide hardware-based enforcement of security policies. This enhancement removes dependencies on IP addresses thus improving scalability and simplifying manageability.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The module has two forwarding modes: dedicated mode or shared mode.

The Cisco Nexus 7000 32-Port 10 GE Module can deliver 8 ports at line rate, or allow up to 32 ports to share 80 Gb/s of bandwidth. The 32 ports are organized into eight groups of 4 ports. In dedicated mode, the first port in each 4-port group is active, delivering line rate performance while the other 3 ports are disabled. In shared mode, all 4 ports in the group are active. The software allows each group to be individually configured for dedicated mode or shared mode.

Shared Mode (Default)

In shared mode all 4 ports are sharing the 10 Gb/s bandwidth channel and all 4 can be active and forward the traffic. When set to this mode the line card presents at maximum 4:1 oversubscription (4* 10 Gb/s = 40 Gb/s, 40 Gb/s/10 Gb/s = 4:1). All 4 ports can have different bandwidth requirements in shared mode but the sum of their throughput required by all four attached devices at any time must not exceed the capacity of the 10 Gb/s channel. This mode is appropriate for attached server ports or aggregating blade servers' connection to the fabric. In shared mode all 32 ports can be enabled and available.

Shared mode = 100 MB ingress, 80 MB egress throughput.

Dedicated Mode

If a port requires dedicated port bandwidth such as inter-switch links or tape devices or storage arrays, then dedicated mode should be enabled on that port. In dedicated mode the 10 Gb/s shared channel bandwidth is allocated and available only to the dedicated port in the port group and other three ports must be disabled. The determination of the forwarding mode is made on a group of 4 ports basis. For each group of 4 ports, the port that is available in dedicated mode is the first port on the block (1, 3, 5, 7 are a group of 4 ports, port 1 is the only active port in dedicated mode). The complete list of dedicated mode ports is 1, 2, 9, 10, 17, 18, 25, and 26.

Dedicated mode = 100 MB ingress, 80 MB egress throughput.

https://t.me/learningnets

Differences Between Dedicated and Shared Mode

In shared mode if only one port of the 4 ports is used, the effective throughput available is close to 10 Gb, but is limited by the mux effect on the port. Dedicated mode removes the MUX, providing up to 10 Gb of line rate.

2:1, 3:1 and 4:1 modes are available by electing to use 2, 3 or 4 of the ports.

Note The line-rate throughput on a single 32-port 10 G module interface is less than 10 Gb/s even though the other 3 ports in the port group are not passing traffic The MUX chip that collapses a group of 4 front-panel interfaces into a single 10 G pipe operates in two different modes: transparent mode and MUX mode. In shared mode, an additional header is added to every frame passing through the chip, consuming a small amount of bandwidth that is therefore not available for "wire" frames. In dedicated mode, this additional header is not used. Therefore, to achieve line rate on a single port on this module, you must configure the appropriate port group for dedicated mode rather than shared mode, which places the MUX chip in transparent mode and disables the other 3 interfaces in the port group.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Hot-Swappable SFPs for 32-port I/O Module

Optics	Media	Distance
10 Gb/s SW, SR SFP+	micron multimode fiber	300m
10 Gb/s SW, LR SFP+	micron single-mode fiber	10km



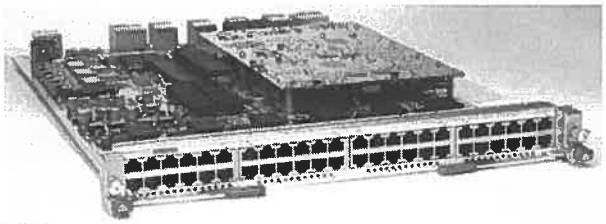
The physical interfaces on the Cisco Nexus 7000 32-Port is 10 G Ethernet module that supports SFP+ pluggable optics to meet the needs of a variety of distances and types of fiber cable.

Two types of SFP+ are used: the short reach (SR) based on MMF with maximum distance of 300m and the long reach (LR) are based on SMF+ and maximum distance of 10km. SFP+ optics are about the same form factor as SFP but they are smaller than SFP to enable high port density.

https://t.me/learningnets

48-Port 1 Gigabit Ethernet I/O Module Features

- 48 1 Gigabit Ethernet 10/100/1000 RJ-45 ports
- 40 G full duplex fabric connectivity
- Fits into data center Access Layer
- Integrated 60Mpps FE for fully distributed forwarding
- VOQs enable controled fabric bandwidth and non-blocking forwarding architecture
- 802.1ae LinkSec on every port
- Buffer: 7.5 MB ingress, 6.2 MB egress
- Queues: 2q4t ingress, 1p3q4t egress
- Blue Locator ID LED for easy location
- Comprehensive set of security features including CTS



The Cisco Nexus 7000 48-Port 10/100/1000 Ethernet module with 40 Gb/s of bandwidth to the fabric is a high-performance, highly scalable module designed for mission-critical Ethernet networks. Populating the 10-slot chassis with this module delivers up to 384 ports of 10/100/1000 Ethernet in a single chassis, or up to 768 ports of 10/100/1000 Ethernet in a single rack. The Cisco Nexus 7000 48-Port 10/100/1000 Ethernet module is ideal for the data center network access layer, where high density, high performance, and continuous system operation are crucial. This module supports OIR. Nexus 7010 Switch uses 24/26 encoding scheme on its interfaces. A sample serial link using 8b/10b encoding could provide 20 GB throughput while the same speed serial link can provide 23 GB throughput using a 24/26 encoding scheme. The forwarding engine is capable of doing 60Mpps and supports VOQs to control fabric bandwidth and for lossless non-blocking forwarding architecture. There are 2 queues per input port and 4 thresholds (2q4t) and 1 strict priority queue, 3 Deficit-Weighted Round-Robin (DWRR) queues, and 4 thresholds (1p3q4t) per output port. Every port supports 802.1ae LinkSec hop-by-hop encryption. The ingress port buffers size is 7.5 MB and there are 6.2 MB of egress port buffers. The ingress and egress buffers and queues are totally independent of VOQs. Unicast traffic is subject to VOQs, whereas multicast is not. Multicast forwarding is built into the I/O module performing egress replication. The M1 forwarding engine also delivers access control list (ACL) filtering, marking, rate limiting, and NetFlow with no effect on performance. Powerful ACL processing supports up to 64,000 entries per module, where entries can address Layer 2, 3, and 4 fields in addition to new Cisco metadata fields that employ security group tags (SGTs).

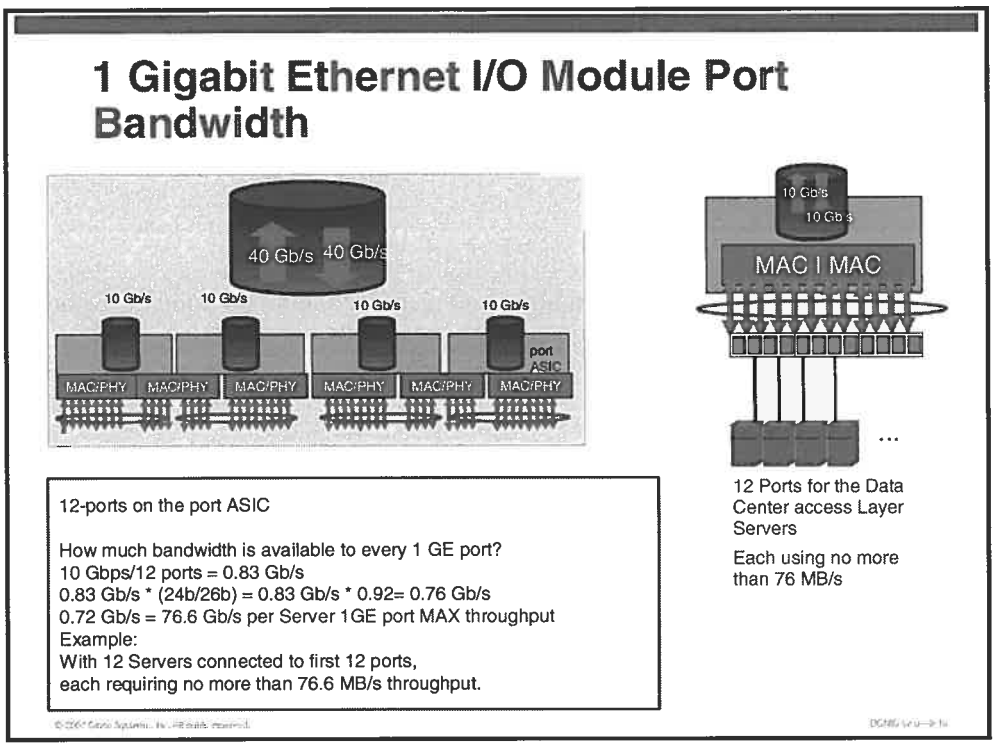
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

48-port 1 Gigabit Ethernet I/O Module Security Features

- Cisco TrustSec–for scalable security policy enforcement based on Security Group Tag (SGT)
- Hop-by-hop ISL encryption 802.1AE LinkSec on every port
  - Based on AES chip
  - 128-bit key
- ACL filtering, marking, rate limiting and Netflow supported without impact on performance.

https://t.me/learningnets



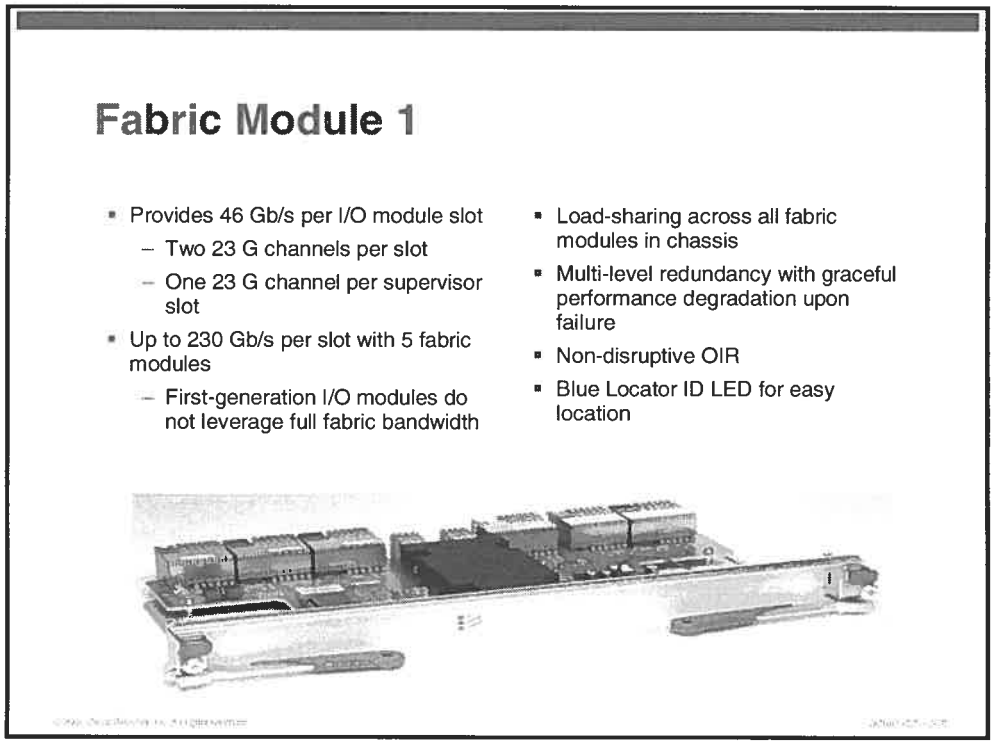
This line card has 46 optimized host ports and fits in the data center access layer. The slide shows the maximum throughput available on every port, assuming that all ports are activated. Ports on the same ASIC may also be ISLs, however 10 Gb/s port ASIC bandwidth will not be dedicated, but shared among activated ports..

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Fabric Modules Architecture

This topic describes the characteristics of the Nexus 7010 Series Switch fabric module architecture and operation.



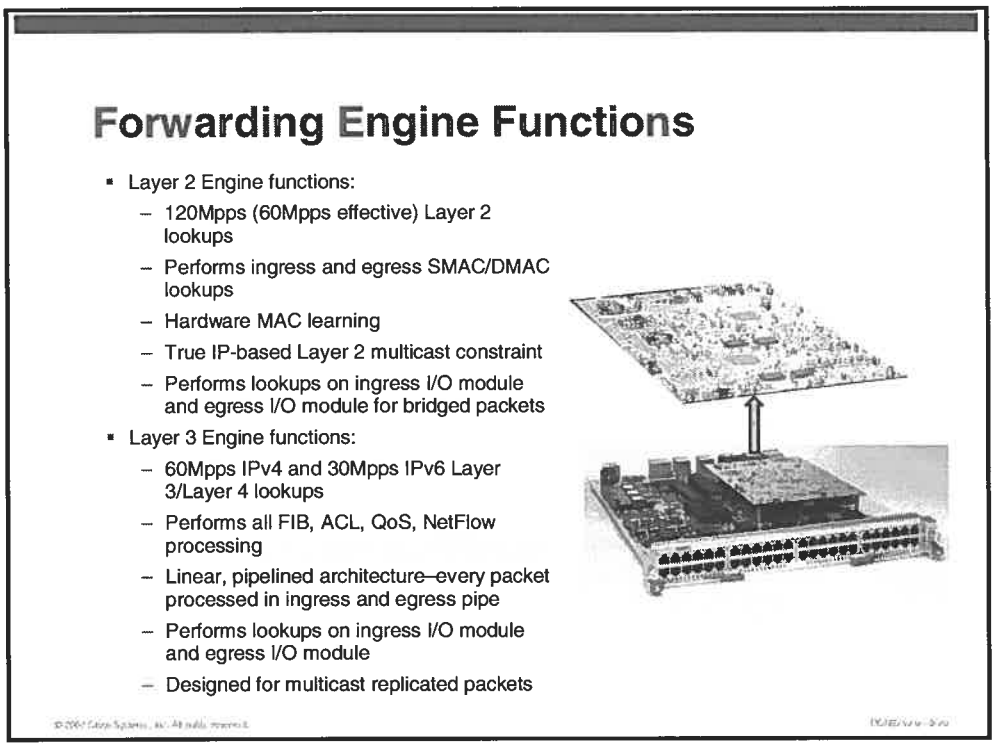
The Nexus 7010 Switch fabric module has two 23 Gb/s channels per slot for each fabric module and one 23 Gb/s channel per supervisor slot. Five installed fabric modules provide 230 Gb/s per slot. Fabric modules are load sharing unicast traffic across all the installed fabric modules in the chassis. Fabric modules provide multi-level redundancy with the possibility of graceful performance degradation upon module failure. Nondisruptive online module insertion and locator ID LEDs are supported.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Forwarding Engine Architecture

This topic describes the characteristics of the Nexus 7010 Series Switch forwarding engine architecture and operation..



The Layer 2 Engine Functions

The Layer 2 forwarding engine (FE) provides 128Mpps but 60 Mpps effective throughput taking into consideration the ingress and egress lookups. Other functions include: hardware MAC learning (the same as Catalyst 65000 Series switches implement the MAC learning) and IP-based Layer 2 multicast constraint. On the EARL 8 card the multicast IP address is being checked, as opposed to the EARL 7 multicast implementation, where the look at the multicast addresses is done based on the MAC address (which introduces the problem of having multiple IP Multicast addresses mapped to the same MAC address). Bridge packet lookups are performed on the ingress I/O modules and on the egress I/O modules.

The Layer 3 Engine Functions

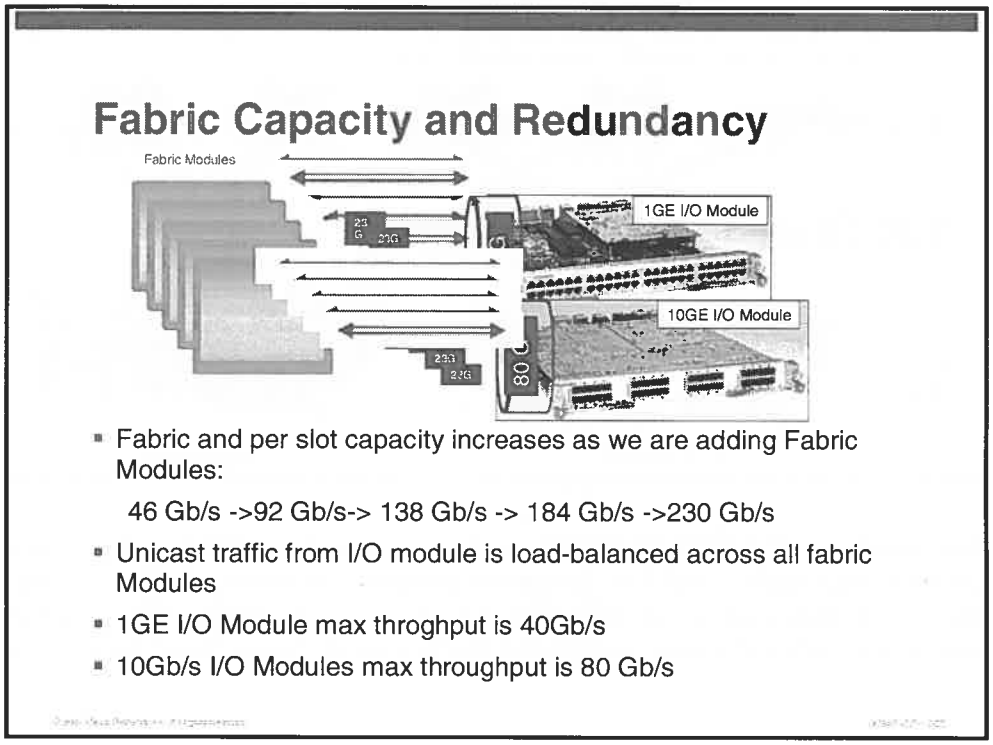
The Layer 3 forwarding engine (FE) supports 60Mpps of IPv4 packets throughput and 30Mppps for IPv6 packets lookups. The FE performs all FIB, ACL, QoS and Netflow processing. It supports TCP flags as well and it supports packet sample Netflow. The entire unicast packets forwarding decision is done on the ingress card, whereas for the multicast packet only egress replicated multicast forwarding is supported; so multicast packets are replicated on ports on the egress line cards. There is no ingress replication mode supported on the Nexus 7010 Series switch.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Fabric Scalable and Performance and Fabric Capacity

This topic describes the scalable fabric capacity and scalable Nexus 7010 Series switch performance, throughput, and the associated redundancy options.



The per-slot bandwidth capacity increases as you add each fabric. One fabric module installed provides 46 Gb/s of fabric bandwidth between the modules. Adding a second module provides 92 Gb/s. Adding a third module provides 138 Gb/s. Adding the fourth module provides 184 Gb/s. Adding the fifth module provides 230 Gb/s. Even though the module has only 40 Gb/s (or 80 Gb/s channel to the fabric), it has a crossbar on the back of the card which will load balance the traffic across all the five fabric modules. Per-slot bandwidth capacity increases with each fabric module installed.

In case one module in described configuration fails, the fabric capacity will decrease from 230 Gb/s to 184 Gb/s. If the second module fails, only 3 modules are operational and the remaining fabric capacity would be 138 Gb/s.

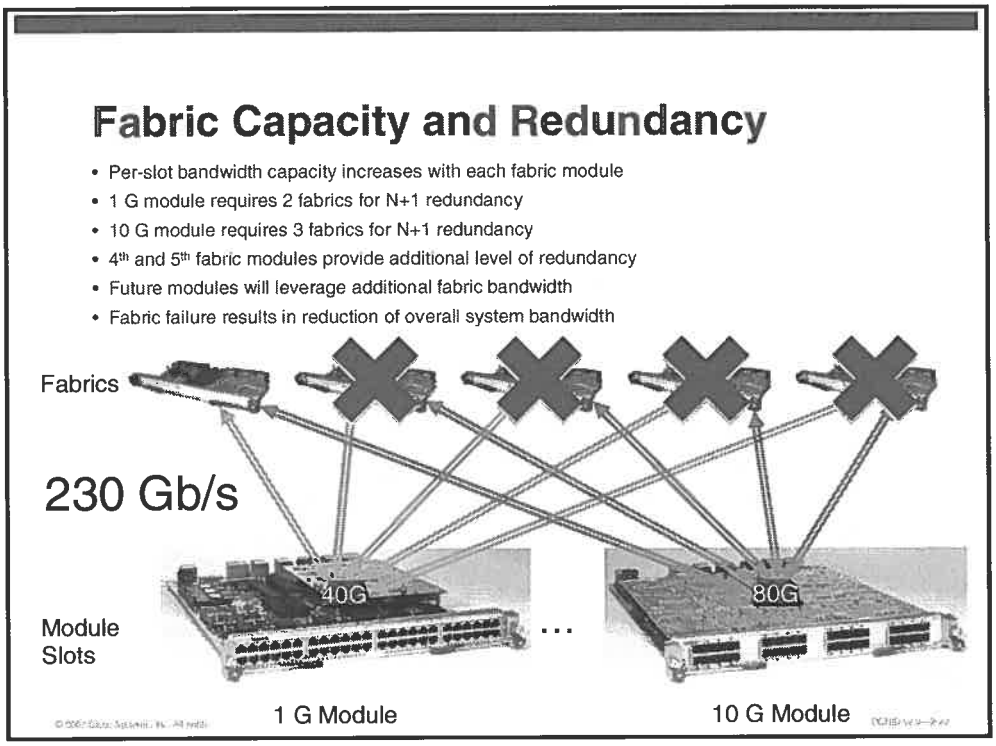
1 Gb/s I/O module requires at least 2 fabric modules installed for N+1 redundancy.

10 Gb/s I/O module requires 3 fabric modules installed for N+1 redundancy.

The 4th and 5th fabric modules provide additional levels of redundancy. Future modules will leverage additional fabric bandwidth. Fabric failure results in reduction of overall system bandwidth.

© 2008 Cisco Systems, Inc.

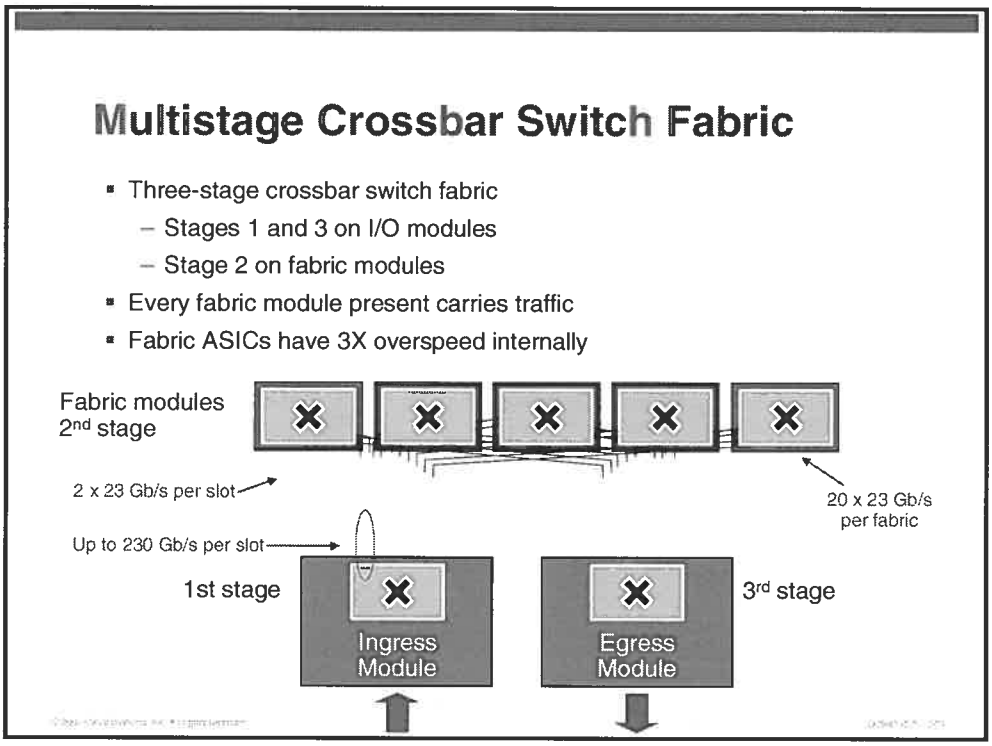
https://t.me/learningnets



The system is implemented as a multi-stage crossbar switched fabric. It is a three stage crossbar, with stages 1 and 3 on the ingress and egress modules and stage 2 on the fabric module. If fabric modules are failing but at least one is still running, the fabric will still be operational. When all fabric modules fail, traffic cannot be transferred across the fabric.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



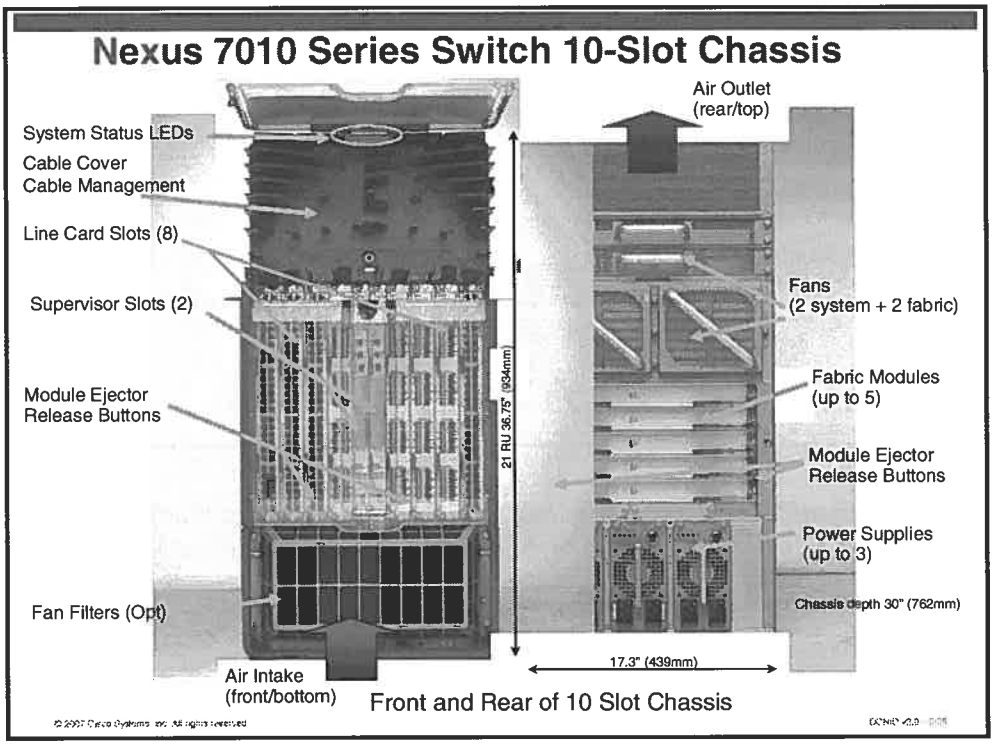
Every fabric module shown in this figure actively carries traffic. Traffic is being sent across all the available fabric modules at all times. Each fabric module internally has three times overspeed inside the ASIC.

Each fabric ASIC has 20 ports of 23 Gb/s per fabric and each line card connects to the backplane with 10 23 Gb/s connections, two to each of the fabric modules. Each fabric module part has 223 Gb/s connections to each of the line card slots except for the supervisor modules, where there is a 123 Gb/s connection.

https://t.me/learningnets

Cisco Nexus 7010 Switch Power Supply, Cooling System, and Cable Management

This topic describes the characteristics of power supplies, redundancy modes, fans, airflow, cooling, cabling, and racks solutions supported by the Nexus 7010 Series switch for optimized and efficient data center resource utilization.

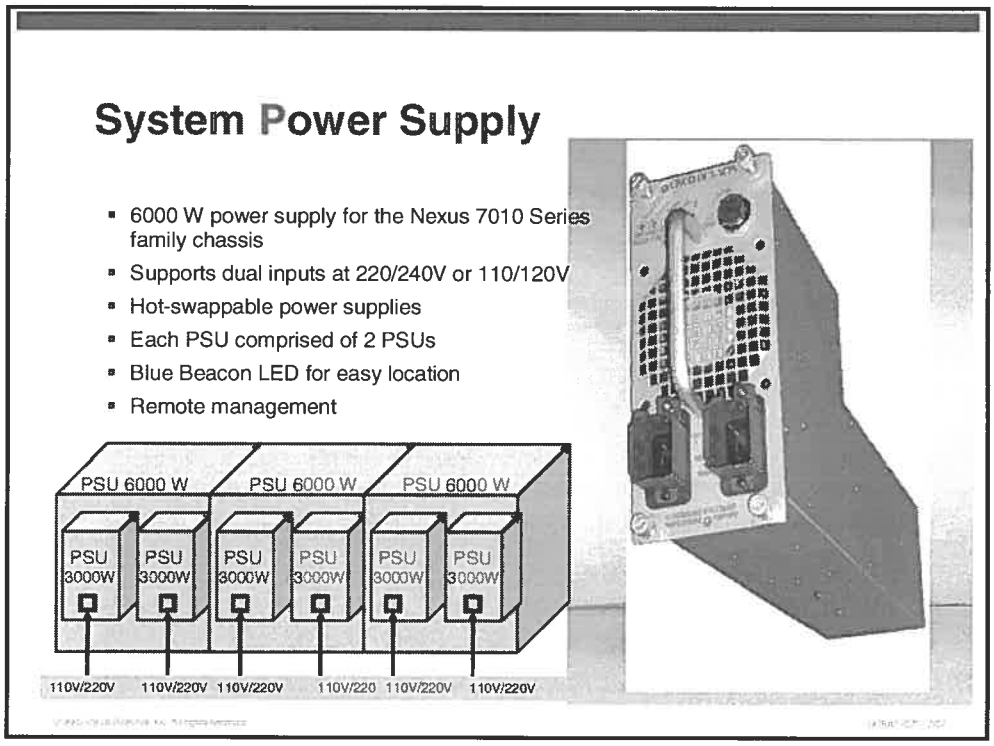


The air enters the chassis at the bottom. The baffle in the bottom splits the air between the fabric at the back and the line cards at the front. The air comes out of the back of the switch chassis.

There are four different fan trays in the Nexus 7010 Series switch. There are two system fan trays used for cooling the front line cards and there are two fabric fan trays located underneath the chassis. Five fabric modules slots are located at the back of the Nexus 7010 Series switch. Every module has ejector buttons. There is an optional fan filter and optional doors for simplified and safer cable management. The two supervisor slots are located in the middle of the chassis. The supervisor slots are different from the other slots; a line card cannot be installed in the supervisor slots. There are three power supplies at the back of the chassis, 6,000 Watts each. Each power supply comprises two 3,000 Watt power supplies. The Nexus 7010 Series switch is the most power efficient switch in the industry in terms of Watts/Tb provided.

© 2008 Cisco Systems, Inc.

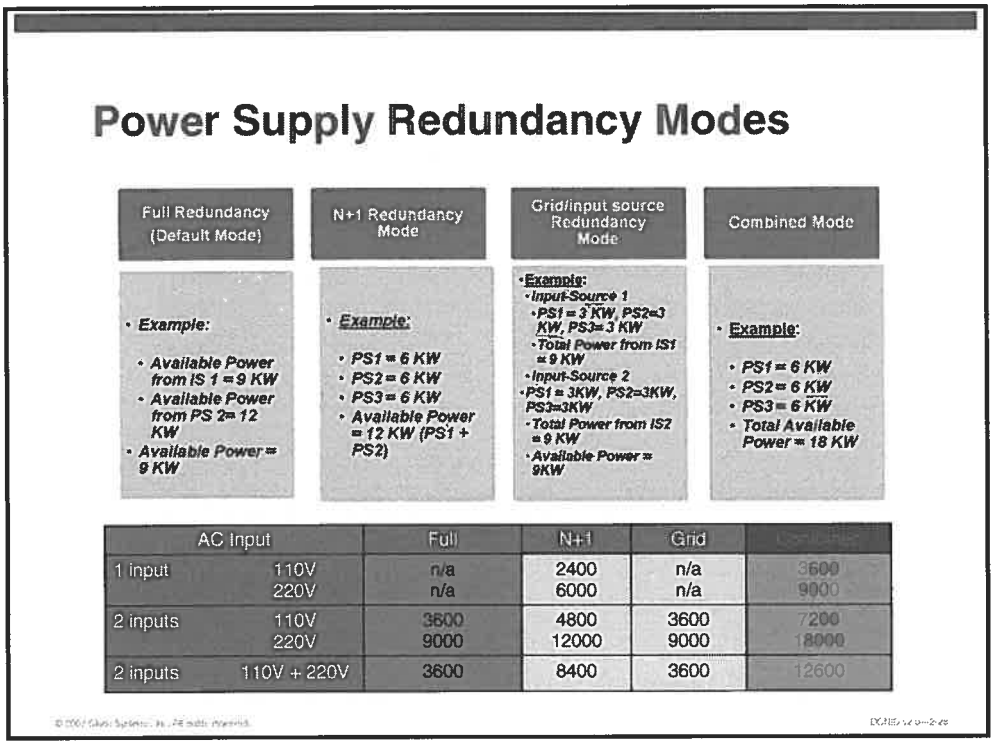
https://t.me/learningnets



Each Nexus 7010 Series switch power supply unit is composed of two internalized isolated power units, giving it effectively two power paths per modular power supply and six paths in total per chassis when fully populated. The system supports the hot-swappable 6000 W redundant power supply with two 20 Amp feeds, each feeding the two 3000 W power supplies installed within the 6000 W power supply. The power supply is flexible to provision circuits with either high-input voltage 200-240V or low-input voltage 100-120V depending on availability and power output needs. The power supply also has the capability to connect one input with a 220V line and the other input with a 110V line. Hot-swappable power supplies enable continuous system operations with no downtime when replacing another power supply (assuming that the remaining power supplies can provide enough power to support the system). The power supplies use a proportional load-sharing method for power distribution to power system components, allowing the efficient use of dissimilar capacity power supplies in the same chassis. Therefore all installed power supplies are active and share the overall load. Additionally, the power subsystem allows the three power supplies to be configured in any one of four redundancy modes. The power supply supports dual universal inputs at the 220/240V or at 110/120V with 50-60 Hz and 20 Amps circuit. Each input (half-module) can connect to a different AC source, 110 or 220V, providing redundancy within the power supply unit; if one input fails, power is still drawn from the other input. The power supply also supports fans at variable speeds for higher cooling redundancy. There is an internal fault monitoring component built into the power supply module, and five LEDs, including the blue Beacon LED, for power supply identification. The monitoring component detects the short circuits and component failures within the power supply unit; if a failure is found, the unit is shut down. Users can remotely power cycle one or all power supplies using the supervisor command-line interface (CLI), enabling remote management and improving operational efficiency (not available at initial software release.)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The 6 kW power supplies support 4 different power redundancy modes: full redundancy, N+1 redundancy and grid redundancy and combined redundancy mode. Essentially, the system needs 9 kW to run fully populated with 32-port cards. The 48-port GE cards require less power. Each of the power supply redundancy modes imposes different power budgeting and allocation models, which in turn deliver varying usable power yields and capacities.

The table on the slide summarizes various Watts of power provided by the Nexus 7010 Series switch with different types of input.

The highest number of PSUs can fail using default redundant mode without impacting the Nexus 7010 switch operation.

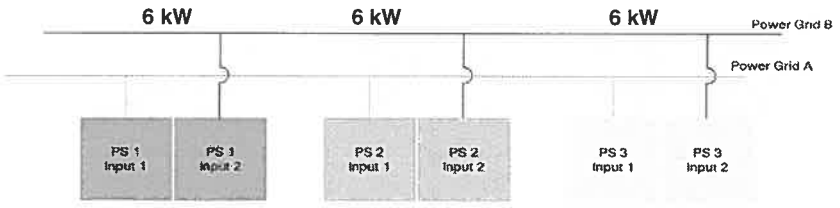
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Combined Mode

- No PSU redundancy
- Offers maximum power
- Power available to the system is the sum of the power outputs of all power supplies in the chassis

Example:
PS1 = 6 kW
PS2 = 6 kW
PS3 = 6 kW
Total Available Power = 18 kW



Nexus 7010 Series switch power supplies combined mode is not redundant power supply mode. In this mode, there is no power supply redundancy. All available power across installed power supplies is combined to provide the sum of all available power to the usable power budget. In this configuration, a failure of a power source or power supply will affect the overall available power to the system. If the current power draw exceeds the postfailure power budget, the system will enter a failure state.

In combined mode the total power available to the system is sum of the power outputs of individual power supplies. In the example shown here, there are three power supplies, all connected to 220 V circuits, each producing 6 kW. If the system is configured in combined mode, the total available power is 18 kW.

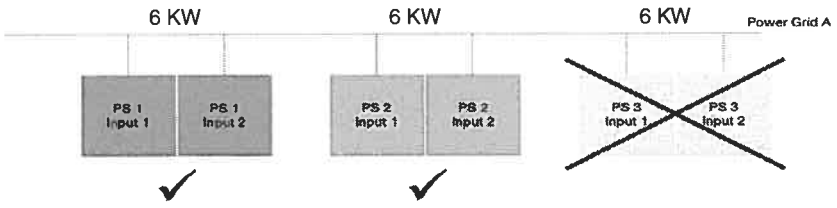
https://t.me/learningnets

Power Supply (N+1) Mode

- Guards against failure of one of the 3 PSUs
- 3 PSUs required for this mode
- Power available to the system is the sum of the two least-rated PSUs

Example:

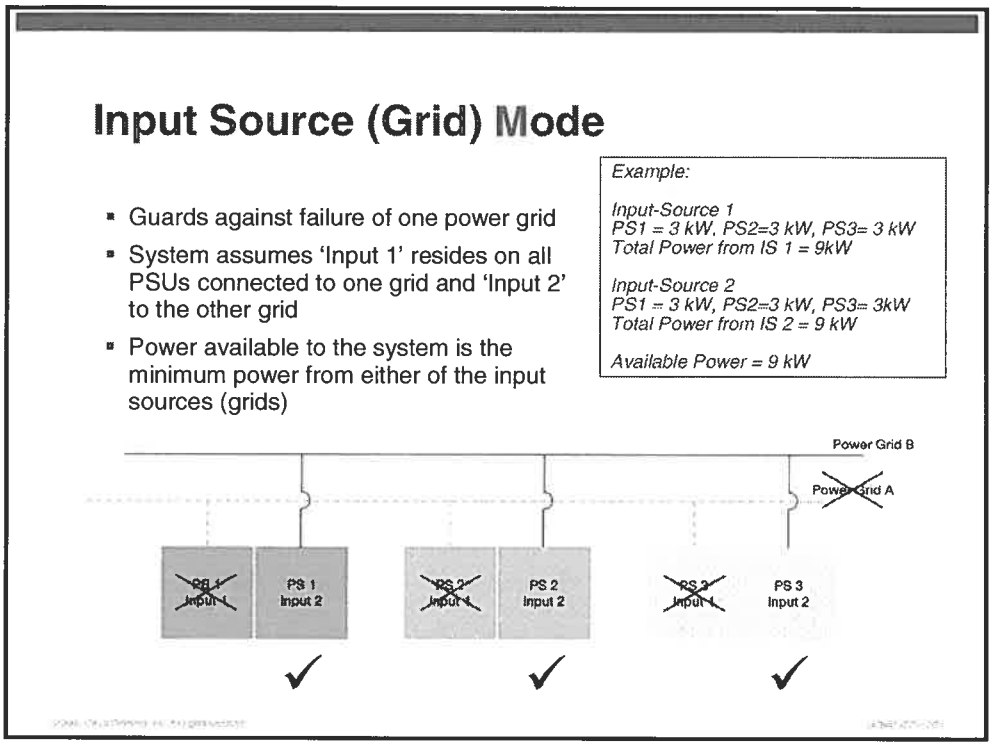
PS1 = 6 KW
PS2 = 6 KW
PS3 = 6 KW
Available Power =12 kW (PS1 + PS2)



In the N+1 PSU redundancy mode, protection against a single power supply unit failure is provided. In the event of a single power supply failure, loads are redistributed using available capacity across remaining functional PSUs. As a result of using this mode, the total available power to the chassis is the sum of all installed power supplies minus that of the largest (for redundancy). A minimum of three installed power supplies is required in order to support N+1 redundancy. If this mode is configured then all three PSUs must be installed in the 10-slot chassis. If only two PSUs are installed the system interprets it as if a failure of PSU has already occurred, and the N+1 mode will no longer be available. The available power in the N+1 mode is the sum of the power outputs of two least rated power supplies. In the example on the slide with three power supplies, each giving 6 kW, the available power is minimum 12 kW. If one of the power supplies fails, the available power remains the same 12 kW. The remaining two power supplies still provide 12 kW and the system remains intact with no line card powered down.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

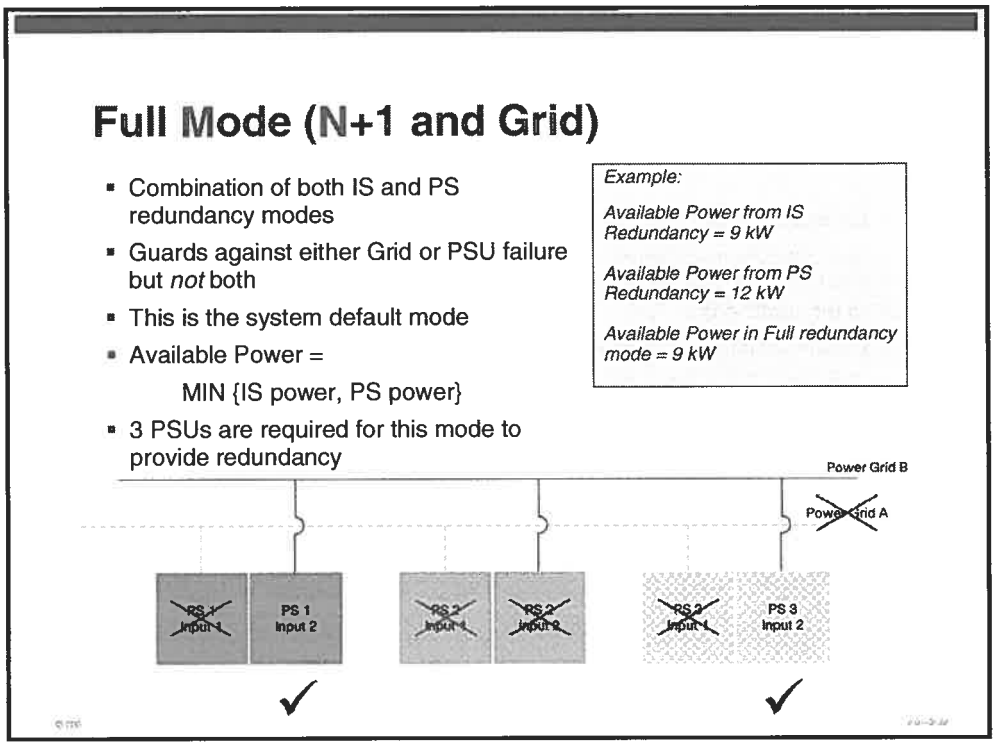


The input grid redundancy mode provides protection against input source failures. In order to implement input grid redundancy, each power supply unit must be connected to two independent power sources (grids). A single power grid failure will allow all installed power supplies to continue functioning through the remaining internalized half still connected to the remaining available power grid. Configuring the system in this mode helps guard against failure of one of the two grids. This mode is appropriate for customers with two or more grids in the data centers that need fault tolerance. If two grids are not connected as represented by the figure, then during fail conditions, the system may end up losing some of the modules.

Because each half of the PSU provides 3 kW of output, the entire grid 1 provides 9 kW and grid 2 provides another 9 kW. The available power in this mode is a minimum of the power provided by the two grids, which is 9 kW. In case of a grid failure (for example grid 2 fails), the remaining grid takes over, providing 9kW of power with the system operation remaining intact.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Full redundancy mode is the default Nexus 7010 Series switch power supply mode. This mode is a combination of N+1 mode and grid power supply redundancy mode. In full redundancy mode the power supply configuration provides protection against a single power source grid failure or a single power supply unit failure. This mode guards against failure of either a grid or a power supply, but not both. The power available in this mode is the minimum power provided by the available two modes: by the input source power supply mode and by N+1 redundancy mode.

If the input source redundancy is configured, the available power would have been 9 kW. With configured N+1 power supply redundancy mode the available power would have been 12 kW. In full redundancy modes, the available power would be 9 kW, which is a minimum of the two. In case of a grid failure 9 kW are still available from grid 2 or in case of a power supply failure, the remaining two power supplies can provide 9 kW. The Nexus 7010 Series switch system will be intact, and no modules will be denied power in this scenario.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Nexus 7010 Series Switch Power Allocation and Efficiency

- System reserves the power for fans, the two supervisors, and for 5 fabric modules while booting
- Cisco Nexus
  - 6000 W PSU is 90.3% efficient
- Higher Efficiency → Lower Power Draw → Lower Heat Dissipation → Lower Cooling Requirements
- Increased PSU efficency reduces TCO
- The power per port:
  - 10 Gigabit Ethernet port of 32-port 10 Gigabit Ethernet module is only 24 W

Component	Power
System & Fans	2184 W
Supervisor (each)	400 W
Fabric (each)	60 W
32-port 10GE (4:1)	750 W
48-port 10/100/1000	400 W

Power Supply Efficiency	Power Draw (for 6 kW output)	Heat Dissipated (BTU)
80%	7500 W	25590
90%	6667 W	22747 (11.1 % less)

The Nexus 7010 software reserves certain power for fans, for both supervisor modules and for 5 fabric modules when it detects hardware components. Even though the system has only three fabric modules installed, the power is reserved for all five, so that another fabric module can be inserted without affecting any existing modules. If the power is insufficient, then the bring-up process is halted. No more I/O modules will be brought up. If the system has sufficient power, then fabric modules come up next. Once the fabric modules are up, then the system determines based on the available power the remaining power and how many I/O modules it can bring up. The system will then bring up all those I/O modules in parallel. The prioritization of powering I/O modules is based on the slot number for the module (from Slot 1 to Slot 10).

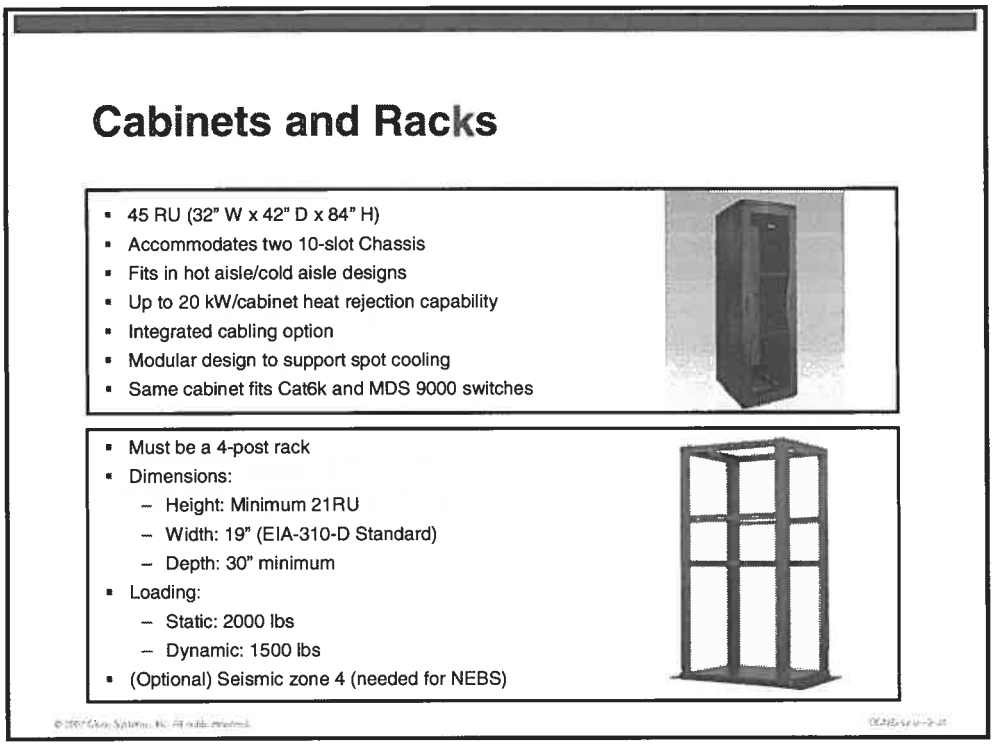
The table in this slide shows the amount of power that Nexus 7010 Series switch allocates to certain hardware modules and components when they are detected. These numbers are based on the absolute maximum power consumption of the components: at peak loads and at extreme—or the highest—temperatures within the acceptable ranges. The actual power consumption would be significantly less than the numbers reported in the table. These numbers are showing how much power the software reserves upon detection of these components. Based on these numbers, the maximum system power draw is 9284 W.

The power efficiency of the 6 kW power supply is 90%, which is among the highest in the industry so far. The PSU efficiency varies with output load. At half-loads, which are around 3000 W, the efficiency of the power supply is approximately 89%. At the load of 4000 W, the efficiency reaches 90%. As the load increases to 6000 W, the efficiency increases to 90.3%. The 6000 W power supply with 90% efficiency delivers at maximum 6667 W of power to the system, whereas an 80% efficient power supply delivers 7500 W.

The higher the power supply efficiency, the lower the power draw would be, which results in lower heat dissipation. And the lower the heat dissipation, the lower the cooling requirement in the data centers would be. Therefore, customers with 90% efficient power supplies can save money both on lower power draw and on lower cooling requirements.

The power per 10 Gigabit Ethernet port on the 24-port I/O module is approximately 24 W. Cisco has a continued commitment to reduce the power per port from green data center perspective. Cisco was able to achieve this without compromising any operational features of the Nexus 7010 Series switch.

https://t.me/learningnets



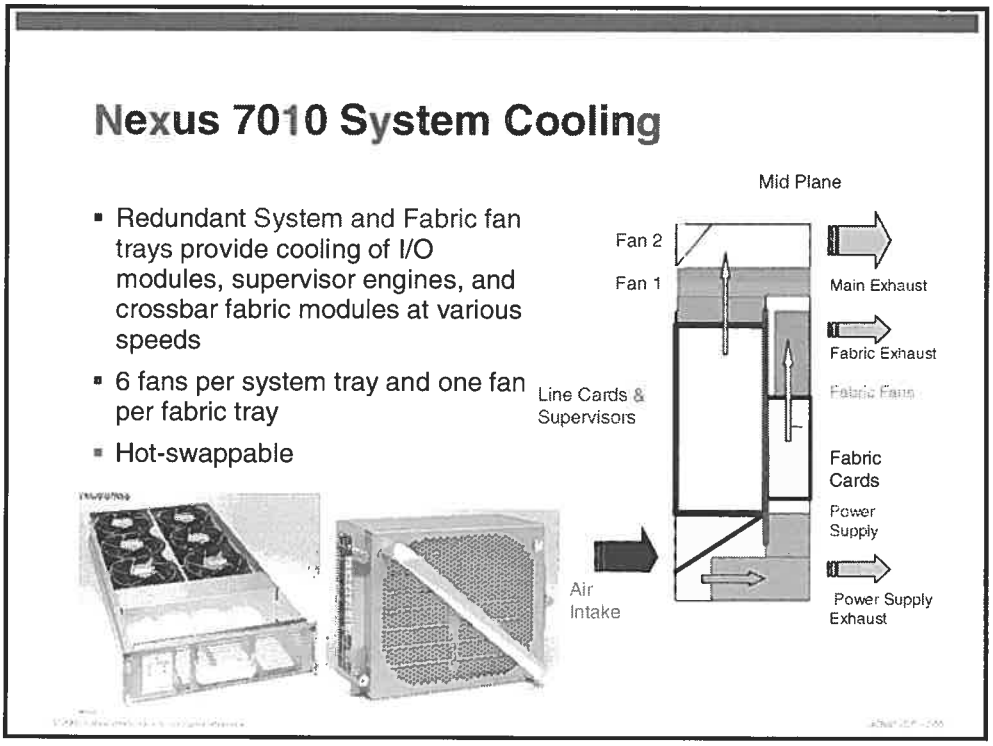
Cisco has verified the cabinets of two vendors, Panduit and Chatsworth are appropriate for the Nexus 7010 Series switches. The first cabinet is 45RU high and 32 inches wide so that there is extra room for the cables on either side of the cabinet. These cabinets can accommodate two 10-slot Nexus 7010 Series switch chassis. The cabinets have an integrated cabling option and some spot cooling mechanisms (optional). If the customer prefers other cabinet vendors they can use them as long as they support the right dimensions to fit the Nexus 7010 Series switch in the rack.

Customers who do not use cabinets can use the 4-post rack solution. The dimensions are listed on the slide. The rack's static and dynamic loading is about 2000 and 1500 pounds respectively.

When ordering the racks and cabinets remember that the Nexus 7010 Series switch is about 30 inches deep. Adding the cable management and handles for the power supply at the back of the cabinet increases the depth to 38 inches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Nexus 7010 Series chassis is designed for front-to-back air-flow which is ideal for hot aisle and cold aisle data center design. There are 2 system fan trays, 2 fabric fan-trays and PSUs with built-in fans. The air enters the chassis on the bottom front side and exits on the back of the chassis. All hardware components, such as the I/O modules, the supervisor modules, the fabrics, and the power supplies, all take air from the front and exhaust at the back. This air flow makes the Nexus 7010 Series chassis an ideal solution for data center hot aisle and cold aisle design. The airflow in the chassis is connected by two system fan trays that cool the supervisors, the line cards, and the two fabric fan trays that cool the fabrics. The power supply has its own fan. The installation of all these components is done at the back of the chassis. Both fan trays are fully redundant and hot-swappable. All fan trays must be present and installed for proper system operation. If a fan tray is removed it must be replaced within 3 minutes or the system will start the shutdown procedure. In case both fan trays of a given group are removed, they must be replaced within 2 minutes or system will be shutdown. System cooling is available at variable speed fans for redundancy reasons. When the fans are running full speed they produce a lot of noise. Each system fan tray has 6 fans and each fabric fan tray has 1 fan. All fans are variable speed 256 speed levels.

Fan speed is a function that depends on the following parameters:

- Measured inlet temperature on supervisor and fabric modules
- Programmed cooling requirement of modules
- Presence of air filter (air filter reduces air-flow and hence fan speed increases)
- Temperature alarms

All the fans support 256 speed levels. If any of these four parameters change, the fan speed adjusts accordingly to provide sufficient cooling to the system.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

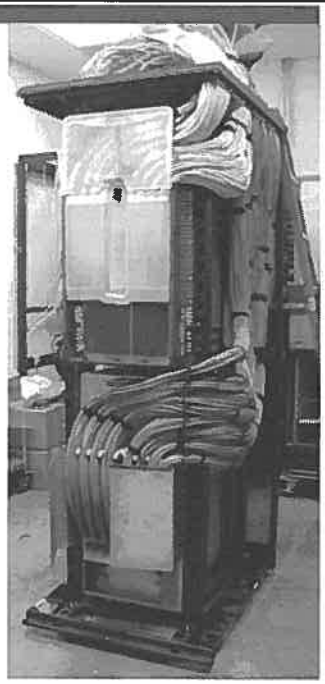
All the fan trays are fully redundant and hot swappable; fully redundant means one system fan tray can cool all the supervisors and I/O modules in the chassis, but the second fan tray still needs to be installed and present in the system for it to work. The same is true for the fabric fan trays. If a fan tray is removed, it has to be replaced within three minutes. If there is no second fan tray, having only two fan trays will create a hole in the chassis, and the fan tray that is pulling the air from the modules will not be able to exert enough pressure and will not be able to cool the system.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Nexus 7010 Series Switch Cable Management

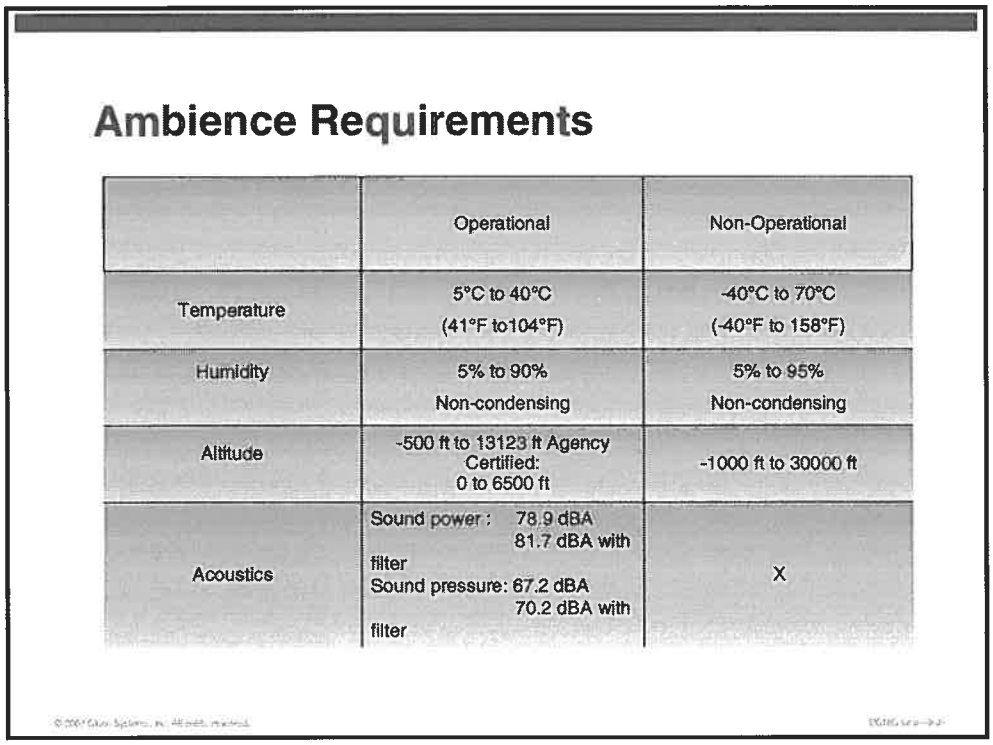
- Integrated cable management tray with straps
- Cable grooming to right, left, or split
- Can route up to 384 Cat6A cables to one side of chassis–worst-case scenario is plus 8 Cat5 cables (from Sup)
- Designed to allow all cables to route to one side or spread left and right



Cable management comes with the Nexus 7010 Series 10-slot chassis. The cable tray is located on the front side of the chassis above the line cards and supervisor modules. The cable tray is designed in such a way that all the cables can be routed to one side of the chassis or split into two, then pulled up to the top of the rack. The cable management tray supports at maximum 384 Cat6A cables. Customers are also free to deploy their own cable management solutions whenever needed, like fingers on the racks, for example.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The ambience requirements for the Nexus 7010 Series switch, in both operational and non-operational modes, are shown in the table.

The noise in the Nexus 7010 Series switch is the caused by the fans spinning and the gush of the air through the chassis. Maintaining a good ambient temperature in the switch is a key to ensuring the noise levels are low. Therefore, the higher the ambient temperature, the higher the fan speed, which results in higher noise. "Sound power" and "sound pressure" are two distinct and commonly confused characteristics of sound. Both share the same unit of measure, the decibel (dB), and the term "sound level" is commonly substituted for each.

Sound Power

Sound power is the acoustical energy emitted by the sound source, and is an absolute value. Sound power is not affected by the environment.

Sound Pressure

Sound pressure is a pressure disturbance in the atmosphere whose intensity is influenced not only by the strength of the source, but also by the surroundings and the distance from the source to the receiver.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

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

Summary

- Nexus 7010 Series Switch design was designed to meet data center network requirements
- Nexus has 2 redundant supervisor modules and 5 fabric modules, and supports two different I/O modules (48-port 1 Gb/s I/O module and 24-port 10 Gb/s I/O module)
- Almost every component is made redundant and hot-swappable (Fans, power supply, CMP, etc..)
- Consolidation and virtualization on Nexus 7010 Series Switch reduces power consumption in data centers
- Redundant supervisor modules have no data plane forwarding functions, they purely provide control plane functions with data plane functions offloaded to the I/O modules

Summary (Cont.)

- Fabric load balances unicast traffic across all available fabric modules for better performance and redundancy
- Redundant supervisor modules support seamless failover
- Nexus 7010 Series Switch has scalable performance and throughput: From 1 to 5 fabric modules can be used for increasing switch capacity and redundancy
- Nexus 7010 Series Switch is based on non-blocking forwarding architecture to prevent head-of-line blocking effect
- Cable management optimzed to support the maximum cabling density with optional doors to prevent accidental cable damage

https://t.me/learningnets

Summary (Cont.)

- Nexus 7010 chassis supports Blue Locator IDs–for operator control to identify every chassis component and module
- Nexus 7010 chassis uses front-to-back airflow
- Nexus 7010 Series Switch supports rear removal of fans, PSU, and fabrics for ease of maintenance
- Nexus 7010 Series Switch supports standalone separated Connectivity Management Processor (CMP) for redunant and highly availaible out-of-band management of the Nexus 7010 Series Switch
- Nexus 7010 10-slot chassis offers many innovations in power and cooling

Summary (Cont.)

- 32-port 10 Gigabit Ethernet I/O module reduces power per 10 Gigabit Ethernet port
- Nexus provides industry-leading power supply efficiency
- Power supply has redundant fans rotating at variable speeds–optimizing power draw
- Nexus 7010 has built-in power supply innovations:
  - 4 redundancy modes
  - Temperature sensors and fault monitoring
  - Remote management
- Parallel bring-up of I/O modules
- Nexus 7010 Series Switch offers Integrated Cable Management

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 2

Cisco Nexus 7010 Software Architecture

Overview

This lesson describes the characteristics of the Nexus 7010 Series Switch NX-OS software components architecture.

Objectives

Upon completing this lesson, you will be able to describe the components and operation of the Cisco Nexus 7010 Switch modular and resilient software architecture.

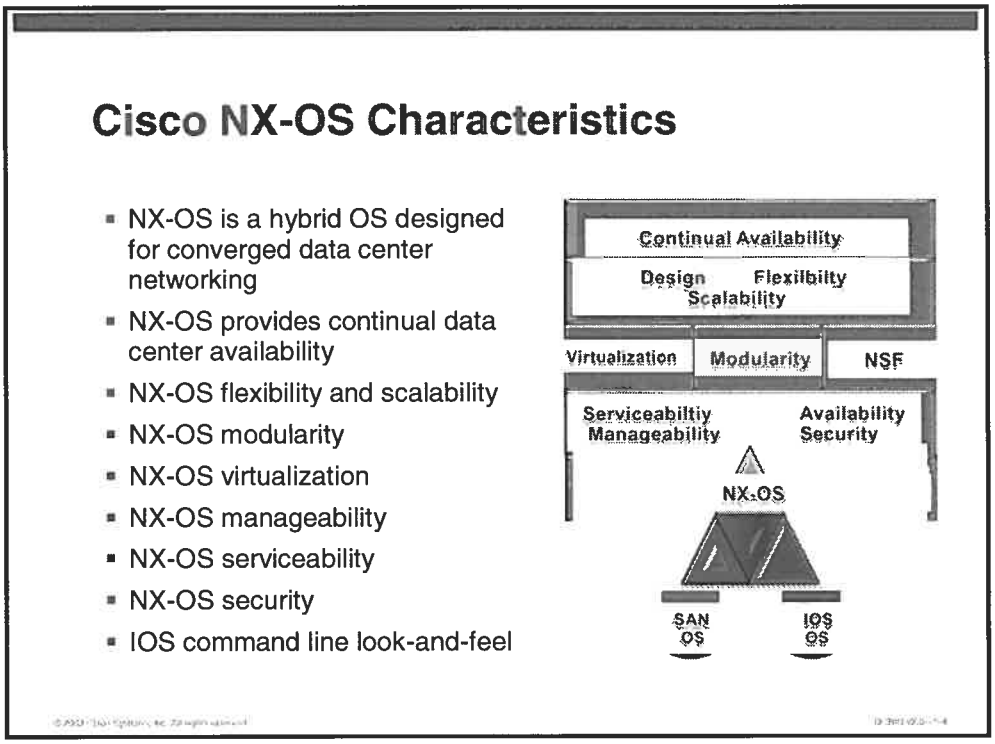
This ability includes being able to meet these objectives:

- Describe the characteristics of Cisco Nexus 7010 Series Switch Software Architecture
- Describe the Cisco Nexus 7010 Series Switch Licensing Models

https://t.me/learningnets

Cisco Nexus 7010 NX-OS Software Architecture Design

This topic describes the Cisco Nexus 7010 Series Switch modular and resilient software architecture with rich-feature set based on Linux kernel. It describes the NX-OS origins, the virtualization features, the Layer 2, Layer 3 and upper layer protocols supported with the focus on providing converged data center functions, service availability and on data center nonstop operation.



Other NX-OS architecture design goals were to provide a granular and modular software component, to protect services and applications within the platform, to provide a degree of fault isolation and for scalability and high-availability.

NX-OS is common software throughout the data center: Cisco NX-OS simplifies the data center operating environment and provides a unified OS designed to run all areas of the data center network including the LAN, SAN and L4-7 network services.

NX-OS Modularity

Modular Cisco NX-OS software design supports distributed multithreaded processing on symmetric multiprocessors (SMPs), multi-core CPUs, and distributed line card processors. Computationally intensive tasks, such as hardware table programming, can be offloaded to dedicated processors distributed across the line cards. Cisco NX-OS modular processes are instantiated on demand each in a separate protected memory space. Thus, processes are started and system resources allocated only when a feature is enabled. The modular processes are governed by a real-time preemptive scheduler that helps ensure the timely processing of critical functions.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

One of the benefits of designing a modular system is the resulting component-based system can easily support granular in-service software upgrades (ISSU) and granular patching. This granularity enables system and application upgrades while the platform is operational. You can, for example, upgrade specific modules or plug-ins as needed.

NX-OS Availability

Continual system operation provided by the Cisco NX-OS includes permitting maintenance, upgrades, and software certification without service interruptions. The combination of process modularity, modular patching, Cisco In Service Software Upgrade (ISSU) capability, and nonstop-forwarding (NSF) graceful restart minimizes the effects of software upgrades and other operations.

NX-OS processes are stable and restartable. Critical processes run in protected memory space and independently from each other and the kernel, providing granular service isolation and fault containment and enabling modular patching and upgrading, and rapid restartability. Individual processes can be restarted independently without loss of state information and without affecting data forwarding, so that after an upgrade or failure, processes restart in milliseconds without negatively affecting adjacent devices or services. Processes with large amounts of state such as IP routing protocols are restarted using standards-based NSF graceful restart mechanisms; other processes use a local persistent storage service (PSS) to maintain their state.

Redundant supervisors support stateful failover upon supervisor failure. Interprocess communication is reliable to help ensure that all messages are delivered and properly acted on during failure and adverse conditions. This communication helps ensure process synchronization and state consistency across processes that may be instantiated on processors distributed over multiple supervisors and I/O modules.

Switched Ethernet out-of-band channels (EOBCs) are redundant. Cisco NX-OS can make full use of redundant EOBCs for communication between control and I/O module processors.

Network convergence or network availability is optimized by providing tools and functions to make both failover and fallback transparent and fast. For example, Cisco NX-OS provides Spanning Tree Protocol (STP) enhancements such as bridge protocol data unit (BPDU) guard, loop guard, root guard, BPDU filters, and bridge assurance to guarantee the health of the STP control plane; Unidirectional Link Detection (UDLD) Protocol; NSF graceful restart of routing protocols; millisecond timers for First-Hop Resiliency Protocols (FHRP); SPF optimizations such as link-state advertisement (LSA) pacing and incremental SPF; and IEEE 802.3ad link aggregation with adjustable timers

NX-OS Serviceability

Cisco NX-OS is built with unique serviceability functions to enable network operators to take early action based on network trends and events, enhancing network planning and improving network operations center (NOC) and vendor response times. Call Home, Cisco Generic Online Diagnostics (GOLD), and Cisco NX-OS Embedded Event Manager (EEM) are some of the features that enhance the serviceability of Cisco NX-OS, Netflow analyzer, packet analyzer, SPAN and others.

https://t.me/learningnets

NX-OS Manageability

The Data Center Network Manager (DCNM) GUI is designed for the management of the Cisco NX-OS product family based on XML-API. Cisco Nexus supports a redundant standalone Connectivity Management Processor (CMP) for lights-out remote management of the platform. The CMP is instrumental in aiding operations by providing an out-of-band access channel to the NX-OS console.

NX-OS Security

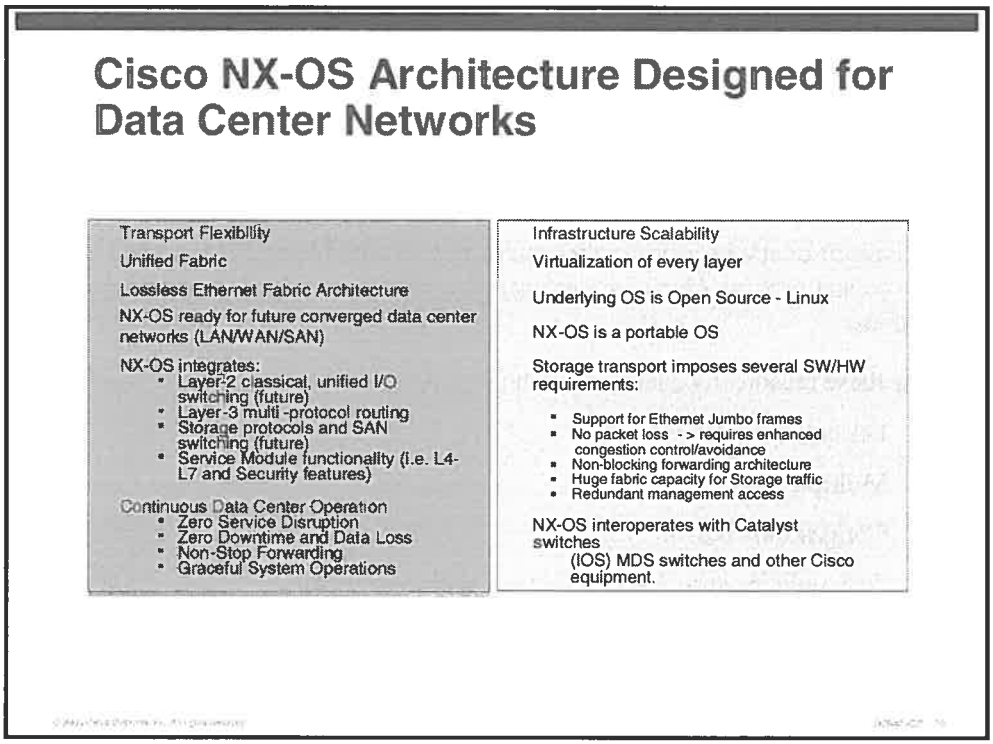
Cisco NX-OS supports comprehensive scalable data center security implementation. In addition to all Catalyst 6500-specific security features, Cisco Nexus supports the Cisco TrustSec suite that provides outstanding data confidentiality and integrity, supporting standard IEEE 802.1AE link-layer hop-by-hop encryption- cryptography with 128-bit Advanced Encryption Standard (AES) cryptography. Link-layer cryptography guarantees end-to-end data privacy while allowing the insertion of security service devices along the encrypted path. Security group access control lists (SGACLs), a new paradigm in network access control, are based on security group tags instead of IP addresses, enabling policies that are more concise and easier to manage due to their topology independence.

Another NX-OS security design goal was the integration of component virtualization. NX-OS supports virtualization at several levels: NX-OS virtualization (Virtual Device Context), Virtual Router Family, Virtual interfaces, virtualization of L2 and L3 processes, etc. Virtualization enables consolidation of different types of data centers and technologies on the common converged hardware and software architecture, simplifies the management of data center networks and reduces the cost of data center deployments due to design flexibility. Virtualization also enables NX-OS high-availability, fault-isolation, enhanced security and better resource utilization.

NX-OS preserves the existing Cisco IOS command line look-and-feel so that minimal effort and investment is needed for the engineers to manage the NX-OS. Nexus 7010 Series Switch management is based on open API using XML-based management database. NX-OS has very scalable control plane processing using symmetrical multiprocessing-based CPUs. NX-OS uses multiple CPUs and multi-core platforms to scale the control plane processing.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Nexus 7000 Series Operating System NX-OS is a modern operating system built with special focus on the software infrastructure and features required for multi-protocol mission critical data center networks. NX-OS is called a multi-purpose OS because it is designed from the ground up to be highly available and highly modular with a focus on integrating these specific data center networking technologies: Data networking (LAN/WAN/MAN), Storage networking (Fibre Channel, iSCSI, FCIP) and Clustering protocols. NX-OS software architecture is based on software component modularity and high-availability enabling zero downtime, zero data loss and nonstop forwarding.

The NX-OS goal was to integrate data center Networking technologies such as Layer 3 routing, Layer 2 switching and bridging technologies, and SAN switching and service module functionalities (such as L4 - L7 and security features). Multi-chassis clustering is not available on NX-OS at this time. The design goals were to provide specific feature sets to the data center networking architecture.

https://t.me/learningnets

This goal imposed several new software and hardware requirements on Nexus 7010 Series architecture:

1. Nexus was designed to support distributed forwarding for Ethernet/SAN traffic.

   It must support all Ethernet frame sizes (64B – Jumbo) Jumbo frames are ~9 KB frames to be able to accommodate future Fibre Channel frame in one single Jumbo frame.

2. In case of oversubscription, there must not be any unintelligent packet drops anywhere in the packet processing chain, because the storage protocols are not supposed to lose any frames.

   For these reasons mechanisms such as:

   - Decent size buffers
   - Multiple queues
   - Congestion control
   - QoS policies associated with each packet or frame
   - Must be supported by NX-OS and hardware.

3. Nexus 7010 Series switch must enable non-blocking forwarding architecture:

   - During oversubscription, traffic from all impacted ports must be treated fairly
   - At no point in the system is traffic from some ports allowed while traffic from other ports is blocked
   - Head-of-line (HOL) blocking must not occur in the system fabric
   - Congestion of one egress linecard should not impact fabric traffic going to other line cards.

4. Nexus must support Multicast/Broadcast over Fabric:

   - Nexus 7010 Series switches must support efficient transport of multicast/broadcast traffic
   - No performance impact when carrying multicast traffic (L2 and L3) over the fabric
   - The same must be supported for L2 traffic flooding (unknown unicast floods, multicast and broadcast).

5. Redundant Management Access

   Connectivity Management Processor (CMP) provides standalone redundant management access to the Nexus 7010 Series switch. It is supported on the supervisor and enables remote reboot of the switch:

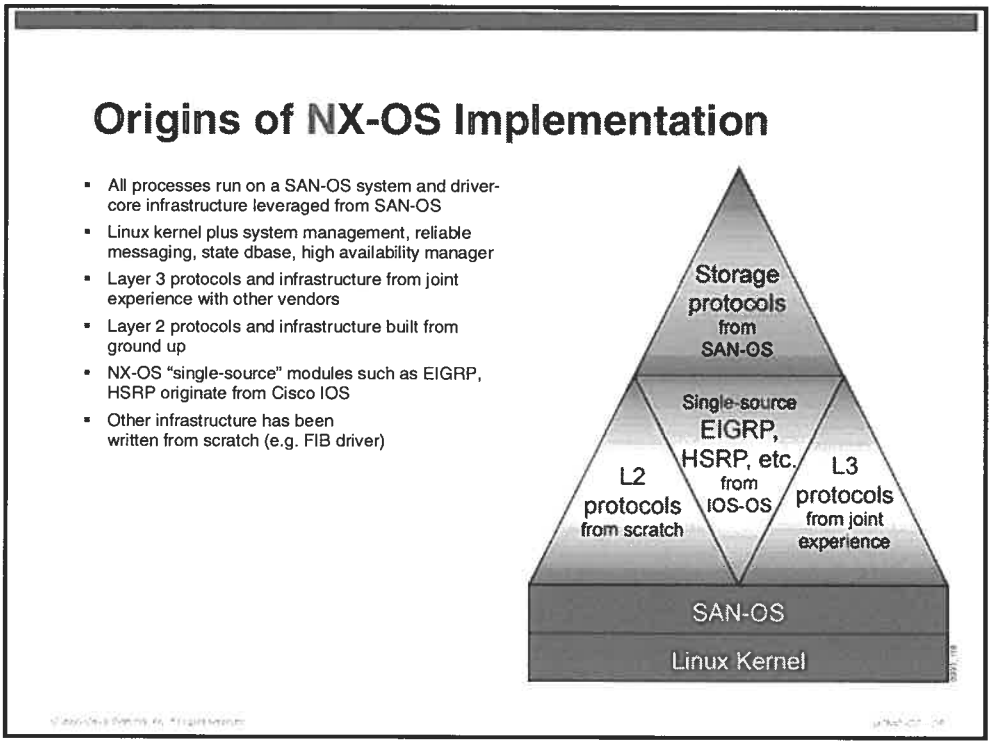
   - Without losing telnet/SSH sessions even when CP/DP (or SP/RP) CPUs are hung
   - View all boot cycle messages
   - Power supply separate from the main chassis power supply to ensure its operation during chassis power cycle (rechargeable battery backup of at least 30 minutes)

NX- OS Software Compatibility

Cisco NX-OS Release 4.0 interoperates with Cisco products running any variant of the Cisco IOS Software operating system. Cisco NX-OS Release 4.0 also interoperates with any networking OS that conforms to networking standards. NX-OS for example interoperates with Catalyst switches running Cisco IOS and MDS switches running SAN-OS.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

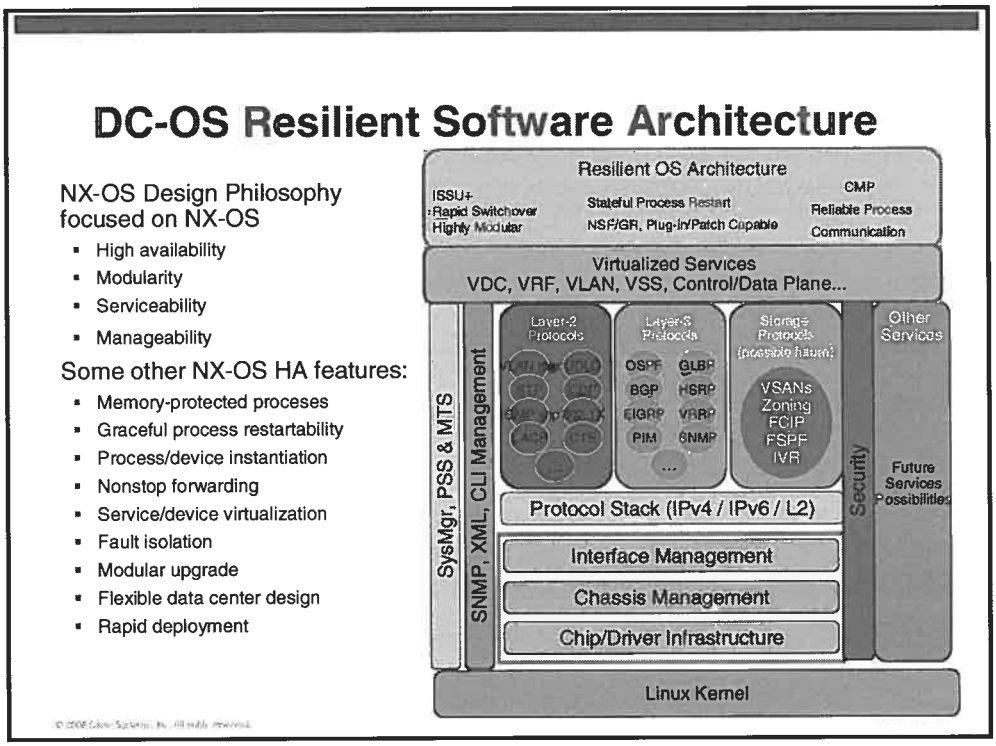


Core NX-OS infrastructure is leveraged from SAN-OS. NX-OS is based on Linux kernel, supports system management, reliable messaging, stateful database and high availability manager Layer 3 protocols. NX-OS is based on a modern, streamlined, scaleable routing implementation. Layer 2 protocols and infrastructure are built from the ground up. The Layer 2 feature set (STP, LACP, and 802.1x) is also data center focused and standards-based. Storage protocols are taken directly from SAN-OS. In the future, when Fibre Channel modules are delivered on NX-OS, storage functionality "plug-ins" will be delivered to the OS. NX-OS "single-source" modules such as EIGRP, HSRP, etc. originate from Cisco IOS. Other infrastructure has been written from scratch (for example FIB driver). The Layer 2 protocols are largely written from scratch using all the Cisco expertise and experience gained from the CatOS to the Cisco IOS implementation over time.

The NX-OS Layer 3 protocols implementation and the IP v4 and IPv6 protocol stack were written as a result of Cisco joint experience in L3 space with other vendors. Certain code application such as EIGRP, HSRP and others were ported over as applications or services within the NX-OS space. Other infrastructure, particularly drivers for hardware had to be written from scratch for the new hardware. All of the mentioned processes run on top of a SAN-OS based framework or system and driver interface, which is based on the Linux kernel and with as many components written into user space as possible to provide protection for the kernel.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The NX-OS is designed very similar to UNIX systems with protected kernel, applications and services running in a user space. The services in user space are modularized to be protected and isolated from each other. The overall architecture has underlying key common components available across modules representing the core fundamental functions across the Nexus 7010 Series Switch platform. For example: Inter-process communications management, interface management and protocol stacks are common across all of the other modules. The services within protocol space are written in a modular fashion to make use of those core components and to interface with them. The benefit to this type of design is fault isolation and protection of services and modules as well as data center scalability and high availability. Another advantage of this architecture is that it is gradually upgradeable. New functional components and new protocols can be implemented, adapted and integrated into the existing NX-OS architecture quickly and easily. Modular design allows granular upgrade and patching. Due to graceful process restart the upgrade can be accomplished without disruption on data center traffic. At the top of the architecture there are storage protocols, Layer 3 protocols and Layer 2 protocols that can be instantiated per Virtual Nexus device and upgraded or a patched based on a service basis. This is basically a next-generation architecture that tries to bring three fundamental technologies into the converged hardware and software platform (LAN, SAN, and L4-L7 networking services).

The NX-OS architecture has been focused on serviceability and modularity with built-in nonstop forwarding features because this platform has been largely designed from the ground up with data center high availability in mind. NX-OS modularity is based on protocols, table managers and different subsystems all run as distinct memory-protected restartable processes. Modularity enables high availability infrastructure which is very important especially for storage networking with zero packet loss requirements. Dual supervisor chassis is the base for nonstop forwarding of all of Storage/L2/L3 protocols. NX-OS implements standardized protocol graceful software process restart procedures. NX-OS modularity provides redundancy. The software is designed to be highly modular to provide fault isolation between services and components that operate in protected memory space. Individual services operate in their own memory space. Services can be instantiated independently.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NX-OS supports device virtualization as part of a modular design that enables deployment of multiple virtual Nexus 7010 Series Switches on the common Nexus 7010 Series hardware and NX-OS architecture. Every instance of VDC runs independently from others. There can be different routing and switching processes instantiated per every VDC on the common NX-OS. For example, there can be multiple instances of OSPF running on the common NX-OS. Each instance of OSPF belongs to different VDC and both OSPF processes are protected from each other and across Virtual Device Contexts (VDC). Virtualization services implemented on top of based services using VDCs and VRFs are delivered across all the different integrated areas such as L2, L3, and security, and operation virtualization per each VDC. The fundamental and virtualization services represent the fundamental base for high-availability of NX-OS. Virtualization and modularity enabled other high-availability NX-OS features: In-Service Software Upgrades(ISSU), stateful supervisor switchover, stateful process restart, graceful restarts for nonstop forwarding of L2, L3 and storage protocols, modular pathing, isolated memory protected processing and redundant standalone CMP management access.

https://t.me/learningnets

Cisco Nexus 7010 NX-OS Software Layer 2, Layer 3 and Routing Protocols

This topic describes the NX-OS 4.0 Layer 2, Layer 3 features and routing protocols supported.

Cisco Nexus NX-OS Features Summary

L2 Features	L3 Features	Security	Management
STP/RSTP	OSPFv2, OSPFv3	Cisco TrustSec	DCNM based on XML-API
STP Enhancements	EIGRP, IS-IS	IP ACL, MAC ACL	RBAC, Network Management Tools
Link Aggregation	RIPv2 (no GR), static	VACL, PACL, RACL	Software Licensing
UDLD	HSRP, GLBP, FHRP, VRRP	Control plane policers	Smart Call Home
CDP	GRE, ECMP, VRF-lite	Traffic storm control	SPAN/RSPAN
Port Security	BGP	Port Security	AAA, Radius, Tacacs+
802.1X	IPv4, IPv6: unicast, multicast	ARP Inspection	SSHv2, SSL, telnet, SNMP (Read-Only)
VLAN	Atomic ACLs programing	DSCP Snooping	L4-L7 Services Management
Private VLAN	Route Policy Manager	IP Source Guard	EEM, Control plane Analyzer
L2 ACL	PBR, uRPF	802.1x, NAC	GOLD, Syslog, Netflow, RMON
Ethernet Jumbo frames	L3 virtualization (VDC, VRF)	MD5, keychain mgmt	

3 licenses
Base - L2
Enterprise L3
Add: VDC
+ Trust Sec

Cisco Nexus NX-OS Features Summary (Cont.)

IP Multicast Features	QoS Features
IGMPv2, IGMPv3 (no GR)	Queuing and scheduling
PIM v6, PIM-Sparse, Bidir-PIM (no GR)	port trust, classification, policing, congestion mgt.
MSDP, MLD(no GR)	priority queuing, bandwidth policing remarking upon classification, egress shaping
	Congestion avoidance:WRED, WRR and management

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Here is a summary of supported NX-OS features:

NX-OS L2 Control Plane Features

- **Ethernet Switching (Jumbo frames)**
- **Spanning Tree Protocols**
  - 802.1W/D [(rapid) spanning tree protocol]
  - 802.1S (multi-instance spanning tree)
  - RPVST+ (Cisco rapid per-VLAN spanning tree)
  - Per-VLAN Rapid Spanning Tree, Multi-Instance Spanning Tree
    - Stateful STP process restart, supervisor switchover, and ISSU
    - Backward compatibility with 802.1D STP using fallback mode
  - Spanning Tree Protocol enables transparent upgrades using ISSU in Spanning Tree Protocol environment
- **Spanning Tree Enhancements**
  - PortFast Guards (Root guard, BPDU guard, BPDU filter)
  - The new advanced features addressing STP and Layer 2 stability are:
    - Bridge assurance
    - Dispute mechanism
  - Port VLAN ID (native VLAN ID) mismatch
  - Per VLAN Spanning Tree (PVST) simulation disable
- **Link Aggregation**
  - EtherChannel
- **Unidirectional Link Detection (UDLD)**
  - Detect optical link misconfiguration, Standard and Aggressive mode
- **Port Security**
  - Source MAC address based authentication
- **802.1X**
  - L2 user/device authentication
- Traffic suppression for unicast, multicast and broadcast traffic

https://t.me/learningnets

NX-OS L2 Data Plane Features

- **VLAN**
  - IEEE 802.1Q VLANs and trunks, 16,000-subscriber VLANs
  - IEEE 802.3ad link aggregation
- **Private VLAN**
  - Restrict communication between hosts on a VLAN
  - Cross-chassis private-VLANs
  - Promiscuous/Isolated/Community PVLANs
- **L2 ACL**
  - Based on MAC Source Address
- **Supports any Ethernet frames, independent of L3 payload (Jumbo frames)**
  - Support for Ethernet frame size 64 B to Jumbo (~9 K) frames

NX-OS IP and Routing Protocols Supported

Cisco NX-OS supports a wide range of IP Version 4 and 6 (IPv4 and v6) services and routing protocols. The implementations of routing protocols are fully compliant with the latest standards, providing modern enhancements and parameters such as 4-byte autonomous system numbers (ASNs) and incremental SPF, while shedding unutilized legacy functionality in favor of a lean implementation that improves feature velocity and system stability. Nonstop Forwarding Graceful Restart (NSF-GR) is supported by all unicast protocols. All protocols support all interface types, including Ethernet interfaces, switched virtual interfaces (SVIs) and subinterfaces, PortChannels, tunnel interfaces, and loopback interfaces. The rich variety of routing protocols and functions is complemented by a broad collection of IP services, including the following:

- Virtual Routing and Forwarding (VRF). All routing protocols and IP services are made VRF aware day one.
- Dynamic Host Configuration Protocol (DHCP) Helper
- Unicast Reverse Path Forwarding (uRPF), up to 16 interfaces
- Equal Cost Multipathing (ECMP) for unicast (up to 16 paths)
- Hot-Standby Routing Protocol (HSRP)
- Virtual Router Redundancy Protocol (VRRP)
- Gateway Load Balancing Protocol (GLBP)
- Enhanced Object Tracking
- Policy-Based Routing (PBR)
- Unicast Graceful Restart for all protocols in IPv4 Unicast Graceful Restart for OPSFv3 in IPv6
- Generic Route Encapsulation (GRE) Tunneling protocol (RFC 2784) (IPv4 Over GRE/IPv4, IPv6 Over GRE/IPv4, IPv4 Over GRE/IPv6, IPv6 Over GRE/IPv6, Tunnel Path MTU discovery (PMTUD), GRE Tunnel Keepalive, GRE Tunneling with IPv4 delivery header, GRE Tunneling with IPv6 delivery header)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NX-OS L3 Control Plane Features

OSPFv2, OSPFv3, EIGRP, IS-IS, HSRP, GLBP, static routing (IPv4 and IPv6), Equal Cost Multipath (ECMP), First Hop Redundancy Protocol(FHRP) Virtual Router Redundancy Protocol(VRRP) and BGP, GRE, RIP v2, RIP (IPv6), VRF, Route Policy Manager, Layer 3 virtualization: VDC, VRF. All routing protocols (OSPF, OSPF, IS-IS, RIP, BGP) support IPv4 and IPv6 but EIGRP which does not support IPv6 on NX-OS. All protocols (OSPF, EIGRP, OSPF, IS-IS, RIP, BGP) are VRF-aware and all protocols (OSPF, EIGRP, OSPF, IS-IS, BGP) but RIP support Graceful Restart.

NX-OS L3 Data Plane Features

IPv4/IPv6 (unicast and multicast), IP ACL, 5-tuple L3-L4 ACL, Netflow (including TCP flags, L2 fields, sampling v9 (and v5) data export versions), Netflow Policy Based Routing (PBR), unicast reverse path forwarding

Layer 3 Configuration Limits

BGP: 300 peers, 150,000 prefixes per peer, 20 million paths total

GLBP: 1024 GLBP groups

HSRP: 2048 HSRP groups, 20 HSRP groups per interface.

OSPF: 200 interfaces, 1000 routers, 300 adjacencies, 200,000 LSAs

Network Security Features

- Cisco TrustSec: New NX-OS data center security implementation that provides data confidentiality and integrity, supporting standard IEEE 802.1AE link-layer cryptography with 128-bit Advanced Encryption Standard (AES) cryptography. Link-layer cryptography guarantees end-to-end data privacy while allowing the insertion of security service devices along the encrypted path. Security group access control lists (SGACLs),
- Datapath Intrusion detection system (IDS) for protocol conformance checks
- Control plane policing (CoPP)
- Message-digest algorithm 5 (MD5) routing protocol authentication
- Cisco integrated security features, including Dynamic Address Resolution Protocol (ARP) inspection (DAI), DHCP snooping, and IP source guard (PACL based), Dynamic ARP Inspection (Binds IP and MAC addresses to prevent APR spoofing), DSCP Snooping (Snooping of DHCP response)
- Authentication, authorization, and accounting (AAA) and TACACS+
- Secure Shell (SSH) Protocol Version 2
- Simple Network Management Protocol Version 3 (SNMPv3) support
- Port security
- AAA, IEEE 802.1x authentication and RADIUS, TACACS+ support
- Layer 2 Cisco Network Admission Control (NAC) LAN port IP

https://t.me/learningnets

- Policies based on MAC and IPv4 addresses supported by named ACLs (port-based ACLs [PACLs], VLAN-based ACLs [VACLs], and router-based ACLs [RACLs], Service Group ACLs [SGACL])
- Traffic storm control (broadcast/multicast/unicast suppression)
- SSHv2, telnet, SSL, Keychain management, Control Plane Policing, Rate Limiters

NX-OS ACL Types

- MAC ACLs (extended), which filter L2 (Ethernet) protocol types or MAC addresses
- IPv4 ACLs (extended IP ACLs), which filter IPv4 traffic, including TCP, UDP, and ICMP traffic
- IPv6 ACLs, which filter IPv4 traffic, including TCP, UDP, and ICMP traffic
- ACL redirect
- PACLs, VACLs, RACLs
- RBACLs based on security group tags
- Policy Based Routing (ACL points to adjacency): Uses ACL redirect mechanism.
- ACL based accounting
- Atomic ACLs

ACL Supported Options

- Source (address/length, address wildcard, host, any)
- Protocol (eigrp, gre, icmp, igmp, ip, ipinip, nos, ospf, tcp, or udp, or an integer in the range from 0 to 255 representing an Internet protocol number)
- Destination (any, host, masked)
- Precedence, ToS, DSCP
- Logging
- Time-range
- Fragments–The access list entry applies to noninitial fragments of packets, this keyword cannot be configured in the same rule with Layer 4 matches, because the context necessary to evaluate them is contained only in initial fragments.
- icmp-type, icmp-code, icmp-message
- igmp-type
- Operator - lt (less than), gt (greater than), eq (equal), neq (not equal), and range (inclusive range)
- Port–The decimal number or name of a TCP or UDP port. A port number is a number from 0 to 65535. TCP and UDP port names are listed below.
- TCP specific options - urg, ack, psh, rst, sync, fin, established

Note Only named ACLs will be supported (all numbered ACLs will be translated and converted to named ACLs)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NX-OS Management and Monitoring Features

Troubleshooting and diagnostics tools supported by Cisco NX-OS are built with unique serviceability functions to enable network operators to take early action based on network trends and events, enhancing network planning and improving network operations center (NOC) and vendor response times. These tools are the following:

- SSHv2, SSL, telnet (browser based, GUI-based)
- Remote Monitoring (RMON)
- Data Center Network Manager (DCNM)
- Software Licensing
- Role based Access Control (RBAC)
- Network Management Tools (Configuration, Inventory, fault management, Performance monitoring), L4-L7 Services management
- Rollback and Session Management
- Smart Call Home, Cisco Generic Online Diagnostics (GOLD), and Cisco NX-OS Embedded Event Manager (EEM) are some of the features that enhance the serviceability of Cisco NX-OS.
- **Embedded packet analyzer**: Cisco NX-OS includes a built-in control plane traffic packet analyzer to monitor and troubleshoot control plane traffic. The packet analyzer is based on the popular Wireshark open source network protocol analyzer.
- **Smart Call Home**: The Smart Call Home feature continuously monitors hardware and software components to provide e-mail-based notification of critical system events. A versatile range of message formats is available for optimal compatibility with pager services, standard e-mail, and XML-based automated parsing applications. Smart Call Home offers alert grouping capabilities and customizable destination profiles. This feature can be used, for example, to directly page a network support engineer, send an e-mail message to a NOC, and employ Cisco AutoNotify services to directly generate a case with the Cisco Technical Assistance Center (TAC). This feature is a step toward autonomous system operation, enabling networking devices to inform IT when a problem occurs and helping ensure that the problem is acted on quickly, reducing time to resolution and maximizing system uptime.
- **Cisco GOLD**: Cisco GOLD is a suite of diagnostic facilities used to verify that hardware and internal data paths are operating as designed. Boot-time diagnostics, continuous monitoring, and on-demand and scheduled tests are part of the Cisco GOLD feature set. This industry-leading diagnostics subsystem allows rapid fault isolation and continuous system monitoring critical in current continuously operating environments.
- **Cisco EEM**: Cisco EEM is a powerful device and system management technology integrated into Cisco NX-OS. Cisco EEM helps customers harness the network intelligence intrinsic to the Cisco software and enables them to customize behavior based on network events as they happen.
- **Cisco Netflow**: Cisco NX-OS implementation of Netflow supports version 5 and version 9 exports as well as the Flexible Netflow configuration model and hardware-based Sampled Netflow for enhanced scalability.

https://t.me/learningnets

IP Multicast Features

Cisco NX-OS provides an industry-leading IP Multicast feature set. The Cisco NX-OS Release 4.0 implementation lays the foundation for the future development of a rich portfolio of multicast-enabled network functions. Similar to the unicast routing protocols, Cisco NX-OS Release 4.0 includes state-of-the-art implementations of the following multicast protocols and functions:

- Protocol Independent Multicast Version 2 (PIMv2)
- Source Specific Multicast (SSM)
- PIM sparse mode (Any-Source Multicast [ASM] for IPv4 and IPv6)
- Bidirectional Protocol Independent Multicast (Bidir PIM)
- Anycast Rendezvous Point (Anycast-RP)
- Multicast NSF for IPv4 and v6
- RP-Discovery using bootstrap router (BSR): Auto-RP and static
- Internet Group Management Protocol (IGMP) Versions 1, 2, and 3 router role
- IGMPv2 host mode
- IGMP snooping
- Multicast Listener Discovery (MLD) Protocol Version 2 (for IPv6)
- Multicast Source Discovery Protocol (MSDP) (for IPv4 only)

None of the multicast services support graceful restart.

Cisco NX-OS eliminates obsolete functions such as PIM dense mode; this is an example of the forward-looking direction of the operating system.

NX-OS QoS Features

Cisco NX-OS supports a rich variety of QoS mechanisms, including classification, marking, queuing, policing, and scheduling. Modular QoS CLI (MQC) is supported for all QoS features. MQC can be used to provide uniform configurations across various Cisco platforms:

- **Queuing and scheduling**: Including policy maps of type queuing, congestion avoidance, congestion management
- **Type QoS policies**: Used to mark, to apply mutations, to set the ingress port trust state, and to police packets)
- **Type queuing policies**: Used to mark, shape, and queue packets, marking is limited to the CoS field and does not support the use of table maps
- **Classification criterias**: ACL, IP precedence, DCSP, protocol, QoS group, discard class, packet length, CoS, RTP, class-maps
- **Policing**: Single-rate, dual-rate, and color-aware policers
  - **Congestion avoidance**: WRED and WRR
- **Congestion management**: Egress management features: priority, scheduling, bandwidth in output queues and bandwidth in input queue
- **Port trust**: Untrusted ingress packet get set CoS field

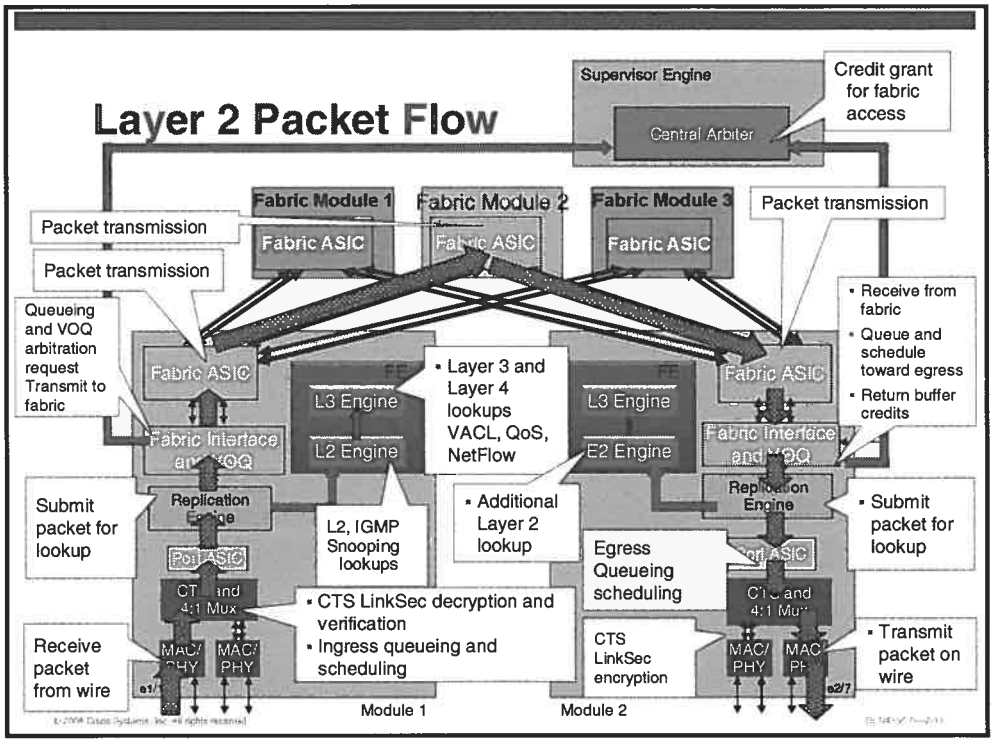
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **1 Gigabit Ethernet Module on Ingress supports**: 2 queues with 4 thresholds per queue
- **1 Gigabit Ethernet Module Egress supports**: 1 strict priority queue and 3 normal queues with 4 thresholds per queue
- **10 Gigabit Ethernet Module on Ingress supports**: 8 queues with 2 thresholds per queue
- **10 Gigabit Ethernet Module on Egress supports**: 1 strict priority queue and 7 normal queues with 4 thresholds per queue

https://t.me/learningnets

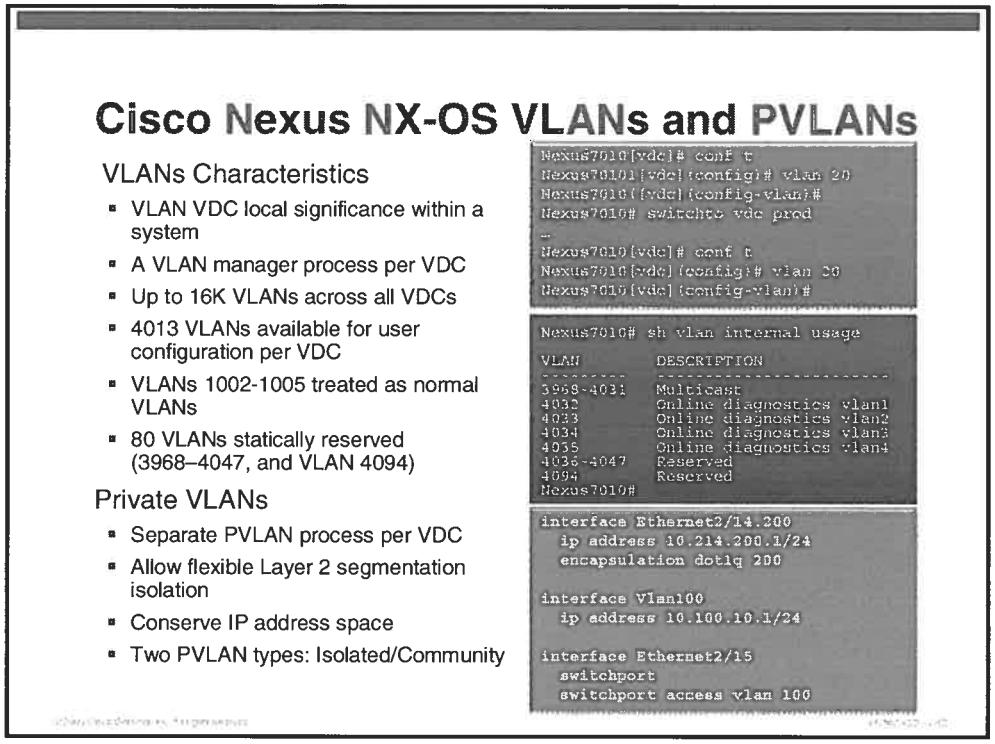
Layer 2 Packet Flow



This diagram represents the Layer 2 packet flow with focus on the Layer 2 functions. First, the packet is extracted by the MAC line card and Cisco TrustSec performs packet decryption if it was encrypted. The Replication engine submits the packet for the lookups. The Layer 2 lookups in the ingress Fast Ethernet are performed on the packet ingress path. The frame is passed to the Layer 3 Fast Ethernet for security reasons, where a VACL or a PACL filtering applies to the packet, an ingress QoS policy, VLAN based policing, marking policy or a L2 port based policy is applied to the frame. If bridged NetFlow is enabled, NetFlow statistics will be connected in the Layer 3 engine on the ingress engine. An additional Layer 2 lookup is performed at the egress Forwarding Engine and then the packet is queued, scheduled, encrypted and transmitted on the wire.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The VLAN Manager is a process that manages the VLAN configuration space within the Nexus system. It runs on a per-VDC basis. In total, there are up to 16 K VLANs available, but there cannot be more than 4 K (4.013) VLANs enabled per each VDC. LIF interfaces are special interfaces in this context because they are equivalent to internal VLANs, but they have no direct relation to VLANs and therefore they are not consuming the available VLAN space. VLAN VDC has local significance within a system and VLANs have interface local significance within a VDC. Up to 16 K VLANs can be configured across all VDCs; 4013 VLANs are available for user configuration per each VDC. Eighty VLANs are statically reserved for multicasting and for other future purposes (3968–4047, and VLAN 4094) and are not user accessible, configurable or changeable. They have the least significant bits of the VLAN IDs set to zero. VLANs range 1002-1005 is treated as normal VLANs – the Token Ring and Fiber Distributed Data Interface (FDDI) reserved VLANs.

NX-OS 4.0 PVLAN Characteristics

Private VLANs allow flexible segmentation and isolation of traffic groups and host communication at Layer 2. Deployment of private VLANs conserves IP address space by reducing need to subnet. There is a separate PVLAN process running per each VDC. PVLAN is conditional service and as such must be manually enabled. PVLAN consists of one or more VLAN pairs:

- primary-vlan, isolated-vlan
- primary-vlan, community-vlan
- primary-vlan, isolated-vlan; primary-vlan, community-vlan-1; primary-vlan, community-vlan-n

https://t.me/learningnets

There are two PVLAN types:

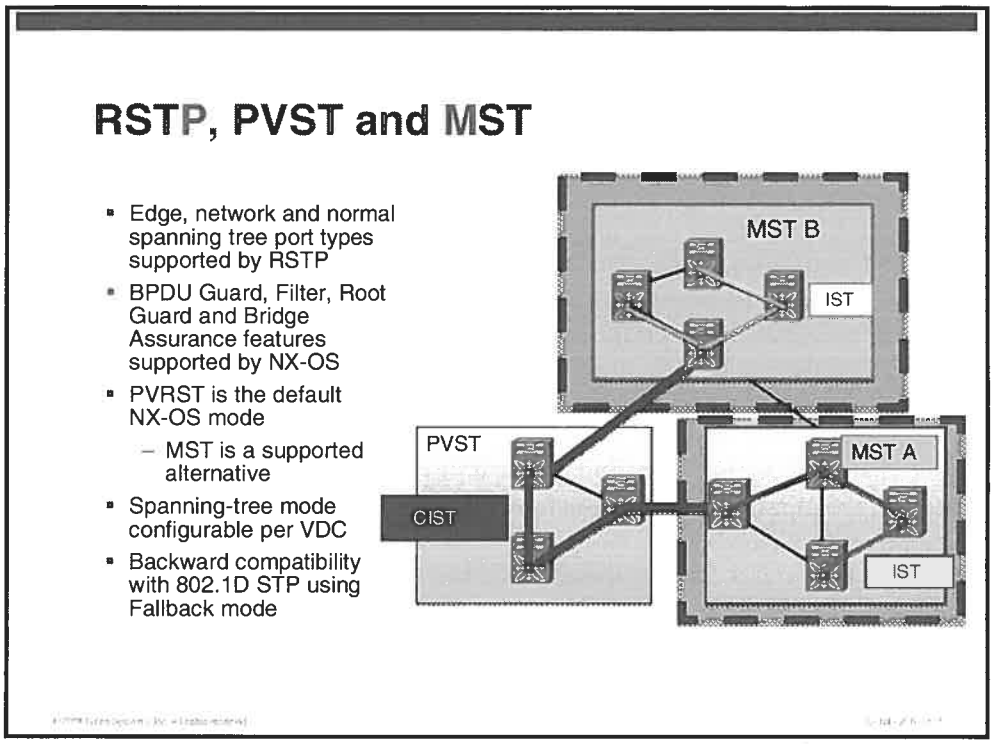
- **Isolated PVLAN**
  Hosts in isolated PVLANs can only communicate through a promiscuous port.
- **Community PVLAN**
  Hosts in Community PVLANs can only communicate with other hosts in the community VLAN and through a promiscuous port.

Note NX-OS supports Unidirectional Link Detection (UDLD). UDLD is a light-weight Layer 2 NX-OS protocol that detects and disables one-way Layer 1 connections. UDLD is a conditional service and must be manually enabled. UDLD speeds up the detection and reaction to the physical level connectivity. When enabled globally all 10 Gb/s (fiber) interfaces run UDLD automatically whereas 1 Gb/s (copper), must enable manually. Per-interface UDLD supports two operational modes:

- Normal mode
  If UDLD neighborship times out, port state marked as undetermined and port behaves according to STP
- Aggressive mode
  If UDLD neighborship times out, UDLD tries to re-establish port state. If unsuccessful, port is put into errdisable state.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



NX-OS introduces new Spanning Tree port types: edge, network and normal ports. Edge and network ports have specific characteristics and behaviors.

Edge port types replace the spanning-tree PortFast feature. Network port types are used for bridge-to-bridge links and normal port types are used for generic ports in the spanning tree topology. The PortFast configuration on the NX-OS is slightly different from the PortFast configuration on the Catalyst Series switches. An edge port can have the BPDU Guard and the BPDU Filter enabled. The BPDU Guard enabled on the edge ports error-disables the edge ports that have received a BPDU. The BPDU Filter enabled on the edge port disables both sending and receiving of BPDUs on the edge ports. When the BPDU filter is installed, then any BPDU that is received by the port is dropped and (a BPDU) is never sent out of that edge port. The Root Guard enabled on the edge port places a port in the root-inconsistent state (error-disabled port) if it receives a superior BPDU. The Root Guard, therefore, protects the perimeter of the spanning tree from superior BPDUs. When the port stops receiving superior BPDUs it will bring the port back into the Spanning Tree Protocol process. The Loop Guard, thus, prevents an alternate or root ports from becoming designated ports due to unidirectional link failure.

NX-OS supports a special function called Bridge Assurance (BA). The Bridge Assurance process ensures that spanning tree fails rather than continues forwarding traffic causing a Layer 2 forwarding loop. Bridge Assurance is a conditional service, enabled globally with spanning-tree bridge assurance. All ports that are network port types send BPDUs regardless of their state (Root ports, Alternate ports, etc.).When a network port stops receiving BPDUs, the port is placed in BA Inconsistent state (blocked) to prevent forwarding loops.

NX-OS 4.0 spanning tree implementation supports both PVRST and MST protocols. These are the current forms of Spanning Tree Protocols according to 802.1D and the 802.1Q standards. The spanning tree mode is configurable per VDC. PVRST is the mode enabled by default. The so-called dispute mechanism supported by the NX-OS PVST protocol detects unidirectional link failures through a mechanism in the BPDUs. This functionality is part of the 802.1D standard. The default PVST path cost method is short, whereas the MST defaults to long path cost spanning tree method. The Cisco proprietary BPDU TLVs are supported to detect PVID

https://t.me/learningnets

mismatches. The PVID mismatch detects a VLAN mismatch between two interfaces when there are BPDUs received from the two different ports and they belong to different VLANS (even if they are not tagged). This feature also helps to check for trunk to access port misconfigurations. The Spanning Tree Protocol can be disabled on a per-VLAN basis. In this case, BPDUs are just flooded through that logical topology. NX-OS supports up to 64 MST instances along with the dispute mechanism supported.

The PVST simulation enables the MST bridge to interoperate with a per-VLAN spanning tree domain. The common spanning tree information is being replicated across all the VLANs of the trunk. Instead of the MST bridge sending MST BPDUs, only per-VLAN BPDUs are being sent. The MST protocol is namely designed to send all the BPDU information in a single BPDU versus per-VLAN Rapid Spanning Tree, which sends a BPDU per-VLAN. The MST bridge (with PVST simulation enabled) instead of sending out one big, fat MST BPDU towards the PVST cloud, the common Spanning Tree information using a standard BPDU is replicated across all the interfaces. In addition to the standard MST implementation, NX-OS supports pre-standard MST. Cisco NX-OS can interoperate with the pre-standard MST BPDUs (with additional hash field). The pre-standard BPDU is not configurable, but is rather detected and enabled automatically upon receiving a pre-standard MST BPDU from the MST bridge. NX-OS supports a Cisco proprietary method that prunes out the MST records in a MST BPDU if a particular VLAN is not active on the trunk ISL. When there is a trunk, the MST sends by default all the information about all the VLANs across the trunk even if those VLANs are not actually active on that trunk. Therefore, it is possible on NX-OS to prune out the propagation of MST BPDUs for VLANs that are not supported by the trunk to improve the efficiency of the transport.

75 k MST instances at maximum. 16 k for RSTP

Scalability for logical ports in the spanning tree: PVRST: 10,000, MST: 50,000 logical ports

NX-OS Spanning Tree Summary

NX-OS Layer 2 protocol features set include hardware forwarding and hardware MAC learning, modern Spanning Tree Protocols (MST/RST), full suite of STP guards, new features like Bridge Assurance and stateful Layer 2 switchover without reconvergence.

In terms of Layer 2 loop prevention mechanisms NX-OS 4.0 supports almost all standards currently supported by Cisco switches. Rapid Spanning Tree Protocol (RSTP) mode is configurable and runs per each VDC. NX-OS supports Per-VLAN Spanning Tree Protocol (PVST) with the following characteristics:

- PVST is a default spanning tree mode for new VLANs
- NX-OS PVST adheres to IEEE 802.1D-2004 standard
- NX-OS PVST supports dispute mechanism to detect unidirectional link failures
- NX-OS PVST default path cost method is "short"
- NX-OS PVST supports PVID and Type inconsistency check
  - PVID inconsistency check to prevent VLAN cross-wiring
  - Type inconsistency to check for trunk/access port interconnects
- Spanning tree can be disabled on per-VLAN basis.
- BPDUs are flooded within the VLAN when spanning tree is disabled

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Multi Spanning Tree Protocol (MST) is supported by NX-OS 4.0 with the following characteristics:

- NX-OS MST supports up to 64 MST instances, each identified by MST-ID in range 1-4094
- NX-OS MST supports a dispute mechanism to detect unidirectional link failures
- NX-OS MST path cost method is always "long"
- NX-OS MST supports PVST simulation and allows MST cloud to interoperate with PVST cloud over an 802.1Q trunk
- NS-OS MST replicates CIST information to all VLANs on trunk
- NX-OS MST interoperates with Cisco pre-standard MST switches upon detection of pre-standard MST BPDU, but user cannot manually configure pre-standard interoperability mode
- NX-OS MST supports MST pruning in MST BPDU based on active VLANs on a port

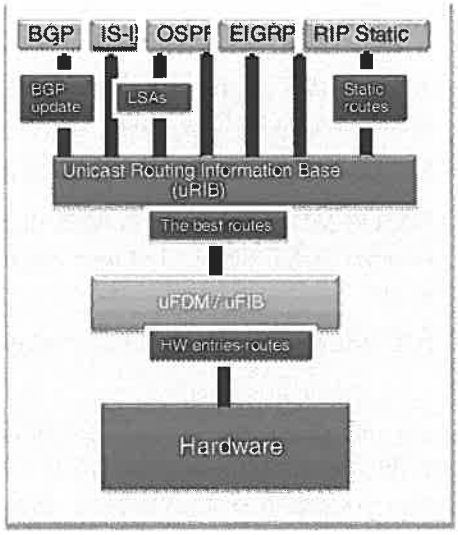
Default spanning tree mode in NX-OS is Per-VLAN Rapid Spanning Tree (PVRST) whereas in Cisco IOS the default mode is PVST. NX-OS has Bridge Assurance by default disabled whereas in Cisco IOS it is enabled. Extended System ID in NX-OS always uses Extended System ID whereas in Cisco IOS it is configurable. PortFast mode in NX-OS is configured with spanning-tree port type edge whereas in Cisco IOS with spanning-tree portfast. NX-OS supports backward compatibility with 802.1D STP using fallback mode.

Note PVRST can support 10.000 logical ports and MST can support 50.000 logical ports participating in the spanning tree topology. Logical ports are just interfaces managed by the Spanning Tree Protocol, whether it is an actual, physical switch port interface or a trunk with multiple VLANs carried on it where each of those VLANs represents a logical port. If the topology supports trunks with a lot of VLANs, then the logical port count increases, even when the number of physical ports is not high. So far there are no per-module restrictions on logical port counts. (Needs to be confirmed by Cisco)

https://t.me/learningnets

Cisco Nexus NX-OS Layer 3 Protocol Characteristics

- Each L3 protocol AS is run as an independent instance.
- Protocols can be enabled/disabled.
- All L3 protocols are capable of controlled restart–some support GR (IETF)
- Layer 3 CP and DP are highly separated which is an enabler for non-disruptive ISSU and patching.
- NSF is enabled by default on all protocols. NSF is a superset of separation of CP and DP which includes controlled restart and GR.
- NX-OS Tunnel Types:
  - Tunnels (Automaticaly created)
  - Explicit Tunnels (Manualy configured)
- NX-OS supports 4 ECMP load-balancing types



Cisco Nexus 7000 Unicast Routing Architecture is based on a unicast Routing Information Base that offers a common API to the control plane routing protocols. Prefixes received from the protocol processes are packed in shared memory buffers on uRIB. uRIB stores routing updates, processes the routing information and selects the best next-hop for every prefix in uRIB. The best route is downloaded to the uFIB and mapped into the hardware data structure format into the Fast Ethernet hardware ASIC.

Unicast RIB provides the best path selection for both unicast forwarding and for multicast RPF. It finds and stores backup paths for both unicast forwarding and multicast RPF. Unicast RIB also performs route redistribution and route watch (a.k.a. Recursive-Next-Hop notifications). Unicast RIB detects routing loops and prevents them ("hiding"). Unicast RIB supports multiple VRFs (for vrf lite only) client convergence notifications and route downloads to FIBs with flow control. Classful routing, unequal-cost load balancing (for EIGRP) and non-0/0 default routes (candidate for defaults) are not supported by Unicast RIB. The maximum number of paths, as well as route filtering, is enforced in clients and not in uRIB/u6RIB. Recursive route resolution is done by clients instead, via the passive server (fully client-driven). The design allows easy insertion of new clients. uFIB is table-based or multi-topology capable with no direct Interface Manager or IP event handling. These events will be reflected in RIB by the client's actions.

Controlled restart is supported for all Layer 3 protocols. Graceful restart is On by default for all protocols that support it. There is limited support for database or route clear commands. Administrators are supposed to use the restart command, instead, for clearing the complete protocol state. The router-id election algorithm is slightly different from the Cisco IOS implementation and the route-map specification is mandatory for route redistribution. There is no support for MPLS or 2547-based L3VPN solutions.

Each protocol instance is an independent service. Protocols can be enabled and disabled. When they are disabled, instances of the protocol are disabled, associated resources released and related protocol configuration removed. NSF is enabled by default on all protocols Except for EIGRP, the GR implementations in NX-OS are NOT compatible with Cisco prestandard IOS NSF implementation. NSF implementations (except for EIGRP) are standards-based and not

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

compatible with Cisco NSF in IOS. Use the IETF version of graceful restart on Catalyst 65000 Series switch (Whitney 1) for interoperability with NX-OS.

Similar to SAN-OS, NX-OS supports 2 types of tunnels: implicit and explicit.

Implicit Tunnels

Implicit tunnels are automatically created tunnels and configured by applications using tunnel APIs. These tunnels cannot be modified by users using implicit tunnel configurations. An example of an implicit tunnel is a CTS tunnel.

Explicit Tunnels

Explicit Tunnels are configured by users. A tunnel manager is a process that creates the tunnels and is a VDC local process. Multiple tunnels in different VDCs can be established with the same name. The tunnel source and the tunnel destination IP-address should be in the same VRF, but the tunnel interface can itself be in a different VRF. Note that the tunnel Path-MTU-Discovery will not work if ICMP packets are blocked by firewalls. A unique source address must be used as the tunnel source address and no internal VLAN is reserved for the tunnel interface in NX-OS, as compared to the Catalyst 65000 Series switch. It implies that NX-OS does not have the following restriction:

```
GRE Tunnels = 4096 - reserved VLANs - Number of Layer 3 ports
- Number of other logical interfaces
```

ECMP Protocol load balances the traffic across up to 16 equal cost paths per prefix. The load-balancing process is done in hardware. There are four hashing options for load balancing across equal cost paths based on:

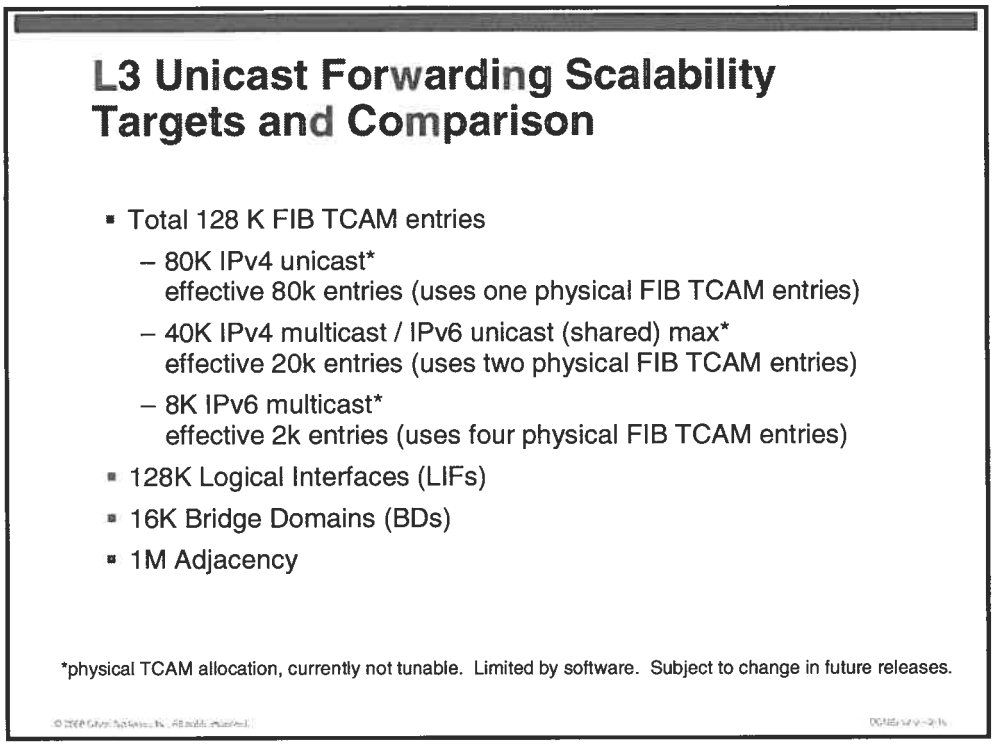
- Layer3 (use only L3 fields)
- Layer3-4 (use both L3 and L4 ports)
- dest-addr (use only destination address)
- dest-addr-port (use destination address and ports)

Load-balancing is configurable only in default VDC, but has global scope. Router ID of default VDC and default VRF are used to randomize the ECMP hash.

FIB Adjacencies and ECMP Relationship

FIB adjacency or ECMP objects are shared. Multiple prefixes in the same or in different tables, but in the same VDC, can point to an adjacency (ECMP) Adjacency is tied to the interface structure. Because a logical interface can appear only in one VDC, an adjacency or ECMP can only be shared by prefixes in a VDC. There is no separate (from adjacency) hardware representation for ECMP in the current hardware. An ECMP object contains a block of adjacency objects with a limit of 16 adjacencies. EMCP FIB objects are not implemented as separate objects in the FIB, they are represented by a set of contiguous objects with a single pointer for a destination to the contiguous block.

https://t.me/learningnets



The table below summarizes the L3 Unicast forwarding scalability comparison of FIB table sizes, MAC address table sizes, and other performance parameters for Catalyst 65000 and Nexus 7010 Series switch.

Table 1: Feature, Earl Versions

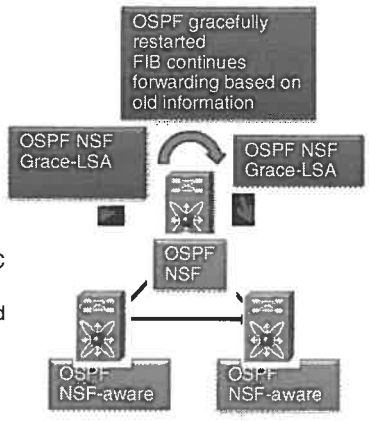
Feature	Cat6k	VSS	Nexus 5700
MAC Address Table size	64K	96K	128K
FIB Table size	1M(XL)	1M(XL)	128K
Forwarding Performance			
Distributed (IPv4)	270Mpps	270Mpps	480Mpps
No. of Physical ports	4K	4K	10K
No. Of Interfaces (VLAN/BD)	4K	4K	16K
ACLs	32K	32K	64K
QoS	32K	32K	
Netflow entries	256K	256K	512K

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS and IOS—Main Differences

- Each NX-OS protocol instance is an independent service
- NSF implementations (except for EIGRP) are standards-based and not compatible with Cisco NSF in IOS
- The IETF RFC-based graceful restart extensions for controlled restart are implemented to leverage the nonstop forwarding in NX-OS
- NSF is enabled by default on all protocols
- Protocols can be enabled and disabled on a per-VDC basis
- Routing protocol configuration is interface-centric and output debugging is enhanced using various filters
- When the protocol is disabled:
  - Disables all instances of the protocol
  - Releases associated resources
  - Removes protocol-related CLI
  - Removes protocol-related config
  - Prevents all potential problems with the process



The general differences between NX-OS and the Cisco IOS are that each protocol instance is an independent service on NX-OS. Therefore, every instance of routing protocols that are run, are run as a separate service on the NX-OS platform. This is referred to as a modularity of software process. Nonstop forwarding (NSF) for routing protocols in Cisco NX-OS software, using the IETF standardized graceful restart function, is described in RFC 3623. Under very specific situations, a switch or router may undergo certain well-known failure conditions that should not affect packet forwarding across the switching platform. NSF capability allows for the forwarding of data packets to continue along routes that are already known, while the routing protocol information is being restored. This capability is useful in cases in which there is a component failure [for example, a Route Processor (RP) crash with a backup RP taking over] or in which there is a scheduled hitless software upgrade.

For example, the RFC 3623 graceful restart nonstop forwarding enhancement to the OSPF routing protocol introduces an OSPF NSF-capable router that is reloading and attempting a graceful restart, originates grace LSAs to notify its neighbors that it will perform graceful restart within the specified amount of time or grace period. During this grace period, the neighboring OSPF routers, called helper routers, continue to announce the restarting router in their LSAs as if it were fully adjacent, as long as the network topology remains static. Prior to RFC 3623, Cisco implemented Cisco-proprietary NSF, referred to as Cisco NSF. The OSPF RFC 3623 Graceful Restart feature allows you to configure IETF NSF in multivendor networks. Configuring IETF NSF increases the availability of your network by allowing OSPF routers to stay on the forwarding path even as their OSPF software is restarted.

The nonstop forwarding and graceful restart implementations are standards-based and are not compatible with what is known as Cisco NSF in IOS. In other words, the IETF RFC-based graceful restart extensions for controlled restart are implemented to leverage the nonstop forwarding in NX-OS. The only exception is the NX-OS EIGRP because there is no IETF standard for graceful restart because there is no standard EIGRP protocol. The EIGRP module and its graceful restart extensions were taken from Cisco IOS and thus are compatible. By default, in all the routing protocols on NX-OS, graceful restart and controlled restart are enabled by default. Many features on the platform such as individual routing protocols can be

© 2008 Cisco Systems, Inc.

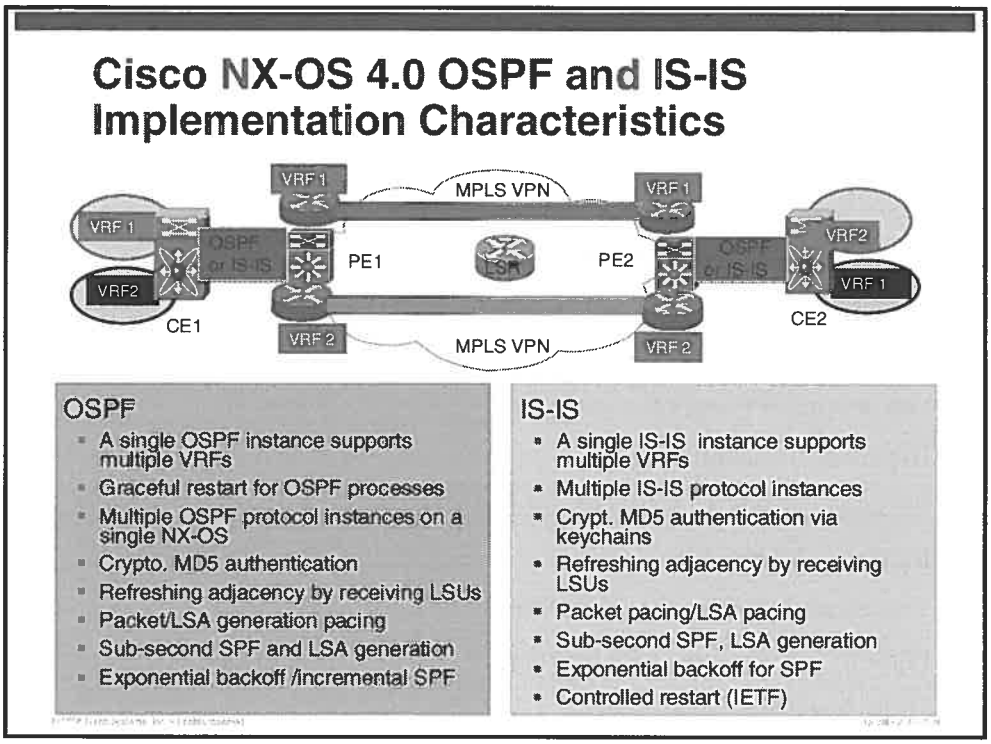
https://t.me/learningnets

enabled or disabled. This capability is useful in a case when the customer is not using a certain protocol. Those processes can be then disabled which releases all the resources related to the protocol being disabled, relinquishes all the resources and disables all instances of the protocol. The protocols can be enabled or disabled on a per-VDC basis. When a protocol such as the OSPF is shut down, all the instances of OSPF are shut down. NX-OS also permits to shut down only a given instance of OSPF process. When the OSPF protocol is completely disabled as a feature, then the related configuration is removed from the configuration file.

Note There are two general classes of controlled restart implementation. Cisco NSF, which was implemented before the IETF standardized on how graceful restart signaling should be done between neighbors. Cisco protocols used this sort of a Cisco proprietary way of notifying neighbors that the routing protocol were restarting but to not stop forwarding to the platform. Packages would still be routed to the platform while the control plane protocol was being restarted. This was the Cisco proprietary NSF operation implementation. The IETF has standardized how graceful restart can be done in certain protocols, for example in OSPF and BGP. NX-OS implements the IETF method. If there is a topology running the former—the pre-standards-based Cisco implementation of nonstop forwarding or graceful restart—they will not be compatible. Therefore, the NX-OS IETF version cannot talk to the Cisco pre-standard-based NSF version. However, Cisco IOS, in certain versions, has implemented the IETF-based graceful restart. Thus, there are versions of Cisco IOS that can talk to NX-OS, using the RFC-based graceful restart extensions. It is a matter of which Cisco IOS version is running if the topology is taking advantage of nonstop forwarding.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Characteristics of the Cisco NX-OS 4.0 OSPF Implementation

The NX-OS OSPF supports graceful restart (IETF). NX-OS can run multiple OSPF protocol instances. A single OSPF process instance can support multiple VRFs. OSPFv3 is also supported, which runs and ships with the OSPFv2. There is a lot of discussion about moving OSPFv3 forward to support IPv4, as well. The software structure is ready to support this functionality in the NX-OS implementation, but the actual protocol handles only IPv6. The subsecond SPF and LSA generation and exponential back-off for SPF are supported by NX-OS OSPF, as well as the incremental SPF for inter-area prefixes and reference bandwidths. The debug filters are available to constrain the OSPF process debug output. Debugging can be filtered based on interface, specific neighbor, LSAs and SPF type. Some of the NX-OS new features are the graceful OSPF restart, according to RFC3623. OSPFv3 graceful restart is actually part of the OSPFv3 specification.

The traffic engineering extensions for OSPF or the multi topology routing (MTR) support for OSPF is not implemented on NX-OS. The Cisco nonstop forwarding and Cisco pre-standards-based graceful restart signaling is not supported. The OSPF multi-area adjacency is not supported currently by NX-OS, but it may be added in the future. As a result, for certain area border routers, depending on the topology and on the link speeds, the virtual links might be an alternative. Also, the lower bandwidth links might be used, instead, as preferred routes because they are inter-area routes. The NX-OS OSPF implementation has replaced the network command with an interface based structure.

In the OSPFv3, the show and configure commands support multiple address families even though OSPFv3 today supports IPv6 routing. But the IPv4 was put in place to be supported by OSPFv3. It has already been taken into account that OSPFv3 would be a multi-address family supporting protocol. Therefore, when OSPFv3 is configured, it is done in the address family sub-mode in order to configure the address family specifics for the OSPFv3.

https://t.me/learningnets

Summary of NX-OS 4.0 OSPF Supported Features

- Graceful Restart for OSPF processes
- Multiple OSPF protocol instances can be enabled on NX-OS so that a single OSPF process instance supports multiple VRFs.
- Cryptographic (MD5) authentication for OSPF routing information exchange.
- LSA refreshing adjacency liveness
- Packet pacing to limit the rate at which OSFP floods the LSA packets
- LSA generation pacing
- Sub-second SPF and LSA generation
- Exponential backoff for SPF
- Incremental SPF for Inter-Area-Prefix, Inter-Area-Router and AS-External LSAs
- Reference bandwidth
- Stub router advertisement for OSPFv2
- Filter for prefixes advertised in Inter-Area-Prefix LSAs between areas of an ABR
- OSPF Version 2 (RFC 2328)
- OSPF for IPv6 (RFC 2740)
- OSPF Stub Router Advertisement (RFC 3137)
- OSPF Version 2 MIB (RFC 4750)
- OSPF Version 3 MIB
- Graceful OSPF Restart (RFC 3623)
- OSPFv3 Graceful Restart
- OSPF Not-So-Stubby-Area Option (RFC3101)
- Alternative Implementations of OSPF Area Border Routers (RFC 3509)
- OSPF Opaque LSA Option (RFC 2370)

Summary of NX-OS 4.0 OSPF Features Not Supported

- Traffic Engineering Extensions to OSPF (RFC 3630)
- Multi-Topology (MT) Routing in OSPF (RFC 4915)
- Authentication/Confidentiality for OSPFv3 (RFC 4552)
- Cisco Nonstop Forwarding
- OSPF Multi-Area Adjacency
- NX-OS OSPF NSF (RFC3623) is NOT compatible with IOS Cisco NSF

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Differences between the NX-OS 4.0 OSPF Protocol Configuration and the IOS-Based OSPF

- NX-OS replaced the network command with the interface mode command, which allows less complex inclusion of networks in a particular OSPF process.
- OSPFv3 commands can support multiple address families
  Address family mode configuration for OSPFv3:
  - router ospfv3 foo
    - router-id 60.60.60.60
    - area 4 virtual-link 50.50.50.50
  - address-family ipv6 unicast
    - area 4 range 2006:2000::/32
  - redistribute static policy ospfv3-static

Characteristics of the NX-OS 4.0 IS-IS Implementation

One of the implementation highlights of the NX-OS IS-IS is the support for the subsecond SPF and the LSA generation. The LSA generation pacing is also supported by the NX OS IS-IS. NX-OS can perform IS-IS process controlled restart or graceful restart. NX-OS IS-IS implementation allows IS-IS multiple protocol instances and a single instance supporting multiple VRFs. IS-IS supports exponential back-off for SPF in NX-OS IS-IS. The IS-IS traffic engineering is not supported by NX-OS. Cisco nonstop forwarding also is not part of NX-OS IS-IS implementation. The NX-OS IS-IS provides no password support for the IS-IS area and IS-IS domain authentication. Instead, IS-IS supports the keychain authentication. Attached bit, hello padding and auto recovery when the LSP is full, condition clears, IS-IS partition avoidance, the incremental SPF, route tags, and the LSP refresh intervals are also not supported by NX-OS IS-IS implementation. The hello interval minimal command is also not supported by NX-OS IS-IS.

Summary of NX-OS 4.0 IS-IS Supported Features

- Cryptographic (MD5) authentication via keychains
- Refreshing adjacency liveness based on receiving LSUs
- Packet pacing to limit the rate at which we flood packets
- LSA generation pacing
- Sub-second SPF and LSA generation
- Exponential backoff for SPF
- Reference bandwidth
- CSNP caching
- Debug filters to constrain debug output
- Controlled restart (IETF)
- Multiple protocol instances
- Single instance supports multiple VRFs
- ISO 10589 Intermediate system to intermediate system intra-domain routing exchange protocol
- RFC 1195 Use of OSI IS-IS for routing in TCP/IP and dual environment

https://t.me/learningnets

- RFC 2763 Dynamic Hostname Exchange Mechanism for IS-IS
- RFC 2966 Domain-Wide Prefix Distribution with Two-Level IS-IS
- RFC 2973 IS-IS Mesh Groups
- RFC 3277 IS-IS Transient Blackhole Avoidance
- RFC 3373 Three-Way Handshake for IS-IS Point-to-Point Adjacencies
- RFC 3567 IS-IS Cryptographic Authentication
- RFC 3784 IS-IS extensions for Traffic Engineering
- Internet Draft Routing IPv6 with IS-IS

Summary of NX-OS IS-IS 4.0 Features Not Supported

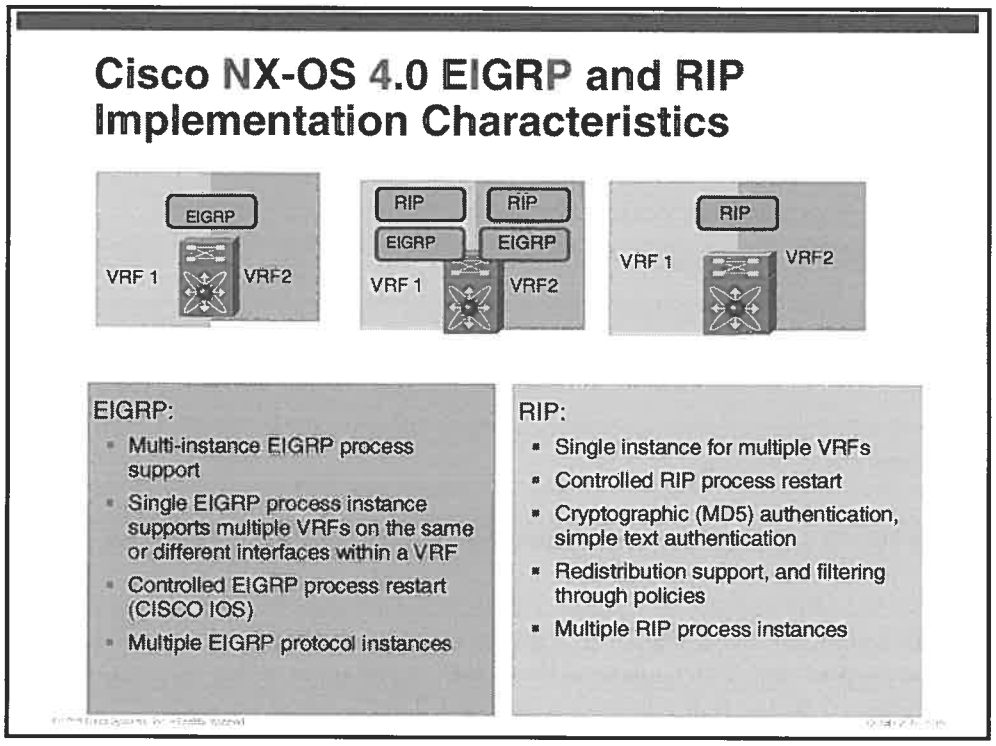
- Traffic engineering extensions to IS-IS
- Multi-topology routing
- Internet Draft Management Information Base for IS-IS
- Cisco Nonstop-Forwarding
- Partition avoidance (stop Level 1-2 border router advertising the Level 1 area prefixes into the Level 2 backbone)
- PRC (partial route calculations)
- iSPF
- route-tags
- The "debug isis rib" CLI command is not supported
- The "lsp refresh-interval" configuration command is not available. Instead set to half the max-lsp-lifetime hello-interval minimal.

Differences between the NX-OS 4.0 IS-IS and the IOS-Based IS-IS Implementation

- Area and domain authentication has no password support but supports keychain authentication
- Hello-Padding on for the first 5 packets only with no enable or disable option
- Set attached bit features has no option to set based on route-map and no support to set across address families
- Supports auto-recovery when LSP full condition clears
- NX-OS IS-IS support for "isis p2p" links is via media type p2p command

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco NX-OS 4.0 EIGRP Implementation Characteristics

NX-OS EIGRP implementation supports multiple instances of EIGPR process running on the same or a different interface and there can be multiple instances of EIGRP on a given interface within a VDC and within a VRF. There are EIGRP specific bandwidth and delay interface commands available for setting the interface metrics. There is an enhanced EIGRP debugging support. EIGRP supports controlled restart and this is compatible with the IOS versions of Cisco NSF for EIGRP, because we have taken the code directly from Cisco IOS. A single instance of EIGRP process can support multiple VRFs. Some of the things supported by EIGRP, graceful restart and Cisco NSF are supported because NX-OS uses the same EIGRP IOS code. NX-OS EIGRP instance can run both the IPv4 and the IPv6. The IPv6 support is being worked on and is not yet ready. NX-OS does not support some of the legacy protocols such as Apple Talk and IPX, as well as some of the newer protocols, like multi topology routing. By default a graceful restart and nonstop forwarding are enabled by default. The EIGRP auto-summarization is disabled by default by NX-OS. The EIGRP configuration has slightly changed in the following ways:

- The router mode commands have been replaced by the interface-based commands and the interface-centric configuration type. IPv4 and IPv6 support is integrated in the same EIGRP process instance.
- The route redistribution process from one EIGRP process to another EIGRP process does not preserve route metrics.

Summary of NX-OS 4.0 EIGRP Supported Features

- MD5 Authentication
- Graceful Restart and NSF (ENG-160004)
- IPv6 routing support, currently in progress
- Single EIGRP instance supports multiple VRFs

https://t.me/learningnets

- Input, output and redistribution route filtering (Route-maps)
- Multi-instance EIGRP process support
- Single EIGRP process instance supports multiple VRFs on the same or different interfaces within a VRF
- Controlled EIGRP process restart (as in CISCO IOS)
- Multiple EIGRP protocol instances

Summary of NX-OS 4.0 EIGRP Features Not Supported

NX-OS 4.0 implementation does not currently support AppleTalk/IPX traffic or Multi-Topology (MT) Routing.

Differences between NX-OS 4.0 EIGRP Implementation and IOS-OS EIGRP Implementation

NX-OS has NSF enabled by default. EIGRP auto-summarization is off by default. NX-OS EIGRP router mode commands are replaced with interface mode commands. NX-OS EIGRP Integrates IPv4 and IPv6 support in the same EIGRP process instance. EIGRP route metric is not preserved during redistribution from one EIGRP process to another EIGRP process, so 'set metric' in a route-map or the default (set by default-metric) is needed to determine route metric value.

NX-OS 4.0 RIP Implementation Characteristics:

RIPv1 is not supported on the NX-OS 4.0 platform. A single instance of RIP supports multiple VRFs on NX-OS and there can be multiple instances of RIP enabled. NX-OS RIP supports controlled restart. The controlled restart of the RIP process does not maintain neighbor adjacencies. There is no signaling among RIP neighbors when a RIP process is restarted on one of them. The protocol restarts, and as long as it is restarting before the dead timer expires, the neighbors will continue to forward data to it. This is not really a graceful restart, but rather a controlled restart. A RIP controlled restart will restart without its neighbors knowing and will come back and refresh its state through the periodic refresh of the protocol itself. The restarted process will not signal to its neighbors that it needs new updates, but will just wait for periodic updates from neighbors and continue to forward.

The RIP process does not support nonstop forwarding. It does not support any type of explicit signaling between a given node and its neighbors that it is entering the nonstop forwarding mode or needs nonstop forwarding to be activated. RIP supports a cryptographic authentication and authentication using keychains. RIPv1, RIP unicast periodic updates and broadcast packet flooding are not supported on NX-OS platform. RIP uses an interface-centric command structure. The commands to see the RIP routes have changed . Split horizons cannot be disabled on the platform; instead poison reverse can be enabled in order to achieve the same effect.

Summary of NX-OS 4.0 Supported Features

- NX-OS RIP supports single instance for multiple VRFs
- NX-OS RIP implements multiple RIP process instances
- RIPv2 and RIPng integrated in same process
- NX-OS supports controlled RIP restart

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- NX-OS RIP implementation supports extensive statistics and debug support with debug filters to constrain debug output per vrf, interface and address family
- NX-OS RIP supports cryptographic (MD5) authentication and simple text authentication and authentication support using keychain
- NX-OS RIP supports route redistribution and route filtering through policies
- RIP Version 2 (RFC 2453)
- RIP for IPv6 (RFC 2080)
- RIPv2 MD5 authentication (RFC 2082)
- RIP Version 2 MIB Extension (RFC 1724)

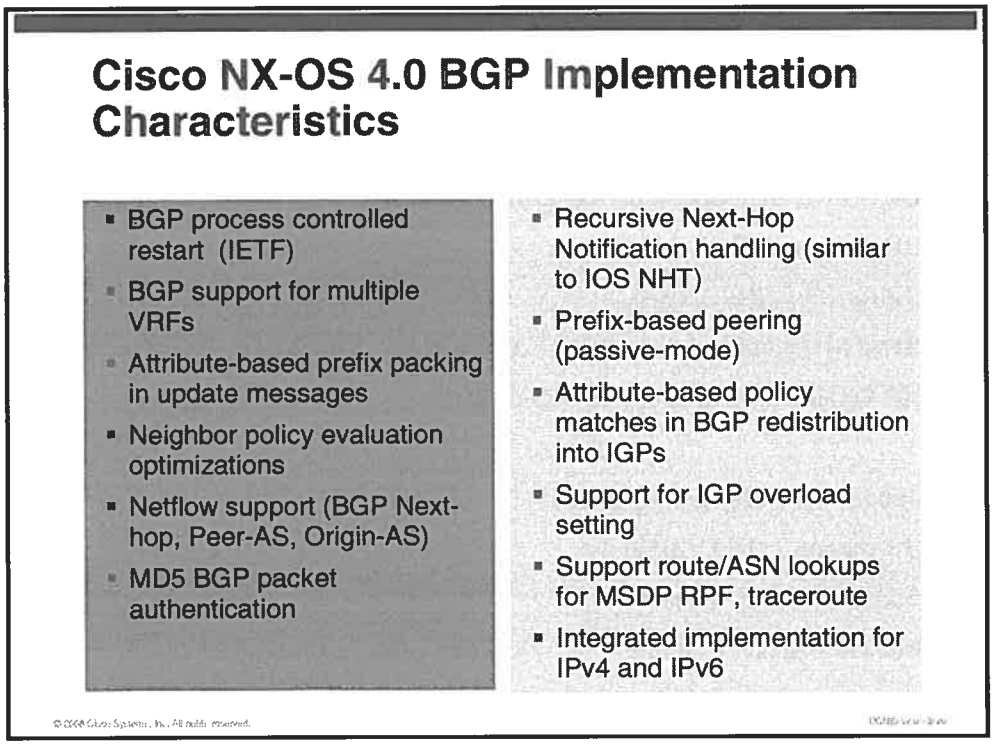
Summary of NX-OS 4.0 RIP Features Not Supported

- RIP version 1 (RFC 1058)
- Unicast periodic updates
- Broadcast packet flooding

Differences between NX-OS 4.0 and IOS-Based RIP Protocol

- NX-OS RIP network command has been replaced with interface mode RIP address-family configuration mode.
- NX-OS RIP commands can support multiple AF. With NX-OS RIP split horizon cannot be disabled. Enabling NX-OS RIP poison reverse process replaces split horizon functionality.

https://t.me/learningnets



Cisco NX-OS 4.0 BGP Implementation Characteristics

BGP itself is designed to integrate IPv4 and IPv6 protocol stacks. NX-OS 4.0 supports a wide range of debug filters to constrain debug output, similar to the OSPF protocol. Debug output can be filtered by the BGP neighbors, by the prefix, by the address family, by the VRFs and by the policy. The NX-OS BGP implementation performs recursive next-hop notification handling similar to Cisco IOS next-hop tracking. For example, an IGP indicates a change to a destination that represents a next-hop interface. The RIB has no means by which it could notify the BGP process that its next hop has changed, but BGP is capable of sending its path information as well as registering next-hops that must be monitored. When the next-hop has changed in the RIB FIB, the RIB can notify the BGP process so that the BPG process can choose the next best path, because the previous next-hop for a certain destination has disappeared.

The implementation of recursive next-hop notification is similar to Cisco IOS next-hop tracking, but operates slightly differently because the overall architecture is different. The RIB and the BGP share a client-server relationship. NX-OS supports multiple VRFs and BGP process controlled restarts, some of which are RFC-based. NX-OS BGP also supports route and AS number look-ups for a multicast source, discovery protocol, RPF, and trace route resolution. NX-OS BGP configuration uses a neighbor-centric configuration model. With NX-OS, BPG peer templates can be created to simplify the configuration of multiple neighbors.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary of NX-OS 4.0 BGP Supported Features

Some of the important things that are supported by NX-OS BGP are:

- BGP graceful restart.
- BGP version 4 (RFC 4271)
- Autonomous System Confederations (RFC 3065)
- Route Reflection (RFC 4456)
- Capabilities Negotiation (RFC 3392)
- BGP Communities Attribute (RFC 1997)
- MD5 Authentication (RFC 2385)
- Route Flap Damping (RFC 2439)
- MultiProtocol Extensions (RFC 2858)
- Route Aggregation (RFC 2519)
- Graceful Restart (RFC 4724)
- Controlled restart (IETF)
- MultiProtocol Extensions for IPv6 (RFC 2545)
- Bestpath transition avoidance (draft-ietf-idr-avoid-transition-05.txt)
- BGP4+ Peering using IPv6 Link-local Address (draft-kato-bgp-ipv6-link-local-00.txt)
- Dynamic Capability (draft-ietf-idr-dynamic-cap-03.txt)
- BGP-4 MIB (RFC 4273)
- Cisco MIB
- draft-ietf-idr-bgp4-mib-15.txt (Peer table objects)
- 4-Byte ASN Support (RFC 4893)
- BGP Cease Subcodes (RFC 4486)
- Integrated implementation for IPv4 and IPv6
- Attribute based prefix packing in update messages
- Neighbor policy evaluation optimizations
- Debug filters to constrain debug output (neighbor, prefix, address-family, VRF, policy filters)
- NetFlow support (BGP Next-hop, Peer-AS, Origin-AS)
- Recursive Next-Hop Notification handling (similar to Cisco IOS NHT)
- BGP support for multiple VRFs
- Prefix based peering - passive-mode
- Attribute based policy matches in BGP redistribution into IGPs
- Support for IGP overload setting
- Support route/ASN lookups for MSDP RPF, traceroute

https://t.me/learningnets

Summary of NX-OS 4.0 BGP Features Not Supported:

MPLS VPN is not supported and NX-OS 4.0 does not support what has been traditionally known as BGP peer groups. Instead of peer groups templates are configured for different sets of groups.

- MPLS VPNs
- QPPB (BGP QOS Policy Propagation)
- Local-AS
- ORFs (Prefix, AS-Path)
- 6PE
- Advertisement-interval
- Extended Communities
- Multi-Topology Routing (MTR)
- BGP Policy Accounting
- GTSM (RFC 3682)
- Peer-groups
- Update-groups

Differences between NX-OS BGP 4.0 Configuration and Cisco IOS

BGP configuration uses neighbor-centric configuration model and nested neighbor mode.

For example:

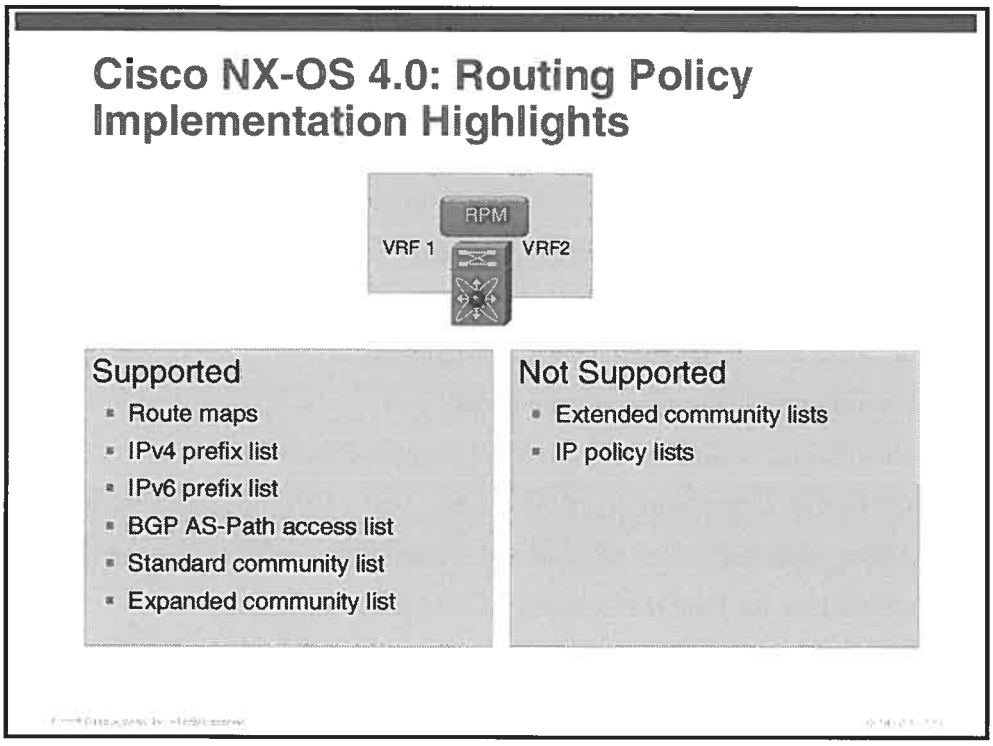
```
router bgp 1
      address-family ipv4 unicast
network 20.20.20.0/24
      neighbor 2.2.2.2 remote-as 2
ebgp-multihop 2
address-family ipv4 unicast
      soft-reconfiguration inbound
```

BGP NX-OS introduces three peer-template types for simpler configuration:

- Session and policy templates
- Neighbor template (similar to neighbor mode)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



NX-OS 4.0 contains a routing policy manager or RPM in the system that handles route policies when they are configured, and helps to apply them to routing protocols. The policy names configured must be strings for the backward compatibility. NX-OS also supports the 4 byte AS numbers. BGP regular expressions are supported by NX-OS routing policy implementation. When a regular expression in BGP is configured for the route policies, NX-OS will check if that expression triggers a CPU hog when the configuration is applied. The routing policy manager itself is a stateful high availability capable component and it does operate within a VDC. The routing policy manager also supports XML configuration for the policies. The BGP can be configured, or routing policies can be configured, using the XML interface. The route policy manager in NX-OS 4.0 support s the route maps, IP prefix lists, IPv6 and IPv4 prefix lists, BGP AS-Path ACLs, standard and expanded community lists, but not Extended community lists or IP policy lists. When configuring NX_OS policies, the routing policy names can be strings and not just numbers. ACLs are not supported for route filtering, redistribution or for route maps. The BGP community is shown using a new format: an AS number followed by a byte number. The BGP will support both decimal and hexadecimal notation for the entry of the new format. There is no limitation of 500 for the maximum number of configured AS-path access lists. Certain match and set commands are not supported in NX-OS implementation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary of the NX-OS 4.0 Routing Policy Implementation

- Routing policy manager (RPM) is a stateful high availability capable component
- Single instance of RPM supports policies for multiple VRFs
- Policy names are assumed to be strings (backward compatible)
- Prefix lists accept new "eq" option besides "le" and "ge" for mask lengths
- 4 byte AS support "match ip[v6] multicast" command for multicast policies
- Policy evaluation statistics in each routing protocol
- Forward reference supported for policies
- Policy change notifications are batched for protocol clients
- BGP regular expressions in policies undergo basic CPU hog check during configuration
- Caching optimization for AS-path and community regex evaluation in BGP policies
- Optimization for BGP prefix lookup during policy evaluation
- XML support for Policy configuration
- IPv4 and IPv6 support provided for policies

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

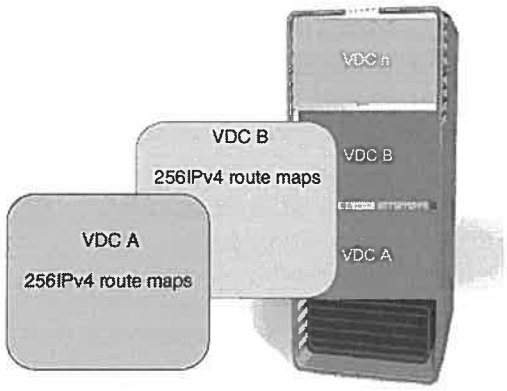
Cisco NX-OS 4.0 Routing Policy Features

Supported Features

- Route maps
- IPv4 prefix list
- IPv6 prefix list
- BGP AS-Path access list
- Standard community list
- Expanded community list

Not Supported Features

- Extended community lists
- IP policy lists



NX-OS supports Route-maps, IPv4 and IPv6 prefix lists, BGP AS-path access lists, standard community lists and expended community lists, whereas extended community lists and IP policy lists are not supported. By default, each VDC can use a maximum of 256 entries of IPv4 route map memory.

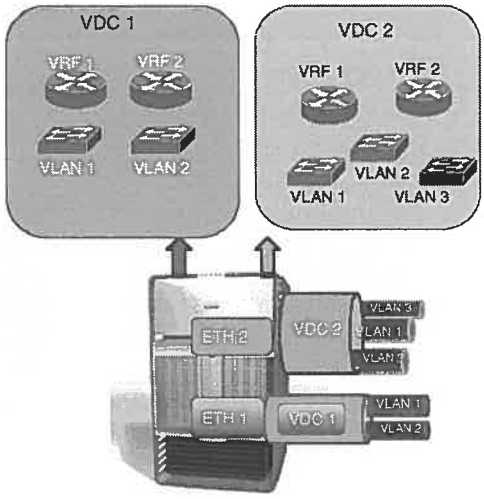
Differences between NX-OS and Cisco IOS Routing Policy Implementation

NX-OS routing policy names can be strings, not just numbers, as in Cisco IOS. ACLs are not supported for route filtering and redistribution. The BGP community is always shown in aa:nn format by NX-Os. Only a new format is supported for configuration that accepts decimal and hexidecimal notation. There is no limitation of 500 on maximum number of configured AS-path access lists. Regex is always shown with double quotes. Some match and set commands supported by Cisco IOS are not supported by NX-OS.

https://t.me/learningnets

Cisco NX-OS Virtualization Types

- Device Virtualization (VDC)
- Virtual Broadcast Domain (VLAN)
- Virtual Routing instance (VRF)
- Data plane virtualization
- Control plane virtualization
- Process virtualization, instantiation
- Interface virtualization (Logical subinterfaces)
- Management plane virtualization
- Hardware virtualization



NX-OS supports several levels of virtualization:

Device virtualization refers to the virtualization of a Nexus switch or the creation of logical devices within the physical switch. This includes virtualization of all processes and databases and interfaces within the devices.

Control plane virtualization refers to the protocols, databases and tables necessary to make forwarding decisions and maintain a functional network topology free of loops. A virtualized device must posses a unique picture of each virtual network it is to handle, hence the requirement to virtualize the control-plane components.

Interface virtualization includes creation of logical sub-interfaces on the top of physical interface, or creation of loopback interfaces.

Process virtualization includes virtualization or instantiation of process per each virtual device. For example, an OSPF process can be running per every VDC and those are referred to as virtualized OSPF processes. Nexus 7010 Series switch supports VRF virtualization that delimits the domain of an IP network within the router. VRF consists of an IP routing table, a derived forwarding table, a set of interfaces that use the forwarding table and a set of rules and routing protocols that determines what goes into the forwarding table.

Data plane virtualization refers to all the processes and tables used to actually forward traffic. The forwarding plane builds forwarding tables based on information provided by the control plane. Similar to the control plane, each virtual network will have a unique forwarding table that needs to be virtualized.

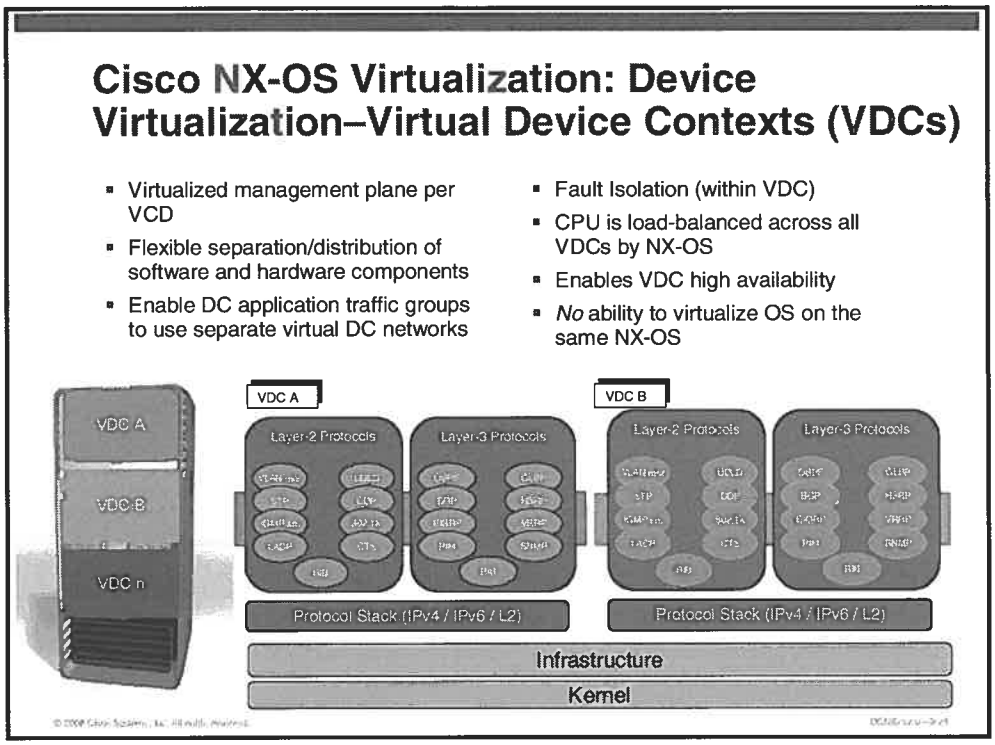
The control plane and forwarding planes can be virtualized at different levels which map directly to different layers of OSI model. For example a device can be VLAN aware and therefore virtualized at Layer 2 but yet have a single routing table, RIB and FIB, which means it is not virtualized at Layer 3.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

For example, sometimes virtualization at Layer 2 is enough. Providing firewall services virtualization requires virtualization at Layers 2, 3 and 4 and the ability to deploy independent management services for each virtual firewall. In that case VRFs within a single VDC would be deployed or different VDC can be deployed for every virtual firewall.

https://t.me/learningnets



NX-OS implements a logical virtualization at the device level known as Virtual Device Context (VDC) allowing multiple instances of the device to operate on the same physical switch simultaneously. Cisco NX-OS offers the capability to segment OS and hardware resources into virtual contexts that emulate virtual devices. Each VDC has its own software processes, dedicated hardware resources (interfaces), and independent management environment. VDCs are instrumental in the consolidation of separate networks onto a common infrastructure, maintaining the administrative boundary separation and fault isolation characteristics of physically separate networks while providing many of the operational cost benefits of a single infrastructure. A VDC virtualizes the device itself, presenting the physical switch as multiple logical devices. Each VDC is isolated from the others, which means that a problem in one VDC does not affect any other VDCs. VDCs are also secure from each other. Each VDC can contain its own unique and independent set of VLANs and VRFs. Each VDC can have assigned to it physical ports, thus allowing for the hardware data plane to be virtualized as well. Within each VDC, a separate management domain can manage the VDC itself, thus allowing the management plane itself to also be virtualized. The definition of a switch control plane includes all those software functions that are processed by the switch CPU (found on the central supervisor). The control plane supports a number of crucial software processes such as the routing information base, the running of various Layer 2 and Layer 3 protocols, and more. The control plane is also responsible for programming the data plane that enables many hardware accelerated features.

In its default state, the switch control plane runs a single device context (called VDC 1) within which it will run approximately 80 processes. Some of these processes can have other threads spawned, resulting in as many as 250 processes actively running on the system at a time depending on the services configured. This single device context has a number of Layer 2 and 3 services running on top of the infrastructure and kernel components of the OS. VDC enables data center design flexibility and enables creation and consolidation of virtual data center networks. Each VDC instance represents and virtualizes a Nexus 7010 Series Switch running its own set of routing, switching, management and security protocols. VDC also allows flexible separation and distribution of hardware resources. NX-OS assigns specific amounts of hardware resources to different VDCs running on the same Nexus 7010 Series hardware

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

platform to limit their ability to saturate the platform. VDC enables management differentiation. Different AAA servers can be assigned to each VDC.

On the other hand the VDC does not allow different OS to run on each virtual device. The software version is global. There is only one version of software on the platform, all the VDCs run the same version. Nexus 7010 Series Switch does not run different versions per each VDC currently. When NX-OS is being upgraded it upgrades the entire platform. The hardware is shared across the VDCs but from the user, configuration, and management perspective, the VDC should appear as a stand alone device

Each VDC is treated as stand alone device with limited resources. Each VDC is uniquely identified by ID or name and has a unique MAC address assigned to identify each VDC. VDCs share processors and line cards but have dedicated interfaces. The NX-OS software takes care of handling the load balancing of CPU resources among VDCs on the backplane. Per-VDC Role Based Management allows per-VDC administration configuration. Management VDCs enable software fault isolation for protocol processes within the VDC. When the system is activated for the first time, it will have a default VDC enabled—this VDC is present at all times during the operation of the switch and users cannot create or delete the default VDC 1.

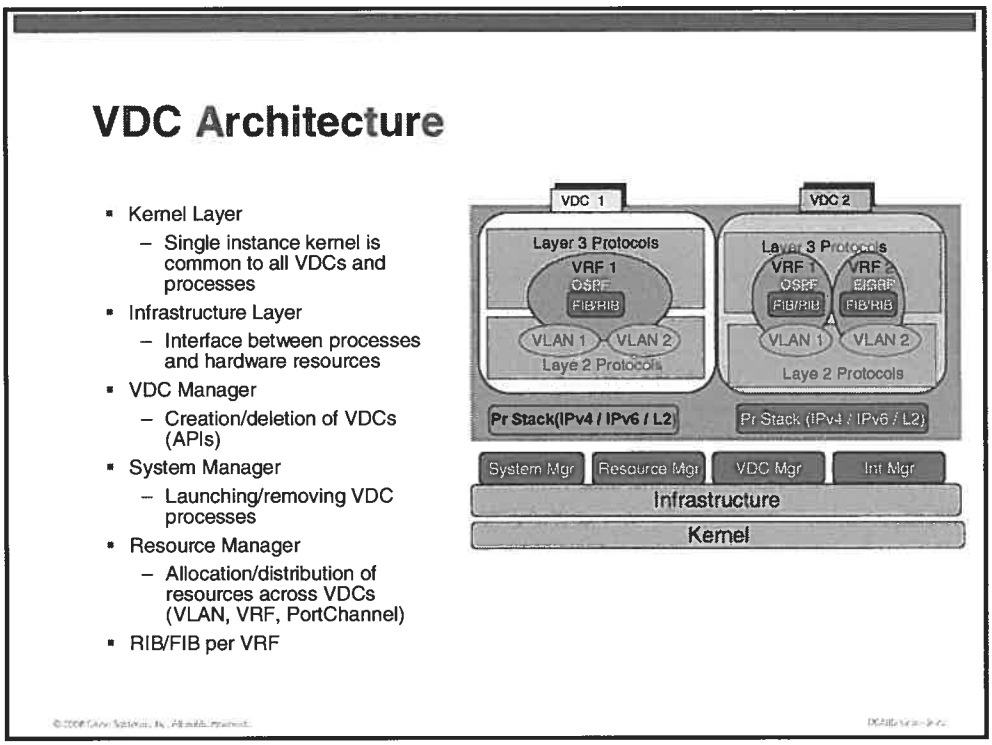
Fault Isolation

The modularity of the software provides resiliency: If the process fails, it should not affect other processes outside of the VDC. There is IPC inter process communication which has a limited boundary within the VDC. If something happens that affects one of the VDCs the IPC blocks the impact of the failed VDC.

VDC High Availability

Each VDC is assigned a default high availability policy that defines what the switch should do in the event of a VDC failure: a reset, a bringdown, or a restart option is possible. The restart option will delete the VDC and recreate it using the running configuration. Restarts will delete the VDC but does not bring VDC back online. The bringdown option issues a reset for the supervisor module if only a single supervisor is present. If dual supervisors are present a supervisor failover is initiated.

https://t.me/learningnets



At the heart of the NX-OS are the kernel layer and infrastructure layer. The kernel supports all processes and all VDCs that run on the switch but only a single instance of the kernel will exist at any one point in time. The infrastructure layer provides an interface between the higher layer processes and the hardware resources of the physical switch (TCAM). Having a single instance of this layer reduces complexity when managing the hardware resources. Having a single infrastructure layer also helps scale performance by avoiding duplication of the system management process. Working under control of the infrastructure layer reside system processes, which also exist as a unique entity.

Processes

Above the infrastructure layer and its associated managers are separate processes that run for each VDC. Each of these processes runs in its own protected memory space, which provides fault isolation. If a process crashes in one VDC, it does affect the processes running in another VDC.

VDC Manager

The VDC manager is responsible for the creation and deletion of VDCs and provides VDC-related APIs for other infrastructure components such as the system manager and resource manager to perform their own related functions. When a VDC is created, the system manager launches all services required for VDC startup that run on a per-VDC basis. When a VDC is deleted, then the system manager tears down all related processed for that VDC.

System Manager

When you create a VDC, the system manager launches all the services required for the VDC startup that exist only for the specific VDC. As you configure new features, the system manager launches the appropriate process. For example, if you enable OSPF in the VDC, the system manager launches an OSPF process for that VDC. If you delete the VDC, the system manager removes all processes related to that VDC.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Resource Manager

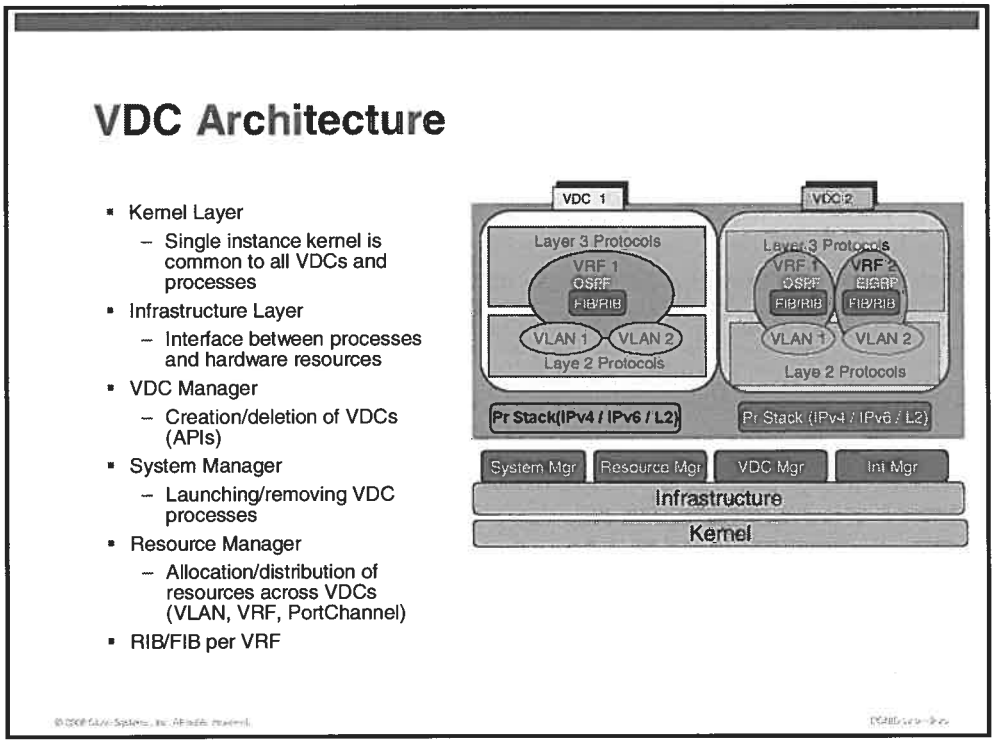
The resources manager is responsible for managing the allocation and distribution of resources between VDCs. Resources such as VLANs, VRFs, PortChannels, and physical ports are examples of resources that are managed by the resource manager.

Sitting above the infrastructure layer and its associated managers are processes that run on a per-VDC basis. Each of these processes runs in its own set of protected memory space. Fault isolation is one of the main benefits derived from the use of a VDC. Should a process fail in one VDC, it will not have an effect on processes running in another VDC.

Inter-VDC Communication

Inter-VDC communication is not facilitated from within the switch. A discrete external connection must be made between ports of different VDCs to allow communication between them.

https://t.me/learningnets



Cisco NX-OS further virtualizes each VDC to support virtual route forwarding (VRF) instances. VRF represents support for Layer 3 virtualization of forwarding and routing tables. VRFs are virtual routing and forwarding tables associated interfaces and routing processes. VRF can be thought of as a virtual routing instance.

The virtualization aspect of the Cisco NX-OS Software platform has been extended to support the notion of virtual device contexts (VDCs). Within that VDC it can contain its own unique and independent set of VLANs and VRFs. Each VDC can have assigned to it physical ports. Multiple VRFs can be enabled in each VDC. Each VRF contains a separate address space with unicast and multicast route tables for IPv4 and IPv6 and each VRF makes routing decisions independent of any other VRF. Each router has a default VRF and management VRF. All Layer 3 interfaces and routing protocols exist in the default VRF as long as it they are not assigned to another VRF.

VLANs are another important resource that has been extended in Cisco NX-OS. There are up to 16,384 VLANs, defined across multiple VDCs, supported in a Cisco Nexus 7000 Series Switch. Each VDC supports a maximum of 4096 VLANs as per the IEEE 802.1q standard. The new extended VLAN support will allow an incoming VLAN to be mapped to a per-VDC VLAN. We will refer to this per-VDC VLAN as a bridge domain for clarity. In this manner, a VDC administrator can create VLANs numbered anywhere in the 802.1q VLAN ID range, and the VDC will map that newly created VLAN to one of the default 16,384 bridge domains available in the switch. This will allows VDCs on the same physical switch to reuse VLAN IDs in their configurations, while at the OS level the VLAN is in fact a unique bridge domain within the context of the physical switch. For example, VDC A and VDC B could both create VLAN 100, which in turn could be mapped to bridge domains 250 and 251, respectively. Each administrator would use show commands that reference VLAN 100 to monitor and manage that VLAN.

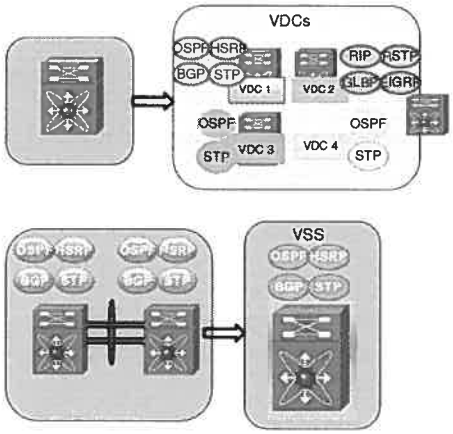
Note If there is VLAN 100 defined in VDC1 then VLAN-100 can be used again in VDC 2. Internally an identifier is created for that VLAN which is internally called a bridge domain. That is the reason there are 16,000 or 4,096 times 4 VLANs available.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS Virtual Switch (VSS) and Virtual Control Plane

- VSS enables Active-Active redundant topology:
  - Multiple switches act like one single switch or
  - Single switch acts like multiple virtual switches
- VSS represents alternative to L2 or L3 redundant design/topology:
  - No blocked ports/paths, no loops
  - Active-Active forwarding
  - Single management point and control plane



The Cisco Virtual Switching System (VSS) is a technology that supports a broad set of data center requirements. VSS is revolutionary in that through innovative enhancements, it provides new architectural options to a data center network environment not currently available. VSS is evolutionary in that it establishes a point of departure for next generation data center architectures from existing architectures, and helps set the direction of technology in support of critical business drivers. VSS technology is available on Cisco Catalyst 6500 and Cisco Nexus 7000 switching technology. The Cisco VSS essentially turns two switches into a single chassis by connecting the control processors together across the Virtual Switch Link (VSL).

VSS technology is characterized through two main functions: virtual switching and virtual port channels. Virtual switching refers to the ability to make multiple switches behave like a single switch (virtual switch), or to have a single switch behave as many switches (virtual switches). The figure presents a virtual switch instance made out of multiple (two) physical switches, and a single physical switch that supports multiple virtual switch instances.

Virtual switch VSS sees all ports, I/O modules and supervisor modules as being part of the same switch. Virtual switches VDC 1 through VDC 4 only see the physical resources assigned to each virtual switch and are unaware of the other virtual switching instances hosted in the same physical switch.

The figure also presents a view of how the control plane is handled on either kind of virtual switch. VSS switch 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. In this case two control planes are acting as one.

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

https://t.me/learningnets

There are inherent benefits to either approach of switch virtualization. On one hand a single logical instance simplifies the management of the network infrastructure and provides non-blocking Layer 2 paths to and from any device attached to VSS and on the other hand, multiple virtual switch instances on a single physical switch that provide control plane and physical topology isolation over a shared network infrastructure.

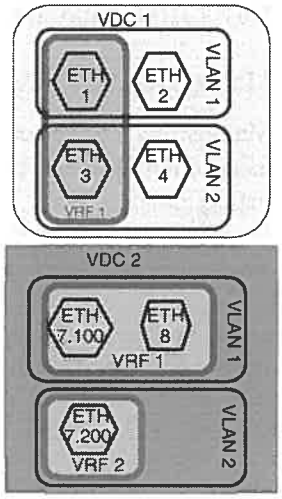
Layer 2 Topology was built by taking advantage of VSS technology that contains no loops, no blocking ports, and no blocked paths. The forwarding path is established by the VSS mechanisms, and because STP is not used to help determine the forwarding topology, it is used strictly as a fail-safe backup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

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS 4.0 Supported Interface Types

- NX-OS interface types
  - Physical
    - Ethernet (10/100/1000, 10 Gb/s), L2 and L3
  - Logical
    - Port-channel
    - Loopback
    - Null
    - SVIs (VLAN interfaces),
    - Tunnels(GRE, PortChannel) sub-interfaces,
    - VSL
  - In-Band
    - Sup-eth0, Sup-core0
  - Out-of-Band
    - Management 0 interface
- By default each interface belongs to VDC 1
  - 1/10 Gigabit Ethernet PortChannel will replace L2 and L3 EtherChannel



NX-OS v4.0 supports physically classical Ethernet (10/100/1000 Mb/s and 10 Gb/s) interfaces and logical interfaces (port channels, loopback interfaces, null interfaces, SVIs, tunnels, and sub-interfaces). In-band interfaces are controlled channel interfaces within the system such as supervisor interfaces and management interfaces, for example two CMP Ethernet interfaces. The physical ports for the management interface are assigned to the default VDC. Logical interfaces can be configured in every VDC. But management zero interface in each VDC has to have its own IP address. There is currently no support in NX-OSv4.0 for Fibre Channel or Data Center Ethernet (Enhanced Ethernet).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **Ethernet Interfaces**

  10/100/1000 Mb/s and 10 Gb/s Ethernet interfaces are supported (access and trunk mode).

- **Management Interface**

  Management Ethernet interface connects the device to a network for remote management using a telnet client, the Simple Network Management Protocol (SNMP), or other management agents.

- **Port Channel Interfaces**

  A port channel is a logical interface that is an aggregation of multiple physical interfaces. Eight links can bundle up into a port channel to improve network stability, bandwidth and redundancy. There are 1 G and 10 G port channels available, which means they aggregate capacity of 1 Gb/s or 10 Gb/s Ethernet links, which can be Layer 2 (access or trunk) or Layer 3 (routed or with sub interfaces). Up to 8 active member ports at maximum can create a single EtherChannel. Up to 256 EtherChannel interfaces are available per system. Port channel uses LACP for channel negotiation and a variety of load-sharing options:

  - Source, destination, source + destination MAC
  - Source, destination, source + destination IP (default)
  - Source, destination, source + destination TCP/UDP ports

  Load-sharing configuration is done per module.

- **Subinterfaces**

  Virtual subinterfaces can be created on a parent interface configured as a Layer 3 interface. A parent interface can be a physical port or a port channel. Subinterfaces divide the parent interface into two or more virtual interfaces on which you can assign unique Layer 3 parameters such as IP addresses and dynamic routing protocols.

- **SVI Interfaces**

  A switch virtual interface (SVI) is a virtual routed interface that connects a VLAN on the device to the Layer 3 router engine on the same device.

- **Loopback Interfaces**

  A virtual loopback interface is a virtual interface with a single endpoint that is always operational. Any packet transmitted over a virtual loopback interface is immediately received by the selfsame interface. Loopback interfaces emulate a physical interface.

- **Tunnel Interfaces**

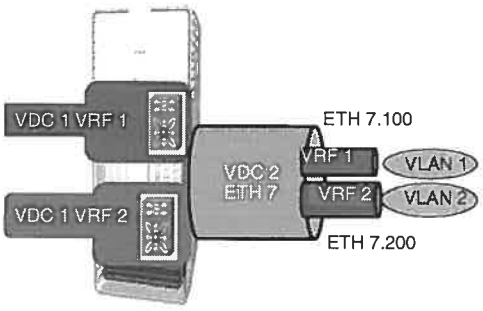
  Tunneling enables encapsulation of arbitrary packets inside a transport protocol. This feature is implemented as a virtual interface to provide a simple interface for configuration. The tunnel interface provides the services necessary to implement any standard point-to-point encapsulation scheme.

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS 4.0 Supported Interface Types (Cont.)

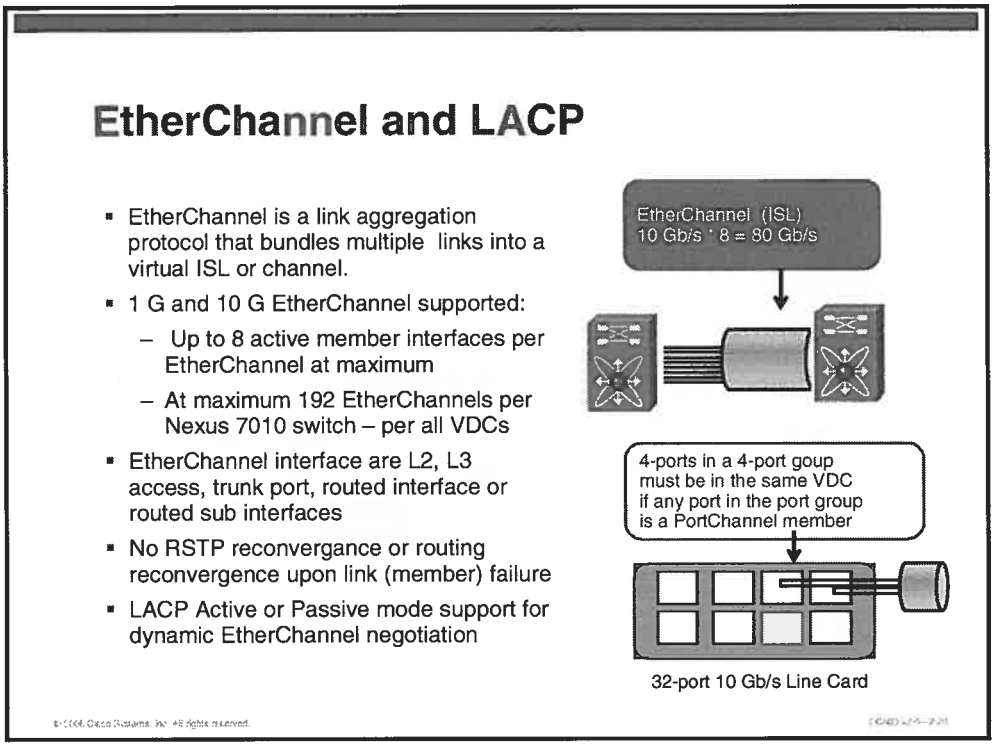
- Each physical port must be assigned to VDC
- All Ethernet interfaces are called "Ethernet"
- No VDC Trunk Link suported in this NX-OS release
- Logical interfaces of the same physical interface belong to the same VDC
- Logical interfaces of the same physical interface can belong to different VRF or VLAN within VRF
- Architected for 10 G now, 40 G, 100 G, and Unified I/O in the future



Physical switch port resources cannot be shared among VDCs. By default, all ports on the switch are assigned to the default VDC (VDC 1). Different port types can be assigned to a VDC. These include Layer 2 ports, Layer 3 ports, Layer 2 trunk ports, and PortChannel (Cisco EtherChannel) ports. Logical interfaces such as SVIs that are associated with the same physical interface cannot be allocated to different VDCs in the current implementation. Thus, it is not possible to virtualize a physical interface and associate the resulting logical interfaces to different VDCs. However, it is possible to virtualize a physical interface and associate the resulting logical interfaces with different VRFs or VLANs. Thus, VDCs can be assigned physical interfaces, while VLANs and VRFs can be assigned logical and physical interfaces.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



NX-OS 4.0 supports the creation of virtual ISLs called EtherChannel. EtherChannel—or the official IEEE term "LACP / 802.3ad"—is a link aggregation protocol that bundles multiple parallel links between a pair of devices, forming a single high-performance channel. Only an ISL with the same characteristics and configuration can be bundled into an EtherChannel. Configuration parameters, such as VDC membership, VLAN membership, speed parameter, etc., on all members of the EtherChannel must match. EtherChannel is a virtual ISL that aggregates the capacity of up to 8 physical ISLs. The primary reason for building EtherChannels is to achieve fabric resiliency to protect against topology/link failure. EtherChannel is represented as a single ISL in the OSPF routing topology database and, as such, does not trigger any topology change upon EtherChannel physical member failure as long as at least one physical EtherChannel ISL is operational. Similarly, RSTP will not trigger reconvergence when the EtherChannel member fails, as long as there is at least one member operational. EtherChannel represents a resilient and highly-available virtual ISL that offers high throughput—aggregated throughput of maximum 8 physical ISLS (all 1Gb/s ISLs or all 10 Gb/s ISLs).

For many of the resources that can be allocated on a per-VDC basis, it is important to note that the resource is typically managed globally. For example, there are 192 Cisco EtherChannel link bundles per chassis that can be apportioned across the active VDCs. If those 192 Cisco EtherChannel link bundles are consumed by two VDCs, then there are no more link bundles available for any other VDC in the chassis.

NX-OS Ether Channel hash is the same as on PFC-3. The load sharing decision is done in the Forwarding Engine. For EtherChannel, as a result of the hash algorithm, a 3-bit hash value is returned that picks one of the links in the EtherChannel. The individual member interfaces have one or more of the bucket values assigned to them.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

EtherChannel and LACP

NX-OS supports a variety of EtherChannel hash options for load balancing:

source-destination-ip (DEFAULT)	Source and Destination IP address
source-ip	Source IP address
destination-ip	Destination IP address
source-destination-port	Source and Destination TCP/UDP port
source-port	Source TCP/UDP port
destination-port	Destination TCP/UDP port
source-destination-mac	Destination TCP/UDP port
source-mac	Source MAC address
destination-mac	Destination MAC address

Load balancing is based on the following hash options, which result in different granularity of load-balanced traffic. Optimal selection of hashing option should be done based on the traffic type.

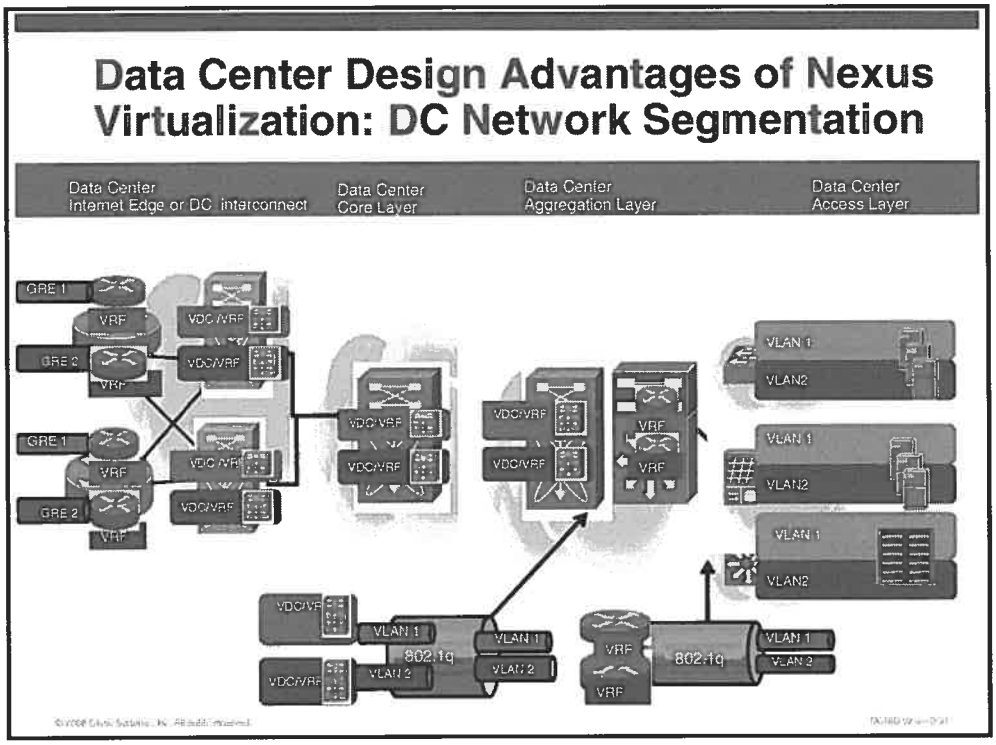
Default option: Source and Destination IP address

- Source IP address
- Destination IP address
- Source and Destination TCP/UDP port
- Source TCP/UDP port
- Destination TCP/UDP port
- Destination TCP/UDP port
- Source MAC address
- Destination MAC address

Note The port cost for port-channel in NX-OS implementation is the sum of configured port-channel member ports, whereas in Cisco IOS the port-channel cost is the sum of operational port-channel member ports. Consequently the NX-OS does not trigger RSTP reconfiguration upon port-channel link failure.

Note On the 32-port 10 Gb/s Ethernet module, you must assign all four interfaces in a port group to the same VDC if you want to use dedicated rate mode or include the interfaces into a port channel.

https://t.me/learningnets



The enterprise campus or data center network is characterized by an abundance of resilient high bandwidth links. The possibility of deploying multiple VRFs drives a highly resilient network design that involves having multiple end-devices connecting each subnet to the routed core across a separate virtual network. Deployment of VRF might require VLAN to VRF mapping that is usually done at the distribution layer switches. Different instances of VRFs from different physical switches can be interconnected using virtual networks or virtual data paths. This is one of the options to keep the traffic from different application groups separate across all data center architecture. The figure represents the logical data center networks on the top of physical data center network architecture. VDC and VRF enable creation of separated traffic groups using separate virtualized routing, networking and data path infrastructure. Separate dedicated application, security and management policies and services can be applied to each VLANs mapped to the virtual networks. The mappings among VLAN, VRFs and VDCs can be done in aggregation layer and carried across the data center core layer and across WAN to the remote department using GRE tunnels per each VRF. Layer 3 virtualization makes a data center networking infrastructure more resilient, enables fault isolation and enables sophisticated treatment of different application traffic groups.

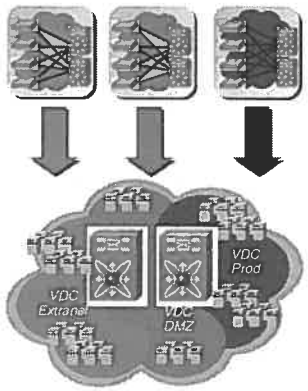
If data center design has to support inter-VRF routing between red and blue VDCs or red and blue VRFs, then red and blue VRFs must be deployed within the same VDC. An alternative is to keep blue and red VRF in different VDCs and to establish inter-VDC connection by connecting two physical interfaces in the respective red and blue VDC using a physical optical ISL with both ends connected to the same physical switch.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

DC Design Advantages of Nexus Virtualization: Consolidated DC Networks Using VDCs

- VDC enables virtualization of Nexus switch on the common Nexus 7010 Series Switch hardware
- VDCs enable consolidation of several data center networks on the common physical infrastructure
- Every virtual VDC of virtual data center network can use different routing and switching, security, management and application set of services
- Each VDC can participate in a separate data center department



Physical network islands are *virtualized* onto common data center networking infrastructure

Data center network design currently uses a siloed or stovepipe approach to the networks. They build separate physical networks with separate devices managed by separate domains or separate organizations in order to preserve some form of organizational characteristics or organizational requirements on their network.

The idea behind Virtual Device Context (VDC) on NX-OS is to enable consolidation of different data center networks on the common converged data center networking infrastructure. Every single VDC is an individual virtualized Nexus platform.

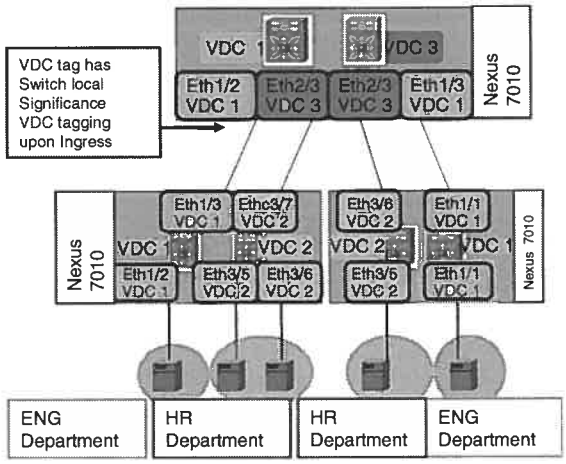
NX-OS enables deployment of several different (virtual) Nexus platforms on the common physical infrastructure in a very efficient manner. Each instance of VDC running on the common Nexus 7010 hardware can participate in a different data center network department with separated dedicated L2, L3 and upper-layer protocols, application services, security and management functions. Data center departments can be built based on the application being supported (Web, E-commerce, SQL database, Synchronous replication); based on protocols they are transporting (LAN, WAN, METRO, SAN), based on enterprise departments (Engineering, HR, Financial) or based on other classification.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Design Advantages of Nexus Virtualization: Other Virtualization Benefits

- Seamless data center network scalability
- Fault isolation per VDC
- Efficient resource allocation
- Data center network condolidatoin/segmentation
- Data center Service differentiation per VDC or virtual data center network
- Converged data center networking
- Data center network resiliency



There are several benefits and reasons for deploying VDC in the data center. VDCs enable efficient computing resource utilization, centralized management of many data center departments, segmentation of data center network infrastructure into several virtual data center networks, data center network consolidation, and deployment of different security policies, management, routing and switching functions per each VDC. VDCs enable future deployment of a converged data center infrastructure consisting of LAN network, WAN network and FC-SAN network all carried across the common hardware and software infrastructure. VDC enable seamless data center scalability. An administrator can deploy a new data center network department on the top of existing data center physical infrastructure by enabling one or more new VDCs on different Nexus devices. All VDCs enabled on the same Nexus 7010 series switch share common hardware resources. A VDC design was made to move as close to true virtualization as possible in order to gain as much fault isolation as possible so that there are as little cross-domain fault impact as possible.

Fault Isolation

When multiple VDCs are created in a physical switch, inherently the architecture of the VDC provides a means to prevent failures within that VDC from affecting other VDCs. So, for instance, a spanning tree recalculation that might be started in one VDC is not going to affect the spanning tree domains of other VDCs in the same physical chassis. An OSPF process crash is another example where the fault is isolated locally to that VDC. Process isolation within a VDC thus plays an important role in fault isolation and serves as a major benefit for organizations that embrace the VDC concept.

The figure represents a fault in a process running in VDC 1 that does not affect any of the running processes in the other VDCs. Other equivalent processes will continue to run uninhibited by any problems associated with the faulty running process.

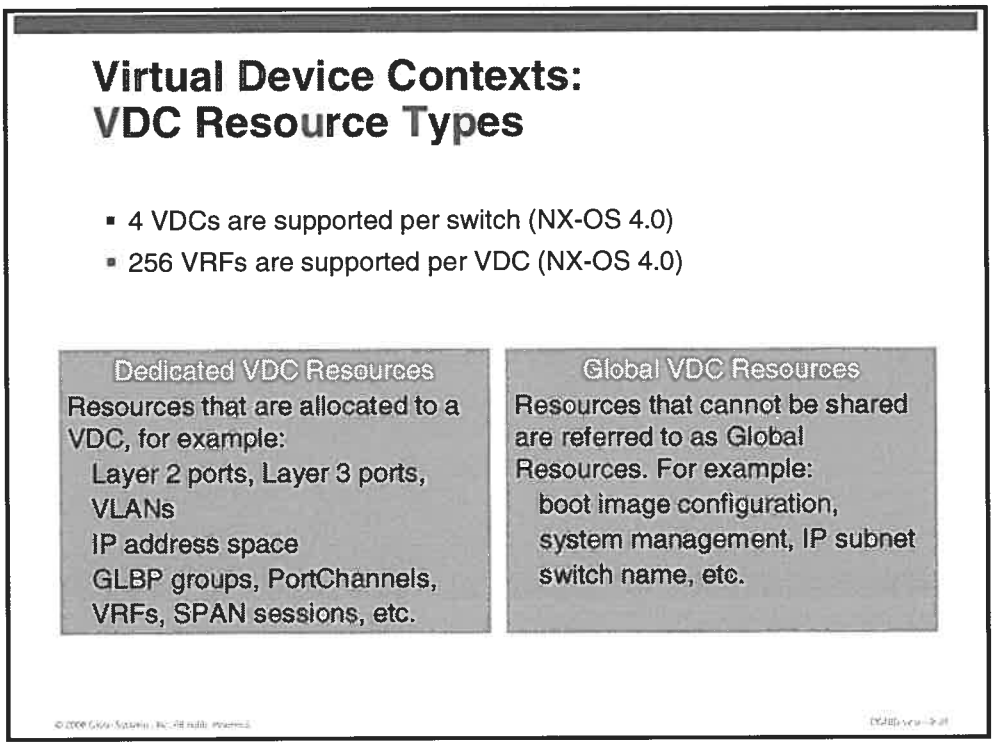
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Fault isolation is enhanced with the ability to provide per-VDC debug commands. Per-VDC logging of messages via syslog is also another important characteristic of the VDC fault isolation capabilities. When combined, these two features provide a powerful tool for administrators in locate problems. The creation of multiple VDCs also permits configuration isolation. Each VDC has its own unique configuration file that is stored separately in NVRAM. There are a number of resources in each VDC whose associated numbers and IDs can overlap between multiple VDCs without having an effect on another VDCs configuration. For example, the same VRF IDs, PortChannel numbers, VLAN IDs, and management IP address can exist on multiple VDCs. More important, configuration separation in this manner not only secures configurations between VDCs but also isolates a VDC from being affected by an erroneous configuration change in another VDC.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



When a VDC is created, selected switch resources can be allocated to that VDC, helping ensure that it has exclusive use of that resource. A resource template controls the allocation of resources and defines how much of that resource can be allocated to a VDC. VDC administrators and users within the VDC cannot change this template. Only the super user is able to change this template. The super user is a user with the highest level of authority to invoke changes to the configuration in the switch. The super user exists within the domain of VDC 1 (the default VDC). Aside from being able to assign physical switch resources between device contexts, the super user has the ability to invoke change in any VDC configuration, create and delete VDCs, and create and delete administrators and users for each VDC. The template is administered separately from the switch configuration, which allows the super user to edit a template without affecting the resource allocation for the VDC that the template has previously been applied to. When the super user wants to apply an updated template to a VDC, that person will have to reapply the template so that it is activated, copied, and merged with the running configuration. For transparent backup purposes, merging templates with the configuration adds the benefit of being able to back up all configurations specific to a VDC by simply backing up the primary configuration. Most, but not all, of the switch resources can be allocated to a VDC.

When creating VDCs, certain resources are shared across VDCs while others must be dedicated to a VDC. There are three types of resources to be found in the whole VDC model:

- **Global resources** are resources that cannot be shared, such as power supplies, backplane, spine cards, and the image file that we use.
- **Dedicated resources** are resources that are allocated to the VDC or ports, mainly. VDC resource allocation mainly focuses on VLAN, SPAN, VRFs, GLBP group and Port Channels. Other resources that are allocated to a VDC are Layer 2 and Layer 3 ports, VLANs, IP address space, etc.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

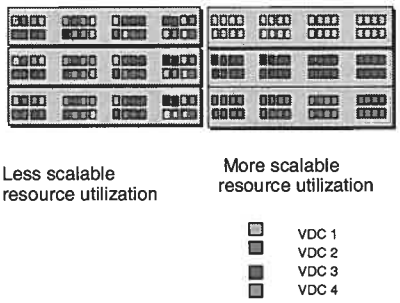
- **Shared resources** are a future option under consideration, which will support sharing resources among VDCs shared trunk port carrying traffic from multiple VDCs. The first thing that could be provided later on is a tagging mechanism to actually differentiate between traffic. To differentiate traffic in one VDC versus the traffic in another VDC, a trunk must be deployed. VDC trunks. Each VDC should be able to carry all 4,096 VLANs in 802.1q space and deploy SLA or QoS on the top of it.

Note When VDCs are created, is not only that CPU and memory that get separated and fault is isolated, but also separate SLAs should be applied to this model. So the initial SLA requirement is to actually have the separate physical interfaces assigned to each VDC, because this is the only way to dedicate a buffer and actually guarantee an SLA to the different VDCs. This must be kept in mind when we design these networks; when the SLAs are to be guaranteed then different interfaces must be assigned to different VDCs.

https://t.me/learningnets

VDC Efficient, Scalable L2 and L3 FE Resource Utilization

- MAC table (TCAM), FIB TCAM and ACL TCAM resource utilization is more scalable when ports on different line cards belong to different VDCs
- When a line card supports ports in many VDCs the MAC, FIB and ACL TCAM utilization is not as scalable as it is in the upper case.



System Capacity	1 VDC on 1 I/O module	1 VDC on 8 I/O modules	8 VDCs, 1 VDC per I/O module
FIB TCAM	128K	128K	1M (8 * 128K)
MAC table	128K	128K	1M (8 * 128K)
Classification TCAM	64K	64K	512K (8 * 64K)
Policers	16K	16K	128K (8 * 16K)

© 2008 Cisco Systems, Inc. All rights reserved.

Efficient Layer 2 MAC Address Table Utilization

The forwarding engine on each line card is responsible for Layer 2 address learning and maintains a local copy of the Layer 2 forwarding table. The MAC address table on each line card supports 128,000 MAC addresses. When a new MAC address is learned by a line card, it will forward a copy of that MAC address to other line cards. This enables the Layer 2 address learning process to be synchronized across line cards. Layer 2 learning is a VDC local process and as such has a direct effect on what addresses are placed into each line card. Distributed Layer 2 learning process is affected by the presence of VDCs. On the first line card, MAC address A is learned from port 1 in VDC 10. This address is installed in the local Layer 2 forwarding table of the first line card. The MAC address is then forwarded to two other line cards 2 and 3. As line card 3 has no ports in VDC 10, it will not install any MAC addresses learnt from the VDC 10 . Line card 2, however, does have a local port in VDC 10, so it will install MAC address A into its local forwarding tables. With this implementation of Layer 2 learning, the Cisco Nexus 7000 Series Switch offers a way to scale the use of the Layer 2 MAC address table more efficiently when VDCs are unique to line cards.

Efficient Layer 3 Forwarding Resources Utilization and Virtual Device Contexts

The forwarding engine on each line card supports 128,000 entries in the forwarding information base (used to store forwarding prefixes), 64,000 access control lists, and 512,000 ingress and 512,000 egress NetFlow entries. When the default VDC is the only active VDC, learnt routes and ACLs are loaded into each line card TCAM tables so that each line card has the necessary information local to it to make an informed forwarding decision.

When physical port resources are split between VDCs, then only the line cards that are associated with that VDC are required to store forwarding information and associated ACLs. In this way, the resources can be scaled beyond the default system limits seen in the preceding example. The effect of allocating a subset of ports to a given VDC results in the FIB and ACL TCAM for the respective line cards being primed with the forwarding information and ACLs for that VDC. This extends the use of those TCAM resources beyond the simple system limit described earlier.

https://t.me/learningnets

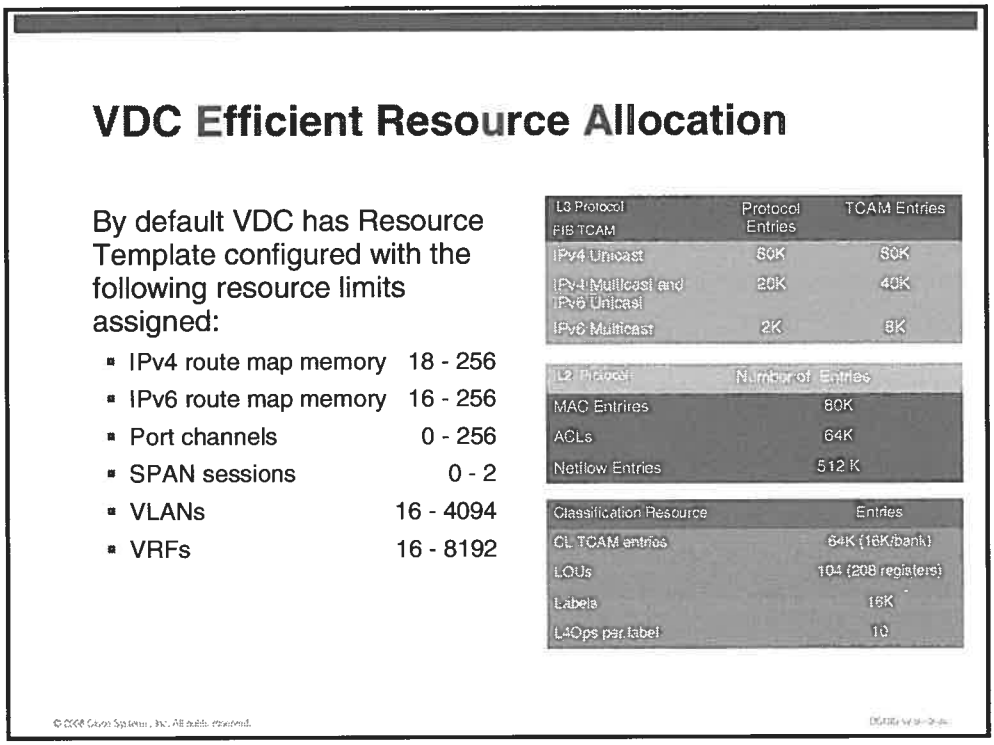
Example

In the preceding example, a total of 180,000 forwarding entries have been installed in a switch with a physical limit of 128,000 forwarding entries. Likewise, a total of 100,000 ACLs have been installed where a system limit of 64,000 Cisco ACE devices is in place. More important, FIB and ACL TCAM space on line cards 2, 5, and 7 is free for use by additional VDCs that might be created. This further extends the use of those resources well beyond the defined system limits noted here. As with the TCAMs for FIB and ACLs, the use of the NetFlow TCAM is more granular when multiple VDCs are active. Let us assume the same setup as before, where we have a set of line cards, each of which belong to a different VDC.

When a flow is identified, a flow record will be created on the local NetFlow TCAM resident on that line card. Both ingress and egress NetFlow are performed on the ingress line card so it is this ingress line card's NetFlow TCAM where the flow will be stored. The collection and export of flows is always done on a per-VDC basis. No flow in VDC X will be exported to a collector that is part of VDC Y. Once the flow is created in a NetFlow TCAM on line card X, it will not be replicated to NetFlow TCAMs on other line cards that are part of the same VDC. In this manner, the use of the TCAM is optimized.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



For many of the resources that can be allocated on a per-VDC basis, it is important to note that the resource is typically managed globally. For example, there are 256 Cisco EtherChannel® link bundles per chassis that can be apportioned across the active VDCs. If those 256 Cisco EtherChannel link bundles are consumed by two VDCs, then there are no more link bundles available for any other VDC in the chassis. Configuring the load balancing option for Cisco EtherChannel is a global configuration and cannot be set on a per-VDC basis. Another example of a globally-managed resource is SPAN. There are two SPAN sessions available for the switch. Both VDC A and VDC B can configure a SPAN session as "monitor session 1"; however, internal to the hardware they are recognized as different SPAN sessions. This is again as a direct result of SPAN being managed as a global resource. As with the link bundle example above, once these two sessions are consumed, there are no more SPAN sessions available for other VDCs.

The creation of a VDC builds a logical representation of the Cisco Nexus 7000 Series Switch, albeit initially with no physical interfaces assigned to it. Physical switch ports are resources that cannot be shared between VDCs. By default, all ports on the switch are assigned to the default VDC (VDC 1). When a new VDC is created, the super user is required to assign a set of physical ports from the default VDC to the newly created VDC, providing the new VDC with a means to communicate with other devices on the network. Once a physical port is assigned to a VDC, it is bound exclusively to that VDC, and no other VDC has access to that port. Inter-VDC communication is not facilitated from within the switch. A discrete external connection must be made between ports of different VDCs to allow communication between them.

Different port types can be assigned to a VDC. These include Layer 2 ports, Layer 3 ports, Layer 2 trunk ports, and PortChannel (Cisco EtherChannel) ports. Logical interfaces such as SVIs that are associated with the same physical interface cannot be allocated to different VDCs in the current implementation. Thus, it is not possible to virtualize a physical interface and associate the resulting logical interfaces to different VDCs.

However, it is possible to virtualize a physical interface and associate the resulting logical interfaces with different VRFs or VLANs. Thus, VDCs can be assigned physical interfaces,

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

while VLANs and VRFs can be assigned logical and physical interfaces. These default configuration parameters provide a definition on the resource consumption that the VDC is allowed to work within. These resources include VLAN, VRF, SPAN, PortChannels, and GLBP group IDs. Resource limit assignments can, however be reconfigured.

For example, Ethernet Ports 1 through 8 belongs to VDC 5 while ports 41 through 48 belong to VDC 10. Within each VDC, ports are further virtualized belonging to either a VLAN or VRF. After a port is allocated to a VDC by the root-admin at default-VDC level, it is up to the VDC administrator to manage (configure and use) the previously assigned port. Running other related commands such as a show interface command, for example, will allow the user to see only those interfaces that have been assigned to the VDC.

Each VDC maintains its own configuration file, reflecting the actual configuration of ports under the control of the VDC. In addition, the local configuration will contain any VDC specific configuration elements such as a VDC user role and the command scope allocated to that user. Having a separate configuration file per VDC also provides a level of security that protects this VDC from operation configuration changes that might be made on another VDC.

VLANs are another important resource that has been extended in Cisco NX-OS. Up to 16,384 VLANs, defined across multiple VDCs, are supported in a Cisco Nexus 7000 Series Switch. Each VDC supports a maximum of 4096 VLANs as per the IEEE 802.1q standard. The new extended VLAN support will allow an incoming VLAN to be mapped to a per-VDC VLAN. We will refer to this per-VDC VLAN as a bridge domain for clarity. In this manner, a VDC administrator can create VLANs numbered anywhere in the 802.1q VLAN ID range, and the VDC will map that newly created VLAN to one of the default 16,384 bridge domains available in the switch. This will allow VDCs on the same physical switch to reuse VLAN IDs in their configurations, while at the OS level the VLAN is in fact a unique bridge domain within the context of the physical switch. For example, VDC A and VDC B could both create VLAN 100, which in turn could be mapped to bridge domains 250 and 251, respectively. Each administrator would use show commands that reference VLAN 100 to monitor and manage that VLAN

The tables summarize the L2 and L3 VDC resources that are available by default:

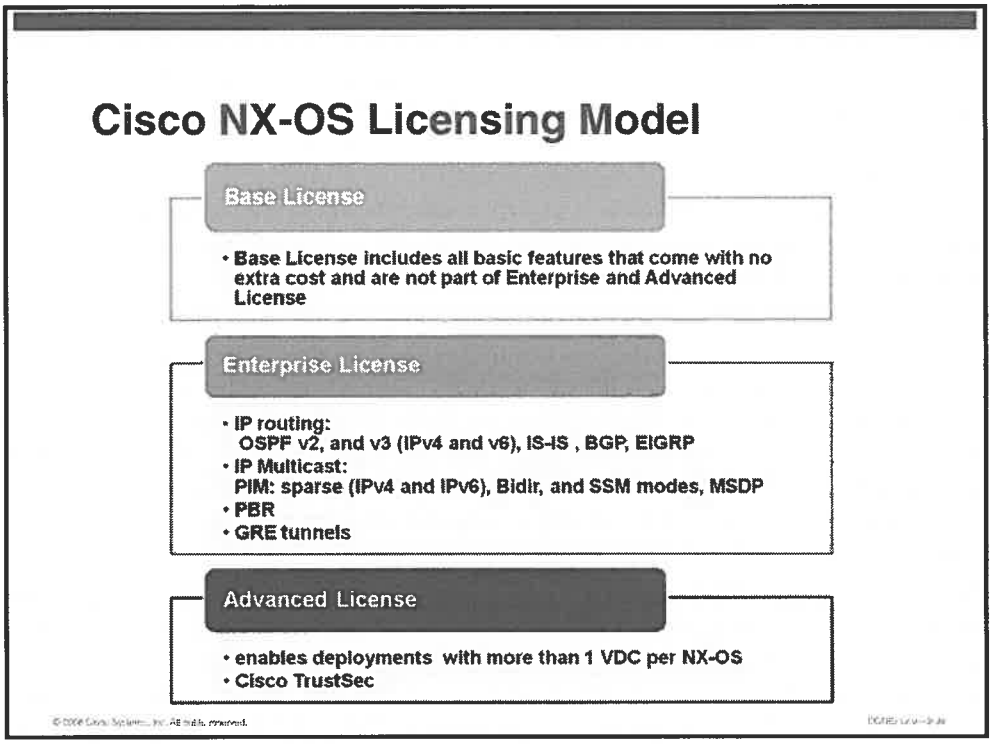
- The **Default VDC** resource template assigns the default number of PortChannels Route maps, VLANs and VRFs to VDC from the pool of all switch VLANs, VRFs, and PortChannels to the VDC enabled.
- The **FIB TCAM table** represents the number of IPv4 and IPv6 FIB TCAM entries available on the FE of each line card.
- The **L2 Protocol table** represents the number of ACLs, Netflow entries and MAC table entries available on each linecard.
- The **Packet Classification Resources table** summarizes the classification entries available on the FE of each line card.

The Nexus 7010 Series switch performs hardware-based packet classification for access control lists (ACLs) and Quality of Service (QoS). Two dual-bank CL TCAMs store entries in the hardware.

https://t.me/learningnets

Cisco Nexus 7010 NX-OS Software Licensing Model

This topic describes the NX-OS licensing characteristics and types of software packages available.



NX-OS licensing is associated to the switch serial number in the serial PROM (SPROM) and does *not* expire after initial purchase. The license key will be provided during license acquisition. This key must be entered or installed in the device to enable a feature set such as Advanced Services for the life of the switch.

There are three license feature sets available:

- IP Base (default)
- Enterprise Services (L3 protocols, multicast, etc)
- Advanced Services (VDC and CTS)

IP Base License (Default License)

The IP Base License is bundled with the hardware at no extra cost. The Enterprise License enables incremental functions applicable in many enterprise deployments, and the Advanced License enables next-generation functions, such as VDCs and Cisco TrustSec.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Enterprise License

The following functions are available only with the Enterprise license of Cisco NX-OS Release 4.0:

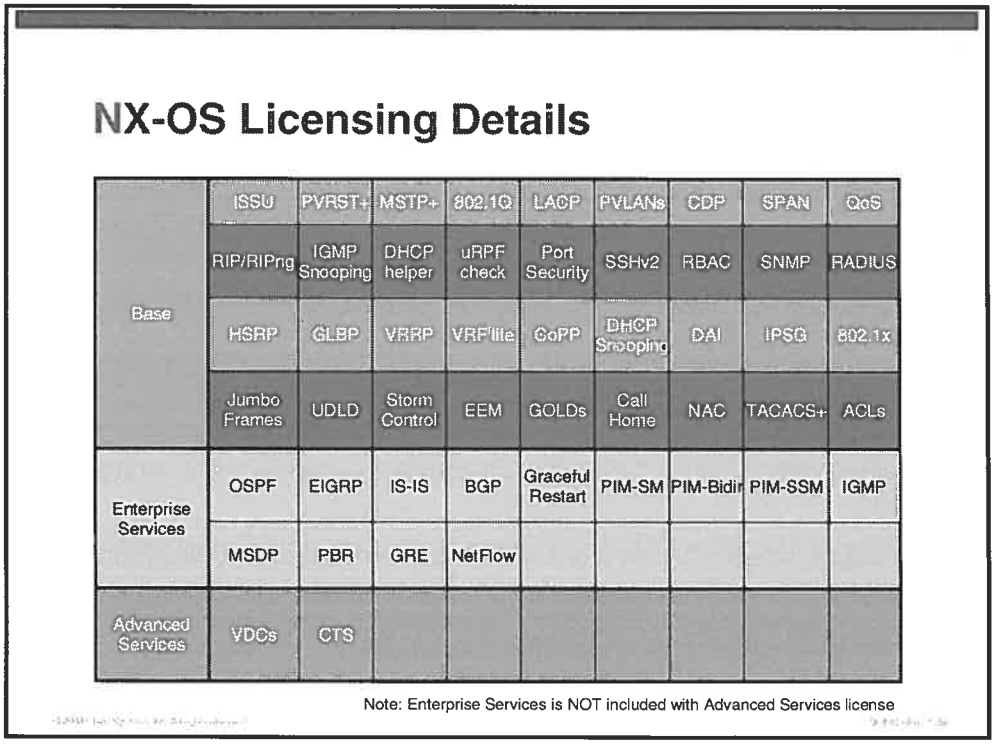
- IP routing and Layer 3 protocols
  - OSPF v2, and v3 (IPv4 and v6)
  - IS-IS
  - BGP
  - EIGRP
- IP Multicast
  - PIM: sparse (IPv4 and IPv6), Bidir, and SSM modes
  - MSDP
- PBR
- GRE tunnels

Advanced License

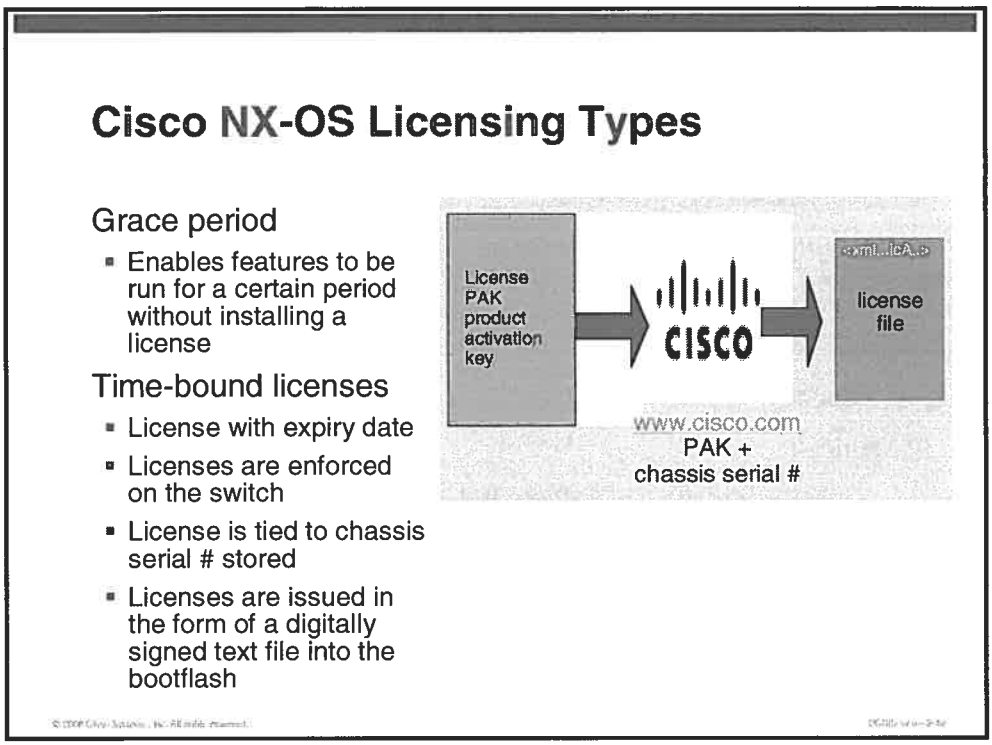
The Advanced license enables use of the following functions in Cisco NX-OS Release 4.0:

- VDCs
- Cisco TrustSec

Enterprise Services and Advanced Services feature sets are orthogonal.



https://t.me/learningnets



The following NX-OS Licensing types are supported:

Grace Period License

If a customer needs to upgrade to an Advanced Services feature set in order to leverage the CTS and VDC features, the switch will allow the new features to be active for a grace period of 120 days (after 60 days log message/SNMP alerts will be triggered to notify regarding the grace period).

- Enables features to be run for a certain period without installing a license
- Allows feature testing/trials without buying a license (e.g. 120 days)
- Periodic syslog, Callhome and SNMP traps warning when grace period nears expiry

Time-Bound License

- License with expiry date
- Currently used in SAN-OS as an emergency when grace period is over and the customer needs time to buy a license
- Expiry date is absolute (for example it expires at a certain time: midnight UTC on expiry date)
- Periodic Syslog, Callhome and SNMP traps warnings and notifications are generated when time bound license nears expiry
- After expiry date feature will continue to run if grace period has not been exhausted

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Summary

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

Summary

- NX-OS has been designed to support misson-critical converged data center networks.
- NX-OS is scalable and resilient software architecture that uses a combination of built-in virtualization and software modularity principles to achieve the highest availability and design flexibility.
- Modularity of NX-OS provides resiliency.
- NX-OS processes are stable, memory-protected and support stateful process restarts.
- NX-OS architecture offers continual availability, nonstop forwarding, and zero downtime.
- NX-OS supports nonstop forwarding for L2, L3 and storage protocol.

Summary (Cont.)

- NX-OS supports virtualization at almost every level: device level (VDC), at Layer 3 (VRF), at Layer 2 (VLAN), etc.
- VDCs provide logical separation between virtual device instances.
- NX-OS device virtualization enables efficient per-VDC resource allocation and efficient utilization.
- Up to 4 VDCs can be configured in a single chassis as of NX-OS v4.0.
- 256 VRFs are supported per VDC. (NX-OS 4.0)
- VDC 1 (default VDC) is always active and *cannot* be deleted.
- The default resource allocation for each VDC can be changed.

https://t.me/learningnets

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Lesson 3

Cisco Nexus 7010 Continual Availability

Overview

This lesson describes the modularity of the Nexus 7010 Switch hardware architecture for data center continual service availability with zero data loss.

Objectives

Upon completing this lesson, you will be able to describe the continual availability of the Cisco Nexus 7010 Switch. This ability includes being able to meet these objectives:

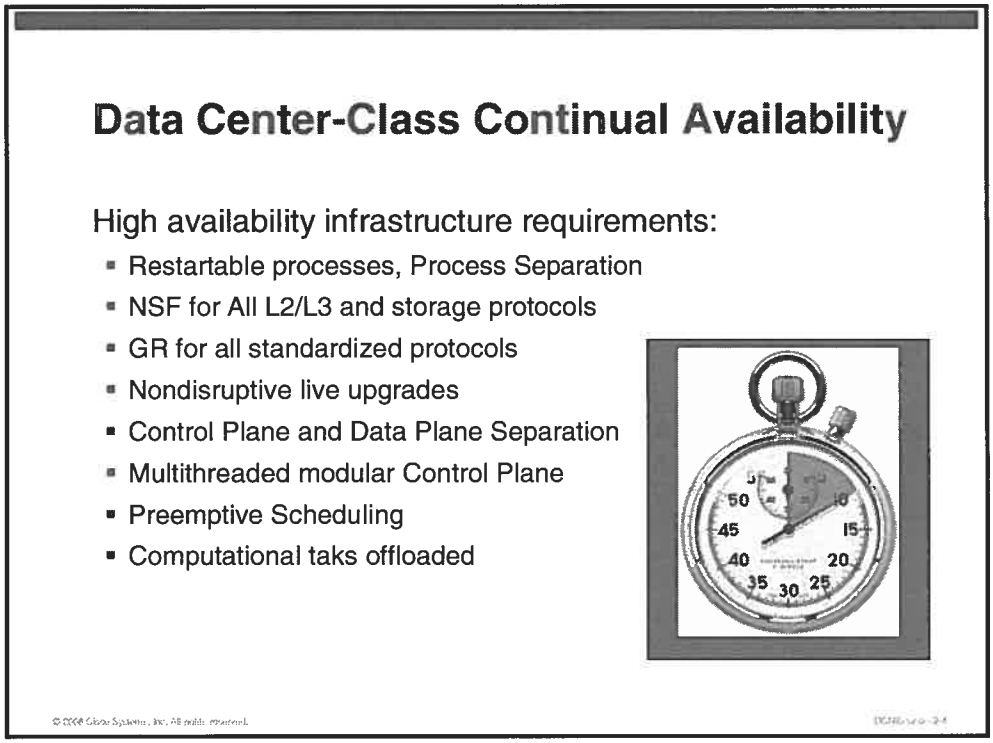
- Describe the characteristics of the high availability features of the NX-OS Layer 2, Layer 3 and routing protocols
- Describe the Nexus 7010 switch redundancy at the hardware, software and logical layer
- Describe the Nexus continual availability based on the platform modularity, virtualization and NSF abilities
- Describe the characteristics of the nonstop forwarding of the data plane traffic during the recovery of a hardware or software control plane failure, stateful supervisor switchover and stateful process/module restart
- Describe the non-disruptive software upgrades
- Describe the methods that provide stable, reliable and efficient memory-protected software processing,
- Describe Scalable Control Plane performance using distributed multiprocessing for fast reconvergence and lower MTTR and multi-threaded CPU

https://t.me/learningnets

Cisco Nexus 7010 Switch Continual Availability

This topic describes the characteristics of the Nexus 7010 Series Switch software and hardware modularity and the high-availability features that provide the data center service continual availability. The following Nexus 7010 Series Switch high availability and network resiliency features are discussed in detail throughout this lesson:

- Nonstop forwarding
- Nondistruptive software upgrades
- Isolated and memory-protected processes
- Stateful Supervisor Switchover
- Stateful/stateless process restarts
- Scalable control plane performance using distributed multiprocessing for fast reconvergence and lower MTTR
- Multithreaded CPU
- Redundant hardware architecture



The main objective in building a highly available data center network design is to avoid TCP session breakage while providing convergence that is unnoticeable, or as fast as possible. Each of the TCP/IP stacks that are built into the various operating systems has a different level of tolerance for determining when TCP will break a session. The least tolerant are the Windows Server and Windows XP client stacks, which have been determined to have approximately 9 second tolerance. Other TCP/IP stacks such as those found in Linux, HP, and IBM are more tolerant and have a longer window before tearing down a TCP session. This does not necessarily mean that the data center network should be designed to converge in less than 9

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

seconds, but it could serve as a guideline to a worst case. The most optimal acceptable failure convergence time is zero, of course. Although each network has its own particular convergence time requirements, many network designers usually break acceptable convergence times into the following categories:

- Minor Failures

  Failures that are expected to happen are configuration errors, link outages because of cable pulls, GBIC failure, and so on. All these failures might be considered minor failures. A minor failure convergence time is usually expected to be in the subsecond to 1 second range.

- Major Failures

  Any failure that can affect a large number of users or applications is considered a major failure. This type of failure could occur because of a power loss, supervisor failure, or module failure. This type of failure usually has a longer convergence time and is usually expected to be less than 3–5 seconds.

The following describes the various recovery times of the components in the Cisco Data Center design:

- **HSRP**: With recommended timers of the Hello timer = 1 second and the Holddown timer=3 seconds, convergence occurs in under 3 seconds. This can be adjusted down to sub-second values, but CPU load must be considered.
- **Routing Protocols**: OSPF and EIGRP can both achieve subsecond convergence time with recommended timer configurations. In general, OSPF convergence can take 6 seconds, whereas EIGRP is faster.
- **Rapid PVST+ Spanning Tree—802.1W**: Permits subsecond convergence time for minor failures when logical ports are under watermarks, and usually 1–2 seconds for major failure conditions.
- **FWSM**: Convergence time is approximately 3 seconds with recommended timers.

The Nexus 7010 Series Switch high availability functionality originates from the SAN-OS. In the SAN space where downtime is less tolerated, there is zero tolerance for the packet loss.

The Nexus 7010 Series Switch supports two cores. The system is designed to scale by leveraging the symmetrical multiprocessing CPUs. When the platform is heavily loaded, however, the preemptive multitasking helps the functions to operate. Modular code with real-time preemptive scheduling enables time-critical functions to run (for example routing protocol keepalives) even while the system is operating at 100% CPU load. Multithreaded and modular control plane enable control plane performance to scale and together with multicore CPUs provide:

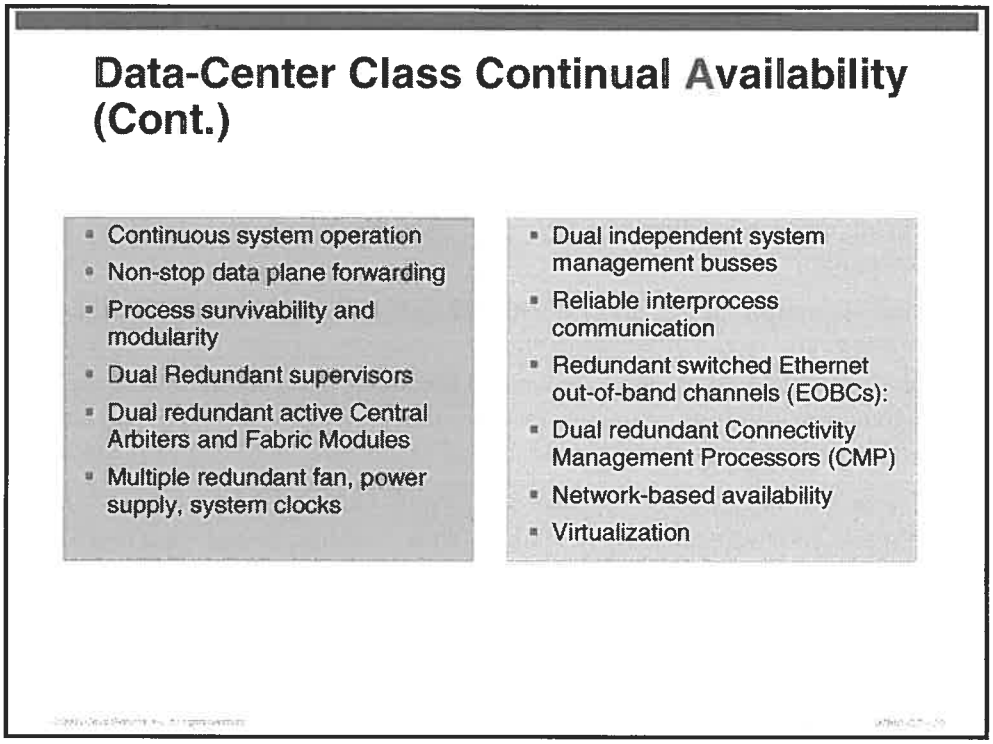
- Faster route reconvergence
- Lower mean-time-to-recovery

https://t.me/learningnets

This is one of those platforms where administrator is able to use the show, debug and other commands while the switch is forwarding the data, because the CPU processing will be preemptively interrupted in order to provide that CLI response. The majority of computationally intensive hardware or table programming tasks are offloaded from the supervisor CPU to the line card local control plane CPUs. The offloading processing provides some scalability in terms of leveraging the hardware resources. Offloading enables the system able to scale with the number of line cards in the system and enables optimal programming of finite hardware resources. There is a complete separation of the control plane and the data forwarding plane. Complete separation of control plane and data plane means there is no such thing as a software forwarding feature. The packets are hardware forwarded. Hardware has very modular code with real-time preemptive multitasking. The multitasking within the OS is real-time preemptive and it is not cooperative multitasking. If a particular process or service is attempting to seize the CPU, the process will get preempted based on the priority within the kernel. The kernel is actually doing the preemption and allows all the processes to be granted an equal processing time. This means that if one of the services freezes up, the routing protocol keepalives or CLI sessions can still function despite the fact that the CPU is running high. This separation allows the platform to continue forwarding while the control plane experiences a failure or upgrade. This separation provides the base for the nonstop forwarding capability and graceful process restart.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Nexus 7010 Series Switch supports the following high-availability features:

- **Continuous system operation**: Cisco NX-OS provides continuous system operation, permitting maintenance, upgrades, and software certification without service interruptions. The combination of process modularity, modular patching, Cisco In-Service Software Upgrade (ISSU) capability, and Nonstop Forwarding (NSF) graceful restart minimizes the effects of software upgrades and other operations.
- **Quick development of enhancements and problem fixes**: The modularity of Cisco NX-OS allows new features, enhancements, and problem fixes to be integrated into the software very quickly. Modular fixes can be developed, tested, and delivered in a very short time span, meeting urgent timelines. These updated images can then be installed without disruption using ISSU.
- **Process survivability and modularity**: Critical processes are run in protected memory space and independently from each other and the kernel, providing granular service isolation and fault containment and enabling modular patching and upgrading, and rapid restartability without disruption to the services. Individual processes can be restarted independently without loss of the process state information and without affecting data forwarding. After an upgrade or failure, processes restart in milliseconds without negatively affecting adjacent devices or services. Processes with large amounts of state such as IP routing protocols are restarted using standards-based NSF graceful restart mechanisms; other processes use a local Persistent Storage Service (PSS) and messages in MTS queue to maintain their state. Process stateful and graceful restartability enable non-disruptive software upgrades.

  Cisco ISSU provides the capability to perform transparent software upgrades on platforms with redundant supervisors, minimizing downtime and allowing customers to integrate the newest features and functions with little or no impact to network operation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- **Dual redundant supervisors**: The Nexus 7010 Series Switch supports the stateful supervisor failover. Redundant supervisors are kept synchronized at all times to enable rapid stateful supervisor failover. Sophisticated checks are in place to help ensure that the state is consistent and reliable throughout the entire distributed architecture after failover occurs.

  Among many other functions they are responsible for packet scheduling and fabric load-balancing.

- **Dual redundant active central arbiters and redundant fabric modules for fabric load-sharing**: The Nexus 7010 Series Switch supports redundant central arbiters controlling fabric capacity and granting the bandwidth among line cards. The unicast traffic is load-shared across all fabric modules installed in the switch when transiting from the ingress to the egress line cards, thus creating a highly-available switching infrastructure.

- **Redundant fans, power supplies and system clocks**: Cisco Nexus 7010 Series Switch supports hot-swappable and redundant fabric modules, power supplies, fans and redundant system clocks.

  Dual independent system management busses are used for the power and the fan temperature monitoring and management purposes.

- **Reliable interprocess communication**: Cisco NX-OS facilitates reliable communication between processes to help ensure that all messages are delivered and properly acted on during failure and adverse conditions. This communication helps ensure process synchronization and state consistency across processes that may be instantiated on processors distributed over multiple supervisors and I/O modules. Control-plane reliable IPC is also used for state synchronization across supervisors or multiples I/O modules.

- **Redundant switched Ethernet out-of-band channels (EOBCs)**: Cisco NX-OS can make full use of redundant EOBCs for communication between control and I/O module processors that provide secure out-of-band management access and standalone management kernel and CPU.

- **Dual redundant Connectivity Management Processors (CMP)**: A standalone Linux kernel with standalone CPU separated from NX-OS kernel provide secure out-of-band management access.

- **Network-based availability**: Network convergence is optimized by providing tools and functions to make the failover and fallback functions transparent and fast, fast routing and switching network convergence.

  Cisco NX-OS provides Spanning Tree Protocol (STP) enhancements such as bridge protocol data unit (BPDU) guard, loop guard, root guard, BPDU filters, and bridge assurance to guarantee the health of the Spanning Tree Protocol control plane. Other protocols that optimize network convergence are Unidirectional Link Detection (UDLD) Protocol, NSF graceful restart of routing protocols; millisecond timers for First-Hop Resiliency Protocols (FHRP), SPF optimizations such as link-state advertisement (LSA) pacing and incremental SPF and IEEE 802.3ad link aggregation with adjustable timers.

- **Virtualization**: Virtualization of networking components enables customers to build highly available and redundant network designs. NX-OS supports virtualization of switches, routers, control planes, interfaces, etc. If an unrecoverable error occurs in a VDC, the NX-OS software provides (high availability) policies that can be specified for each VDC. Other kinds of virtualization already discussed are supported by the Nexus 7010 Series Switch.

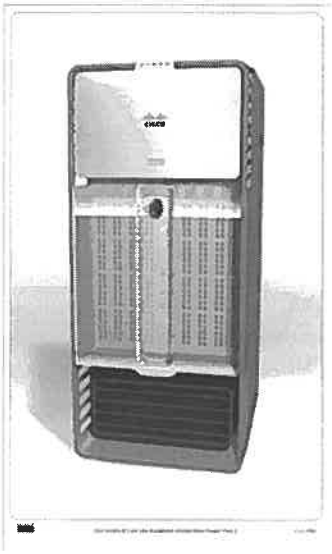
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data-Center Class Continual Availability (Cont.)

Nexus high availability is implemented at almost every level to deliver data center high availability with zero service disruption:

- Hardware level
- Software level
- Logical level



The Cisco Nexus 7000 platform is modular in its design, with an emphasis on redundant critical components throughout all of its subsystems. A highly modular and compartmentalized approach to the systems design is applied to all facets of the platform, spanning across the physical, environmental power, and system software aspects of its architecture. Additionally, a distinct functional separation between control plane and forwarding data plane is emphasized in its design in order to allow continuous operation and zero data loss during planned or unplanned control-plane events or failures. The Cisco Nexus 7010 Series Switch incorporates high availability features at the hardware, software and logical layer.

Hardware high availability features are mainly redundancy of hardware components that are hot-swappable and support seamless failover without impact on data center operation. Hardware system resiliency is based on redundant hardware components, supervisor redundancy, and dual supervisor modules running in active standby mode, helping to ensure operational continuity in the event of a supervisor failure. Switch fabric redundancy with up to five switch fabric modules can be deployed to provide N+1 redundancy of the fabric. The fabric modules can be added to or removed from the system without service disruption.

- **Redundant power supplies**: Three redundant power supplies supporting N+1 redundancy, input grid redundancy, and full redundancy provide variable usable power and load sharing.
- **Redundant out-of-band channels**: Gigabit Ethernet–based switched out-of-band channels between the supervisor and line cards enable independent control and management.
- **Connectivity Management Processor (CMP)**: A separate CPU complex on each supervisor module provides automated remote (lights out) management of the system through a dedicated 10/100/1000 Ethernet connection.

https://t.me/learningnets

Software high availability features offer multi-layered, modular, multi-faceted resiliency, minimizing the impact on the data plane should the control plane fail or a switchover occur. The different high availability service levels provide data plane protection, including service stateful process restarts, graceful restarts, stateful supervisor module switchovers, and In-Service Software Upgrades (ISSUs). Comprehensive management with integrated monitoring troubleshooting features are considered part of the Nexus 7010 Series Switch high availability features, because they enable prediction of malicious operation and potential failure. The prediction leads to prevention, fast detection or recovery from the network operation failure.

NX-OS implements a virtualization at several levels, for example at the device level, allowing multiple instances of the device to operate on the same physical switch simultaneously. These logical operating environments are known as virtual device contexts, or VDCs. Virtual device contexts provide logically separate device environments that can be independently configured and managed. This degree of isolation between virtual devices provides a very valuable level of fault isolation in addition to the security and administrative benefits. Human error or failure conditions due to configuration are naturally isolated within a given virtual device. Though the Nexus 7010 Series Switch virtualization features are primarily in function of logical segregation of physically separated resources, they also provide high availability to the data center networking infrastructure. Cisco NX-OS 4.0 allows virtualization of the control and management plane, creating separate operating environments for independent configuration and management, thus separating the fault domains to improve system availability and avoid cascading service disruptions.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Nexus 7010 Switch Software, Hardware and Kernel Failure Recovery

- Stateful/stateless process restart
  - is triggered upon a failed software process on the (same) active supervisor module
- If the process restart is not succesfull then System Manager can trigger:
  - SSO (dual supervisors) or
  - Reset of active supervisor (single supervisor)
- In case of a kernel process failure, SSO is triggered
- In case of a hardware failure, SSO is triggered
- Each service defines its own high availability policy
- Cascade high availability policies
  - If one fails (lightweight recoveries), the next high availability policy is tried
  - An example: System Manger fails to restart 3 times and tries stateless restart or SSO.

Control plane services will be restarted instantly when they crash or hang. The crash and restart of the service is transparent to other services. After the service restarts, the service recovers its states from the Persistent Storage Services (PSS) and resumes processing all the pending transactions from the Messaging Transaction Service (MTS) queue. If a service keeps crashing after multiple restarts, supervisor reset or switchover can be triggered based on the high availability policies. Process stack and core data can be collected for debugging purposes with an option to transfer core files to the remote location. The Cisco Nexus 7000 platform provides 1+1 redundant supervisor modules that can perform a Supervisor Switchover (SSO) in critical failure situations. The switchover of active to standby supervisor is non-disruptive to the forwarding plane and can therefore provide nonstop forwarding data plane traffic during the switchover. Supervisor module switchover is very fast and happens within in the range of miliseconds. Services checkpoints their runtime states to PSS database. System Manager monitors the health of services using heartbeats, and triggers restart of the process on the same supervisor module and reset of the only supervisor module or switchover to the standby in the presence of redundant supervisor switch configuration. A switchover may be triggered based on one or more criteria that contribute to the determination of an unrecoverable failure:

- A software process has failed and has been repeatedly restarted within the "restart timer" stability window on the active supervisor module without success. System Manager starts supervisor switchover.
- A hardware failure has occured. System Manager triggers switchover to the standby supervisor module.
- A kernel process has crashed. This event immediately triggers switchover to the standy supervisor module.

https://t.me/learningnets

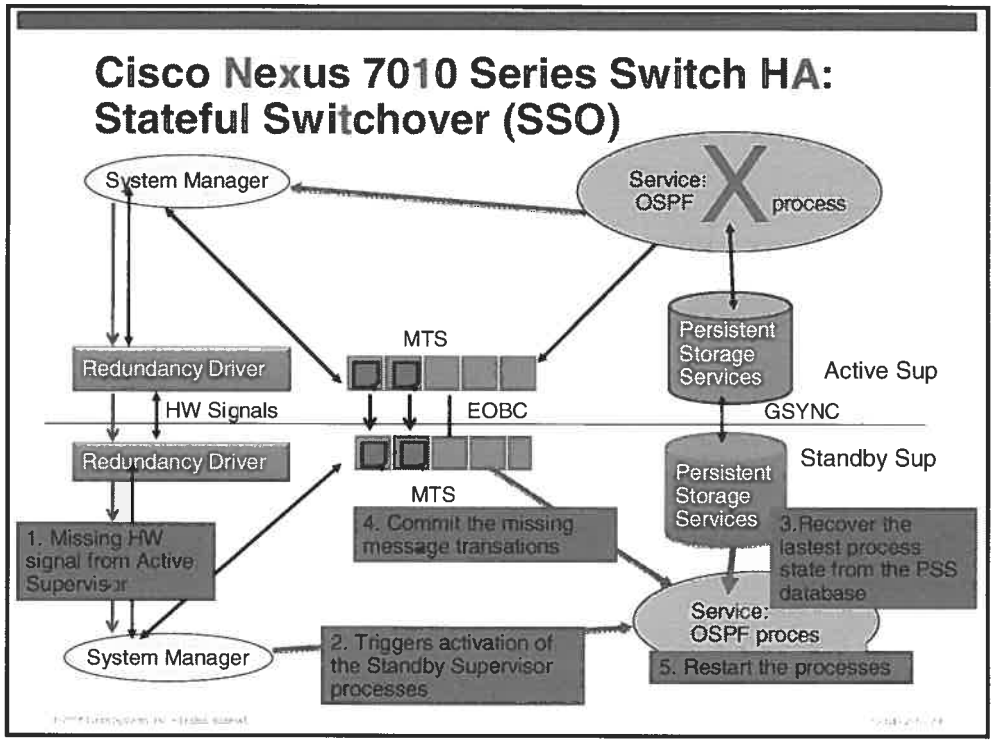
Not all software processes can support stateful process restart after supervisor module switchover. Stateful process restart uses the information from the PSS and MTS to recover the most recent process state. Some of those processes that do not support stateful restart, support nonstop forwarding (NSF) graceful restart or a fresh restart, instead. In both cases the data plane keeps forwarding the data (based on the old FIB, RIB, MIB, TCAM) while the control plane is restarting or during the SSO. The new incarnation of service restarted using fresh process restart loses the failed old process information state. The majority of routing protocols support NSF graceful restart to prevent route flapping and neighbor flapping scenario upon service fauilure. NSF graceful restart has similary minimal final impact on data routing and it has zero impact on the data plane traffic.

Summary of reaction to a process failure:

Failure Scenario	**Reaction**
Service/process exception	Service restart
Service/process crash	Service restart
Unresponsive service/process	Service restart
Repeated service failure	Supervisor reset (single)/switchover (dual)
Unresponsive system manager	Supervisor reset (single)/switchover (dual)
Supervisor hardware failure	Supervisor reset (single)/switchover (dual)
Kernel Crash	Supervisor reset(single)/switchover (dual)
Unresponsive System	Supervisor reset(single)/switchover (dual)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Nexus 7010 Series Switch stateful switchover (SSO) is supported with dual supervisors in the chassis. During the course of normal operation, the primary supervisor will constantly exchange and synchronize its process state with the redundant supervisor. A software process called watchdog is used to monitor the responsiveness of the active (primary) supervisor. Should the primary supervisor fail, a fast switchover is enacted by the system. Failover occurs at both the control plane and data plane layers. At supervisor switchover, the data plane continues to use the Layer 2 and Layer 3-derived forwarding entries simply by maintaining the state written into the hardware. For the control plane, the graceful restart process that is part of nonstop forwarding (NSF) standard is used to provide failover for Layer 3 processes. For Layer 2 processes, the control plane is maintained by locally stateful PSS mechanisms.

This process provides for the following:

- Uninterrupted data plane forwarding during a failover
- Rapid recovery from the failure to a stable operating state
- A nondisruptive recovery mechanism that will not render the network unstable during the recovery process

The SSO feature maintains stateful protocol and application information, user session information is maintained during a switchover, and line cards continue to forward network traffic with no loss of sessions, providing improved network availability. SSO provides a faster switchover relative to high system availability (HSA), Router Processor Redundancy (RPR), and RPR+ by fully initializing and fully configuring the standby supervisor module and by synchronizing state information, which can reduce the time required for routing protocols to converge. Initial process state synchronization (GSYNC) is done when application is started on the standby module. The application retrieves a snapshot of persistent context from the active peer which includes runtime configuration and debug configuration and the runtime database. MTS is informed when to start delivering synchronization messages. Standby service receives only updates not already included in the snapshot and no updates are missing that were not included in the snapshot. Subsequent synchronization of persistent context, persistent and log

https://t.me/learningnets

messages occurs via synchronized (MTS) events when active application sends, receives or drops messages. Synchronization can also occur via the PSS only (so-called slower recovery).

NX-OS leverages three key core infrastructure services to provide overall high availability functionality:

- System Manager
- Persistent Storage Service (PSS) database
- Message and Transaction Service (MTS) queue

System Manager

System Manager monitors services and system health, maintains overall high availability states and enforces high availability policies. System manger is semi-equivalent to the init daemon on the UNIX platform with the following functions:

- Spawns application process upon configuration (conditional)
- Identifies application crash failure
- Restarts failed process by spawning a new process
- Enforces high availability policies on the system

The System Manager orchestrates overall system function, service management, and system health monitoring. It is also responsible for maintaining overall high availability states, enforcing high availability policies, and managing system configuration. It is the System Manager that is responsible for launching, stopping, monitoring, and restarting services on active or standby supervisor module. System Manager is also leveraged for initiating and managing the syncing of service states and intersupervisor states for stateful switchover. It is also the System Manager that will initiate a Supervisor Switchover if it determines that the current supervisor has undergone an unrecoverable failure or if key core services are undergoing errors and cannot be restarted reliably. In addition, being the overall control and monitoring process, the System Manager is responsible for "punching" or triggering the keepalive indicator for the hardware-based watchdog timer on the supervisor. The absence of this periodic heartbeat from the System Manager within the keepalive timeout period of the watchdog timer will indicate a nonresponsive System manager, which in turn triggers a hardware-based supervisor reset (single supervisor) or switchover (dual supervisors). The reason the term services is used is because every process is associated with one instance of the running process id. Therefore System manger function is ready to kill those initial processes and to manage the link among the service to its processes. If a service does need to be restarted, the System Manager will find the right process associated with that particular service incidence and will kill it and restart it. System Manager is also watching service operations. Most of the other processes that were started by the System Manager run as child processes to the System Manager. When a given service, such as OSPF, exits, those exit codes and heartbeats, and other process outputs are sent back to the System Manager. System Manager is actually able to monitor the health of each of these services. System Manager is also responsible for restarting these processes and for enforcing high availability policies in the system. High availability policies are a set of configurations that come with services and dictate how the services should be treated in the event of a failure. High availability policies are hard coded into the system and are not configurable by the end-user. A high availability policy, for example, instructs the system to initiate a stateful restart three times when a service freezes and if it that is not successful then the policy triggers a stateless restart. System Manager is thus responsible for reading the policies and implementing them based on service state.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Persistent Storage Service (PSS) Database

Persistent Storage Services (PSS) store service configurations and runtime data. PSS provides reliable persistent storage used by components to 'checkpoint' their state and internal data structures for non-disruptive restart. PSS database facility supports key/value pair database, single database index and atomic read/write guarantee on a single transaction limits. PSS uses bootflash for underlying storage and has 40 MB of bootflash (1 GB). L3 routing protocols can have too much data that cannot fit in the PSS, hence those services typically use NSF graceful restart instead. PSS database provides an API to all other services to allow them to checkpoint their state. It takes in key/value pairs from the services and allows a service to write in its current runtime database or runtime state into its database. PSS allows the service to send its runtime configuration and the current state. Every time there is a change in the service state the service just checkpoints that change in the state to the PSS. The PSS is actually implemented in two areas: nonvolatile and volatile memory (RAM). The PSS provides an API call to allow servers to write to nonvolatile space when it has to survive a kernel restart or a supervisor restart. RAM space is taken up by PSS for things that can be written to a volatile space that need to be accessed quicker for faster reads and writes that are highly transaction intensive. PSS can leverage both either volatile or nonvolatile memory. A majority of state information is actually written to a volatile memory because it provides faster access to the data. Highly event-driven processes or high transaction rates are better suited towards volatile memory. The nonvolatile memory is actually used by a few processes for things that we know have to persist across kernel restarts or across supervisor restarts. The choice between volatile or nonvolatile memory is actually made by the service itself. When the service makes the call to write to PSS it calls a different API based on whether it requires volatile or nonvolatile memory. So the service developers themselves can determine which one is more applicable based on the service. PSS allows services to subsequently recover to the last known operating state preceding a failure, thereby allowing for a stateful restart. This state recovery capability is available to NX-OS services in both single- and dual-supervisor configurations, and helps enable a service to transparently return to operation with zero impact to data plane traffic, neighboring devices, or other internal services. For example, even in a single supervisor configuration, the PSS enables the stateful restart of a service such as spanning-tree without impacting the overall spanning-tree topology or stability.

Message and Transaction Service (MTS) Queue

Message and Transaction Services (MTS) are high-performance interprocess communications (IPCs) with high availability semantics like message persistency, event syncing, etc. MTS stores the transaction messages that cannot be processed by a failed process but will be used by the restarted process instance. MTS provides reliable interprocess communication such as:

- The support for unicast and multicast type messages
- The support for interprocess communication across different processors using AIPC and EOBC
- The support for both communication and notification signals

MTS is highly available because persistent messages survive in the receiver queue during a crash and clients do not see server crashes, as the receiving queue is always alive. MTS functionality is analogous to the commit and rollback type of RDBDS functionality. For example: when the OSPF process needs to talk to another process or service running, it will communicate to it via the MTS. There are a number of reasons for implementing the MTS instead of permitting direct interprocess communication. One of the reasons is that there is a structure that allows messages to be not unicast and multicast from one service to another service. A service can put out a message, destined to multiple destinations, to MTS and MTS

https://t.me/learningnets

ensures that message gets delivered to the respective destinations. The service itself is not responsible for the message routing. MTS runs on each CPU or local CPU. So there is an MTS process on the supervisor and then one on each line card as well. MTS provides buffered memory space for the messages in the receive queues buffered up for each service. This provides the ability to keep messages sent to the failed services in the MTS queues to prevent loosing them while the failed process is restarting. MTS buffers messages sent to the failed service and when they are restarted messages are still held in the MTS queues and can be delivered to the destination service. It is analogous to a relational database commit-and-rollback type functionality. MTS also enables communications of process across different line cards over the EOBC. Process from one line card sends a message destined to a process on another line card to the MTS and uses the IPC or AIPC stack to package that message and send it across the Ethernet out-of-band channel. For example, a message sent from the supervisor to a line card is first stored in the MTS and sent across the Ethernet out-of-band channel to the destination line card, and then off to the receiving MTS on that line card. MTS is monitored by the System Manager and when MTS fails it is restarted like any other service and can be recovered statefully. System Manager itself is monitored by a kernel module. There is a watchdog timer that runs on a kernel module that System Manager has to ping periodically. If the System Manager does not ping the watchdog and the kernel module times out it restarts System Manager or triggers a switchover. The System Manager process runs in the user space. The NX-OS Message and Transaction Service (MTS) is a high performance IPC message broker that specializes in high availability semantics. MTS handles message routing and queuing between services on and across line cards (including across all contents supervisors). MTS facilitates the exchange of messages such as event notification, synchronization, and message persistency between system services and system components. MTS also facilitates and manages message queuing between services and, as a result, can maintain persistent messages and logged messages in queues for access even after a service restart. MTS contains messages that were sent to routing or other process while processes failed and could not process the messages. Each of these system infrastructure component services plays a role in providing overall system availability. Individual functional system services (for example, authentication, authorization, and accounting (AAA), Spanning Tree Protocol (STP), Network Time Protocol (NTP), command parser, and so on) are responsible for using these system-level redundancy infrastructure components to checkpoint their own states, recover the states upon restart, and synchronize the states to their standby counterparts using the PSS and MTS facilities.

Services

Services checkpoints its own states, recovers the states upon restart from PSS and synchronizes the states to its standby counterpart. A process may be killed and restarted as another process with new process ID, but both processes are representing the same service. Each process instance can be treated as a separate service; a service is tied to many processes depending on how many times it has been restarted. In a redundant configuration, such as when dual supervisor modules are in operation, mirrored services run on each supervisor module, with configuration and operating state synchronized between them. One of those supervisors will operate as the active supervisor while the other operates in a standby facility until activated in a switchover.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Stateful Supervisor Switchover of the Nexus 7010 Series Switch includes the following steps:

Step 1 The System Manager detects a failure and triggers the switchover

Step 2 MTS of the active supervisor module is queuing the messages that were not processed by the failed protocols on the active supervisor module

Step 3 The active and standby MTS queues are synchronized across an Ethernet out-of-band channel (EOBC). The active and standby PSS databases are also synchronized using global synchronization (GSYNC). The active supervisor unit sends the current operating configuration and the process state to the standby PSS database.

Step 4 The standby System Manager triggers starting of the services on the standby module

Step 5 Process on the standby module first recovers the PSS database information, commits the pending transaction in the standby synchronized MTS queue and starts its operation

Step 6 Stateful L3 components handle Supervisor transition to the standby supervisor.

Step 7 Non-stateful L3 components get launched by the System Manager

Step 8 Routing protocols get restarted based on the CLI configuration

Step 9 L3 protocols reacquire state in one of the following two methods:

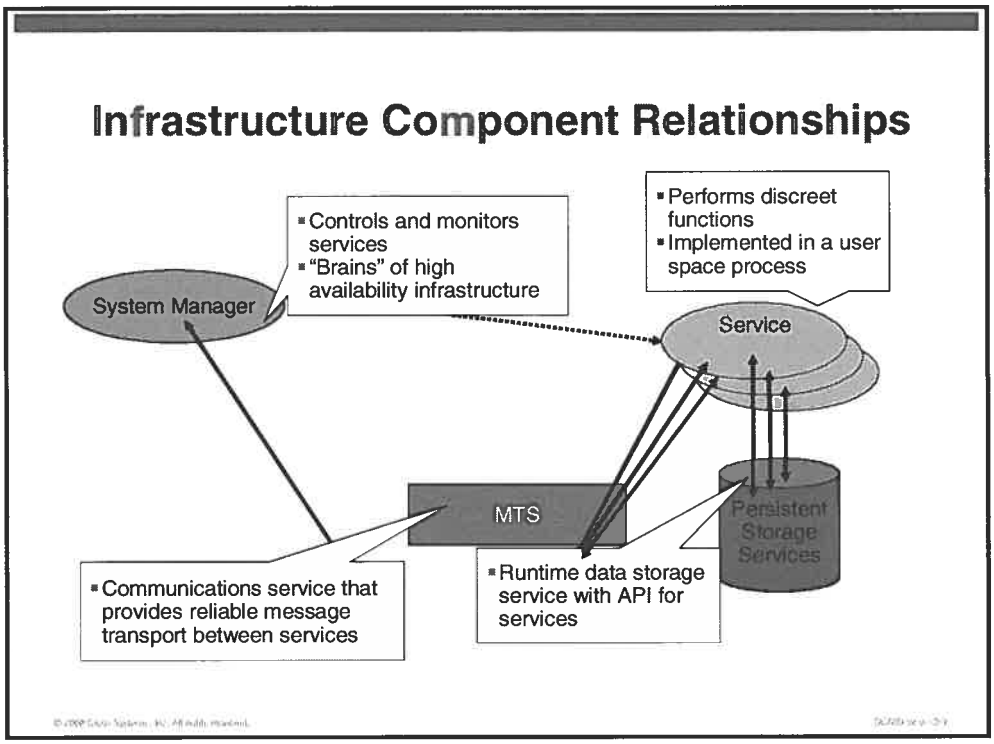
- Graceful restart-capable protocols acquire state from peers
- Rest rely on periodic protocol refresh

Step 10 Protocols add all the routes to RIB

Step 11 RIBs add all the routes to FIB on the I/O module Fast Ethernet

Step 12 FIB on the I/O module deletes the stale routes to reflect the new routing topology

https://t.me/learningnets



There are multiple services running on the Nexus 7010 Series Switch platform. Each of those services represents some type of functionality, such as a routing protocol or STP. The System Manager launches the services and monitors, controls their state and implements high availability policies.

Here is an example of a group of steps representing a failed process recovery:

Step 1 Service checkpoints state during normal operation.

Step 2 System Manager detects service failure (no heartbeat).

Step 3 System Manager restarts the service.

Step 4 The runtime database is populated: Process is reading the information from the persistent database.

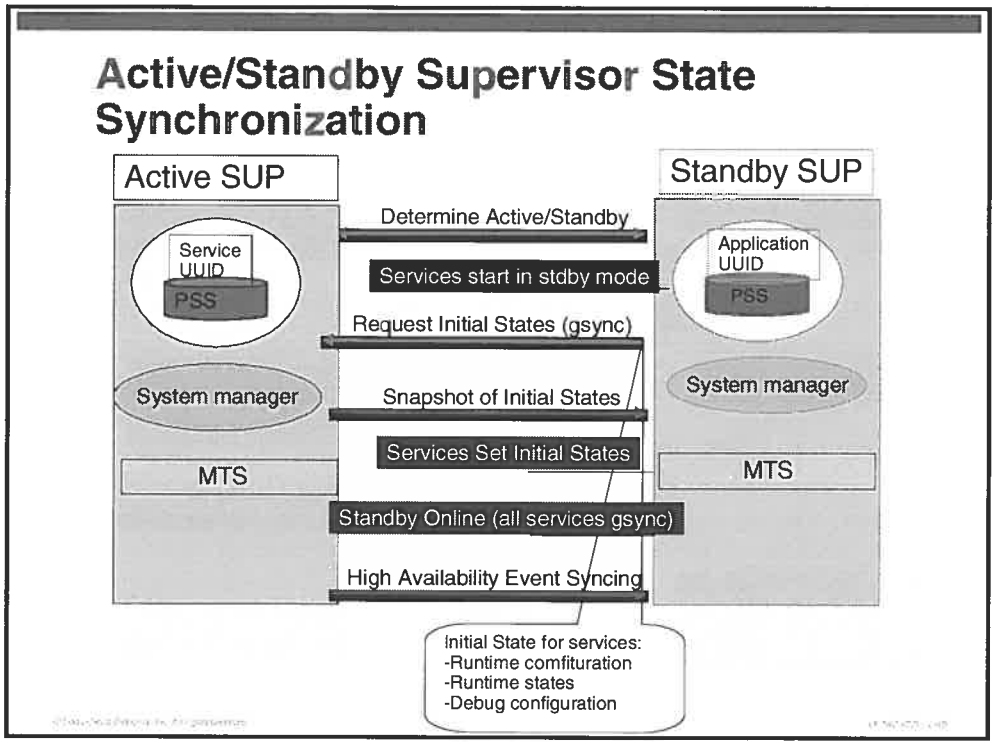
Step 5 The process reads from the MTS queue to retrieve pending transactions.

Step 6 The inconsistencies are fixed of the persistent and runtime database, rolling-forward the related pending transactions is performed. The pending transactions from MTS are committed by the new incarnation of the process, where applicable to start processing from the most recent process state.

Step 7 The MTS starts to de-queue messages from its queues.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

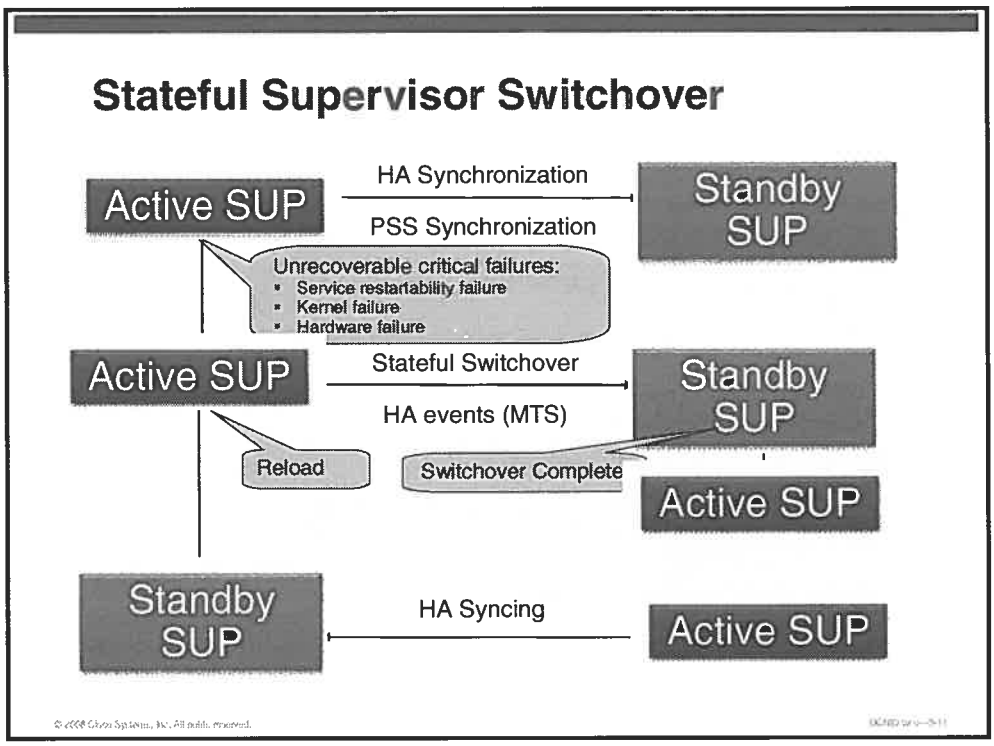


The active and standby supervisor modules in the Cisco Nexus 7010 platform remain synchronized in state and configuration using a two-stage process. The first time a supervisor enters a fully operational state, it will determine whether it is the active or the standby supervisor. If it is the standby unit, it will perform the first stage of the synchronization by initiating a global synchronization (GSYNC) with its counterpart using PSS. To accomplish this synchronization, the standby supervisor will issue a GSYNC request to the active supervisor. The active supervisor module will send a snapshot of all the initial process states to the standby supervisor module to synchronize their initial states using the System manager support. The Active supervisor unit, in return, will begin a complete exchange of current operating configuration and process state to the standby unit. Once this first stage complete information exchange has taken place, the supervisors enter the second stage of SSO synchronization, in which event driven and state change triggered synchronization updates are sent via messages through the MTS facility. A service that is updating or checkpointing its state will issue a message to MTS (as well as writing to PSS). MTS will route the subsequent synchronization message to the peer MTS process on the standby supervisor through MTS messages over the EOBC. The MTS process on the standby supervisor will notify the corresponding process on the standby supervisor in order for it to update and checkpoint its own state to match that of the peer process on the active supervisor. Additionally, the System Manager of the active supervisor will monitor the state and operational readiness of the standby supervisor. If the standby supervisor has entered an unexpected state, or if MTS synchronization to the standby fails, the standby supervisor is reset in an effort to facilitate effective and consistent state between the active and standby supervisors. In the event of the detection of an unrecoverable error, the System manager, depending on the configured high availability policy, may initiate a Supervisor Switchover from the current active supervisor over to the current standby unit.

Note First, supervisors are synchronized with the active supervisor module services. From that point on, as soon as the process state changes on the active supervisor, high availability event synchronization messages will be sent from active supervisor MTS to the standby supervisor MTS.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



A single supervisor module failure on the Cisco Catalyst 65000 switch used to result in a minimum of a few seconds' convergence time. This convergence time will most likely be more because of the tables that have to be rebuilt (ARP, CAM tables, and so on), so maximum convergence time can approach the 9 second limit of the Windows TCP/IP stack. This convergence time and possible lost session can be avoided by using dual supervisors with Nonstop Forwarding with Stateful Supervisor Switchover (NSF SSO) on the primary data center switch. Nexus 7010 Series Switch supervisor modules support NSF SSO and thus can dramatically improve the convergence time in a supervisor failure condition. NSF with SSO is a supervisor redundancy mechanism on the supervisor module that provides intra-chassis stateful switchover.

The recommended data center design that uses service modules has a minimum convergence time of ~6–7 seconds primarily because of Catalyst switch service modules. With NSF SSO, the service modules, for example in the Catalyst 65000, do not converge. This alone represents a large reduction in convergence time, making dual supervisors with NSF SSO a tool for achieving increased high availability in the data center network.

NSF SSO can provide a very robust solution to data centers that require a very high level of resiliency and are willing to use available slots to achieve it. The processes and functions that are involved for NSF SSO to work correctly are very complex and have many inter-related dependencies. Cisco recommends testing NSF SSO to ensure that it works as expected in a specific customer design.

Let us assume that the active supervisor either has a hardware failure, kernel failure or has repeated service restart failures and requires a switchover. A switchover will be triggered and standby supervisor will become the active supervisor. Once the Supervisor Switchover is completed, which happens fairly quickly because the supervisors are already synchronized, the active supervisor is reloaded. If the active supervisor reloads and passes through the diagnostics, it will then become the new standby supervisor and the process will reverse itself. There will be a global synchronization to the standby supervisor and then event synchronization will happen in the opposite direction. State recovery is minimal because services are already

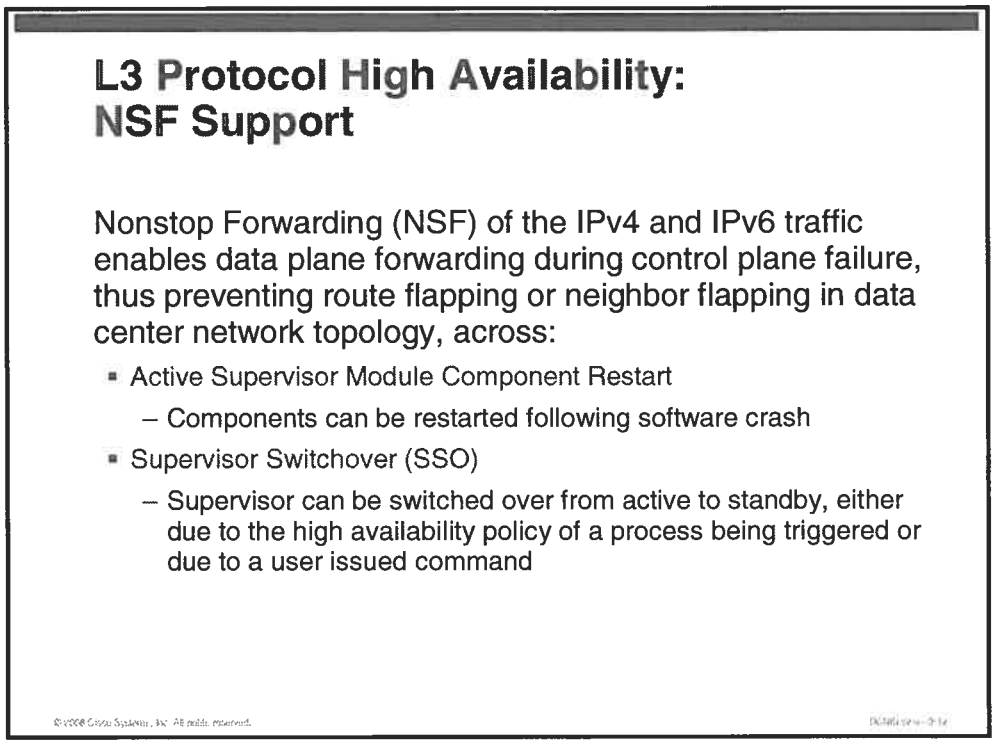
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

"ready to run" (provided basic event synchronization). Recovery can take longer if other layers of synchronization are added on top of basic syncing e.g. PSS Synchronization. When the standby supervisor comes up, an initial global synchronization is done and the application is started on the standby module and from that point on, subsequent synchronization of the context between them is done by the MTS through high availability event messages. Since the PSS is actually a service as well, the PSS can be synchronized. Some services are designed to leverage and recover their state on the standby supervisor using the PSS database. Either we can use high availability event messages to synchronize services or just synchronize the PSSs and allow them to recover their state when they switch over from the PSS. It may take longer to recover the process state from PSS database using the PSS synchronization upon the switchover, rather than just sending several messages for high availability events. Although PSS is an option for synchronization and is a part of the API, most of the processes use high availability event synchronization.

Note It is possible that interior gateway protocol (IGP) timers can be tuned low enough such that NSF SSO is defeated, because the failure is detected by adjacent nodes before it is determined to be an SSO stateful switchover.

https://t.me/learningnets



NX-OS nonstop forwarding is supported for all of storage, Layer 2, and Layer NX-OS protocols for IPv4 and IPv6 stack. Nonstop Forwarding (NSF) is an important technology to implement high availability of network equipment. Nonstop forwarding is supported without requiring NSF-aware neighbors. As the data center network develops, data center raises the higher requirements for the network high availability. NSF is a high availability technology that can ensure normal data forwarding when the control plane of a router fails, thus protecting key services on the network. Currently the NSF is seeing wider and wider applications.

NSF Principles

NSF guarantees nonstop data forwarding in the case when the control plane of the Nexus 7010 Series Switch fails. One of the most typical examples of a switch failure is a restart of the control plane of the Nexus 7010 Series Switch, accompanying active and standby supervisor switchover. There might be multiple causes that result in a failure of the control plane, for example, a Nexus switch fault or restart due to software or hardware failures, or active-standby switchover triggered by the configuration command during the software upgrade.

In general, when the Nexus 7010 series switch control plane fails, the neighbors at the routing protocol plane detect that their neighborship is down or that it is coming back again up after a delay. This process is called a neighborship flapping. The neighborship flapping finally leads to the route flapping. In this case, a route black hole appears at the restarting Nexus switch so that the neighbor routers bypass the restarting router with the data forwarding, which greatly lowers network high availability.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NSF is developed to cope with the route flapping, but the following requirements must be met:

Hardware Requirement

Two main supervisor module redundancies are available:

- Active control plane on the active supervisor module
- Standby control plane on the standby supervisor module.

Once the active supervisor restarts, the standby supervisor module takes over. The system is built based on a distributed architecture, with data forwarding and control plane separated. In addition, the system provides a dedicated interface board for data forwarding.

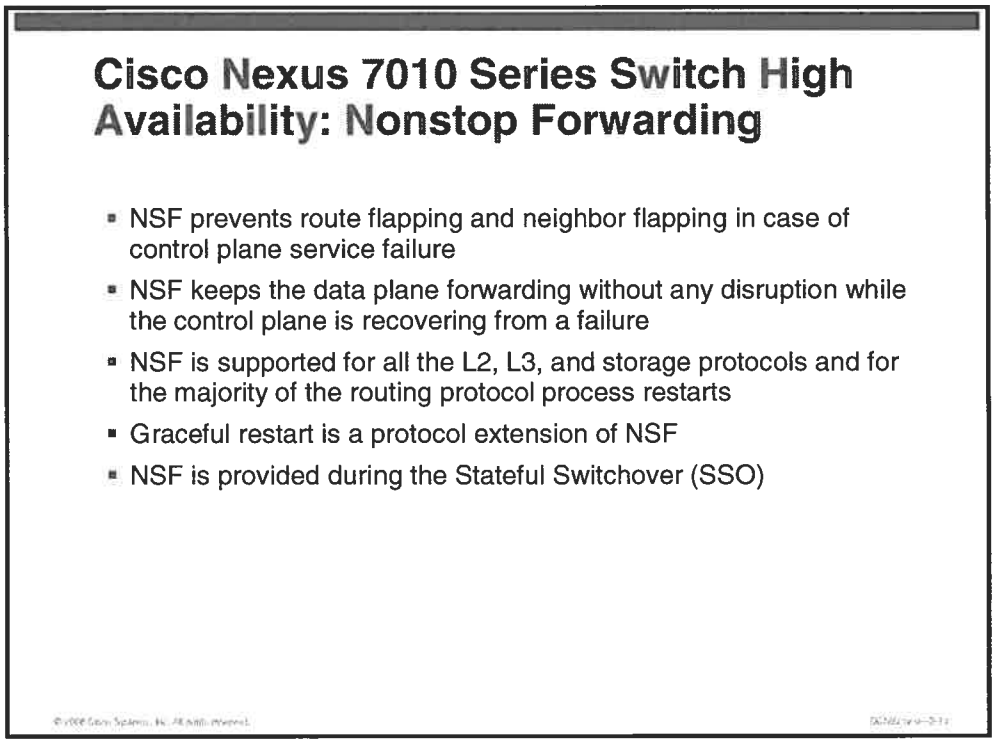
Software Requirement

During normal operation, the active supervisor module backs up the configuration information and interface status information to the standby supervisor module. During active/standby switchover, the interface board does not restart; the interface remains UP and the forwarding table (FIB) on the I/O module Fast Ethernet is not removed.

Protocol Requirement

The capability of graceful restart is available to the routing protocols such as OSPF, IS-IS, and BGP and other protocols, for example: LDP and RSVP.

https://t.me/learningnets



In the distributed model, the Nexus 7010 Series Switch is composed of a main board and an interface board. The main board is responsible for the control and management processes, for example, learning routes and calculating routes. The interface board is responsible for processing the services, for example, forwarding IP messages, exchanging MPLS messages, etc. The Nexus 7010 Series Switch has redundant design with two main supervisor modules, active and standby. During normal operation, the active supervisor module backs up all static configuration information and part of dynamic information to the standby supervisor module, so that the standby module has the same process configuration as the active. When the active one becomes faulty due to hardware or software failure, the standby takes over and restarts the control plane, management plane, and other control plane services.

Supervisor Switchover (SSO)

The Cisco Nexus 7010 Series Switch platform provides 1+1 redundant supervisor modules that can perform a Supervisor Switchover (SSO) in critical failure situations. The switchover of active to standby supervisor is nondisruptive to the forwarding plane and can therefore provide nonstop forwarding of data during a supervisor switchover. Due to the separation of the control plane from the data plane forwarding, the interface Fast Ethernet functions will not change/restart during the active/standby switchover. The forwarding tables in the Fast Ethernet on the I/O module are maintained to keep the data forwarding.

NSF Terminology

Nonstop Forwarding is possible because the data forwarding plane and the control plane are independent and the data can be normally forwarded while the control plane goes wrong.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Graceful Restart

Graceful restart is an extension that supports NSF. Based on the device different NSF abilities, the routers and switches are classified as follows:

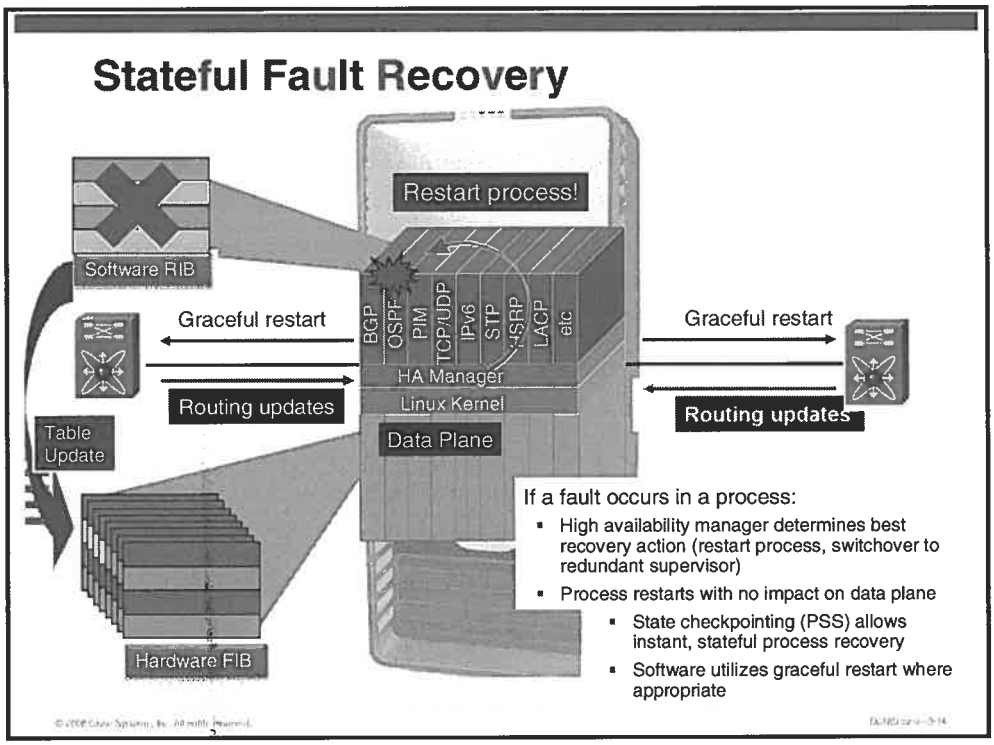
- **NSF-Capable Router**: The router is capable of NSF. It is configured with two main supervisors as redundancy configuration and the routing protocols have the graceful restart capability.
- **GR-Capable Router**: The router has the graceful restart capability.
- **GR-Aware Router**: The router is able to detect that its neighbor is experiencing graceful restart and to support the neighbor to perform the graceful restart. It is obvious that a graceful restart-Capable Router is also a graceful restart-Aware Router.
- **GR-Unaware Router**: The router is not able to detect that its neighbor is experiencing graceful restart and not able to support the neighbor to perform the graceful restart.

Based on the role during the graceful restart, the routers are classified as follows:

- **GR-Restarter Router**: The router implements protocol graceful restart.
- **GR-Helper Router**: The router helps implement protocol graceful restart.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



If a software failure occurs, the high availability manager determines the best recovery action: process restart, switchover to redundant supervisor or single supervisor module reset. The process restarts with no impact on the data plane forwarding. State checkpointing (PSS) allows instant, stateful process state recovery. Software processes utilize graceful restart where appropriate and where process recovery using PSS and MTS is not available.

The Cisco Nexus 7010 platform has the ability to program its forwarding hardware so that once the forwarding hardware is programmed packets can still flow through the system, regardless of what is happening with the control plane.

NX-OS services (for example HSRP) can checkpoint their runtime state to the PSS database. (They write their runtime state to the PSS). If that particular process crashes, the System Manager can restart it. A restarted process can then retrieve the state from the PSS and MTS to resume its most recent state of operation. All of these actions happen within milliseconds. Since the platform has been programmed for forwarding, it is still forwarding in hardware on the interfaces without interruption while the control plane is restarting the process. Certain processes are restarted gracefully (for example, Layer 3 routing protocols) using routing information from its neighbors for the state recovery instead of using the PSS database. When the routing updates are retrieved from its neighbors, the state is recovered locally by that routing process, the RIB is updated and the FIB is updated accordingly, all while the forwarding is still taking place. The recovery of process state using the routing neighbors through the graceful restart does not impacting the forwarding of a platform during operation.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

L3 High Availability

Type of Service Restart	Components Using GR Extensions	Components Using Periodic Protocol Refresh	Stateful Components	Non-Stateful Components
Protocols	OSPFv2	PIM, PIMv6	RPM	uRIB
	OSFPv3	IGMP, MSDP,	CLIS NetFlow	mRIB
	IS-IS EIGRP	MLD ICMPv6		u6RIB
	BGP	RIP	NetStack SMM	m6RIB

Based on the facts listed about process restarts, NX-OS L3 components can be categorized as follows:

- Components using graceful restart extensions
  - ospf, ospfv3, isis, eigrp, bgp
- Components relying on periodic protocol refresh
  - pim, pim6, igmp, msdp, mld (icmpv6), rip
- Stateful components
  - rpm, clis, netflow, netstack, smm
- Non-stateful components
  - urib, mrib, u6rib, m6rib

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Series Switch High Availability Policies for VDC

VDC HA Policies	Action (Single SUP)	Action (Dual SUP)
Restart	VDC is deleted then re-created with running config	Same
Bring down (TBD)	VDC is deleted	Same
Reset	Reset the Supervisor module	Switchover to the standby supervisor module

© 2008 Cisco Systems, Inc. All rights reserved.

Should a control plane failure occur, the administrator has a set of options that can be configured on a per-VDC basis, defining what action will be taken regarding that VDC. There are three actions that can be configured:

- Restart

  The restart option will delete the VDC and then re-create it with the running configuration. This configured action will occur regardless of whether there are dual supervisors or a single supervisor present in the chassis.

- Bringdown

  The bringdown option will simply delete the VDC. Subsequent VDCs created will have a default value of bringdown assigned to them. This value can be changed under configuration Control.

- Reset

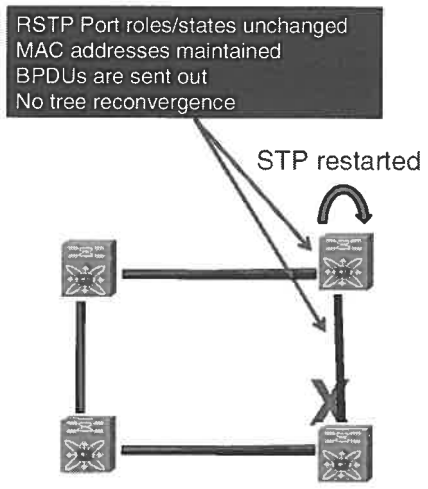
  The reset option will issue a reset for the active supervisor when there is only a single supervisor in the chassis. If dual supervisors are present, the reset option will force a supervisor switchover. The default VDC always has a high availability option of reset assigned to it.

 © 2008 Cisco Systems, Inc.

https://t.me/learningnets

Spanning Tree High Availability

- RSTP supports stateful process restart and SSO
- Port state and per-port BPDU information maintained across restarts, switchovers, ISSU
- No spanning-tree reconvergence upon service failure, ISSU, or process restart
- Counters, statistics, etc. not synchronized across restarts



NX-OS Spanning Tree Protocol (STP) supports stateful restart of the spanning tree service. If the STP fails or crashes, it will recover and return with all the recent port states intact. The STP reconvergence is not triggered upon the service failure or control plane failure or Supervisor switchover, and the MAC table is not flushed and relearned. A stateful switchover from the active to standby supervisor is supported, keeping the same spanning tree port information without any tree reconvergence. However, counters and statistics would not be synchronized and thus would be lost. (The spanning tree ports will continue to receive and send BPDUs, and will maintain the MAC TCAM entries during the stateful restart.)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Spanning Tree Edge Port Guards

- BPDU Guard
  - Error-disables edge ports that receive a BPDU
- BPDU Filter
  - Disables both sending and receiving of BPDUs on edge ports
- Root Guard
  - Places port in "root-inconsistent" state if it receives a superior BPDU
- Loop Guard
  - Prevents alternate or root ports from becoming designated ports due to unidirectional link failure

© 2008 Cisco Systems, Inc. All rights reserved.

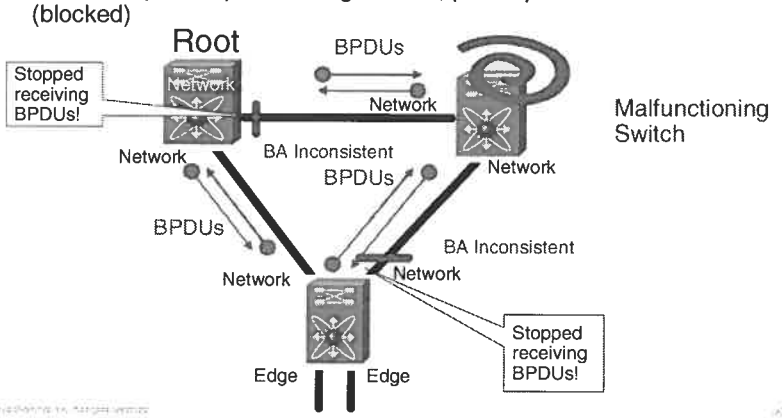
The following guard features are available on the spanning tree edge port: the BPDU Guard, the BPDU Filter, the Root Guard and the Loop Guard can be enabled. The BPDU Guard error-disables a port that receives a BPDU. The BPDU Filter has a milder reaction to a received BPDU on the edge port. It will simply drop any BPDUs that are received by the edge port and will not send any BPDUs out. The Root Guard feature enabled on the edge port protects the perimeter of the spanning tree from the superior BPDUs. When a superior BPDU is received o the Root Guard-enabled edge port, the port will be error-disabled and will enter a root-inconsistent state. When the superior BPDUs are not received anymore it will bring the port back into the spanning tree process. The Loop Guard is simply a unidirectional link failure mechanism.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Spanning Tree High Availability: Bridge Assurance (BA)

- Bridge Assurance ensures that spanning tree fails "closed" rather than "open"
- Bridge Assurance is a conditional service, enabled globally
- All ports with "network" port type send BPDUs regardless of role (Root port, Alternate port)
- If network port stops receiving BPDUs, port is placed in BA Inconsistent state (blocked)



Network ports are those that are connected to another bridge. The network ports have the Bridge Assurance service enabled. The Bridge Assurance feature blocks the network port that does not receive a BPDU to prevent a forwarding loop. When a network port with Bridge Assurance enabled comes up, it will immediately starts forwarding traffic out, because it transits from the blocking to the forwarding state instantly. If network port stops receiving BPDUs on any of those network interfaces then the interface is blocked or in other words the port is set into BA inconsistent state, which is a blocked state. Because the BPDUs have not arrived on the network port, the port is kept blocked. Let us imagine a topology with a switch that has ports participating in the spanning tree topology and do not have the Bridge Assurance service enabled because they are not configured network interfaces. When one of the switches has a malfunctioning control plane but the data plane continues forwarding, then this Layer 2 network topology will end up with a loop. Bridge Assurance is useful when a switch control plane is not functioning properly but the data plane is still operational and forwarding the traffic, because it will prevent the loop when the control plane cannot prevent it.

Note Bridge Assurance enabled network ports always originate and send out BPDUs. This is not the case for the ports with Bridge Assurance service disabled—BPDUs are not sent out of a blocked port or of a root port. When a network port is set to Bridge Assurance inconsistent state by the Bridge Assurance, the network is partitioned rather than accepting the normal spanning tree reaction that would first open up a loop and start forwarding all the traffic.

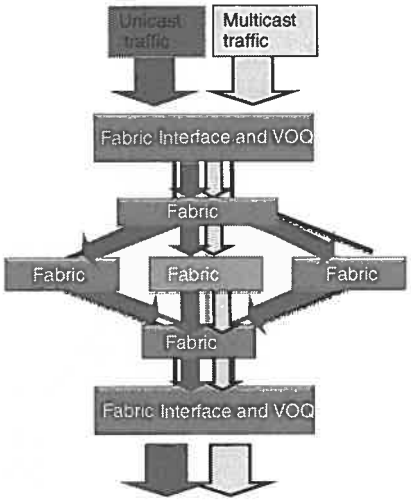
The dispute mechanism is applicable rather than Bridge Assurance as long as the current standard RSTP is run. Cisco Catalyst switch 65000 is targeting to support the Bridge Assurance feature for the Whitney 2 release. Bridge Assurance is planed to be a Cisco-wide switching feature. Bridge Assurance is a Cisco-specific feature and should not be used on the ports connected to a third-party switch. If switch is connected to a third-party switch, the port should not be configured as an edge or as a network port, but must be treated as a normal spanning tree port. If the BPDU has not been received on that port, there would be no mechanism, relative to Bridge Assurance, to detect the BPDU.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Network Resiliency: Fabric Load-Sharing

- Ingress fabric interface knows all active paths (all fabric modules) to every physical output queue (as mapped to VOQs)
- Unicast traffic is load-balanced across all the fabric modules to the egress line card
- Multicast traffic flow uses a single path (fabric module) across the fabric determined by the line card Fast Ethernet EtherChannel hashing algorithm



In addition to control plane and management availability through redundant supervisor modules, the Cisco Nexus 7000 platform also provides switching fabric availability through redundant switch fabric module implementation. A single Cisco Nexus 7010 chassis can be configured with one or more spine cards, up to a maximum of five for capacity as well as redundancy. Each I/O module installed in the system will automatically connect to and use all functional installed switch fabric modules. A failure of a switch fabric module will trigger an automatic reallocation and balancing of traffic across the remaining active switch fabric modules. Replacement of the failed fabric module reverses this process. Once the replacement fabric module is inserted and online, traffic is again redistributed across all installed fabric modules, thereby restoring the redundancy level.

The Cisco Nexus 7010 Series Switch ingress fabric interface dictates the path packets take through the fabric to the egress fabric interface.

- The **unicast traffic** flows from the ingress fabric interface and is load-balanced using the round-robin load balancing method across all fabric modules' active paths to the egress fabric interface.
- The **multicast traffic** flows from the fabric interface, selects one of the active fabric modules or paths to the egress fabric interface for the packet based on EtherChannel hash driven by the forwarding engine, and sends the multicast traffic along the same path.

Note The successful delivery of traffic across the fabric modules is not acknowledged and in case of a fabric module failure those frames that are in transit (across fabric crossbars) will be lost and the rest of the traffic is delivered across the rest functional fabric modules.

Nexus 7010 Series Switch network resiliency is also provided by the virtual output queuing

(VOQ) is explained in the QoS lesson. A central arbitration mechanism and virtual output queues for buffer availability at the egress help ensuring the delivery of the traffic across the fabric.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Series Switch Modularity: NetStack

The complete network stack runs in the user space:

- L2 packet management, ARP
- IPV4, IPV6
- ICMPv4, ICMPV6 (MLD, ND)
- TCP/UDP, socket library
- Adjacency management
- IP configuration

Benefits:

- Modularity
- Licensing
- High availability: ISSU, restart, switchover
- Decouple net stack from underlying OS
- Platform independence

Cisco NX-OS supports another functional component called the NetStack. It is basically an IP stack that runs in the user space. NetStack is used by the control plane. For example, routing protocol neighbor-to-neighbor communications leverage the NetStack. It was written from scratch to be task-specific for NX-OS operation. The NetStack is not a Linux network stack but is a completely modular or self-contained stack within the user space. The stack therefore can be upgraded whenever necessary because it is not tied to the Linux kernel. The fact that it is not a kernel loadable module eliminates the inherent risk of loading and unloading modules in the kernel because it runs completely in the user space and provides platform independency. NetStack enables the control plane to efficiently decouple network stacks from the underlying OS, so that, again, it is treated like a functional upgradable plug-in.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

ISSU Upgrade: Cisco NX-OS Image Sizes

Kickstart compressed image:

- To boot the Linux kernel
- Capability for 'system' image to bootstrap
- Provide a bare-bones CLI shell

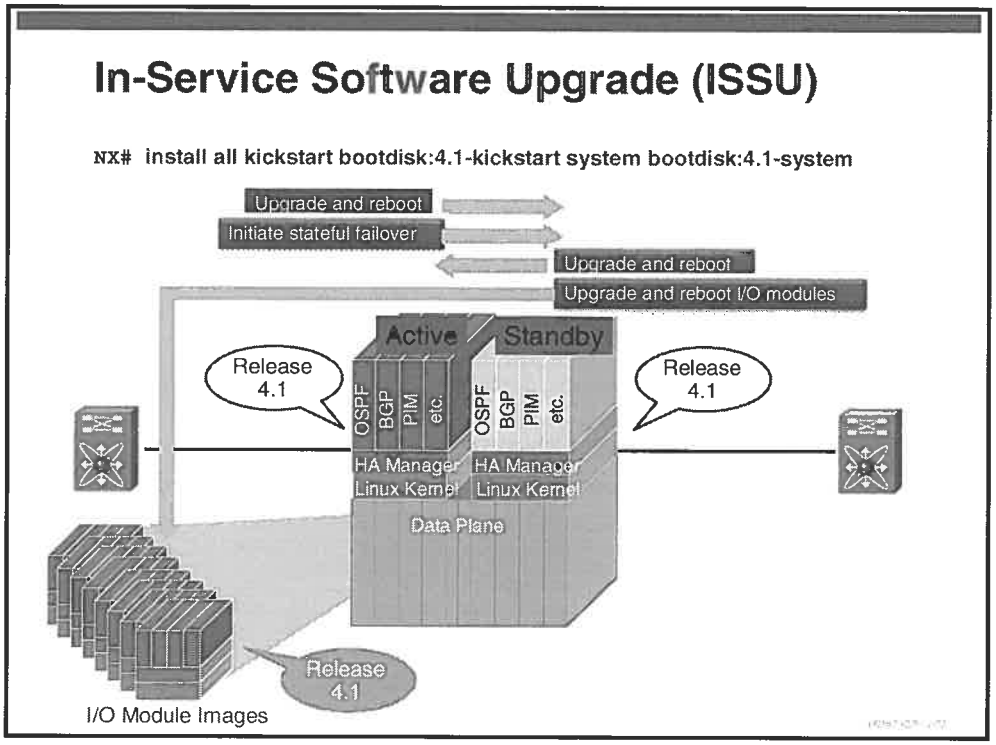
System compressed image:

- Loads line card drivers
- Loads NX-OS functionality that runs on the Supervisor and on the line cards
- Contains one or more plug-ins (Core, Ethernet, etc.)

There are two basic image types for the Cisco NX-OS. There is the kickstart image, which is essentially the Linux kernel that needs to be loaded along with drivers in order to be able to load the system image NX OS. At the system boot the kickstart image is loaded first, and then the system image is decompressed and loaded. The system images are the functional modules or plug-ins. The system image includes the code and the line card drivers that are downloaded from the supervisor to the line cards during the upgrade.

© 2008 Cisco Systems, Inc.

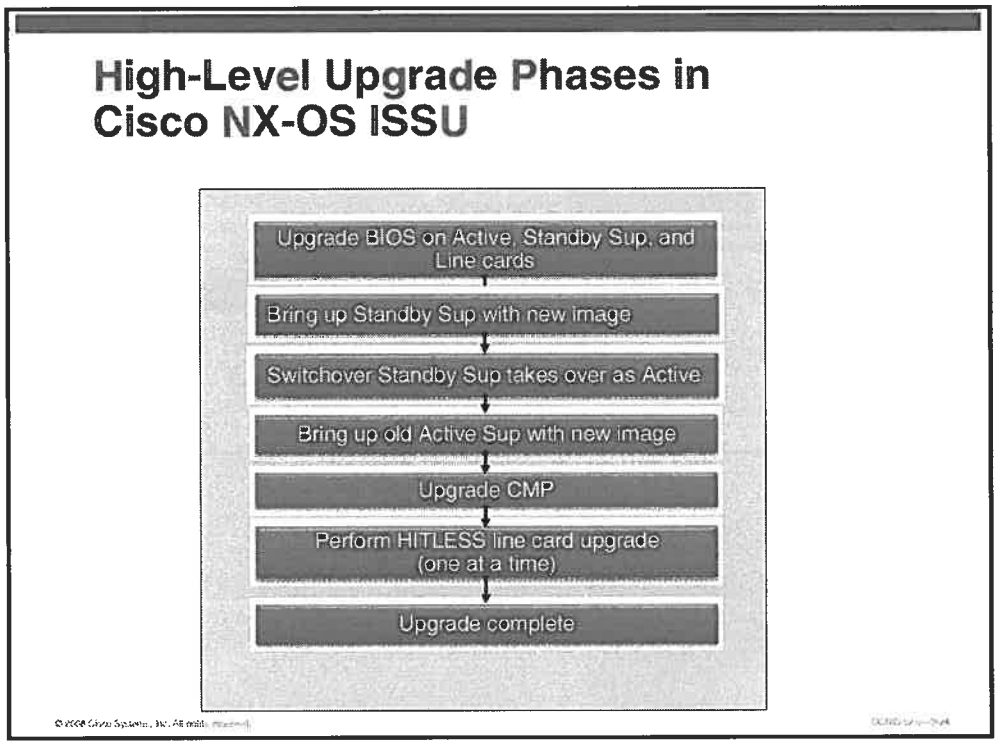
https://t.me/learningnets



The NX-OS platform has the ability to perform hitless upgrades—the ability to upgrade or downgrade from a software release to another major software release without affecting data forwarding. When the administrator triggers an In-Service Software Upgrade (ISSU) using the respective commands the currently standby supervisor is upgraded and rebooted without impacting the control plane or the forwarding plane. The stateful failover is initiated. After the reboot the switch performs process state synchronization (PSS) and initiates a failover. The former standby supervisor becomes a new active supervisor module. The other supervisor module which was formally the active supervisor becomes the standby supervisor and gets upgraded and rebooted. At that stage both supervisor modules are synchronized and both are running the latest software release. After the supervisors upgrade the CMPs and line cards are upgraded. Linecards are upgraded in a rolling mode: one by one, using the system software compressed image code on the supervisor module. The image code is downloaded to the line cards, one after the other. There is a waiting window (20–25 seconds) between each line card upgrade to make sure that the line card's are actually backed up and in stable state before the system upgrades the next line card. In case of an error detected, the ISSU process is aborted and the system reverts back to the original state. The line card upgrades are hitless; while the control plane of the line cards is upgraded the forwarding hardware and data plane is not touched and has been programmed to forward the data.

Note At this juncture, ISSU cannot be applied on a per-VDC basis. The installed software on the chassis is applicable for all active VDCs.

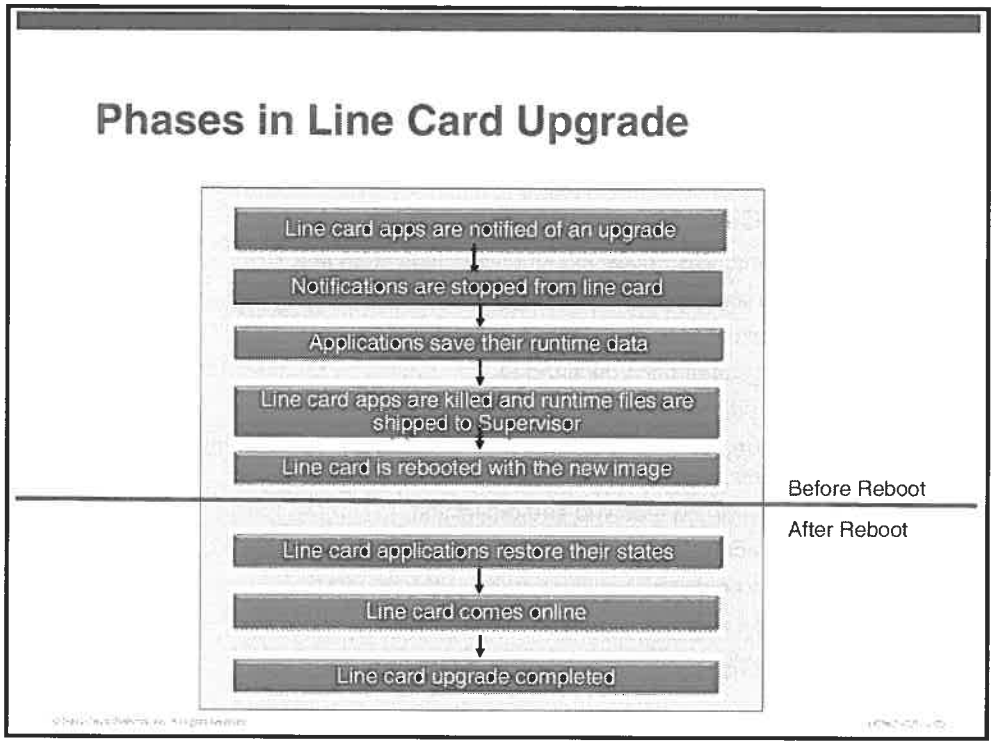
https://t.me/learningnets



The ISSU on the supervisor upgrades the standby supervisor with the new image, reboots the standby supervisor with the new image, and forces the stateful switchover. The active supervisor becomes the new standby supervisor and the standby becomes the new active supervisor running the new image. The old active supervisor is reloaded with the new image and the CMP is upgraded. The CMP has its own image and the CMP upgrade is performed separately. Once the CMP is successfully upgraded, the line card upgrades are performed one at a time with a wait period in between. When the line cards are upgraded the upgrade is complete. If an error occurs during this process, the upgrade will stop. One of the conditions during the ISSU upgrade is that the network has to be at a stable topological state, such that no topology change and no link or device failure should occur. If the ISSU is initiated while the topology change occurs, the ISSU process will be canceled due to the current unstable topology which could lead to a disruptive upgrade. During the line card upgrade, which goes through a number of processes itself, the applications or services of the line card are notified that an upgrade is being triggered. Applications in turn stop their communications and their notifications from the line card to the supervisor. All the local applications on the line card save their runtime data locally into a zip file and send that to this supervisor storage in nonvolatile memory in order to save their state.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



When the runtime data is sent to the supervisor for storage the line card is reloaded. The line card does not store its image locally. It boots the image or net boots it from the supervisor module. The line card retrieves its image, restarts its applications and comes online. The line card restores its state from the supervisor when it performs the upgrade. If an administrator was to insert a new line card or replace a line card while the line card is upgrading, the line card online insertion and removal (OIR) is put on hold until the line card upgrade is completed. A new line card that is inserted during an ISSU will not come up until the ISSU is complete or unless the ISSU is aborted. ISSU is done without affecting the traffic on the line card, because all the hardware on the line card has been programmed and we are not touching that programming while we are upgrading the control plane, instead taking advantage of the complete separation between control plane and the forwarding plane or data plane.

Note ISSU restarts the control plane without touching the forwarding hardware. The line card is not powered down and then powered up, but it is a kind of kernel restart performed locally on the line card, which does not directly affect the hardware that is already programmed.

Note The two supervisors can run with different versions and can coexist for that period of time during the upgrade, but it is not recommended to run different versions as an operational state. During the upgrade the two supervisors can coexist.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

ISSU Impact on Spanning Tree

During the Cisco NX-OS ISSU:

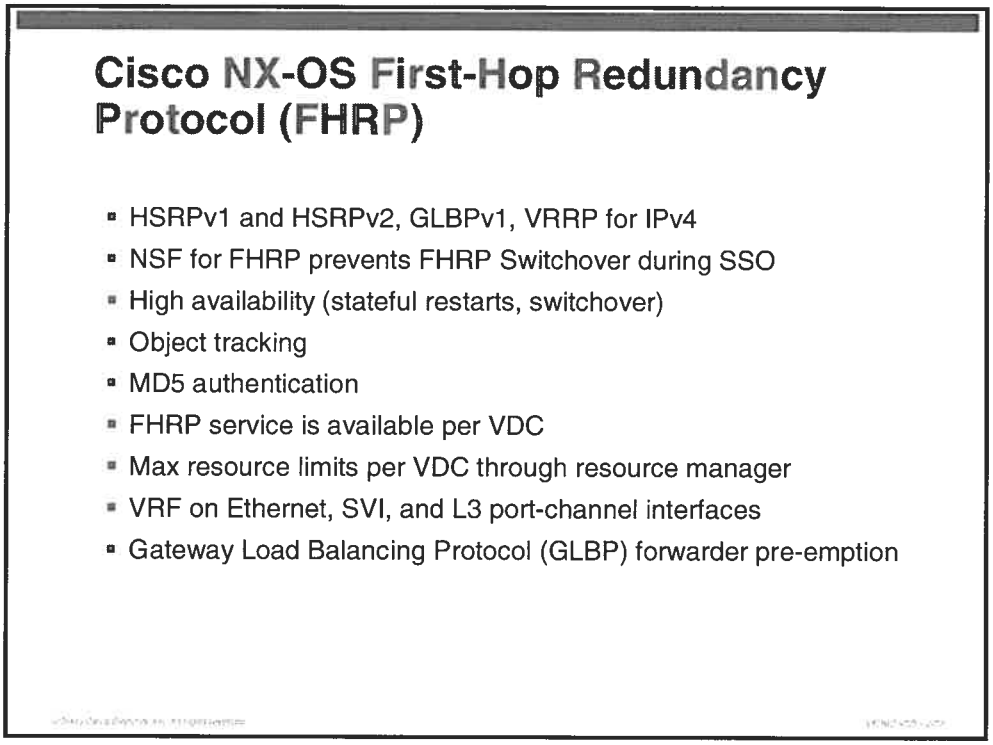
- Sophisticated ISSU code manages topology changes
- Topology must be stable
- Tree reconvergence not triggered
- RSTP port states/roles unchanged
- I/O module data plane keeps forwarding data
- Spanning tree control software continues running to prevent forwarding loops or tree reconvergence
  - BPDUs can be received and processed
  - BPDUs can be transmitted
- Spanning tree control software offline with regard to:
  - Interface port state cannot be changed
  - MAC table cannot be flushed, synchronized, etc.

© 2008 Cisco Systems, Inc. All rights reserved.

The network topology must be stable when the ISSU upgrade is performed. Topology change or link flapping should not occur to make topology stable. Stable topology is expected from the routing perspective as well as the Spanning Tree or switching perspective. During the ISSU upgrade the data plane continues forwarding the traffic using the existing MAC table entries all in the hardware. The spanning tree software can continue to process and send BPDUs. It is critical for the ports participating in the spanning tree protocol to continue sending or receiving BPDUs during the ISSU in order to prevent tree reconvergence or forwarding loops. During the ISSU the interface state changes are not possible nor is it possible to flush the MAC table entries based on topology changes. The whole synchronization process is essentially suspended until the ISSU is completed for that line card and for the supervisor module.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



First-Hop Routing Protocols (FHRP) such as HSRP and GLBP are provided by the Cisco NX-OS Software platform. These services offer default gateway redundancy for attached hosts. The FHRP service is available on a per-VDC basis. As part of creating the VDC, the administrator can define the number of FHRP groups that are available to each VDC. The first-hop redundancy protocols or the default gateway redundancy protocols supported by the NX-OS are HSRP, GLBP and VRRP. NX-OS FHRP currently includes HSRPv1 and HSRPv2 as well as GLBPv1 and VRRP, all of which support IPv4 specifically, but not IPv6. The first-hop redundancy protocols in DCOS are stateful. All FHRP protocols on NX-OS support the stateful restarts. There is also support for object tracking through another service called Object Tracking Manager which provides object tracking information and notification to the FHRP client code. The HSRP, GLBP and VRRP are implemented as clients to the Object Tracking Manager. FHRP timers are millisecond timers and all FHRP protocols provide the MD5 authentication. There are special accommodations to support the nonstop forwarding within the FHRP implementations on the NX-OS. IP redundancy API, ICMP redirect for HSRP or FHRP on the management interface are not supported by the NX-OS.

NSF for FHRP (GLBP and HSRP)

The failover time of the HSRP service is approximately 3 seconds whereas the VRRP needs 20 seconds of failover time. With the current IOS 12.2.(18) release, the HSRP state was not maintained so far by NSF SSO on the Catalyst 65000 Series Switches. The HSRP static MAC address was statefully maintained between supervisors but the state of HSRP between aggregation nodes was not. This means that if the primary Sup720 failed, the SSO-enabled secondary Sup720 took over and continued to forward traffic that was sent to the HSRP MAC address, but the HSRP hellos that are normally communicated to the standby HSRP instance on the second aggregation switch are not communicated. This means that during a switchover on the first aggregation switch, for example, the second aggregation switch HSRP instances takes over as primary during the SSO control plane recovery. Because the HSRP MAC address was statefully maintained on the first aggregation switch standby Sup720 module, the sessions continued to flow through the first aggregation switch, regardless of the active state that appeared on the second aggregation switch. After the control plane came up on the first

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

aggregation switch, the HSRP hello messages began to flow, and preemptively moved the active state back to the first aggregation switch. The control plane recovery took approximately 2 minutes. With looped access layer topologies (triangle and square) that align HSRP, STP primary root, and active service modules on the first aggregation switch, this does not create an issue because the access layer to the aggregation layer traffic flow continued to be directed to the first aggregation switch. If a loop-free access layer design was used, the active HSRP default gateway instance on the second aggregation switch responded to the ARP requests.

The NX-OS provides NSF for HSRP and GLBP to avoid the FHRP failover discussed previously. In the environments where small hold-timers are used for first hop redundancy protocols, such as subsecond timers, it is entirely possible that during a stateful switchover for a supervisor, if the supervisor switchover time takes longer than the FHRP hold-timers or the dead timer permit, a FHRP switchover will be triggered. To avoid the FHRP switchover in that particular case, the NX-OS offers a temporary hold-timer. That can be enforced by the supervisor when it is going to switchover. The temporary hold-timer will prolong the FHRP hold-time expiration only temporarily in order for the supervisor switchover to occur without FHRP switchover. After the switchover, the timer reverts back to the original sub-second value.

Global commands enable the administrator to set the timers on a per-protocol basis—different temporary hold-timers for HSRP and for GLBP can be set. There is a per-group command to enable or disable the hold-timer settings for the NSF HSRP and NSF GLBP protocols. For example, all HSRP groups in a certain VDC can use a hold time of 10 seconds during a graceful switchover or the extended-hold timer feature can be disabled for a specific group.

Note With small values for FHRP hold-timers (i.e. subsecond timers) unnecessary, FHRP protocol failovers may occur during operations such as the software upgrade or a stateful supervisor switchover. The solution to the problem is to temporarily increase the FHRP protocol hold-time to prevent protocol failover from happening.

NX-OS FHRP Scalability

Minimum Hello or Hold timer: 250ms/750 ms (recommended minimum is 1/3 of a second)
There will be 2000 HSRP groups available on the NX-OS with a maximum of 20 HSRP groups per interface and with one object-tracking object per group with and hello and hold timers of 250ms and 750ms.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

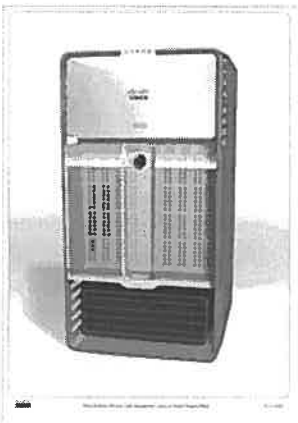
Other High Availability Features

High availability software architecture

- Online diagnostics (GOLD) continually monitors all the hardware components
- Control Plane Policing (CoPP)
- Roles-based access control provides granular management controls
- Configuration verification and rollback helps minimize disruptions caused by operator errors
- Embedded Event Manager (EEM) scripts triggered based on failures or management events
- EtherChannels

High availability hardware architecture

- Call Home for system problem notification
- Dual redundant Connectivity Management Processors (CMP) provide secure out-of-band management access
- Dual redundant Ethernet Out Of Band (EOBC) for control plane signaling to line cards
- Dual redundant active central arbiters for packet scheduling



Other high availability features are supported by the NX-OS, such as monitoring, troubleshooting and management features that enable the administrators to predict the failure condition and to prevent or simplify the troubleshooting process. For example, generic online diagnostics, control plane policing, role-based access control, configuration verification and rollback, Embedded Event Manager (EEM), Call Home notifications, dual connectivity management processors, dual redundant EOBCs and dual active central arbiters will be discussed in the Management lesson.

https://t.me/learningnets

Summary

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

Summary

- A deeper examination of the Cisco Nexus 7000 platform reveals that the entire platform is truly designed from the bottom up with a goal towards highly-available uninterrupted service under almost any condition.
- By strategically employing fault isolation, fault detection, and fault recovery on many levels in all areas, the Cisco Nexus 7000 platform is capable of delivering critical enterprise-class service for current and future business needs while preserving capital investment as well as the data path.
- Cisco Nexus 7010 Series Switch has almost every single component redundant. Redundancy exists at the software, hardware, and logical layer in order to provide service resiliency, high-availability, and nonstop forwarding with zero data loss.
- Hardware provides redundancy at every component level: supervisors, fabric modules, power supplies, fans, etc.
- Software offers multi-layered, multi-faceted resiliency: Stateful process restarts, graceful restart for routing protocols, SSO and ISSU.

© 2008 Cisco Systems, Inc. All rights reserved.

Summary (Cont.)

- Hardware and software combine to deliver data center-class high availability with zero service disruption.
- Nonstop data plane forwarding is guaranteed by the memory-protected statefuly/statelessly restarted processes triggered by the software or hardware control plane failure. The data plane can forward the traffic while the control plane is restarting. The majoritiy of routing processes that do not support stateless restart support graceful restart, which enables NSF of data plane.
- NSF graceful restart enables process state recovery using the neighbor routing information.
- Redundant supervisor modules support Stateful switchover with nonstop forwarding of the data plane traffic.
- NX-OS supports Stateful L2 switchover without triggering the spanning tree reconvergence:
  - Cisco Nexus can perform ISSU that enables seamless upgrade on the Cisco Nexus 7010 Series Switch with the dual supervisor modules. The ISSU restarts the control plane without touching the forwarding hardware. The linecards are not powered down during the upgrade but perform a kind of kernel restart.

© 2008 Cisco Systems, Inc. All rights reserved.

https://t.me/learningnets

Lesson 4

Cisco Nexus 7010 Switch Network Management

Overview

This lesson describes the Cisco Nexus 7010 Switch Network Management architecture and features, as well as the management tools available.

Objectives

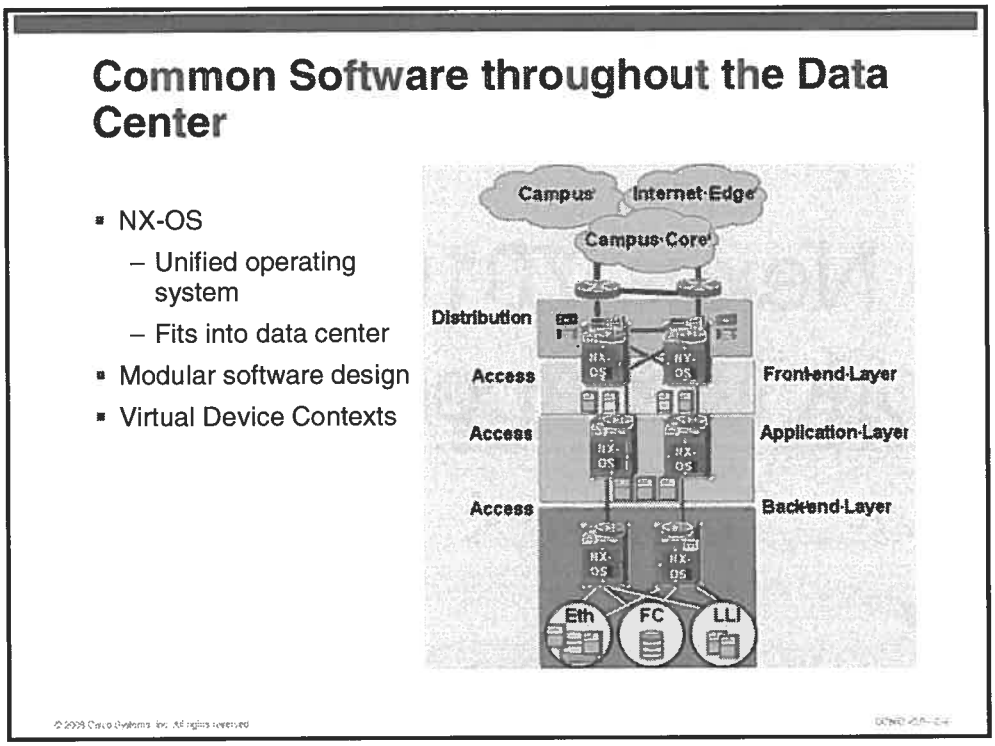
Upon completing this lesson, you will be able to meet the following objectives:

- Describe the Cisco Nexus 7010 Switch Configuration Methods
- Describe the Cisco Nexus 7010 Switch Management using DCNM
- Describe the Cisco Nexus 7010 Switch Operating System Serviceability
- Describe the Cisco Nexus 7010 Switch Operating System Manageability

https://t.me/learningnets

Cisco Nexus 7010 Switch Configuration Methods

This topic describes the Cisco Nexus 7010 Switch configuration methods.



Cisco NX-OS simplifies the data center operating environment by providing a unified operating system designed to run all areas of the data center network including the LAN, SAN and L4-7 network services. The diagram shows how NX-OS fits into a data center, integrating technologies such as Ethernet, L4-L7 services (such as firewall services, via deep packet inspection, classification), Storage Area Network, and Virtualization.

Modular Software Design

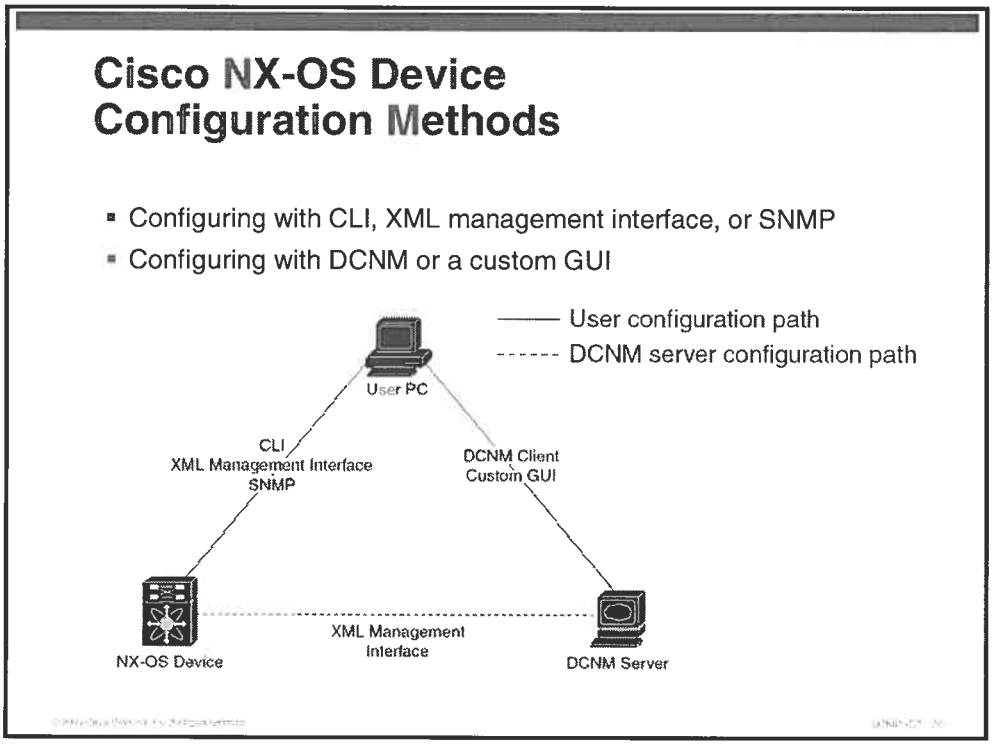
Cisco NX-OS supports distributed multithreaded processing on symmetric multiprocessors (SMPs), multi-core CPUs, and distributed line card processors. Computationally intensive tasks, such as hardware table programming, can be offloaded to dedicated processors distributed across the line cards. Cisco NX-OS modular processes are instantiated on demand, each in a separate protected memory space. Thus, processes are started and system resources allocated only when a feature is enabled. The modular processes are governed by a real-time preemptive scheduler that helps ensure the timely processing of critical functions.

Virtual Device Contexts

Cisco NX-OS can segment OS and hardware resources into virtual contexts that emulate virtual devices. Each virtual device context (VDC) has its own software processes, dedicated hardware resources (interfaces), and independent management environment. VDCs are instrumental in the consolidation of separate networks onto a common infrastructure, maintaining the administrative boundary separation and fault isolation characteristics of physically separate networks while providing many of the operational cost benefits of a single infrastructure.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Data Center Management Using CLI, XML Management Interface or SNMP Protocol

Cisco NX-OS devices can be configured using the command line interface (CLI), the XML management interface over Secure Shell (SSH), or Simple Network Management Protocol (SNMP) as follows:

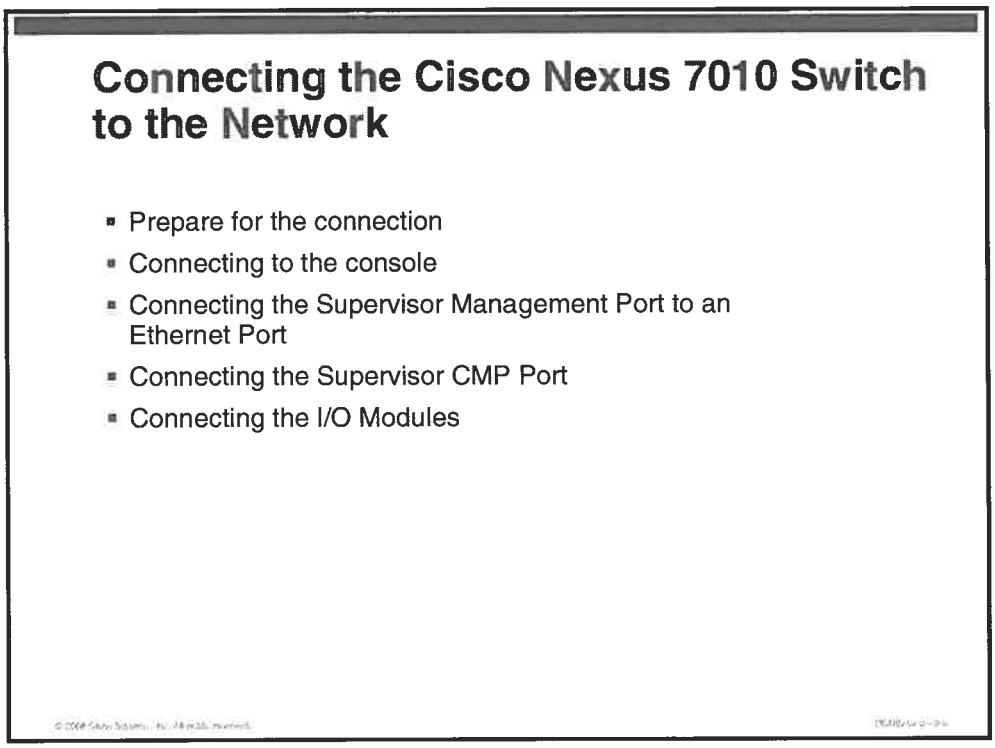
- **CLI from an SSH session, a Telnet session, or the console port**: Devices can be configured using the CLI from an SSH session, a Telnet session, or the console port. SSH provides a secure connection to the device.
- **XML management interface over SSH**: Enabled devices can be configured using the XML management interface, which is a programmatic method based on the NETCONF protocol that complements the CLI functionality.
- **SNMP**: Devices are configured using management information bases (MIBs).

Configuring with DCNM or a Custom GUI

Cisco NX-OS devices can be configured using the Data Center Network Manager (DCNM) client or from a custom GUI as follows:

- **DCNM Client**: Devices can be configured using the DCNM client, which runs on clients local PC and uses web services on the DCNM server. The DCNM server configures the device over the XML management interface.
- **Custom GUI**: Custom GUI can be created to configure devices using the DCNM web services application program interface (API) on the DCNM server. Simple object access protocol (SOAP) can be used to exchange configuration messages with the DCNM server. The DCNM server configures the device over the XML management interface.

https://t.me/learningnets



After successful installation of the Cisco Nexus 7010 Switch device in its rack or cabinet, operator is ready to connect it to a console, connect its management interface, and connect its Ethernet ports to the network.

The following steps are required to finish the connection:

Step 1 Preparing for Connections

During preparation of the site for network connections to the Cisco Nexus 7010 Switch, consider the following for each type of interface:

- Cabling required for each interface type
- Distance limitations for each signal type
- Additional interface equipment needed

Step 2 Connecting to the Console

Local management connection between a Cisco Nexus 7010 Switch is required before connecting the device to the network. This connection lets you perform the following initial configuration functions: set the IP address for the switch, configure the device using the command line interface (CLI), monitor network statistics and errors, configure SNMP agent parameters, and download software updates.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

This management connection is made between the asynchronous serial port on each Cisco Nexus 7010 Switch supervisor module and console devices capable of asynchronous transmission, such as a computer terminal (the computer terminal must support VT100 terminal emulation). Each Nexus 7000 Series supervisor has the following two asynchronous serial ports:

- **Console Serial Port**: Direct connection to an asynchronous device
- **COM1/AUX Serial Port**: Connection to the asynchronous device through a modem. Default port characteristics are the following: 9600 baud, 8 data bits, 1 stop bit, and no parity. RJ-45 rollover cable must be used for console connection.

Step 3 Connecting the Supervisor Management Port to an Ethernet Port

The Cisco Nexus 7010 Switch supervisor management port provides out-of-band management, which enables the operator to use the CLI or the DCNM interface to manage the switch by its IP address. This port uses a 10/100/1000 Ethernet connection with an RJ-45 interface.

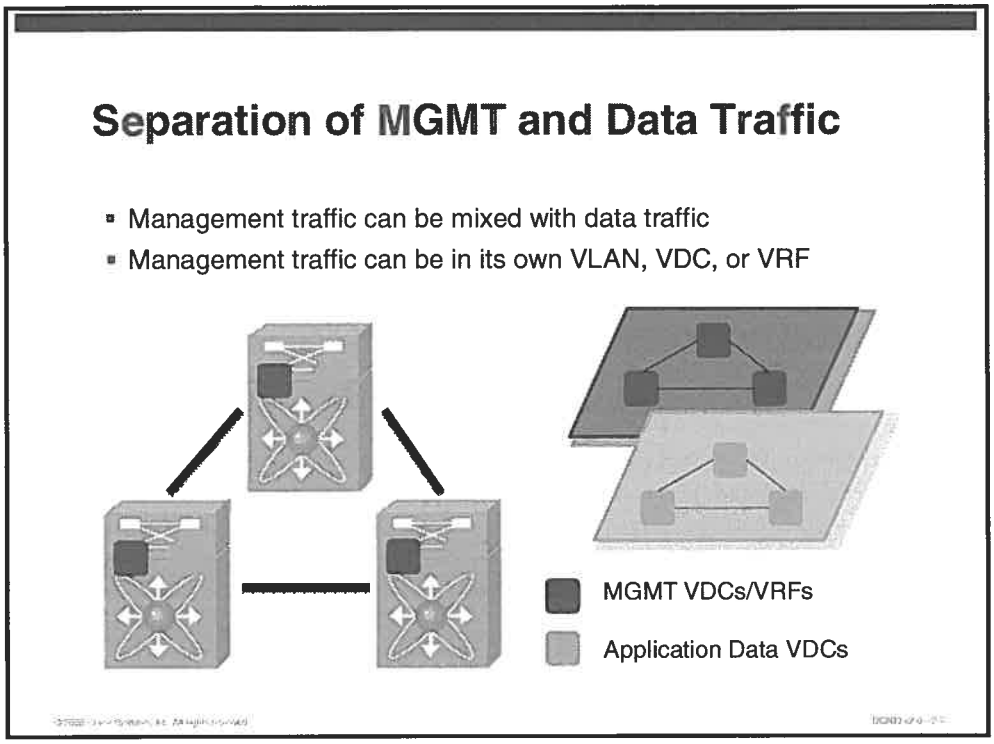
Step 4 Connecting the Supervisor CMP Port

The Connectivity Management Process (CMP), which is included on the Cisco Nexus 7010 Switch supervisor module, is a secondary, lightweight processor that provides a second network interface to the system for use even when the control plane is not reachable. Administrators can access the CMP to perform operations, such as taking over the control plane console, restarting the control plane, or restarting a particular I/O module.

Step 5 Connecting the I/O Modules

First the local management connection and the out-of-band management connection must be completed. Than the I/O modules can be connected to the network. The 48-port 10/100/1000 Ethernet module is using copper cables for the connection, and the 32-port 10 Gigabit Ethernet Ethernet module is using transceivers and fiber-optic cables for the connection.

https://t.me/learningnets



The Management network must be separated from the data center application traffic. The main reason for having a separate standalone management logical infrastructure is to be able to troubleshoot the data center network when application logical network fails and can not provide any management access to the switches to troubleshoot the network failure. Data center network security and manageability can be provided in a better way. When separation of the traffic is implemented the following options are available:

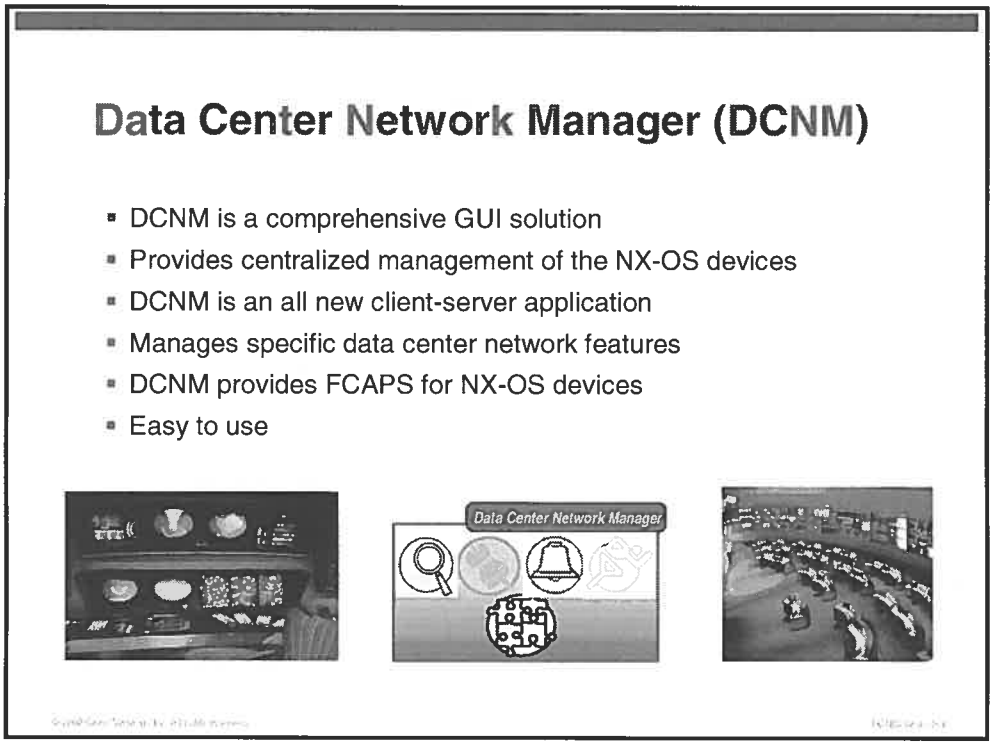
- Management traffic can be separated using a dedicated VLAN for the management traffic
- Management traffic can be in its own VDC/VRF

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Management using DCNM

This topic describes the Cisco Nexus 7000 Series switch management infrastructure using DCNM. The DCNM management components are discussed along with the DCNM client-server architecture.



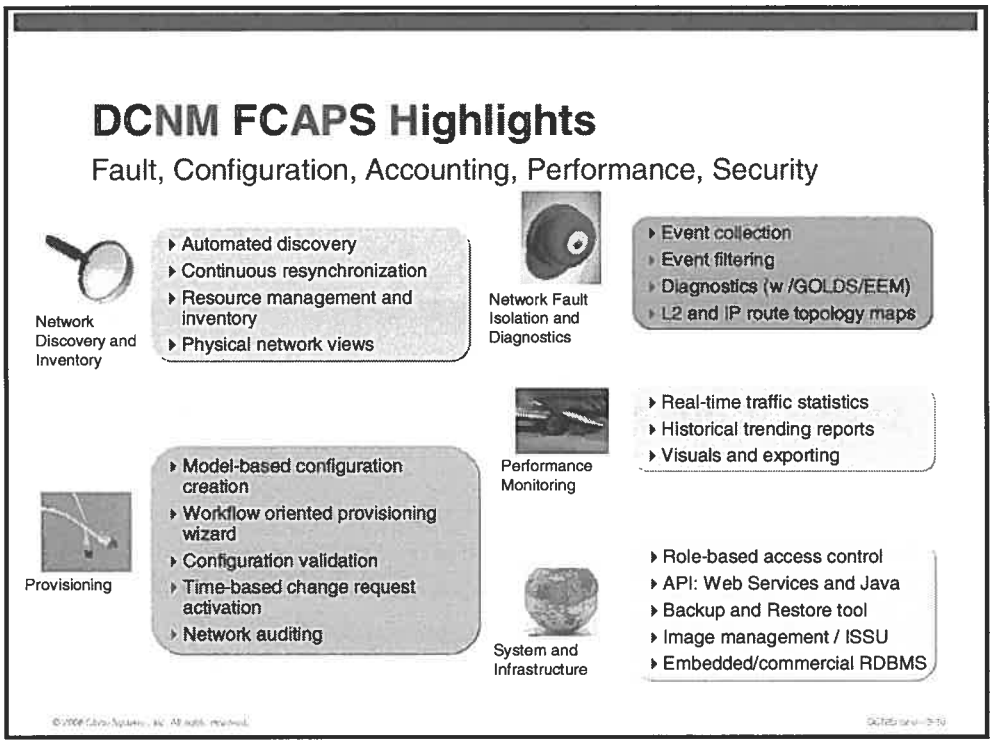
Data Center Network Manager (DCNM) is an easy way of utilizing the management application for the Cisco Nexus Operating System (NX-OS) based devices only. It is designed to provide centralized management of NX-OS based data center networking devices.

DCNM is a brand new product designed from the ground up. It is being developed under the same management that developed the Cisco Fabric Manager SAN-OS management software, which is the management solution for the Cisco MDS switches. DCMN is evolving and would like to be an FCAPS for all Nexus operating system (NX-OS) based devices.

DCNM provides fault management, configuration management, accounting, performance and security management functions (FCAPS). DCNM is following the NX-OS releases.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



DCNM provides the following management functions as part of the FCAPS management suite:

Network Discovery and Inventory

- Automated network topology discovery
- Continuous DCNM server database resynchronization with the XML clients
- Resource management and inventory
- Physical network views

Provisioning

- Model-based configuration creation
- Workflow oriented provisioning wizard
- Configuration validation
- Time-based change request activation
- Network auditing

Network Fault Isolation and Diagnostics

- Event collection functions
- Event filtering functions
- Network Diagnostics (w/GOLDS/EEM)
- L2 and IP route topology maps

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Performance Monitoring

- Real-time traffic statistics
- Historical trending reports
- Visualizing the performance reports (using graphs, etc.) and performance report exporting

System and Infrastructure management functions

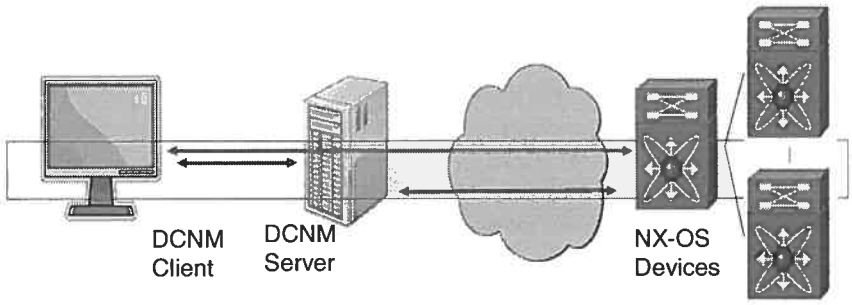
- Role Based Access Control
- API: Web Services and Java
- Database backup and restore tools
- NX-OS image management and centralized ISSU upgrade
- Embedded commercial RDBMS

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

DCNM Solution Components

- DCNM is a client server solution
- DCNM server communicates with the NX-OS devices
- DCNM client communicates with the DCNM server or with NX-OS devices directly



From an end user point of view DCNM is a pretty simple client-server oriented architecture. The main components of the DCNM solution are:

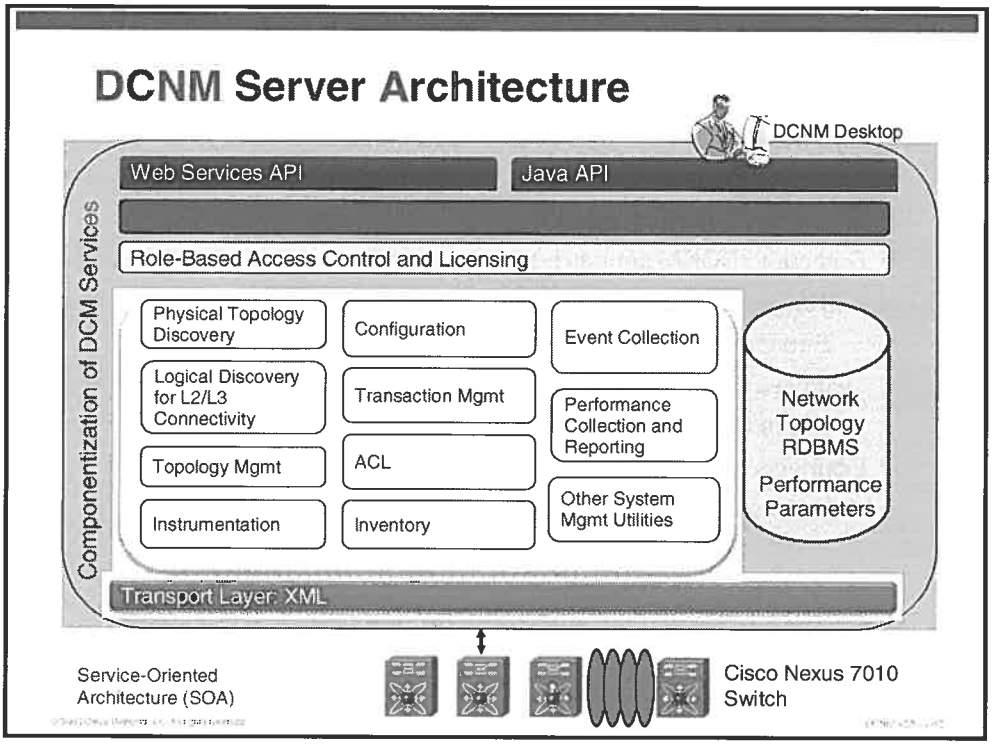
- NX-OS devices
- DCNM server
- DCNM clients

NX devices are building the network infrastructure and the topology is based on the network requirements. The server is a dedicated machine–a standalone server that performs centralized discovery of the network topology. The server also does the performance management functions and therefore is up and running non-stop. The DCNM client is a Java based client that could be run from any Java enabled workstation. Client is an administrator laptop that retrieves the discovered topology from the DCNM server and that manages the NX-OS devices by directing sending SET/GET configuration parameters to the NX-OS.

NX-OS switches support XML database and the XML API exposed to the DCNM clients to retrieve or to set the Nexus 7010 Series switch configuration parameters. The DCNM clients can therefore communicate configuration requests to the NX-OS switches (via DCNM server). DCNM server uses the NX-OS XML API to discover the network topology that consists of NX-OS-based networking devices and the related network parameters, and to retrieve performance parameters from the Nexus switches. DCNM manager will be able to discover the SNMP-based networking devices such as IOS-based routers and switches and SAN-OS switches using SNMP GET requests, but will not be able to manage those devices (SNMP Read-only access).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The diagram shows the architectural block view of the DCNM application. The server component of DCNM talks to the Cisco Nexus 7010 Switches using the XML API interface, which is the only means of interface between DCNM server and the Cisco Nexus 7010 Switches. The Cisco Nexus 7010 Series Switch provides the XML database and the XML API for accessing the XML database parameters. The big pink block in the diagram is showing the SOA components—or the services components—of the data center network that represent the base for the DCNM network management.

Functionally, components can be broken down into their own small service components within the server application. This breakdown offers a more structured way of adding new components into the application and is also an easier way of managing components because they can be structured to work independently of the other services. RDBMS is supported for storing database performance and for the data center network topology parameters. Initially only PostgreSQL 8.1 and 8.2 are supported. On top of that there exists a unified role-based access control and licensing implementation followed by a web server. The web server provides the interface for an administrator to use API to talk to the DCNM server. A set of XML APIs are published and available and are independent of the DCNM server APIs. The result is an open API for the DCNM server.

Therefore, third-party applications could talk to the DCNM server, rather than talking to the individual Cisco Nexus 7010 Switches. This way, third party applications get aggregated information out of DCNM server. Published APIs will be in two formats (Java API and Web Services API). Both APIs are providing the same options, as first Java API is created and then converted to the Web Services API.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

DCNM Server

- Java-based application
- Dedicated DCNM server required
- DCNM server technologies:
  - Service-oriented architecture (SOA)
  - Java 2 Platform Enterprise Edition / J2EE / Java EE
  - Jboss for DCNM clustering
- Connects to the NX-OS switch using SSH session
- XML is used for communication

DCNM server is a Java application. A dedicated server is required for DCNM and the server component runs as a background Java process. SSH is used by the server to connect to the Cisco Nexus 7010 Switches. NX-OS should have SSH enabled by default. DCNM server publishes API in two formats (Java API, Web Services API) and supports only PostgreSQL 8.1 database.

Several server technologies are included in DCNM, providing many benefits, as follows:

- Service-oriented architecture (SOA)
  - DCNM server technology is based on SOA. SOA plus other building blocks enables DCNM to offer redundancy and scalability, two required critical components of any data center environment. SOA unifies business processes by structuring large applications in smaller modules called services and new applications built from a mix of services exhibit greater flexibility and uniformity. Each function within the application can be a service by itself. While SOA is executed, one of the services can be taken out and it can be put into a different application. It should work as it is because each service has its own API interfaces and services can talk between themselves using the APIs. This is the reason why new functionalities can be added easily.
- Java 2 Platform Enterprise Edition / J2EE / Java EE
  - J2EE platform is one of the first ones to adhere to the SOA concepts. It is a set of coordinated technologies and practices that enable solutions for developing, deploying, and managing multi-tier, server-centric applications.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

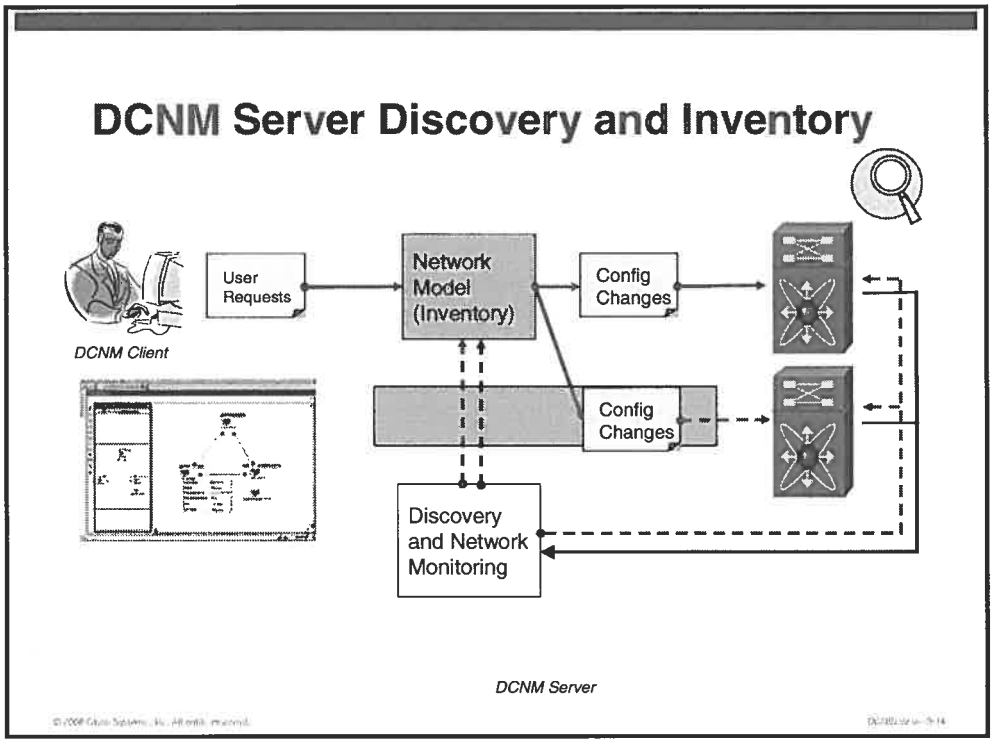
- Jboss
  - JBoss Enterprise Middleware is enterprise-class open source software that enables software developers to build, deploy, integrate, orchestrate, and present applications and services in a SOA. It is one of the most widely used client server platforms, which provides an instant infrastructure for a client server application to scale. Jboss implements the SOA architecture underneath it, using Java. It provides functionality like clustering at no cost to the application. JBoss technologies can be leveraged to make DCNM a clustered application and to have higher performance and to scale better.

Besides hardware requirements (suggestions–dual core processor or dual processor, RAM size 6 GB, 50 GB of HDD) the DCNM server requires one of the following two operating systems:

- Windows Server 2003 Standard Edition Service Pack 1
- Red Hat Enterprise Linux AS release 4

Communication between DCNM server and Cisco Nexus 7010 switches is using XML only.

https://t.me/learningnets



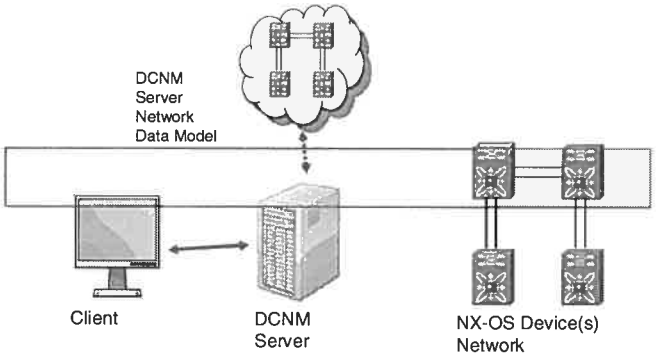
Once the DCNM server is up and running, it starts the discovery process. The DCNM server discovers all Cisco Nexus 7010 switches and all adjacent devices where CDP is enabled. During the discovery phase, a DCNM server performs an extensive collection of physical inventory as well as its configuration details. A huge amount of data about the device and how the device is configured is available. The next step is building a virtual model of the physical network in its memory. After the DCNM server discovery process, the SSH sessions to the server are still up for regular gathering of information at specified intervals. Initial discovery is looking for a File System and for the running configuration, and then the ongoing resynchronization takes place. DCNM relies on Syslog and Accounting Log information from the NX-OS devices to synchronize the DCNM server with the network.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

DCNM Server Network Model

- DCNM server builds network data model
- Client GET request from DCNM server
- Client SET request through the DCNM server (validation)



DCNM server builds an intelligent network data model in its memory representing the physical data center network configuration. This structure in the memory enables the server to intelligently serve user requests. DCNM server performs an exhaustive discovery of the Cisco Nexus 7010 Switches, and that enables DCNM server to respond to the client "Get" requests without having to go back to the device. "Get" requests from DCNM clients are served by the DCNM server. Therefore, rather than going out to the device and doing another query, the DCNM server serves the "Get" request from its network model.

The network is a kind of master; changes in the network model are driven or triggered by the actual network changes. DCNM relies on the Syslog and Accounting Log information from the NX-OS devices to synchronize DCNM server with the network. A Syslog message payload provides a rich set of information and DCNM server polls the NX-OS devices every 60 seconds and retrieves the latest Syslog information. A large percentage of users are utilizing CLI and they will make changes via CLI, irrespective of how successful DCNM is. It is important that DCNM captures those changes via CLI into its network model to keep itself truthful. It also retrieves the accounting log. The accounting log captures details of every CLI command executed on the NX-OS device. Accounting log messages also enable DCNM to learn about changes made either by CLI or third party applications. This knowledge enables DCNM to keep its model in synchronization with the device. By using the accounting log and the syslog, DCNM has 100% reliability of being able to represent the real world network within its environment.

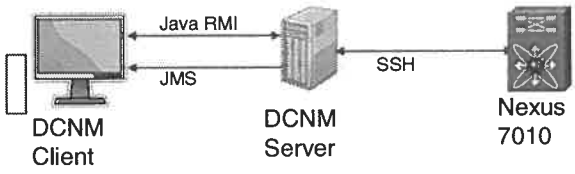
Set requests from DCNM client make changes to the device configuration. The request is first applied to DCNM server's own network model in its memory to validate the request. Validation rules applied are the same as what NX-OS does to validate CLI commands. The request is then forwarded to device(s) only if validation is successful. However, if DCNM server finds that the request is invalid, such as DCNM server can find a duplicate IP address, it cannot configure the secondary IP address if the primary IP address is not configured yet, it cannot change the port mode, or if the interface is a member of a port channel, etc.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

DCNM Client/Server Communications

- DCNM server connects to the NX-OS devices over SSH
- DCNM client communicates to the DCNM server over Java RMI
- DCNM server notifies DCNM client of asynchronous events as JMS messages
- DCNM client is a Java application
- DCNM client is downloaded from the DCNM server using Java Web Start technology



No direct communication between DCNM client and the NX-OS devices exist. DCNM server connects to the NX-OS devices through SSH (single SSH connection). This is the only connection between DCNM server and NX-OS devices. If a firewall is used in the network then the suggested design puts the firewall between DCNM client and DCNM server. By best practice, the server must be deployed closer to the NX-OS devices. The communication between the server and client is via Java RMI. In addition to Java RMI from the server to the client, there is JMS messaging going on. However, that is a one-way message, so the server updates the clients of any events that it is used back from the switches. Those are the modes of communication and besides this, there are no other ports opened.

The server can be running on the laptop computer and does not have to be connected to the devices all the time. Start the DCNM server when connected to the NX-OS device(s). Initially the server will perform discovery of NX-OS devices. After the discovery process, the server might be disconnected (when no performance management exists) and customer demos can be done off-line. All the functionality of DCNM server is available, except the configuration changes to the NX-OS devices cannot be done (because you do not have connectivity). From the UI point of view, and from launching the DCNM client application and the DCNM server application, everything is up and running.

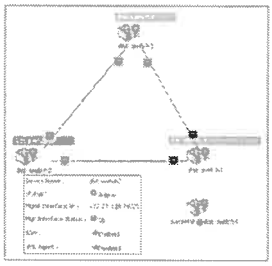
DCNM client is a pure Java application. It is a thin client—all the business logic is handled by the server, and not by the client. The first time that a user wants to launch a client, the user must connect to the server in order to download the client software to the desktop, where the client gets launched. Java Web Start technology provides this functionality. No installation of DCNM client is needed. After the initial download of the client software, the users just double click the icon on the desktop in order to start the application. Java Web Start technology will check if the client is the same version than the one on the DCNM server. They must be synchronized. If the server is running a newer version, then the client software will be automatically upgraded.

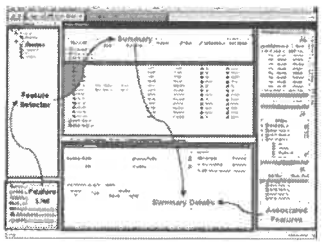
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Topology View and GUI

- Discovery for NX-OS devices only
- Discovery for adjacent NX-OS devices if CDP is enabled
- DCNM 4.0 supports switch and VDC level topology view
- Cisco look and feel
- GUI flow is network feature-centric





DCNM server is performing network discovery, and network topology is stored in the DCNM server's memory. Discovery is possible for NX-OS devices and for adjacent NX-OS devices with CDP enabled. Overlay views like VLAN topology are not supported in DCNM 4.0.

The GUI converged to the Cisco look and feel, and its flow is network-feature centric rather than device centric. (A similar approach is used on Cisco Transport Planner (CTP) software, Vframe, etc.) This feature makes all relevant information available on the screen, empowering the user to configure from a single screen without having to search for information about the network. Despite the feature-centric approach, the users can go to the device view, as well.

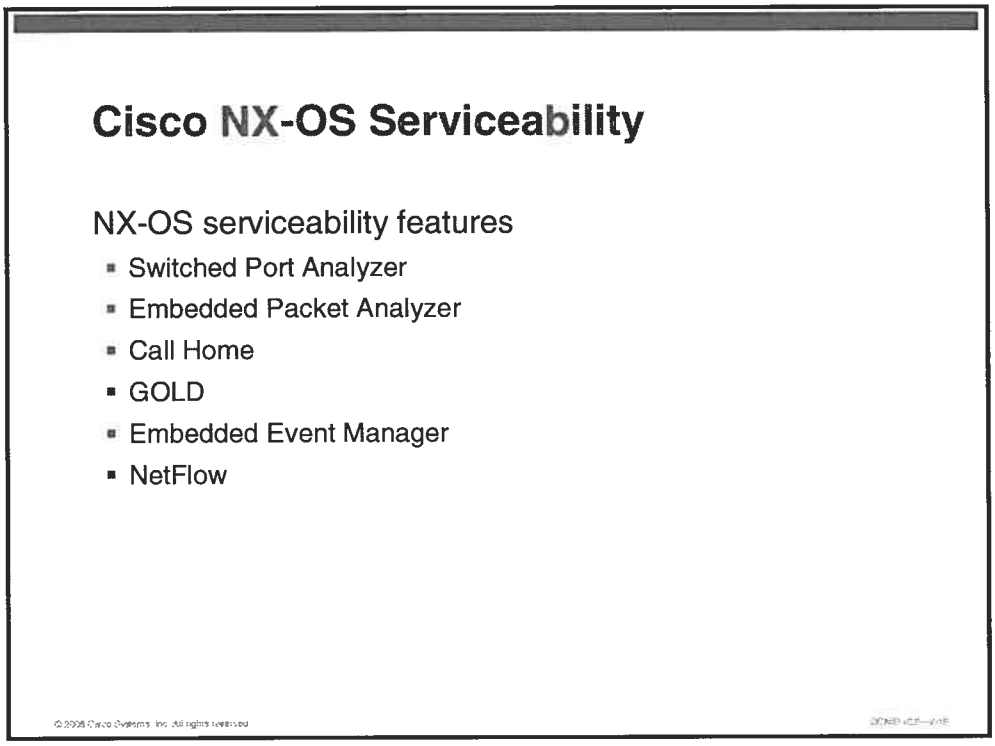
The diagram shows the screen shot of the DCNM client. On the bottom left it is a feature selector where the list of features supported by the application is shown. Once the feature is selected, the further feature details available will pop up. Then an item under the feature must be selected to get a summary view on the middle top. From the summary view, more details are presented in the bottom, middle portion of the screen. The right part of the screen is very unique to DCNM and it is called associated panels. When configuring a feature, the system suggests what you should look at next or what the associated information or features are that might be looked at in the next step.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Operating System Serviceability

This topic describes the Cisco Nexus 7010 Switch Operating System Serviceability features such as: Switched Port Analyzer, Embedded Packet Analyzer, Call Home, GOLD, Embedded Event Manager, and Netflow.



Cisco NX-OS is built with serviceability functions that let you take early action based on network trends and events. These features help in network planning and improving response times. The following features are included:

Switched Port Analyzer

The Switched Port Analyzer (SPAN) feature performs real-time network analyses of the traffic from any source port (called the SPAN source ports) by non-intrusively directing the SPAN session traffic from the interesting ports to the SPAN destination port that has an external analyzer attached to it. It performs control and data plane traffic analysis.

Embedded Packet Analyzer

Cisco NX-OS includes a built-in packet analyzer to monitor and troubleshoot control plane traffic only. The packet analyzer is based on the popular Wireshark open source network protocol analyzer.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Call Home

The Call Home feature continuously monitors hardware and software components to provide e-mail-based notification of critical system events. A versatile range of message formats is available for optimal compatibility with pager services, standard e-mail, and XML-based automated parsing applications. It offers alert grouping capabilities and customizable destination profiles. This feature can be used, for example, to directly page a network support engineer, send an e-mail message to a NOC, and employ Cisco AutoNotify services to directly generate a case with the Cisco Technical Assistance Center (TAC). This feature is a step toward autonomous system operation, enabling networking devices to inform IT when a problem occurs and helping ensure that the problem is acted on quickly, reducing time to resolution and maximizing system uptime.

GOLD

Cisco GOLD is a suite of diagnostic facilities to verify that hardware and internal data paths are operating as designed. Boot-time diagnostics, continuous monitoring, and on-demand and scheduled tests are part of the Cisco GOLD feature set. GOLD allows rapid fault isolation and continuous system monitoring.

Embedded Event Manager

Cisco Embedded Event Manager (EEM) is a device and system management technology built into NX-OS. Cisco EEM helps customers customize behavior based on network events as they happen.

Netflow

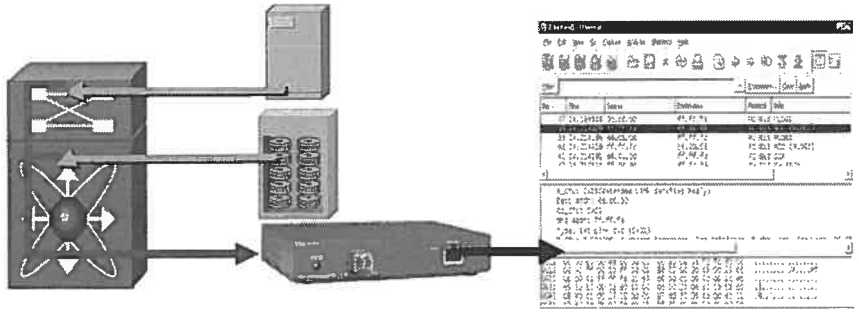
The Cisco NX-OS Netflow implementation supports version 5 and version 9 exports. It also supports the Flexible Netflow configuration model and hardware-based Sampled Netflow for enhanced scalability.

https://t.me/learningnets

SPAN Overview

SPAN is a proactive monitoring tool for debugging protocols and data analysis

- Non-intrusive
- Copies traffic to SPAN port



The Switched Port Analyzer (SPAN) utility can be used to perform detailed troubleshooting or to take samples of traffic from a particular application host for proactive monitoring and analysis. This utility is most helpful when you have a protocol analyzer available and you are monitoring user traffic between clients and servers. When there is a problem in a storage network that you cannot solve by fixing the device configuration, troubleshoot the protocol level. Debugging commands report control traffic between the switch and nodes, but to view all traffic originating from or destined to a particular end node (such as a host or disk), use a protocol analyzer to capture protocol traces. Filters can be applied to capture specific ingress or egress traffic.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch SPAN Summary

- 2 bidirectional SPAN sessions per Nexus switch
- Session configuration through sub-SPAN mode
- Support for VLANs, L2, and L3 interfaces
- Mix of source interface types permitted for a single span session
- Support for virtual SPAN using trunks with allowed VLAN lists
- VLAN filters supported

The Switched Port Analyzer (SPAN) feature allows the administrator to analyze all traffic between ports by non-intrusively directing the traffic to a SPAN destination port that has an external analyzer attached to it. Cisco Nexus 7010 Switch is now supporting:

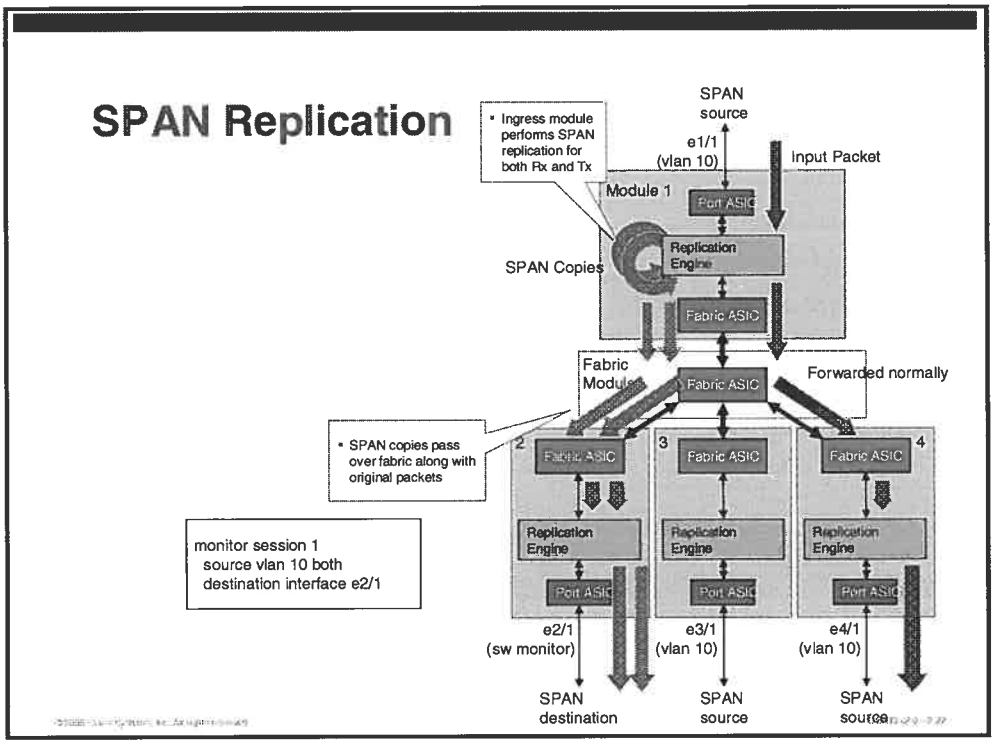
- 2 bidirectional SPAN sessions per system (globally, not per-VDC), where SPAN source and destination exist on the same switch (if it is not RSPAN). Traffic monitored (captured) can be interface traffic that is in transmit, in receive, or in both directions.
- Session configuration is available through the sub-mode—SPAN configuration mode.
- Support for VLANs, L2 and L3 interfaces
- Mix of source types permitted in a single session, where for example SPAN source of VLAN 10 and interface Ethernet 1/1 are in same session. Possible SPAN sources are Switchports (Access ports, Trunk ports, Private VLAN ports, Port-channels), Routed interfaces (Physical interfaces, Port-channels), VLANs and PVLANs, Supervisor inband interface (permitted from default VDC only; will show traffic for all VDCs.)
- Support for virtual SPAN using trunks with allowed VLAN lists. Allows per-VLAN filtering to group of SPAN destinations. Supports single SPAN session with list of VLANs as source or multiple destination ports. When each destination is configured as trunk with different allowed VLAN list, a subset of SPANned VLANs is sent to each destination.

SPAN destinations must be explicitly configured as monitor destination ports. Up to n destination ports are supported per SPAN session. SPAN destination can be Ethernet interface or port-channel interface, where only "on-mode" port-channel is supported for SPAN destination. SPAN destination can be configured as 802.1q trunks to support virtual SPAN.

VLAN filters allow monitoring a subset of VLANs on trunk ports. The filter specifies a list of VLANs to capture and the traffic for other VLANs is not sent to SPAN destination (for example, the filter can capture just VLAN 55 and 56 traffic from trunk on interface e2/4).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



SPAN involves creating a copy of each packet from the SPAN source. The ingress module performs SPAN replication for both Rx and Tx SPAN. Replication engines on the ingress module perform replication. Replicated copies consume forwarding engine and fabric bandwidth. Replicated copies and original traffic are subject to drop independently.

Note The CRC corrupted frames are not copied to the span destination ports, therefore SPAN method is not appropriate for discovering the low-level protocol problems but is effective for upper layer protocol analysis. For that kind of analysis, the RMON features can be utilized.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

RSPAN

- RSPAN source/destination sessions not supported
- RSPAN VLANs can be created
- RSPAN traffic from other switches can be received on trunk ports
- RSPAN VLANs can be source for a SPAN session
- No ERSPAN support yet

Remote SPAN (RSPAN) allows you to remotely monitor traffic for one or more SPAN sources that are distributed in one or more source Cisco Nexus 7010 Switches. The RSPAN feature is non-intrusive and does not affect network traffic switching for any SPAN source ports. Traffic captured on the remote switch is tunneled across the Nexus network to the RSPAN source/destination sessions not supported, but RSPAN VLANs can be created. RSPAN traffic from other switches can be received on trunk ports and trunking must be enabled on all ISLs in the tunnel transmission path. RSPAN VLANs can be source for a SPAN session.

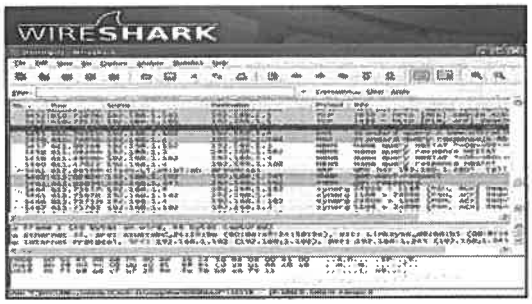
Note ERSPAN is not currently supported by Cisco Nexus 7010 Switch.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NX-OS Wireshark Overview

- Used to be known as Ethereal®
- Wireshark® is the world's foremost network protocol analyzer
- NX-OS offers an integrated packet capture tool called Ethanalyzer
- Ethanalyzer cannot capture data plane traffic being forwarded in hardware



Wireshark® used to be known as Ethereal (open source network protocol analyzer). It is the world's foremost network protocol analyzer and is the standard across many industries and educational institutions. NX-OS leverage on these characteristics and offers an integrated packet capture tool for control plane packets only. The system was built on top of Wireshark and is called Etheranalyzer. Etheranalyzer is a wrapper over Tshark, the terminal-based version of Wireshark. NX-OS runs on top of a Linux Kernel, which supports packet capturing using libpcap library. Wireshark decodes packets captured by the libpcap library.

Etheranalyzer can:

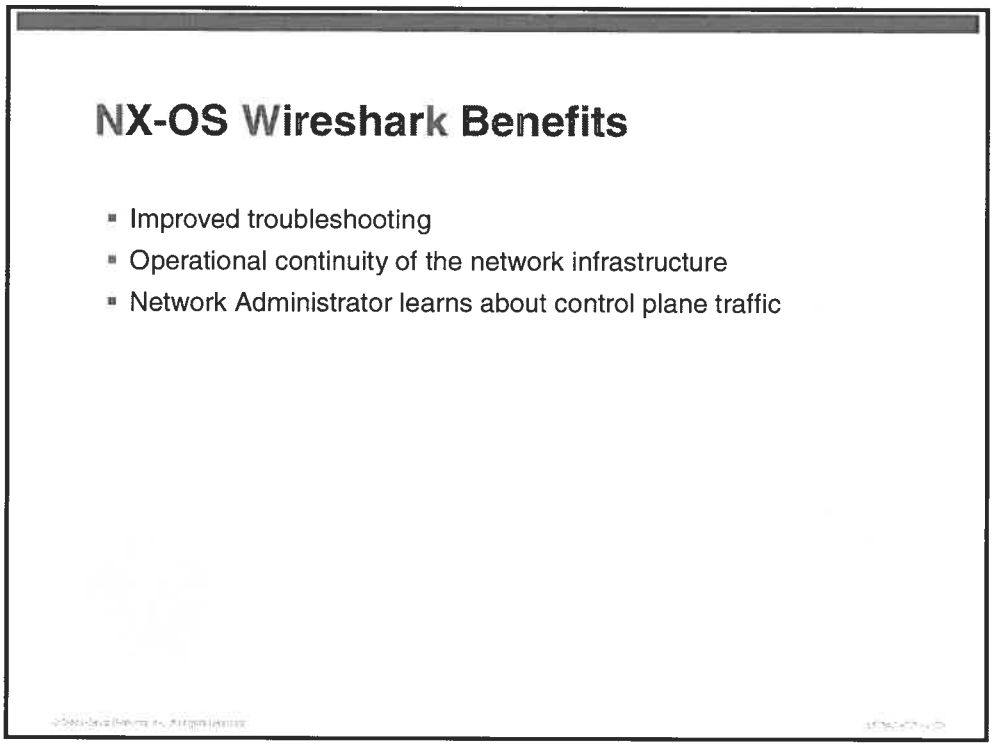
- Capture packets sent/received by the supervisor module (no data packets)
- Open and/or save packet data captured
- Decode the internal Cisco Nexus 7010 switch header of control packet

Etheranalyzer cannot capture the data plane hardware forwarded traffic.

The packets captured can be displayed in a very detailed way or just in one line per each packet with the most important fields of the packet. The capture can be saved in one of the storage device that are available in the system and can be read later on. Filtering of packets captured and/or displayed with very flexible filter expressions is available.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The main reasons for using Etheranalyzer from a customer stand point are:

- Etheranalyzer, as part of our integrated manageability toolset, improves troubleshooting, thus reducing time-to-resolution
- Etheranalyzer helps preserve and improve the operational continuity of the network infrastructure
- Etheranalyzer achieves the above two statements by allowing the network administrators to learn about the nature of the control plane traffic flowing through their network.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

NX-OS Wireshark at Work

```
NX-OS-1(config)# ethanalyzer local limit-captured-frames 1
Frame 1 (60 bytes on wire, 60 bytes captured)
    Arrival Time: Nov 11, 2007 12:27:44.73
    [Time delta from previous captured frame: 0.12 seconds]
    Frame Number: 5
    Frame Length: 60 bytes
    Capture Length: 60 bytes
    [Frame is marked: False]
    [Protocols in frame: eth:arp]
Ethernet II, Src: 00:c0:b7:78:d7:55 (00:c0:b7:78:d7:55), Dst: ff:ff:ff:ff:ff:ff (ff:ff:ff:ff:ff:ff)
    Destination: ff:ff:ff:ff:ff:ff (ff:ff:ff:ff:ff:ff)
        Address: ff:ff:ff:ff:ff:ff (ff:ff:ff:ff:ff:ff)
        .... ...1 .... .... .... .... = IG bit: Group address (multicast/broadcast)
        .... ..1. .... .... .... .... = LG bit: Locally administered address (this is NOT the factory default)
    Source: 00:c0:b7:78:d7:55 (00:c0:b7:78:d7:55)
        Address: 00:c0:b7:78:d7:55 (00:c0:b7:78:d7:55)
        .... ...0 .... .... .... .... = IG bit: Individual address (unicast)
        .... ..0. .... .... .... .... = LG bit: Globally unique address (factory default)
    Type: ARP (0x0806)
    Trailer: 4B58EA4A50140000DE8F0000020405B40101
Address Resolution Protocol (request)
    Hardware type: Ethernet (0x0001)
    Protocol type: IP (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (0x0001)
    Sender MAC address: 00:c0:b7:78:d7:55 (00:c0:b7:78:d7:55)
    Sender IP address: 172.25.45.41 (172.25.45.41)
    Target MAC address: 00:c0:b7:78:d7:55 (00:c0:b7:78:d7:55)
    Target IP address: 172.25.45.1 (172.25.45.1)
```



The diagram shows a detailed view of the output from the NX-OS device. By specifying a brief keyword the details can be observed. At the same time very powerful filtering expressions can be used to limit the output.

The output shows one frame with its arrival time, difference in time with the previous packet (0.12 seconds), frame number (5), length (60 bytes), and protocol in frame (Ethernet frame - ARP). In ARP frame source (00:c0:b7:78:d7:55) and destination (ff:ff:ff:ff:ff:ff) MAC address is shown followed by the source (172.25.45.41) and destination (172.25.45.1) IP address.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Call Home Overview

- Also known as "Phone Home"
- First introduced by IBM
- E-mail-based notification of system critical events
- Automated communication between software or hardware
- MDS was the first Cisco platform
- Catalyst 6500 took the Call Home concept from MDS
- Interaction with GOLD and EEM



Call Home is also known as "Phone Home", as it was traditionally implemented through POTS before the Internet became the prevailing notification tool. The feature first arrived in the IBM world on their mainframe systems many years ago. Call Home provides automated communication between software and hardware installed in the customer's environment and their vendors or manufacturer. The first Cisco platform with Call Home capabilities built-in were Cisco MDS switches. Cisco Catalyst 6500 (with Whitney1 release) took the Call Home concept from MDS, and worked with CA to create backend message parsing rules and defined corresponds for different event types to make it "Smart".

In the storage world downtime and availability are critical issues. The Call Home feature is trying to minimize these issues. Most of the software and hardware in the PC world (Microsoft, Adobe, etc.) actually has this feature. After a crash the error log is created and the users are asked if they want to send it back to the developers. Because Cisco MDS and Catalyst 6500 began using the feature it became bundled as part of SmartNET. This is where the Smart Call Home name originated.

The notification transport is email only. There are three types of message notification formats supported:

- XML format (for XML parsing applications)
  - Used for communication back to Cisco backend, providing the messages in Adaptive Messaging Language (AML) format.
- Full Text format
  - Used if you want to email to yourself.
- Short Text (human readable format)
  - Used for a pager, where user can be notified with a short message.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

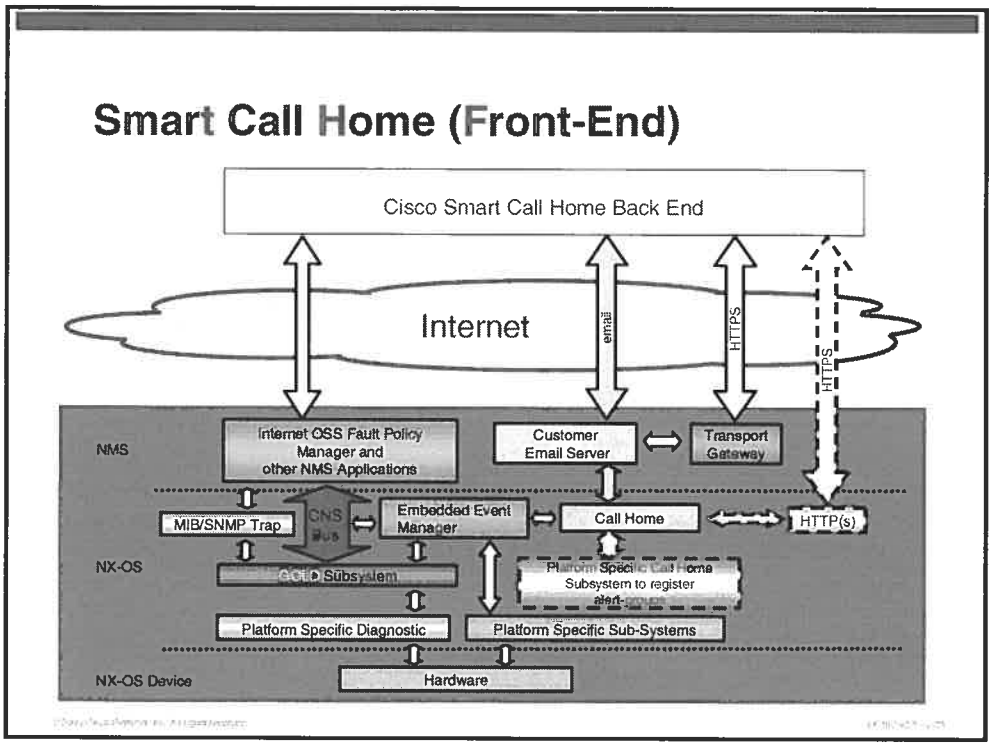
Note Call Home is tightly coupled with GOLD and EEM. Most of the hardware system level diagnosis will contact the Call Home through EEM. All the software components supporting Call Home feature also have a tie-in with the Call Home subsystem to notify if there is any problem or even a software process crash.
Call Home is also integrating the Syslog and the RMON messages.

Features not yet supported in NX-OS:

- User configurable periodic configuration messages
- 1-to-1 mapping between Alert-Group and Message Type
- Notification sent through HTTPS
- Proxy support for HTTPS

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



A hardware-specific subsystem is monitored by EEM and constantly looks at system problems that might not be covered by GOLD. GOLD is also running different tests to monitor the hardware components and proactively tell EEM there may be a problem. For example, a module might be overheating and an ASIC might experience some parity errors as a result.

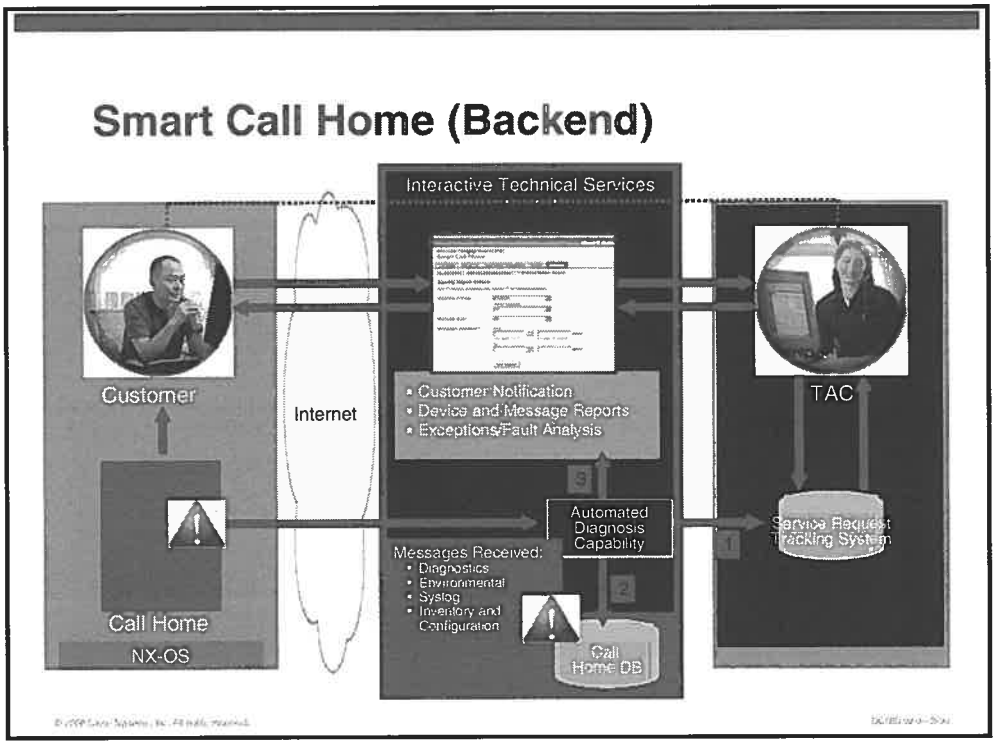
EEM communicates all of these events to the Call Home. There is a platform-specific subsystem representing each subsystem (routing, other software subsystems that actually register through Call Home, etc.). If a process crashes, it can notify the Call Home as well to send the information to the user. Even though there may be noticeable on the data plane, the administrator may want to take a look into the information. Through the information provided by the system, the administrator may choose to contact the TAC to see if there is any preventative activity needed.

Currently, email is only mechanism to send this information out of the NX-OS system and, as shown on the diagram, the transport gateway can be used to avoid direct email notification to Cisco.

The Transport Gateway is a piece of software running on any PC. The NX□OS system can email all its information to that transport gateway and from there all the information can be sent through secure HTTP to the Cisco backend. Additional future improvements will allow the system to directly send HTTPS information to the Cisco backend, as well as provide an HTTP proxy inside the Call Home feature..

© 2008 Cisco Systems, Inc.

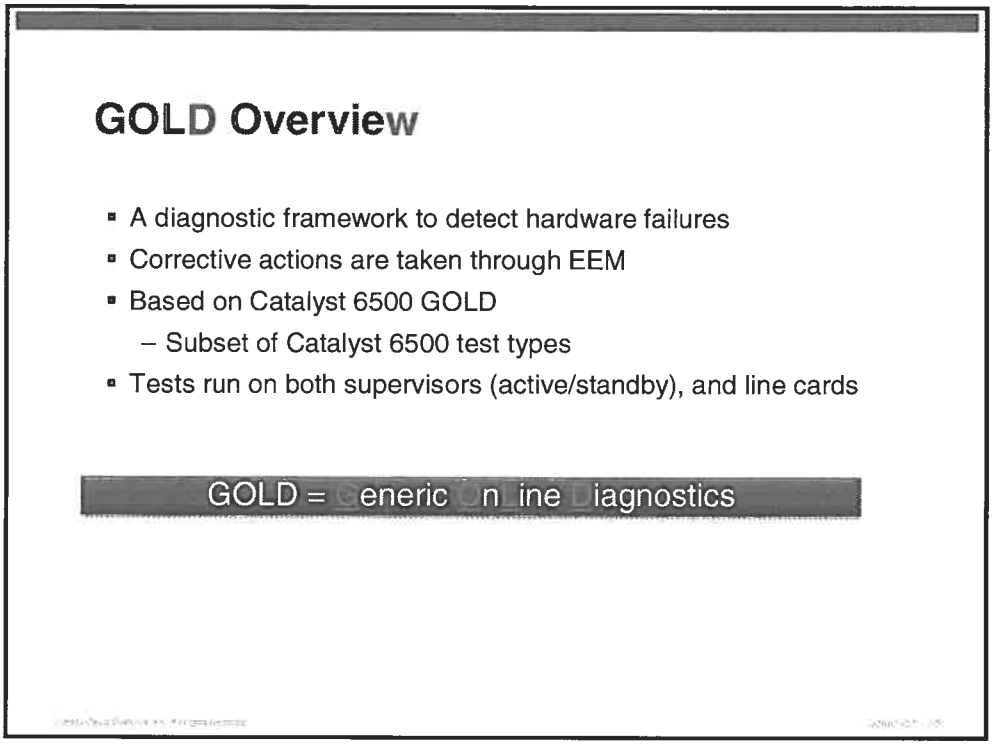
https://t.me/learningnets



From a user perspective, it is important to find out if something is going on with the NX-OS device and in case of a problem or some system level error (being either a process crash, a hardware component failure, a fan failure, a power supply failure, etc.) a user has the choice to either email himself (through a pager or direct email) or to contact Cisco backend. The backend has an automatic diagnostic system. The backend server parses through the information, identifying the severity and possibly opening a TAC case automatically for the customer. On the other hand, if it is a minor issue the solution is known and the customer can fix it quickly. The built-in database is available, which the backend server uses for an information comparison. This is the way to find out what kind of corrective action the user might need to take or double-check before further troubleshooting an issue. This information is sent back to the user as well, where all the past events are logged with the known serial number. The historically logged events on the backend server are available through the user CCO ID. TAC can go through the same tool to see what has been happening on the device.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Generic Online Diagnostics (GOLD) for NX-OS is a diagnostic framework used to detect hardware failures while the system is online and operational. It runs many tests (during boot and during the switch operation) and is largely based on the Cisco Catalyst 6500 GOLD, where the biggest difference is the amount of test types. The NX-OS array of tests is just a subset of those available in Cisco Catalyst 6500 and more tests will be added in the future. Test types currently available are Bootup, Health Monitoring, On-demand and Scheduled. All the tests are run to verify that the hardware is operating normally. Corrective actions are taken through Embedded Event Manager (EEM) policies, where thresholds are tunable by the administrator. The tests run on both supervisors—both the active and the standby—as well as on the line cards. GOLD is not supported on the CMP which is important to know. There is no online diagnostic for the CMP itself. GOLD is shipped with default EEM policies. Common default action for all GOLD EEM policies is to disable the test (if possible at device/instance level) and to trigger corrective action (for example, error should disable ports or reload the card). GOLD commands must be used in default VDC only.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

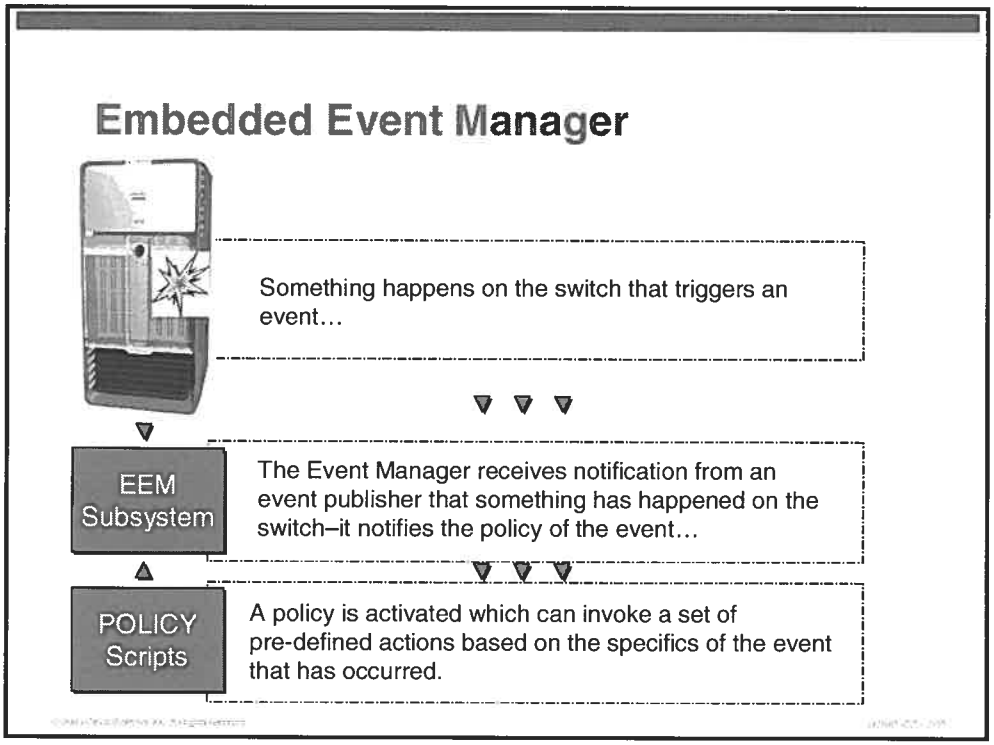
GOLD Call Home Integration

- Integration with Call Home to generate diagnostic alerts
- Call Home triggered when consecutive-fail-threshold for a test matches the configured GOLD failure threshold
- Diagnostic Call Home can also be triggered through CLI

© 2008 Cisco Systems, Inc. All rights reserved.

As previously mentioned, GOLD also has integration in NX-OS with the Call Home facility which has the ability to send outbound messages based on those diagnostic results. There are separate GOLD failure thresholds that are used for trigging Call Home, and those can actually be set and modified. However, with those different thresholds you can have one set of thresholds to take an action and a different set of thresholds to send a Call Home message. Call Home is triggered when consecutive fail thresholds for certain tests are exceeded, and these are independent of EEM policies. EEM policies dictate the action taken when this happens in GOLD. Call Home can also be used as an action within a GOLD EEM script. Diagnostic Call Home can also be triggered through CLI.

https://t.me/learningnets



Embedded Event Manager (EEM) is a centralized event management tool where events received can be managed by applying centralized policies to the events. That is the goal of an EEM. EEM in NX-OS is distributed across all the modules. One EEM process is running on the supervisor and one EEM process is running across each of the modules (an EEM client). EEM client allows events to be handled locally on the respective line card module and frees up Supervisor resources from managing event publication and executing policies across the chassis. This "distributed" architecture further improves (minimizes) latency of time to invoke EEM policy execution.

The standby supervisor can take actions, as well. These actions are limited to the actions that are generated out of the standby supervisor, but if it is something beyond that, then it has to go to the main supervisor.

The steps of operation are:

Step 1 The event happens on a chassis on one of the modules

Step 2 The module where the event gets generated forwards it out to the event manager on the chassis.

Step 3 Event Manager can apply a policy, which can be a combination of actions that must be taken with respect to the event received.

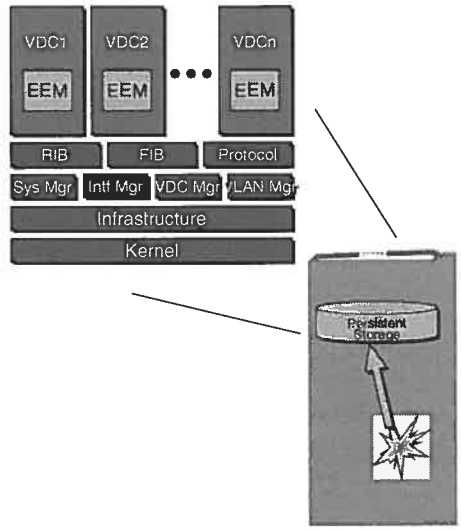
Sometimes some services done on the event might be sending a customized SYSLOG message out on an SNMP event or doing something more specific relevant to the event.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Embedded Event Manager (Cont.)

- Global EEM (default VDC)
- Each VDC has its own EEM
- Events are written to persistent storage

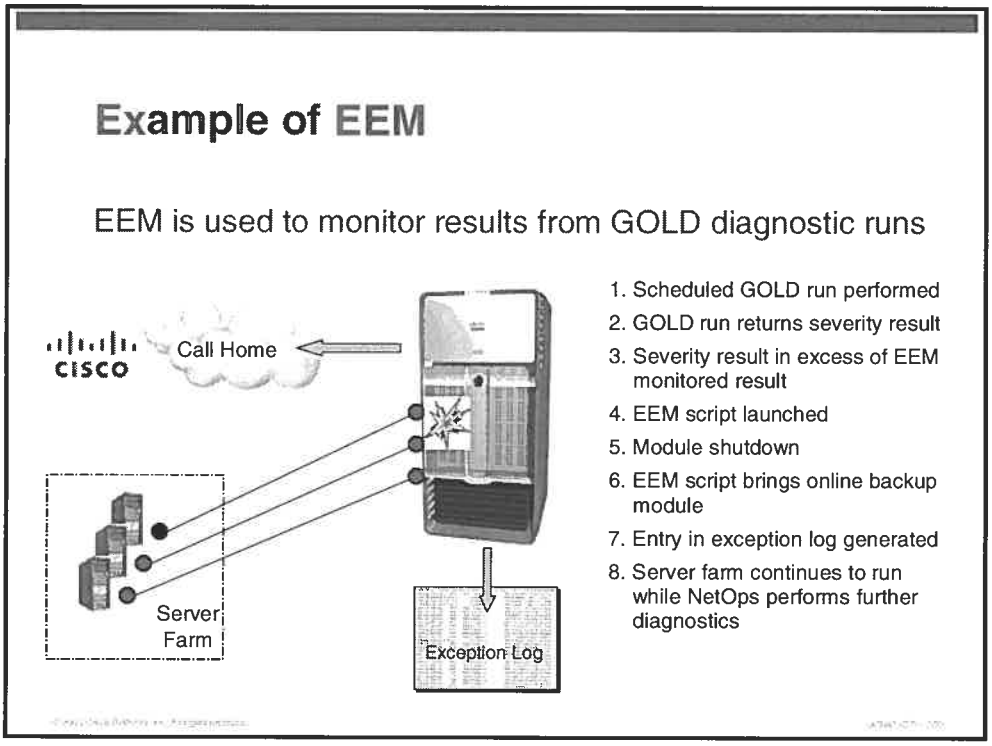


The EEM infrastructure in NX-OS that runs on the Supervisor is truly virtualized. EEM policies work with VDCs. Each VDC has its own instance of EEM. Some EEM publishers can be global (default) VDC aware and can share resources across multiple VDCs, like GOLD. Each line card has a single instance of EEM running. The line card modules can support multiple VDCs, but the events that rise out of modules are mostly of a global nature. Therefore, they only have one VDC running on each line card.

When events happen it is written into the persistent storage, where all the EEM events are archived. The size of the storage allocated for EEM on the NX-OS device will be 50 MB or 100 MB, which translates to about 1000 to 2000 messages. If more space is needed the administrator can perform a backup of this information periodically to a location outside the box. The 50 MB or the 100 MB of storage is not in one contiguous file; the information is broken into multiple files. Each file is no larger than 5 MB in size, which gives the administrator the ability to backup smaller chunks of files. There is no need to move large files out of the box every time.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Example

EEM is used to monitor results from GOLD diagnostic runs. In the event of a GOLD run returning a severity result in excess of a stated value, it will kick off a series of actions.

The steps are as follows:

Step 1 Scheduled GOLD run performed

Step 2 GOLD run returns severity result

Step 3 Severity result in excess of EEM monitored result

Step 4 EEM script launched

Step 5 EEM script gracefully shuts module down

Step 6 EEM script brings online backup module

Step 7 EEM script generates entry in exception log–sends call home to Cisco TAC

Step 8 Server farm continues to run while NetOps performs further diagnostics

Administrator performs a GOLD run. An event happens which surpasses the threshold limits, and the event is received by the EEM subsystem. Once the EEM subsystem receives the event, it launches the script. For this particular event, we want to shut down a module where we are having a problem and we want to start using the backup module. In the list of events, one of the things EEM does in action is shut down the malfunctioning module, and then it can also bring up a backup module in its place. The next step might be sending a Call Home message. Whatever events EEM collects over the period, it makes a log of those events and the administrator can look back to see what happened. The EEM log is different from the SYSLOG that is available on the switch. It is a completely different set of information, and custom SYSLOG messages can be sent out to the SYSLOG server. While EEM is in action and further diagnostics are performed, the server farm continues to run.

https://t.me/learningnets

NetFlow

- Flexible NetFlow is the latest Cisco NetFlow paradigm
- An optimal flow record for application
- A subset of fields implemented in the hardware table
- Monitors a wider range of packet information
- Enhanced network anomaly and security detection
- NetFlow terms:
  - Flow Monitor
  - Flow Record
  - Flow Exporter
  - Version 5 and 9



Flexible NetFlow is the latest Cisco NetFlow paradigm. It is a very flexible way of configuring NetFlow in the network or to define a flow record that is optimal for a particular application. Definition is done by selecting the keys from a large collection of pre-defined fields. All of the fields are not supported. A subset of Flexible NetFlow key and non-key fields based on the support is provided by the Forwarding Engine. Therefore, only subsets of fields, which are implemented in the hardware table, are supported.

The Flexible NetFlow-based configuration model includes:

- Create flow record
- Create exporter
- Combine flow record and exporter in flow monitor
- Tie monitor to interface

NetFlow offers the ability to monitor a wider range of packet information, producing new information about network behavior. Enhanced network anomaly and security detection is available, as well. NetFlow is configured on the interface, not globally, which is the best way to check just a couple of interfaces.

The NetFlow terms are:

Flow Monitor

A Flexible Flow Monitor is essentially a NetFlow cache. The Flow Monitor has two major components: the Flow Record and the Flow Exporter. The Flow Monitor can track both ingress and egress information. The Flow Record contains the information that is being tracked by NetFlow (ie, IP address, ports, protocol etc.). The Flow Exporter describes the NetFlow export. Flow Monitors may be used to track IPv4 traffic, IPv6 traffic, multicast or unicast, MPLS, or bridged traffic. Multiple Flow Monitors can be created and attached to a specific physical or logical interface. Flow Monitors can also include packet sampling information if sampling is required.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Flow Record

A Flexible NetFlow Flow Record defines what information NetFlow will track. The Flow Record may be user defined or a pre-defined scheme available in Cisco IOS Software. The Flow Record is defined as a set of key and non-key fields. Typical NetFlow key fields are IP addresses and ports, and if the set of key fields are unique, a new flow is created. The non-key field information is collected and attached to the flow. Typical non-key fields include timestamps, packet and byte counters and TCP flag information. Essentially, the Flow Record tells NetFlow what information to obtain from the packets being forwarded.

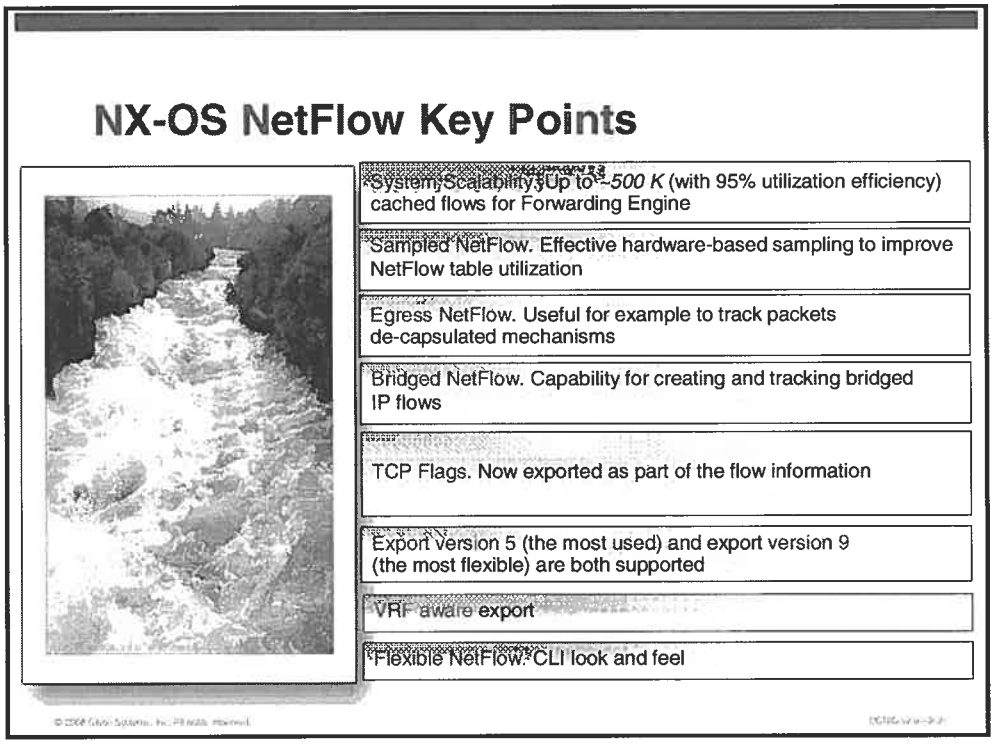
Flow Exporter

A Flexible NetFlow Flow Exporter describes information about the NetFlow export that is sent to the reporting server or NetFlow collector. The NetFlow exporter includes the destination address of the reporting server, the type of transport (i.e., UDP or SCTP), and the export format (i.e., version 9). There can be multiple exporters per Flow Monitor. The flexible NetFlow Exporter is QoS aware. The export stream will be prioritized with other traffic based on its class of service or DSCP value.

Version 5 and 9

A NetFlow exports information to reporting servers in various formats including NetFlow version 5 and version 9. NetFlow version 5 is used with traditional NetFlow and is a fixed export format with a limited set of information being exported. NetFlow version 9 is a flexible and extensible NetFlow format used by Flexible NetFlow. NetFlow version 9 includes a template to describe what is being exported and the export data. The template is periodically sent to the NetFlow collector telling it what data to expect from the router or switch. The data is then sent for the reporting system to analyze. NetFlow version 9 is extensible and flexible and therefore any data available in the device can theoretically be sent in NetFlow version 9 formats. Flexible NetFlow allows the user to configure and customize the information that is exported using NetFlow version 9. NetFlow version 9 is the basis for the IETF standard IPFIX associated with the IP Flow and Information working group in IETF.

https://t.me/learningnets

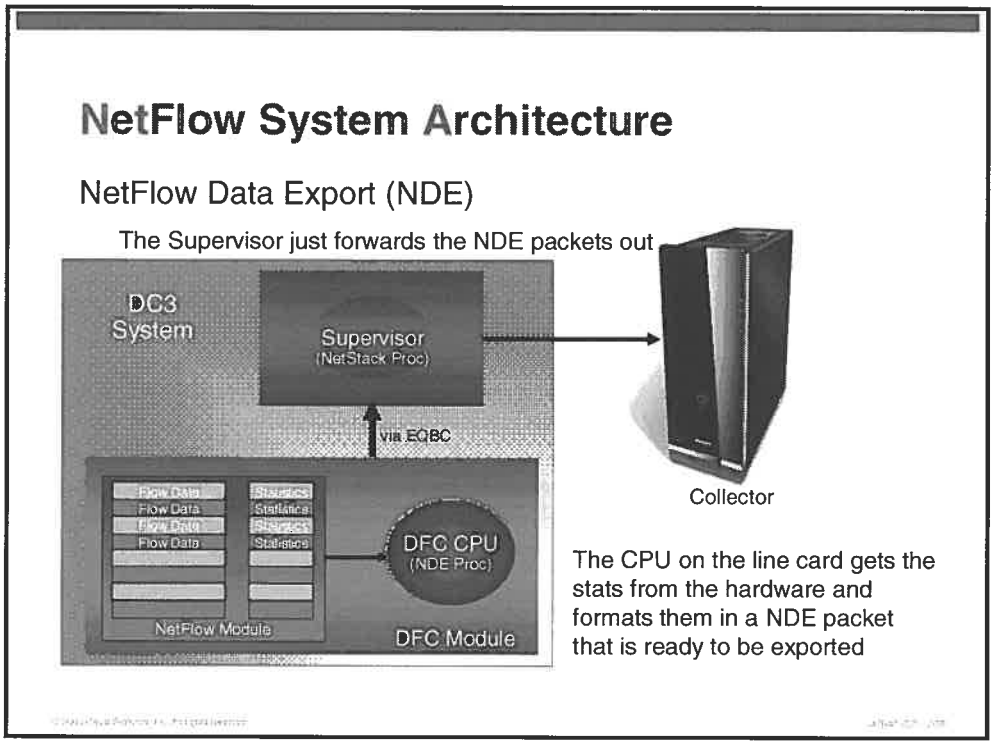


Systems scalability is up to 500 K cached flow per forwarding engine. With a fully-loaded chassis the Cisco Nexus 7010 Switch is able to cache four million flows. This is a big step compared to Sup720, where 128 K entries were supported. In the NetFlow table, the system stores only those packets that are sampled, so the NetFlow table will not be populated with a lot of useless information–effective hardware-based sampling to improve NetFlow table utilization. This is one of the advantages compared to Cisco Catalyst 6500, where all the flows go into the table and just those flows that are sampled are exported,.

Egress NetFlow and bridged NetFlow are supported, where egress NetFlow is used to track decapsulated packets and bridged NetFlow is used to create and track bridged IP flows. Finally TCP flags are supported, where they are exported as part of the flow information. This information is very useful to understand TCP flow directions and to detect Denial of Service (DOS) attacks. In terms of export version format, export version 5 and 9 are supported. These are the most used (version 5) and the most flexible (version 9) formats. Export version 5 is the default version. NetFlow exporting is VRF-aware. Specified destinations can be sent for the export and it is possible to define a VRF where the administrator wants to send the export. The programming method is important as well. The way to program NetFlow is completely different. It is based on flexible NetFlow which is a new paradigm in the Cisco NetFlow progression.

© 2008 Cisco Systems, Inc.

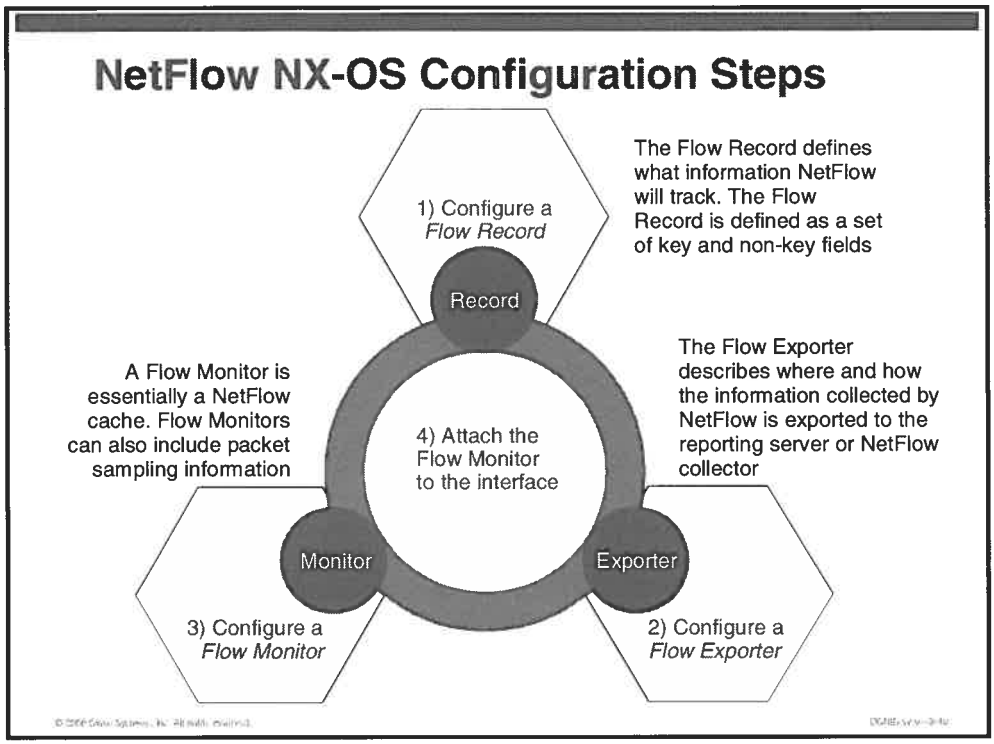
https://t.me/learningnets



The system flushes a NetFlow table. NetFlow Data Export (NDE) packets are formatted at the line card level, which helps the CPU on the supervisor. Then they are sent from line card hardware to the supervisor using the EOBC channel. The supervisor forwards the packets out of the system. The goal for the next release is to be able to export directly from the line card.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Steps needed to program the NetFlow in NX-OS:

Step 1 At the beginning, the administrator must define what must be captured in the NetFlow table. The Flow Record is defined, where the Flow Record is just a set of key fields and non-key fields. A key field is what the system is going to look for in a packet to see if it belongs to a certain flow. A non-key field is the additional information that the system is going to export as part of that flow where the system is going to send the flow to a NetFlow collector.

Step 2 The next step is a Flow Exporter configuration. The main questions are, "What does the administrator need do with these flows?" and "How do the flows get sent the NetFlow collector?"

Step 3 After successful configuration of the Flow Record and the Flow Exporter, a Flow Monitor must be defined. A Flow Monitor puts together this information: what kind of packets and flows must be collected and where to send the information.

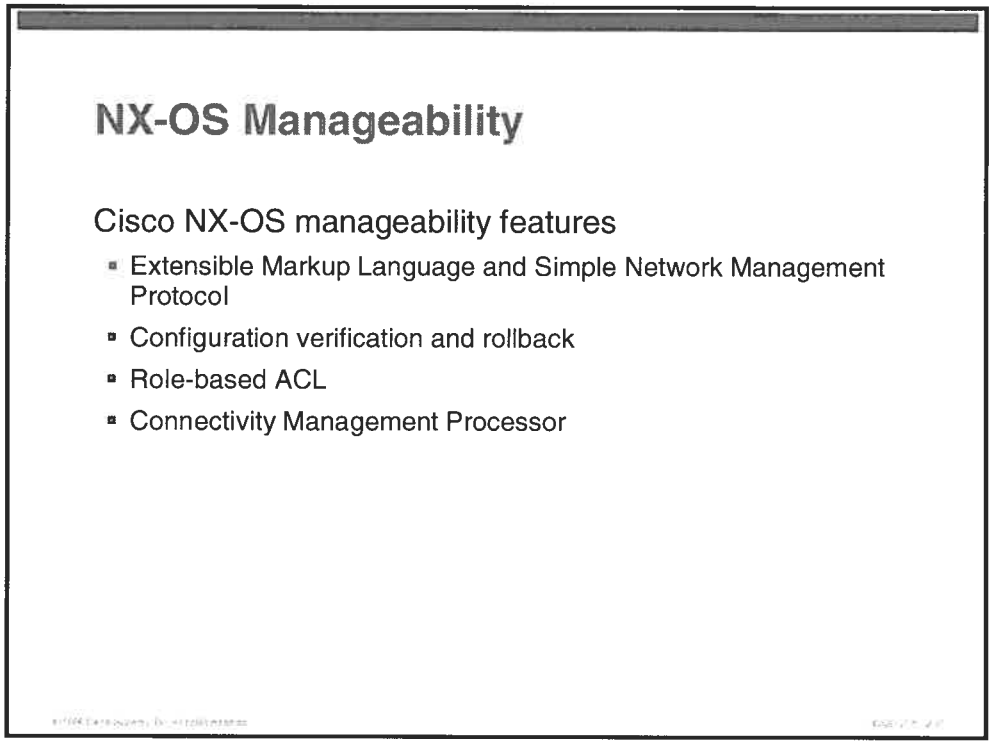
Step 4 The last step is to attach this Flow Monitor to an interface. What I mean is I go to an interface, attach it from a monitor that says what kind of packets I want to look at, what kind of flow I want to look at and where I need to send the information that I collected with the NetFlow.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Operating System Manageability

This topic describes the Cisco Nexus 7010 Switch Operating System Manageability.The object of this topic is to describe the Cisco NX-OS Manageability features as follows: Extensible Markup Language (XML) and Simple Network Management Protocol (SNMP), Configuration Verification and Rollback, role-based access control (RBAC) and the Connectivity Management Processor (CMP).



Cisco NX-OS is built with manageability functions that let you manage the Cisco Nexus 7010 system platform. The following features are included:

Extensible Markup Language and Simple Network Management Protocol

Cisco NX-OS is compliant with Simple Network Management Protocol (SNMP) version 1, version 2, and version 3. A rich collection of Management Information Bases (MIB) is supported. The same results can be achieved through XML-based configuration through NetConf. XML in NX-OS is called NetConf XML.

Configuration Verification and Rollback

With Cisco NX-OS, you can verify the consistency of a configuration and the availability of necessary hardware resources prior to committing the configuration. A device can be preconfigured and the verified configuration applied at a later time. Configurations also include checkpoints, to allow operators to roll back to a known good configuration as needed.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Role-Based Access Control

With role-based access control (RBAC), Cisco NX-OS enables you to limit access to switch operations by assigning roles to users. Administrators can customize access and restrict it to the users who require it.

Connectivity Management Processor

Cisco NX-OS supports the use of a Connectivity Management Processor (CMP) for lights-out remote platform management. The CMP provides an out-of-band access channel to the NX-OS console.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

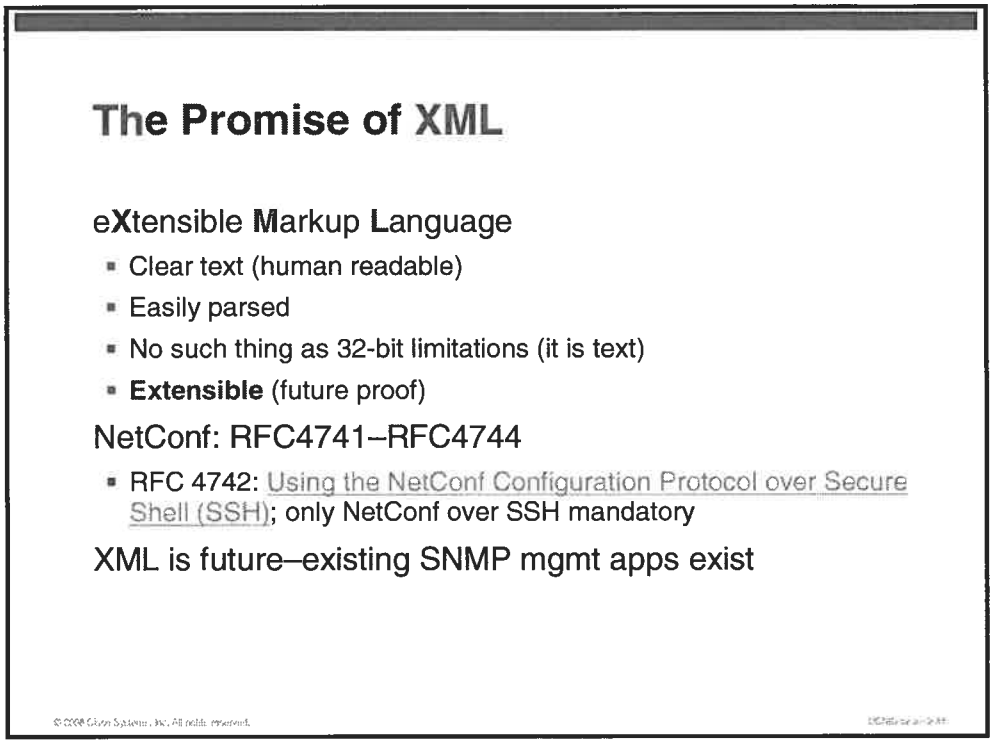
SNMP

- SNMP corrects the CLI scraping issues
- SNMP MIBs supported by NX-OS
- Historically, SNMP was not secure
  - SNMPv3 with AES addresses most of the security concerns (at the cost of CPU load)
- SNMP notificiation and traps are supported

SNMP was meant to fix the issues associated with CLI scraping. Supported SNMP MIBs are publicly available for download (http://tools.cisco.com/ITDIT/MIBS/servlet/index) and many management tools use them. Historically, SNMP was not secure—no confidentiality, integrity or strong authentication. In SNMP v1 and SNMP v2, all the commands and the outputs send back messages through plain text. There is no encryption and no security provided until SNMP v3. In SNMP v3, even though it provided the security through AES, it also put additional load on a CPU. CPU is encrypting and decrypting the messages. SNMP typically does not scale. UDP-based "get next/response" (even with bulk "gets") causes excessive CPU load and it is hard for a client to collect many statistics simultaneously. NX-OS platform code supports sending SNMP traps and notifications. At this stage, minimal SNMP traps are implemented.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The content of Extensible Markup Language (XML) is sent through clear text. Because of clear text, it is readable by humans. At the same time, XML is easy to parse because the benefit of XML is that every output field is tagged. Therefore, based on the tag, the administrator can easily find the information needed rather than looking into MIB definition first.

Since the output is acquired directly from the CLI, there is no 32-bit counter limitation. By defining a new tag, the administrator can basically expand the XML itself indefinitely. Therefore, it is really future-proof. Additionally, there has been a lot of effort going on in terms of RFC NetConf. It is actually the way we deliver XML-based schema to implement the network management through XML. There are actually a couple of NetConf RFCs, as follows:

- **RFC 4741**: NETCONF Configuration Protocol
- **RFC 4742**: Using the NETCONF Configuration Protocol over Secure Shell (SSH)
- **RFC 4743**: Using NETCONF over the Simple Object Access Protocol (SOAP)
- **RFC 4744**: Using the NETCONF Protocol over Blocks Extensible Exchange Protocol (BEEP)

Only NetConf over SSH is mandatory, so that is essentially how we address the security part of the issue. These days, SSH is implemented on most platforms. Through NetConf, you can basically have a XML-based polling to the device, and at the same time maintain the data security by not by putting overhead on the protocol itself. Therefore everything in the CLI is actually XML NetConf capable. XML NetConf is the management API moving forward. Any task in Config-mode can be done in XML and most tasks in Exec-mode can be done in XML, as well.

While even the biggest SNMP fans would probably agree that XML is the future, for better or worse, there are existing management applications which use SNMP today. NX-OS does not have feature parity with Cisco IOS MIBs. Many target customers indicated the MIB they want supported, which are not present today. Likewise, many management tools (Tivoli, Opsware, Smarts etc.) are expecting MIBs that are not there. There will be some short-term pain in either supporting the SNMP MIBs customers/partners expect and/or assisting them in migrating to the XML/Netconf interface.

https://t.me/learningnets

The Promise of XML + NetConf in Cisco NX-OS

- Anything in CLI is in XML/NetConf
- Screen-scraping CLI output no longer necessary
- Cisco NMTG future direction is XML/NetConf
- Sample output:

```
<rpc message-id="101"
   xmlns="urn:ietf:params:xml:ns:netconf:base:1.0">
  <get>
    <filter type="subtree">
      <t:top xmlns:t="foo">
        <t:interfaces>
          <t:interface t:ifName="Ethernet1/1"/>
          <t:interface t:ifName="Ethernet1/2"/>
          [..]
        </t:interfaces>
      </t:top>
    </filter>
  </get>
</rpc>
```

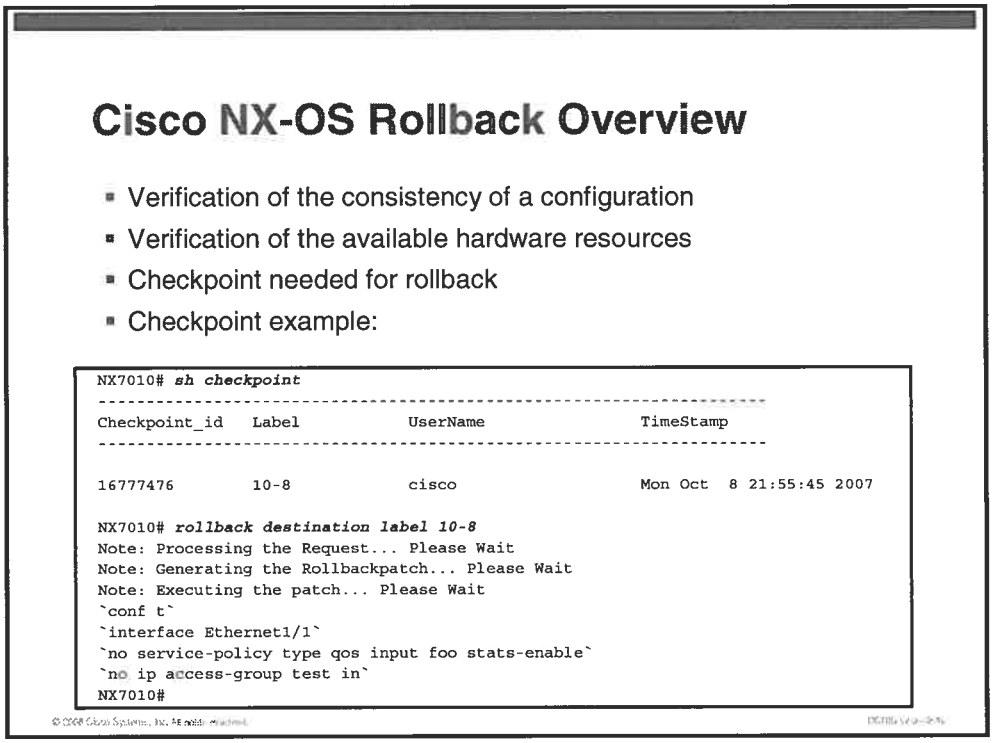
Anything in CLI is in XML/NetConf and the screen scraping output should no longer than necessary. There are still some improvements going on as some "show" commands might not have the right tag for every field, but at least the whole output is wrapped in a XML tag. Therefore, in that sense, you can get almost all the output on the NX-OS through XML.

From the NMTG perspective of view in the future the systems will gradually migrate from SNMP to the XML-based. In this way a consistent and scalable way to do the remote network management through a different protocol is provided. Thus from operation perspective, we can get a lot of the data very easily.

The sample output shows what data we can get. For everything there is a filter or a tag, which tells you exactly what the following information is. It is clearer now what stands behind human readable format. The sample output shows the tag being interfaced. For interface, its name is Ethernet1/1 and this representation is much easier then reading the SNMP OIB part of the information. NetConf sessions are based on TCP compared to the SNMP sessions, which are going though UDP, so this method is more reliable in the area of communication.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



With Cisco NX-OS, the administrator can verify the consistency of a configuration and the availability of necessary hardware resources prior to committing the configuration. A device can thus be preconfigured and the verified configuration applied at a later time. Configurations also include checkpoints, to allow operators to roll back to a known good configuration as needed. Configuration must be checkpointed to allow rollbacks when needed.

Checkpoint Example

NX7010# sh checkpoint

Checkpoint_id Label UserName TimeStamp

16777476 10-8 Cisco Mon Oct 8 21:55:45 2007

NX7010# rollback destination label 10-8

Note

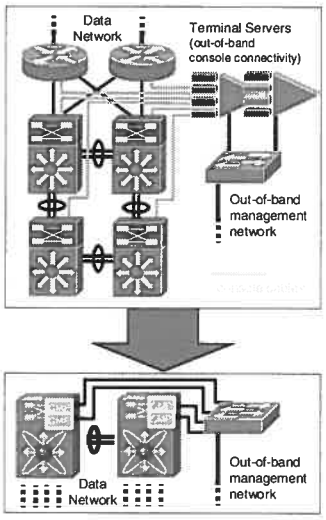
Processing the Request... Please Wait
Generating the Rollbackpatch... Please Wait
Executing the patch... Please Wait
`conf t`
`interface Ethernet1/1`
`no service-policy type qos input foo stats-enable`
`no ip access-group test in`
NX7010#

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Connectivity Management Processor (CMP)

- Standalone, "always on" management processor on supervisor engine
- Provides management connectivity via 10/100/1000 Ethernet interface
- Remote connection via telnet or ssh
- Runs lightweight Linux kernel and network stack
- Completely independent of NX-OS on main CPU



The purpose of a Connectivity Management Processor (CMP) is a standalone, always-on management processor on the Supervisor Engine. One CMP exists on the primary engine and one exists on the backup supervisor engine. The CMP provides secure lights-out management and is a 10/100/1000 interface, where either SSH or Telnet can be used to attach to it. CMP is not running a full version of NX-OS, it is running a lightweight Linux separate kernel and network stacks, which means that CMP is just dedicated to management activity tasks. It is important that CMP is completely independent of the NX-OS on the main CPU.

One of the CMP benefits is to remove the need for a separate terminal service for out-of-band management out of the box. CMP has a direct Ethernet connection from the management network switch to the CMP management interface and it can be used to watch the switch reboot and all other activities of the main CPU. Terminal service can still be used if needed as there is a console port on the Supervisor for doing that. The CMP is an example of how servers are used today, with the out-of-band management interface.

The primary functions of the CMP are:

- To provide remote connectivity to the Cisco Nexus 7010 Switch Supervisor card and aid troubleshooting if the system is otherwise not reachable (somebody erases a flash image by accident and after reboot the device did not come up again).
- To monitor the health of the Central CPU complex on the Cisco Nexus 7010 Switch Supervisor card.

The functions supported from within CMP are:

- Capability to monitor (view) the output on the Supervisor console port
- Capability to log the output on the Supervisor console port
- Capability to take full control of the Supervisor console port
- Monitor Supervisor health and reset Supervisor control plane if it hangs

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

- If Supervisor is reset/non-operable:
  - Read the Supervisor OBFL (on board failure logs)
  - Capability to access line card console(s)

CMP is a separate unit in the Supervisor engine. It has its own dedicated CPU and Linux OS, it has its own Ethernet port, its own DRAM, own flash and resides in its own power domain. CMP can also access the system controller and do certain things on a system.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Capabilities of CMP

- The CMP is a standalone processor on each Supervisor module with:
  - RAM and boot flash
  - Front panel 10/100/1000 Ethernet port
  - Access to the I/O (console/serial) ports and power mgmt
- Only "admin" user access is available
- CMP complex is part of the Supervisor module but in a separate power-domain–always On
- CMP can monitor Supervisor console port
- CMP can access all line card modules

The CMP is a processor on each Cisco Nexus 7010 Switch Supervisor. It has its own RAM and boot flash, front panel 10/100/1000 Ethernet port and its own access to the I/O (console/serial) ports and power management. It does not have its own console port, but it has access to each Cisco Nexus I/O and console serial ports of the route processor, control processor and all the line cards and it can do power management for the system. The primary functions are to provide remote connectivity to the Supervisor to aid in troubleshooting if the Cisco Nexus 7010 is not reachable. CMP is able to monitor the health of the central CPU (control plane) complex on the Supervisor card.

Only "admin" user access is available on CMP and only simple username/password authentication is used. CTS are not supported. Authentication process checks if the central CPU of the Supervisor is available. In this case, the CMP forwards the authentication request to the control plane AAA module. If the central CPU does not respond, the CMP uses its local authentication database to authenticate the user.

CMP complex is part of the Supervisor module but in a separate power-domain. It is always On. CMP can power-cycle, via the Power Management FPGA (PM-FPGA), the entire Supervisor module and not power-cycle itself. Because the CMP is in a different power domain, it does not get reset. The Supervisor power status can be monitored by the CMP via the PM-FPGA.

The CMP can monitor the Supervisor Console via the Mux. The Mux normally operates in passive mode and allows the CMP to monitor all output from the Supervisor Console. The Mux can take control of the Supervisor Console for input/output, so you can logically detach the Supervisor from the front panel console port and redirect it to the CMP and then reroute console I/O into an SSH session. The CMP cannot access consoles of the standby CMP and the standby Supervisor because it cannot get to them.

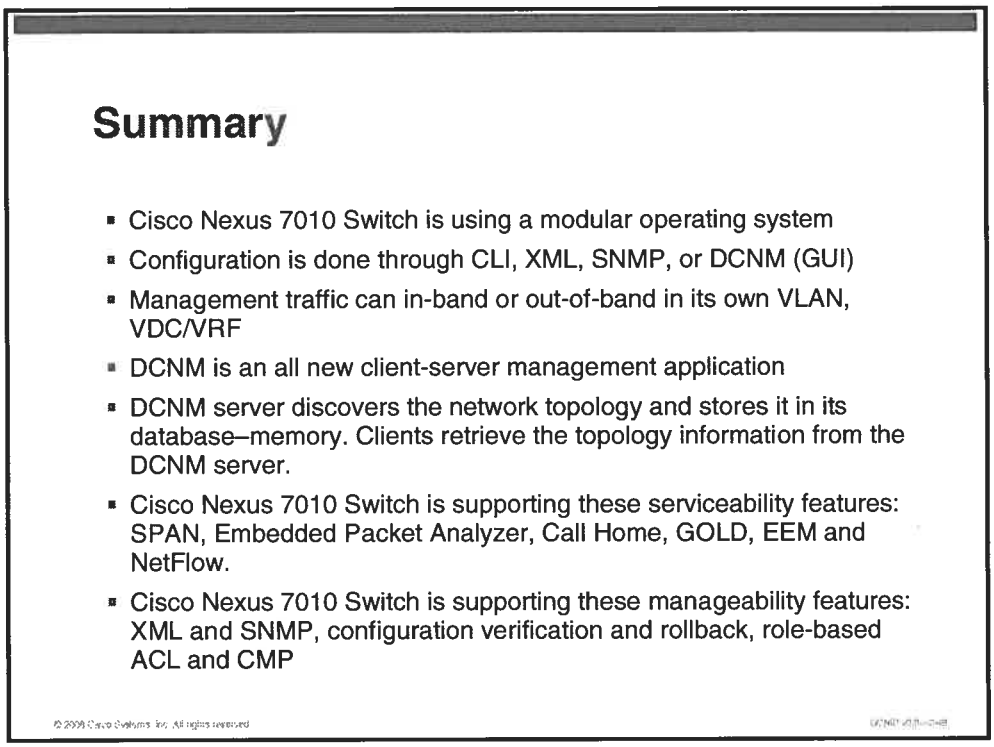
The CMP can access all the line card modules via the mid-plane bus, similar to the control plane. The normal way to do this is to access line card modules via control plane, but in case of problems on line cards, CMP has an opportunity to go directly via the mid-plane bus.

© 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

Lesson 5

Cisco Nexus 7010 Switch Security

Overview

This lesson describes the Cisco Nexus 7010 Switch Security features and implementation.

Objectives

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

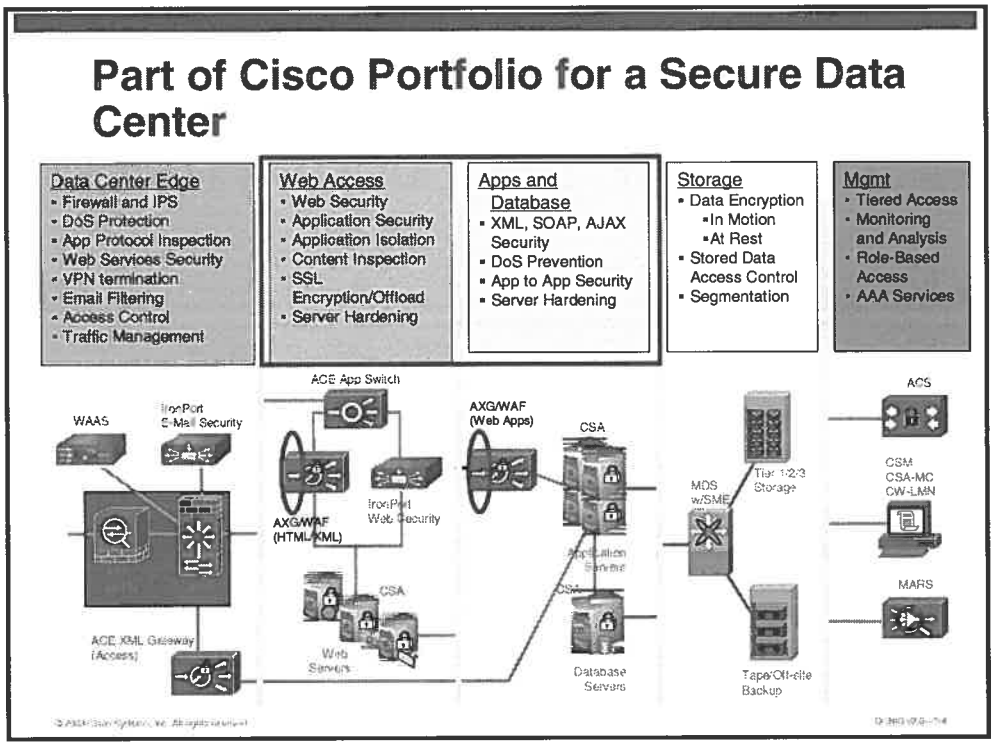
- Describe the Cisco Nexus 7010 Switch Security features

https://t.me/learningnets

Cisco Nexus 7010 Switch Security Overview

The objective of this topic is to describe the two configuration options that are available for Cisco Nexus 7010 Switch.

Cisco Secure Data Center



© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The diagram is presenting few big areas in NX-OS devices. This lesson addresses Cisco Nexus 7010 Switch Security and the following security features that exist under integrated security:

- CoPP
- Packet sanity checks
- uRPF
- DHCP-Snooping
- Dynamic ARP Inspection
- IP Source Guard
- Port Security
- Broadcast Suppression
- ACLs
- Cisco TrustSec

We have an extensive set of "sanity" checks done on the packet field headers, uRPF, DHCP-Snooping, Dynamic ARP Inspection, IP Source Guard, Port Security, and Broadcast Suppression.

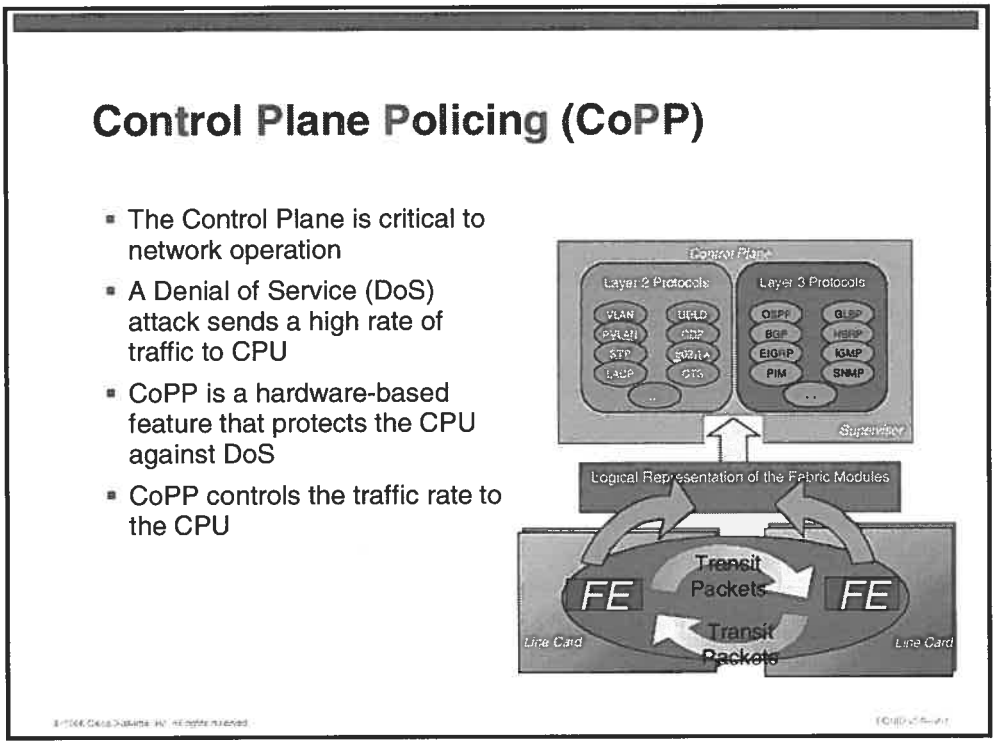
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco Nexus 7010 Switch Integrated Security Features

This topic describes the Cisco Nexus 7010 Switch Integrated Security Features.

Control Plane Policing (CoPP)



The Control Plane Policing (CoPP) is critical to network operation. Denial of Service (DoS) attack targeting the Control-Plane can be devastating to the network stability and availability, leading to business-impacting network outages. DoS targeting the control plane usually involves high rates of packets that are destined for the CPU. CoPP is a hardware-based feature that protects the Supervisor by controlling the rate at which the packet actually reaches the CPU. Control/Management Plane is not necessary perpetrated maliciously. CoPP is preventing any kind of high rates of traffic (destined for the CPU) that result in excessive CPU utilization.

Control Plane Policing leverages several functions that are provided by the Forwarding Engine. To protect the Supervisor, the CoPP:

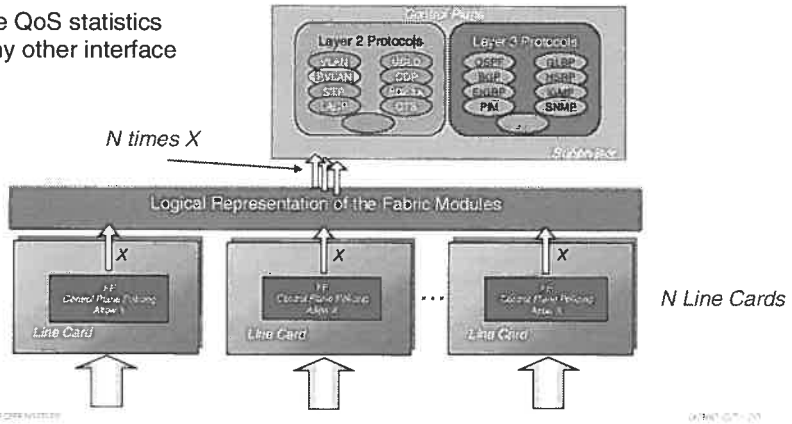
- Separates data plane packets from control plane packets (Exception Logic)
- Identifies DoS attack packets from valid packets (Classification)
- Once identified, marks, drops, or polices them (Service Policies)
- Provides DoS protection mechanisms independently from existing interface capabilities (global independent control-plane interface)
- Is based on MQC configuration

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS Implementation of CoPP

- CoPP in NX-OS is hardware based
- Total traffic is $n * x$ (n = number of line cards, x = max traffic per class)
- CoPP will support IPv4, IPv6, ARP and MAC ACLs
- Same QoS statistics as any other interface



NX-OS introduces new important enhancements for CoPP. Besides supporting the usual IPv4 and IPv6 "*receive*" traffic, CoPP will be capable of delivering the following functionality:

- Multicast Traffic (preserving the bridging on the ingress VLAN)
- ARP packets including ARP Broadcast Storms
- Layer2 Broadcast packets
- IP unicast with DMAC having I/G bit set
- Redirection to CPU of packets with IP Options
- Redirection to CPU of DHCP-Snooping, DAI, Port-Security packets

CoPP will support IPv4, IPv6, ARP and MAC ACLs and it will be able to match on packets generating exceptions and redirections. The rate in the policy-map can be configured as packet per second (pps); however, the statistics will still be shown in bytes per second (bps)

NX-OS implementation of CoPP is hardware-based only. There will not be any Control Plane Policing done in software by the Supervisor. When configuring the various rates for the policy-map the administrator must take into consideration the number of line cards in the system. The total traffic received by the Supervisor will be *N* times *X*, where *N* is the number of line cards and *X* is the rate allowed for the particular class.

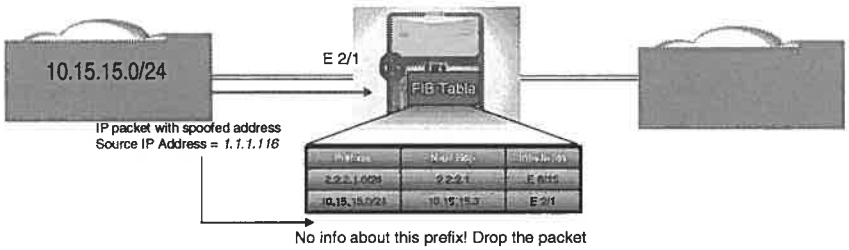
The CoPP support the same statistics that are available for any service policy on any other interface. If administrator performs show ACL command and then put the name of the ACL included in the class map, there will be the counters for each entry inside the ACL.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Unicast Reverse Path Forwarding (uRPF)

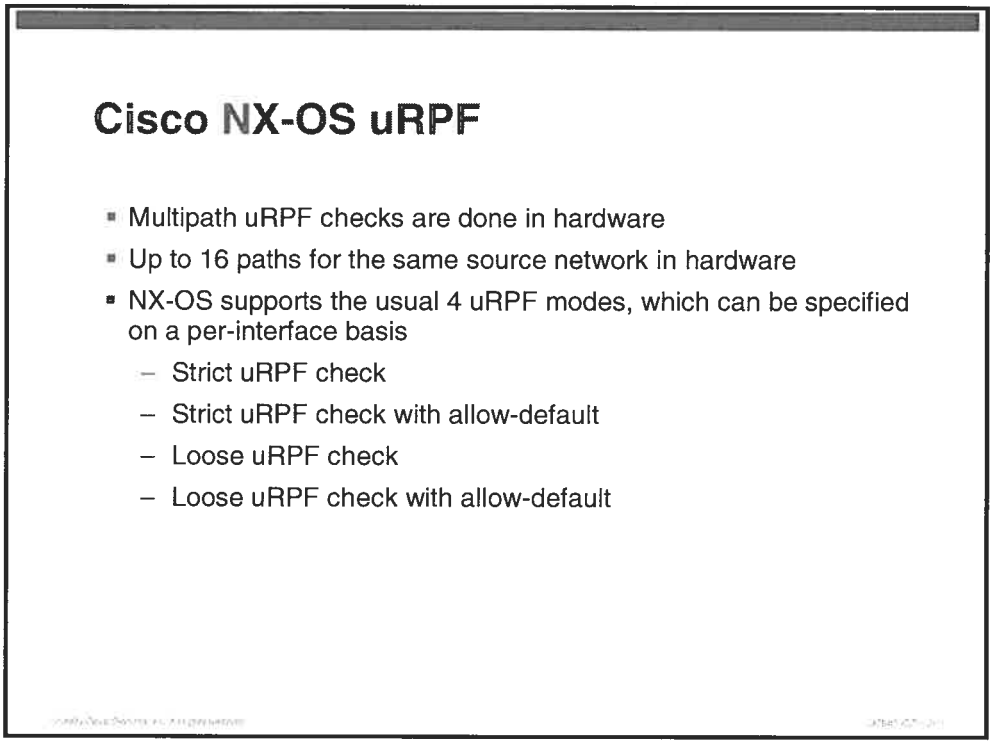
- Prevents against (DoS) attacks, including Smurf and Tribal Flood Network (TFN)
- Packet source addresses must be valid and consistent with the IP routing table
- Packets with unknown IP source addresses are dropped



Unicast Reverse Path Forwarding (uRPF) is a very well known feature. It takes care of attacks that leverage IP address spoofing or attacks where the IP address changes rapidly. An example of these attacks are (DoS) attacks, including Smurf and Tribal Flood Network (TFN), which can take advantage of rapidly changing source IP addresses to allow attackers to thwart efforts to locate or filter the attacks. uRPF forwards only those packets that are known by the system, where the IP address of those packets is consistent with information contained have in the IP routing table. When you forward the packets you look at the IP destination address of the packet to understand where the next-hop is. But with uRPF the system looks at the IP source address to make sure that this packet is supposed to be received from that interface and is not somebody that just came up with it here. uRPF performs a reverse path lookup for packets to ensure that their source IP addresses are known and installed in the local forwarding table. Packets with unknown IP source addresses are dropped.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Cisco Nexus 7010 Switch can perform multipath uRPF checks in hardware. The device can check up to 16 paths for the same source network in hardware. The number 16 derives from the number of ECMP the system can support. NX-OS supports the usual 4 uRPF modes and they can be specified on a per interface base:

- Strict uRPF Check
- Strict uRPF Check with allow-default
- Loose uRPF Check
- Loose uRPF Check with allow-default

Since 16 is also the maximum number of ECMP supported, there is no longer need for "punt", "pass" or "interface-group" modes. The packets failing the RPF check will be punted to the CPU. CoPP protects from high rate of such packets.

Strict uRPF Check

The Strict uRPF check verifies that traffic received by the system is sourced from a prefix that exists in the routing table *and* that the prefix is reachable through the interface the packet is coming from. Packets that do not meet these criteria are discarded. Strict uRPF check provides the greatest level of protection against spoofed or invalid source IP addresses.

Strict uRPF Check with allow-default

The Strict uRPF with allow-default adds an additional check against the valid default route (if it exists). The packets pass the uRPF check provided they are received on one of the reverse-path interfaces for the default route. If there is no default route present, this mode behaves the same as the strict uRPF check.

https://t.me/learningnets

Loose uRPF Check

The Loose uRPF Check method verifies only that traffic received by the system is sourced from a prefix that exists in the routing table, regardless of the interface on which the traffic arrives. Packets that do not meet these criteria are discarded

Loose uRPF Check with allow-default

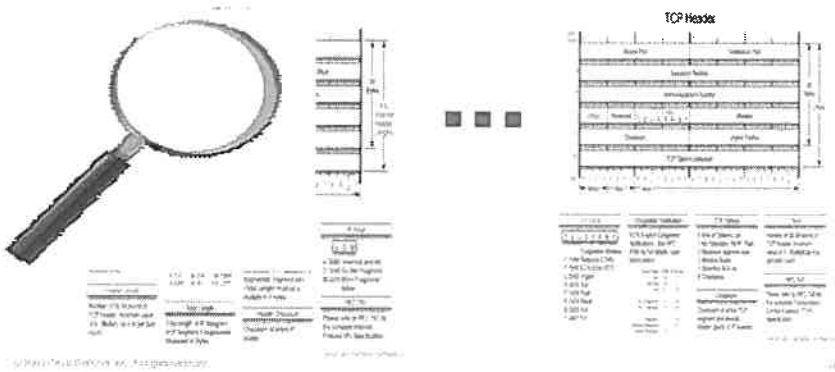
The Loose uRPF check with allow-default method is the same as not having uRPF check enabled at all, as long as there is a default route. If no default route exists in the routing table, the behavior is the same as loose uRPF check, in that if the source prefix does not exist, the traffic is discarded.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS Packet Sanity Checks

- Forwarding Engine performs sanity checks on the header fields of IPv4 and IPv6 packets
- Protection of the network and the system from "illegal" packets
- The IP sanity checks are enabled by default
- The packets failing the sanity checks are dropped



Cisco Nexus 7010 Switch Forwarding Engine performs sanity checks on the header fields of IPv4 and IPv6 packets to make sure that we do not allow to reach the Supervisor with packets that are illegal (for example, IP packets with version 10). NX-OS packet sanity checks protect the network and the system from "illegal" packets. The IP sanity checks are enabled by default and can be individually disabled. The packets failing the sanity checks will be dropped. Counters for each of these sanity checks are available as well. If the system is receiving some type of traffic and the administrator does not know about it, there is a problem with an application.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Main Sanity Checks on IPv4 Packets



IOS Check	Description	Supported	Service Module
IPv4 checksum	IPv4 checksum is verified for frames without options and frames with options.	YES	FWSM/ACE
IP header length minimum	The IPv4 header length (IHL) should be greater than or equal to 5, and less than or equal to the IPv4 Total Length.	YES	FWSM/ACE
Ethernet Frame length minimum	The Ethernet frame length should be greater than or equal to the IPv4 Total Length plus (14 + 4) octets, where the 4 is due to the CRC-32.	YES	FWSM/ACE
Fragment length maximum	The IPv4 payload length + (IPv4 Fragment Offset * 8) should be less than or equal to 65536 bytes.	NO	FWSM/ACE
Unexpected fragment	If the IPv4 DF flag is active, then the Fragment Offset should be 0	NO	FWSM
IP version	The version field should be 4 for an IPv4 Ethernet type.	YES but not elegant	FWSM
UDP length maximum	The UDP length should be less than or equal to the IPv4 payload length.	NO	FWSM/ACE
TCP length maximum	The (TCP Data Offset * 4) should be less than or equal to the IPv4 payload length, and the TCP Data Offset should be greater than or equal to 5.	NO	FWSM/ACE
TCP tiny fragment	If the IPv4 Fragment Offset is 0 and the IPv4 Protocol is TCP, the IPv4 payload length should be greater than or equal to a programmable minimum value. The programmable minimum value default is 16.	NO	ACE

Sanity checks performed on IPv4 packets:

- **IPv4 checksum**: IPv4 checksum is verified for frames with and without options.
- **IP header length minimum**: The IPv4 header length (IHL) should be greater than or equal to 5, and less than or equal to the IPv4 Total Length.
- **Ethernet frame length minimum**: The Ethernet frame length should be greater than or equal to the IPv4 Total Length plus (14 + 4) octets, where the 4 is due to the CRC-32.
- **Fragment length maximum**: The IPv4 payload length + (IPv4 Fragment Offset * 8) should be less than or equal to 65536 bytes.
- **Unexpected fragment**: If the IPv4 DF flag is active, then the Fragment Offset should be 0
- **IP version**: The version field should be 4 for an IPv4 Ethernet type.
- **UDP length maximum**: The UDP length should be less than or equal to the IPv4 payload length.
- **TCP length maximum**: The (TCP Data Offset * 4) should be less than or equal to the IPv4 payload length, and the TCP Data Offset should be greater than or equal to 5.
- **TCP tiny fragment**: If the IPv4 Fragment Offset is 0 and the IPv4 Protocol is TCP, the IPv4 payload length should be greater than or equal to a programmable minimum value. The programmable minimum value default is 16.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Other sanity checks performed on IPv4 packets:

- **Broadcast Source IP Address**: Source IP address is 255.255.255.255
- **Reserved IP Addresses**: Source IP address or destination IP address is 127.x.x.x.
- **Identical IP Destination and IP Source Addresses**: Source IP address is equal to Destination IP address, and they are not 0.0.0.0.
- **Destination IP Address is zero**: Destination IP Address is 0.0.0.0
- **Source IP is a Class D Address**: Source IP address belongs to this range: 224.0.0.0 - 239.255.255.255
- **Source IP or Destination IP is a class E**: Source IP address or Destination IP address belong to this range: 240.0.0.0 - 255.255.255.255

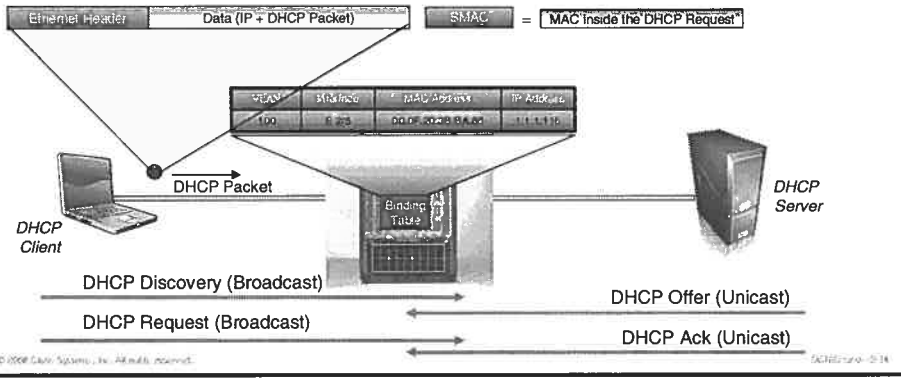
Sanity checks performed on IPv6 Packets:

- **Ethernet version 2 Frame length minimum**: The Ethernet frame length should be greater than or equal to the IPv6 layer length, plus IPv6 payload length, plus (14 + 4) octets, where the 4 is due to the CRC-32.
- **Fragment length maximum**: The (IPv6 Payload Length - IPv6 Extension Header Bytes) + (Fragment Offset * 8) should be less than or equal to 65536 bytes.
- **IP version:** The version field should be 6 for an IPv6 Ethernet type.
- **UDP length maximum**: The UDP length should be less than or equal to the IPv6 Payload Length.
- **TCP length maximum**: The (TCP Data Offset * 4) should be less than or equal to the IPv6 Payload Length, and the TCP Data Offset should be greater than or equal to 5.
- **TCP tiny fragment**: If the IPv6 fragment offset is 0 and the IPv6 protocol is TCP, the IPv6 payload length, should be greater than or equal to a programmable minimum value. The programmable minimum value default is 16.

https://t.me/learningnets

Cisco NX-OS DHCP Snooping

- DHCP is described in the RFC 2131
- The DHCP snooping validates DHCP messages received from untrusted sources
- The DHCP snooping builds and maintains a binding database
- DHCP snooping will also check the MAC address
- DCHP relay is supported



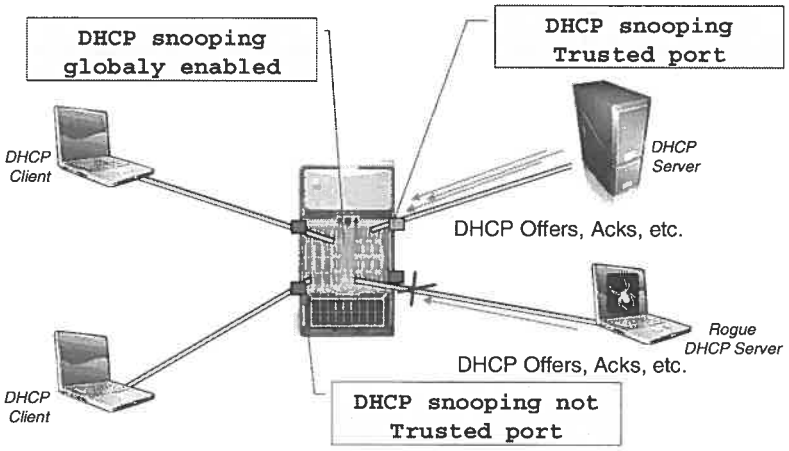
Dynamic Host Configuration Protocol (DHCP) is a protocol used to obtain IP addresses and other parameters such as the default gateway, subnet mask and DNS address from a DHCP server. DHCP is described in the RFC 2131. DHCP Snooping builds and maintains a binding database between the IP addresses and the MAC addresses that are in the DHCP request. The database contains information about untrusted hosts with leased IP addresses and is also used by other features like Dynamic ARP Inspection. By default NX-OS implementation of DHCP Snooping will also check that the MAC address inside the DHCP request matches the MAC address of the Ethernet frame carrying the packet. These are sanity checks that we do on DHCP packet and are a part of DHCP-Snooping feature. NX-OS supports DHCP Relay as well. It helps in relaying DHCP requests/replies between client and server when they are on different subnets. DHCP Server address needs to be configured on the SVI/L3 interface on which the client is connected.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS DHCP Snooping (Cont.)

- DHCP snooping assigns a level of trust to a switch port
- DHCP snooping validates DHCP message

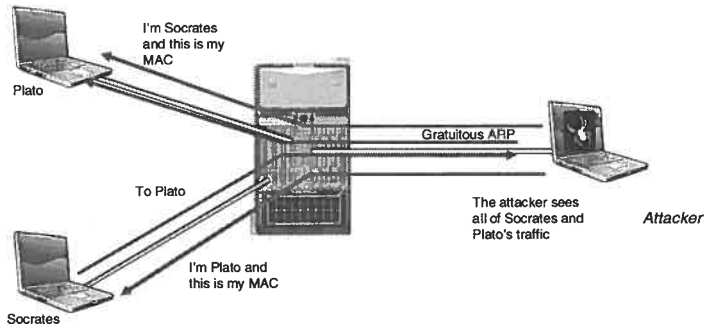


DHCP Snooping is designed to assign a level of trust to a switch port and validates DHCP messages. The DHCP snooping validates DHCP messages received from untrusted sources and filters out invalid messages. Example can be to make sure that that you do not receive a DHCP message that is supposed to be sent by a server; instead you receive it by a client. The DHCP-Snooping is a conditional service and the feature must be enabled followed by DHCP-Snooping service. The diagram shows that DHCP-Snooping is enabled globally and then there is a trusted port connecting the DHCP server, and the other ports are not trusted.

https://t.me/learningnets

Cisco NX-OS Dynamic ARP Inspection (DAI)

- DAI takes care of Man in the Middle attack
- Man in the Middle attack called ARP poisoning type of attack
- The DAI sets the rate of ARP packets for untrusted interfaces

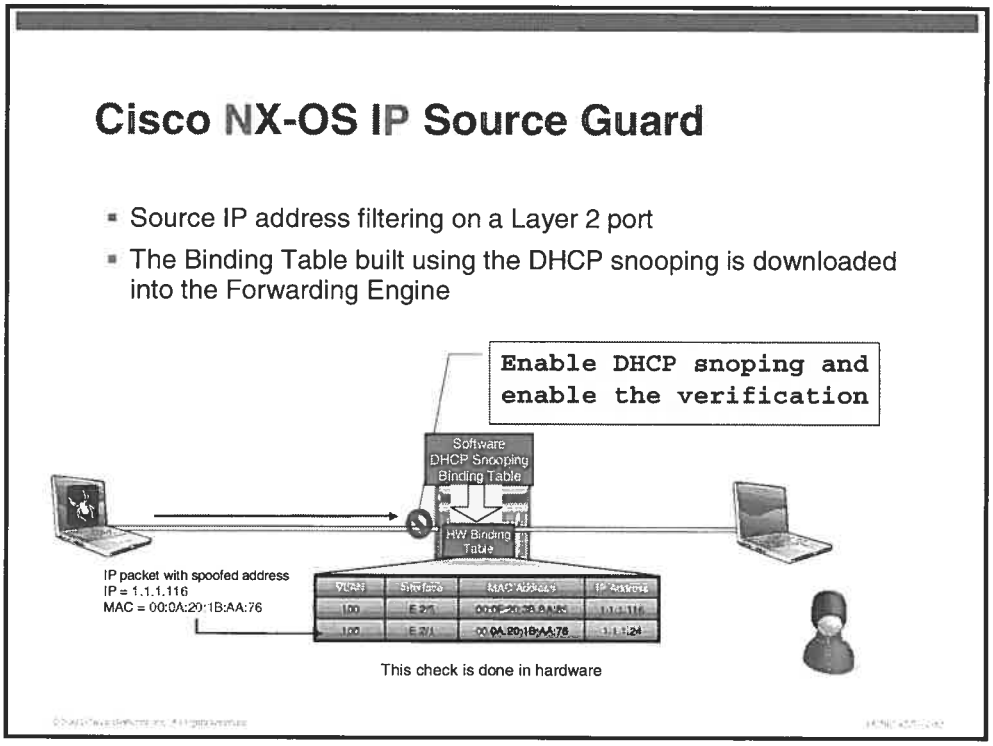


Dynamic ARP Inspection (DAI) takes care of Man in the Middle attack is where a device assumes the identity of another device allowing it to receive traffic that it should not normally have access to.

The diagram shows what happens after poisoning the ARP cache of each device. Communication between Plato and Socrates is sent via the attacker who stitches the communication together allowing Plato and Socrates to continue to talk while snooping on *all* conversations. In this case, Plato believes that the attacker is Socrates and Socrates believes that the attacker is Plato. Thus, when Socrates has to send some packet to Plato, the attacker is going to see the packets. DAI intercepts, logs, and discards ARP packets with invalid IP-to-MAC address bindings (using the Binding Table built by the DHCP Snooping). DAI is very useful against ARP spoofing attacks and ARP cache poisoning by intercepting and validating all ARP requests and responses on untrusted ports. DAI sets the rate of ARP packets for untrusted interfaces to 15 packets per second (pps). When the rate of incoming ARP packets exceeds the configured limit, the port is put into error-disabled state. DAI can also filter ARP packets based on ACLs which take precedence over the DHCP binding and DAI can validate the destination MAC address, the sender and target IP addresses, and the source MAC address in the ARP packet

© 2008 Cisco Systems, Inc.

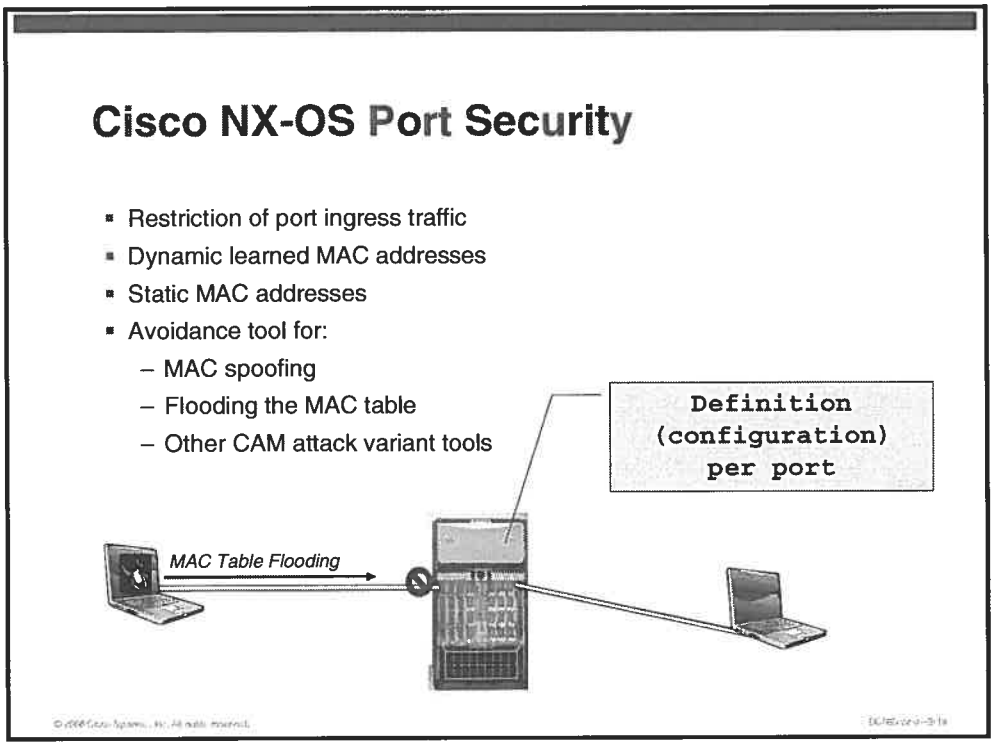
https://t.me/learningnets



IP Source Guard provides source IP address filtering on a Layer 2 port to prevent a malicious host from impersonating a legitimate host by assuming the legitimate host's IP address. The Binding Table built using the DHCP snooping is downloaded into the Forwarding Engine. DHCP-Snooping takes care of the validating the DHCP request. Dynamic ARP Inspection validates the ARP request. IP Source Guard validates the IP packets. IP Source Guard makes sure that the IP and MAC binding (Support dynamic and static DHCP binding) is consistent with what we have, that is, what the DHCP-Snooping has built. IP Source Guard is performed in hardware (using ACL TCAM) so the DHCP binding table that was built by the DHCP-Snooping is downloaded into the hardware, and the check on the binding of the IP MAC is done in hardware. To have the IP Source Guard, you still have to enable the DHCP-Snooping feature as a service and then you apply like you used to do on the Cat6k. IP Source Guard is not supported on Trunk Ports.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Port Security takes care of stopping MAC spoofing and flooding of the MAC table or any other CAM attack variant tool. You can use port security with dynamically learned and static MAC addresses to restrict port ingress traffic by limiting the MAC addresses that are allowed to send traffic into the port. Port Security allows you to set a MAC address for a port or set a max number of MAC addresses it can learn. NX-OS also supports Violation Actions and Aging for MAC address tables.

Example

Administrator can specify max number of MAC addresses the Cisco Nexus 7010 Switch can learn. Let us enter the static max of three–no more than three MAC addresses allowed on port. When port receives the fourth packet, fourth different MAC addresses is going to shut down the port by default.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS Port Security (Cont.)

- The actions supported for violation actions are:
  - Shutdown (default)
  - Protect (traffic dropped in hardware)
  - Restrict (traffic dropped in software)
- The violation condition is triggered when:
  - The MACs seen on a port pass the 'max' threshold
  - MAC moves from one secured port to another secured port
- "MAC moving"–the hardware supports the following capabilities:
  - Forward the traffic seamlessly
  - Notify the software
  - Drop



There are several actions supported for Violation. The device can "shut down" the port, which is the default. The device can "protect", where the traffic is dropped in hardware, and the device can "restrict", where the packet sent to software is dropped. The configuration condition is triggered by events like the number of MAC addresses seen on the port passes the maximum threshold that has been set with Port Security or when, for example, a MAC moves from one secured port to another secured port, when the MAC is seen on different ports. This may be a problem with VMware. The MAC moving gives the ability to forward the packets even though the MAC has moved. When movement happens within a secured port, the administrator can still decide not to drop the packet. Moving might also create notifications.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Cisco NX-OS Broadcast Port Suppression

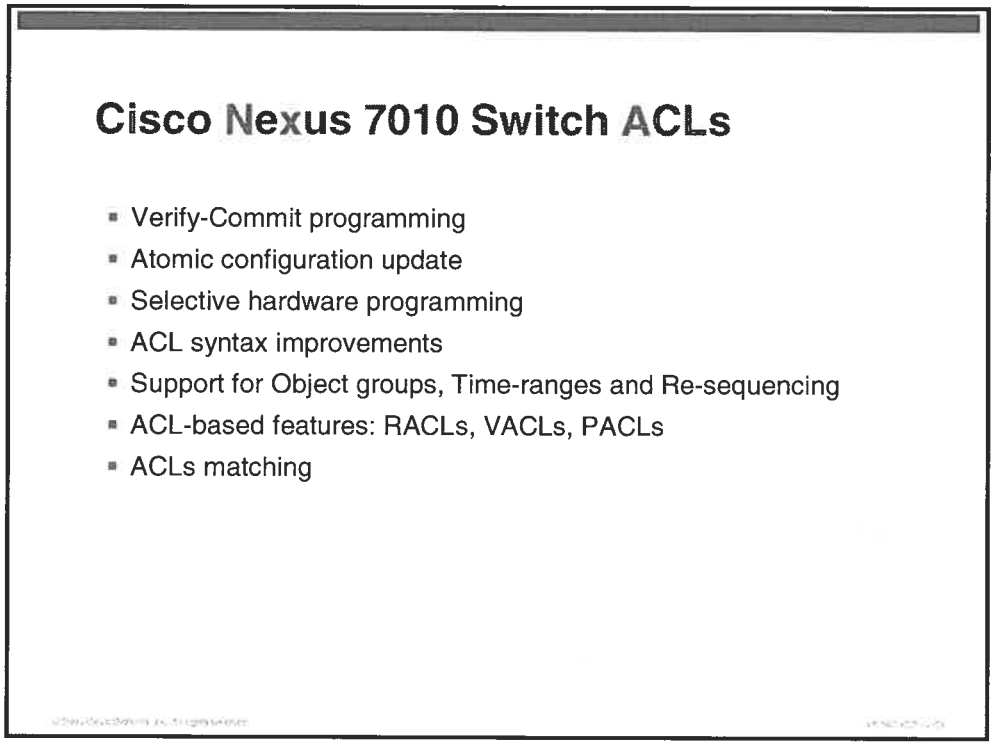
- Broadcast traffic can impact bandwidth availability and network performance
- Traffic storm control needed
- Timer to control the amount of traffic is 1 second interval
- Excess traffic might be dropped
- Suppression mechanism is the same on both line cards



As a percentage of the total bandwidth of the port, high volumes of broadcast traffic can impact bandwidth availability and impact network performance. Limitation of this traffic type is required. Traffic Storm Control allows controlled amount of "storm" traffic to be forwarded out of a target port. The Cisco Nexus 7010 Switch monitors outgoing "storm" traffic at 1 second intervals comparing the volume of storm traffic with the configured level that this port can forward. Traffic in excess of the configured limit is dropped. The suppression mechanism will be the same on both 1 G and 10 G line cards and we have double digit granularity.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

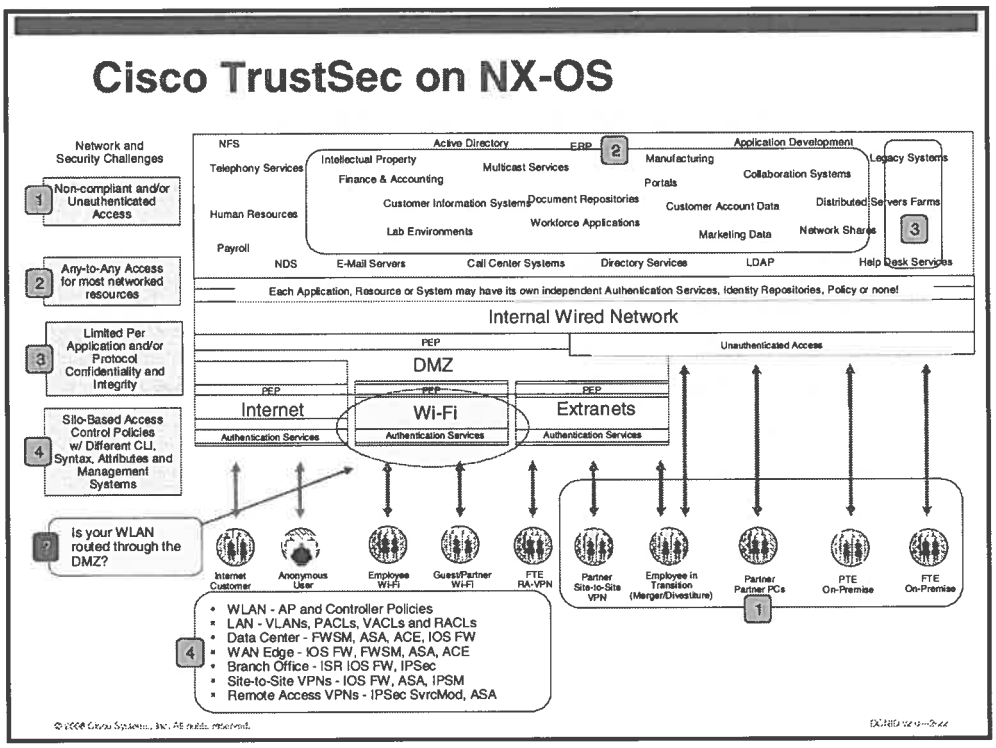


Several key points exist with NX-OS Access Control Lists (ACLs):

- Verify-Commit programming paradigm for better usability and manageability
- Atomic configuration update with no traffic interruption for continuous operations
- Selective hardware programming for better scalability and resources utilization
- ACL syntax improvements for better usability and manageability:
  - Slash notation for IP addresses
  - No standard/extended and named/numbered ACLs
- Support for Object groups, Time-ranges and Re-sequencing
- ACL-based features in 4.0: RACLs, VACLs, PACLs and PBR...
- SGACLs are also supported but they are out of the scope of this presentation
- ACL-based features in 4.0: RACLs, VACLs, PACLs and PBR...
- SGACLs are also supported but they are out of the scope of this presentation

© 2008 Cisco Systems, Inc.

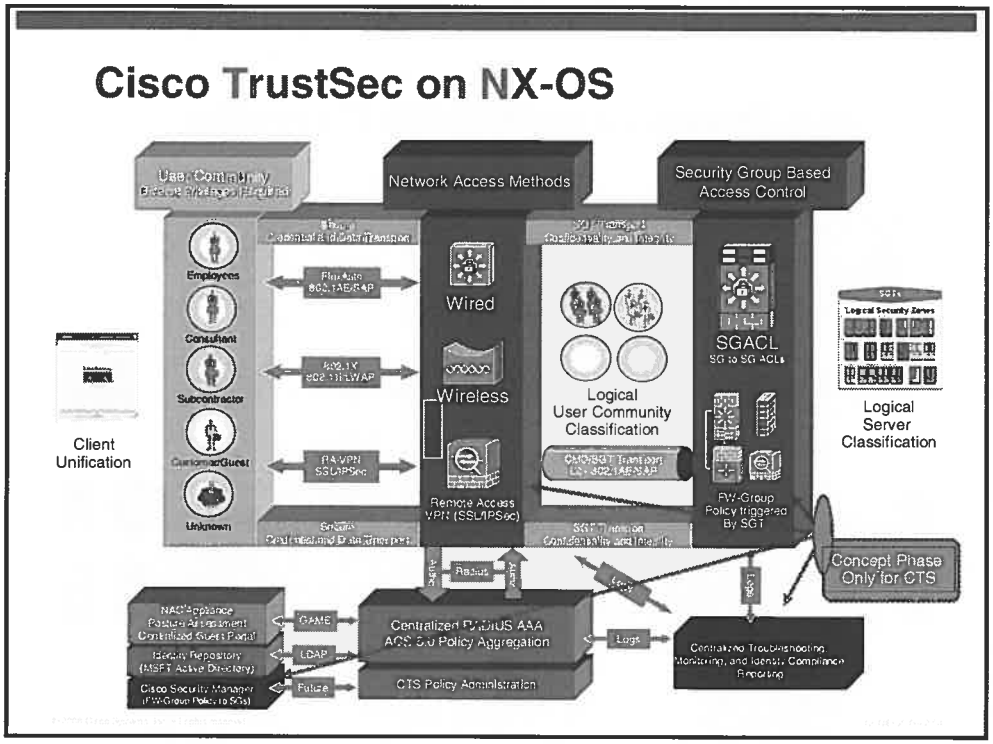
https://t.me/learningnets



Many different standalone security parts or areas exist in current networks. For example, in wireless networks there are many ways to do wireless security; in wired networks there are many ways to do wire-based security, and in the remote access part of the network there are certain ways of doing remote access security using remote access VPNs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



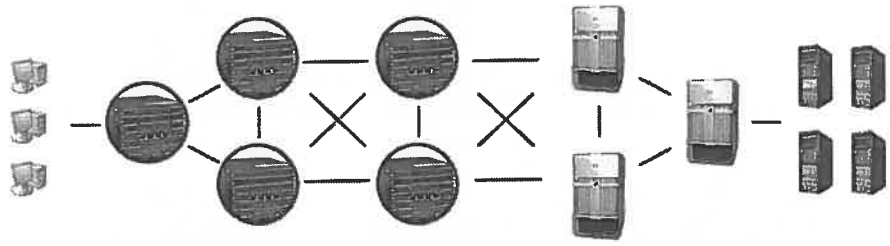
When TrustSec—or Cisco TrustSec—as a model is described, it is described as overarching security architecture. The idea is to combine wired, wireless and remote access into one common place of management and configuration, and then have the network perform the actual enforcement of some of these policies for us. As shown in the diagram, there are some common access points coming in, whether they are wired, wireless or remote access. The administrator wants to access some common resources, as well. CTS allows the system to determine, based on what user profile the user has, its role and other attributes, which provide the user with the relevant level of access onto the network.

https://t.me/learningnets

Cisco TrustSec on NX-OS Overview

Four major areas of CTS:

1. Network Device Admission Control
2. Endpoint Admission Control
3. Packet Confidentiality and Integrity
4. Security Group Access Control Lists (SGACLs)

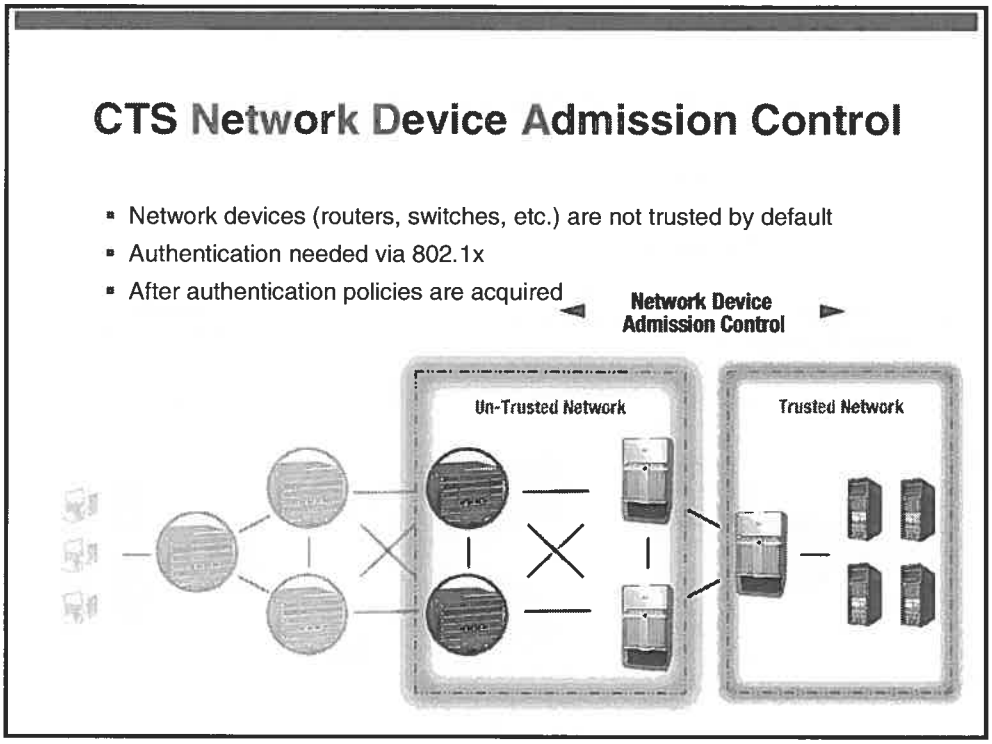


Cisco TrustSec (CTS) is split into a couple of different areas of operation. Four main points exist:

- Network Device Admission Control is dealing with how network devices can join a network.
- Endpoint Admission Control deals with how endpoints join a network.
- Packet Confidentiality and Integrity deals with packet confidentiality and integrity
- Security Group Access Control Lists (SGACLs) discusses security group ACLs.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Network devices (routers, switches, etc.) are not trusted by default and will need to authenticate themselves to a trusted node via 802.1X with EAP-FAST infrastructure prior to any network connectivity becoming available. Once they are authenticated, polices are then acquired, as well as any applicable session keys or cipher suites. Admission control is divided into Network Device Admission Control (NDAC) and Endpoint Admission Control (EAC).

NDAC means that currently, once we have a router that is configured or switches which are configured or access points which are configured appropriately, these devices basically have full access onto the network. Much like before, we had dot-1x, users, once devices were plugged into a network port, they had complete access to network resources. The devices may not be able to pull down a file from the file server because they did not have the correct passwords, but they have access to network devices.

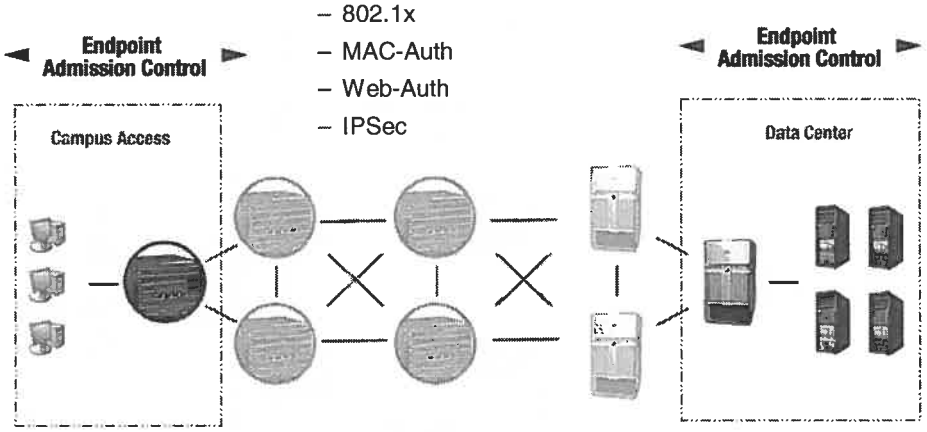
Network device admission control is designed to enforce a trust relationship between network devices and their trusted peers. So as soon as we implement network device admission control, we assume that network devices themselves are not trusted, which then involves a certain amount of authentication on the part of network devices.

What this means is that once we do have authentication, we need the switches and routers themselves to effectively implement some form of authenticator and supplicant functions in IOS software- using dot-1x terminology.

https://t.me/learningnets

CTS Endpoint Admission Control

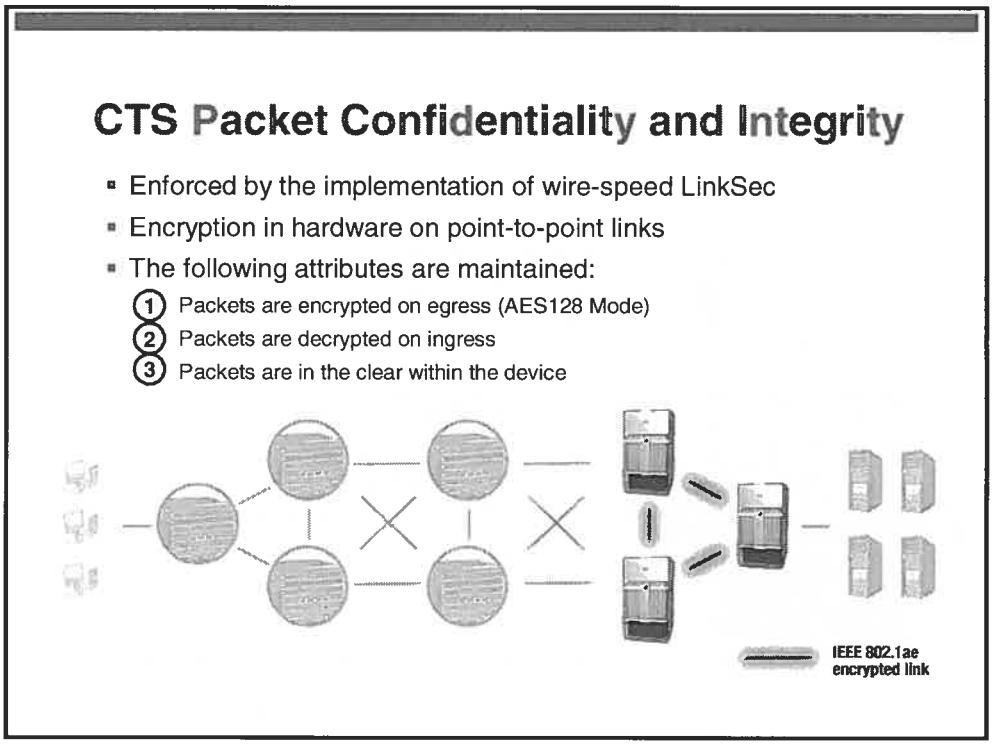
- Provides unified TrustSec policy for IBNS, NAC, WLAN and remote access
- Many authentication mechanisms supported, such as:
  - 802.1x
  - MAC-Auth
  - Web-Auth
  - IPSec



Endpoint admission control is supported by Nexus 7010 switches. CTS allows us to authenticate not just with dot-1x but also to fall back to, for example, using MAC authentication and using MAC-Auth Bypass and/or using Web Authentication. Flexible authentications—or flex-auths—are coming in the near future, which will allow the administrator to go from dot-1x and fall back to things like MAC Auth Bypass and Web-Auth in order to obtain some sort of policy or some sort of user identification and then apply policy. This is important because once we have the identity of the particular end user, we can then perform a certain level of access control on the user, such as whether to allow them onto the network, and then when we get to the actual resources themselves, we can use their identity to perform some sort of roles-based access control or roles-based enforcement at the actual data center itself.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



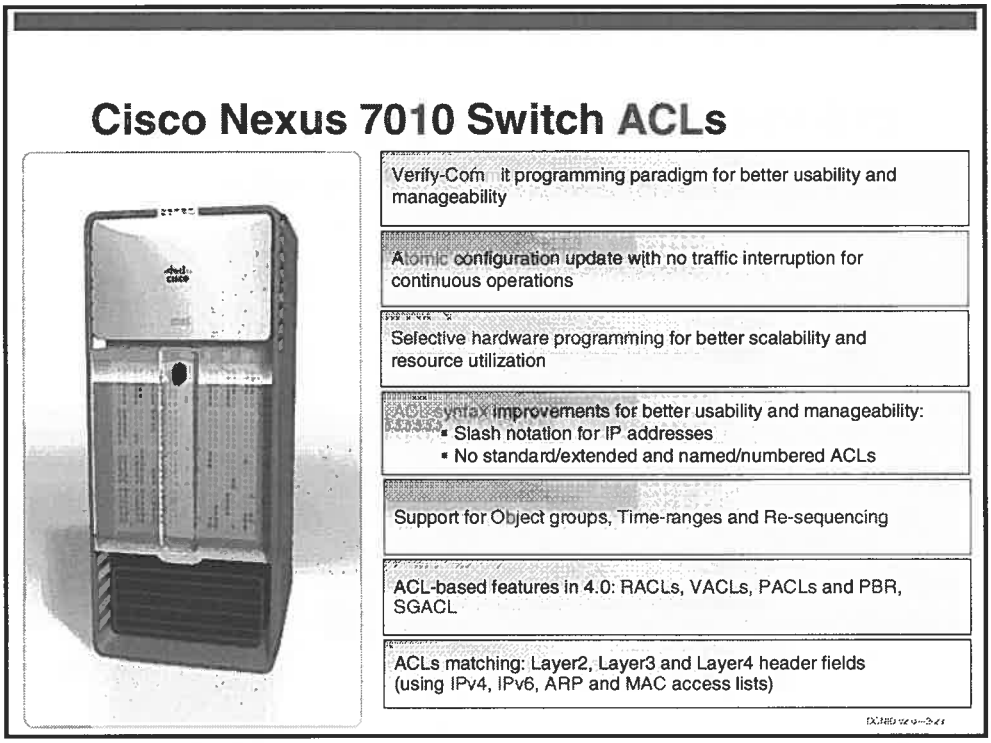
Packet confidentiality and integrity has been enforced by the implementation of wire-speed LinkSec IEEE 802.1ae Layer 2 encryption in hardware on point-to-point links. This follows a "bump-in-the-wire" model where the following attributes are maintained:

- Packets are encrypted on egress (AES128 Mode)
- Packets are decrypted on ingress
- Packets are in the clear within the device

Packet confidentiality and integrity involves a couple of things. The system is leveraging 802.1ae or LinkSec to be able to perform what is called hop-by-hop authentication and encryption. When packets go out of the switch, they are encrypted at line rate on the actual switchboard itself using ASICs, which are custom-built to do line rate encryption. As the packets flow across the link, they are in encrypted format so it prevents any sort of tampering or spoofing. And as they enter the network device, its next hop network device (its next adjacent device), and those packets will be unencrypted. Within the actual network device itself packets will be flowing in the clear. This is called a hop-by-hop encryption mechanism in the entire architecture.

Additionally integrity check is available. It is not just looking at the actual checksum to make sure that it is a valid Ethernet packet. The NX-OS system also has an integrity check that makes sure that the frame was not tampered with, as the system basically dequeues the frame as it receives the frame. Even without encryption an integrity check is performed to make sure that the frame was not tampered with.

https://t.me/learningnets

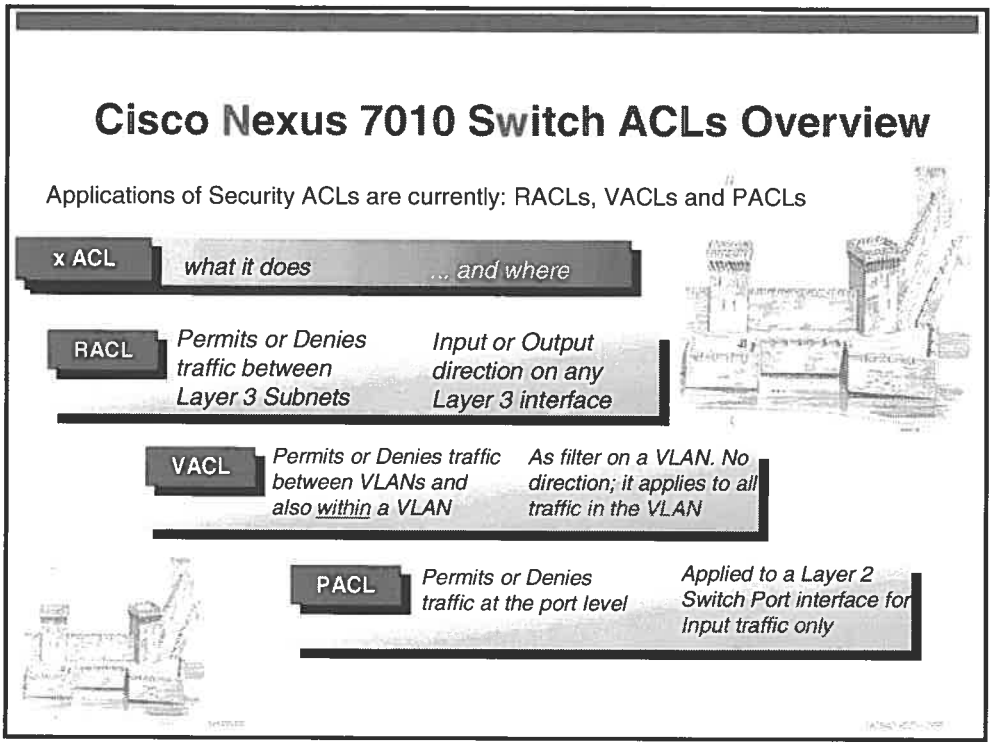


The slide presents the key point regarding the Cisco Nexus 7010 Switch Access Control Lists (ACLs).

- **Verify-commit** programming paradigm is a new way of programming the ACL in our system. It is provides the possibility to dry-run the configuration against the availability of the hardware resources without touching the hardware.
- **Atomic** configuration update is a way of updating the ACL policy on an interface without interrupting the traffic that is flowing on that interface.
- All ACL TCAM will be programmed in the same way on each forwarding engine. **Selective hardware programming** is providing better scalability and resources utilization. ACL will be only programmed where it needs to be programmed, helping to scale system resources.
- Additional improvements are done concerning **ACL syntax**, where the slash notation for the IP addresses is finally used. All ACL types (standard, extended, named or numbered ACLs) are now only one type and called simply ACLs.
- Additional support is added for **object groups**, **time-ranges**, and re-sequencing
- Support for the following ACL-based features: Routing ACL (RACL), VLAN ACL (VACL), Port ACL (PACL), and Policy Based Routing (PBR).
- **Matching** of packets can be done on Layer 2, Layer 3, and Layer 4 header fields by using IPv4, IPv6, ARP, and MAC access list.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

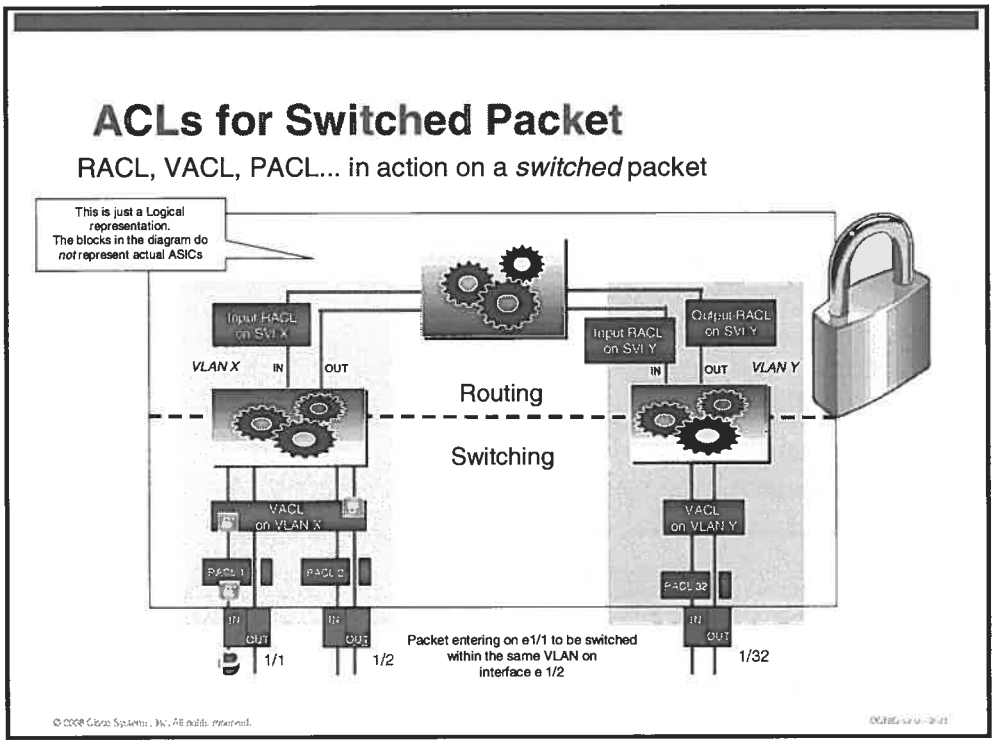


Cisco Nexus 7010 Switch supports several types of Access Control Lists (ACLs):

- **Router ACL (RACL)**: Permits or denies traffic between Layer 3 subnets and can be applied to any Layer 3 interface (input or output direction).
- **VLAN ACL (VACL)**: Permits or denies traffic between VLANs and between VLANs. It is applied as filter on a VLAN – No direction as it applies to all traffic in the VLAN.
- **Port ACL (PACL)**: Permits or denies traffic at the port level and can be applied to Layer 2 Switch Port interface for input traffic only.
- **Security Group ACL (SGACL)**: This ACL is mentioned in this chapter.

© 2008 Cisco Systems, Inc.

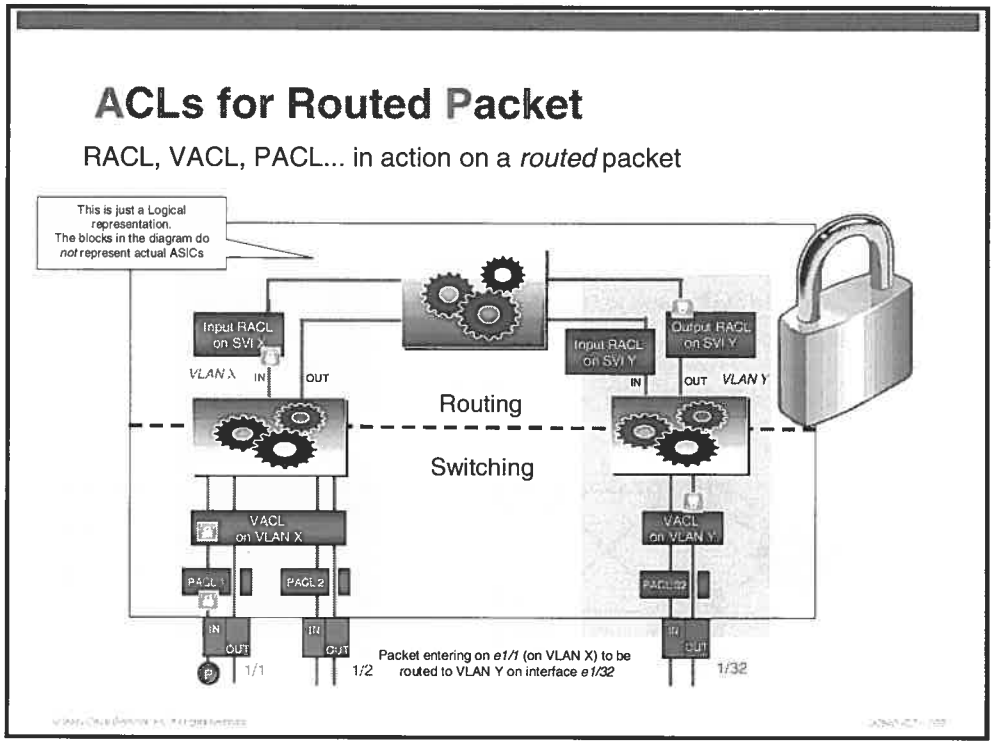
https://t.me/learningnets



The diagram represents the position of different ACLs, where PACL is at port level, VACL is at Layer 2 and RACL is at Layer 3. The animation shows the switched packet entering the system, where PACL and VACL are applied to the packet. Because the packet is switched, there is no need for passing the RACL. The packet in the animation enters the system on e1/1 port and is switched within the same VLAN on interface e1/2. The packet passes PACL 1 on the way into the system (incoming port e1/1), then VACL on VLAN x entering the switching engine and again VACL on VLAN x going toward the outgoing port (e1/2).

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

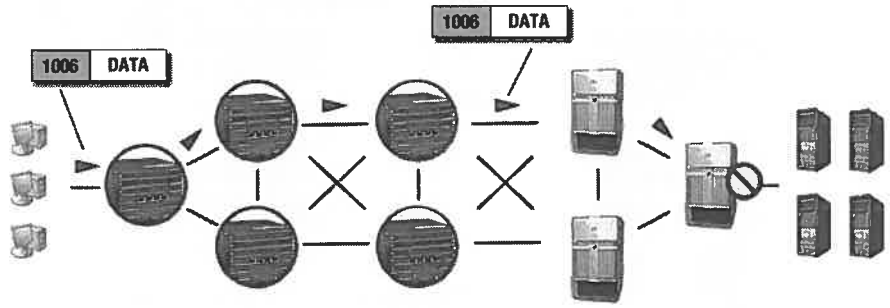


The diagram represents the position of different ACLs, where PACL is at port level, VACL is at Layer 2 and RACL is at Layer 3. The animation shows the switched packet entering the system, where PACL, VACL and RACL are applied to it. Because the packet is routed, there is no need for passing the RACL. The packet in the animation enters the system on e1/1 port (on VLAN x) and it is routed to VLAN y on interface e1/32. The packet passes PACL 1 on the way into the system (incoming port e1/1), then passes VACL on VLAN x entering the switching engine. Before the packet is routed, it goes through input RACL (on SVI x) and after the routing engine the packet passes output RACL (on SVI y) and lastly the packet goes through VACL on the VLAN y and toward the outgoing port (e1/32).

https://t.me/learningnets

CTS Security Group Access Control

- SGACLs used to scale enforcement rules
- SGACLs used to decouple network topology from ACLs
- User associated to a Role
- Role ID is (SGT) used to tag frames from users
- Security Group-based policies applied at a defined point in the network



TrustSec implements Security Group Access Control Lists (SGACLs) to scale enforcement rules and to decouple network topology from ACLs. This task is achieved by associating users to roles, tagging frames from users with a Role Identifier, or Security Group Tag (SGT), and then allowing the network administrator to apply Security Group-based policies at a defined point in the network.

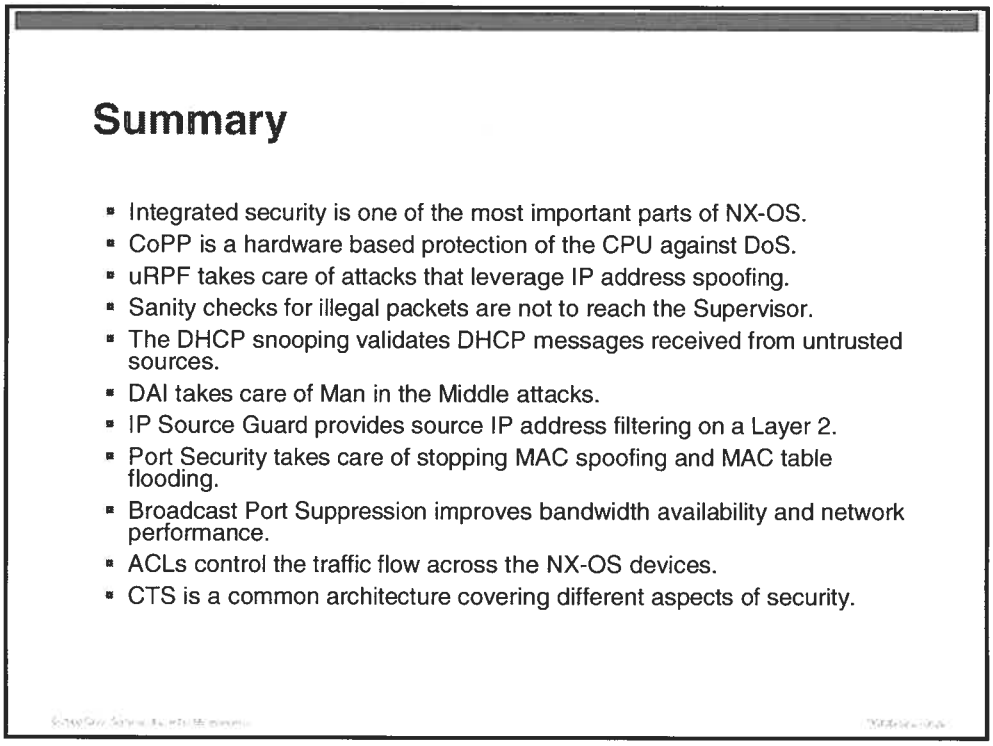
As individual end workstations are authenticated and authorized to access a network, they have a full suite of dot-1x tools that they can call upon. VLANs can be assigned and a certain number of other attributes can be downloaded, as well (dot-1x accounting, Port security like voice VLAN, etc.). After that, the administrator programs down an SGT value–a tag ID, which is used to bind that particular unique user. Every user can have many different roles and can be members of more than one group. For a particular time of day or day of week, and depending also on location, the administrator can associate this group with a particular security group tag. All this information is configured on the ACS itself. Through the authentication and authorization phases, users get all the information necessary to join the network and access network resources.

© 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 6

Cisco Nexus 7010 QoS Implementation

Overview

This lesson describes the Cisco Nexus 7010 Series Switch QoS implementation and its application in the data center network architecture.

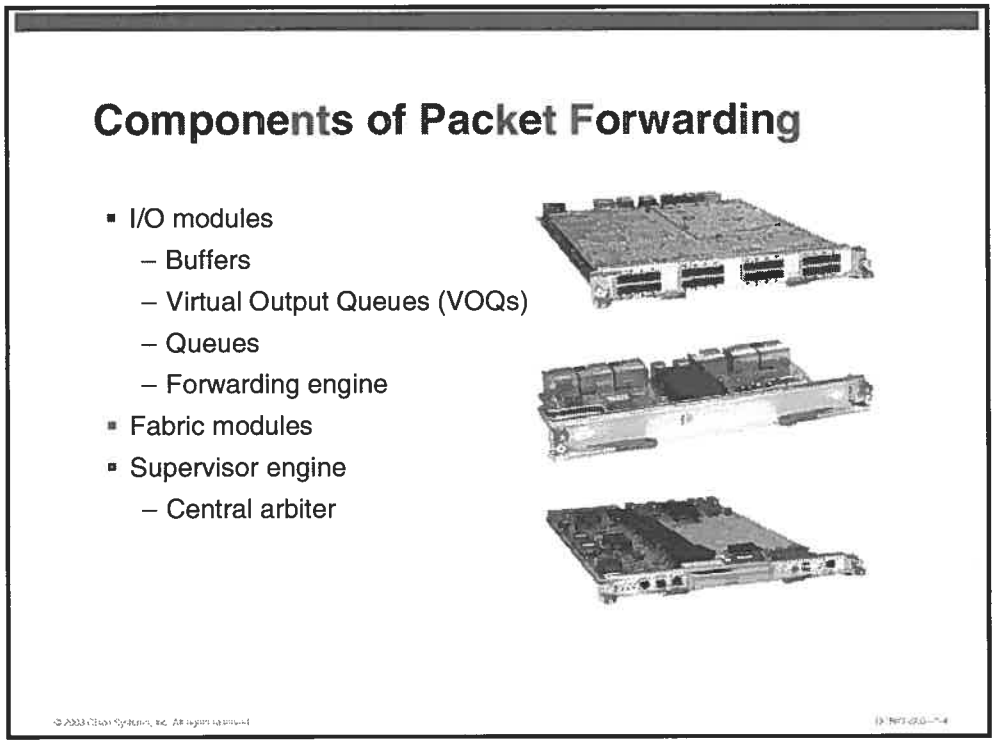
Objectives

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

- Cisco Nexus 7010 Packet Forwarding and VOQ
- Cisco Nexus 7010 QoS Components and Features
- Describe the Data Center Network QoS

https://t.me/learningnets

Cisco Nexus 7010 Packet Forwarding and VOQ



Nexus 7010 series switch components of packet forwarding are the following:

I/O Modules

I/O Modules with

- Buffers to temporarily store the packets crossing the module
- Virtual output queues (VOQs) that ensure fair access to fabric bandwidth and non-blocking architecture
- Queues to classify traffic (packets)
- Advanced hardware forwarding engine used for faster processing of packets as some functions (including some QoS) are processed in hardware

Fabric Modules

Fabric modules provide switching (bandwidth) capacity for I/O modules. Fabric bandwidth is represented by the VOQs. The systems are supporting the use of more modules, providing redundancy and load/sharing capabilities.

Supervisor Engine

The Supervisor Engine controls access to fabric bandwidth using central arbitration and prevents Head-of-line blocking effect.

The Cisco Nexus 7000 Series operating system NX-OS is a modern operating system built with special focus on the software infrastructure and the features required for multi-protocol, mission-critical data center networks. NX-OS is called a multi-purpose OS since it was

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

designed from the ground up to be highly available and highly modular with a focus on integrating specific data center networking technologies:

- Data networking (LAN/WAN/MAN)
- Storage networking (FC, iSCSI, FCIP)

NX-OS software architecture is based on software component modularity and high-availability, enabling zero downtime, zero data loss and nonstop forwarding. NX-OS interoperates with Catalyst switches running Cisco IOS and MDS switches running SAN-OS. NX-OS will also interoperate with other Cisco equipment. The NX-OS goal was to integrate Data Center Networking technologies, such as Layer 3 routing; Layer 2 switching and bridging technologies; SAN switching; and service module functionalities, such as L4 - L7 and security features. NX-OS was targeted initially to run on the Nexus 7000 Hardware platform, but it is also designed to be portable so that it can be ported to other platforms in the future, if necessary. The design goals were to provide specific feature sets to the data center networking architecture. This fact imposed several new software and hardware requirements on Nexus 7010 Series architecture:

1. Nexus was designed to support distributed forwarding for Ethernet/SAN traffic. It must support all Ethernet frame sizes (64b – Jumbo). Jumbo frames are approximately 9 KB frames defined by IEEE to be able to accommodate future Fibre Channel frames in one single Jumbo frame.

2. In the case of over-subscription, there must not be any unintelligent packet drops anywhere in the packet processing chain, since the storage protocols are not supposed to lose any frames. For these reasons, mechanisms such as:

   - Decent size buffers
   - Multiple queues
   - Congestion control
   - QoS policies associated with each packet or frame must be supported by NX-OS and hardware.

3. Nexus 7010 Series switch must enable non-blocking forwarding architecture:

   - During over-subscription, traffic from all impacted ports must be treated fairly.
   - The system is designed to either allow all traffic or block all traffic, but not a combination of both.
   - Head-of-line (HOL) blocking must not occur in the system fabric.
   - Congestion of one egress line card should not impact fabric traffic going to other line cards.

4. Nexus must support Multicast/Broadcast over Fabric:

   - Nexus 7010 Series Switches must support efficient transport of multicast/broadcast traffic
   - No performance impact when carrying multicast traffic (L2 and L3) over the fabric
   - The same functionality must be supported for L2 traffic flooding (unknown unicast floods, multicast and broadcast).

https://t.me/learningnets

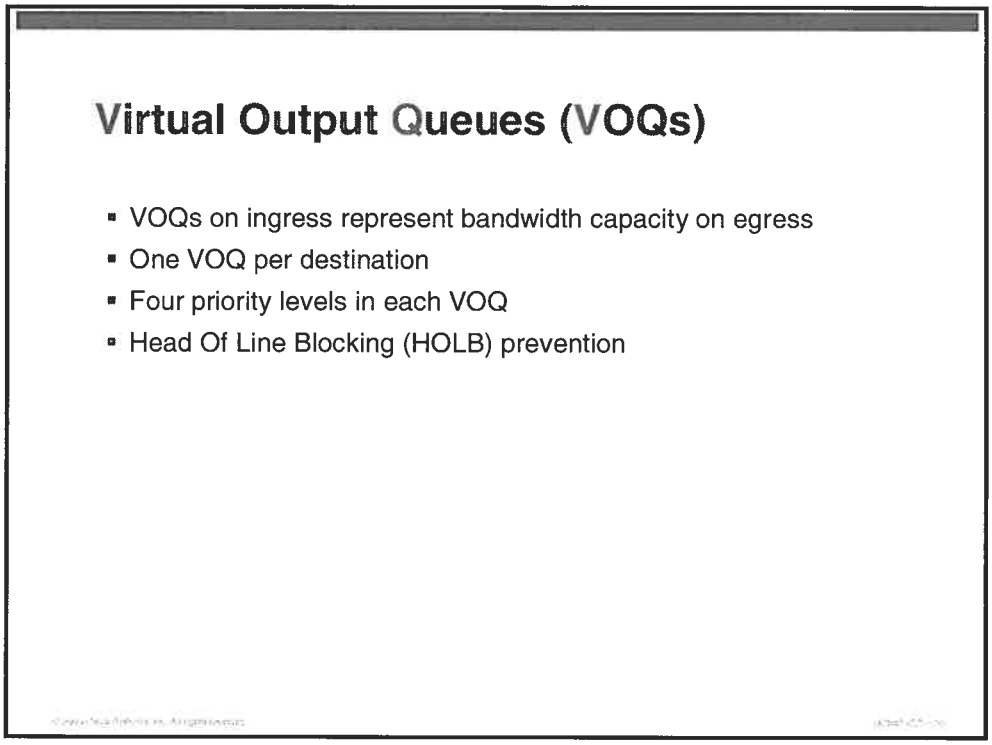
5. Redundant Management Access

Connectivity Management Processor (CMP) provides standalone redundant management access to the Nexus 7010 Series Switch. CMP is supported on the supervisor and enables remote reboot of the switch:

- Without losing telnet/SSH session even when CP/DP (or SP/RP) CPUs are hung
- View all boot cycle messages
- Power supply is separate from the main chassis power supply to ensure its operation during chassis power cycle.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



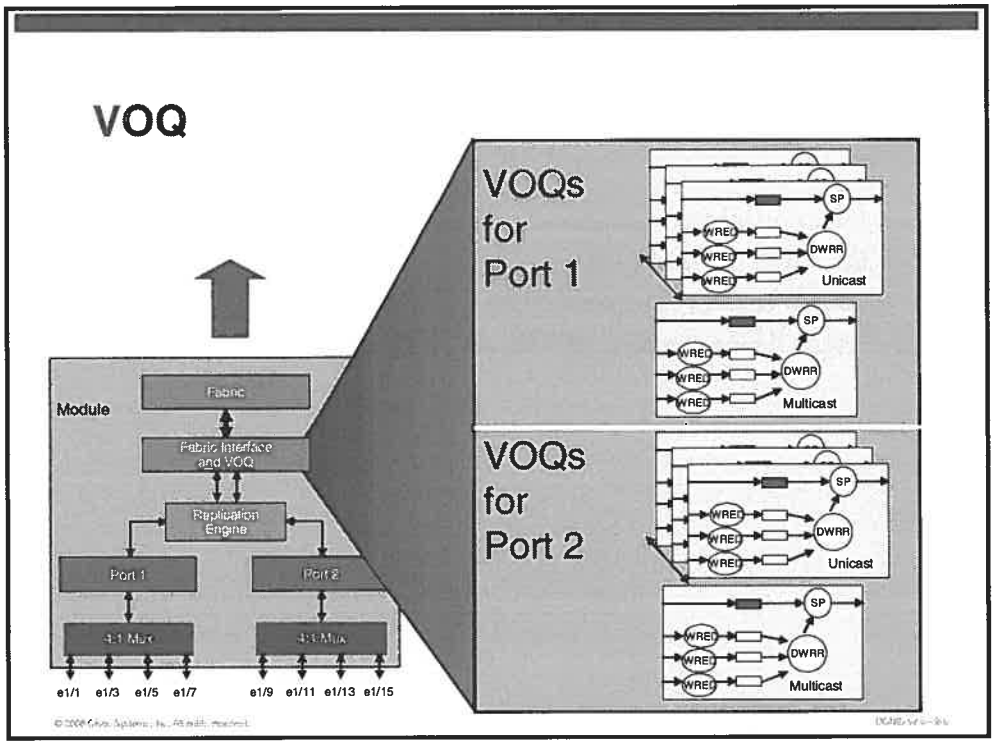
Virtual Output Queues (VOQs) on ingress modules represent bandwidth capacity on egress modules. They guarantee delivery to egress module for arbitrated packets entering fabric. If VOQ is available on ingress line card than the respective bandwidth capacity exists on the egress line card. VOQ is *not* equivalent to ingress or egress port buffers or queues. It relates only to ASICs on the ingress and on the egress line cards to fabric. It represents egress capacity but resides on an ingress module.

For every destination port-group (or port) on the egress modules, each ingress module has corresponding VOQ with four priority levels. VOQ destinations do not necessarily map directly to front panel ports on egress modules. Multiple VOQs exist for different destinations. On initially shipping modules, a VOQ with four priority levels serves one of the following destinations on the egress module:

- One 10 G port (dedicated mode) or
- Four 10 G ports (shared mode) or
- Twelve 1 G ports

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



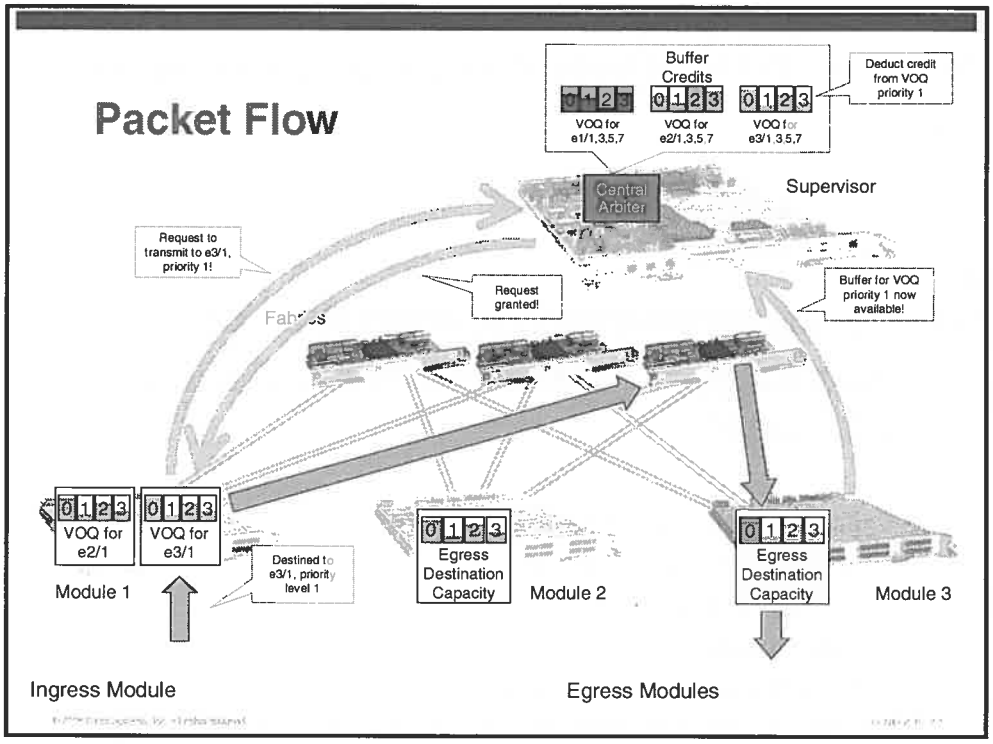
Where are VOQs in the ingress module? They reside on the ASIC located in front of the Fabric ASIC that is sending the frames to the switch fabric. As incoming ports on the ingress module are grouped together into several port ASICs, we have a set of VOQs per port ASIC. It is clear that we have many VOQs for unicast traffic and just one VOQ for multicast traffic.

Example

If the Cisco Nexus 7010 Series Switch is populated with 8 10 Gb/s line cards all enabled in a dedicated port mode, then for non-blocking architecture 32 * 8 = 256 destination ports exist in the Nexus chassis and there will be 256 VOQs needed to accomplish non-blocking forwarding architecture. To enable prioritization for every destination, another 256 * 4 = 1024 VOQs are needed in this architecture. Therefore, if one port is sending data to all the other 255 ports, there will be 1023 VOQs available on the ingress line card for non-blocking architecture.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Packet forwarding is controlled by the supervisor module through the central arbitration. The central arbiter ensures fair access to bandwidth for multiple ingress ports transmitting to one egress port. Through VOQ central arbiter prevents congested egress ports from blocking ingress traffic destined to other ports. Priority traffic is sent first and the system is engineered to support Unified I/O, which can provide no-drop service across fabric for future FCoE interfaces.

Packet Flow Example

- Packet is coming to the ingress module 1 and the destination is port 1 in the destination egress module 3. The priority of the packet is "level 1".
- Module 1 requests transmission of this packet to the destination module/port with level 1 - VOQ for level 1 priority. .
- The central arbiter is granting requests if there is space available on destination.
- Packet goes from switching fabric to destination port through switching fabric.
  - When the packet is sent out of the egress port, the egress module 3 reports that buffer for VOQ level 1 is available.

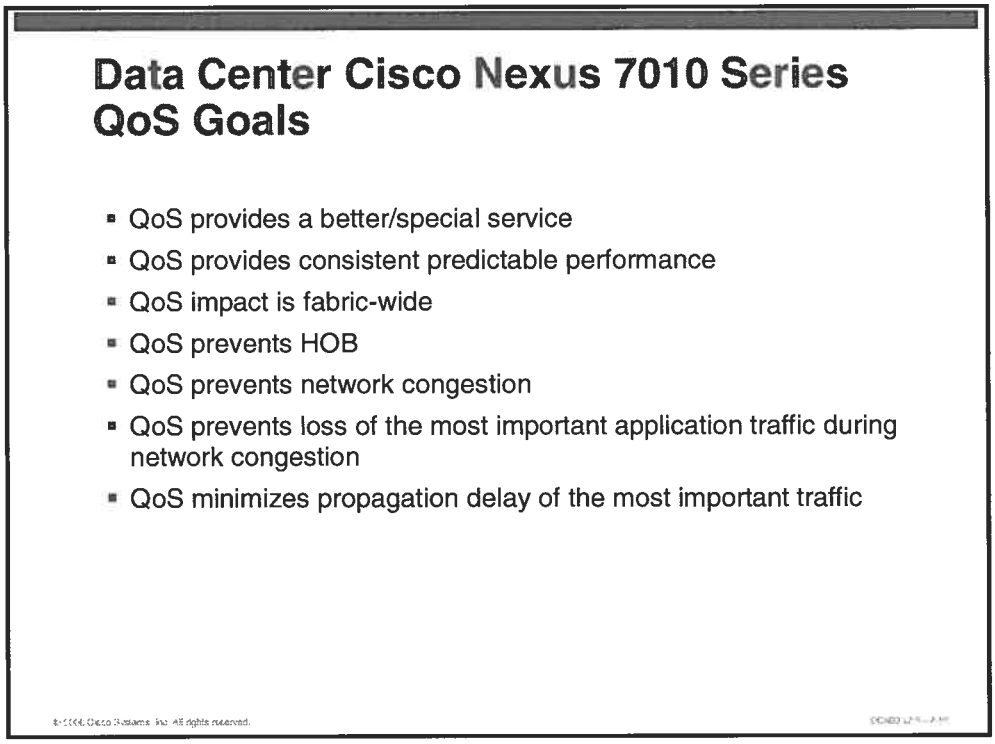
Note The central arbiter is QoS-aware which makes the final QoS impact in the data center fabric-wide, as opposed to local (per switch) QoS impact.

https://t.me/learningnets

Cisco Nexus 7010 QoS Components and Features

This topic describes the Nexus 7010 Series Switch QoS components and features.

QoS Goals



QoS is the ability of the network to provide better or "special" service to selected users and applications. QoS cannot create bandwidth but QoS can manage available bandwidth and assure that some packets get priority or high priority treatment.

Cisco QoS features enable network administrators to control and predictably service a variety of networked applications and traffic types, thus allowing network managers to take advantage of a new generation of media-rich and mission-critical applications.

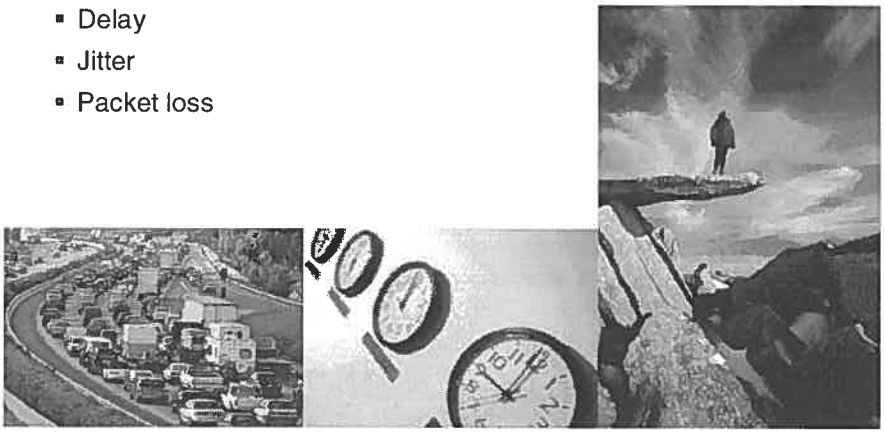
The goal of QoS is to provide better and more predictable network service by providing dedicated bandwidth, controlled jitter and latency, and improved loss characteristics. NX-OS 7010 QoS features achieves these goals by providing tools for managing network congestion, shaping network traffic, and setting traffic policies across the network. QoS offers intelligent network services that, when correctly applied, help to provide consistent, predictable performance.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center QoS Goals (Cont.)

- Bandwidth
- Delay
- Jitter
- Packet loss



Converged data center networks that support voice, video, application and storage traffic create new requirements for managing network traffic. QoS meets those requirements. Converged networks suffer from different quality issues, including lack of adequate bandwidth, end-to-end and variable delay, and lost packets.

Networks must provide:

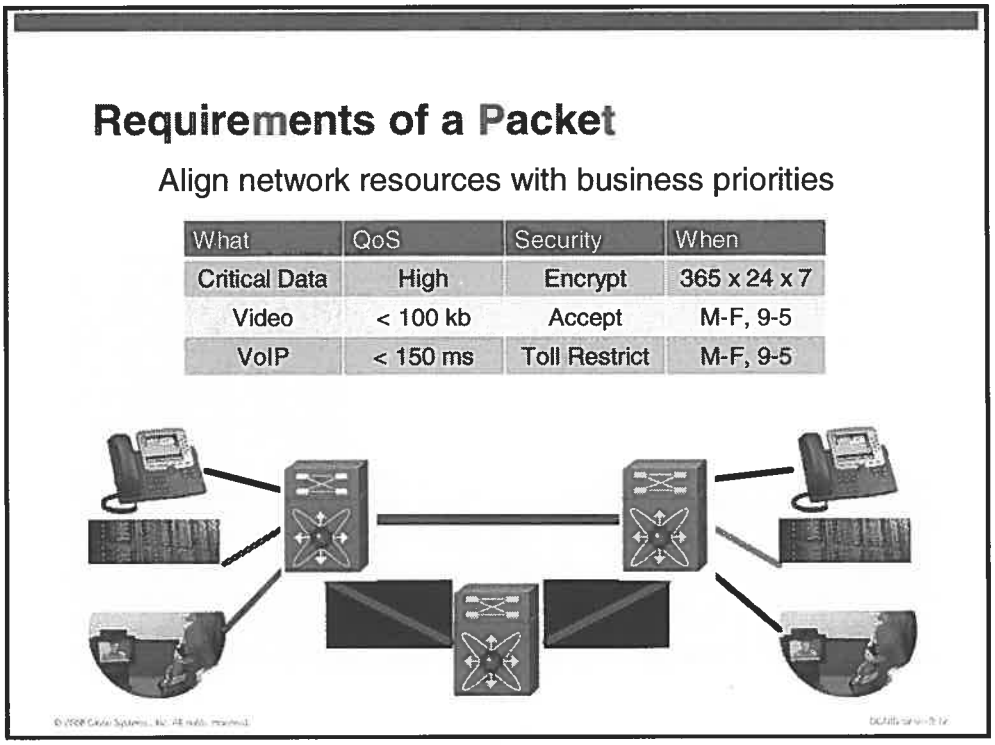
- Enough bandwidth for critical applications. In some cases the network infrastructure must guarantee this bandwidth as well.
- Low and constant delay for delay sensitive applications. Important delay-sensitive packets must be sent first or before best effort traffic.
- Low or no packet loss. Some applications can adapt to packet loss by slowing down the transmit rate, some applications do not care, and some of them require no packet loss.

Storage applications such as SCSI I/O reads and writes are suffering from losing a single data frame. In case of a lost Fibre Channel frame the whole SCSI read or write Fibre Channel exchange will be restarted.

The first step before applying QoS is to evaluate the traffic patterns or traffic types in the network. Then the proper configuration of the network equipment is required to meet the application requirements for bandwidth, delay and packet loss.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



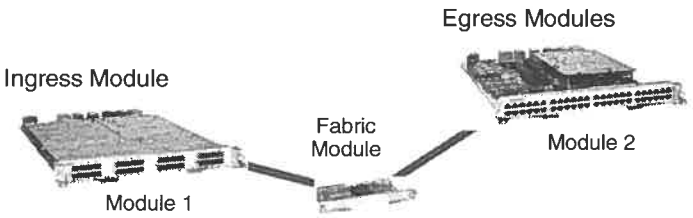
Every data center application has an application profile. The application profile defines the amount of bandwidth needed and the acceptable propagation delay, the I/O size and the throughput needed. The graphic illustrates how a QoS policy could be defined for a network. Critical data and database replication applications have a high QoS priority and must be available all the time. Video applications are guaranteed 100 kb/s of bandwidth, but can operate only between the hours of 9:00 a.m. to 5:00 p.m. on weekdays. Voice traffic is guaranteed less than 150 ms delay in each direction but is limited to the hours of 9:00 a.m. to 5:00 p.m. on weekdays; toll calls are completely restricted to avoid personal long-distance calls.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Where to Apply QoS in Cisco Nexus 7010 Switch

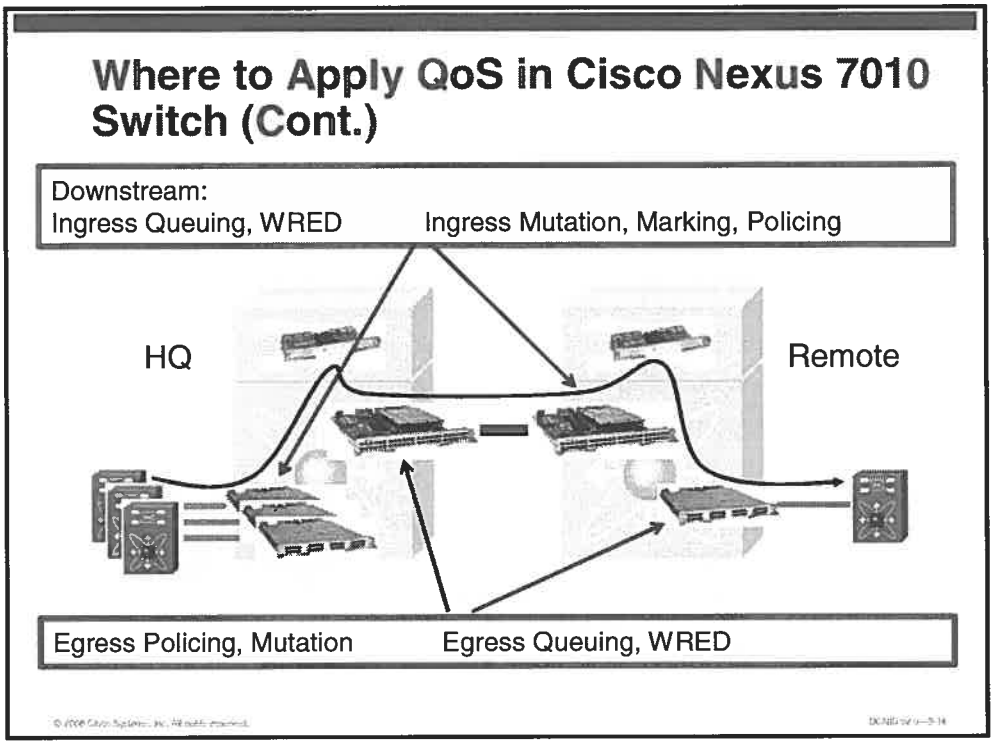
- Ingress
  - ingress port
  - ingress forwarding engine (ingress pipe)
- Egress
  - egress forwarding engine (egress pipe)
  - egress port



Cisco NX 7010 platform provides QoS via the ingress and egress modules. Packets go through the system by entering the ingress port, passing through the ingress forwarding engine, the fabric module, next egress forwarding engine, and lastly exits the egress port. This design allows the system to provide both QoS ingress and egress feature sets. These QoS features are also available for the Cisco Nexus 7010 switch without any additional licensing requirements.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Downstream (HQ to Remote)

Frames are coming from the user (in HQ) to the Ingress module in an NX 7010 device. The Ingress module can provide Ingress Queuing, WRED, Mutation, Marking and Policing. Then the frame is sent out the Egress module (where the frame goes out of the NX 7010 system). The Egress module can provide Egress Policing, Mutation, Queuing and WRED. Queuing is used to treat different frames with different priority and to give different classes of traffic different guaranties. WRED is used for intelligent dropping. A default dropping mechanism used is tail drop, where all the frames are dropped if the queue is full. With WRED, the NX 7010 system is monitoring the usage of the queue and randomly dropping some frames to slow down the traffic. Mutation is used to remark the frames prior to the classification. We have the ability to control the marking and adapt the marking to our classification rules. Marking itself is used to mark the frames with the several markers (Layer 2, Layer 3 and local markers). When marking is used, then later the network classification is easier as we can trust the marking of packets and we do not need to apply CPU-time-consuming ACLs. If the rate of incoming frames is higher than the outgoing bandwidth, there is a need for a rate limiting action called policing. The same is true if there are more incoming interfaces and only one outgoing interface. It is likely that speed mismatch will happen. Again, the policing action can limit the traffic rate by dropping excess traffic.

Upstream (Remote to HQ)

Upstream traffic flow is the opposite of downstream traffic flow. It can be written that the same rules apply, but there is a difference in bandwidth and interfaces. If many interfaces are sending traffic from HQ to the Remote site then it is more likely that congestion will happen at the egress side of the NX 7010 system. Conversely, there is a fixed bandwidth of traffic sent back to the users and the only problem is the speed of the user connection. However, in case of multiple users, the high speed link on Remote side is sharing the bandwidth to the users and congestion is not likely to happen. Proper marking and queuing is required to prioritize important traffic.

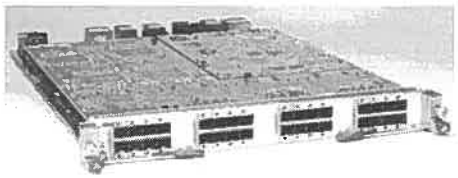
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Port QoS—32-Port 10 G Module

Buffers

- 1 MB ingress
- 80 MB egress
  - Dedicated mode: per port
  - Shared mode: per port-group



Queue structure

- 8q2t ingress
  - The 4 ingress buffer changes between shared and dedicated modes
- 1p7q4t egress
  - Dedicated mode: per port
  - Shared mode: per port-group

Cisco NX 7010 system offers two I/O modules:

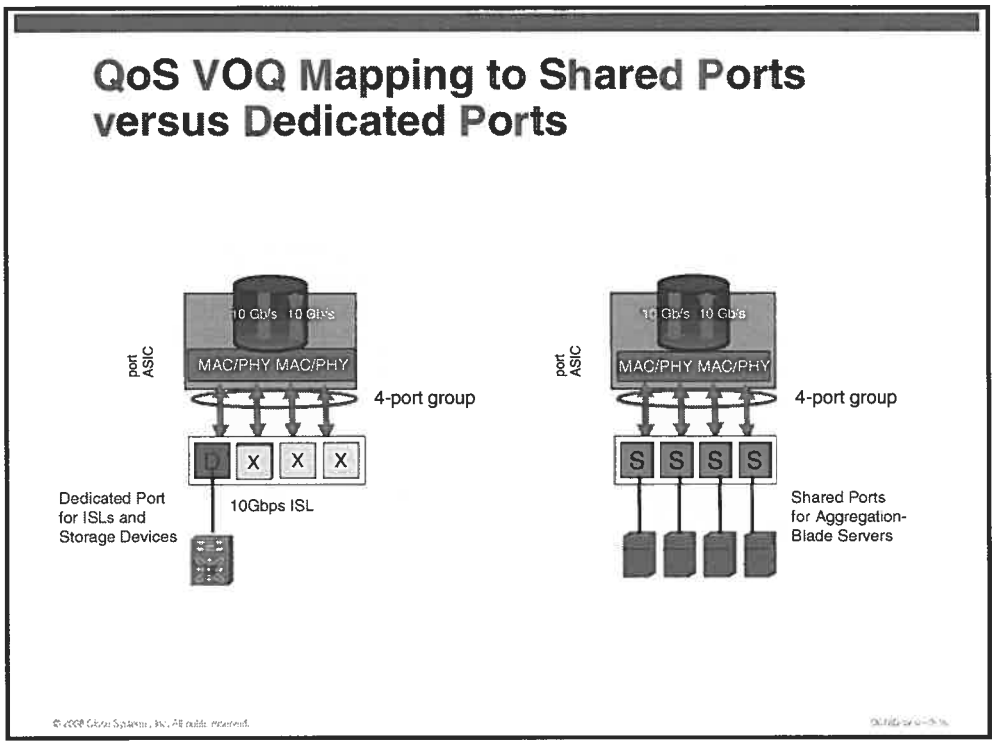
- 32-Port 10 Gigabit Ethernet Module
- 48-Port 10/100/1000 Ethernet Module

On the 32-port 10 G module, there are 1 Mb buffers available on ingress. On egress, 80 Mb of buffers is available regardless of whether you are in dedicated or shared mode. In shared mode, 80 Mb is shared among the 4-ports in the port-group, whereas, in dedicated mode, you get per-port egress buffer. The queue structure is essentially the same, regardless of dedicated or shared mode. On ingress, eight queues with two thresholds each are available–per port queue structure. For the eight ingress queues, no priority queue is supported. Two thresholds for each queue are for tail drop thresholds.

The 4 ingress buffer switches between shared mode and dedicated mode.

On the egress side, 1p7q4t means one priority queue available. The other seven are additional DWRR queues, where each has four thresholds for Weighted Red (WRED) or tail drop. Another option on egress of this card is shaped round-robin tail drop. In this case, priority queue is not available. This applies to any port configured for shaped round-robin. Therefore, shaped round-robin and priority queuing on a per-port basis are mutually exclusive.

https://t.me/learningnets



The module has two forwarding modes and different egress and ingress queuing policies:

- Dedicated mode
- Shared mode

Shared Port Mode (Default)

Shared mode is used for aggregation (for example, Blade servers). In shared port mode on the 48-port line card, all 4 ports are grouped together and they share the 10 Gb/s bandwidth. They share VOQ from the ASIC, as well, and they all represent a single destination from the aspect of the VoQs. In this situation, it is clear that VOQs are available and common to all four ports in the 4-port group. If shared mode is used then all ports in the group must be in the same VDC.

Dedicated Port Mode

Dedicated mode is used for the dedicated connection of one device (for example, ISL links or connection for storage devices). If dedicated mode is used then all other three ports belonging to the same 4-port group (or ASIC) are not used—they are disabled. In this situation, QoS VOQs are available per every single 10 Gb/s port.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

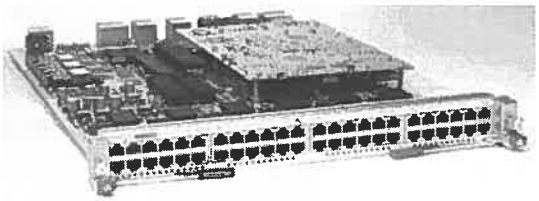
Port QoS—48-Port 10/100/1000

Buffers

- 7.56 MB ingress per port
- 6.15 MB egress per port

Queue structure

- 2q4t ingress
- 1p3q4t egress



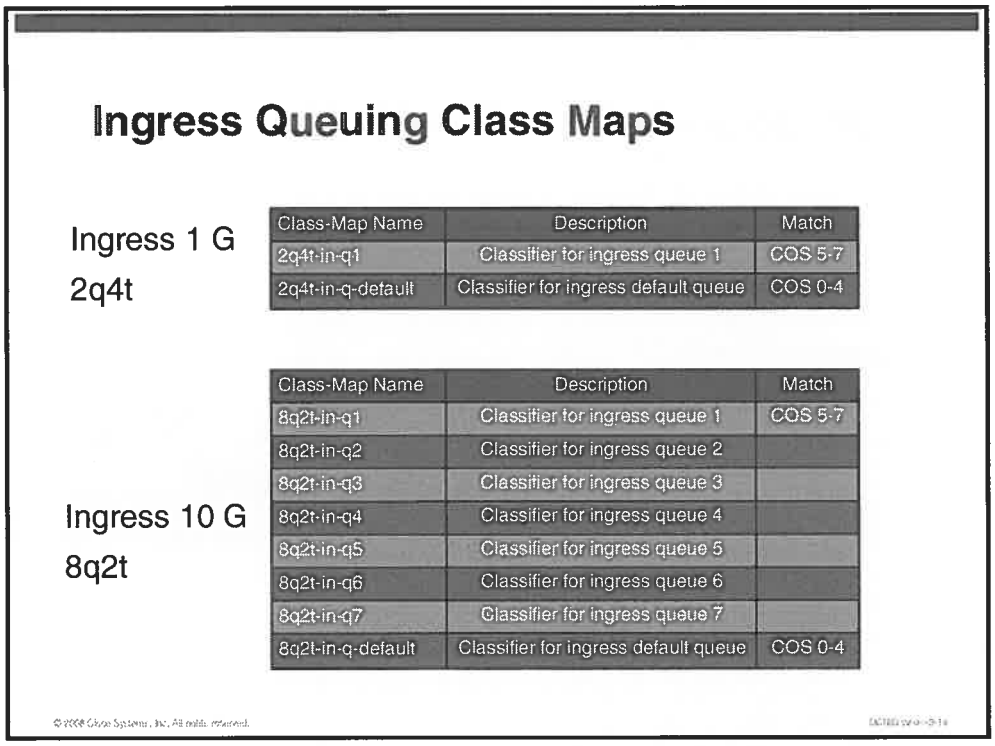
Cisco NX 7010 system offers two I/O modules:

- 32-Port 10 Gigabit Ethernet Module
- 48-Port 10/100/1000 Ethernet Module

On the 48-Port 10/100/1000 Ethernet Module, all buffers are per-port. 7.56 Mb of buffer is available on ingress and 6.15 Mb of buffer Is available on egress. The queue structure is 2q4t on ingress and 1p3q4t on egress. On ingress, two queues with four thresholds each are available - per port queue structure. For those two ingress queues, no priority queue is supported. On the egress side, 1p3q4t means one priority queue is available. The other three are additional DWRR queues, where each has four thresholds for Weighted Red or tail drop. Another option on egress on this card is shaped round-robin. In this case, priority queue is not available. This applies to any port configured for shaped round-robin. Therefore, shaped round-robin and priority queuing on a per-port basis are mutually exclusive.

© 2008 Cisco Systems, Inc.

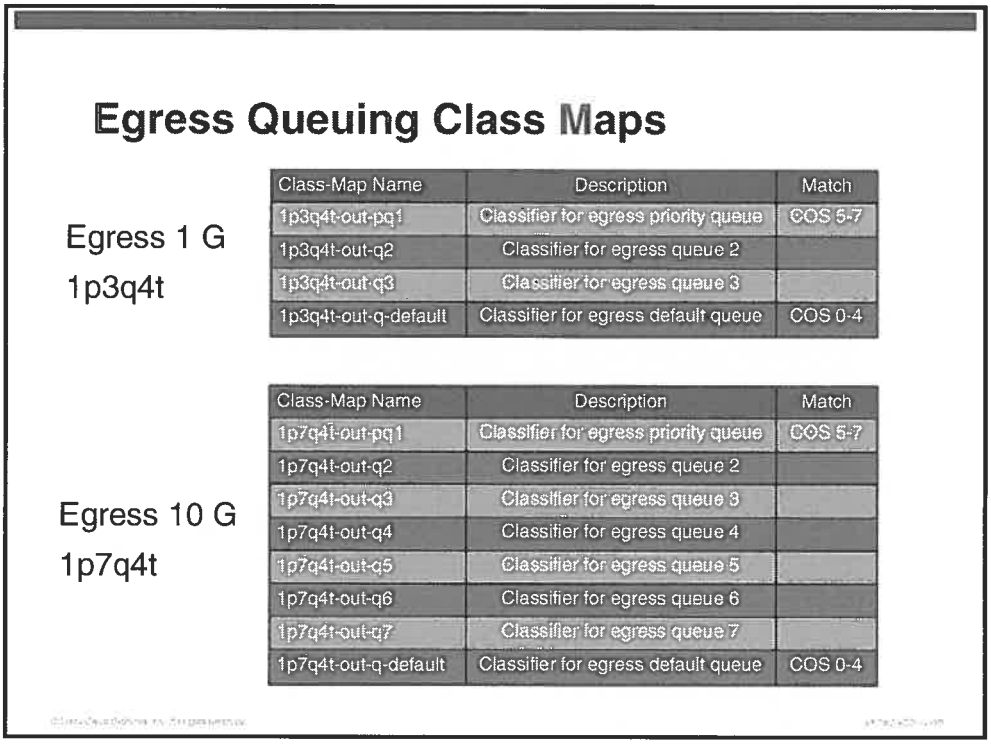
https://t.me/learningnets



On both user modules, ingress queues are available and the default COS-to-queue mapping exists. A 1 G card has 2 queues and a 10 G card has 8 queues. Regardless of which user module is used, the COS 0 to 4 goes into the queue default. The rest of COS values (COS 5 to 7) goes to the highest queue available (second queue on a 1 G card and the eight queue on a 10 G card). Therefore, the COS 5 to 7 traffic goes to the priority queue or just the highest priority queue. In case of ingress, there is no priority queue and the default configuration trusts the marking of the traffic coming into the NX 7010 system, so if the queuing policy on an ingress port is not changed, the default is to trust the values that are received.

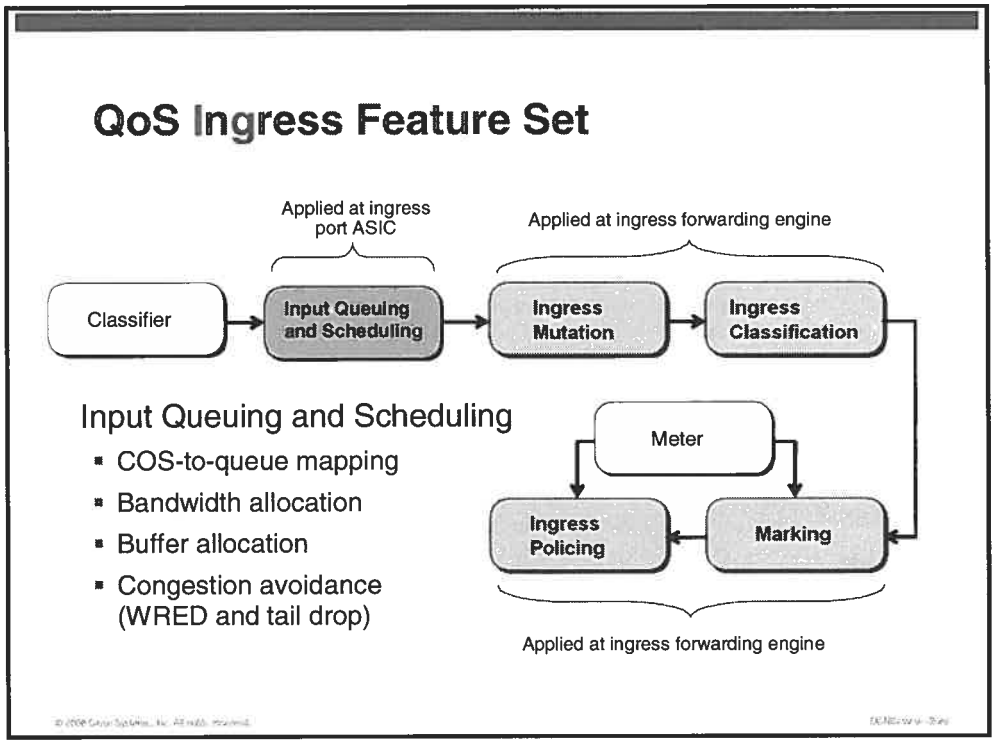
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Like on the ingress side of both user modules, egress queues are available and the default COS-to-queue mapping exists, too. A 1 G card has 4 queues with 1 priority queue and a 10 G card has 8 queues with one priority queue, as well. Regardless of which user module is used, the COS 0 to 4 goes into the queue default. The rest of COS values (COS 5 to 7) goes to the highest queue available (the fourth queue on a 1 G card and the eight queue on a 10 G card). Therefore, the COS 5 to 7 traffic goes to the priority queue or just the highest priority queue, so if the queuing policy on an ingress port is not changed, the default is to trust the values that are received.

https://t.me/learningnets



Cisco Nexus 7010 switch, from the data center perspective, has all the required QoS functionality. From a scheduling and queuing standpoint, Nexus 7010 switching platform is providing all the expected features – ingress classification, marking, and policing.

The sequencing of actions in a QoS policy varies for packets coming into the system from packets going out of the system. The order of ingress QoS actions is as follows:

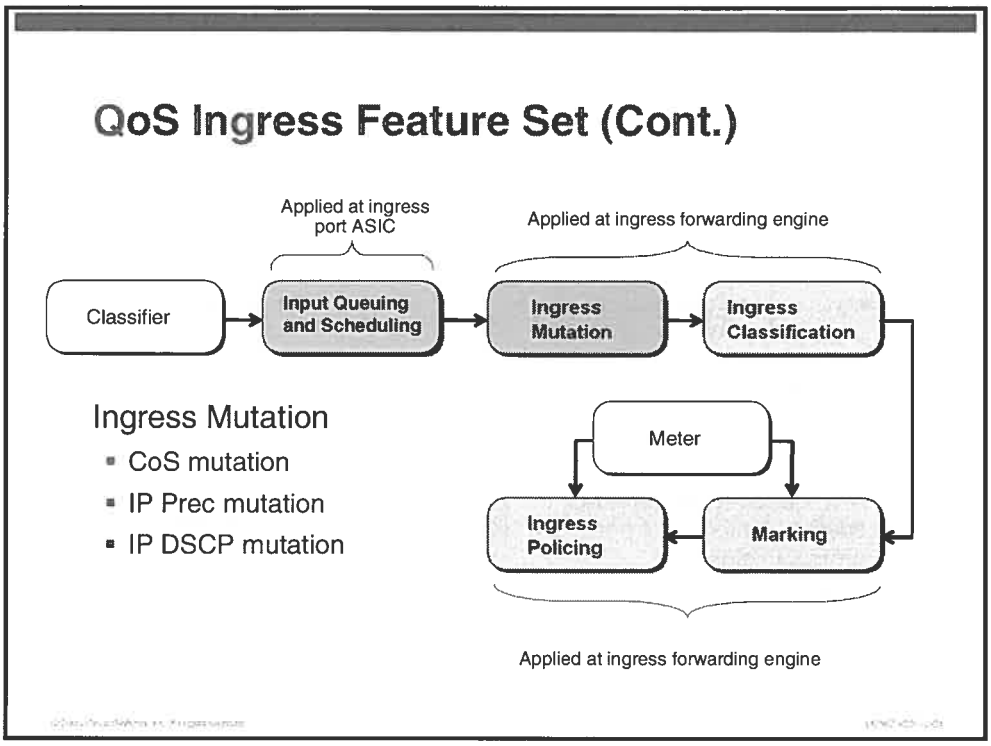
- Queuing and Scheduling (port ASIC)
  - Mutation (used to modify QoS fields in the ingress packet before traffic classification occurs)
  - Classification and determination of port trust state
  - Marking
  - Policing

The summary is providing the list of features, which can be used at the ingress port ASIC and at the ingress forwarding engine (ingress pipe):

- Input Queuing and Scheduling
  - COS-to-queue mapping
  - Bandwidth allocation
  - Buffer allocation
  - Congestion avoidance (WRED and tail drop)

© 2008 Cisco Systems, Inc.

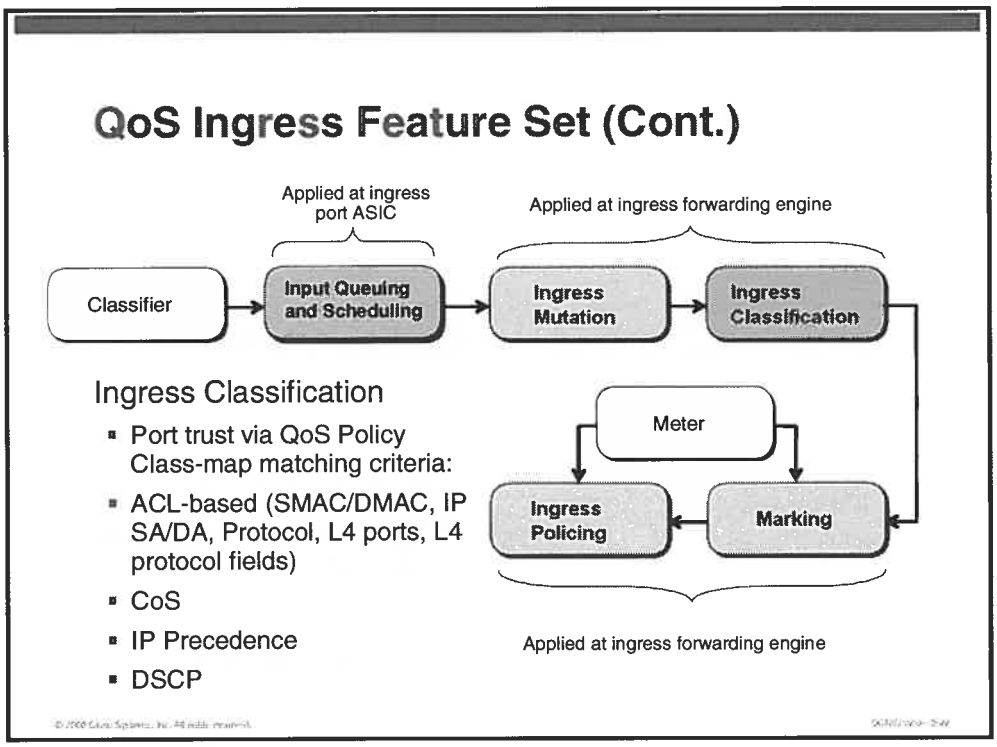
https://t.me/learningnets



The summary is providing the list of features, which can be used at ingress port ASIC and at ingress forwarding engine (ingress pipe):

- Ingress Mutation
  - CoS mutation
  - IP Prec mutation
  - IP DSCP mutation

https://t.me/learningnets

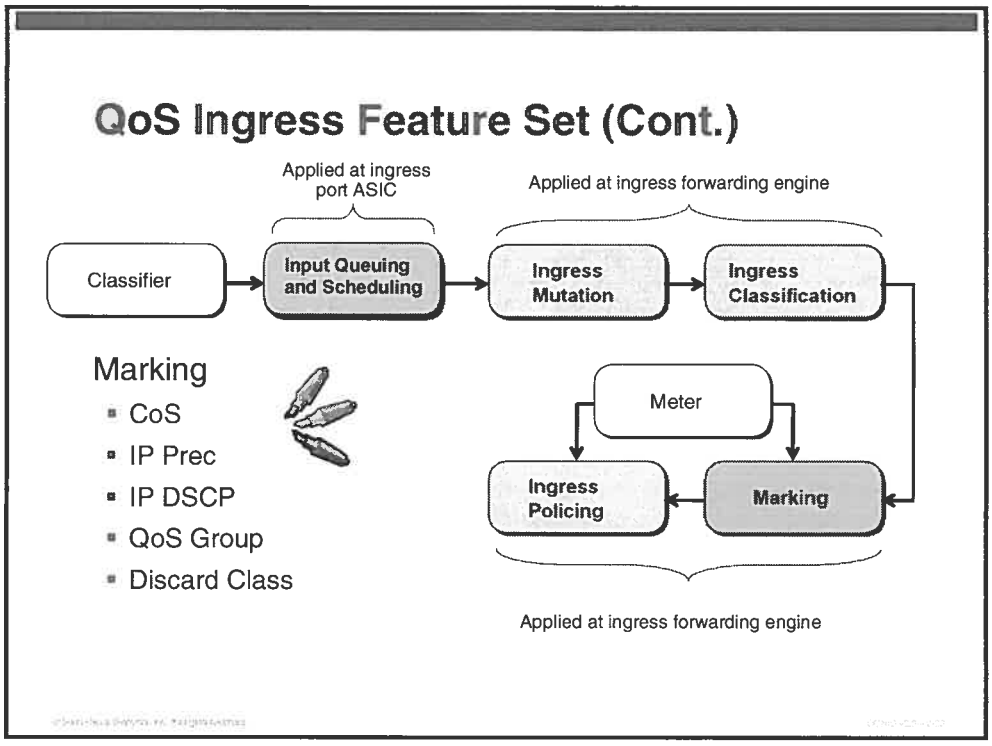


The summary is providing the list of features, which can be used at ingress port ASIC and at ingress forwarding engine (ingress pipe):

- Ingress Classification
  - Port trust via QoS Policy
  - Class-map matching criteria: ACL-based (SMAC/DMAC, IP SA/DA, Protocol, L4 ports, L4 protocol fields), CoS, IP Precedence, DSCP

© 2008 Cisco Systems, Inc.

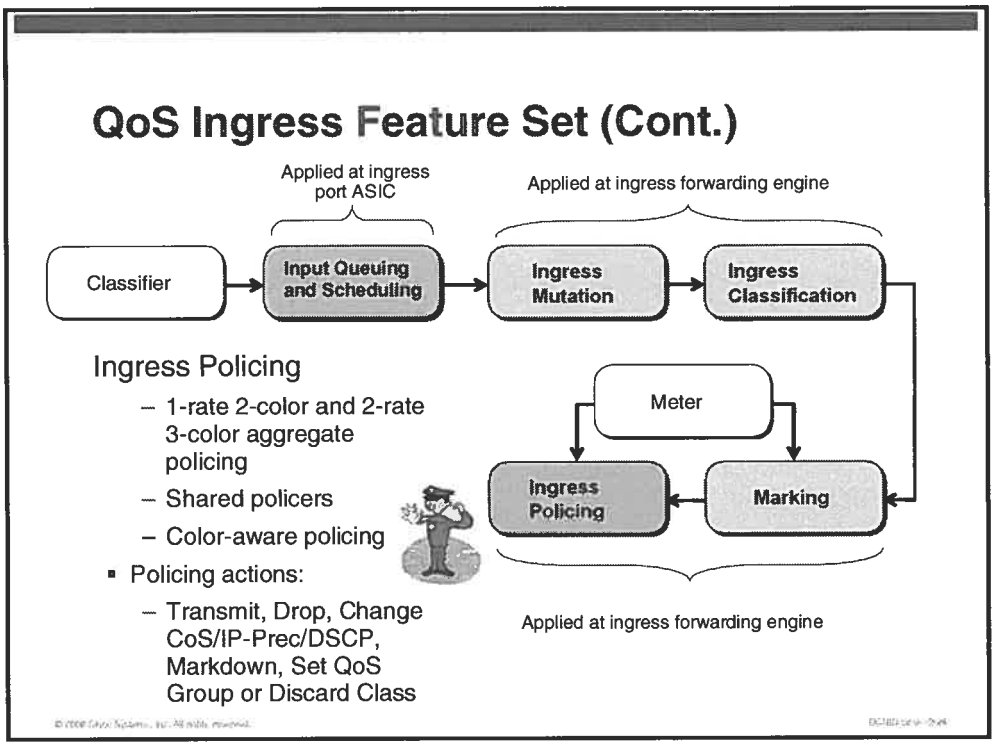
https://t.me/learningnets



The summary is providing the list of features, which can be used at ingress port ASIC and at ingress forwarding engine (ingress pipe):

- Marking
  - CoS
  - IP Prec
  - IP DSCP
  - QoS Group
  - Discard Class

https://t.me/learningnets

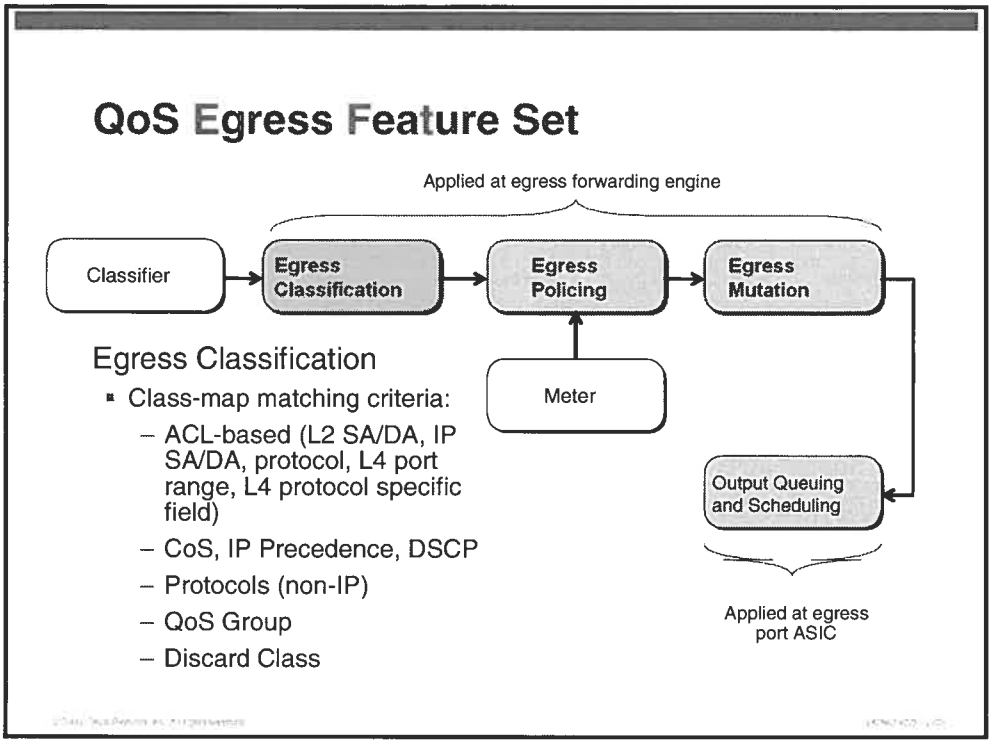


The summary is providing the list of features, which can be used at ingress port ASIC and at ingress forwarding engine (ingress pipe):

- Ingress Policing
  - 1-rate 2-color and 2-rate 3-color aggregate policing
  - Shared policers
  - Color-aware policing
  - Policing actions: Transmit, Drop, Change CoS/IP-Prec/DSCP, Markdown, Set QoS Group or Discard Class

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



On the egress side NX 7010 platform meet data center requirements too. On the egress side classification, policing, and then egress queuing and scheduling, including priority queue and shaped round robin is available.

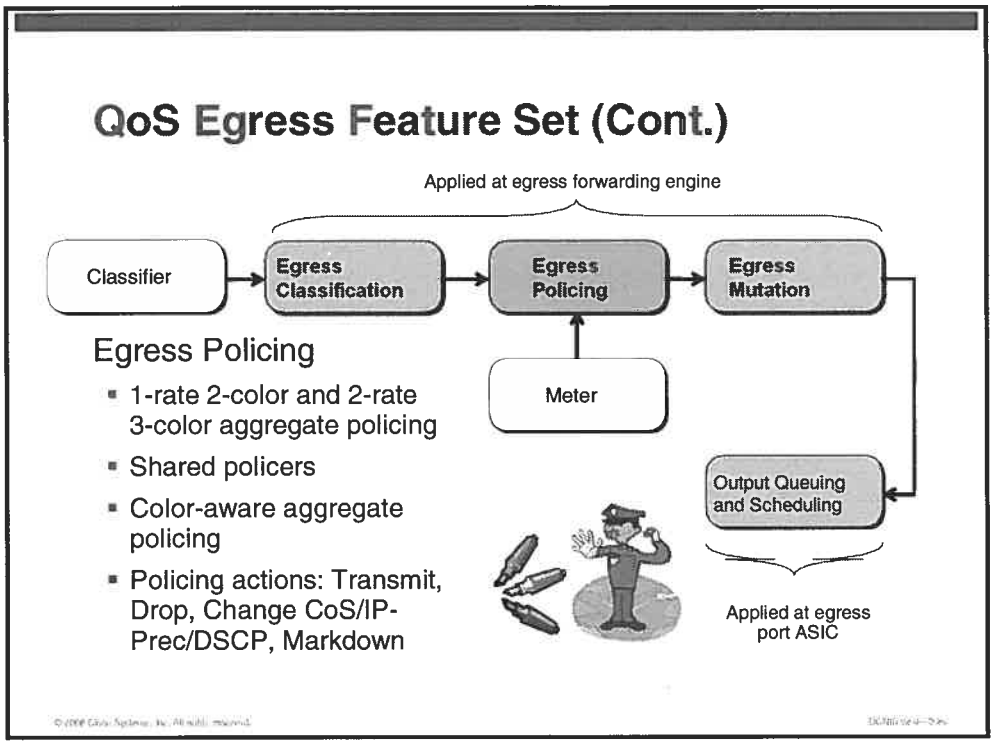
The sequencing of actions in a QoS policy varies for packets coming into the system from packets going out of the system. The order of egress QoS is as follows:

- Classification
- Policing
- Mutation (used to modify QoS fields in the egress packet before traffic classification occurs)
- Queuing and Scheduling (port ASIC)

The summary is providing the list of features, which can be used at egress forwarding engine (egress pipe) and at egress port ASIC:

- Egress Classification
  - — Class-map matching criteria: ACL-based (L2 SA/DA, IP SA/DA, Protocol, L4 port range, L4 protocol specific field), CoS, IP Precedence, DSCP, Protocols (non-IP), QoS Group, Discard Class

https://t.me/learningnets

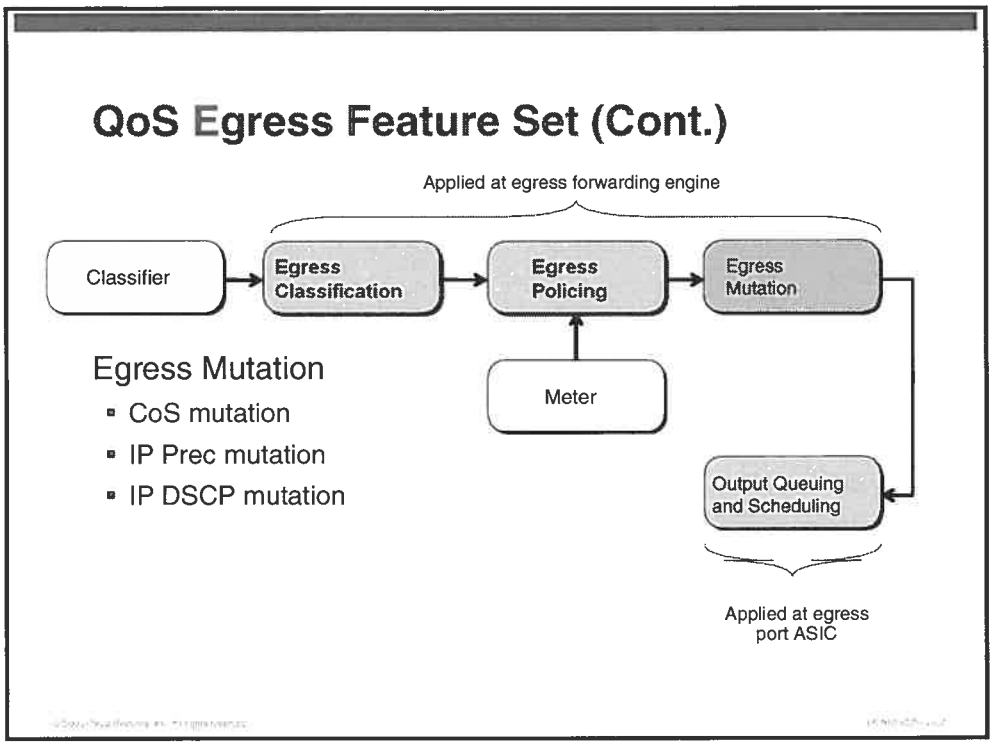


The summary is providing the list of features, which can be used at egress forwarding engine (egress pipe) and at egress port ASIC:

- Egress Policing
  - 1-rate 2-color and 2-rate 3-color aggregate policing
  - Shared policers
  - Color-aware aggregate policing
  - Policing actions: Transmit, Drop, Change CoS/IP-Prec/DSCP, Markdown

© 2008 Cisco Systems, Inc.

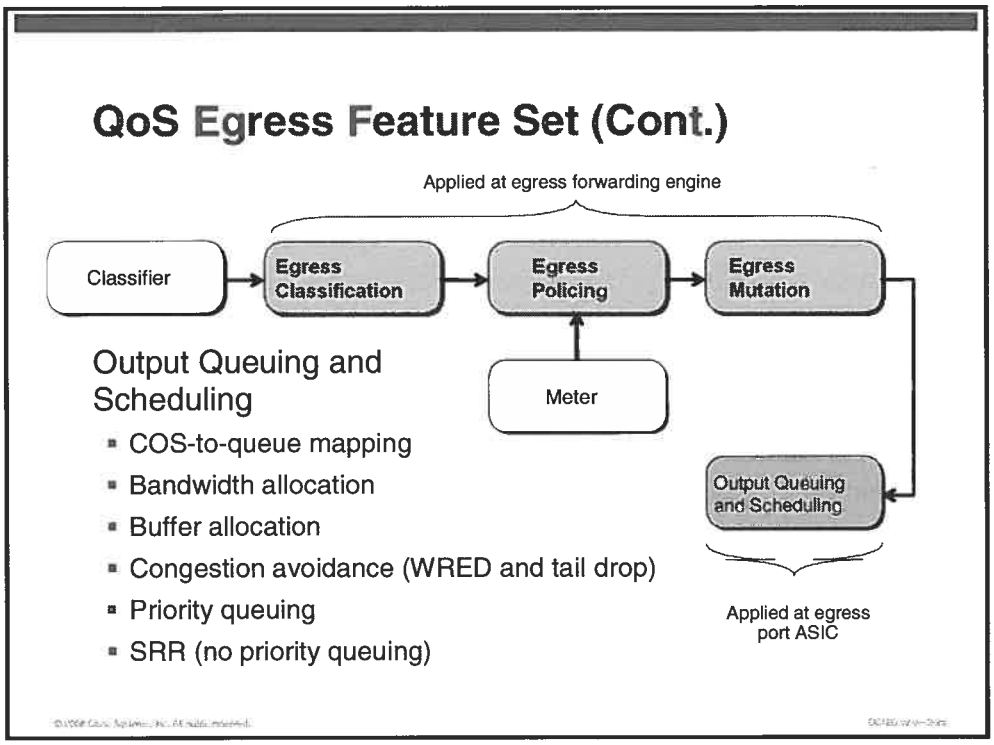
https://t.me/learningnets



The summary is providing the list of features, which can be used at egress forwarding engine (egress pipe) and at egress port ASIC:

- Egress Mutation
  - CoS mutation
  - IP Prec mutation
  - IP DSCP mutation

https://t.me/learningnets



The summary is providing the list of features, which can be used at egress forwarding engine (egress pipe) and at egress port ASIC:

- Output Queuing and Scheduling
  - COS-to-queue mapping
  - Bandwidth allocation
  - Buffer allocation
  - Congestion avoidance (WRED and tail drop)
  - Priority queuing
  - SRR (no priority queuing)

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

QoS Mechanisms Defined

- Classification: Identifying and splitting traffic into different classes
- Marking: Used to mark packets based on classification and/or metering
  - Mutation: Packet marking applied prior to classification
- Congestion Management: Queuing mechanism to prioritize transmission of packets
- Congestion Avoidance: Used to drop packets early in order to avoid congestion later in the network
- Policing/Shaping: Used to enforce a rate limit based on the metering

Classification

Classification is the separation of packets on an interface into traffic classes that can be referenced from multiple QoS policies. You create class maps to represent each traffic class of interest. You create classes of traffic by matching packet characteristics with the classification criteria (ACL, IP precedence, Differentiated Services Code Point (DSCP), protocol, QoS group, discard class, spanning tree BPDU, VLAN ID, packet length, CoS, RTP, TCP or UDP ports, class map, flow).

Marking

Marking is a method you use to modify QoS fields of incoming and outgoing packets. The QoS fields you can mark are CoS in Layer 2, and IP precedence and DSCP in Layer 3. QoS group and discard class are two labels local to the system that you can assign intermediate marking values that you can then use in determining the final values marked in a packet.

Mutation

Mutation is a special case of marking that you apply on all packets prior to classification into traffic classes. You can apply mutation to any of the packet fields supported by the Marking mechanism.

https://t.me/learningnets

Congestion Management

Congestion management mechanisms (queuing algorithms) use the marking on each packet to determine in which queue to place packets. Different queues are given different treatment by the queuing algorithm based on the class of packets in the queue. Generally, queues with higher priority packets receive preferential treatment. All output interfaces in a QoS-enabled network use some kind of congestion management (queuing) mechanism to manage the outflow of traffic. Each queuing algorithm was designed to solve a specific network traffic problem and has a particular effect on network performance. The NX 7010 software features for congestion management or queuing include:

- WRR (Weighted Round Robin, Default)
- SRR (Shaped Round Robin)
- Strict Priority Queuing (Enabled by default, and can only be used with WRR)

Congestion-Avoidance

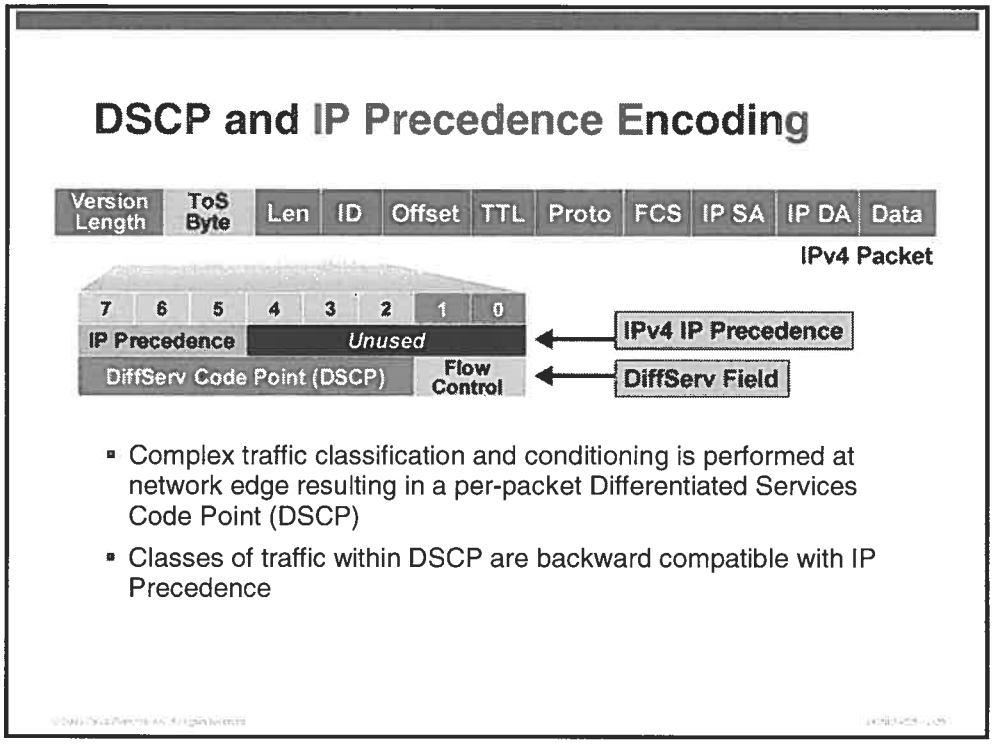
Congestion-avoidance mechanisms monitor network traffic loads in an effort to anticipate and avoid congestion at common network bottlenecks. Congestion avoidance is achieved through packet dropping. Congestion avoidance mechanisms are typically employed on output interfaces wherever a high-speed link or set of links feed into a lower-speed link (that is, a LAN feeding into a slower WAN link.) This ensures that the WAN is not instantly congested by LAN traffic. Tail drop is default behavior of the Nexus 7010 switch. WRED is a Cisco primary congestion-avoidance technique. WRED increases the probability that congestion is avoided by dropping low-priority packets rather than high-priority packets.

Policing

Policing mechanism is used to condition traffic before transmitting traffic to a network or receiving traffic from a network. Policing is the ability to control traffic bursts and conform traffic to ensure that certain types of traffic get certain types of bandwidth. Policing drops or marks packets when predefined limits are reached. Policing mechanisms can be set to first drop traffic classes that have lower QoS priority markings. Policing mechanisms can be used at either input or output interfaces.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The DiffServ architecture is based on a simple model where traffic entering a network is classified and possibly conditioned at the boundaries of the network. The traffic class is then identified with a DSCP or bit marking in the IP header.

The DiffServ model uses the DiffServ field in the IP header to mark packets according to their classification. The DiffServ field occupies the same eight bits of the IP header that were previously used for IP Precedence and called ToS byte inside IPv4 packet header.

There are three IETF standards describing the purpose of those eight bits:

- RFC 791 includes specification of the ToS field where the high-order three bits are used for IP precedence. The other bits are used for delay, throughput, reliability, and cost.
- RFC 1812 modifies the meaning of the ToS field by removing meaning from the five low-order bits (those bits should all be zero). This method gained widespread use and became known as the original IP precedence.
- RFC 2474 replaces the ToS field with the DiffServ field where the six high-order bits are used for the DSCP. The remaining two bits are used for explicit congestion notification.

Each DSCP value represents the class of traffic. Each traffic class is assigned a per-hop behavior (PHB). Each PHB is implemented using the appropriate QoS mechanism or a set of QoS mechanisms.

https://t.me/learningnets

QoS Configuration

- QoS Modular QoS CLI (MQC) is used for configuration
- MQC defines a three-step configuration model:
  - Define match criteria—class-map
  - Associate action with each defined class—policy-map
  - Attach policy to interface or VLAN in a particular direction—service policy

One of the main advantages of the Nexus 7010 platform versus other switching platforms is support for the IOS Modular QoS CLI model (MQC). MQC is object oriented configuration model which is easy to use, has very good fine tuning capabilities and excellent modularity. Rather than configuring "raw" CLI commands interface by interface, you develop a uniform set of traffic classes and QoS policies, which can be applied on interfaces. The use of the MQC allows the separation of traffic classification from the definition of QoS policy. This enables easier initial QoS implementation and maintenance as new traffic classes emerge and QoS policies for the network evolve. This configuration model will make Nexus 7010 platform consistent with all other IOS platforms used at Cisco.

MQC is a three-step configuration model where you define:

Step 1 Class-maps, which identify the traffic that you want to do something with.

Step 2 Policy-maps, which define actions to take on configured classes.

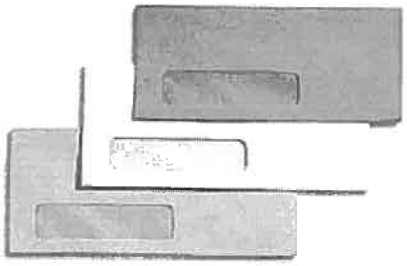
Step 3 Service policy. Once you have the policy-map defined, you can attach it as a service policy either to a physical interface or a VLAN interface or sub-interface.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Class Types

- Queuing classes identify traffic based on COS
- QoS traffic classes identify traffic according to match criteria



Classification is the separation of packets on an interface into traffic classes. Classification has the following guidelines and limitations:

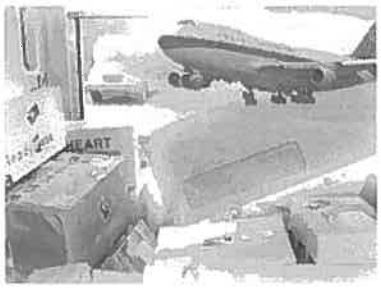
- You can specify a maximum of 1024 match criteria in a class-map.
- You can configure at most 4096 classes for use in a single policy-map.
- For Layer 2 interfaces, traffic can be classified based on the port or VLAN ID of the incoming packet, but not both. If a policy-map includes both VLAN- and port-based polices, the system will act on the VLAN-based policies, while ignoring the port-based policies.

QoS classes for Nexus 7010 switch are not different than existing classes used in 6500 Catalyst switch other than some additional capabilities. The big difference is the queuing classes. Rather than having WRRQ and the receive queue port-based configurations for configuring the queuing and scheduling, it is now required to configure queuing classes and then queuing policies. This is how the queue structures, the cost of queue mappings, and the scheduling queue sizes are defined, as well as WRED thresholds.

https://t.me/learningnets

Policy Types

- Queuing policies are utilizing port ASIC
- QoS marking and policing policies mark or police traffic utilizing CL TCAM



As we have queuing and QoS classes, we have also queuing and QoS policies. So once you classify the traffic for queuing or for QoS functions, you then reference those classes within a policy of the same type. So if you have a class-map that is a queuing class-map, you will use a queuing policy and reference those queuing classes to take some action on them. QoS marking and policing policies are not different than the ones used in Cisco IOS QoS implementation, where you are going to have a mark down or police. Remarking policy based on the classes is available as well.

The Nexus 7010 switch is using ASIC capabilities and CL TCAM for QoS. Queuing policies leverage port ASIC capabilities to map traffic to queues and schedule packet delivery. QoS marking and policing policies mark or police traffic, leveraging CL TCAM to match packets.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

QoS Actions per Queue

- Priority defines a queue as the priority queue
- Bandwidth defines WRR weights for each queue
- Shape defines SRR weights for each queue
  - Note: Enabling shaping disables priority queuing support for
- Queue-limit defines queue size and defines tail drop thresholds
- Random-detect sets WRED thresholds for each queue

QoS mechanisms are techniques to manage network resources. The assumption that a high capacity non-blocking switch with multigigabit backplanes never needs QoS is incorrect. Many networks or individual network elements in data center are oversubscribed and it is easy to create scenarios in which congestion can potentially occur and that there require some form of QoS treatment. QoS actions are applied per traffic class during the policy configuration.

Priority queuing is required to prioritize important traffic to best effort or other traffic. "Priority" QoS action is taking care that the priority queue is used. Classification must be configured to classify the important traffic into the priority queue. In the case that bandwidth guarantee is required; again classification is the first step, where the correct traffic is classified into the traffic class, which requires the bandwidth guarantee. Limiting the traffic is related to policing and/or shaping. If priority and bandwidth commands are providing some low delay and/or bandwidth guarantees, then limiting the traffic is dropping or delaying the packets, which are exceeding the configured allowed rate. In all queuing action queue size can be defined. This definition is related to the amount of buffers available and to the treatment we would like to assure. Small buffers are reacting faster, but every burst can trigger the action. Bigger queues are reacting with some delay, which is good, but the queuing can give you different performance. When buffers are full, the default action is tail drop. All the additional packets are dropped regardless of the priority and importance. WRED is providing more intelligent drop of the packets where packets are randomly dropped and the system is slowing down only a few TCP sessions and, as has been statistically proven, these sessions are more aggressive sessions.

https://t.me/learningnets

Trusted/Untrusted Port Policy

- Default policies "trust" received QoS values
- Default policy cannot be modified
- Non-defaults need new policy
  - For untrusted port, use queuing policy to place all traffic in q-default and set all frames to COS 0

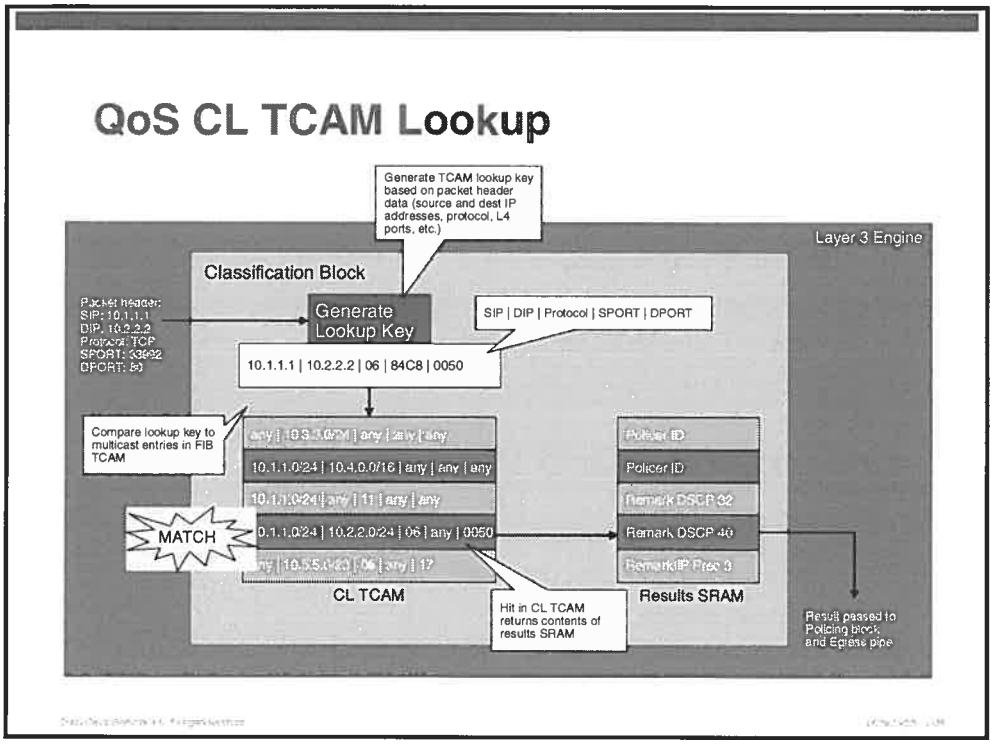
© 2008 Cisco Systems, Inc. All rights reserved.

Default queuing policies "trust" received QoS values. Default policy cannot be modified. To use non-defaults, create a new policy and apply it to the desired interfaces.

The default is essentially to trust the values and preserve them as we go across a switch. This being a data center switch, the expectation is that you have control over your data center and what is actually being transmitted and received by the data center. If you do not have control over your data center, you can always create an untrusted policy for your ports, but doing so is not the default. In case you do not trust the port, you can put all of the traffic into one class (class-default) and apply COS 0 to all frames in that class. There is no MLS QoS port trusted or whatever the CLI is on the 6 K. There is no per-interface config command to do that. You have to define an untrusted policy and then apply that untrusted policy to the ports that you want to be untrusted.

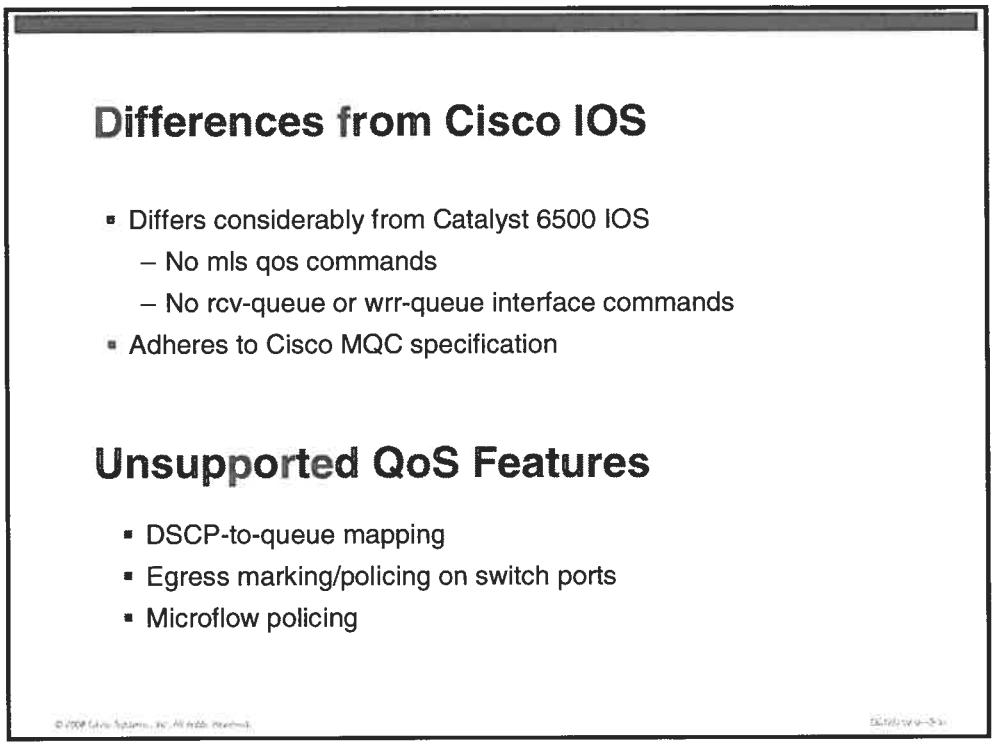
© 2008 Cisco Systems, Inc.

https://t.me/learningnets



We have a classification block in Cisco Nexus 7010 switch modules. When a packet comes in, a large amount of information is taken out of the packet header and a lookup key is generated based on this information. The diagram shows a source and destination IP protocol and ports, which are compared to all the entries in the classification TCAM. FIB TCAM term is used because it is exactly the same process. But rather than returning a result that points to an adjacency or an RPF interface, the system points to either a policer index or just a remark value. Since DSCP is only 64 values, and the system can easily contain that within the result SRAM, it is a small value. So Cisco Nexus 7010 Switch gives the remark value right there in the result field and this tells us what QoS value to remark the packet to. If a more sophisticated process like policing is done, then this policer index is used later in the policing block to identify the policer that this packet belongs to, and then the policing block takes care of tokens, etc.

https://t.me/learningnets



Cisco Nexus 7010 operating system is not the same as Cisco IOS. It differs considerably from Catalyst 6500 IOS. All the port MLS QoS port configuration commands are not present. The only exception to the MQC rule here is the aggregate policers are being configured globally. Aside from that, everything QoS-related is integrated into the class-map, policy-map, or service policy model.

DSCP-to-queue mapping might be the most important feature, but it is currently unavailable.. Next one is microflow policing, interesting for COPP more than for any other QoS-type application in the data center. Microflow policing is not used very often in a data center. For a residence hall in a university, microflow policing and UVRL makes sense. The last important QoiS feature missing is also egress marking/policing on switchports. This is a continuing limitation of the Forwarding Engine, which is, you cannot do an egress policy on a per-Layer 2 switch port basis. So egress polices are always tied to a Layer 3 entity, be it a routed interface, routed sub-interface or a VLAN, but not a per-Layer 2 switch port basis. It is working on ingress only.

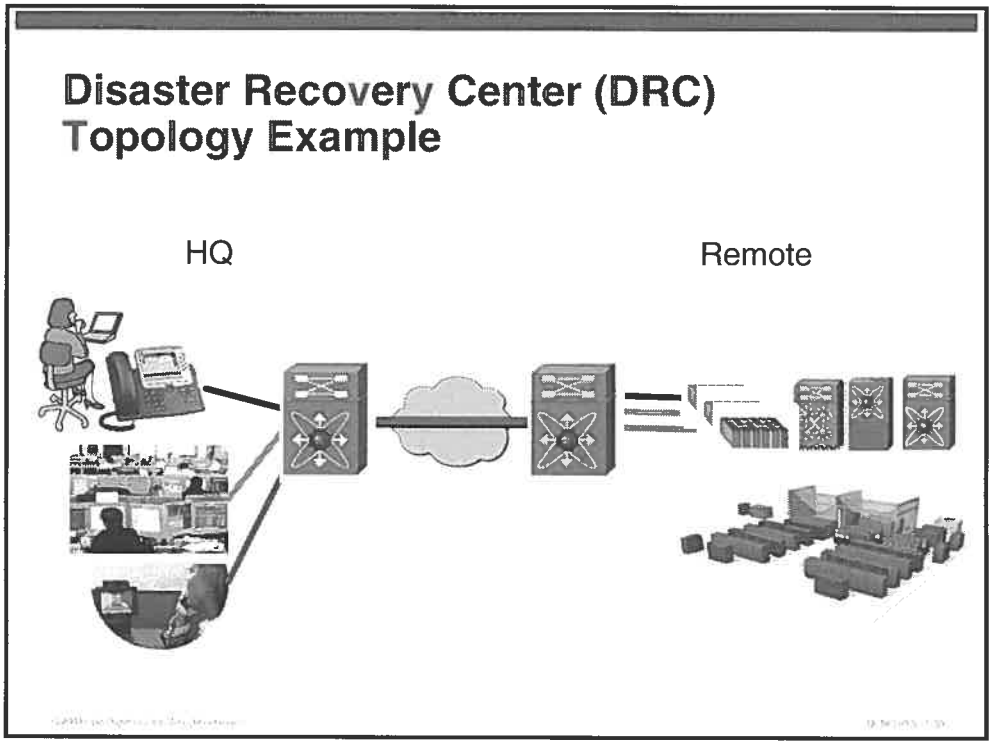
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

Data Center Network QoS

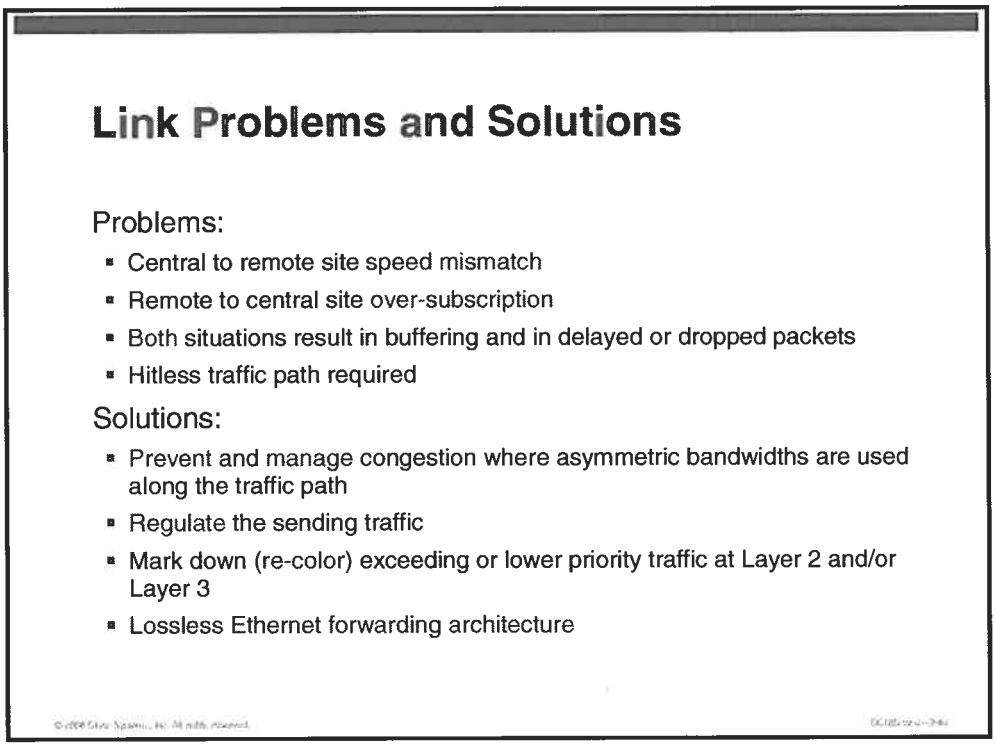
This topic describes the QoS and its solutions for the data center environment.

Disaster Recovery Center (DRC) Topology



In this example, the company HQ network is connected to the remote site using Nexus 7010 Series equipment. The link between the locations is provided by ISP offering a dedicated link between the sites via optical fiber. The remote site is used as a Disaster Recovery Center (DRC) and provides additional equipment for redundancy of voice and video services. The following slides will describe possible problems and solutions related to QoS.

https://t.me/learningnets



QoS mechanisms are managing network resources and avoiding problems the user might face.

A few common problems are:

- Central to remote site speed mismatch, where central site is using higher bandwidth link and remote site is using lower bandwidth link.
- Remote to central site over-subscription, where many remote sites are sending traffic to the central site and the bandwidth of the link at the central site becomes too low for all the remote sites.
- Both of these situations result in buffering and in delayed or dropped packets as speed mismatch and over-subscription might happen at the same time.
- Hitless traffic path is required, where we cannot lose even one Fiber Channel frame.

There are many solutions we might provide:

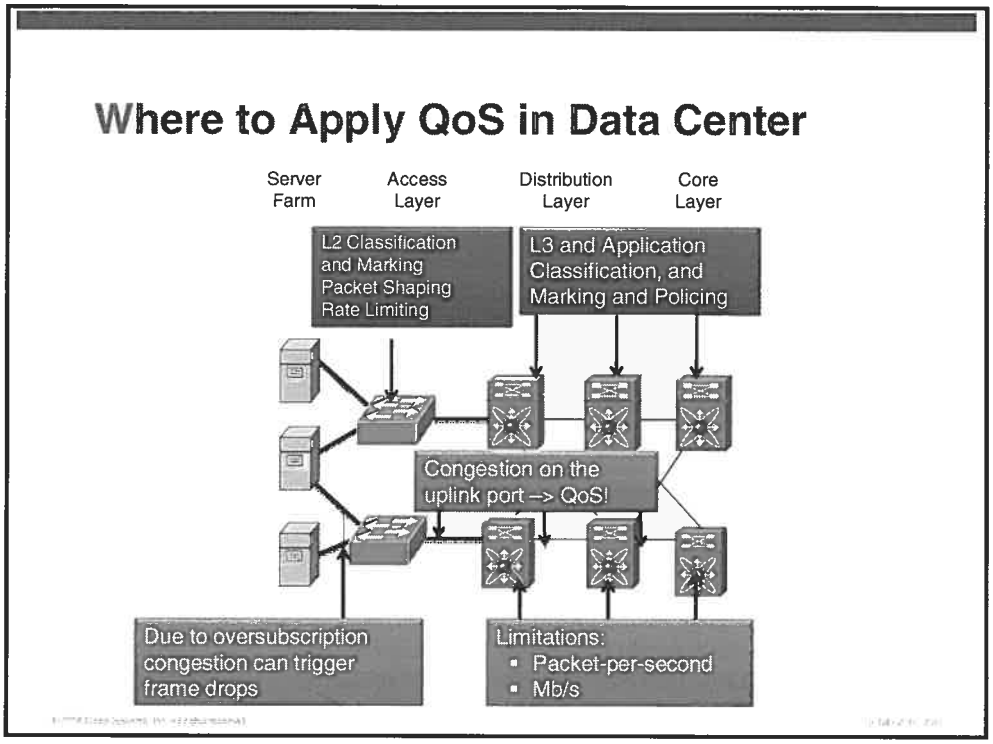
Network devices involved in the interconnection must prevent and manage congestion as well as regulate sending traffic by limiting the amount of traffic sent into the network from the source. When relegating the traffic the limiting action can be applied or network administrator can decide to remark the frames. Important traffic is tagged with a different marker than best effort traffic. Later on in the network important traffic gets better treatment than best effort traffic.

Where hitless traffic path is required Cisco Nexus 7010 system provides lossless Ethernet forwarding architecture using VOQs and the non-blocking capacity of switch fabric module(s).

Depending on the traffic flow and uplink over-subscription, bandwidth is managed with QoS mechanisms on the access, distribution, and even on the core layer switches.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Data center networks provide secure, predictable and measurable guaranteed services. They are achieving the required QoS by managing the delay, the delay variation, the bandwidth, and the packet loss parameters on the network. This is the key to a successful end-to-end business solution.

Many networks or individual network elements in the data center might be oversubscribed, or it is easy to create scenarios in which congestion can potentially occur and that requires some form of QoS treatment. The sum of the bandwidth on all ports on the switch where end devices are connected is usually greater than that of the bandwidth using an uplink port when the access ports are fully used. Congestion on the uplink port is unavoidable. Uplinks from the data center access to the data center distribution layer or from distribution to the core layer must often require QoS. Switch over-subscription occurs when a switch allows more ports (bandwidth) in the chassis than the switch hardware or the common channel can transfer smoothly through the internal structure. The over-subscription is the ratio between the potential bandwidth and throughput that can be sent by the servers to the switch ports and the real throughput capacity of the switch architecture. Depending on the traffic flow and uplink over-subscription, bandwidth is managed with QoS mechanisms on the access, distribution and even on the core layer switches.

Congestion avoidance techniques monitor network traffic loads so that congestion can be anticipated and avoided before it becomes a problem and requires packet drop, smaller TCP window size, or finally restarted TCP sessions or HTTP sessions, or SCSI I/O operations. If congestion avoidance techniques are not used, then when an interface queue becomes full, packets trying to enter the queues are discarded regardless of what traffic they hold. This process is called tail drop. The key point here is that congestion-avoidance techniques allow packets from the streams identified to be dropped when the queue is getting full.

© 2008 Cisco Systems, Inc.

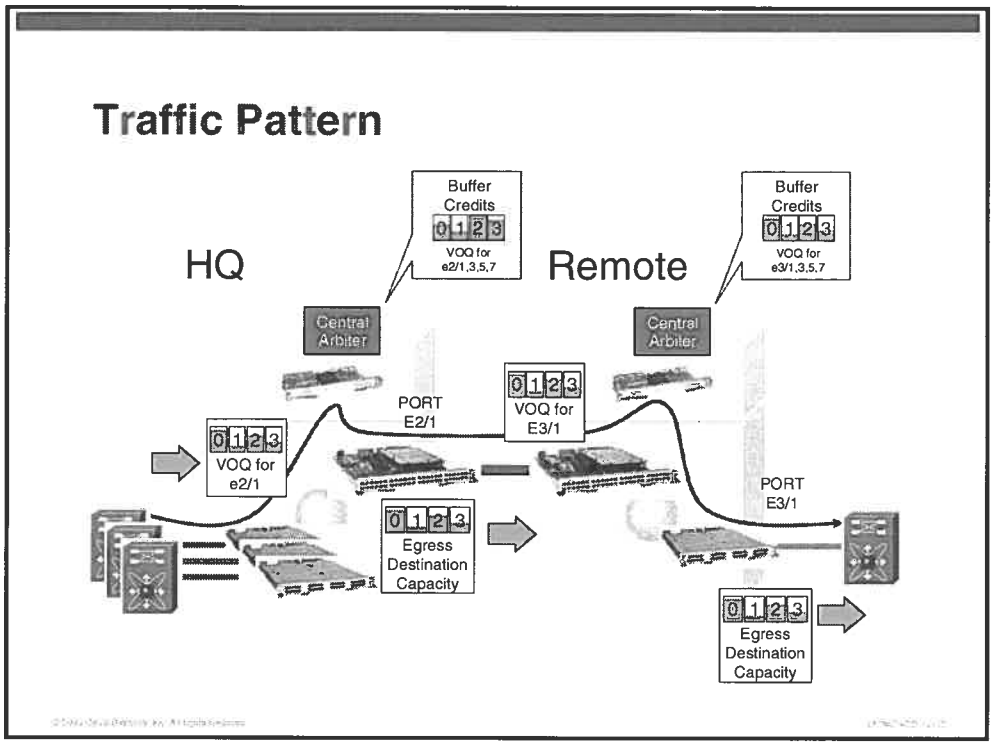
https://t.me/learningnets

Within a data center server is a high-capacity building block that is attached to the Campus Core network. The main concern with the Server Farm module is that it receives the majority of the traffic from the entire campus. Random frame drops can occur because the uplink ports on the switch are over-subscribed. To guarantee that no random frame drops occur for the business-critical applications, the network designer should apply QoS mechanisms to the server links. On access layer L2 classification, marking and limiting of the packets is applied. Classification is always applied prior to any other QoS action. For traffic classification, many different CPU intensive steps might be required. To allow easier classification in the distribution and/or core layer, traffic is marked at the access layer. Therefore, marking is one of the actions applied to the traffic class. The next action used at the access layer is limiting the traffic to avoid too much traffic entering the network.

On the distribution layer, additional classification might be used as well as limiting of the traffic and intelligent dropping (Cisco WRED mechanism). If the distribution layer is still using many QoS mechanisms then the core layer is designed for fast traffic transport, and it is not recommended to apply CPU-intensive QoS steps.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

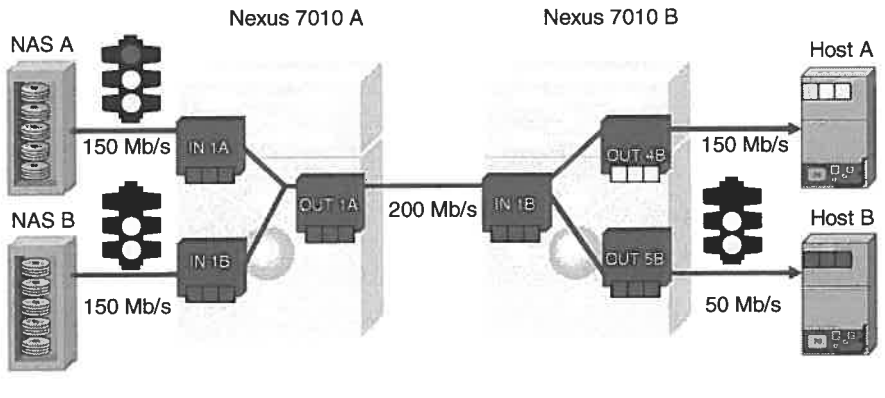


Frames are coming from the user (in HQ) to the ingress module in the Cisco Nexus 7010 system. Ingress modules contain VOQ for the destination ports on the egress module showing the free space. Ingress module asks the supervisor module central arbiter for the frame to be grated permission to be sent. As bandwidth is available the frames can be sent out to the link between HQ and the remote site. When a frame is sent out, the released buffers are reported to supervisor module (to the central arbiter). On the remote site, the situation is the same. The ingress card asks the central arbiter if the egress card has some free bandwidth available. As bandwidth becomes available, the frame is sent to the egress card and out to the backup server.

https://t.me/learningnets

Congestion in Cisco Nexus 7010 Switch

Congestion can occur when multiple ports compete for access to a shared resource.



Congestion in Cisco Nexus 7010 switch can occur when multiple sources compete for access to a shared data path. For example, the diagram shows a network where two hosts retrieve data from two NAS (Network Access Storage) servers over a 200-Mb/s link:

- Both NAS A and NAS B can send data at 150 Mb/s, and Host A can receive data at 150 Mb/s, but Host B can receive data at only 50 Mb/s.
- Because NAS B is sending data faster than Host B can receive that data, port OUT5B will eventually run out of buffers and will detect congestion.
- When port OUT5B runs out of buffers, port IN1B will be buffered by mainly the frames destined for both OUT5B. To prevent lost frames, the switch will hold frames as long as possible and drop them if they are kept for more than 500 miliseconds. Eventually, IN1B also runs out of buffers.
- Congestion proceeds backwards from IN1B to OUT1A via the Ethernet pause mechanism.
- When OUT1A fills the buffers, the congestion is detected, and both IN1A and IN2A begin to buffer frames and sooner or later run out of buffers as well.

As we talk about Ethernet frames exchanged between NASs and Hosts we have multiple TCP sessions running. When buffers are full, the TCP frames and the TCP window size will be reduced and the frames will be tail dropped.

NASs will retransmit the frames which were dropped. At the same time Ethernet ports supporting flow control will send PAUSE frames back to the NAS servers to the neighboring port, sending the frames to the congested port. The PAUSE frames are sent closer and closer to the source of traffic until they reach the NAS source of the traffic. At the same time, the TCP session is retransmitting frames and the whole system is providing lossless traffic stream.

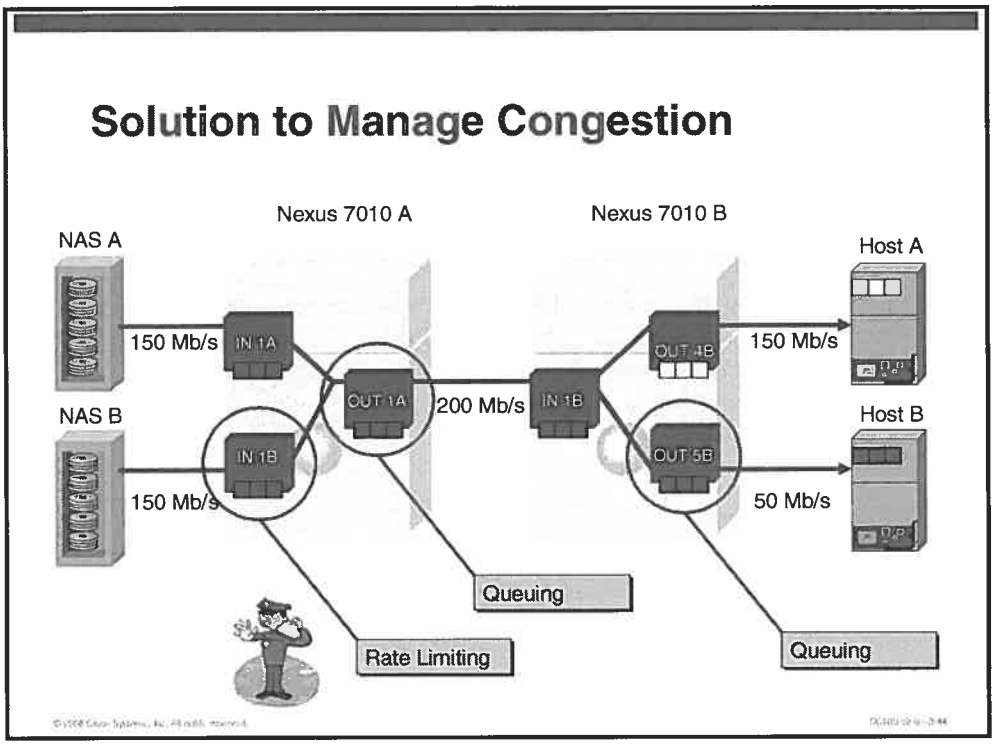
© 2008 Cisco Systems, Inc.

https://t.me/learningnets

In this scenario, NAS B causes congestion by sending data faster than host B can receive that data. When IN1B runs out of buffers, IN1B can no longer receive data, which means that no nodes can use the link between Nexus 7010 A and B. Bandwidth between port IN1B and OUT1A will be available but unusable, and data transfer between NAS A and Host A will be unnecessarily impacted.

The final impact of the wrong design is that both NAS servers will slow down the data transport, at both levels at the TCP level, the Window size will be reduced, TCP retransmissions will occur and eventually TCP sessions will be dropped and restarted. The congestion will spread from Host B back to the switches and to the NAS servers. At the Ethernet level the Ethernet pause frames will be sent back to the NAS servers. TCP sessions between NAS servers and hosts will be restarted, and torn down if the congestion is not released within 9 seconds.

https://t.me/learningnets

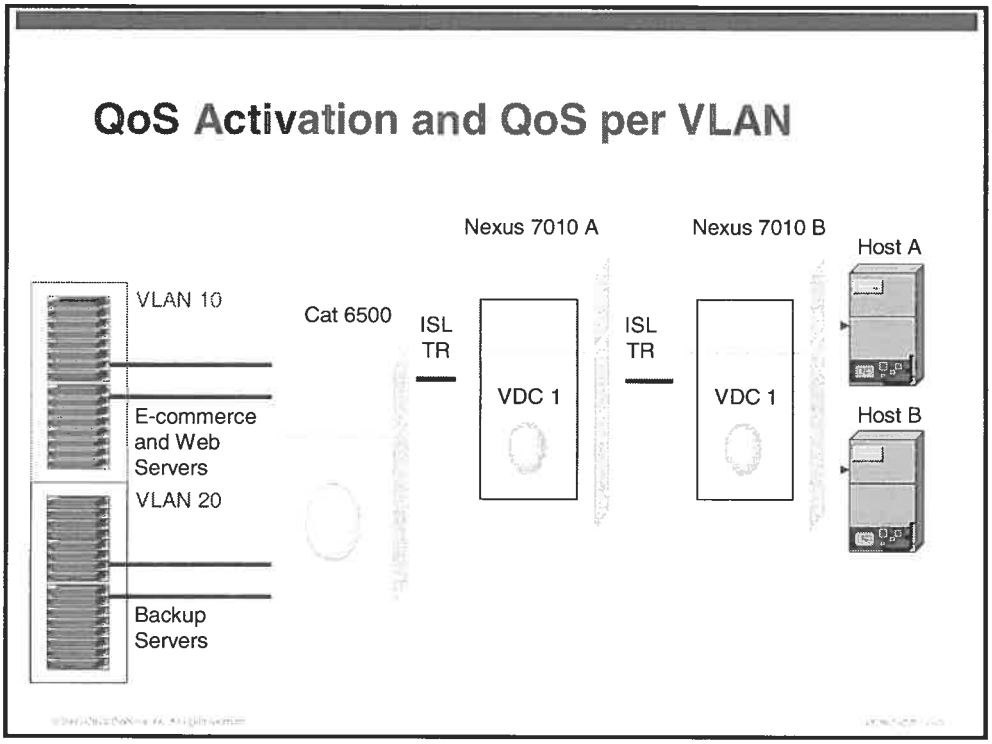


Interesting example from the real network is a set of NAS devices or server serving the hosts. Congestion can happen on many parts of the network and QoS is required to remove all the possible problems related to congestion and priority treatment of the traffic.

When the source of the traffic is sending too much traffic into the system the design rule is to rate limit the traffic as close to the source as possible. The main reason to rate limit is because traffic which is generating congestion must be removed or limited as soon as possible to prevent congesting the links and using CPU power for switching those frames. Besides congestion, we need to change the priority treatment to the traffic going to different destinations as well as to the traffic going to the same destination. In the second case, the Cisco Nexus 7010 system is treating some traffic as higher priority traffic.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



Scenario

In the example shown above, the company is using several servers providing E-commerce, Web and Backup services. All the servers are connected to the Cisco Catalyst switch 6500 using different VLANs. E-commerce and Web servers are using VLAN 10, and Backup servers are using VLAN 20. Cisco Catalyst switch is connected to the Cisco Nexus 7010 system via ISL trunk sending all the traffic into the single VDC inside (VDC1). The first Cisco Nexus 7010 Switch is receiving the traffic and will require QoS configuration. The next Cisco Nexus 7010 Switch is then serving different hosts. The topology can be different from case to case.

In this scenario, the company is interested in defining different priorities to traffic in different VLANs. The company is also interested in providing more bandwidth to the E-commerce and Web servers during working hours. During the remaining part of the day, traffic from Backup servers takes over priority in the Cisco Nexus 7010 system.

Solution

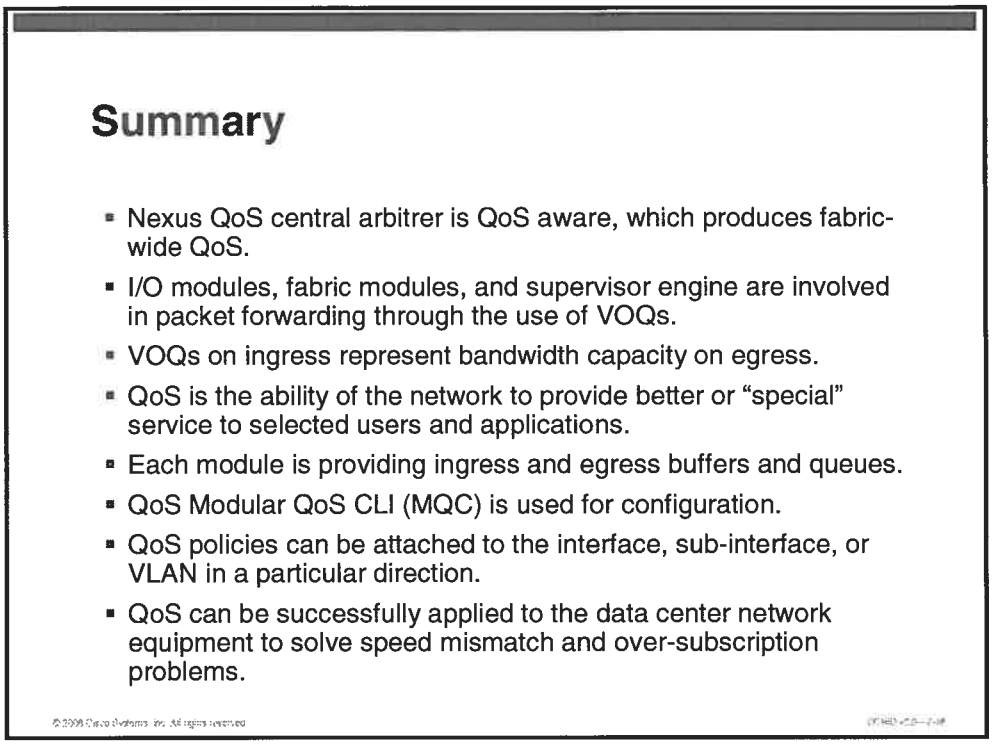
The Cisco Catalyst switch is receiving all the traffic from the servers. All the frames are tagged with the appropriate VLAN and all the traffic is sent toward the first Cisco Nexus 7010 device. In the first device, classification is required at the beginning. The next step is to configure correct actions for the traffic (limiting the traffic, prioritization, etc.). Because the company has different policies for different groups of servers, the policy defined is applied to VLANs. Each group of servers belongs to a specific VLAN and this is the best way to apply the policy.

The next important consideration is the scheduling of the policies. The company requires different policies applied to the traffic at different times of the day. The Cisco Nexus 7010 system is supporting QoS activation where one service policy is applied for the working hours and another policy for the rest of the day. The service policy during working hours is prioritizing the VLAN 10 traffic as this is important during the working hours and this service policy gives more bandwidth to E-commerce and Web applications. During the rest of the day Backup traffic is important for database replication and more bandwidth is required for VLAN 20 traffic. Therefore, for the rest of the day the QoS activation inside Cisco Nexus 7010 is changing the policy applied per VLAN.

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

Lesson 7

Cisco Nexus 7010 Switch Positioning in the Data Center

Overview

This lesson positions the Nexus 7010 Series Switch in the data center networking architecture.

Objectives

Upon completing this lesson, you will be able to describe how the Cisco Nexus 7000 Series Switches fit into the data center architecture. This ability includes being able to meet these objectives:

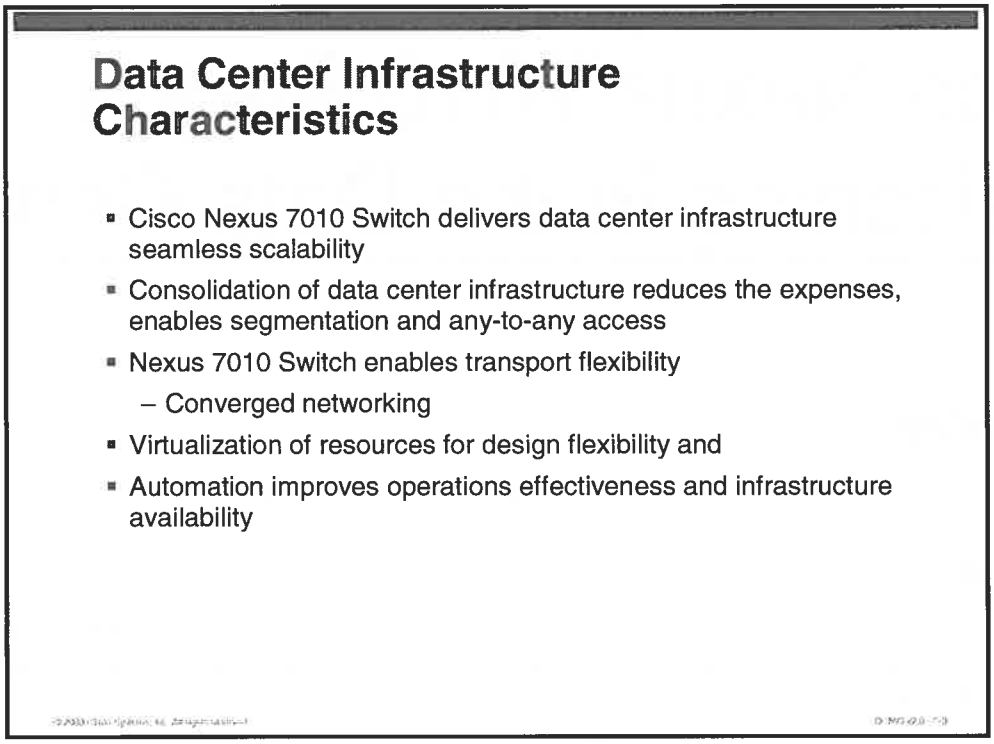
- Describe the vision behind the Nexus 7000 Series Switches in the new data center architecture
- Describe where the Nexus 7010 Series Switch fits in the new data center converged network architecture
- Clarify the difference in the role of the Nexus 7010 Series Switches as opposed to the role of other data center switches, such as Catalyst 6500 and 4900 Series Switches and others in the data center architecture

https://t.me/learningnets

Cisco Nexus 7010 Switch Data Center Positioning

This lesson explains how to position the Nexus 7010 Series Switches in the data center infrastructure.

Characteristics of the Data Center infrastructure

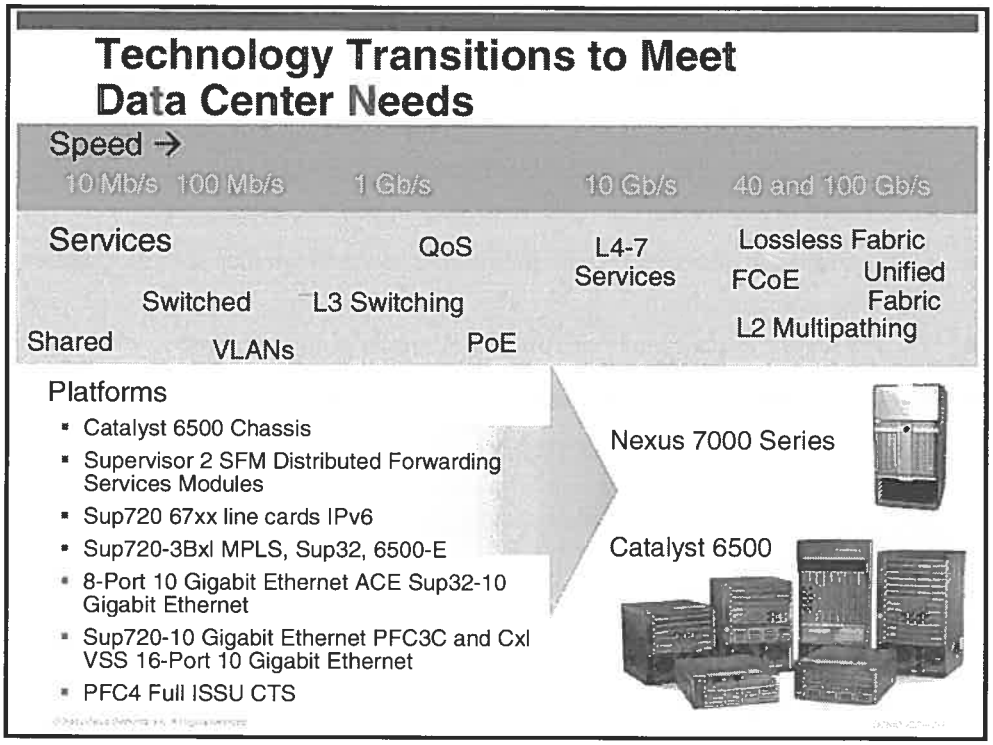


Data center infrastructure characteristics are the following:

- Cisco Nexus 7010 Switch delivers data center infrastructure seamless scalability
- Consolidation of data center infrastructure reduces expenses and enables segmentation and any-to-any access
- Cisco Nexus 7010 Switch enables transport flexibility
  - Converged networking
- Virtualization of resources for design flexibility and further data center automation
- Cisco Nexus 7010 Switch provides operational continual availability through a "Zero Service Loss" architecture
- Automation improves operations effectiveness and infrastructure availability

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The Cisco Nexus 7000 Series Switch positioning is a market transition from Catalyst 5000 to the Catalyst 6500 Series Switches where 1 Gb, 10 Gb, and Layer 4 through 7 services were integrated into the platform. There are going to be continued innovations on the Catalyst 6500 Series Switches for the next several years. The Nexus 7000 Series Switch is designed to support higher throughput with ports from 10 Gb, 40 and 100 Gb/s port speeds, providing multi-protocol support on the same chassis in a lossless fabric, Fibre Channel over Ethernet (FCoE) interfaces, the ability to actually scale the Layer 2 deployments using Layer 2 multipathing, and the consolidation, virtualization, and automation of the data center networks. Consolidation through scalability brings the 1 Gb/s, 10 Gb/s, 40 Gb/s, or 100 Gb/s port speeds. Virtualization is enabled through I/O consolidation which provides centralized storage management along with the ability to rapidly provision an I/O, whether it is Ethernet or FCoE, and then being able to virtualize both the server and the I/O. This virtualization is a significant value proposition. Cisco Nexus 7000 Series Switches support enhanced manageability of the platform such as XML APIs and NetFlow. Nexus 7000 Series Switches are supposed to solve many other modern data center problems such as:

- Power and cooling poised to outpace server capital expenditures (CapEx)
- Storage is expected to grow at a 40% – 70% CAGR
- Resource virtualization increases demands on the network
- Market transitions to 10 Gigabit Ethernet, 40 Gigabit Ethernet, and 100 Gigabit Ethernet
- New application architectures drive new network requirements
- Expectation of 24x7 application availability
- More than 50% of network downtime is caused by human error
- Ongoing regulatory requirements

© 2008 Cisco Systems, Inc.

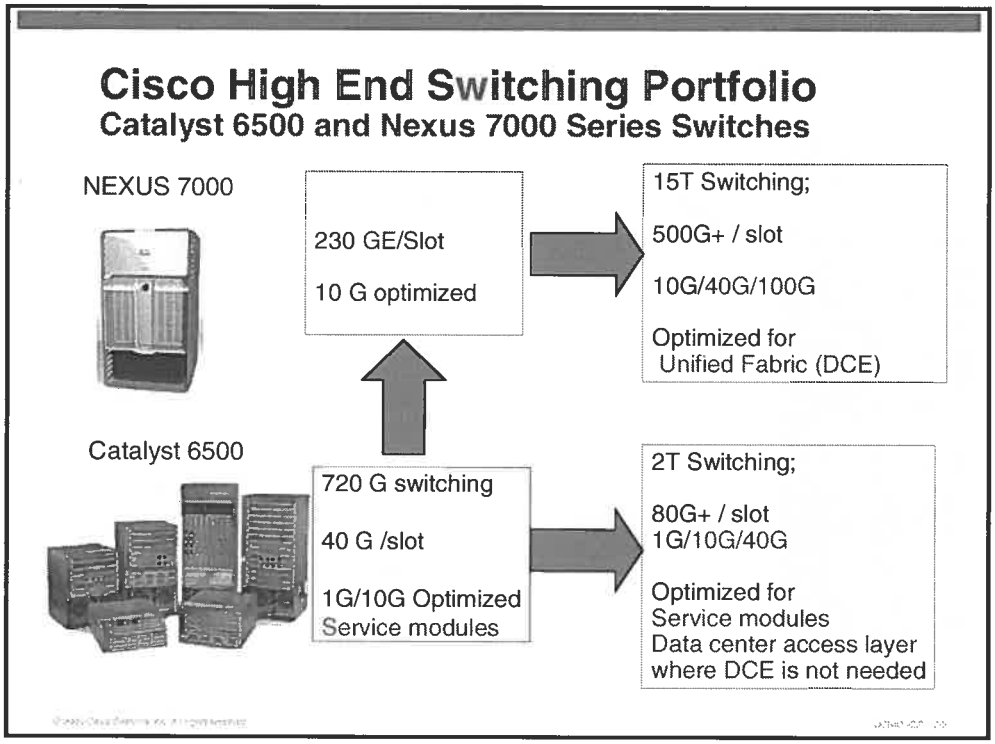
https://t.me/learningnets

The impact of the Nexus solution on the listed problems is at least the following:

- Density and capacity paired with virtualization allows utilization close to actual capacity—defers need to add infrastructure
- High-density/high-utilization deployment reduces power/port
- Unified fabric can drive 5% – 8% reduction in power consumption
- Unified fabric reduces CapEx by facilitating move to smaller servers with fewer host adaptors
- Zero Service Loss Architecture eliminates single point of failure
- Integrated management tools reduce time-to-resolution
- Graceful operations reduce "human error"

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

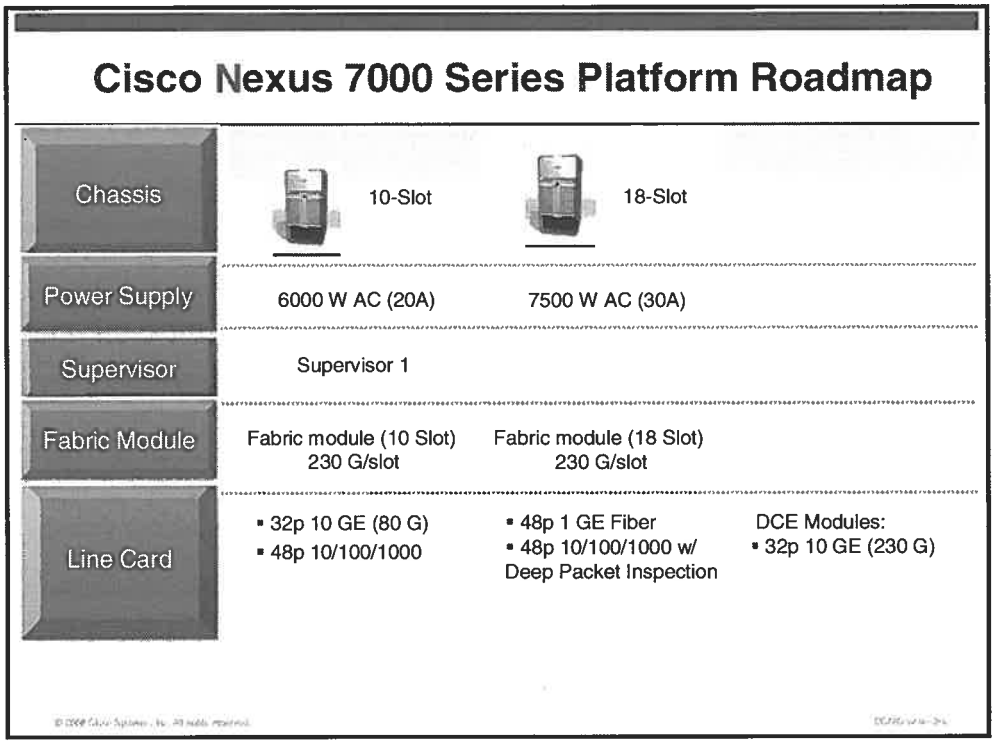


Catalyst 6500 and Nexus 7000 Series Switches

The Catalyst 65000 Series Switches are going to continue to focus on the campus network and also on the access layer of the data center infrastructure, where unified I/O is not necessarily a requirement. Cisco is planning the platform from the current 2 Tb/s aggregated platform which basically assumes the new 13-slot chassis and the 80 Gb per slot Supervisor in the 2010 time period. From a design perspective, the Catalyst 6500 Series Switch is optimized for 1 Gb/s and 10 Gb/s, whereas Nexus 7000 Series Switches are supposed to be more optimized for 10, 40 and 100 Gb/s connectivity. A Nexus 7000 series switch can be used to provide a 1 Gb/s connection, but it is an expensive 1 Gb/s connection that might be better served by the Catalyst 6500, depending on the connection requirements. Service modules are not supported by the Nexus 7010 Series Switch. This is the area in which the Catalyst 6500 Series Switches are continuing to invest. For consolidation and virtualization purposes, the Nexus switch family is going to focus increasingly on the converged I/O for the converged data center networks.

© 2008 Cisco Systems, Inc.

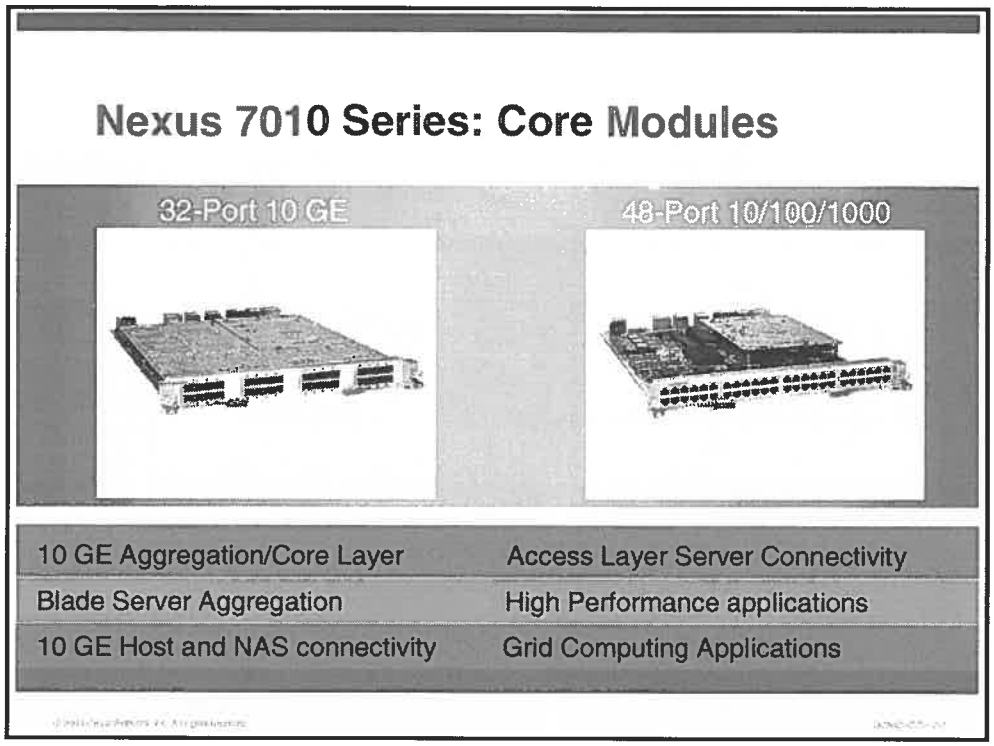
https://t.me/learningnets



Cisco has begun to ship the Nexus 7000 Series Switch 10-slot chassis. The fabric modules are different for the 10-slot and the 18-slot Nexus 7000 Series Switch because the 18-slot chassis has more slots to support. Therefore, the two chassis are not interchangeable. Each chassis has its own dedicated fabric modules, so the 10-slot ships with up to 5 fabric modules. The 18-slot chassis comes with its own set of spine cards. The 18-slot chassis will have a 1 Gigabit Ethernet fiber line card. There is a 48-port 10/100/1000 line card but with built-in deep packet inspection capability. The built-in packet inspection is bringing in a new internal technology which gives the ability to look at any offset inside the packet at the line rate and to take actions based on the packet inspection. The primary focus of the 18-slot chassis is in the data center, which is a 230Gb/s per slot switch with 15+ Terabits of aggregated bandwidth throughput.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



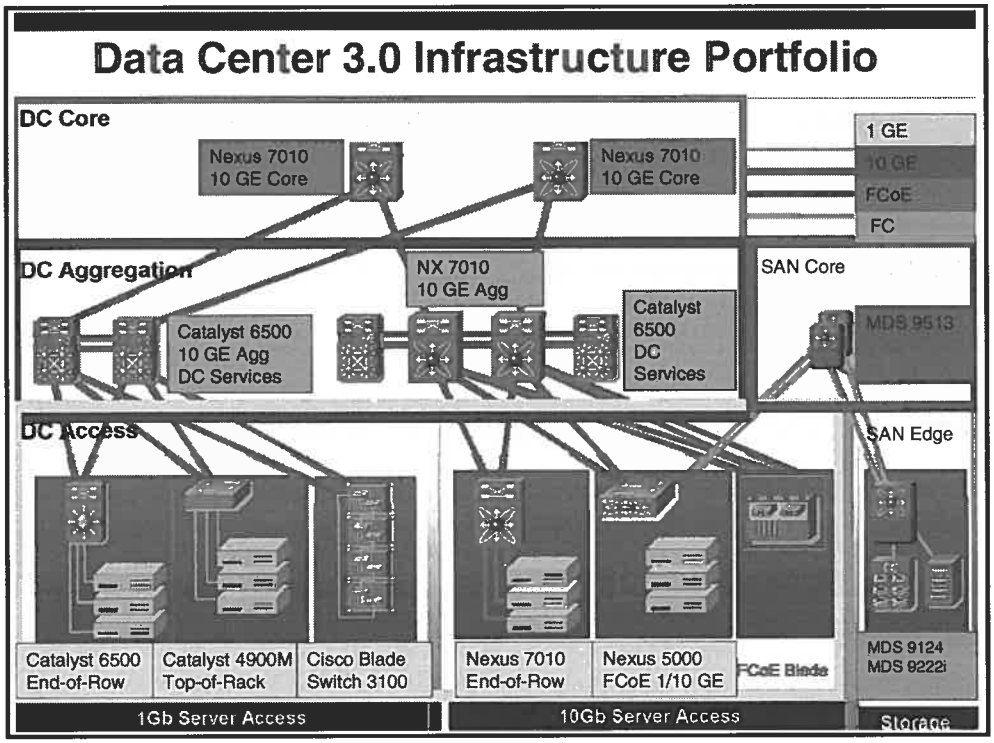
Cisco sells a 32-port 10 Gigabit Ethernet line card that has an 80 Gb/s switching fabric connection in the back. The port or line card over-subscription is 4:1. It fits in the aggregation or in the core data center layer where 10 Gb/s uplinks from access to the aggregation layer or from aggregation to the core layer are needed. Blade Servers can also utilize the 10 Gb/s uplinks to the 32-port line card.

Cisco also sells also a 48 port 10/100/1000 line card with a 40 Gb/s switching fabric in the back. All line cards in Nexus 7000 have a distributed forwarding engine built into the line card. These are server ports providing a 1 Gb/s connection to the access layer. They can be used for HPC as well as for grid computing application server uplinks to the access layer.

Access Layer 1 Gb/s or 10 Gb/s server connectivity can be provided by Catalyst 6500 line cards. For example, a Blade Server that requires 10 Gb/s can be attached to the 10 Gb/s Catalyst 65000 Switch line card ports, as well as to the Nexus 7010 Switch I/O module 10 Gb/s port. Catalyst has all the access-layer features needed to provide server connectivity to the data center fabric.

© 2008 Cisco Systems, Inc.

https://t.me/learningnets



The 1 Gb/s server connectivity to the access layer data center is provided by the Catalyst 6500 Switch, with the 6748 line cards and the Sup720 with the 10 Gb uplinks coming to the data center aggregation layer. The 48-port 1 Gb/s line card can be used for the 1 Gb/s servers, where the 10Gb/s uplinks to the aggregation layer use the 32-port 10 Gb/s line card. This is a more expensive solution. The Catalyst Switch provides all services needed at the access layer along with service modules, whereas the Nexus 7000 and Nexus 5000 Series Switch provides higher performance and unified I/O access. On average, a server uses 20MBps application throughput which can be provided by a 1 Gb/s Ethernet link offering 100MBps throughput. The most demanding data center servers such as SQL database servers, Tuxedo servers or Video servers can saturate a 1 Gpbs link, so 100 MBps application throughput will be required and provided by the 1 Gb/s Ethernet connections or 10 Gb/s unified I/O link from Nexus 5000 Switch. 10 Gb/s unified I/O channel has 4 Gb/s dedicated to FC using a built-in rate limiter and the rest 6 Gb/s of bandwidth is available to the Ethernet traffic. While FC traffic is not using the available 4 Gb/s bandwidth of the unified I/O channel the Ethernet traffic can use that bandwidth.

For Server top-of-rack designs, the 4900M Series Switches can be used with 10/100/1000 Gb/s ports for or 10 Gb/s 1:1 or 2:1 oversubscribed ports for more demanding servers or for the uplinks to the aggregation layer. The Catalyst 4948 Series Switch with 1 Gb/s 1 RU is also appropriate for connecting 1 Gb/s servers to the fabric. Catalyst 3700 and 2900 Series Switches are also available, but the primary platform of focus is the Cisco 4900 Series Switch.

For blade servers, Cisco Blade Switches are appropriate. Cisco Blade Switches with the 10Gb uplink, for all the three major blade server vendors—IBM, HP and Dell—are available. The 10Gb density requirements depends on how many of these you are collapsing, or how many of these you are connecting to a distribution box. In most cases, the Catalyst 65000 located in the distribution layer will meet the 10 Gb/s uplinks requirements from the access layer to the aggregation layer. When the aggregation layer interconnects the heterogeneous access layer or unified I/O access layer, many 10 Gb/s uplinks to the distribution layer are needed and can be provided by the Nexus 7010 Series Switch 10 Gb/s I/O module populated in the aggregation network. The Nexus 7010 Series Switch in the distribution layer is used mainly because of the 10Gb density and the scalability that it brings to the distribution layer. There are no services

© 2008 Cisco Systems, Inc.

https://t.me/learningnets

modules available in the Nexus 7010 Series Switch, hence, an external Catalyst Switch is used for the service module integration. The Access layer would consist of the top-of-rack servers with unified I/O capabilities and the blade Switches with unified I/O. For a very demanding 10 Gb/s access layer, the Nexus is an appropriate platform in the distribution layer. When unified I/O is not needed, the new 4900M is available, which provides the top-of-rack capability with 24 10 Gb/s ports. The Fibre Channel over Ethernet links are connecting the Nexus 5000 with the servers in the access layer or Fibre Channel links are connecting the top of the rack Nexus 5000 Switch and the blade servers as shown on the figure. In the future, when the MDS switches might have FCoE ports, the FCoE connections will be extended closer to the storage arrays. Because of application-to-application latency—about 50 to 60 microseconds, the Ethernet stack might be too slow for storage traffic. Therefore, at least until the Ethernet stack latencies comes down significantly, using the Fibre Channel protocol will continue to be interesting. So far the Nexus 7010 Series Switch does not have service modules slots due to the bandwidth per slot available at the Nexus 7010 Series Switch. Instead, linkages between the Nexus and an external appliance or a node are planed to be created so that the services would be accelerated. The design direction is to accelerate the data plane in Nexus 7010 Series Switch and to create linkages to the Catalyst 6500 Series Switch for what cannot be accelerated in the data plane. The Nexus 5000 Series Switch is the top-of-rack switch with 20 or 40 10 Gb/s ports. The servers are attached to the Nexus switch using the converged adapter that transmits Fibre Channel and Ethernet protocol onto the 10Gb Ethernet unified I/O channel. The server has a special Ethernet NIC capable of supporting Fibre Channel over Ethernet and muxing the Fibre Channel on the Ethernet inside. This adapter supports Ethernet protocol with the SCSI stack inside it. The unified I/O is coming from the server into the top-of-rack switch. Continuing with the unified I/O into the aggregation cloud, or at the access layer, the unified I/O channel is broke up into the Ethernet and Fibre Channel channels.

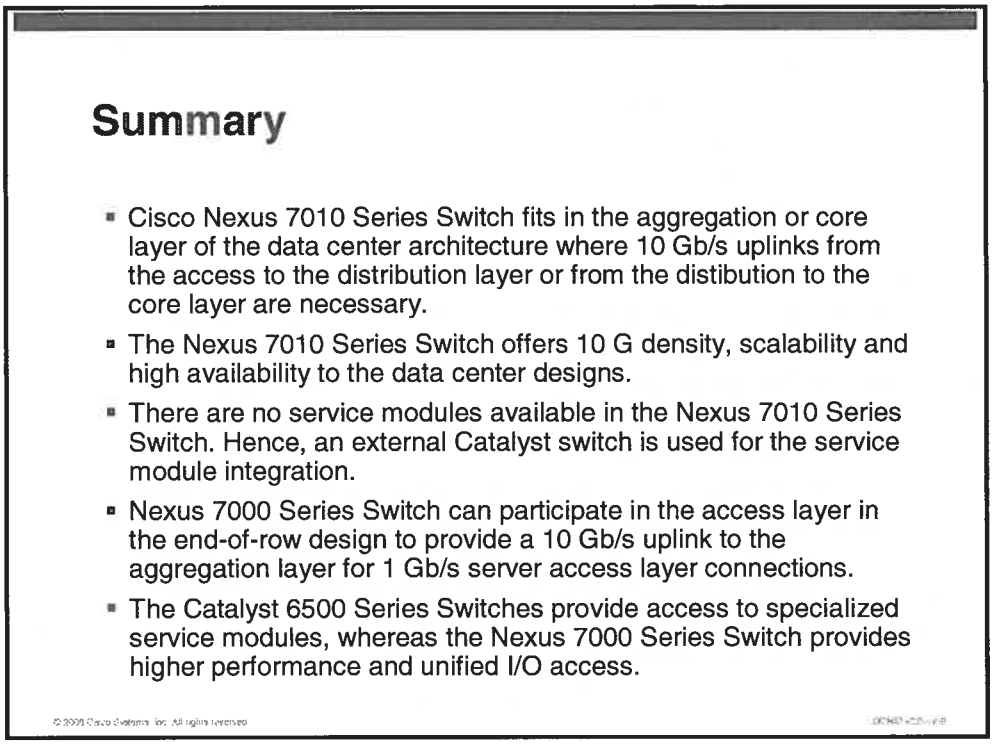
The data center core layer is supported by the Nexus 7000 Series Switch, since it provides the highest performance and the high availability to the data center core along with the support for unified I/O.

© 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