1

Borderless Campus Design and Deployment Models

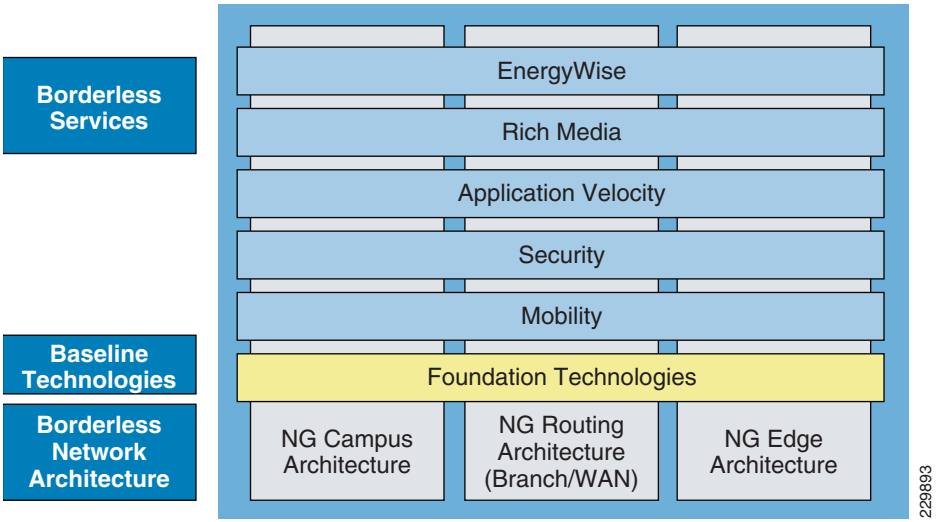
1 Executive Summary

Enterprises are making a fundamental shift, with employees no longer confined to physical offices, geographical locations, and time zones. In today's globalized workplace, work can occur anywhere in the world and enterprise information needs to be virtually accessible, on-demand, through every part of the network. These requirements drive the need to build next-generation networks that are secure, reliable, and highly available. Adding to the impetus for next generation networks is the transformation of the network from traditional data and voice transport to super-highways for video, building systems, physical security systems, and other non-traditional systems that now use IP networks to communicate. As more and more systems converge on the network, network complexity increases and the capability to handle traffic in a responsible manner is essential as the network becomes even more mission critical.

This network transformation presents many challenges to enterprise IT staff. They must be able to transform their networks to allow secured network access from anywhere, but enforce security policies based on how users are accessing the network. They must allow multiple systems or services to simultaneously traverse the network, while allocating sufficient network resources to ensure those systems do not negatively impact each other. The network must be agile to adapt to future requirements without requiring an overhaul of the existing network. And of course the scalable, resilient, highly available, service differentiating, adaptable network they create must be cost effective and protect investments in the network.

The Cisco Borderless Network architecture is designed to directly address these IT and business challenges. It offers a seamless user experience with a next-generation network that allows different elements of the network, from access switches to wireless access points, to work together and allow users to access resources from anyplace at anytime. The Cisco Borderless Network uses a tiered approach to virtually collapse the network as a single borderless network. It also integrates key services into the network fabric while increasing reliability and security and decreasing service time. For such an infrastructure, the enterprise network must be developed with an architectural approach that embeds intelligence, simplifies operations, and is scalable to meet future demands. The Cisco Borderless Network is a next-generation network architecture that combines several innovations and architectural design considerations to offer a new workspace experience. The Cisco Borderless Network is composed of several modular components, as illustrated in Figure 1.

Figure 1 Cisco Borderless Network Framework



Each building block in the Cisco Borderless Network framework is designed to offer the following components:

- Network Infrastructure—Builds enterprise campus, WAN, and edge networks as an open platform that can provide secure and intelligent services at the network edge, aggregation scalability, and a high-performance backbone solution to enable end-to-end borderless services and applications.
- Foundation Technologies—Common baseline technologies that are integrated across various enterprise architectures to optimize service delivery, intelligently differentiate between various applications, and build the highly-available network infrastructure.
- Borderless Services—Enables the end-to-end borderless user experience to provide ubiquitous connectivity with security, reliability, and sustainability to the enterprise workspace users and the network edge elements. Empowers network architects to leverage the network as a platform to offer rich services to reduce business operational costs, increase efficiency through green practices, and much more.

2 Borderless Campus Network Design

The Borderless Campus Network architecture is a multi-campus design, where a campus consists of multiple physical buildings with a wide range of network services that offer the capability for anyone to securely access network resources from anywhere at anytime, as shown in Figure 2.

Figure 2 Borderless Campus Network Design

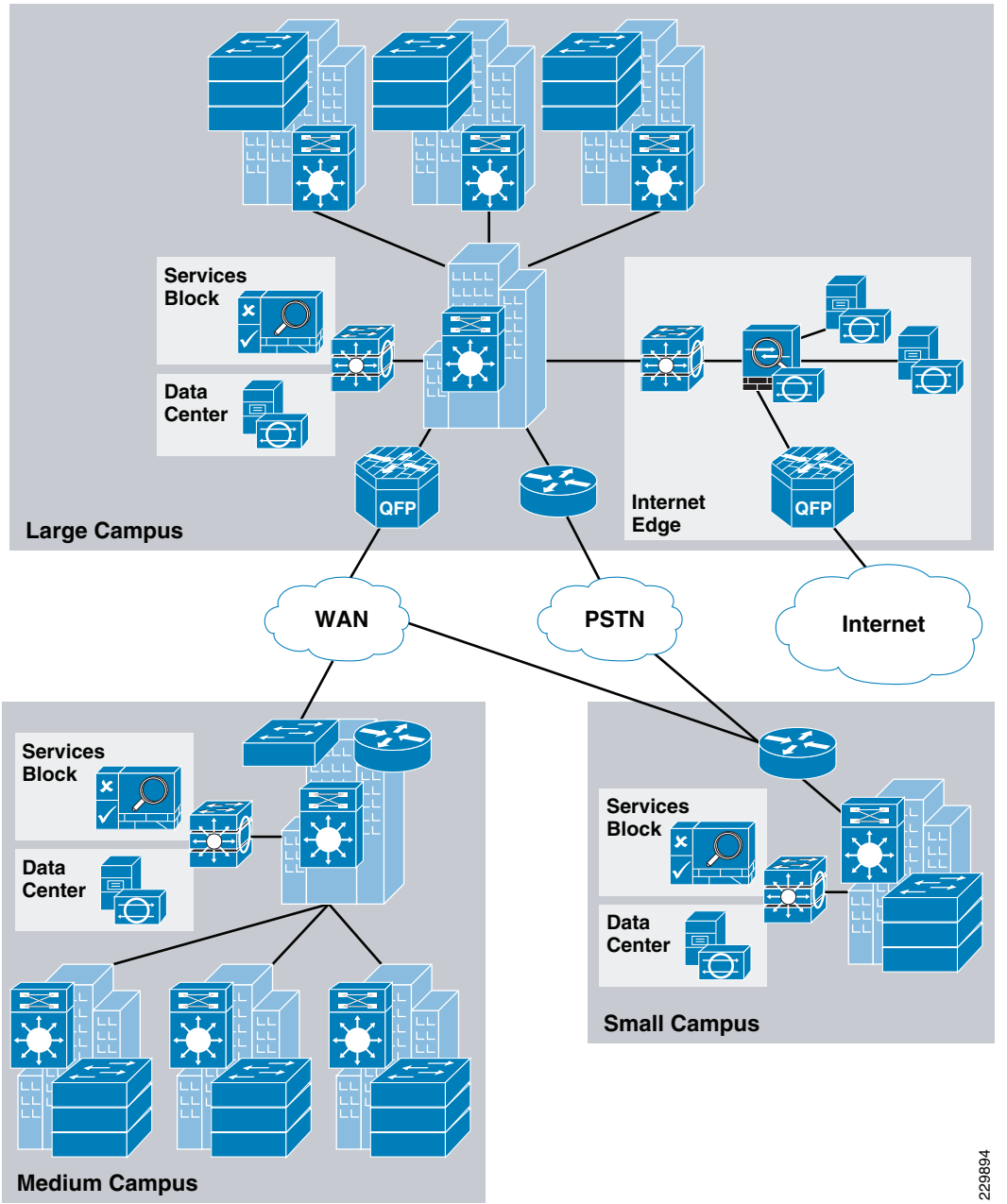
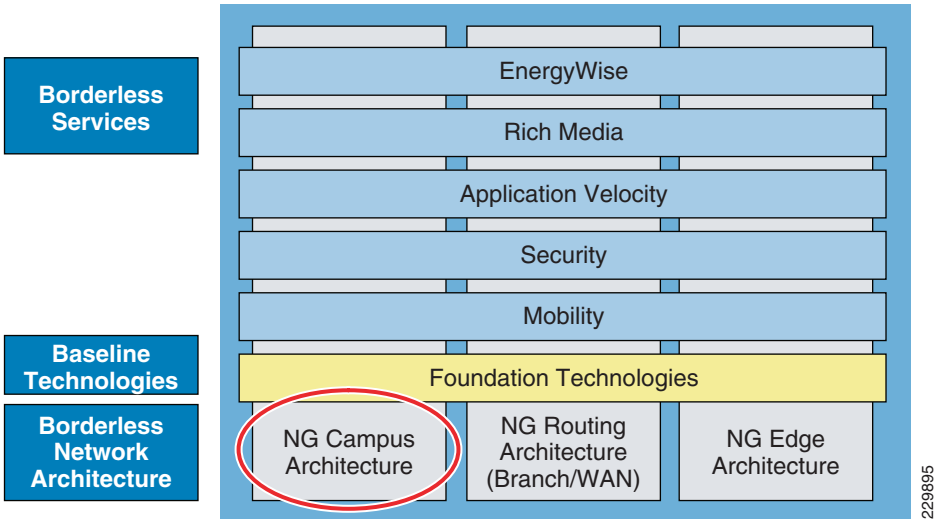

Services Block
Data Center
QFP
Large Campus
Internet Edge
QFP
WAN
PSTN
Internet
Services Block
Data Center
Medium Campus
Services Block
Data Center
Small Campus
229894

Figure 3 shows the service fabric design model used in the Borderless Campus.

Figure 3 Borderless Campus Architecture



This document describes the campus framework and network foundation technologies that provide a baseline of routing, switching, and several key network services guidelines. The campus design interconnects several other infrastructure components, such as endpoints at the network edge, data center, WAN, and so on, to provide a foundation on which mobility, security, video, and unified communications (UC) can be integrated into the overall design.

This campus design provides guidance on building the next-generation enterprise network, which becomes a common framework along with critical network technologies to deliver the foundation for the service fabric design. This chapter is divided into the following sections:

- *Campus design principles*—Provides proven network design choices to build various types of campus infrastructure.
- *Campus design model for the enterprise*—Leverages the design principles of the tiered network design to facilitate a geographically-dispersed enterprise campus network made up of various elements, including networking role, size, capacity, and infrastructure demands.
- *Considerations of a multi-tier campus design model for enterprises*—Provides guidance for the enterprise campus LAN network as a platform with a wide range of next-generation products and technologies to seamlessly integrate applications and solutions.
- *Designing network foundation services for campus designs in the enterprise*—Provides guidance on deploying various types of Cisco IOS technologies to build a simplified and highly-available network design to provide continuous network operation. This section also provides guidance on

designing network-differentiated services that can be used to customize the allocation of network resources to improve user experience and application performance and to protect the network against unmanaged devices and applications.

3 Borderless Campus Network Design Principles

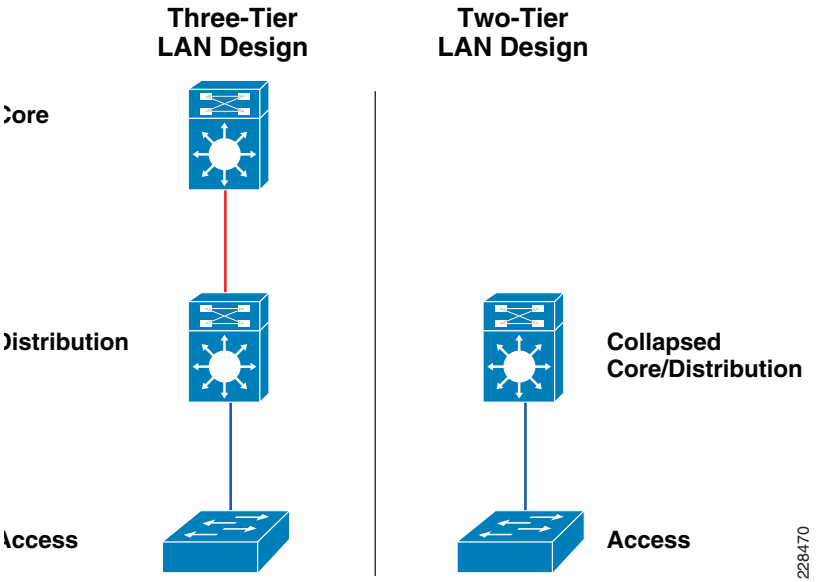
Designing the borderless campus requires that sound network design principles are used to ensure maximum availability, flexibility, security, and manageability. The use of sound network design ensures that the network will deliver on current requirements as well as be well prepared for future services and technologies. This document provides design guidelines that are built upon the following principles to allow the enterprise network architect to build a geographically-dispersed borderless network:

- *Hierarchical*
  - Facilitates understanding the role of each device at every tier
  - Simplifies deployment, operation, and management
  - Reduces fault domains at every tier
- *Modularity*—Allows seamless network expansion and integrated service enablement on an on-demand basis
- *Resiliency*—Satisfies user expectations for keeping the network always on
- *Flexibility*—Allows intelligent traffic load sharing by using all network resources

These are not independent principles. The successful design and implementation of a campus network requires an understanding of how each of these principles applies to the overall design. In addition, understanding how each principle fits in the context of the others is critical in delivering the hierarchical, modular, resilient, and flexible networks required by enterprises.

Designing the Borderless Campus network in a hierarchical fashion creates a flexible and resilient network foundation that allows network architects to overlay the security, mobility, and unified communication features essential to the service fabric design model. The two proven, time-tested hierarchical design frameworks for campus networks are the three-tier layer and the two-tier layer models, as shown in Figure 4.

Figure 4 Three-Tier and Two-Tier Campus Design Models



The key layers are access, distribution, and core. Each layer can be seen as a well-defined structured module with specific roles and functions in the campus network. Introducing modularity into the campus hierarchical design further ensures that the campus network remains resilient and flexible to provide critical network services as well as to allow for growth and changes that may occur over time.

- *Access layer*

  The access layer represents the network edge, where traffic enters or exits the campus network. Traditionally, the primary function of an access layer switch is to provide network access to the user. Access layer switches connect to distribution layer switches, which perform network foundation technologies such as routing, quality of service (QoS), and security.

  To meet network application and end-user demand, the next-generation Cisco Catalyst switching platforms no longer simply switch packets, but now provide more converged, integrated, and intelligent services to various types of endpoints at the network edge. Building intelligence into access layer switches allows applications to operate on the network more efficiently, optimally, and securely.

- *Distribution layer*

  The distribution layer interfaces between the access layer and the core layer to provide many key functions, including:

  - Aggregating large-scale wiring closet networks
  - Aggregating Layer 2 broadcast domains and Layer 3 routing boundaries

  - Providing intelligent switching, routing, and network access policy functions to access the rest of the network
  - Providing high availability through redundant distribution layer switches to the end-user and equal cost paths to the core, as well as providing differentiated services to various classes of service applications at the edge of network

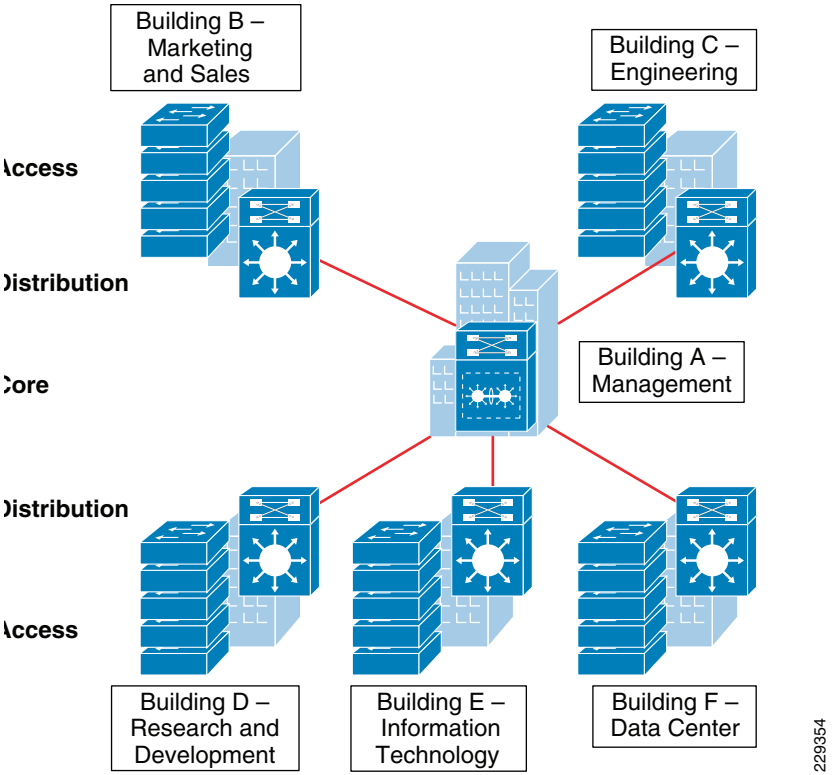
- *Core layer*

  The core layer is the network backbone that hierarchically connects several layers of the campus design, providing for connectivity between end devices, computing and data storage services located within the data center and other areas, and services within the network. The core layer serves as the aggregator for all o fthe other campus blocks and ties the campus together with the rest of the network.

Note For more information on each of these layers, see the enterprise class network framework at: http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/campover.html.

Figure 5 shows a sample three-tier campus network design for enterprises where the access, distribution, and core are all separate layers. To build a simplified, scalable, cost-effective, and efficient physical cable layout design, Cisco recommends building an extended-star physical network topology from a centralized building location to all other buildings on the same campus.

Figure 5 Three-Tier Campus Network Design Example

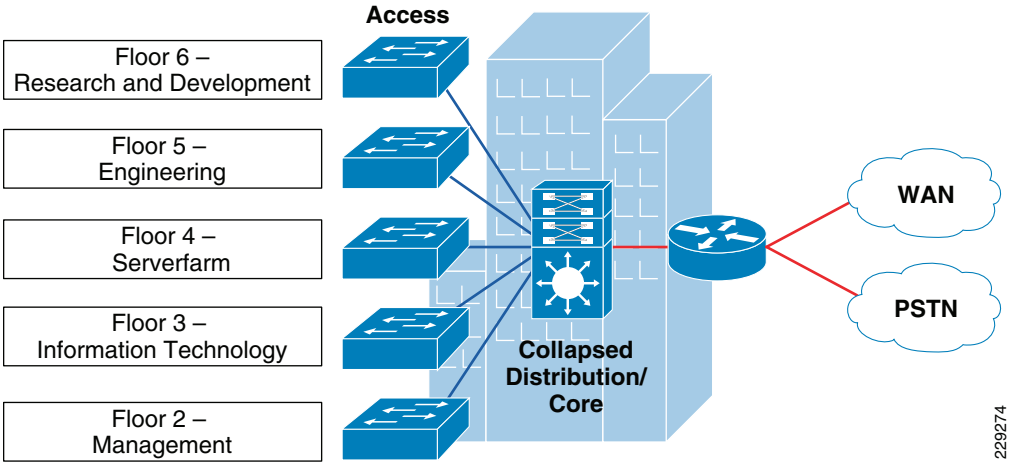


The primary purpose of the core layer is to provide fault isolation and high-speed backbone connectivity with several key foundational services. Isolating the distribution and core into separate layers creates a clean delineation for change control activities affecting affecting end devices (laptops, phones, and printers) and those that affect the data center, WAN, or other parts of the campus network. A core layer also provides for flexibility in adapting the campus design to meet physical cabling and geographical challenges. If necessary, a separate core layer can use a different transport technology, routing protocols, or switching hardware than the rest of the campus, providing for more flexible design options when needed.

In some cases, because of either physical or network scalability, having separate distribution and core layers is not required. In smaller campus locations where there are fewer users accessing the network or in campus sites consisting of a single building, separate core and distribution layers may not be needed. In this scenario, Cisco recommends the alternate two-tier campus network design, also known as the collapsed core network design.

Figure 6 shows a two-tier campus network design example for an enterprise campus where the distribution and core layers are collapsed into a single layer.

Figure 6 Two-Tier Network Design Example



If the small-scale collapsed campus core design is used, the enterprise network architect must understand network and application demands so that this design ensures a hierarchical, modular, resilient, and flexible campus network.

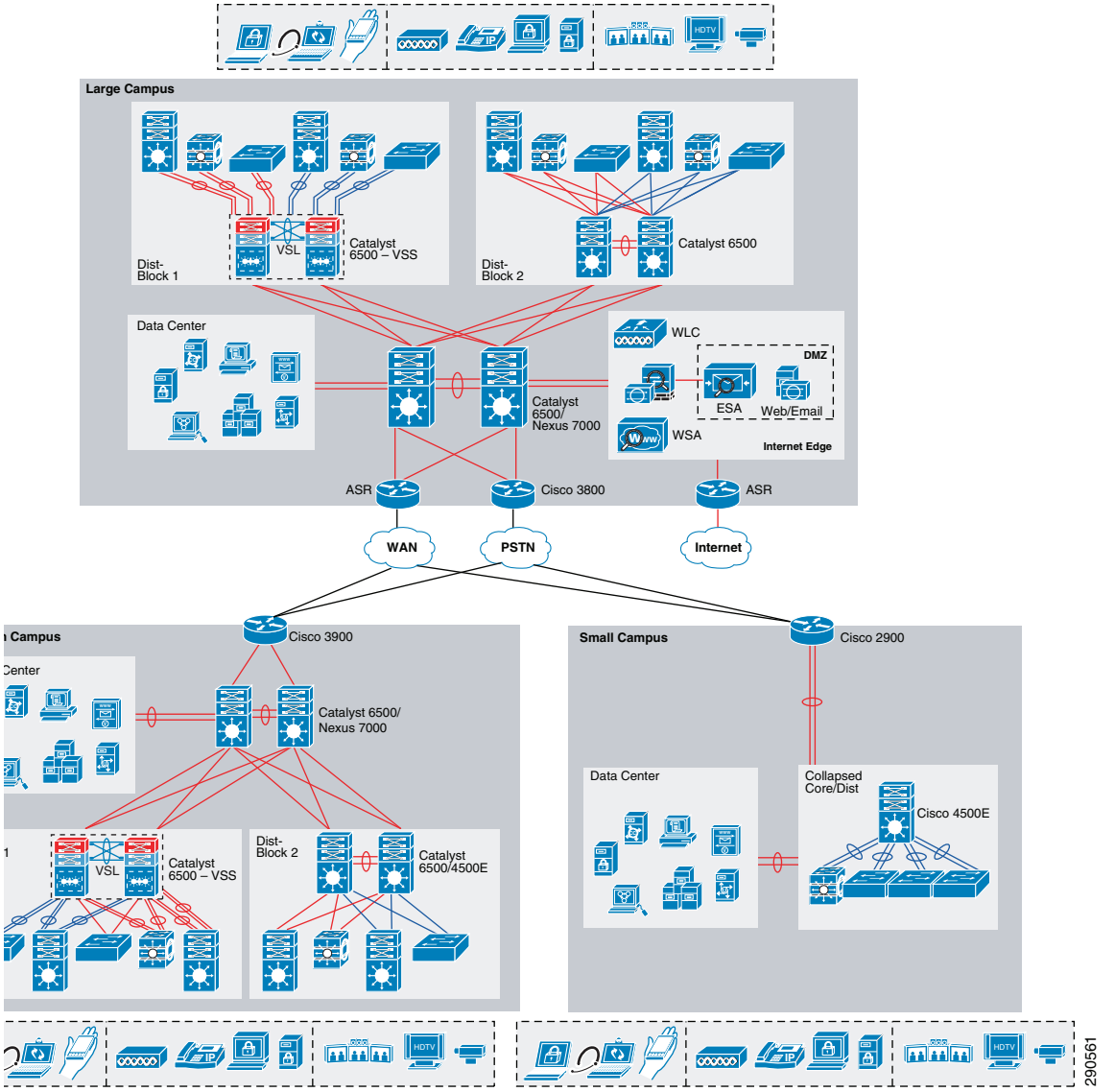
4 Borderless Campus Network Design Models

Both campus design models (three-tier and two-tier) have been developed with the following considerations:

- *Scalability*—Allowing for network speeds from 100mb to 10gb, the ability to scale a network based on required bandwidth is paramount. The network provides investment protection by allowing for upgradability as bandwidth demand increases.
- *Simplicity*—Reducing operational and troubleshooting cost by the use of network-wide configuration, operation, and management.
- *Resiliency*—Ability to provide non-stop business communication with rapid sub-second network recovery during abnormal network failures or even network upgrades.
- *Cost-effectiveness*—Integrated specific network components that fit budgets without compromising design principles and network performance.

As shown in Figure 7, multiple campuses can co-exist within a single enterprise system that offers borderless network services.

Figure 7 Borderless Campus Network Design Model



Depending on the medium and small campus office facility, the number of employees and the networked devices in remote campuses may be equal to or less than the large campus. Hence compared to the large campus network, the medium and small campus sites may have alternate network designs that can provide network services based on overall campus network capacity.

Using high-speed WAN technology, several medium and small enterprise campuses can interconnect to a centralized large campus that provides protected shared data and network services to all employees independent of their physical location.

Table 1 shows a summary of the Borderless Campus Network design models as they are applied in different overall enterprise network designs.

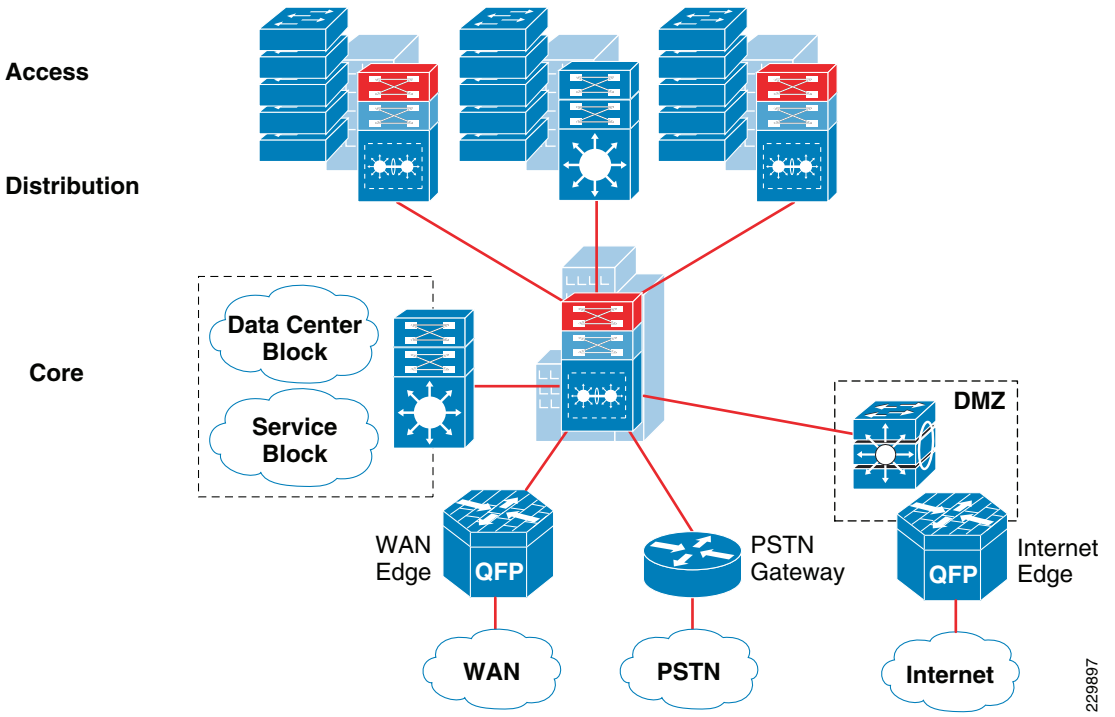
Table 1 Enterprise Recommended Campus Design Models

Enterprise Location	Recommended Campus Design Model
Large campus	Three-tier
Medium campus	Three-tier
Small campus	Two-tier

Large Campus Network Design

The large campus in the enterprise design consists of a centralized hub campus location that interconnects medium and small campuses of several sizes to provide end-to-end shared network access of resources and borderless services. The large campus typically consists of various sizes of building facilities and various organizational and departmental groups. The network scale in the large campus is higher than the medium and small campus networks and includes end users, IP-enabled endpoints, servers, security, and network edge devices. Multiple buildings of various sizes exist in one location, as shown in Figure 8.

Figure 8 Large Campus Reference Design



The three-tier campus design model for the large campus meets all key technical aspects to provide a well-structured and strong network foundation. The modularity and flexibility in a three-tier campus design model allows easier expansion and integration in the large campus network and keeps all network elements protected and available.

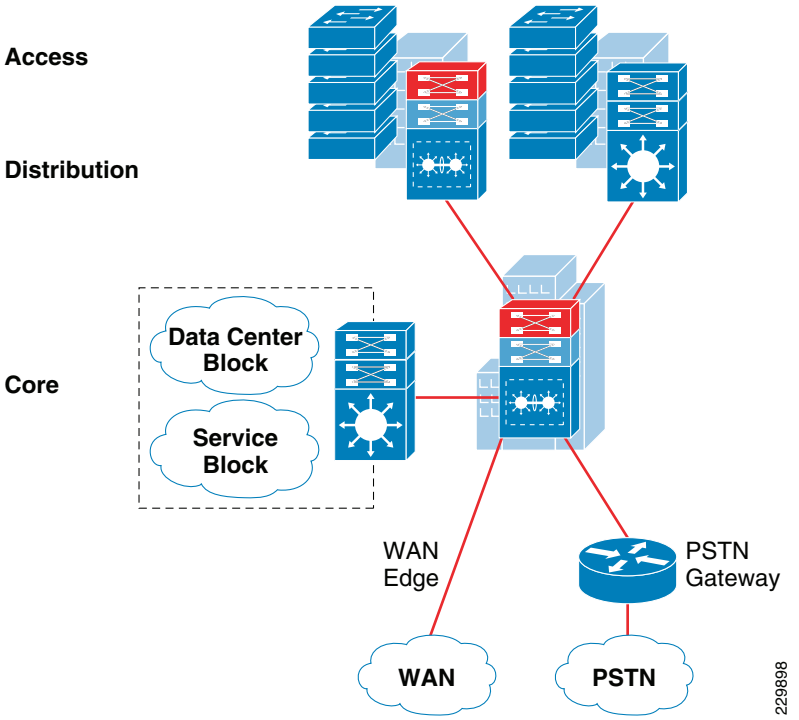
To enforce external network access policies for end users, the three-tier model in the large campus also provides external gateway services to employees for accessing the Internet.

Medium Campus Network Design

From a location, size, and network scale perspective, the medium campus is not much different than the large campus. Geographically, it can be distant from the large campus and require a high-speed WAN circuit to interconnect both campuses. The medium campus can also be considered as an alternate campus to the large campus, with the same common types of applications, endpoints, users, and network services. Similar to the large campus, separate WAN devices are recommended to provide application delivery and access to the large campus given the size and number of employees at this location.

Similar to the large campus network design, Cisco recommends the three-tier campus design model for the medium campus, as shown in Figure 9.

Figure 9 Medium Campus Reference Design



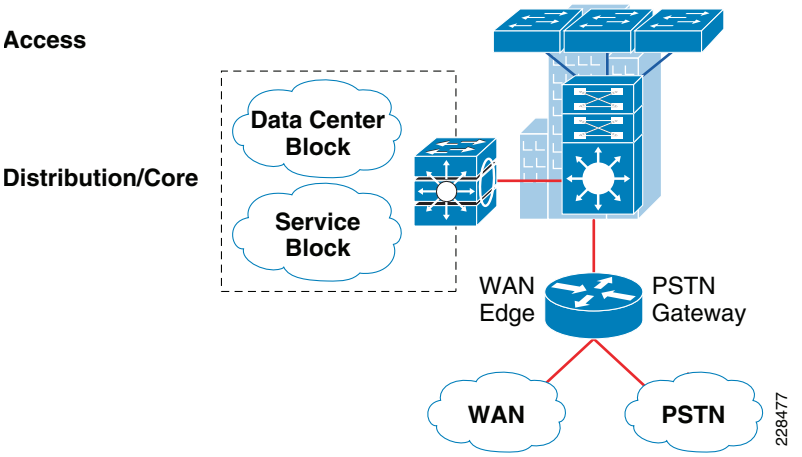
Small Campus Network Design

The small campus is typically confined to a single building that spans multiple floors with different organizations. The network scale factor in this design is reduced compared to other large and medium campuses. However, application and borderless services demands are still consistent across the enterprise geographical locations.

In such smaller scale campus network deployments, the distribution and core layer functions can be alternatively collapsed into the two-tier campus model without compromising basic network design principles. Prior to deploying the collapsed core and distribution system, network architects must consider the scale, expansion, and manageability factors which may reduce overall operational efficiency.

WAN bandwidth requirements must be assessed appropriately for the remote small campus network design. Although the network scale factor is reduced compared to other larger campus locations, sufficient WAN link capacity is needed to deliver consistent network services to users. A single Cisco platform in a highly-redundant configuration mode can provide collapsed core and distribution LAN layers. This alternate and cost-effective network design model is recommended only in smaller locations; WAN traffic and application needs must be considered. Figure 10 shows the small campus network design in more detail.

Figure 10 Small Campus Reference Design



5 Multi-Tier Borderless Campus Design Models

The previous section discussed various recommended campus design models for each enterprise location. This section provides more detailed network infrastructure guidance for each tier in the campus design model. Each design recommendation is optimized to keep the network simplified and cost-effective without compromising network scalability, security, and resiliency. Each campus design model for an enterprise location is based on the three parts of the campus network architecture—core, distribution, and access layers.

Campus Core Layer Network Design

As described in the previous section, the core layer is the center point of the network and becomes a high-speed transit point between multiple distribution blocks and other systems that interconnect to the services block, the WAN, and the campus edge. The common design in large networks is to build a high-performance, scalable, reliable, and simplified core.

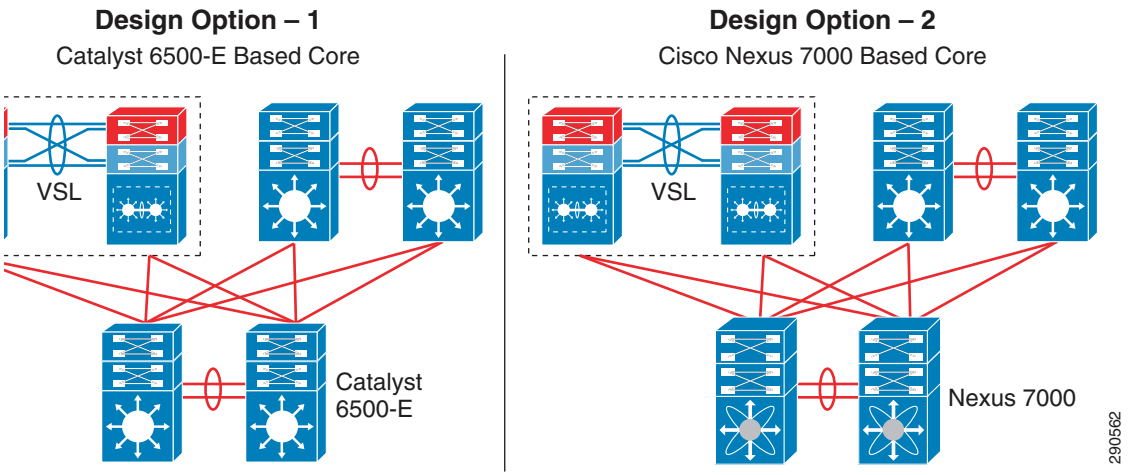
When network architects are designing a campus core, it becomes imperative to take into consideration network scalability, capacity, and reliability to allow for high-performance, end-to-end borderless services. Quantifying the core layer scalability and performance may be challenging as it varies depending on the needs of the enterprise. In campus core design, large enterprise networks are largely built with highly-resilient systems and high-speed 10Gbps links. Network architects must proactively foresee the expansion, evolution, and advancement of devices and applications on the network that may impact the core.

Cisco recommends building the next-generation borderless campus core with the following principles. The architecture should be:

- Designed to support modern technologies that enable advanced networking and integrated services to solve key business problems.
- Scalable to adapt to enterprise network needs and able to provide intelligent borderless network services.
- Flexible, with design options that maximize return on investment (ROI) and reduce total cost of ownership (TCO).

These design principles are important when designing the core network so that the core is capable of addressing current and future borderless network demands. Cisco recommends the Cisco Catalyst 6500-E and Nexus 7000 switching platforms for the core of the next generation borderless campus. These multi-terabit switching platforms are designed with a robust hardware architecture that exceeds the foundational borderless campus requirements. Figure 1-11 illustrates core designs for building the next-generation Borderless Campus core.

Figure 1-11 Core Layer Design Model Options



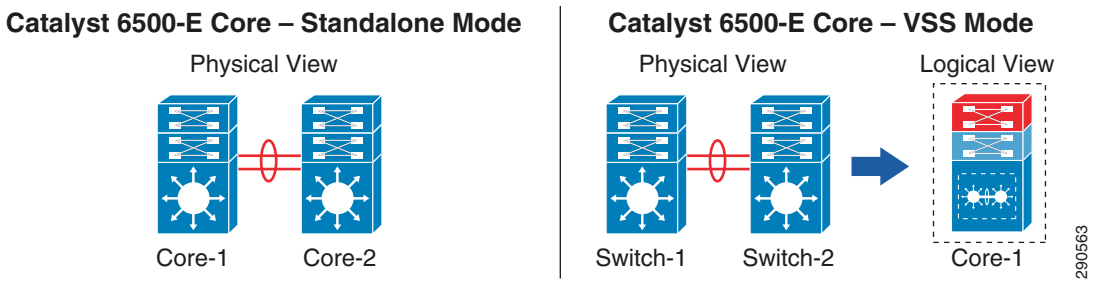
Cisco Catalyst 6500-E

The industry-leading, widely-deployed Cisco Catalyst 6500-E series platform has advanced hardware and software innovations that make it the preferred system to build an enterprise-class borderless campus core network. Cisco Catalyst 6500-E switches have a flexible architecture that enables a rich set of features and advanced technologies, along with the high-speed interfaces needed for the borderless campus. In the large and medium campuses, bandwidth intensive and latency sensitive applications—such as real-time IP-based voice and video—are ubiquitous, so network architects must take this into consideration when selecting the appropriate core platform. As networks expand, the management and troubleshooting of the infrastructure increases, however administrators can leverage Cisco's system virtualization technology to ease those burdens.

To provide mission-critical network services, it is recommended that the core layer be highly resilient. Deploying resilient, dual Cisco Catalyst 6500-E systems provides constant network availability for business operations during faults and also provides the ability to load share high-speed network traffic between different blocks (e.g., the distribution and service blocks). A redundant core network design can be deployed in a traditional standalone model or in a Virtual Switching System (VSS) model. The campus core layer network design and operation broadly differ when the core layer is deployed as a standalone, which operates all three planes (forwarding, control and data planes) in isolation. However with Cisco VSS technology, two core systems are clustered into a single logical system and the control and management planes get combined on the systems to produce a single logical Catalyst 6500-E core system.

The Standalone/VSS Physical and Operational View is shown in Figure 12.

Figure 12 Standalone/VSS Physical and Operational View



Note For more detailed VSS design guidance, see the *Campus 3.0 Virtual Switching System Design Guide*: http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/VSS30dg/campusVSS_DG.html.

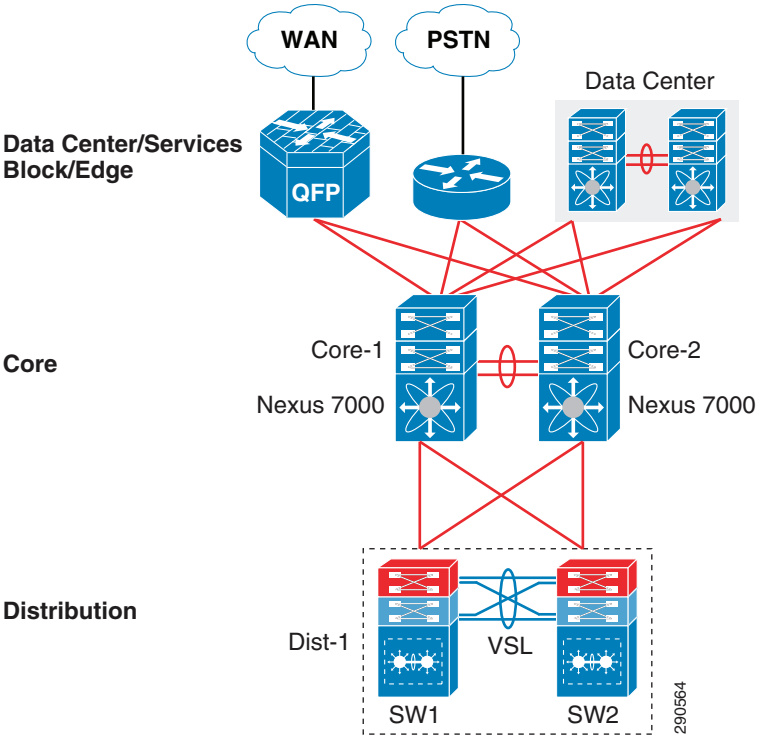
Cisco Nexus 7000

In high-speed and dense networking environments, enterprises require a simplified network architecture that expands the infrastructure's scalability, performance, and reliability. With this in mind, Cisco developed a powerful, multi-terabit switching platform, the Cisco Nexus 7000, to deliver these fundamental requirements. Next-generation data center architectures are built on the Cisco Nexus product family and the Cisco Nexus 7000 series platform leads in data center aggregation and in the data center core networking role.

Because of its unique architecture, technical advantages, and ability to deliver a baseline of campus core requirements, the Cisco Nexus 7000 series can be an alternative platform for deployment in the campus core. In the campus core environment, the Cisco Nexus 7000 offers un-paralleled 10G density to aggregate distribution blocks. It enables low-latency and wire-speed backbone connectivity between the service block and campus edge. The Nexus 7000 utilizes Cisco NX-OS as its operating system, which is a highly-evolved, multithreaded, and modular operating system to deliver core class networking services. NX-OS offers resilient network communication, system virtualization, and several other technical innovations that enable enterprises to have the capabilities needed for the next-generation Borderless Campus network. The Nexus 7000 platform operates in a standalone configuration that locally maintains the control, distributed forwarding, and management planes. For a resilient and mission critical campus core design, the Cisco Nexus 7000 system should be deployed with redundant hardware components that maintain backbone switching capacity and service availability during planned upgrades or un-planned network outages.

Figure 13 illustrates core network design options with the Cisco Nexus 7000 peering with other Cisco platforms to enable end-to-end business communication:

Figure 13 Cisco Nexus 7000 Campus Core Design



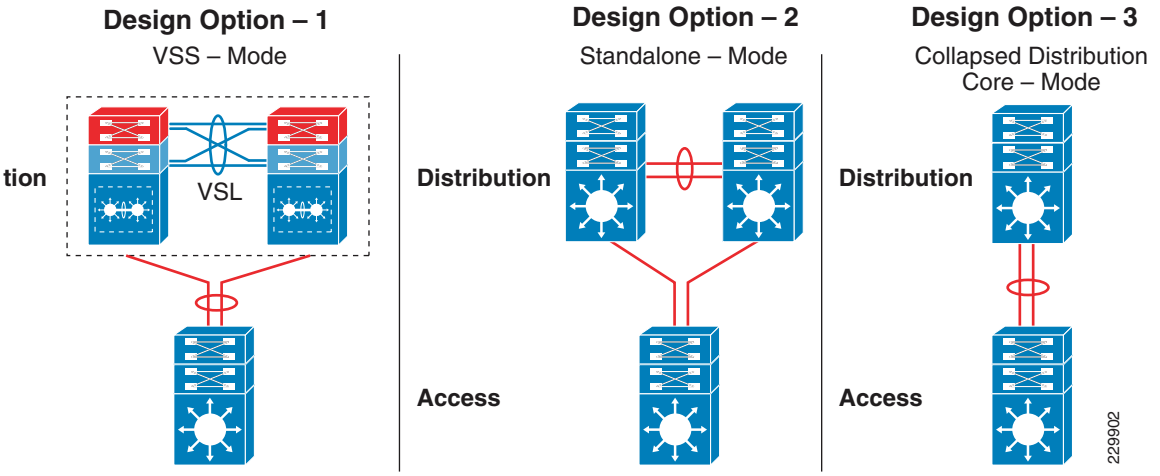
Campus Distribution Layer Network Design

The distribution or aggregation layer is the network demarcation boundary between wiring closet switches and the campus core network. The framework of the distribution layer system in the enterprise design is based on best practices that reduce network complexities, increase reliability, and accelerate network performance. To build a strong campus network foundation with the three-tier model, the distribution layer has a vital role in consolidating networks and enforcing network edge policies.

The distribution layer design options provide consistent network operation and configuration tools to enable various borderless network services. Three simplified distribution layer design options can be deployed in large, medium, and small campus locations, depending on network scale, application and borderless services demands, and cost, as shown in Figure 14. All distribution design models offer consistent network foundation services, high availability, expansion flexibility, and network scalability. However each enterprise network is different, with unique business challenges that require

a cost-effective aggregation solution, scalability, high-speed network services, virtualized systems, etc., that can be enabled with advanced technologies. Depending on network designs and key technical requirements, network architects must make appropriate aggregation layer design choices to enable end-to-end borderless network services.

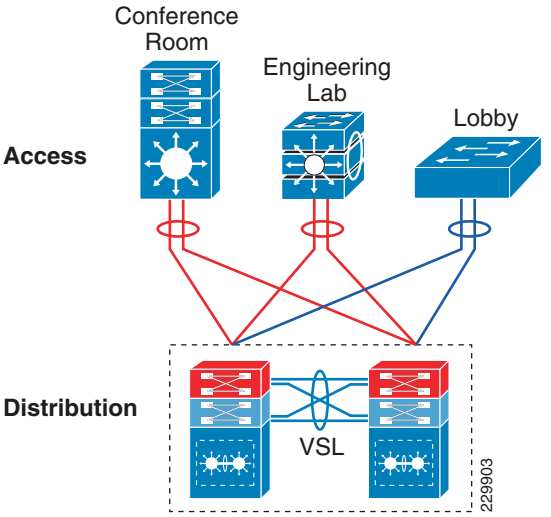
Figure 14 Distribution Layer Design Model Options



Distribution Layer Design Option 1—VSS Mode

Distribution layer design option 1 is intended for the large and medium campus network design and it is based on deploying the Cisco Catalyst 6500-E Series switches using Cisco VSS, as shown in Figure 15.

Figure 15 VSS-Enabled Distribution Layer Network Design

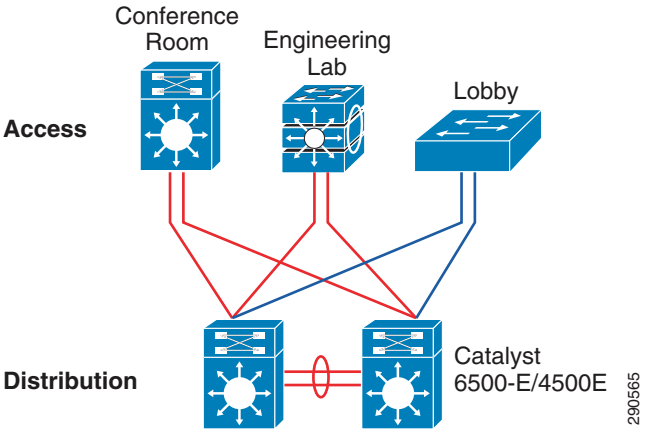


Distribution Layer Design Option 2—Standalone Mode

The distribution layer option 2 is a traditional and proven network design in the enterprise campus network. It can be deployed with redundant Cisco Catalyst 6500 or 4500E systems to operate in standalone mode. This is an alternative distribution network deployment model if there is no desire or capability to virtualize the aggregation layer switches using Cisco VSS technology. In the large campus, the Cisco Catalyst 6500 with non-VSL capable supervisor modules can be deployed in standalone mode, whereas in the medium campus, network administrators can deploy the Catalyst 6500 or the alternative Catalyst 4500E system in standalone mode.

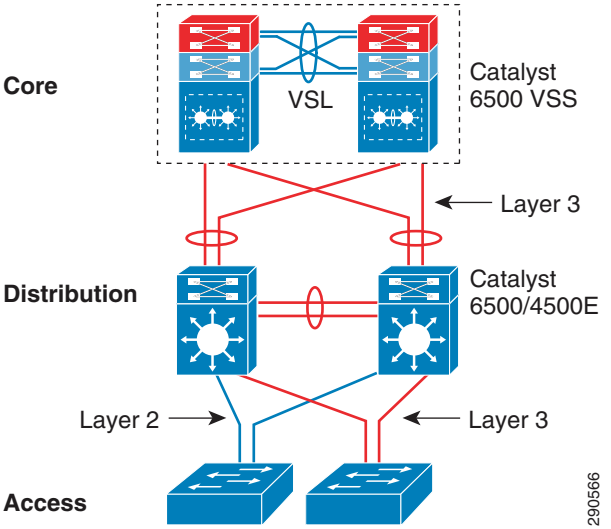
The two single-chassis standalone mode distribution layer design options are shown in Figure 16.

Figure 16 Standalone Mode Distribution Layer Network Design



In the standalone mode, each Catalyst distribution system operates in independent mode and builds local network adjacencies and forwarding information with access and core layer peering devices. Layer 2 and Layer 3 protocol operation is done over each physical interface between the standalone distribution switches and the access layer switches. Since the core layer network in large and medium campus networks is simplified using Cisco VSS technology, the network administrator can simplify the core network topology by bundling Layer 3 interfaces into a logical EtherChannel, as shown in Figure 17.

Figure 17 Network Design with Distribution in Standalone Mode



This network design does not raise any significant concerns in Layer 3 network designs. Each standalone distribution system will establish Layer 3 adjacencies with core and access layer (routed access) devices to develop routing topologies and forwarding tables. The traditional multilayer network design faces the following challenges when the access layer switches communicate with two distinct distribution layer switches:

- The multilayer network uses simple Spanning-Tree Protocol (STP) to build Layer 2 loop-free network paths, which results in a sub-optimal and asymmetric forwarding topology.
- It requires per-VLAN virtual gateway protocol operation between aggregation systems to provide high availability. For large networks, First Hop Redundancy Protocol (FHRP) protocols may limit network scalability and consume more system and network resources.
- For a stable, secure, and optimized multilayer network, each distribution and access layer system will require advanced network parameters tuning.
- Layer 2 network recovery becomes protocol type- and timer-dependent. The default protocol parameters could result in network outages for several seconds during faults. Protocol timers can be tuned aggressively for network recovery within a second range, however it cannot meet the high-availability baseline for business-class video applications like Cisco TelePresence.

Cisco innovated VSS technology to mitigate such challenges. Hence it is recommended to deploy a Cisco VSS-based distribution layer infrastructure that simplifies the multilayer network and increases network capacity and performance, resulting in a highly-reliable network that provides consistent and deterministic network recovery. The traditional standalone-mode distribution layer network is an alternative solution that does not introduce any fundamental design changes. Deployment guidance

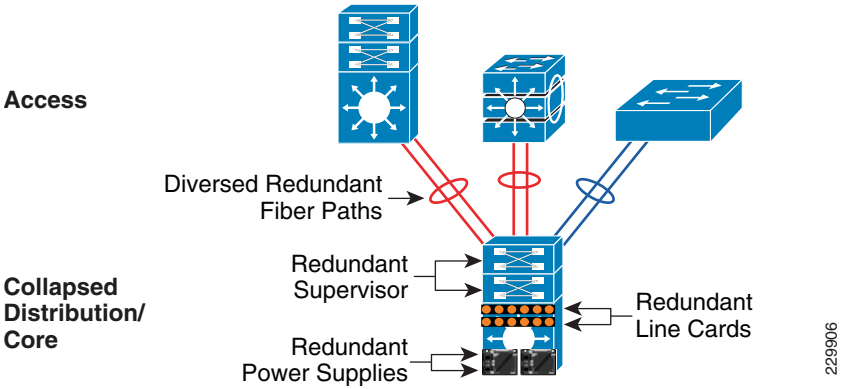
documented in various design guide remains consistent and can be leveraged, hence the standalone distribution network design is not fully re-validated in this document. For more information on configuring and deploying standalone-mode distribution layer Catalyst switches, see the *Campus Network for High Availability Design Guide*: http://www.cisco.com/en/US/partner/docs/solutions/Enterprise/Campus/HA_campus_DG/hacampusdg.html.

Distribution Layer Design Option 3—Collapsed Distribution/Core Mode

The small remote campus location may have several departments working on various floors within a building. Network administrators can consider collapsing the core function into the distribution layer switch for such a small campus where there may only be a single distribution block. The collapsed distribution/core system can provide network services to a small number of wiring closet switches and directly connect to the WAN edge system to reach a large campus for centralized data and communication services. Deploying a two-tier network model for a single distribution block in a small campus does not break the three-tier campus design principles required in large and medium campus networks. This solution is manageable and cost effective as it meets the needs of network users and the number of endpoints is not as large as large or medium enterprise campuses.

The collapsed distribution/core network can be deployed with two redundant systems as recommended in Distribution Layer Design Option 1—VSS Mode or alternatively in standalone mode as described in Distribution Layer Design Option 2—Standalone Mode. In a space-constrained small campus environment, a single Cisco Catalyst 4500E series platform can be deployed with multiple redundant hardware components. Building a single, highly-available, collapsed distribution/core system will ensure the network performance, availability, and reliability required to run borderless services. With various redundant hardware components, this solution can provide 1+1 in-chassis protection against various types of hardware and software failure. Deploying the network in a recommended design will provide consistent sub-second network recovery. A single Cisco Catalyst 4500E with multiple redundant system components can be deployed as shown in Figure 18.

Figure 18 Highly Redundant Single Collapsed Distribution/Core Design

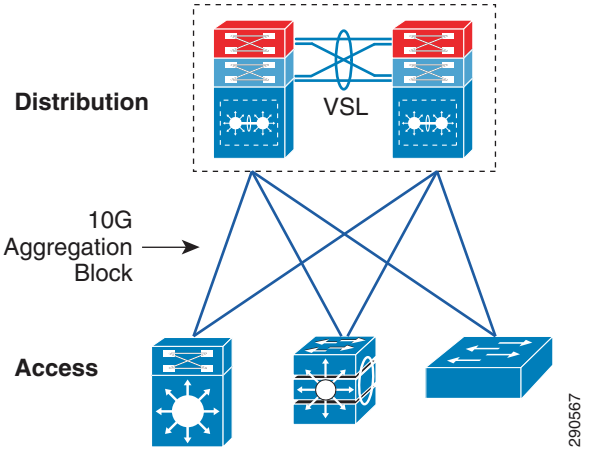


Campus Access Layer Network Design

The access layer is the first tier or edge of the campus, where end devices such as PCs, printers, cameras, Cisco TelePresence, etc. attach to the wired portion of the campus network. It is also the place where devices that extend the network out one more level, such as IP phones and wireless access points (APs), are attached. The wide variety of possible types of devices that can connect and the various services and dynamic configuration mechanisms that are necessary make the access layer one of the most feature-rich parts of the campus network. Not only does the access layer switch allow users to access the network, the access layer switch provides network protection so that unauthorized users or applications do not enter the network. The challenge for the network architect is determining how to implement a design that meets this wide variety of requirements—the need for various levels of mobility, the need for a cost-effective and flexible operations environment, etc.—while being able to provide the appropriate balance of security and availability expected in more traditional, fixed-configuration environments. The next-generation Cisco Catalyst switching portfolio includes a wide range of fixed and modular switching platforms, each designed with unique hardware and software capabilities to function in a specific role.

Enterprise campuses may deploy a wide range of network endpoints which all have different requirements on the network; low-latency, link speed, and low-jitter rates are just some of those requirements. The network architect must consider network requirements, as well as the planned growth of network resources, when determining bandwidth requirements for the access layer to distribution uplinks. To build a high-performance distribution-access block, Cisco access layer switching platforms are designed with 10Gbps uplinks to provide borderless network services at wire-rate.

Figure 19 High-Performance Distribution-Access Block

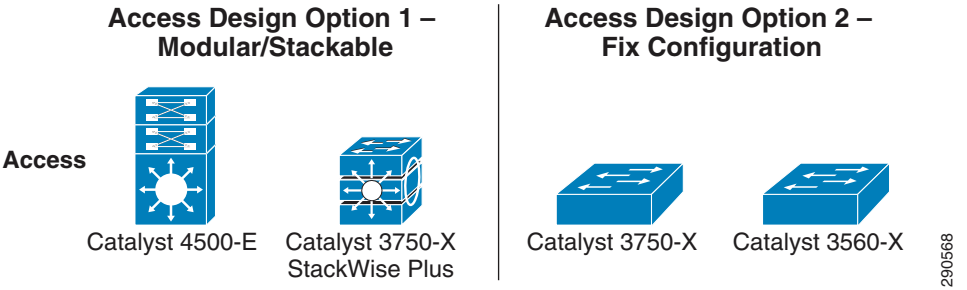


Building a 10Gbps distribution-access block provides the following benefits:

- Increased throughput—Increases network bandwidth capacity ten-fold on a per-physical-port basis. The oversubscription bandwidth ratio in a high-density wiring closet falls within the recommended range.
- High performance—Accelerates application performance by multiplexing a large number of flows onto a single high-speed connection instead of load-sharing across multiple slow aggregate links.
- Reduced TCO—The cost of access switches becomes less per port; it reduces additional cost by deploying fewer cables and connectors when building parallel paths between two systems.
- Simplified design—Single high-speed link to manage, operate, and troubleshoot instead of multiple individual or aggregated bundled connections.

Based on the broad range of business communication devices and endpoints, network access demands, and capabilities, two access layer design options can be deployed, as shown in Figure 20.

Figure 20 Access Layer Design Models

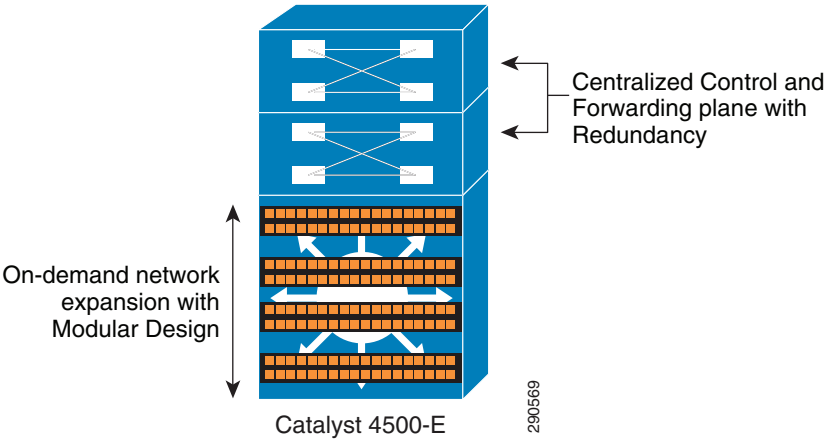


Access Layer Design Option 1—Modular/StackWise Plus Access Layer Network

Access layer design option 1 is intended to address network modularity, performance, scalability, and availability for IT-managed, critical voice and video communication edge devices. To accelerate the user experience and campus physical security protection, these devices require low latency, high performance, and a constantly-available network switching infrastructure. Implementing a modular and stackable Cisco Catalyst switching platform provides the flexibility to increase network scalability in the densely-populated campus network edge.

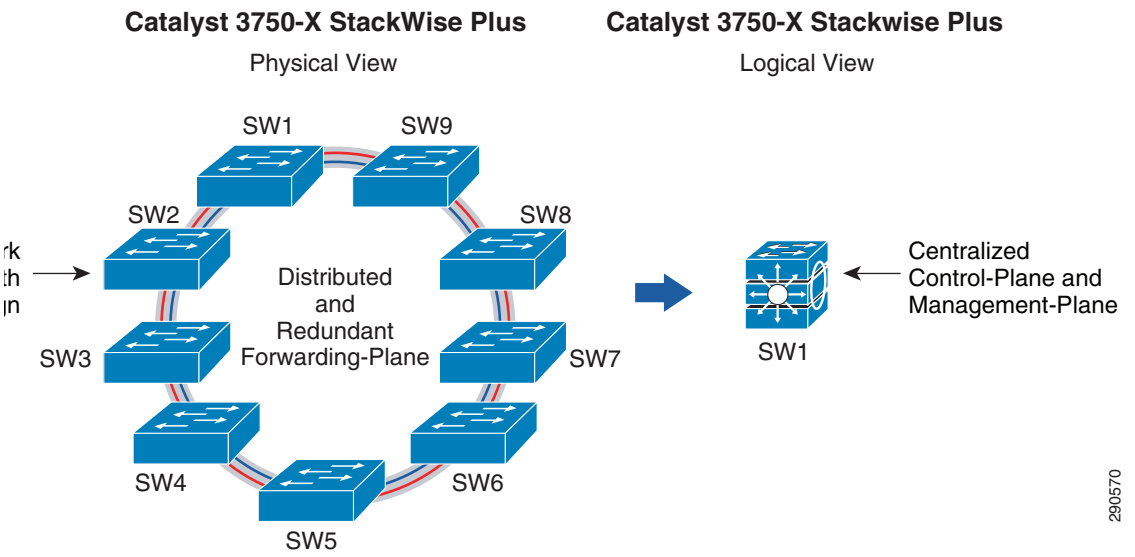
In large and medium campus deployments, the ubiquitous Cisco Catalyst 4500E Series platform provides a scalable, high-speed, and robust network solution. In a high-density access environment, it is imperative to simplify the management of hundred of end points through a single chassis. It is also essential during hardware or software failures to provide wire-speed network performance without compromising network reliability by using a non-stop forwarding architecture. The next-generation hardware architecture of the Cisco Catalyst 4500E in the campus access layer leverages new Cisco IOS software capabilities to enable several borderless network services at the campus network boundary.

Figure 21 Network Edge Expansion with Modular Design



The Cisco Catalyst 3750-X Series is the alternative Cisco access layer switching platform. Using Cisco StackWise Plus technology provides flexibility and availability by clustering multiple Cisco Catalyst 3750-X Series Switches into a single high-speed stack ring that simplifies operation and allows incremental access layer network expansion or contraction. Catalyst 3750-X switches deployed in Cisco StackWise Plus mode alters network operation compared to standalone mode. When deployed in StackWise plus mode, the switches become a single logical access layer switch, the control plane processing becomes centralized, and because of the distributed forwarding architecture, all the hardware resources gets fully utilized across all stack member switches (see Figure 1-22). Cisco StackWise Plus provides high-speed multigigabit switching capacity for network traffic switching within the stack-ring and the distribution-access block can be built with multiple parallel 10Gbps uplink paths for load sharing and network resiliency. The network is optimized and simplified when the cross-switch uplink ports are bundled into a single logical interface using EtherChannel technology. This network design provides non-stop network communication in case of the failure of an individual stack member switch.

Figure 1-22 Network Edge Expansion with StackWise Plus Design



Access Layer Design Option 2—Fixed Configuration Access Layer Network

This entry-level access layer design option is widely chosen for enterprise environments. The fixed configuration Cisco Catalyst switching portfolio supports a wide range of access layer technologies that allow seamless service integration and enable intelligent network management at the edge. Fixed configuration Cisco Catalyst switches in standalone mode are an ideal design choice for a small size wiring closet to provide consistent borderless network services for up to 48 endpoints.

The next-generation fixed configuration Cisco Catalyst 3750-X and 3560-X Series are commonly deployed platforms for wired network access that can be in a mixed configuration with critical devices, such as Cisco IP phones, and non-mission critical endpoints, such as library PCs, printers, and so on. For non-stop network operation during power outages, the Catalyst 3560-X must be deployed with an internal or external redundant power supply solution using the Cisco RPS 2300. Increasing aggregated power capacity provides the flexibility to scale with enhanced Power-over-Ethernet (PoE+) on a per-port basis. With its wire-speed 10G uplink forwarding capacity, this design reduces network congestion and latency to significantly improve application performance.

To provide a consistent end-to-end enhanced user experience, the Cisco Catalyst 3750-X and 3560-X Series platforms support critical network control services to secure the network edge and intelligently provide differentiated services to various class-of-service traffic, as well as simplified management. The Cisco Catalyst must leverage the dual uplink ports to interconnect the distribution system for increased bandwidth capacity and network availability.

Both design options offer consistent network services at the campus edge to provide differentiated, intelligent, and secured network access to trusted and untrusted endpoints. The distribution options recommended in the previous section can accommodate both access layer design options.

6 Summary

As enterprises make a fundamental shift in their networks to meet the new demands of employees, customers, and partners, network infrastructure design decisions are critical. The Borderless Campus 1.0 CVD describes the design decisions required for an enterprise network campus. The Borderless Campus 1.0 architecture provides an architecture that showcases Cisco's best practices. This chapter discuses the design options for each layer in the Borderless Campus 1.0 Architecture. The remaining chapters describe implementation and deployment options.

2

Deploying Network Foundation Services

After designing each tier in the model, the next step in enterprise network design is to establish key network foundation technologies. Regardless of the applications and requirements that enterprises demand, the network must be designed to provide a consistent user experience independent of the geographical location of the user or application. The following network foundation design principles or services must be deployed in each campus location to provide resiliency and availability so all users can access and use enterprise applications:

- Implementing campus network infrastructure
- Network addressing hierarchy
- Network foundation technologies for campus designs
- Multicast for applications delivery
- QoS for application performance optimization
- High availability to ensure business continuity in the event of a network failure

Design guidance for each of these six network foundation technologies is discussed in the following sections, including where they are deployed in each tier of the campus design model, the campus location, and capacity.

1 Implementing Campus Network Infrastructure

Chapter 1, "Borderless Campus Design and Deployment Models," provided various design options for deploying the Cisco Catalyst and Cisco Nexus 7000 platforms in a multi-tier, centralized large campus and remote medium and small campus locations. The Borderless Enterprise Reference network is designed to build simplified network topologies for easier operation, management, and troubleshooting independent of campus location. Depending on network size, scalability, and reliability requirements, the Borderless Enterprise Reference design applies a common set of Cisco Catalyst and Nexus 7000 platforms in different campus network layers as described in Chapter 1, "Borderless Campus Design and Deployment Models."

The foundation of the enterprise campus network must be based on Cisco recommendations and best practices to build a robust, reliable, and scalable infrastructure. This subsection focuses on the initial hardware and software configuration of a wide-range of campus systems to build hierarchical network designs. The recommended deployment guidelines are platform- and operational mode-specific. The initial recommended configurations should be deployed on the following Cisco platforms independent of their roles and the campus tier in which they are deployed. Implementation and deployment guidelines for advanced network services are explained in subsequent sections:

- Cisco Catalyst 6500-E in VSS mode
- Cisco Nexus 7000
- Cisco Catalyst 4500E
- Cisco Catalyst 3750-X Stackwise Plus
- Cisco Catalyst 3750-X 3560-X in standalone mode

Deploying Cisco Catalyst 6500-E in VSS Mode

All the VSS design principles and foundational technologies defined in this subsection remain consistent when the Cisco Catalyst 6500-E is deployed in VSS mode at the campus core or the distribution layer.

Prior to enabling the Cisco Catalyst 6500-E in VSS mode, the enterprise network administrator should adhere to Cisco recommended best practices to take complete advantage of the virtualized system and minimize the network operation downtime when migrating in a production network. Migrating VSS from the standalone Catalyst 6500-E system requires multiple pre- and post-migration steps to deploy the virtual system that includes building the virtual system itself and migrating the existing standalone

network configuration to operate in the virtual system environment. Refer to the following document for the step-by-step migration procedure:
http://www.cisco.com/en/US/products/ps9336/products_tech_note09186a0080a7c74c.shtml

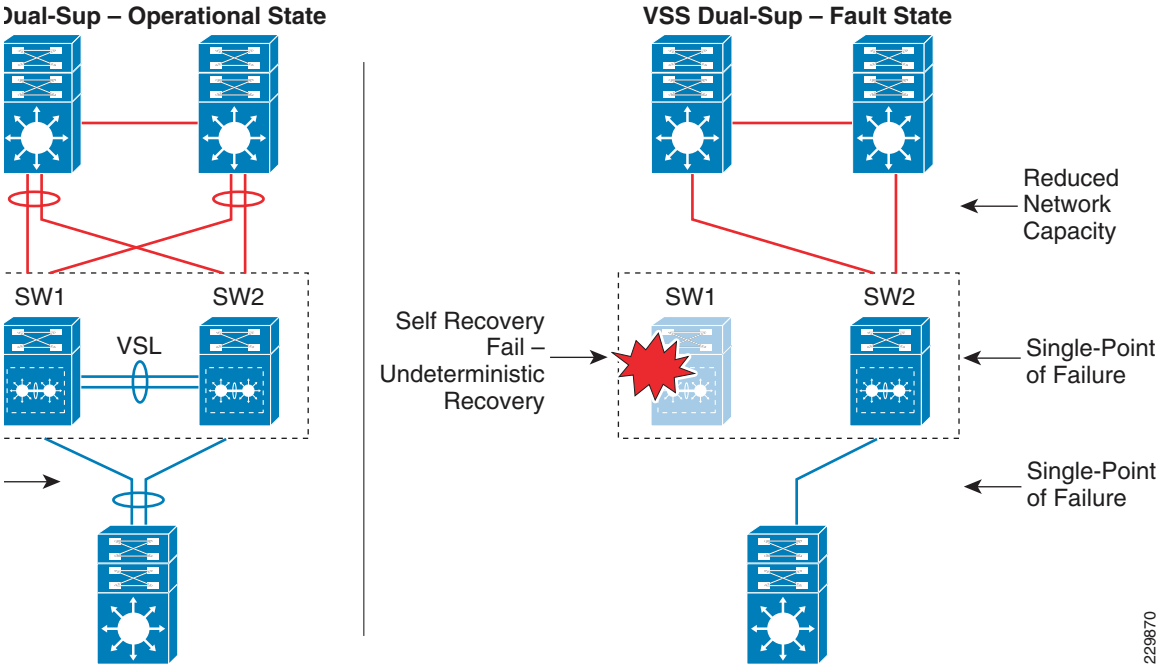
The following subsections provide guidance on the procedures and mandatory steps required when implementing VSS and its components in the campus distribution and core:

- VSS High-Availability
- VSS Identifiers
- Virtual Switch Link
- VSL Design Consideration
- Unified Control-Plane
- VSL Dual-Active Detection and Recovery
- VSL Dual-Active Management
- VSS Quad-Sup Migration

VSS High-Availability

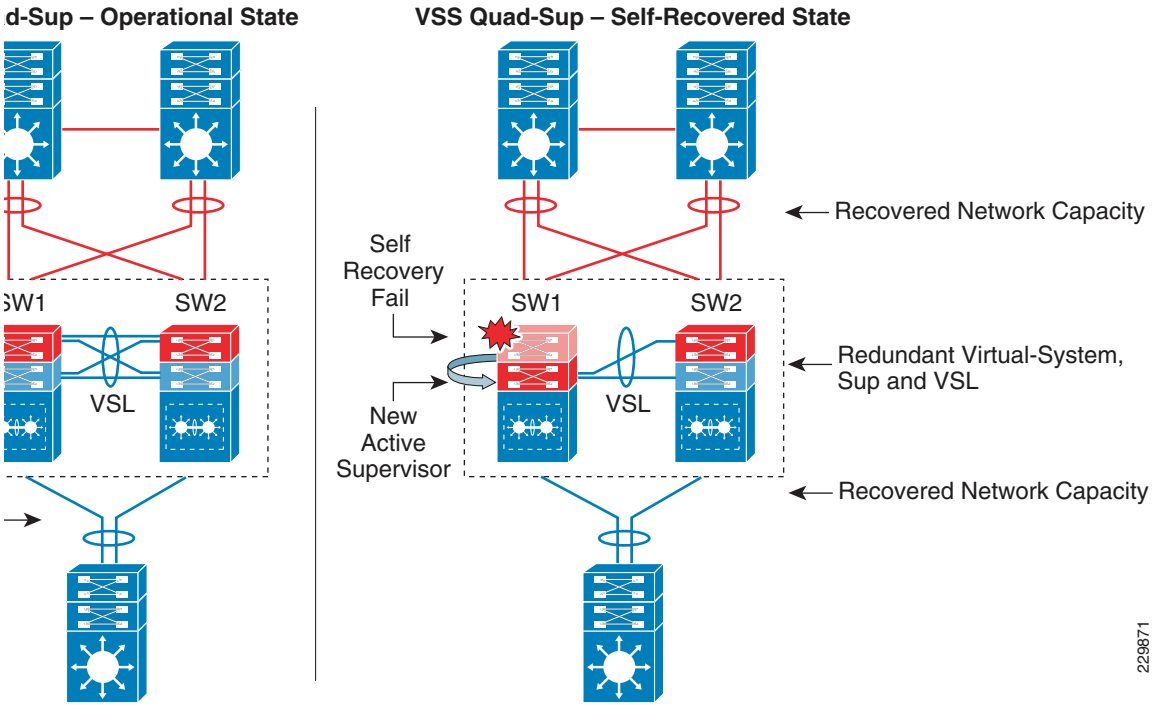
The Cisco Catalyst 6500-E simplifies the control and management plane by clustering two systems deployed in the same role and running them in the same campus layers. Along with multiple innovations to build a unified system with a distributed forwarding design, Cisco VSS technology also leverages the existing single chassis-based NSF/SSO redundancy infrastructure to develop an inter-chassis redundancy solution. Hence it allows for a redundant two-supervisor model option which is distributed between two clustered chassis, instead of a single-standalone redundant chassis. While the dual-sup design solved the original challenge of simplifying network topology and managing multiple systems, in the early phase it was not designed to provide supervisor redundancy within each virtual switch chassis. Hence, the entire virtual switch chassis gets reset during supervisor failure or may remain down if it cannot bootup at all due to faulty hardware or software. In either case, during the fault state the campus network faces several challenges, as illustrated in Figure 1-1.

Figure 1-1 Dual-Sup Failure—Campus Network State



Starting with software release 12.2(33)SXI4 for the Cisco Catalyst 6500-E system running in virtual switching mode, additional features allow for two deployment modes—redundant and non-redundant. Cisco VSS can be deployed in a non-redundant dual-supervisor option (one per virtual switch chassis) and a redundant quad-supervisor option (two per virtual switch chassis). To address the dual-supervisor challenges, the Catalyst 6500-E running in VSS mode introduces innovations that extend dual-supervisor capability with a redundant quad-supervisor to provide intra-chassis (stateless) and inter-chassis (stateful) redundancy, as shown in Figure 1-2.

Figure 1-2 Quad-Sup Failure—Campus Network Recovered State



When designing the network with the Catalyst 6500-E in VSS mode, Cisco recommends deploying VSS in quad-supervisor mode, which helps build a more resilient and mission critical campus foundation network. Deploying VSS in quad-supervisor mode offers the following benefits over the dual-supervisor design:

- Maintains inter-chassis redundancy and all other dual-supervisor benefits
- Increases virtual switch intra-chassis redundancy
- Offers deterministic network recovery and availability during abnormal supervisor failure
- Maintains in-chassis network services module availability
- Minimize single point link or system failure conditions during a fault state
- Protects overall network capacity and reliability

VSS Identifiers

This is the first pre-migration step to be implemented on two standalone Cisco Catalyst 6500-Es in the same campus tier that are planned to be clustered into a single logical entity. Cisco VSS defines the following two types of physical node identifiers to distinguish a remote node within the logical entity, as well as to set the logical VSS domain identity to uniquely identify beyond the single VSS domain boundary.

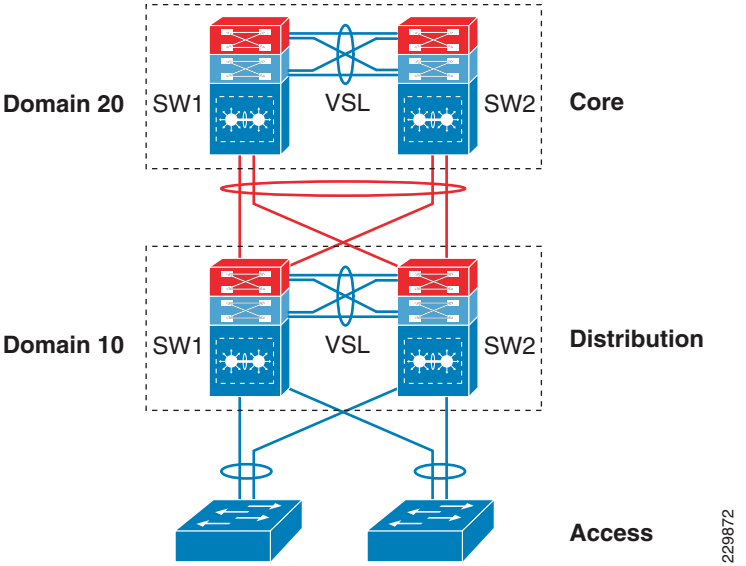
Domain ID

Defining the domain identifier (ID) is the initial step in creating a VSS with two physical chassis. The domain ID value ranges from 1 to 255. The Virtual Switch Domain (VSD) comprises two physical switches and they must be configured with a common domain ID. When implementing VSS in a multi-tier campus network design, the unique domain ID between different VSS pairs prevents network protocol conflicts and allows for simplified network operation, troubleshooting, and management.

Switch ID

Each VSD supports up to two physical switches to build a single logical virtual switch. The switch ID value is 1 or 2. Within VSD, each physical chassis must have a uniquely-configured switch ID to successfully deploy VSS. From a control plane and management plane perspective, when the two physical chassis are clustered after VSS migration, it creates a single large system. Therefore all distributed physical interfaces between the two chassis are automatically appended with the switch ID (i.e., <*switch-id*>/<*slot#*>/<*port#*>) or TenGigabitEthernet 1/1/1. The significance of the switch ID remains within VSD; all the interface IDs are associated to the switch IDs and are retained independent of control plane ownership. See Figure 3.

Figure 3 VSS Domain and Switch ID



The following sample configuration shows how to configure the VSS domain ID and switch ID:

Standalone Switch 1:

```
VSS-SW1(config)# switch virtual domain 20
VSS-SW1(config-vs-domain)# switch 1
```

Standalone Switch 2:

```
VSS-SW2(config)# switch virtual domain 20
VSS-SW2(config-vs-domain)# switch 2
```

Switch Priority

When the two switches boot, switch priority is negotiated to determine control plane ownership for the virtual switch. The virtual switch configured with the higher priority takes control plane ownership, while the lower priority switch boots up in redundant mode. The default switch priority is 100; the lower switch ID is used as a tie-breaker when both virtual switch nodes are deployed with the default settings.

Cisco recommends deploying both virtual switch nodes with identical hardware and software to take full advantage of the distributed forwarding architecture with a centralized control and management plane. Control plane operation is identical on each of the virtual switch nodes. Modifying the default switch priority is an optional setting since each of the virtual switch nodes can provide transparent operation to the network and the user.

Virtual Switch Link

To cluster two physical chassis into single a logical entity, Cisco VSS technology enables the extension of various types of single-chassis internal system components to the multi-chassis level. Each virtual switch must be deployed with direct physical links, which extend the backplane communication boundary (these are known as Virtual-Switch Links (VSL)).

VSL can be considered Layer 1 physical links between two virtual switch nodes and are designed to operate without network control protocols. Therefore, VSL links cannot establish network protocol adjacencies and are excluded when building the network topology tables. With customized traffic engineering on VSL, it is tailored to carry the following major traffic categories:

- Inter-switch control traffic
  - Inter-Chassis Ethernet Out Band Channel (EOBC) traffic— Serial Communication Protocol (SCP), IPC, and ICC
  - Virtual Switch Link Protocol (VSLP) —LMP and RRP control link packets
- Network control traffic
  - Layer 2 protocols —STP BPDU, PagP+, LACP, CDP, UDLD, LLDP, 802.1x, DTP, etc.
  - Layer 3 protocols—ICMP, EIGRP, OSPF, BGP, MPLS LDP, PIM, IGMP, BFD, etc.
- Data traffic
  - End user data application traffic in single-home network designs
  - An integrated service module with centralized forwarding architecture (i.e., FWSM)
  - Remote SPAN

Using EtherChannel technology, the VSS software design provides the flexibility to increase on-demand VSL bandwidth capacity and to protect network stability during VSL link failure or malfunction.

The following sample configuration shows how to configure VSL EtherChannel:

Standalone Switch 1:

```
VSS-SW1(config)# interface Port-Channel 1
VSS-SW1(config-if)#description SW1-VSL-EtherChannel
VSS-SW1(config-if)# switch virtual link 1

VSS-SW1(config)# interface range Ten 1/1 , Ten 5/4 - 5 , Ten 6/4 -5
VSS-SW1(config-if)#description SW1-VSL-Dual-Sup-Uplink+6708
VSS-SW1(config-if)# channel-group 1 mode on
```

Standalone Switch 2:

```
VSS-SW2(config)# interface Port-Channel 2
VSS-SW1(config-if)#description SW2-VSL-EtherChannel
VSS-SW2(config-if)# switch virtual link 2
```

```
VSS-SW2(config)# interface range Ten 1/1 , Ten 5/4 - 5 , Ten 6/4 -5
VSS-SW1(config-if)#description SW2-VSL-Dual-Sup-Uplink+6708
VSS-SW2(config-if)# channel-group 2 mode on
```

VSL Design Consideration

Implementing VSL EtherChannel is a simple task, however it requires a properly optimized design to achieve high reliability and availability. Deploying VSL requires careful planning to keep system virtualization intact during VSS system component failure on either virtual switch node. Reliable VSL design requires planning in three categories:

- VSL Links Diversification
- VSL Bandwidth Capacity
- VSL QoS

VSL Links Diversification

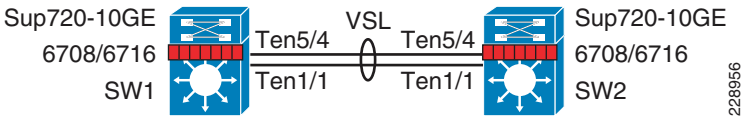
Complete VSL link failure may break system virtualization and create network instability. Designing VSL link redundancy by using diverse physical paths on both systems prevents network instability, eliminates single point-of-failure conditions, and optimizes the bootup process.

VSL is not supported on all Catalyst 6500-E linecards due to the special VSL encapsulation headers that are required within the VSL protocols. The next-generation specialized Catalyst 6500-E 10G-based supervisor and linecard modules are fully capable and equipped with modern hardware ASICs to support VSL communication. VSL EtherChannel can bundle 10G member links with any of following next-generation hardware modules:

- Sup720-10G
- WS-X6708
- WS-X6716-10G and WS-X6716-10T (must be deployed in performance mode to enable VSL capability)

When VSS is deployed in dual-sup and quad-sup designs, Cisco recommends a different VSL design to maintain network stability. In a dual-sup VSS configuration, Figure 4 shows an example of how to build VSL EtherChannel with multiple diverse physical fiber paths using the supervisor 10G uplink ports and VSL-capable 10G hardware modules.

Figure 4 Recommended Dual-Sup VSL Links Design

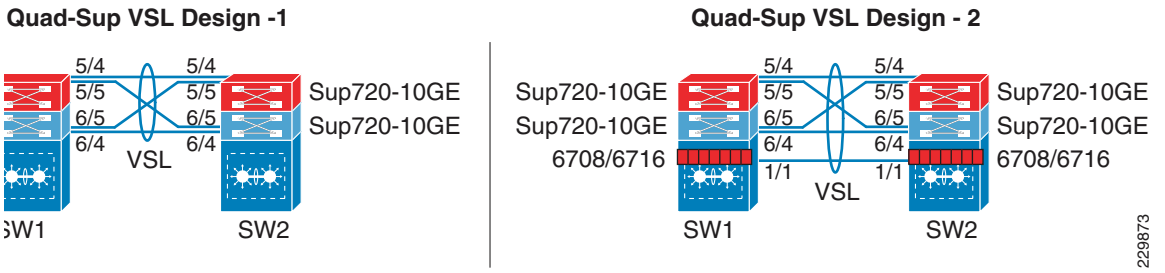


Deploying VSL with multiple, diversified VSL ink designs offers the following benefits:

- Leverage non-blocking 10G ports from supervisor modules.
- Use 10G ports from VSL-capable WS-X6708 or WS-X6716 linecard module to protect against abnormal failures on the supervisor uplink port (e.g., GBIC failure).
- Reduces single point-of-failure probabilities as it is rare to trigger multiple hardware faults on diversified cables, GBIC, and hardware modules.
- A VSL-enabled 10G module boots up more rapidly than other installed modules in the system. The software is designed to initialize VSL protocols and communication early in the bootup process. If the same 10G module is shared to connect other network devices, then depending on the network module type and slot bootup order, it is possible to minimize traffic loss during the system initialization process.
- Use a four-class, built-in QoS model on each VSL member link to optimize inter-chassis communication traffic, network control, and user data traffic.

Since VSS quad-sup increases chassis redundancy, network architects must redesign VSL to maintain its reliability and capacity during individual supervisor module failures. Figure 5 shows two different approaches to building VSL EtherChannel with multiple diverse and full-mesh fiber paths between all four supervisor modules in the VSS domain. Both designs are recommended to increase VSL reliability, capacity, and diversity during multiple abnormal faults conditions. When migrating from a dual-sup VSS design (Figure 4) to a quad-sup VSS design, VSL re-design is necessary, as illustrated in Figure 5.

Figure 5 Recommended Quad-Sup VSL Links Design



Deploying a diverse and full-mesh VSL in a VSS quad-sup design offers the following benefits:

- Both quad-sup VSL designs leverage all of the key advantage of a dual-sup VSS design.
- Both designs use the non-blocking 10G uplink ports from all four supervisors to build full-mesh VSL connection.
- VSS quad-sup design option # 1 offers the following benefits:
  - A cost-effective solution to leverage all four supervisor modules to build well-diversified VSL paths between both chassis. It provides the flexibility to continue to use a non-VSL capable linecard module for other network functions.

  - Increases the overall bandwidth capacity to enable more hardware-integrated borderless network and application services at the campus aggregation layer.
  - Maintains up to 20G VSL bandwidth for traffic redirection during individual supervisor module failure.
  - Increases network reliability by minimizing the probability of dual-active or split-brain conditions that de-stabilize the VSS domain and the network.
- VSS quad-sup design option # 2 offers following benefits:
  - The primary difference between option 1 and 2 is option #2 introduces additional 10G ports from VSL-capable WS-X6708, WS-X6716-10G, or WS-X6716-10T linecard modules.
  - This design option provides additional VSL link redundancy and bandwidth protection against abnormal failure that prevents redundant supervisor modules in both virtual switch chassis from booting.
  - Accelerates VSL-enabled 10G module bootup time, which allows for rapid build up of network adjacencies and forwarding information.
  - Based on EtherChannel load sharing rules, this design may not provide optimal network data load sharing across all five VSL links. However, this VSL design is highly suited to protect VSS system reliability during multiple supervisor fault conditions.

VSL Bandwidth Capacity

From each virtual switch node, VSL EtherChannel can bundle up to eight physical member links. Therefore, VSL can be bundled up to 80G of bandwidth capacity; the exact capacity depends on the following factors:

- The aggregated network uplink bandwidth capacity on per virtual switch node basis, e.g., 2 x 10GE diversified to the same remote peer system.
- Designing the network with single-homed devices connectivity (no MEC) forces at least half of the downstream traffic to flow over the VSL link. This type of connectivity is highly discouraged.
- Remote SPAN from one switch member to other. The SPANed traffic is considered as a single flow, hence the traffic hashes only over a single VSL link that can lead to oversubscription of a particular link. The only way to improve traffic distribution is to have an additional VSL link. Adding a link increases the chance of distributing the normal traffic that was hashed on the same link carrying the SPAN traffic, which may then be sent over a different link.
- If the VSS is carrying the borderless services hardware (such as FWSM, WiSM, etc.), then depending on the service module forwarding design, it may be carried over the VSL bundle. Capacity planning for each of the supported services blades is beyond the scope of this design guide.

For optimal traffic load sharing between VSL member-links, it is recommended to bundle VSL member links in powers of 2 (i.e., 2, 4, and 8).

VSL QoS

The service and application demands of next-generation enterprise networks require a strong, resilient, and highly-available network with on-demand bandwidth allocation for critical services without compromising performance. Cisco VSS in dual- and quad-sup designs are designed with application intelligence and automatically enable QoS on the VSL interface to provide bandwidth and resource allocation for different class-of-service traffic.

The QoS implementation on VSL EtherChannel operates in restricted mode as it carries critical inter-chassis backplane traffic. Independent of global QoS settings, the VSL member links are automatically configured with system-generated QoS settings to protect different class of applications. To retain system stability, the inter-switch VSLP protocols and QoS settings are fine tuned to protect high priority traffic with different thresholds, even during VSL link congestion.

To deploy VSL in non-blocking mode and increase the queue depth, the Sup720-10G uplink ports can be configured in one of the following two QoS modes:

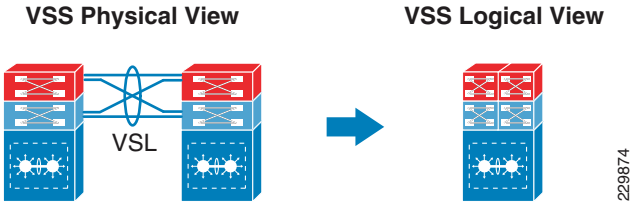
- *Non-10G-only mode (Default)*—In this mode, all supervisor uplink ports must follow a single queuing mode. Each 1G/10G supervisor uplink port operates with a default queue structure (1p3q4T). If any 10-Gbps uplink port is used for the VSL link, the remaining ports (10 Gbps or 1Gbps) follow the same CoS mode of queuing for any other non-VSL connectivity because VSL only allows CoS-based queuing. This default VSL QoS structure is designed to automatically protect critical inter-switch communication traffic to maintain virtual system reliability and to preserve original QoS settings in data packets for consistent QoS treatment as a regular network port. The default mode also provides the flexibility to utilize all 1G supervisor uplink ports for other network applications (e.g., out-of-band network management).
- *10G-only mode*—In this mode, the 10G supervisor uplink ports can double the egress queue structure from 1p3q4t to 1p7q4t. To increase the number egress queues on 10G uplink port from a shared hardware ASIC, this mode requires all 1G supervisors uplink ports to be administratively disabled. Implementing this mode does not modify the default CoS-based trust (ingress classification), queueing (egress classification) mode, or the mapping table.

  The 10G-only QoS mode can increase the number of egress queue counts but cannot provide the additional advantage to utilize all queue structures since it maintains all other QoS settings the same as default mode; hence it is recommended to retain the VSL QoS in default mode.

Unified Control-Plane

Clustering two Cisco Catalyst 6500-E chassis into a logical chassis builds a unified control plane that simplifies network communications, operations, and management. Cisco VSS technology centralizes the control and management plane to ease operational challenges and utilizes all network resources with intelligent distributed forwarding decisions across the virtual switch (see Figure 6).

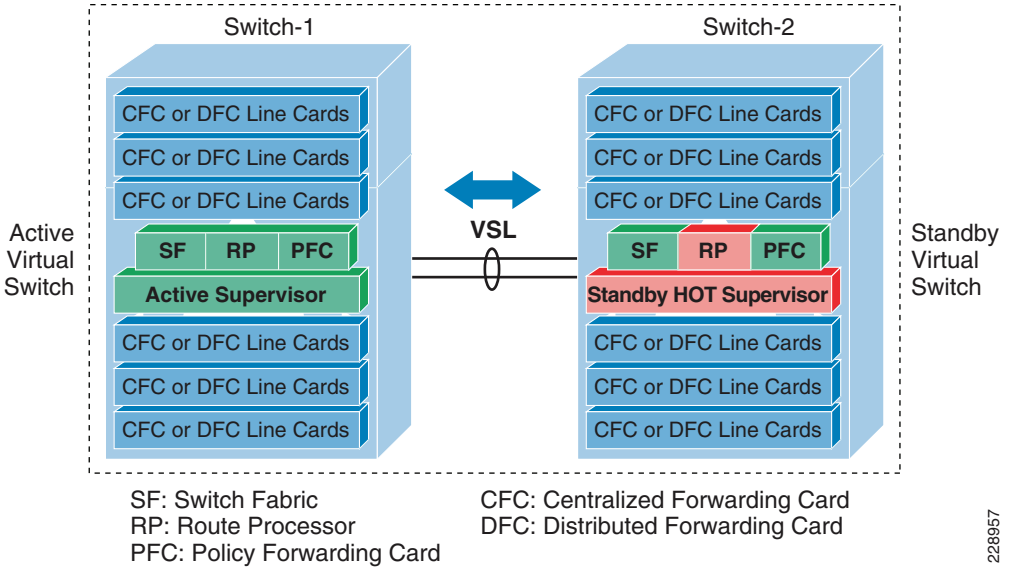
Figure 6 VSS Quad-sup Physical versus Logical View



Deploying redundant supervisor with common hardware and software components into a single standalone Cisco Catalyst 6500-E platform automatically enables the Stateful Switch Over (SSO) capability, providing in-chassis supervisor redundancy. The SSO operation on the active supervisor holds control plane ownership and communicates with remote Layer 2 and Layer 3 neighbors to build distributed forwarding information. The SSO-enabled active supervisor is tightly synchronized with the standby supervisor and synchronizes several components (protocol state machine, configuration, forwarding information, etc.). As a result, if an active supervisor fails, a hot-standby supervisor takes over control plane ownership and initializes graceful protocol recovery with peer devices. During the network protocol graceful recovery process, forwarding information remains non-disrupted, allowing for nonstop packet switching in hardware.

Leveraging the same SSO and NSF technology, Cisco VSS in dual-sup and quad-sup designs supports inter-chassis SSO redundancy by extending the supervisor redundancy capability from a single-chassis to multi-chassis. Cisco VSS uses VSL EtherChannel as a backplane path to establish SSO communication between the active and hot-standby supervisor deployed in a separate chassis. The entire virtual switch node gets reset during abnormal active or hot-standby virtual switch node failure. See Figure 7.

Figure 7 Inter-Chassis SSO Operation in VSS—Dual-Sup and Quad-Sup



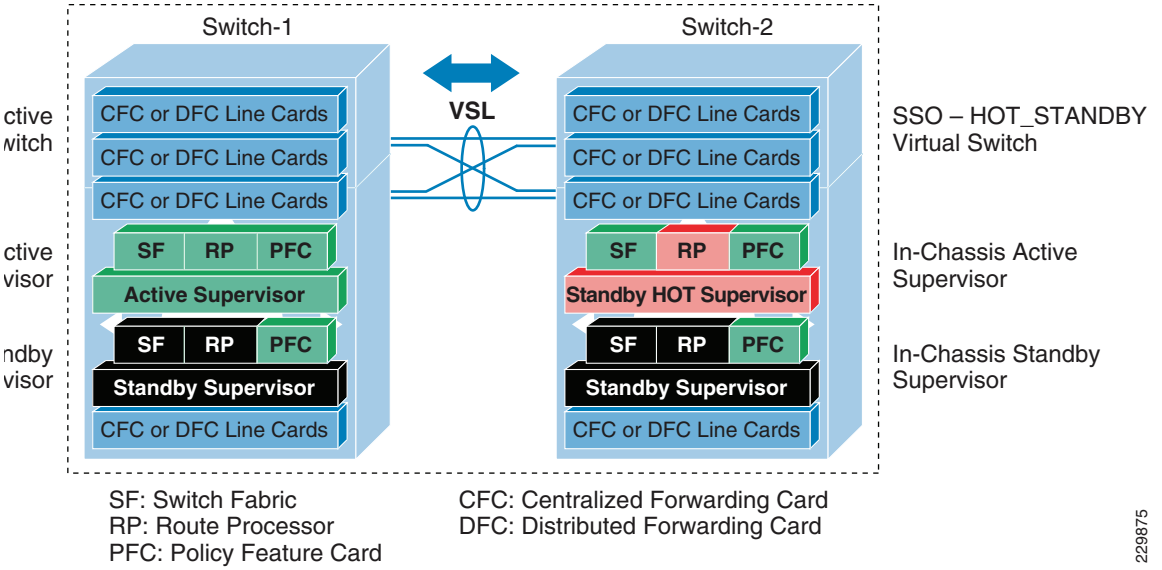
To successfully establish SSO communication between two virtual switch nodes, the following criteria must match between both virtual switch nodes:

- Identical software versions
- Consistent VSD and VSL interface configurations
- Power mode and VSL-enabled module power settings
- Global PFC Mode
- SSO and NSF-enabled

During the bootup process, SSO synchronization checks all of the above criteria with the remote virtual system. If any of the criteria fail to match, it forces the virtual switch chassis to boot in an RPR or cold-standby state that cannot synchronize the protocol and forwarding information with the current SSO ACTIVE virtual switch chassis.

While inter-chassis SSO redundancy provides network simplification and stability, it cannot provide chassis-level redundancy when a supervisor in either of the virtual switches fails and cannot self-recover. Cisco VSS with quad-sup retains the orginal inter-chassis design as the dual-sup design, however it addresses intra-chassis dual-sup redundancy challenges. To support quad-sup redundant capability, the software infrastructure requires changes to simplify the roles and responsibilities to manage the virtual switch with a distributed function. Figure 8 illustrates two supervisor modules in each virtual switch chassis to increase redundancy, enable distributed control and forwarding planes, and intelligently utilize all possible resources from quad-sups.

Figure 8 VSS Quad-Sup Redundancy



The quad-sup VSS design divides the role and ownership between two clustered virtual switching systems and allows the in-chassis redundant supervisors to operate transparently and seamlessly of SSO communication:

- SSO ACTIVE—The supervisor module that owns the control and management plane. It establishes network adjacencies with peer devices and develops distributes forwarding information across the virtual switching system.
- SSO HOT_STANDBY—The supervisor module in HOT_STANDBY mode that synchronizes several system components, state machines, and configurations in real-time to provide graceful and seamless recovery during SSO ACTIVE failure.
- In-Chassis ACTIVE (ICA)—The in-chassis active supervisor module within the virtual switch chassis. The ICA supervisor module could be in the SSO ACTIVE or HOT_STANDBY role. The ICA module communicates with and controls all modules deployed in the local chassis.
- In-Chassis STANDBY (ICS)—The in-chassis standby supervisor module within the virtual switch chassis. The ICS supervisor module communicates with the local ICA supervisor module to complete its WARM bootup procedure and keeps system components in synchronization to take over ICA ownership if the local ICA fails and cannot recover.

The Cisco Catalyst 6500-E running in VSS quad-sup mode introduces Route-Processor Redundancy-WARM (RPR-WARM) to provide intra-chassis redundancy. Cisco's new RPR-WARM supervisor redundancy technology is a platform-dependent innovation that enables co-existence of

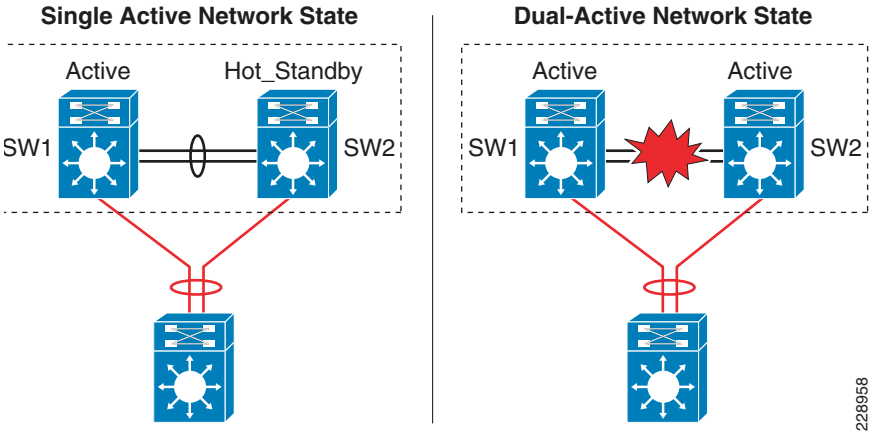
multiple supervisor redundant modes in a virtual switch. Only a compatible ICS supervisor module may bootup in RPR-WARM mode. The term RPR-WARM combines following hybrid functions on ICS supervisor modules:

- Supervisor—RPR-WARM extends the capability of legacy RPR cold-state redundancy technology and synchronizes certain system configurations such as the startup-configuration, VLAN database, and BOOT variable. The in-chassis standby supervisor does not synchronize any hardware or software state machines to provide graceful recovery during local ICA module failure. Refer to Chapter 4, "Deploying High Availability in Campus," for more details.
- Distributed linecard—RPR-WARM enables a unique function on the ICS supervisor module. During the bootup process ICS supervisor initializes as the regular supervisor, but gets WARM upgraded with the new Sup720-LC IOS software, which allows the ICS supervisor to operate as a distributed linecard. The Sup720-LC IOS software is packaged within the 12.2(33)SXI4 IOS software release. When the ICS module is operating in distributed linecard and RPR-WARM mode, it enables all its physical ports for network communication and is synchronized with hardware information for distributed forwarding.

VSL Dual-Active Detection and Recovery

The preceding section described VSL EtherChannel functions as an extended backplane link that enables system virtualization by transporting inter-chassis control traffic, network control plane, and user data traffic. The state machine of the unified control plane protocols and distributed forwarding entries gets dynamically synchronized between the two virtual switch nodes. Any fault triggered on a VSL component leads to a catastrophic instability in the VSS domain and beyond. The virtual switch member that assumes the role of hot-standby maintains constant communication with the active switch. The role of the hot-standby switch is to assume the active role as soon as it detects a loss of communication with its peer via all VSL links without the operational state information of the remote active peer node. Such an unstable network condition is known as *dual-active*, where both virtual switches get split with a common set of configurations and each individually takes control plane ownership to communicate with neighboring devices. The network protocols detect inconsistency and instability when VSS peering devices detect two split systems claiming the same addressing and identification. Figure 9 depicts the state of a campus topology in a single active-state and a dual-active state.

Figure 9 Single Active and Dual-Active Campus Topology



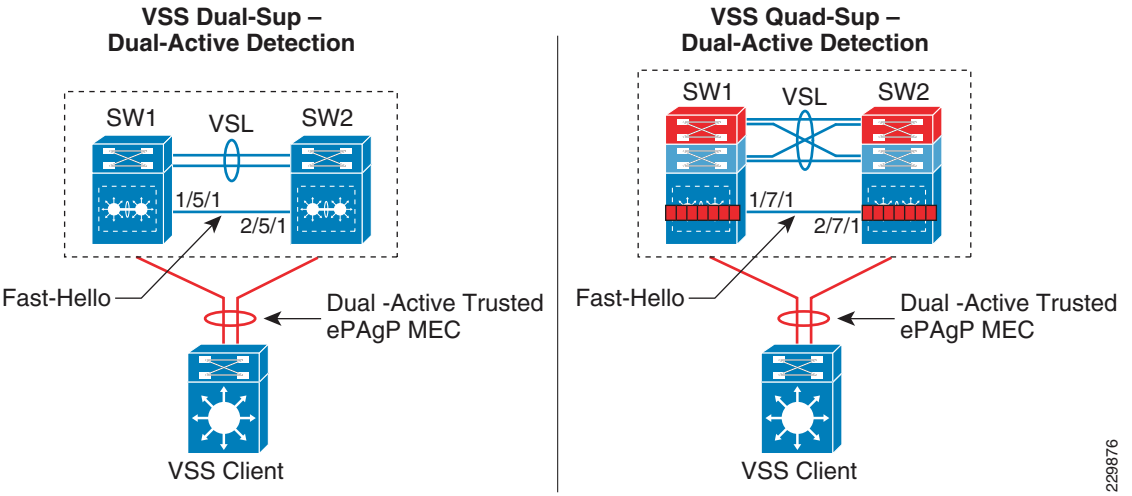
System virtualization is affected during the dual-active network state and splits the single virtual system into two identical Layer 2/Layer3 systems. This condition can destabilize campus network communication, with two split system advertising duplicate information. To prevent such network instability, Cisco VSS introduces the following two methods to rapidly detect dual-active conditions and recover by isolating the old active virtual switch from network operation before the network becomes destabilized:

- Direct detection method—This method requires an extra physical connection between both virtual switch nodes. Dual-Active Fast-Hello (Fast-Hello) and Bidirectional Forwarding Decision (BFD) protocols are specifically designed to detect the dual-active condition and prevent network malfunction. All VSS-supported Ethernet media and modules can be used to deploy this method. For additional redundancy, VSS allows the configuration of up to four dual-active fast-hello links between virtual switch nodes. Cisco recommends deploying Fast-Hello in lieu of BFD for the following reasons:
  - Fast-Hello can rapidly detect a dual-active condition and trigger the recovery procedure. Independent of routing protocols and network topology, Fast-Hello offers faster network recovery.
  - Fast-Hello enables the ability to implement dual-active detection in a multi-vendor campus or data center network environment.
  - Fast-Hello optimizes the protocol communication procedure without reserving higher system CPU and link overheads.
  - Fast-Hello supersedes a BFD-based detection mechanism.
  - Fast-Hello links do not carry network control or user data traffic, so they can be enabled on regular Gigabit-Ethernet links.

- Indirect detection method—This method relies on an intermediate trusted Layer2/Layer3 MEC Cisco Catalyst remote platform to detect the failure and notify the old-active switch about the dual-active detection. Cisco extended the capability of the PAgP protocol with extra TLVs to signal the dual-active condition and initiate the recovery procedure. Most Cisco Catalyst switching platforms can be used as a trusted PAgP+ partner to deploy the indirect detection method.

All dual-active detection protocols and methods can be implemented in parallel. As depicted in Figure 10, in a VSS network deployment peering with Cisco Catalyst platforms, Cisco recommends deploying Fast-Hello and PAgP+ methods for rapid detection to minimize network topology instability and to retain application performance. In a dual-sup VSS design, Fast-Hello can be implemented between supervisor or linecard ports. Fast-hello and PAgP+ detection methods do not operate differently on VSS quad-sup. However to ensure Fast-Hello links are available during supervisor failures and minimize the number of Fast-Hello ports used, it is recommended to deploy Fast-Hello on the linecard modules instead of the supervisor.

Figure 10 Recommended Dual-Active Detection Method



The following sample configuration illustrates the implementation of both methods:

- Dual-Active Fast-Hello

```
cr23-VSS-Core(config)#interface range Gig1/7/1 , Gig2/7/1
cr23-VSS-Core(config-if-range)# dual-active fast-hello

! Following logs confirms fast-hello adjacency is established on
! both virtual-switch nodes.
%VSDA-SW1_SP-5-LINK_UP: Interface Gi1/7/1 is now dual-active detection capable
%VSDA-SW2_SPSTBY-5-LINK_UP: Interface Gi2/7/1 is now dual-active detection capable
```

```
cr23-VSS-Core#show switch virtual dual-active fast-hello
Fast-hello dual-active detection enabled: Yes
Fast-hello dual-active interfaces:
Port       Local StatePeer Port    Remote State
-------------------------------------------------------------------------------
Gi1/7/1    Link up        Gi2/7/1      Link up
```

- PAgP+

  Enabling or disabling dual-active trusted mode on Layer 2/Layer 3 MEC requires MEC to be in an administrative shutdown state. Prior to implementing trust settings, network administrators should plan for downtime to provision PAgP+-based dual-active configuration settings:

```
cr23-VSS-Core(config)#int range Port-Channel 101 - 102
cr23-VSS-Core(config-if-range)#shutdown

cr23-VSS-Core(config)#switch virtual domain 20
cr23-VSS-Core(config-vs-domain)#dual-active detection pagp trust channel-group 101
cr23-VSS-Core(config-vs-domain)#dual-active detection pagp trust channel-group 102

cr23-VSS-Core(config)#int range Port-Channel 101 - 102
cr23-VSS-Core(config-if-range)#no shutdown
```

```
cr23-VSS-Core#show switch virtual dual-active pagp
PAgP dual-active detection enabled: Yes
PAgP dual-active version: 1.1

Channel group 101 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
          Dual-Active     Partner              Partner   Partner
Port      Detect Capable  Name                 Port      Version
Te1/1/2   Yes              cr22-6500-LB         Te2/1/2   1.1
Te1/3/2   Yes             cr22-6500-LB         Te2/1/4   1.1
Te2/1/2   Yes             cr22-6500-LB         Te1/1/2   1.1
Te2/3/2   Yes             cr22-6500-LB         Te1/1/4   1.1

Channel group 102 dual-active detect capability w/nbrs
Dual-Active trusted group: Yes
          Dual-Active     Partner              Partner   Partner
Port      Detect Capable  Name                 Port      Version
Te1/1/3   Yes             cr24-4507e-MB        Te4/2     1.1
Te1/3/3   Yes             cr24-4507e-MB        Te3/1     1.1
Te2/1/3   Yes             cr24-4507e-MB        Te4/1     1.1
```

```
    Te2/3/3    Yes             cr24-4507e-MB        Te3/2      1.1
```

VSL Dual-Active Management

Managing the VSS system during a dual-active condition becomes challenging when two individual systems in the same network tier contain a common network configuration. Network instability can be prevented with the dual-active detection mechanism; the old ACTIVE virtual-switch chassis disables all physical and logical interfaces. By default this recovery mechanic blocks in-band management access to the system to troubleshoot and analyze the root cause of dual-active. The network administrator should explicitly configure VSS to exclude disabling the network management interface during the dual-active recovery state. Any Layer 3 physical management ports can be configured to be excluded; since dual-active breaks system virtualization, logical interfaces like SVIs and Port-Channels cannot be excluded. To minimize network instability, it is highly recommended to exclude only network management ports from both virtual-switch chassis. The following is a sample configuration to exclude the Layer 3 network management ports of both virtual-switch chassis:

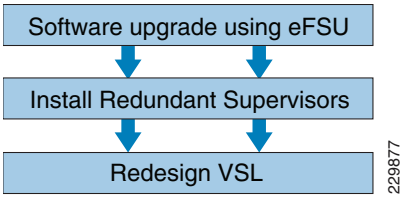
```
Dist-VSS(config)#switch virtual domain 1
Dist-VSS(config-vs-domain)#dual-active exclude interface Gi1/5/2
Dist-VSS(config-vs-domain)#dual-active exclude interface Gi2/5/2

Dist-VSS#show switch virtual dual-active summary
Pagp dual-active detection enabled: Yes
Bfd dual-active detection enabled: Yes
Fast-hello dual-active detection enabled: Yes

Interfaces excluded from shutdown in recovery mode:
  Gi1/5/2
  Gi2/5/2
In dual-active recovery mode: No
```

VSS Quad-Sup Migration

As in new deployments, migrating from a dual-sup to a quad-sup VSS design is simplified and can be performed without network downtime. The existing dual-sup VSS systems must be migrated to quad-sup-capable IOS software on the ICA and ICS modules to deploy quad-sup.

Note ICS supervisor modules must not be inserted into virtual switching until the ICA supervisor is deployed with VSS quad-sup software capability. Inserting the ICS supervisor module prior to upgrading the software is not supported and may create adverse effects on the system and network.

Figure 11 VSS Quad-Sup Migration Procedure



Cisco recommends the following migration guidelines to gracefully deploy quad-sup:

Step 1 Upgrade Cisco IOS software on the ICA supervisor module to 12.2(33)SXI4 and later releases. This upgrade procedure must leverage Enhanced Fast Software Upgrade (eFSU) technology to gracefully upgrade both virtual switch chassis. Verify that the network is stable and fully operational. eFSU is supported starting with 12.2(33)SXI3 and later releases. For more details, refer to Catalyst 6500-E VSS eFSU Software Design and Upgrade Process in Chapter 4, "Deploying High Availability in Campus."

Step 2 Physically insert the ICS supervisor module; it must bootup with the same Cisco IOS software as the ICA module. The ICS supervisor module must meet the following criteria when deploying VSS in quad-sup mode:

- Identical supervisor module types
- ICA and ICS are running identical 12.2(33)SXI4 and later releases software and license versions.

The ICS supervisor role and current status on both virtual switch chassis can be verified upon completing a successful bootup process:

```
cr23-VSS-Core#show switch virtual redundancy | inc Switch|Current Software

My Switch Id = 1
Peer Switch Id = 2

Switch 1 Slot 5 Processor Information :
Current Software state =

Switch 1 Slot 6 Processor Information :
Current Software state = RPR-Warm

Switch 2 Slot 5 Processor Information :
Current Software state = STANDBY HOT (switchover target)

Switch 2 Slot 6 Processor Information :
Current Software state = RPR-Warm
```

Step 3 Pre-configure full-mesh VSL connections between the ICA and ICS modules and bundle into existing VSL EtherChannel, as illustrated in Figure 5. Connect new VSL fiber links fibers and make sure they are in an operational state. Finally move the VSL cable when required to build full-mesh VSL connections.

Deploying VSS Quad-Sup with Mismatch IOS Version

The intra-chassis ICA/ICS role negotiation procedure and parameters are different than the SSO role negotiation that occurs between ICA on two virtual switch chassis. During the bootup process, the ICA and ICS go through several intra-chassis negotiation processes— role, software compatibility check, etc. If any of the criteria fail to match with ICA, then the ICA forces ICS to fallback in (ROMMON) mode.

Deploying quad-sup with mismatched IOS versions between ICA and ICS becomes challenging when installation is done remotely or the VSS system is not easily accessible to install compatible IOS software on local storage of the ICS supervisor module. In such cases, Cisco IOS software provides the flexibility to disable version mismatch check and allow ICS to bootup with a different IOS version than ICA. However ICS must boot with the Cisco IOS software that includes quad-sup capability—12.2(33)SXI4 and later releases.

The network administrator must execute following step prior to inserting the ICS supervisor to mitigate IOS mismatch challenge between ICA and ICS module:

Step 1 Disable IOS software mismatch version check from global configuration mode:

```
cr23-VSS-Core(config)#no switch virtual in-chassis standby bootup mismatch-check
```

Step 2 Physically insert the ICS supervisor module in virtual switches SW1 and SW2. During intra-chassis role negotiation, ICA will report the following message, however it will allow ICS to complete the bootup process in RPR-WARM mode:

```
%SPLC_DNLD-SW1_SP-6-VS_SPLC_IMAGE_VERSION_MISMATCH: In-chassis Standby in switch 1 is
trying to boot with a different image version than the In-chassis Active
    cr23-VSS-Core#show module switch 1 | inc Supervisor
      5    5  Supervisor Engine 720 10GE (Active)    VS-S720-10G        SAD120407CT
      6    5  Supervisor Engine 720 10GE (RPR-Warm)  VS-S720-10G        SAD120407CL
```

Step 3 Copy the ICA-compatible IOS software version on the local storage of the SW1 and SW2 ICS supervisor modules:

```
cr23-VSS-Core#copy <image_src_path> sw1-slot6-disk0:<image>
cr23-VSS-Core#copy <image_src_path> sw2-slot6-disk0:<image>
```

Step 4 Re-enable IOS software mismatch version check from global configuration mode. Cisco recommends to keep version check enabled; running mismatched IOS versions between ICA and ICS may cause SSO communication to fail during the next switchover process, which will result in the virtual switch entering RPR mode and keeping the entire chassis in a non-operational state.

```
cr23-VSS-Core(config)#switch virtual in-chassis standby bootup mismatch-check
```

Step 5 Force ICS supervisor module reset. In the next bootup process, the ICS module will now bootup with an ICA-compatible IOS software version:

```
cr23-VSS-Core#hw-module switch 1 ics reset
Proceed with reset of Sup720-LC? [confirm] Y

cr23-VSS-Core#hw-module switch 2 ics reset
Proceed with reset of Sup720-LC? [confirm] Y
```

Table 1 summarizes the ICS operational state during bootup process in each virtual switch chassis with compatible and incompatible Cisco IOS software versions.

Table 1 VSS Quad-sup Software Compatibility Matrix and ICS State

SW1-ICA – IOS (SSO-ACTIVE)	SW1-ICS – IOS	SW2-ICA – IOS (SSO-STANDBY)	SW2-ICS – IOS	SW1 ICS State SW2 ICS State
12.2(33)SXI4	12.2(33)SXI4	12.2(33)SXI4	12.2(33)SXI4	SW1 ICS – RPR SW2 ICS – RPR
12.2(33)SXI4	12.2(33)SXI4 a	12.2(33)SXI4	12.2(33)SXI4 a	SW1 ICS – ROM SW2 ICS – ROM
12.2(33)SXI4	12.2(33)SXI4 a	12.2(33)SXI4	12.2(33)SXI4	SW1 ICS – ROM SW2 ICS – RPR
12.2(33)SXI4	PRE 12.2(33)SXI4	12.2(33)SXI4	PRE 12.2(33)SXI4	SW1 ICS – ROM SW2 ICS – ROM

1. Software version compatibility check can be disabled to allow ICS to bootup with mismatch version
2. Disabling Software version compatibility check is ineffective in this case as ICS is attempting to boo software that does not support VSS Quad-sup capability.

Virtual Routed MAC

The MAC address allocation for the interfaces does not change during a switchover event when the hot-standby switch takes over as the active switch. This avoids gratuitous ARP updates (MAC address changed for the same IP address) from devices connected to VSS. However, if both chassis are rebooted at the same time and the order of the active switch changes (the old hot-standby switch comes up first and becomes active), then the entire VSS domain uses that switch's MAC address pool. This means that the interface inherits a new MAC address, which triggers gratuitous ARP updates to all Layer 2 and Layer 3 interfaces. Any networking device connected one hop away from the VSS (and any networking device that does not support gratuitous ARP) will experience traffic disruption until the MAC address of the default gateway/interface is refreshed or timed out. To avoid such a disruption, Cisco recommends using the configuration option provided with the VSS in which the MAC address for Layer 2 and Layer 3 interfaces is derived from the reserved pool. This takes advantage of the virtual switch domain identifier to form the MAC address. The MAC addresses of the VSS domain remain consistent with the usage of virtual MAC addresses, regardless of the boot order.

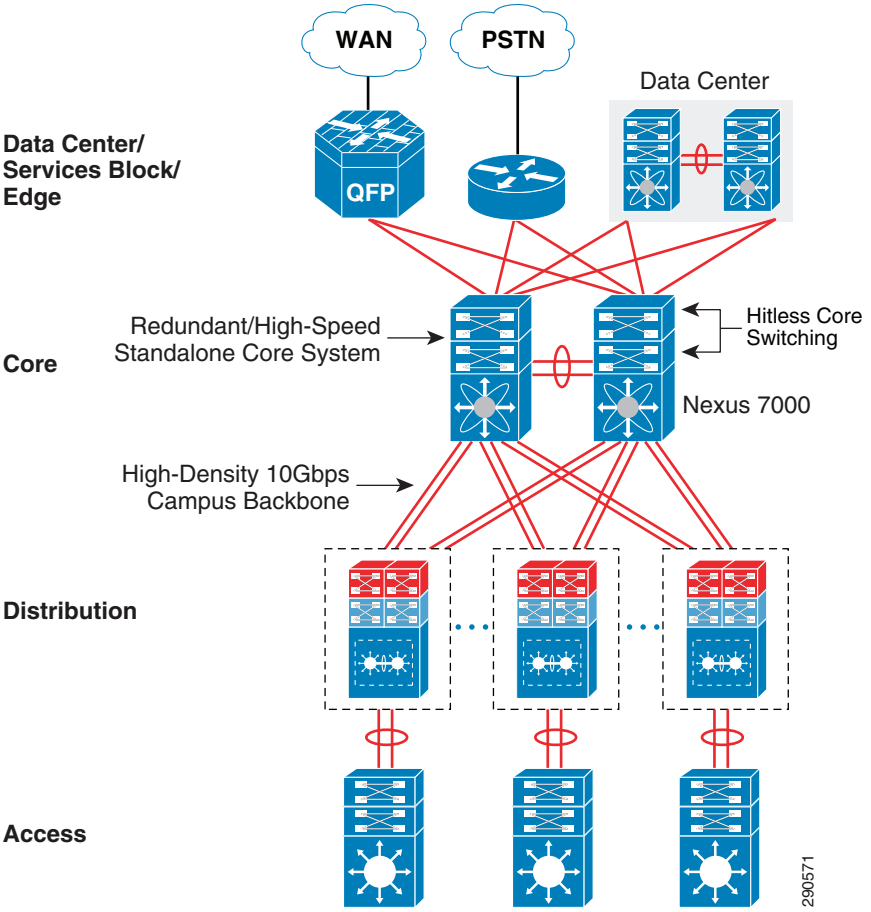
The following configuration illustrates how to configure the virtual routed MAC address for the Layer 3 interface under switch-virtual configuration mode:

```
cr23-VSS-Core(config)#switch virtual domain 20
cr23-VSS-Core(config-vs-domain)#mac-address use-virtual
```

Deploying Cisco Nexus 7000

The campus core can be deployed in an alternative standalone network design to a virtualized campus core with Cisco VSS technology. In the large and medium enterprise campus network, architects may need a solution with a high-performance network backbone to handle data traffic at wire-speed and future proof the backbone to scale the network for high density. Cisco recommends deploying the Cisco Nexus 7000 system in the campus core, as its robust system architecture is specifically designed to deliver high-performance networking in large-scale campus and data center network environments. With advanced hardware, lossless switching architecture, and key foundational software technology support, the Cisco Nexus 7000 is equipped to seamlessly integrate in the enterprise campus core layer. The Nexus 7000 is designed to operate using the next-generation unified Cisco NX-OS operating system that combines advanced technology and software architectural benefits from multiple operating systems. Cisco NX-OS can interoperate in a heterogeneous vendor campus network using industry-standard network protocols.

Figure 12 Cisco Nexus 7000-based Campus Core design



The campus core layer baseline requirement to build a high-speed, scalable, and resilient core remains intact when deploying the Cisco Nexus 7000. It is highly recommended to deploy redundant system components to maintain optimal backbone switching capacity and network availability for non-stop business communication. This subsection provides guidelines and recommendations for initial system setup with the following redundant components:

- Implementing Redundant Supervisor Module
- Distributed Forwarding with Crossbar Fabric Module
- Deploying Layer 3 Network I/O Module

Implementing Redundant Supervisor Module

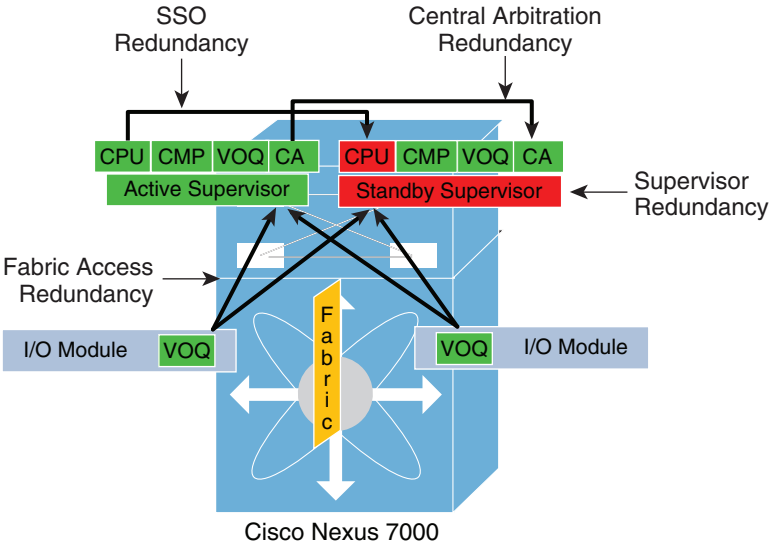
The supervisor module maintains the centralized network control and management plane of the Nexus 7000 system. The robust architecture of the Sup-1 module is uniquely designed with multiple components and decouples the control, management, and data plane operation. All Layer 3 routing functions are handled centrally by the supervisor module to communicate with peering devices and build the dynamic routing and forwarding (FIB) table. The supervisor dynamically synchronizes FIB to the forwarding engine on every I/O for distributed forwarding to optimize switching performance. To optimize the performance, scalability, and reliability of the Nexus 7000 system, it is critical to understand its internal architecture and the function of each of its components. This section provides a brief explanation of some of the key supervisor internal components:

- CPU—The network control plane processing is handled by the dual-core CPU that offers flexibility to scale control plane capacity in a large campus network environment. To maintain control plane redundancy, the Nexus 7000 system must be deployed in a redundant supervisor network configuration. The system configuration, network protocols, and internal state machines are constantly synchronized between the active and standby supervisor modules.
- CMP—Connectivity Management Processor (CMP) is a dedicated "light-weight" CPU running independent of the operating system on the active and standby supervisor modules for out-of-band management and monitoring capability during system disaster recovery.
- Central Arbiter—A special ASIC to control the crossbar switch fabric with intelligence to access data communication between I/O modules. By combining the central arbitration function on the supervisor with distributed Virtual Output Queue (VOQ) on distributed I/O modules, it offers multiple benefits:
  - The central arbiter ensures that the distributed traffic switched through the switch backplane gets fair fabric access and bandwidth on egress from the I/O module (e.g., multiple ingress interfaces sending traffic to a single upstream network interface or I/O module).
  - Intelligently protects and prioritizes the high-priority data or control plane traffic switched through the system fabric.
  - VOQ is a hardware buffer pool that represents the bandwidth on an egress network module. The central arbiter grants fabric access to the ingress I/O module if bandwidth on the egress module is available. This software design minimzes network congestion and optimizes fabric bandwidth usage.
  - Provides 1+1 redundancy and hitless switching architecture with active/active central arbitration on active and standby supervisor modules.

Deploying redundant supervisor modules is a base system requirement to provide non-stop business communication during any hardware or software abnormalities. With graceful protocol capability and lossless switching architecture, the Nexus 7000-based campus core can be hitless during soft switchover of the active supervisor module. The Cisco Nexus 7000 supports NSF/SSO for Layer 3 protocols to provide supervisor redundancy; the active supervisor module synchronizes the NSF-capable protocol state machines to the standby supervisor to take over ownership during

switchover. The I/O modules remain in a full operational state and continue to switch network data traffic with distributed FIB, while the new supervisor gracefully recovers protocol adjacencies with neighbor devices. Figure 13 summarizes the distributed system components and redundant plane architecture in the Cisco Nexus 7000 system.

Figure 13 Cisco Nexus 7000 Supervisor Redundancy Architecture



CPU: Central Processing Unit
CMP: Connectivity Management Processing
VOQ: Virtual Output Queue
CA: Central Arbiter
290572

The Nexus 7000 system by default gets configured in HA or SSO configuration mode by adding a redundant supervisor module. Even the Layer 3 protocol is by default in NSF-capable mode and does not require the network administrator to manually configure NSF capability in a redundant system.

```
cr35-N7K-Core1#show system redundancy status
Redundancy mode
---------------
      administrative:   HA
         operational:   HA

This supervisor (sup-1)
-----------------------
    Redundancy state:   Active
    Supervisor state:   Active
      Internal state:   Active with HA standby

Other supervisor (sup-2)
```

```
-----------------------
    Redundancy state:    Standby
    Supervisor state:    HA standby
      Internal state:    HA standby
```

Distributed Forwarding with Crossbar Fabric Module

As described earlier, Cisco Nexus 7000 is a fully-distributed system that decouples the control and data plane functions between different modular components. While the supervisor module handles control processing centrally and I/O modules perform distributed forwarding, the Nexus 7000 system has a multi-stage and modular crossbar switch fabric architecture. The crossbar fabric module enables multi-terabit backplane switching capacity between high-speed I/O modules through a dedicated fabric channel interface. For high fabric bandwidth capacity and fabric module redundancy, up to five crossbar fabric modules can be deployed in a single Nexus 7000 system. It is imperative to understand the internal crossbar fabric module function in order to design the Nexus 7000 system in the campus core for better core network switching capacity and resiliency.

Fabric Bandwidth Access

As described earlier, the fabric ASIC and VOQ on each distributed I/O module request fabric access from the central arbiter located on the supervisor. If the destination or egress module has sufficient bandwidth, the request gets granted and the data can be sent to the fabric module. If the ingress and egress ports are located on the same I/O module, then the central arbitration process does not get involved. This internal crossbar switching architecture provides multiple benefits as described in the previous section.

Fabric Module Bandwidth Capacity

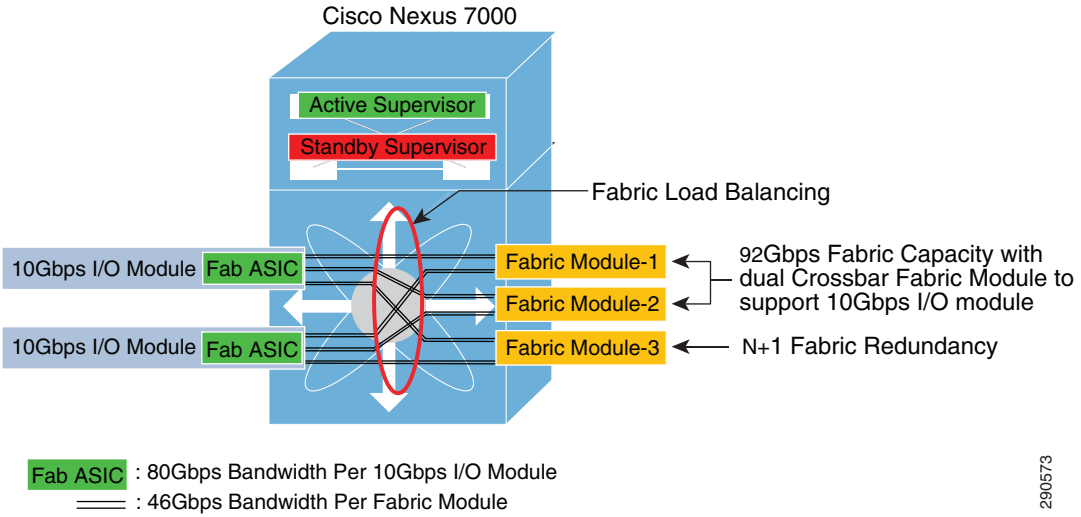
With the parallel fabric switching architecture, each I/O module can achieve up to 46 Gbps of bi-directional backplane capacity from each individual fabric module. To access the fabric bandwidth between I/O modules, the supervisor continues to provide central arbitration to optimally and intelligently utilize all fabric switching capacity. Deploying additional fabric modules can increase the aggregated fabric switching capacity and fabric redundancy on the Nexus 7000 system. For wire-speed per-port performance, the 10G I/O module supports up to 80 Gbps per slot. Hence to gain complete 10Gbps I/O module capacity, it is recommended to deploy at least two fabric modules in the system.

Fabric Module Load Balancing

The I/O module builds two internal uni-directional parallel switching paths with each crossbar fabric module. With the multi-stage fabric architecture, the ingress module determines the fabric channel and module to select prior to forwarding data to the backplane. The Nexus 7000 system is optimized to intelligently load share traffic across all available crossbar fabric modules. The unicast data traffic gets

load balanced in round-robin style across each active fabric channel interface and multicast traffic leverages a hash algorithm to determine a path to send traffic to the fabric module. This internal load share design helps optimize fabric bandwidth utilization and increase redundancy.

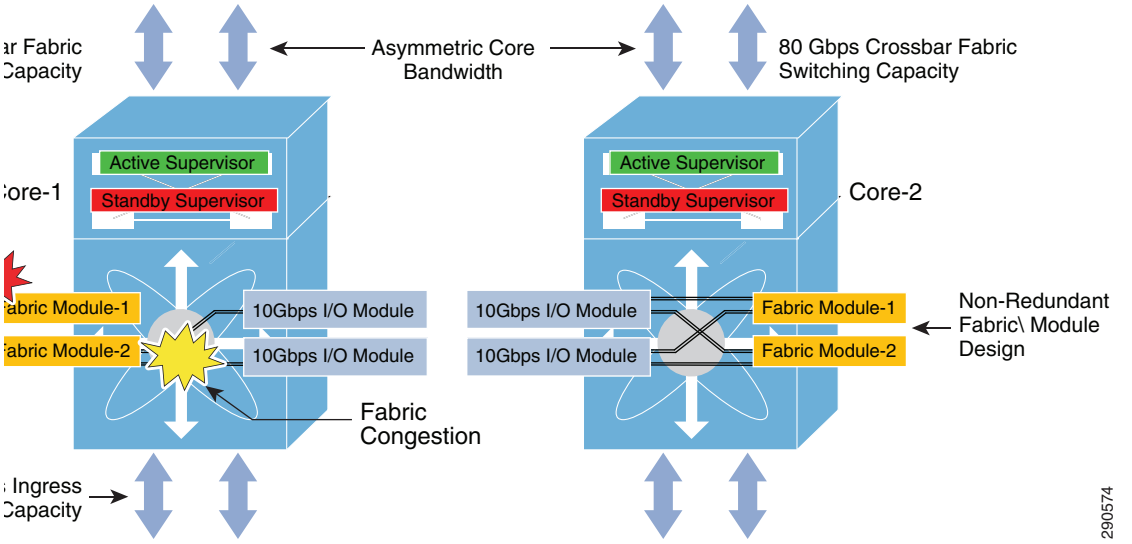
Figure 14 Nexus 7000 Crossbar Fabric Module Architecture



Fabric Module Redundancy

Maintaining Nexus 7000 backplane switching capacity is as important as core layer network switching capacity. With all network paths in an operational state, the 80Gbps per slot I/O module may not operate at full switching capacity due to a lack of available internal switch fabric bandwidth during fabric module failure. It is highly recommended to deploy at least three fabric modules in each Nexus 7000 core chassis to provide N+1 fabric module redundancy. Deploying fabric module redundancy maintains consistent internal backplane switching capacity and allows the mission-critical core layer system to seamlessly operate during abnormal fabric failure. Adding two additional fabric modules can future-proof the Nexus 7000 by increasing backplane bandwidth and further increasing fabric module redundancy. Application and network services may be impacted due to asymmetric core layer switching capacity and may cause internal fabric congestion in a non-redundant campus core system during individual fabric module failure (see Figure 1-15).

Figure 1-15 Crossbar Fabric Module Failure—Campus Network State



Deploying a crossbar fabric module in the Cisco Nexus 7000 system does not require additional configuration to enable internal communication and forwarding of data traffic. The network administrator can verify the number of installed crossbar fabric modules and their current operational status on a Cisco Nexus 7000 system with this command:

```
cr35-N7K-Core1# show module fabric | inc Fab|Status
Xbar  Ports  Module-Type                    Model              Status
1     0      Fabric Module 1               N7K-C7010-FAB-1    ok
2     0      Fabric Module 1               N7K-C7010-FAB-1    ok
3     0      Fabric Module 1               N7K-C7010-FAB-1    ok
```

Deploying Layer 3 Network I/O Module

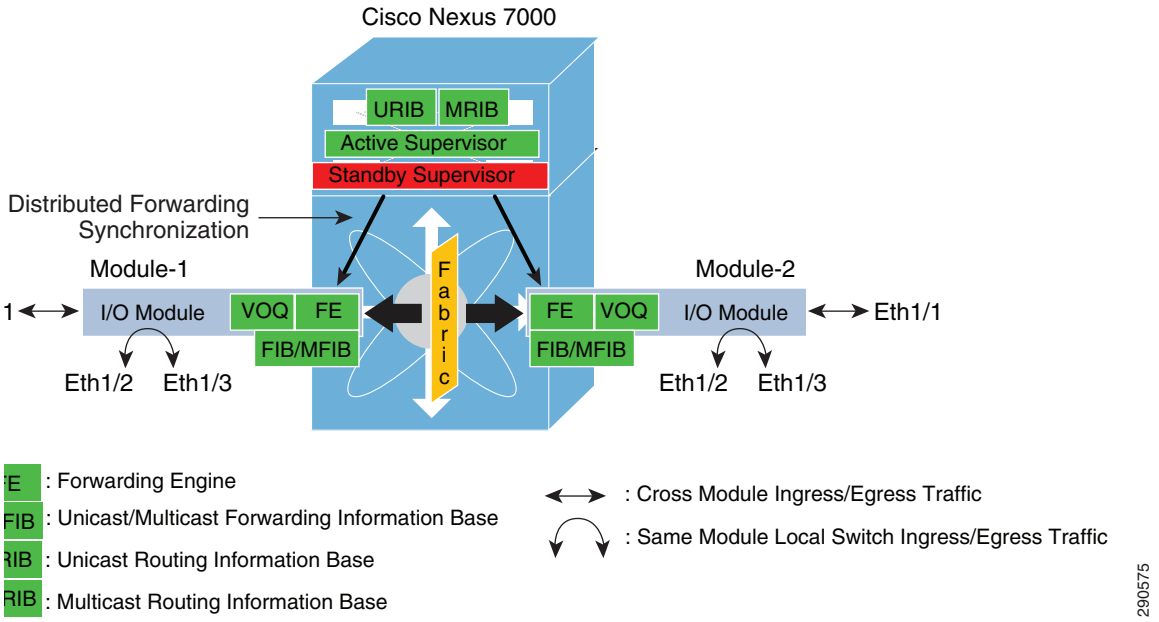
To build a high-speed, non-blocking 10Gbps campus core network, it is recommended deploy a high-performance, scalable, and Layer 3 services-capable 10G I/O network module. The Cisco Nexus 7000 system supports the M1 series 10G I/O module with advanced Layer 3 technologies in campus network designs. The Nexus 7000 system supports the 8 port 10G M108 I/O module (N7K-M108X2-12L) and the 32 port 10G M132 I/O module (N7K-M132XP-12L); the architecture and capability of each I/O module is specifically designed to address a broad set of networking and technology demands.

The 8 port M108 10G network I/O module is specifically designed for deployment in a high-scale campus core network that provides wire-speed switching performance to optimize quality and integrity of applications, services performance, and advanced Layer 3 technology. The module is equipped with dual-active forwarding engines to minimize switching delay and reduce latency by

actively performing egress forwarding lookups, increasing capacity and performance with distributed and intelligent QoS, ACL, etc. The 32 port 10G M132 module is designed for deployment in a high-density, high-performance aggregation layer to connect a large number of access switches. The backplane capacity of the M132 is the same as the M108, however with increased port density it operates at a 4:1 oversubscription rate that may not be an ideal hardware design for a high-speed campus backbone network. Hence it is recommended to deploy the M108 I/O module in the Cisco Nexus 7000 system in the campus core layer.

The active supervisor module establishes the protocol adjacencies with neighbors to build the unicast routing information base (URIB) or multicast routing information base (MRIB). The software-generated routing information remains operational on the supervisor module. To provide hardware-based forwarding, each I/O network module builds local distributed forwarding information tables. The distributed FIB table on the I/O module is used for local destination address and path lookup to forward data with new egress information without involving the supervisor module in the forwarding decision. The central arbitration and crossbar fabric data switching is involved when the egress port is on a different I/O network module. The network data traffic can be "local-switch" if the ingress and egress path is within the same module based on local FIB information. If the Nexus 7000 core system is deployed with multiple M108 I/O modules, then designing the ingress and egress physical paths may help optimize performance and increase network redundancy.

Figure 16 Cisco Nexus 7000 Distributed I/O Forwarding Architecture



The I/O network module is plug-n-play like the crossbar fabric module and does not require any specific user provisioning to enable internal system communication. The operational status of the module can be verified using the same syntax:

```
cr35-N7K-Core1# show module  | inc Gbps
1    8       10 Gbps Ethernet XL Module       N7K-M108X2-12L     ok
2    8       10 Gbps Ethernet XL Module       N7K-M108X2-12L     ok
```

Deploying Cisco Catalyst 4500E

In the Borderless Campus design, the Cisco Catalyst 4500E Series platform can be deployed in different roles. In large campus networks, the Catalyst 4500E must be deployed in a high-density access layer that requires wire-speed network services with enterprise-class network reliability for constant business operation during abnormal faults. In medium campus networks, the Catalyst 4500E series platform can be considered as an alternative aggregation layer system to the Catalyst 6500. The Catalyst 4500E can also be deployed in a collapsed distribution/core role for a small, space-constrained campus location. The next-generation Cisco Catalyst 4500E Series switch is a multi-slot, modular, highly-scalable, multi-gigabit, high-speed, and resilient platform. A single Catalyst 4500E Series platform in an enterprise design is built with redundant hardware components to be consistent with the Catalyst 6500-E VSS-based design. For Catalyst 4500E in-chassis supervisor redundancy, network administrators must consider Catalyst 4507R+E or 4510R+E slot chassis to accommodate redundant supervisors and use LAN network modules for core and edge network connectivity.

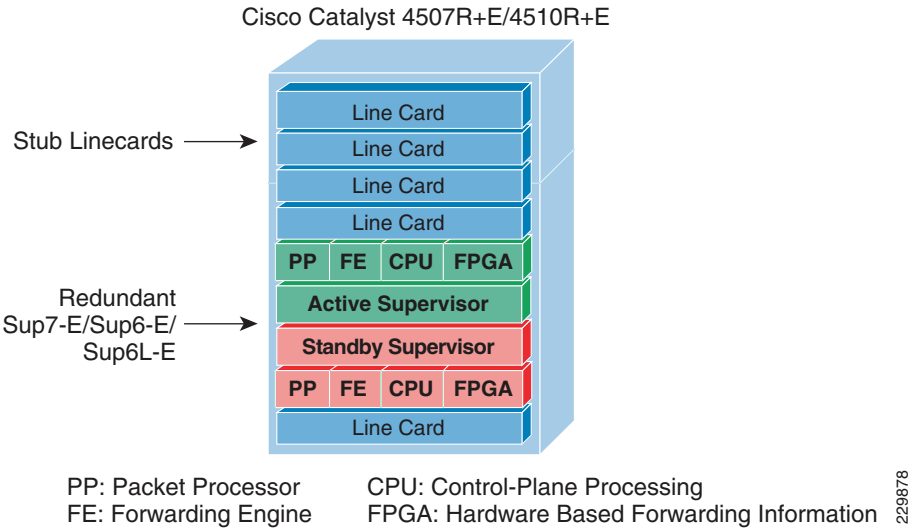
The Cisco Catalyst 4500E Series supports a wide-range of supervisor modules designed for high-performance Layer 2 and Layer 3 network connectivity. This reference design recommends deploying the next-generation Sup7-E to support borderless services in the campus access and distribution layers. The flexible and scalable architecture of the next-generation Sup7-E supervisor is designed to enable multiple innovations in the campus access and distribution layers for borderless services. The Catalyst 4500E with Sup7-E supervisor module is ideal for large enterprise networks that have dense wiring closets with a large number of end points that require power (PoE), constant network availability, wire-rate throughput, and low-latency switching performance for time-sensitive applications such as high-definition video conferencing. The Sup7-E operates on a new IOS software infrastructure and is designed to offer multiple hardware-assisted advanced technology benefits without compromising system performance.

Alternatively, the current-generation Sup6-E and Sup6L-E can also be deployed as they support hardware switching capabilities, scalability, and performance for various types of applications and services deployed in the campus network.

Implementing Redundant Supervisor

The Cisco Catalyst 4507R+E and 4510R+E models support intra-chassis or single-chassis supervisor redundancy with dual-supervisor support. Implementing a single Catalyst 4507R+E in highly-resilient mode at various campus layer with multiple redundant hardware components protects against different types of abnormal failures. This reference design guide recommends deploying redundant Sup7-E, Sup6-E, or Sup6L-E supervisor modules to deploy full high-availability feature parity. Therefore, implementing intra-chassis supervisor redundancy and initial network infrastructure setup will be simplified for small campus networks. Figure 17 illustrates the Cisco Catalyst 4500E-based intra-chassis SSO and NSF capability.

Figure 17 Intra-Chassis SSO Operation



During the bootup process, SSO synchronization checks various criteria to ensure both supervisors can provide consistent and transparent network services during failure events. If any of the criteria fail to match, it forces the standby supervisor to boot in RPR or a cold-standby state in which it cannot synchronize protocol and forwarding information from the active supervisor. The following sample configuration illustrates how to implement SSO mode on Catalyst 4507R+E and 4510R+E chassis deployed with Sup7-E, Sup6-E, and Sup6L-E redundant supervisors:

```
cr24-4507e-MB#config t
cr24-4507e-MB (config)#redundancy
cr24-4507e-MB (config-red)#mode sso

cr24-4507e-MB#show redundancy states
```

```
my state = 13 - ACTIVE
peer state = 8  - STANDBY HOT
    < snippet >
```

Deploying Supervisor Uplinks

Every supported supervisor module in the Catalyst 4500E supports different uplink port configurations for core network connectivity. The next-generation Sup7-E supervisor module offers unparalleled performance and bandwidth for the premium-class campus access layer. Each Sup7-E supervisor module can support up to four 1G or 10G non-blocking, wire-rate uplink connections to build high-speed distribution-access blocks. The current-generation Sup6-E and Sup6L-E supervisor modules support up two 10G uplinks or can be deployed as four different 1G uplinks using Twin-Gigabit converters. To build a high-speed, low-latency campus backbone network, it is recommended to leverage and deploy 10G uplinks to accommodate bandwidth-hungry network services and applications operating in the network.

All Cisco Catalyst 4500E Series supervisors are designed with unique architectures to provide constant network availability and reliability during supervisor resets. Even during supervisor soft-switchover or administrative reset events, the state machines of all deployed uplinks remain operational and with the centralized forwarding architecture they continue to switch packets without impacting any time-sensitive applications like high-definition video conferencing. This unique architecture protects bandwidth capacity while administrators perform supervisor IOS software upgrades or an abnormal event in software triggers a supervisor reset. Cisco recommends building diversified, distributed, and redundant uplink network paths as such designs offer the following benefits:

- Improve application performance by increasing aggregated network capacity with multiple high-speed 10Gbps uplinks in the access-distribution block.
- Enhance bi-directional traffic engineering with intelligent network data load sharing across all uplink physical ports.
- Improve system and application performance by utilizing the distributed architecture advantage of hardware resources—buffers, queue, TCAM, etc.
- Protect network-level redundancy and minimize congestion between distributed aggregation systems caused during a major outage at the access or distribution layer.

Sup7-E Uplink Port Design

Non-Redundant Mode

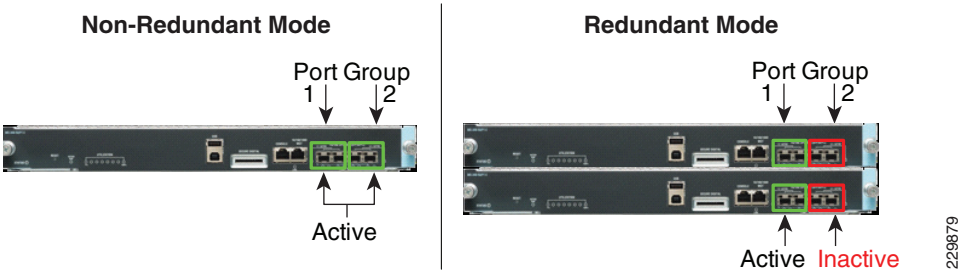
In non-redundant mode, there is a single Sup7-E supervisor module deployed in the Cisco Catalyst 4500E chassis. The four 1G/10G uplink ports on the Sup7-E modules are divided into two port groups—port group 1 and 2. Independent of 1G or 10G modes, both port groups and all four uplink ports are in an active state to use for uplink connections. All four uplink ports are non-blocking and can provide wire-rate performance of 4G/40G.

Redundant Mode

In redundant mode, the Catalyst 4507R or 4510R chassis are deployed with dual Sup7-E supervisor modules in a redundant configuration. Port group 2 becomes automatically inactive on both supervisor modules when the Catalyst 4500E system detects redundant modules installed in the chassis. It is recommend to utilize all four uplink ports from both supervisors if the uplink connections are made to a redundant system like the Catalyst 6500-E VSS. Both supervisor modules can equally diversify port group 1 with the redundant upstream system for the same consistent bandwidth capacity, load-balancing, and link redundancy as non-redundant mode.

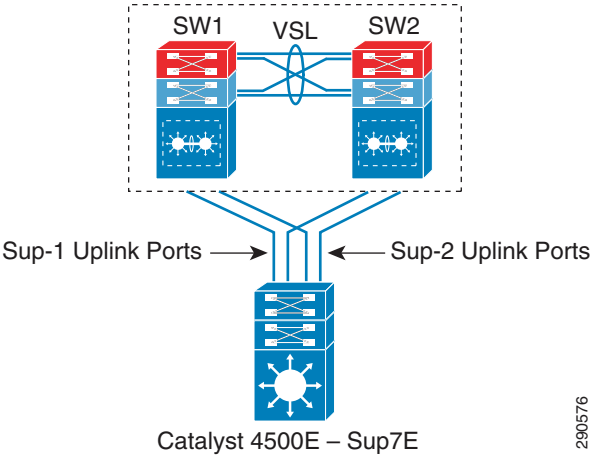
Figure 18 summarizes 1G/10G uplink port support on the next-generation Sup7-E in non-redundant and redundant deployment scenarios.

Figure 18 1G/10G Uplink Port Support on Sup7-E—Non-Redundant and Redundant Deployments



In a redundant mode configuration, the active interface from port group 1 from both supervisor modules should be utilized to build distributed and redundant uplink connections to the aggregation system. Diversified physical paths between redundant chassis and supervisor modules yields a resilient network design for coping with multiple fault conditions. Figure 19 illustrates the recommended uplink port design when the Cisco Catalyst 4500E is deployed with a redundant Sup7-E supervisor module.

Figure 19 Recommended 4500E Redundant-Sup Uplink Network Design



Sup6-E Uplink Port Design

Non-Redundant Mode

In non-redundant mode, there is a single supervisor module deployed in the Catalyst 4500E chassis. In non-redundant mode, by default both uplink physical ports can be deployed in 10G or 1G with Twin-Gigabit converters. Each port operates in a non-blocking state and can switch traffic at wire-rate performance.

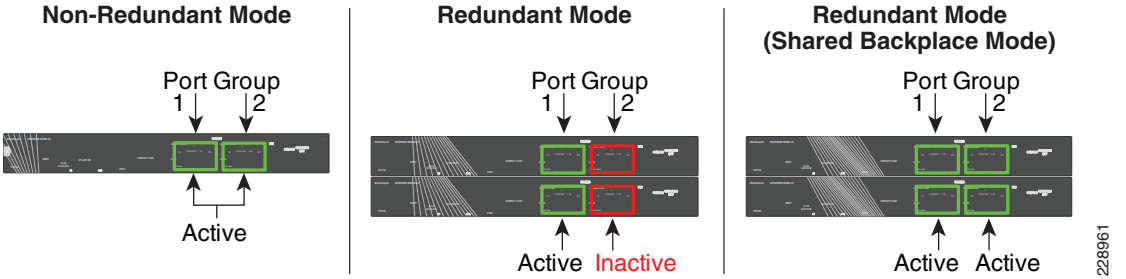
Redundant Mode

In the recommended redundant mode, the Catalyst 4500E chassis is deployed with dual supervisors. To provide wire-rate switching performance, by default port group 1 in both the active and hot-standby supervisors is in active mode and port group 2 is placed in an in-active state. The default configuration can be modified by changing the Catalyst 4500E backplane settings to sharing mode. Shared backplane mode enables the operation of port group 2 on both supervisors. Note that sharing the 10G backplane ASIC between the two 10G ports does not increase switching capacity; rather it creates 2:1 oversubscription. If the upstream device is deployed with chassis redundancy (i.e., Catalyst 6500-E VSS), then it is highly recommended to deploy all four uplink ports for the following reasons:

- Helps develop full-mesh or V-shape physical network topology from each supervisor module.
- Increases high availability in the network during an individual link, supervisor, or other hardware component failure event.
- Reduces latency and network congestion when rerouting traffic through a non-optimal path.

Figure 20 summarizes uplink port support on the Sup6-E module in non-redundant and redundant deployment scenarios.

Figure 20 Catalyst 4500E Sup6-E Uplink Mode



The next-generation supervisor Catalyst 4500E Sup7-E must be considered for aggregated wire-rate 40G performance. The following sample configuration provides instructions for modifying the default backplane settings on the Catalyst 4500E platform deployed with Sup6-E supervisors in redundant mode. The new backplane settings will be effective only after the chassis is reset; therefore, it is important to plan the downtime during this implementation:

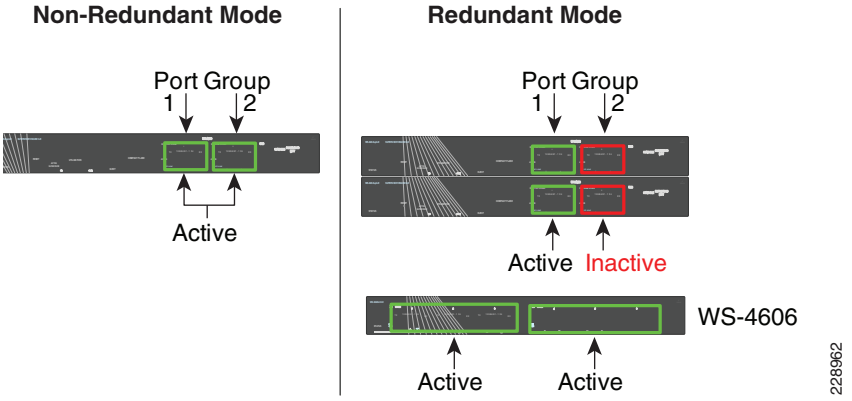
```
cr24-4507e-MB#config t
cr24-4507e-MB(config)#hw-module uplink mode shared-backplane

!A 'redundancy reload shelf' or power-cycle of chassis is required
! to apply the new configuration

cr24-4507e-MB#show hw-module uplink
 Active uplink mode configuration is Shared-backplane



cr24-4507e-MB#show hw-module mod 3 port-group
Module Port-group ActiveInactive
------------------------------------------------------------------------
   3         1      Te3/1-2Gi3/3-6

cr24-4507e-MB#show hw-module mod 4 port-group
Module Port-group ActiveInactive
------------------------------------------------------------------------
   4         1      Te4/1-2Gi4/3-6
```

The physical design of the Sup6-E uplink port should be same as that recommended for the Sup7-E (see Figure 19).

Sup6L-E Uplink Port Design

The Sup6L-E uplink port functions the same as the Sup6-E in non-redundant mode. However, in redundant mode the hardware design of the Sup6L-E differs from the Sup7-E, as the Sup6-E currently does not support a shared backplane mode that allows the active use of all uplink ports. The Catalyst 4500E deployed with the Sup6L-E may use the 10G uplinks of port group 1 from the active and standby supervisors when the upstream device is a single, highly-redundant Catalyst 4500E chassis. If the upstream device is deployed with chassis redundancy, (i.e., Cisco VSS), then it is recommended to build a full-mesh network design between each supervisor and each virtual switch node. For such a design, the network administrator must consider deploying the Sup7-E or Sup6-E supervisor module that supports four active supervisor uplink forwarding paths or deploying the 4500-E with Sup6L-E supervisor by leveraging the existing WS-4606 Series 10G linecard to build a full-mesh uplink. Figure 21 illustrates the deployment guidelines for a highly-resilient Catalyst 4500E-based Sup6L-E uplink.

Figure 21 Catalyst 4500E Sup6L-E Uplink Mode



Deploying Cisco Catalyst 3750-X StackWise Plus

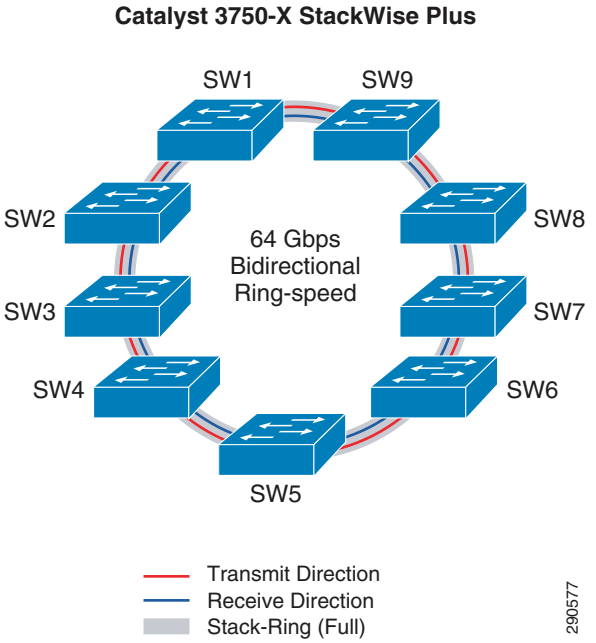
As the campus network edge expands, it becomes challenging for network administrators to manage and troubleshoot large numbers of devices and network access systems. In a best practice network design, wiring closet switches are deployed with a common type of end point so the administrator can implement consistent configurations on access network resources, e.g., user policies, global system settings, etc. Cisco Catalyst switches significantly simplify access layer management and operation with Cisco StackWise Plus technology, which is based on a high-speed stack ring that builds a hardware-based bi-directional physical ring topology.

The next-generation Layer 2/Layer 3 Cisco Catalyst 3750-X switch support Cisco StackWise Plus technology. Cisco StackWise Plus offers flexibility to expand access layer network scalability by stacking up to nine Catalyst 3750-X series switches into a single, logical access switch that significantly reduces control and management plane complexities. The StackPorts are system links to stack switches and are not traditional network ports, hence they do not run any Layer 2 network protocols (e.g., STP). To develop a virtual switch environment, each participating Cisco Catalyst 3750-X in a stack ring runs the Cisco proprietary stack protocol to keep network protocol communication, port state machine, and forwarding information synchronized across all the stack member switches.

Stack Design

Cisco StackWise Plus technology interconnect multiple Catalyst switches using a proprietary stack cable to build a single, logical access layer system. The Catalyst 3750-X series platform has two built-in stack ports to bi-directionally form a pair with stack member switches in a ring. The stack ring architecture is built with high-speed switching capacity; in full stack-ring configuration mode, the Cisco StackWise Plus architecture can provide maximum bi-directional attainable bandwidth up to 64 Gbps (32 Gbps Tx/Rx side). See Figure 22.

Figure 22 Recommended Cisco StackWise Plus Ring Design



Cisco Catalyst 3750-X - StackWisePlus Mode

```
cr37-3750X-1#show switch stack-ring speed
Stack Ring Speed        : 32G
Stack Ring Configuration: Full
Stack Ring Protocol     : StackWisePlus
!Displays unidirectional StackRing speed, current Ring state - Full/Half and Stack Mode
```

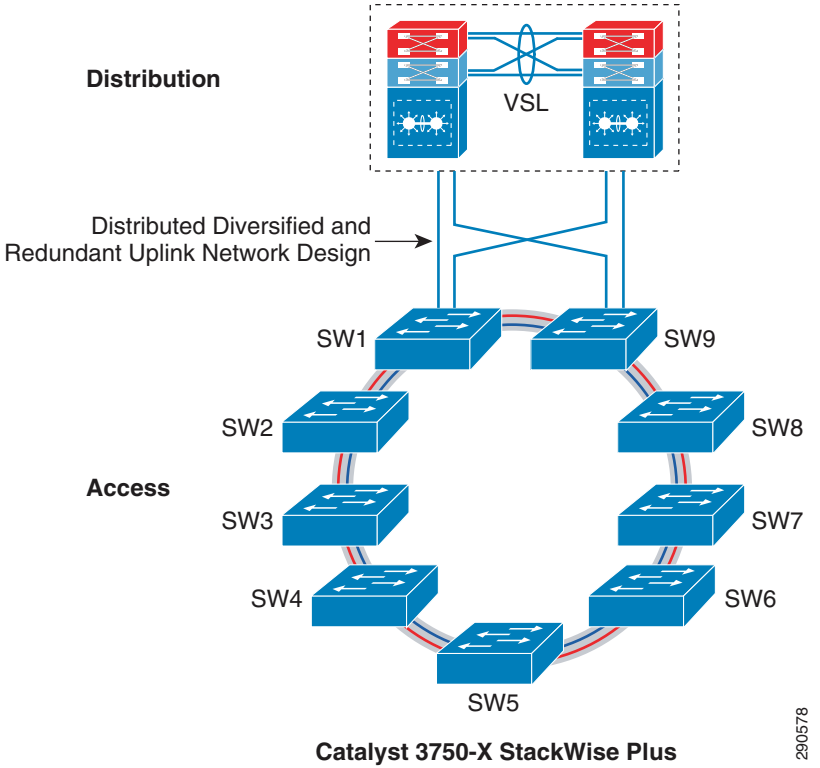
Uplink Port Design

The physical design in the distribution access block is important to build a high-speed, reliable, and scalable access layer network. In any best practice enterprise network design, network architects deploy redundant aggregation systems to load share network traffic and provide network redundancy during failures. Each access layer system physically builds redundant physical paths to each aggregation layer system. Cisco recommends maintaining these network fundamentals, even when Cisco system virtualization technologies like VSS or StackWise Plus are deployed. During distribution-access link failure, the alternate re-routing path between two distribution systems may become congested. Deploying a diverse and redundant physical network design in the access and distribution layer systems minimizes network congestion caused by the re-routing of data traffic.

Cisco Catalyst 3750-X series switches support two 10Gbps uplinks; it is recommended to utilize both uplinks ports even when these switches are deployed in stack configuration mode. The network administrator must identify and use the uplink port from the first switch of the stack-ring, i.e., Switch-1, and last switch of stack-ring, i.e., Switch-9, to physically connect each distribution layer system. Additional uplink ports from a stack member can be deployed if additional bandwidth capacity, load sharing, or path redundancy is required (see Figure 23). Cisco recommends building diversified, distributed, and redundant uplink network paths as this offers:

- Improved application performance by increasing aggregated stack switching capacity with multiple distributed high-speed 10Gbps uplinks between stack member Catalyst switches.
- Enhanced bi-directional traffic engineering with intelligent network data load sharing within the stack ring and across all distributed uplink physical ports.
- Improved system and application performance by utilizing the distributed forwarding architecture advantage of hardware resources—buffers, queue, TCAM, etc.
- Protect stack and network level redundancy and minimize congestion between distributed aggregation systems caused during a major outage at the access or distribution layer.

Figure 23 Recommended Cisco StackWisePlus Uplink Port Design



Unified Control Plane

The hardware architecture and internal software design of both stack technologies are different, however both offer consistent system design and deployment options. Cisco StackWise Plus provides a robust distributed forwarding architecture through each stack member switch and a unified, centralized control and management plane to simplify operation in a large-scale wiring closet network design. From the stack ring a single switch is elected into the master role and manages the centralized control plane process for all of the member switches. However each stack member switch provides distributed switching and network services likes QoS, security, etc. This distributed software design increases system resource capacity, prevents overload processing on the master switch, and optimizes stack ring bandwidth capacity. Refer to Chapter 1, "Borderless Campus Design and Deployment Models," Figure 1-22, for a physical versus logical view of a system in stack configuration mode.

Since the Cisco StackWise Plus solution offers high redundancy, it allows for a unique centralized control and management plane with a distributed forwarding architecture. To logically appear as a single virtual switch, the master switch manages all management plane and Layer 3 control plane operations (IP routing, CEF, PBR, etc.). Depending on the implemented network protocols, the master switch communicates with rest of the Layer 3 network through the stack ring and dynamically develops the global routing table and updates all member switches with distributed forwarding information.

Unlike the centralized Layer 3 management function on the master switch, the Layer 2 network topology development is completely based on a distributed design. Each member switch in the stack ring dynamically discovers MAC entries from the local port and uses the internal stack ring network to synchronize the MAC address table on each member switch in the stack ring. Table 2 lists the network protocols that are designed to operate in a centralized versus distributed model in the Cisco StackWise Plus architecture.

Table 2 Cisco StackWise Plus Centralized and Distributed Control-Plane

Protocols		Function
Layer 2 Protocols	MAC Table	Distributed
	Spanning-Tree Protocol	Distributed
	CDP	Centralized
	VLAN Database	Centralized
	EtherChannel - LACP	Centralized
Layer 3 Protocols	Layer 3 Management	Centralized
	Layer 3 Routing	Centralized

Using the stack ring as a backplane communication path, the master switch updates the Layer 3 forwarding information base (FIB) on each member switch in the stack ring. Synchronizing to a common FIB for member switches allows for a distributed forwarding architecture. With distributed forwarding information in the StackWise Plus software design, each stack member switch is designed to perform local forwarding information lookup to switch traffic instead of relying on the master switch, which may cause a traffic hair-pinning problem.

SSO Operation in 3750-X StackWise Plus

Device level redundancy in StackWise mode is achieved via stacking multiple switches using Cisco StackWise Plus technology. The Cisco StackWise Plus provides a 1:N redundancy model at the access layer. The master switch election in the stack ring is based on internal protocol negotiation. During the active master switch failure, the new master is selected based on a reelection process that takes place internally through the stack ring.

The Cisco StackWise Plus solution offers network and device resiliency with distributed forwarding, but the control plane is not designed in a 1+1 redundant model. This is because Cisco Catalyst 3750-X StackWise Plus switches are not SSO-capable platforms that can synchronize the control plane state machines to a standby switch in the ring. However, it can be configured in NSF-capable mode to gracefully recover from a master switch failure. Therefore, when a master switch failure occurs, all the Layer 3 functions that are deployed on the uplink ports may be disrupted until a new master election occurs and reforms Layer 3 adjacencies. Although the new master switch in the stack ring identification is performed within 0.7 to 1 second, the amount of time for rebuild the network and forwarding topology depends on the protocol's function and scalability.

To prevent Layer 3 disruptions in the network caused by a master switch failure, the elected master switch with the higher switch priority can be isolated from the uplink Layer 3 EtherChannel bundle path and use physical ports from switches in the member role. With the Non-Stop Forwarding (NSF) capabilities in the Cisco StackWise Plus architecture, this network design helps to decrease major network downtime during master switch failure.

Implementing StackWise Plus Mode

As described earlier, the Cisco Catalyst 3750-X switch dynamically detects and provisions member switches in the stack ring without any extra configuration. For a planned deployment, the network administrator can pre-provision the switch in the ring with the following configuration in global configuration mode. Pre-provisioning the switch in the network provides the network administrator with the flexibility to configure future ports and enable borderless services immediately when they are deployed:

```
cr36-3750x-xSB(config)#switch 3 provision ws-3750x-48p

cr36-3750x-xSB#show running-config | include interface GigabitEthernet3/
interface GigabitEthernet3/0/1
interface GigabitEthernet3/0/2
```

Switch Priority

The centralized control plane and management plane is managed by the master switch in the stack. By default, the master switch selection within the ring is performed dynamically by negotiating several parameters and capabilities between each switch within the stack. Each StackWise-capable member switch is by default configured with switch priority 1.

```
cr36-3750x-xSB#show switch
Switch/Stack Mac Address : 0023.eb7b.e580
    H/W Current
Switch#     Role    Mac Address        Priority    Version     State
-----------------------------------------------------------------------------------
* 1         Master  0023.eb7b.e580     1           0           Ready
   2        Member  0026.5284.ec80     1           0           Ready
```

As described in a previous section, the Cisco StackWise architecture is not SSO-capable. This means all of the centralized Layer 3 functions must be reestablished with the neighbor switch during a master switch outage. To minimize control plane impact and improve network convergence, the Layer 3 uplinks should be diverse, originating from member switches instead of the master switch. The default switch priority must be increased manually after identifying the master switch and switch number. The new switch priority becomes effective after switch reset.

```
cr36-3750x-xSB (config)#switch 1 priority 15
Changing the Switch Priority of Switch Number 1 to 15
cr36-3750x-xSB (config)#switch 2 priority 14
Changing the Switch Priority of Switch Number 2 to 14

cr36-3750x-xSB # show switch
Switch/Stack Mac Address : 0023.eb7b.e580

    H/W Current
Switch#     Role    Mac Address        Priority    Version     State
-----------------------------------------------------------------------------------
* 1         Master  0023.eb7b.e580     15          0           Ready
2           Member  0026.5284.ec80     14          0           Ready
```

Stack-MAC Address

To provide a single unified logical network view in the network, the MAC addresses of Layer 3 interfaces on the StackWise (physical, logical, SVIs, port channel) are derived from the Ethernet MAC address pool of the master switch in the stack. All Layer 3 communication from the StackWise switch to the endpoints (such as IP phones, PCs, servers, and core network system) is based on the MAC address pool of the master switch.

```
cr36-3750x-xSB#show switch
Switch/Stack Mac Address : 0023.eb7b.e580
    H/W Current
Switch#     Role    Mac Address        Priority    Version     State
-----------------------------------------------------------------------------------
* 1         Master  0023.eb7b.e580     15          0           Ready
2           Member  0026.5284.ec80     14          0           Ready

cr36-3750s-xSB #show version
```

```
.   .   .
Base ethernet MAC Address       : 00:23:EB:7B:E5:80
.   .   .
```

To prevent network instability, the old MAC address assignments on Layer 3 interfaces can be retained even after the master switch fails. The new active master switch can continue to use the MAC addresses assigned by the old master switch, which prevents ARP and routing outages in the network. The default **stack-mac timer** settings must be changed in Catalyst 3750-X StackWise switch mode using the global configuration CLI mode as shown below:

```
cr36-3750x-xSB (config)#stack-mac persistent timer 0
cr36-3750x-xSB #show switch
Switch/Stack Mac Address : 0026.5284.ec80
Mac persistency wait time: Indefinite
                                           H/WCurrent
Switch#    Role    Mac Address     Priority   Version    State
-----------------------------------------------------------------------------
* 1        Master  0023.eb7b.e580     15          0          Ready
2          Member  0026.5284.ec80     14          0          Ready
```

Deploying Cisco Catalyst 3750-X and 3560-X

The Borderless Campus design recommends deploying fixed or standalone configuration Cisco Catalyst 3750-X and 3560-X Series platforms at the campus network edge. The hardware architecture of access layer fixed configuration switches is standalone and non-modular in design. These switches are designed to go above the traditional access layer switching function to provide robust next-generation network services (edge security, PoE+ EnergyWise, etc.).

Deploying Cisco Catalyst switches in standalone configuration mode does not require any system-specific configuration as they are ready to be deployed at the access layer with their default settings. All recommended access layer features and configuration are explained in the following sections.

Designing the Campus LAN Network

In this reference design, multiple parallel physical paths are recommended to build a highly-scalable and resilient foundation for an enterprise campus network. Depending on the system design and implementation, the default network configuration requires each network configuration, protocol adjacencies, and forwarding information on a per-interface basis to load share traffic and provide network redundancy.

Distributed Physical Path Network Design

As a general best practice to build resilient network designs, it is highly recommended to interconnect all network systems with full-mesh diverse physical paths. This network design automatically creates multiple parallel paths to provide load-sharing capabilities and path redundancy during network fault events. Deploying a single physical connection from any standalone campus system to separate redundant upstream systems creates a "triangle"-shaped physical network design that is an optimal and redundant design, in contrast to the non-recommended partial-mesh "square" physical network design. Utilizing the hardware resources for network connectivity varies based on the system type and the capabilities that are deployed in each campus layer.

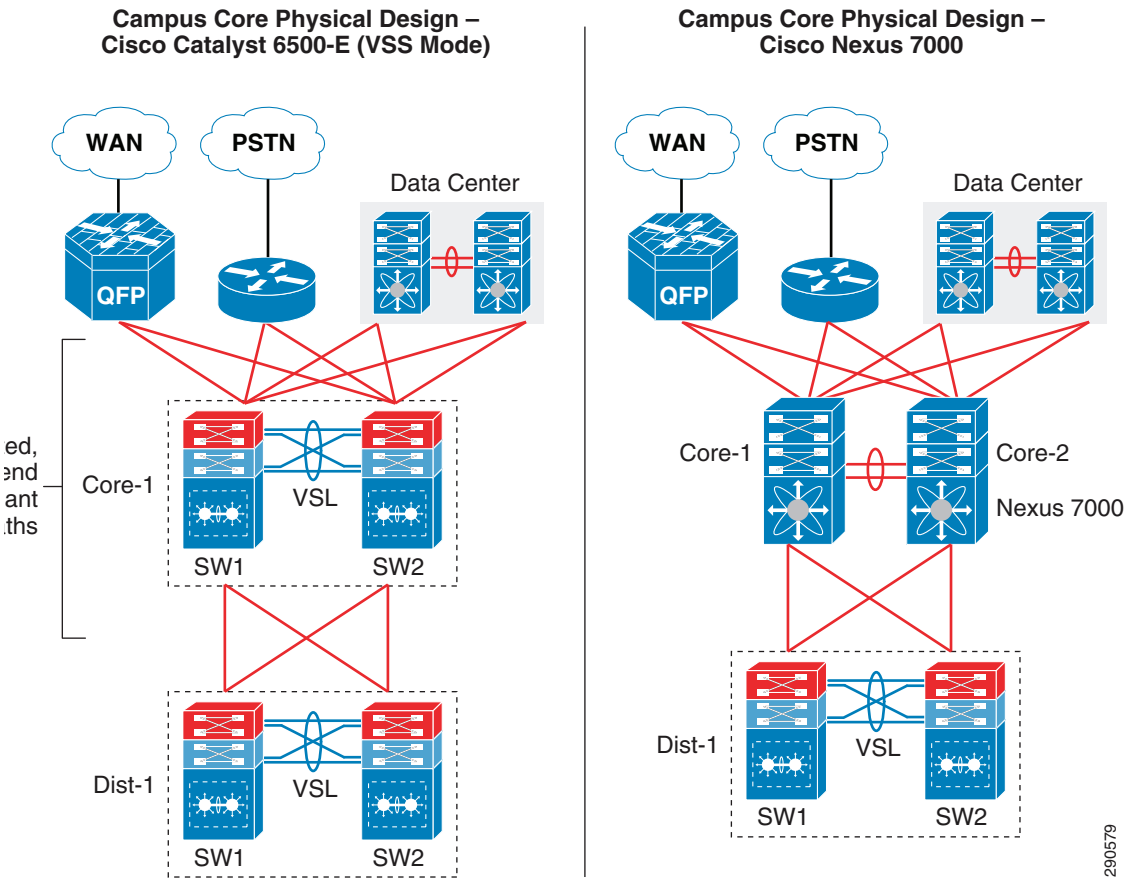
Access Layer

The access layer Cisco Catalyst switching systems provide flexibility to scale network edge ports and have multiple high-speed 1G/10G uplink physical ports for network connectivity. The number of uplink ports may vary based on multiple factors, e.g., system type (standalone versus stack mode or redundant versus non-redundant supervisor). Independent of system type and deployment mode, Cisco recommends building full-mesh redundant physical paths in each design. Refer to Deploying Cisco Catalyst 4500E and Deploying Cisco Catalyst 3750-X StackWise Plus for uplink port design recommendation.

Distribution Core Layer

Campus distribution and core layer systems typically have modular hardware with centralized processing on a supervisor module and distributed forwarding on high-speed network modules. To enable end-to-end borderless network services, the campus core layer system interconnects to several sub-blocks of the network—data center, WAN edge, services block, etc. The fundamentals of building a resilient campus network design with diversified, distributed, and redundant physical paths does not vary by role, system, or deployed configuration mode. Figure 1-24 illustrates full-mesh redundant physical connections between a wide range of Cisco devices deployed in different campus, edge, and data center network tiers:

Figure 1-24 Redundant Campus Core and Edge Physical Path Design



Optimizing Campus Network Operation

While multiple redundant paths to redundant systems provides various benefits, it also introduces network design and operational challenges. Each system builds multiple protocol adjacencies, routing topologies, and forwarding information through each individual Layer 3 physical path. In a multilayer design, the distribution access layer network may operate at reduced capacity based on a loop-free Layer 2 network topology. Based on the recommended full-mesh campus LAN design and Cisco's system virtualization technique, this guide provides guidance to simplify overall campus network operation.

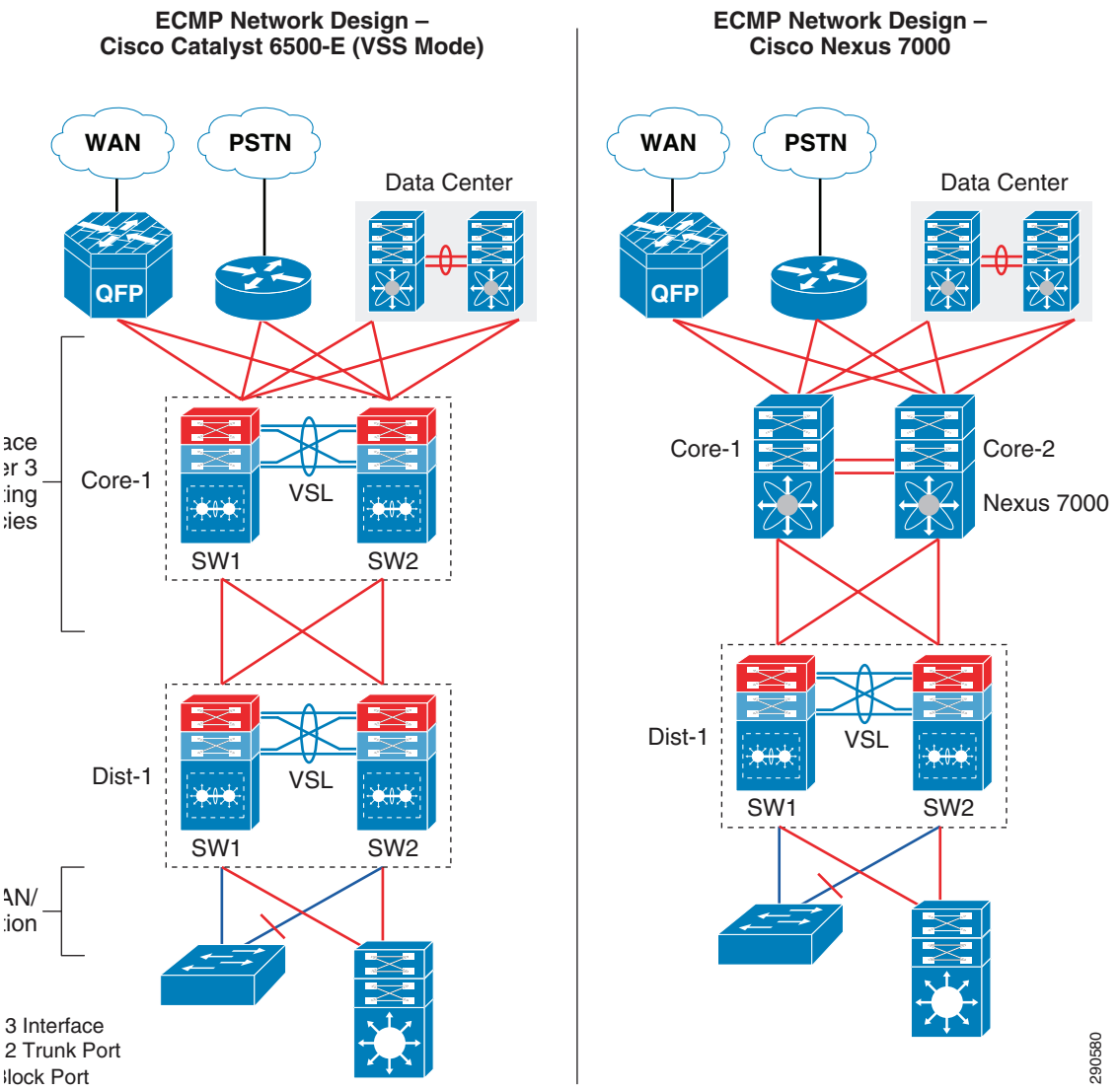
Based on the campus network layer, system capability, and physical layout, the campus network can be designed and deployed in two different models—Equal Cost Multi Path (ECMP) and EtherChannel technology. Both deployment models can co-exist in a campus network, however Cisco recommends simplifying the campus network with EtherChannel when possible. Equal Cost Multi Path Network Design provides design and deployment guidance for both modes.

Equal Cost Multi Path Network Design

In simple terms, when a campus system develops best-path routing and forwarding tables to the same destination address through multiple next-hop devices, it is known as ECMP. In such a network design, the campus backbone network provides data traffic load balancing and path redundancy over full-mesh physical connections between each system. An ECMP network requires its own set of network policies, configuration, and tuning to develop distributed forwarding information across each physical connection.

The campus distribution and core layer systems build parallel Layer 3 adjacencies to build a redundant routing topology and forwarding databases to load share traffic and provide redundancy. At the distribution-access layer block, the Layer 2 network design is recommend only in standalone and non-redundant access and distribution layer network designs. Deploying a Layer 2 network design in redundant (dual-sup) or virtual-switch (VSS/StackWise Plus) mode may build a sub-optimal loop-free network topology and operate at reduced capacity (see Figure 1-25). Figure 1-25 demonstrates the default Layer 2 and Layer 3 network design with redundant and complex control plane operation with an under-utilized Layer 2 forwarding plane design.

Figure 1-25 ECMP-based Campus Network Design



The design in Figure 1-25 be optimized with the recommended EtherChannel-based campus design to solve the following challenges for different network modes:

- *Layer 3*—Multiple routing adjacencies between two Layer 3 systems. This configuration doubles or quadruples the control plane load between each of the Layer 3 devices. It also adds more overhead by using more system resources, such as CPU and memory to store redundant dynamic routing information with different Layer 3 next-hop addresses connected to the same router. It develops Equal Cost Multi Path (ECMP) symmetric forwarding paths between the same Layer 3 peers and offers network scale-dependent, Cisco CEF-based network recovery.
- *Layer 2*—Multiple parallel Layer 2 paths between the STP Root (distribution) and the access switch will create a network loop. To a build loop-free network topology, the STP blocks the non-preferred individual link path from entering into a forwarding state. With a single STP root virtual switch, such network topologies cannot fully use all the network resources and creates a non-optimal and asymmetric traffic forwarding design.
- *VSL Link Utilization*—In a Cisco VSS-based distribution network, it is highly recommended to not create a hardware or network protocol-driven asymmetric forwarding design (e.g., single-home connection or STP block port). As described in Deploying Cisco Catalyst 6500-E in VSS Mode, VSL is not a regular network port; it is a special inter-chassis backplane connection used to build the virtual system and the network must be designed to switch traffic across VSL-only as a last resort.

ECMP Load-Sharing

In an ECMP-based campus network design, each campus system load shares the network data traffic based on multiple variables to utilize all symmetric forwarding paths installed in the routing information base (RIB). To load share traffic in a hardware path, each switch independently computes hash based on a variety of input information to program forwarding information in the hardware. Cisco Catalyst switches running IOS software leverage the Cisco Express Forwarding (CEF) switching algorithm for hardware-based destination lookup to provide wire-speed switching performance for a large-scale network design. Independent of IGP type, CEF uses the routing table to pre-program the forwarding path and build the forwarding information base (FIB) table. FIB installs multiple adjacencies if there are multiple paths in RIB. For wire-speed packet switching, the CEF also uses an adjacency table that contains Layer 2 egress header and rewrite encapsulation information.

Per-Flow Load Sharing

To efficiently distribute network traffic across all ECMP paths, Cisco IOS and NX-OS builds deterministic load share paths based on Layer 3 addresses (source-destination flow) and provides the ability to include Layer 4 port input information. By default the Cisco Catalyst and the Nexus 7000 system perform ECMP load balancing on a per-flow basis. Hence based on the locally-computed hash result, a flow may always take the same physical path to transmit egress traffic unless a preferred path from hash result is unavailable. In a large enterprise campus network design, it is assumed there is a good statistical mix of IP hosts, networks, and applications that allows aggregation and core layer systems to optimally load share traffic across all physical paths. While per-destination load sharing

balances network traffic across multiple egress paths, it may not operate in deterministic mode to evenly distribute traffic across each link. However the load sharing mechanism ensures the flow pattern and path is maintained as it traverses through each campus network hop.

As every enterprise campus network runs a wide-range of applications, it is complex to unify and recommend a single ECMP load-balancing mechanism for all network designs. The network administrator may modify default settings to include source or destination IP address and Layer 4 port information for ECMP load balance hash computation if the campus system does not load balance traffic across all paths:

Catalyst 6500

```
Dist-VSS(config)#mls ip cef load-sharing full
```

Catalyst 4500E

```
cr19-4500-1(config)#ip cef load-sharing algorithm include-ports source destination
cr19-4500-1#show cef state | inc include
 include-ports source destination per-destination load sharing algorithm
```

Nexus 7000

By default Cisco Nexus 7000 includes Layer 3 and Layer 4 port information in ECMP load sharing and it can be verified with following **show** command:

```
cr35-N7K-Core1# show ip load-sharing
IPv4/IPv6 ECMP load sharing:
Universal-id (Random Seed): 3839634274
Load-share mode : address source-destination port source-destination
```

Per-Packet Load Sharing

Cisco CEF per-packet load sharing ensures that each ECMP path is evenly utilized and minimizes the under and overload link conditions caused by certain heavy data flows. The Cisco Catalyst switching system does not support per-packet load sharing. The Cisco Nexus 7000 can perform per-packet load sharing on a per-flow basis, i.e., same source and destination address. Per-packet load sharing cannot be done on every egress packet that contains a different source or destination network address. If a flow takes a separate core switching path, it is possible to receive packets out-of-order, which may impact network and application performance. As the campus core system switches data traffic for a wide-range of business critical communications and applications, it is recommended to retain default per-destination ECMP load sharing to minimize application and network instability.

EtherChannel Network Design

Traditionally campus networks were designed with standalone networks systems and ECMP, which did not, provided the flexibility to simplify network design with redundant devices or paths to logically act as a single entity. Campus network designs are evolving with Cisco's system virtualization innovation in Catalyst switching platforms, such as VSS, StackWise Plus, or FlexStack, with redesign opportunities in all three tiers. While each of these virtualization techniques clusters multiple physical

systems into a single large and unified logical system, the distributed multiple parallel physical paths can now be bonded into logical point-to-point EtherChannel interfaces between two systems. The principle of building a full-mesh physical campus network should not be changed when a campus device or link are implemented to operate in logical mode.

Designing a multilayer or campus backbone network with EtherChannel between two systems offers multiple benefits:

- Simplified—Bundling multiple ECMP paths into logical EtherChannel reduces redundant protocol adjacencies, routing databases, and forwarding paths.
- Optimize—Reduces the number of control plane operations and optimizes system resources, such as CPU/memory utilization, to store single upstream and downstream dynamic routing information instead of multiple versions without ECMP. Provides flexible Layer 2 to Layer 4 variables to intelligently load share and utilize all resources for network data traffic across each bundled interface.
- Reduce complexity—Simplifies network operational practice and reduces the amount of network configuration and troubleshooting required to analyze and debug problems.
- Increase capacity—Eliminates protocol-driven restrictions and doubles switching capacity in multilayer designs by utilizing all resources (bandwidth, queue, buffer, etc.) across all bundled Layer 2 uplinks.
- Resilient—Provides deterministic hardware-driven network recovery for seamless business operation. Minimizes routing database re-computation and topology changes during minor network faults, e.g., link failure.

In a standalone EtherChannel mode, multiple and diversified member links are physically connected in parallel between two physical systems. All the key network devices in the Borderless Campus design guide support EtherChannel technology. Independent of location in the network, whether in a campus layer, data center, or the WAN/Internet edge, the EtherChannel fundamentals and configuration guidelines described in this section remain consistent.

Multi-Chassis EtherChannel Fundamentals

Cisco's Multi-Chassis EtherChannel (MEC) technology is a breakthrough innovation that creates logical point-to-point EtherChannels distributed across mulitple physical switches, which allows for a highly-resilient virtual switch in the VSS domain. Deploying Layer 2 or Layer 3 MEC with VSS introduces the following benefits:

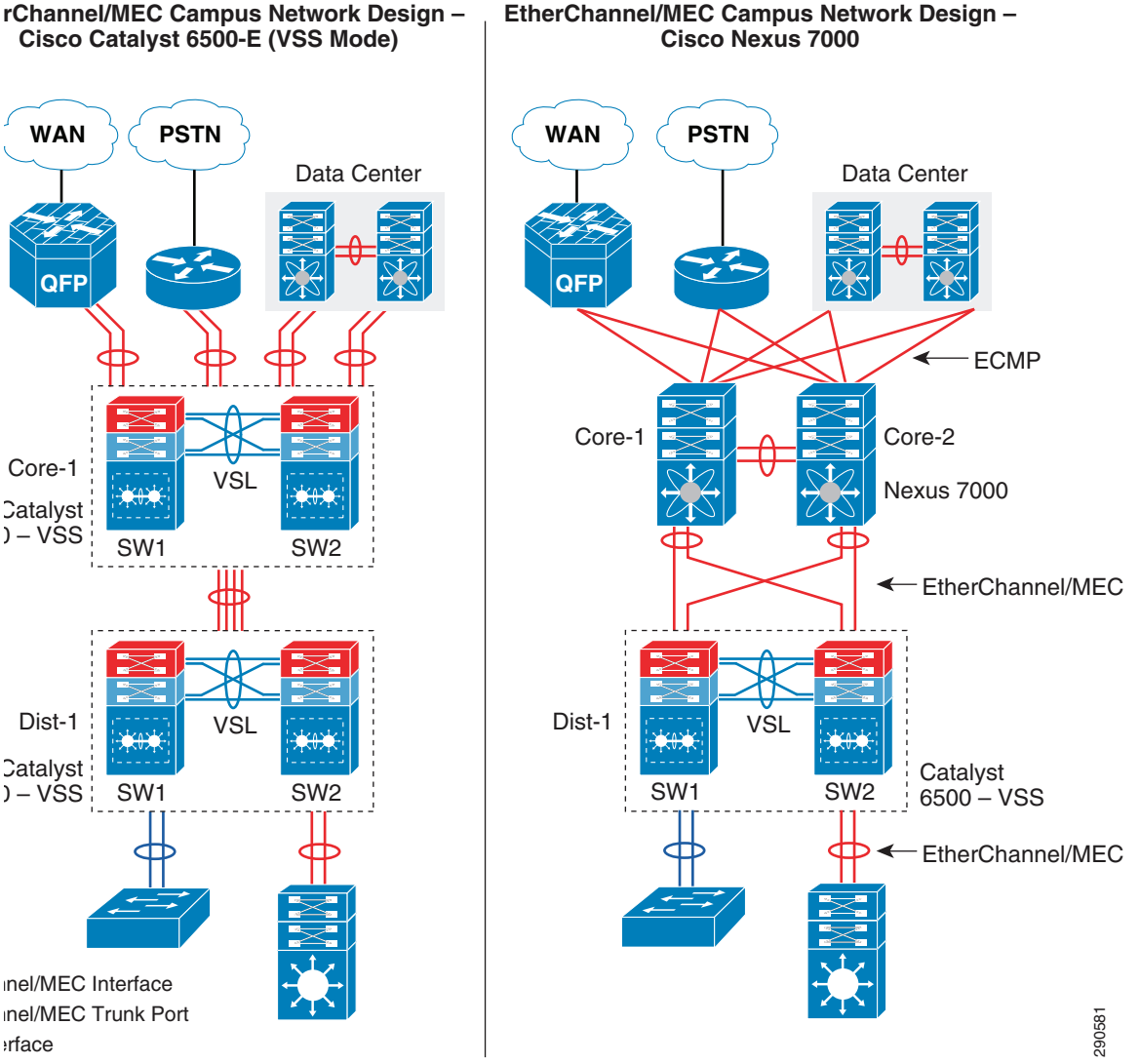
- In addition to all EtherChannel benefits, the distributed forwarding architecture in MEC helps increase network bandwidth capacity.
- Increases network reliability by eliminating the single point-of-failure limitation compared to traditional EtherChannel technology.
- Simplifies the network control plane, topology, and system resources within a single logical bundled interface instead of multiple individual parallel physical paths.

- Independent of network scalability, MEC provides deterministic hardware-based subsecond network recovery.
- MEC technology on the Catalyst 6500 in VSS mode remains transparent to remote peer devices.

Cisco recommends designing the campus network to bundle parallel paths into logical EtherChannel or MEC in all layers when possible. The campus network can be deployed in a hybrid EtherChannel and ECMP network design if any device cannot logically bind interfaces.

Figure 1-26 illustrates the recommended end-to-end EtherChannel/MEC and hybrid-based borderless campus network design that simplifies end-to-end network control plane operation.

Figure 1-26 Recommended EtherChannel/MEC-based Campus Network Design



Implementing EtherChannel

In standalone EtherChannel mode, multiple and diversified member links are physically connected in parallel between two physical systems. All the key network devices in the Borderless Campus design support EtherChannel technology. Independent of campus location and network layer—campus, data center, WAN/Internet edge—all the EtherChannel fundamentals and configuration guidelines described in this section remain consistent.

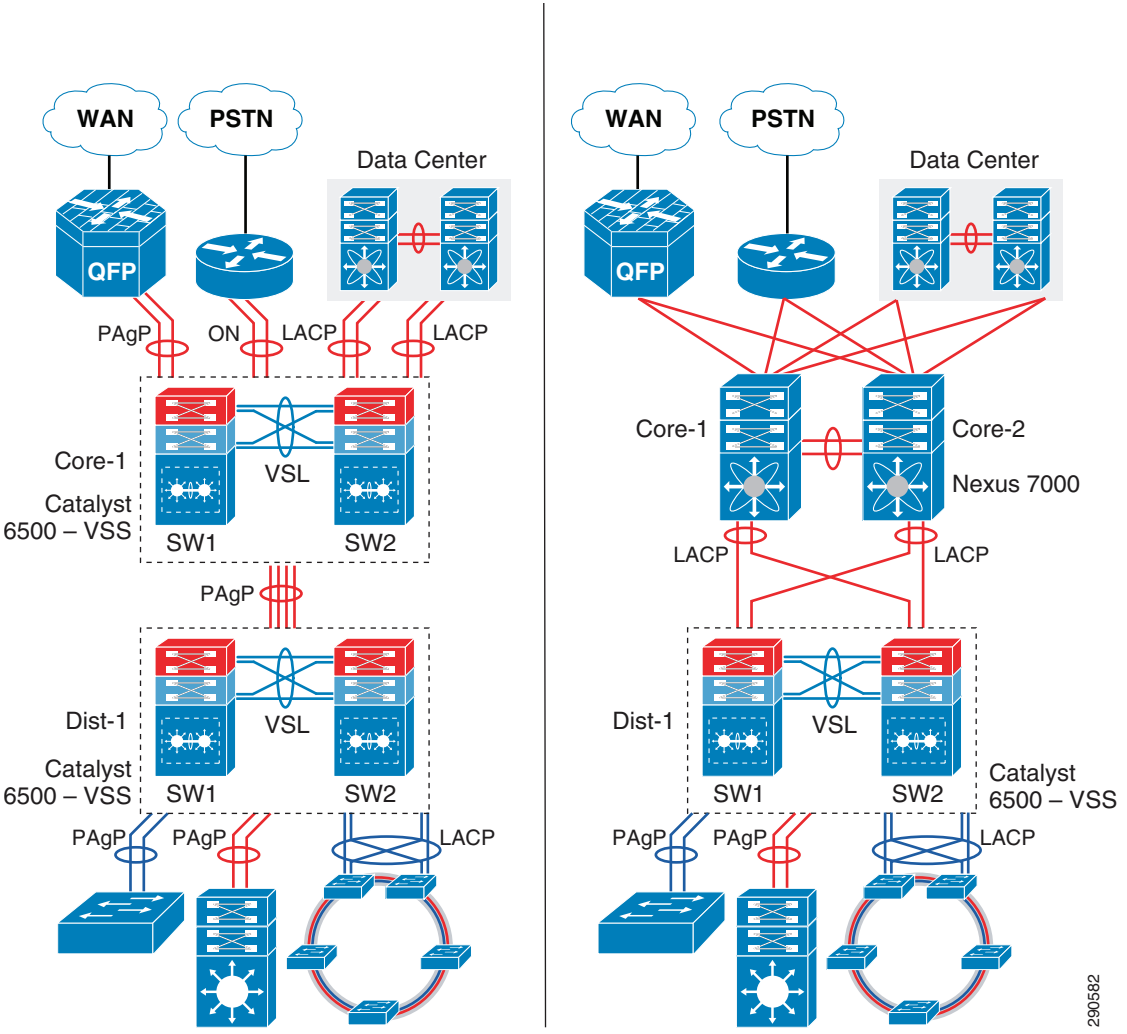
Port-Aggregation Protocols

The member links of EtherChannel must join the port channel interface using Cisco PAgP+ or industry-standard LACP port aggregation protocols. Both protocols are designed to provide consistent link capabilities, however the Cisco PAgP+ protocol provides an additional solution advantage, dual-active detection. Implementing these protocols provides the following additional benefits:

- Ensures link aggregation parameter consistency and compatibility between two the systems.
- Ensures compliance with aggregation requirements.
- Dynamically reacts to runtime changes and failures on local and remote Etherchannel systems.
- Detects and removes unidirectional links and multidrop connections from the Etherchannel bundle.

All Cisco Catalyst switching platforms support Cisco PAgP+ and industry-standard LACP protocols. If Cisco or any other vendor campus system does not support PAgP, it should be deployed with the industry-standard LACP link bundling protocol. Figure 27 illustrates configuration guidelines for Cisco PAgP+ or LACP protocols between two systems based on their software capabilities.

Figure 27 Network-Wide Port-Aggregation Protocol Deployment Guidelines



Port aggregation protocol support varies on the various Cisco platforms. Depending on each end of the EtherChannel device types, Cisco recommends deploying the port channel settings specified in Table 3.

Table 3 MEC Port-Aggregation Protocol Recommendation

Port-Agg Protocol	Local Node	Remote Node	Bundle State
PAgP+	Desirable	Desirable	Operational
LACP	Active	Active	Operational
None [1]	ON	ON	Operational

1. None or Static Mode EtherChannel configuration must be deployed in exceptional cases when remote nodes do not support either of the port aggregation protocols. To prevent network instability, the network administrator must implement static mode port channel with special attention that ensures no configuration incompatibility between bundled member link ports.

The implementation guidelines to deploy EtherChannel and MEC in Layer 2 or Layer 3 mode on all campus Cisco IOS and NX-OS operating systems are simple and consistent. The following sample configuration illustrates implementation of point-to-point Layer 3 EtherChannel or MEC on the Cisco Catalyst and the Nexus 7000 series systems:

- Cisco IOS

```
cr23-VSS-Core(config)#interface Port-channel 102
cr23-VSS-Core(config-if)# ip address 10.125.0.14 255.255.255.254
! Bundling single MEC diversed physical ports and module on per node basis.
cr23-VSS-Core(config)#interface range Ten1/1/3 , Ten1/3/3 , Ten2/1/3 , Ten2/3/3
cr23-VSS-Core(config-if-range)#channel-protocol pagp
cr23-VSS-Core(config-if-range)#channel-group 102 mode desirable

cr23-VSS-Core#show etherchannel 102 summary | inc Te
102    Po102(RU)       PAgP      Te1/1/3(P)     Te1/3/3(P)     Te2/1/3(P)   Te2/3/3(P)
cr23-VSS-Core#show pagp 102 neighbor | inc Te
Te1/1/3   cr24-4507e-MB        0021.d8f5.45c0   Te4/2       27s SC      10001
Te1/3/3   cr24-4507e-MB        0021.d8f5.45c0   Te3/1       28s SC      10001
Te2/1/3   cr24-4507e-MB        0021.d8f5.45c0   Te4/1       11s SC      10001
Te2/3/3   cr24-4507e-MB        0021.d8f5.45c0   Te3/2       11s SC      10001
```

- Cisco NX-OS

```
cr35-N7K-Core1(config)# interface port-channel101
cr35-N7K-Core1(config-if)#description Connected to Dist-VSS
cr35-N7K-Core1(config-if)#ip address 10.125.10.1/31

cr35-N7K-Core1(config-if)# interface Eth1/2 , Eth2/2
cr35-N7K-Core1(config-if-range)#channel-group 101 mode active
! Bundling single EtherChannel diversed between I/O Modules.
```

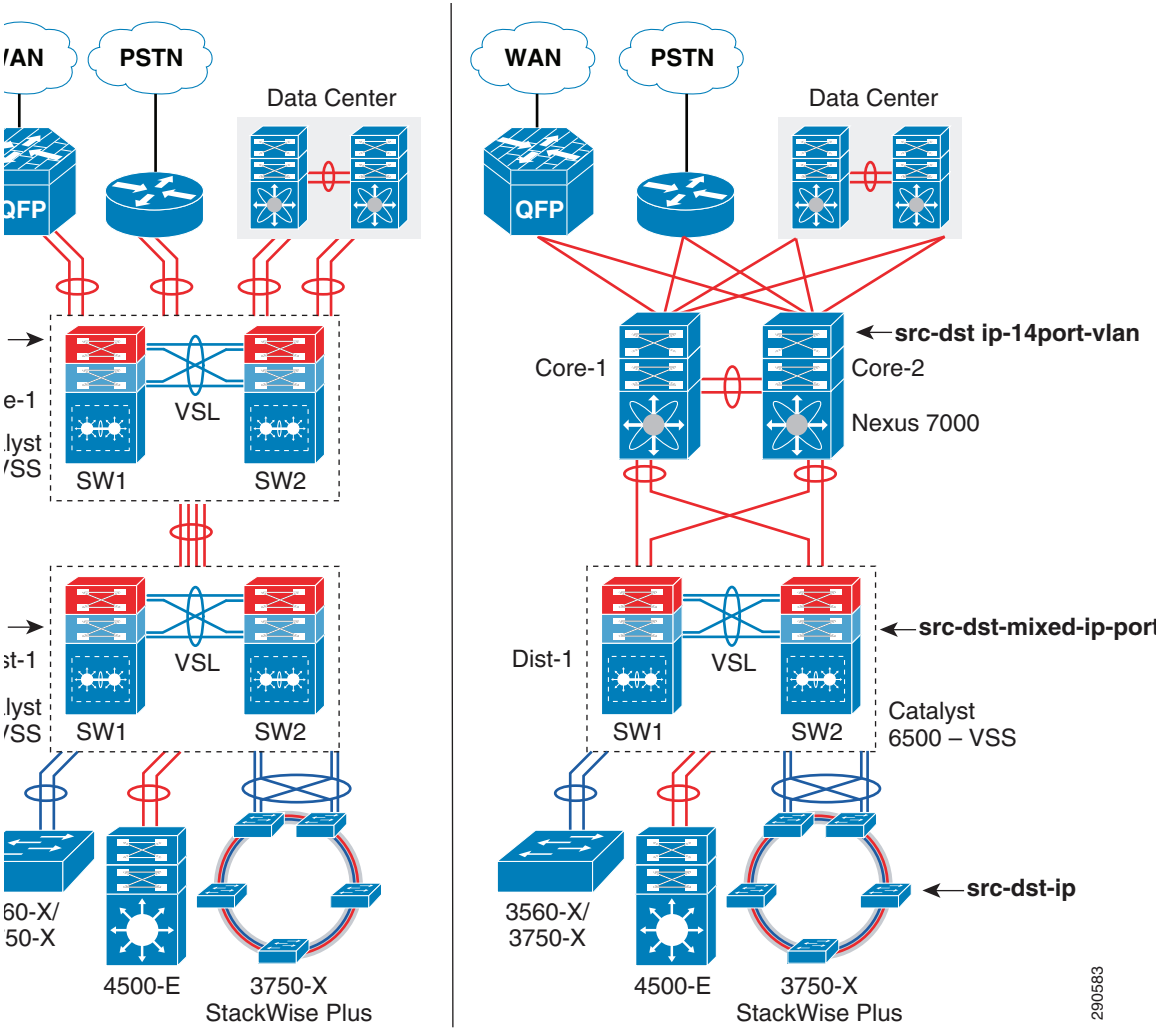
```
cr35-N7K-Core1#show port-channel summary interface port-channel 101 | inc Eth
101   Po101(RU)   Eth      LACP      Eth1/2(P)    Eth2/2(P)

cr35-N7K-Core1#show lacp port-channel interface port-channel 101
port-channel101
  System Mac=f8-66-f2-e-17-41
  Local System Identifier=0x8000,f8-66-f2-e-17-41
  Admin key=0x64
  Operational key=0x64
  Partner System Identifier=0x8000,2-0-0-0-0-1
  Operational key=0x3
  Max delay=0
  Aggregate or individual=1
  Member Port List=2
```

EtherChannel Load-Sharing

The number of applications and their functions in a campus network design is highly variable, especially when the network is provided as a common platform for business operation, campus security, and open accessibility. It is important for the network to become more intelligence-aware, with deep packet inspection and load sharing traffic by fully using all network resources.

Fine tuning EtherChannel and MEC adds extra computing intelligence to the network to make protocol-aware egress forwarding decisions between multiple local member links paths. For each traffic flow, such tuning optimizes the egress path selection procedure with multiple levels of variable information that originated from the source host (i.e., Layer 2 to Layer 4). EtherChannel load-balancing methods vary on Cisco Catalyst and Nexus 7000 switching and routing platforms. Figure 1-28 illustrates the recommended end-to-end Etherchannel load balancing method.

Figure 1-28 Recommended EtherChannel Load Balancing Method



Note For optimal traffic load sharing between Layer 2 or Layer 3 member links, it is recommended to bundle the number of member links in powers of 2 (i.e., 2, 4, and 8).

Implementing EtherChannel Load-Sharing

EtherChannel load sharing is based on a polymorphic algorithm. On a per-protocol basis, load sharing is done based on source XOR destination address or port from Layer 2 to 4 header and ports. For higher granularity and optimal utilization of each member link port, an EtherChannel can intelligently load-share egress traffic using different algorithms.

All Cisco Catalyst switching and Cisco Nexus 7000 systems must be tuned with optimal EtherChannel load sharing capabilities similar to the following sample configuration:

Catalyst 3xxx and 4500-E

```
cr24-4507e-MB(config)#port-channel load-balance src-dst-ip
cr24-4507e-MB#show etherchannel load-balance
EtherChannel Load-Balancing Configuration:
        src-dst-ip
```

Cisco Nexus 6500-E

```
cr23-VSS-Core(config)#port-channel load-balance src-dst-mixed-ip-port
cr23-VSS-Core#show etherchannel load-balance
EtherChannel Load-Balancing Configuration:
        src-dst-mixed-ip-port vlan included
```

Cisco Nexus 7000

```
cr35-N7K-Core1(config)#port-channel load-balance src-dst ip-l4port-vlan
cr35-N7K-Core1#show port-channel load-balance
Port Channel Load-Balancing Configuration:
System: src-dst ip-l4port-vlan
```

Implementing MEC Load-Sharing

The next-generation Catalyst 6500-E Sup720-10G supervisor introduces more intelligence and flexibility to load share traffic with up to 13 different traffic patterns. Independent of virtual switch role, each node in VSD uses the same polymorphic algorithm to load share egress Layer 2 or Layer 3 traffic across different member links from the local chassis. When computing the load-sharing hash, each virtual switch node includes the local physical ports of MEC instead of remote switch ports; this customized load sharing is designed to prevent traffic reroute over the VSL. It is recommended to implement the following MEC load sharing configuration in global configuration mode:

```
cr23-VSS-Core(config)#port-channel load-balance src-dst-mixed-ip-port

cr23-VSS-Core#show etherchannel load-balance
EtherChannel Load-Balancing Configuration:
        src-dst-mixed-ip-port vlan included
```

Note MEC load-sharing becomes effective only when each virtual switch node has more than one physical path in the same bundle interface.

MEC Hash Algorithm

Like MEC load sharing, the hash algorithm is computed independently by each virtual switch to perform load sharing via its local physical ports. Traffic load share is defined based on the number of internal bits allocated to each local member link port. The Cisco Catalyst 6500-E system in VSS mode assigns eight bits to every MEC; 8-bit can be represented as a 100 percent switching load. Depending on the number of local member link ports in a MEC bundle, the 8-bit hash is computed and allocated to each port for optimal load sharing results. Like the standalone network design, VSS supports the following EtherChannel hash algorithms:

- *Fixed*—Default setting. Keep the default if each virtual switch node has a single local member link port bundled in the same Layer 2/Layer 3 MEC (total of two ports in MEC).
- *Adaptive*—Best practice is to modify to adaptive hash method if each virtual switch node has greater than or equal to two physical ports in the same Layer2/Layer 3 MEC.

When deploying a full-mesh V-shaped, VSS-enabled campus core network, it is recommended to modify the default MEC hash algorithm from the default settings as shown in the following sample configuration:

```
cr23-VSS-Core(config)#port-channel hash-distribution adaptive
```

Modifying the MEC hash algorithm to adaptive mode requires the system to internally reprogram the hash result on each MEC. Therefore, plan for additional downtime to make the new configuration effective.

```
cr23-VSS-Core(config)#interface Port-channel 101
cr23-VSS-Core(config-if)#shutdown
cr23-VSS-Core(config-if)#no shutdown

cr23-VSS-Core#show etherchannel 101 detail | inc Hash
Last applied Hash Distribution Algorithm: Adaptive
```

2 Network Addressing Hierarchy

Developing a structured and hierarchical IP address plan is as important as any other design aspect of the borderless network. Identifying an IP addressing strategy for the entire network design is essential.

Note This section does not explain the fundamentals of TCP/IP addressing; for more details, see the many Cisco Press publications that cover this topic.

The following are key benefits of using hierarchical IP addressing:

- *Efficient address allocation*
  - Hierarchical addressing provides the advantage of grouping all possible addresses contiguously.
  - In non-contiguous addressing, a network can create addressing conflicts and overlapping problems, which may not allow the network administrator to use the complete address block.
- *Improved routing efficiencies*
  - Building centralized main and remote campus site networks with contiguous IP addresses provides an efficient way to advertise summarized routes to neighbors.
  - Route summarization simplifies the routing database and computation during topology change events.
  - Reduces the network bandwidth used by routing protocols.
  - Improves overall routing protocol performance by flooding fewer messages and improves network convergence time.
- *Improved system performance*
  - Reduces the memory needed to hold large-scale discontiguous and non-summarized route entries.
  - Reduces CPU power required to re-compute large-scale routing databases during topology change events.
  - Becomes easier to manage and troubleshoot.
  - Helps in overall network and system stability.

3 Network Foundational Technologies for LAN Design

In addition to a hierarchical IP addressing scheme, it is also essential to determine which areas of the campus design are Layer 2 or Layer 3 to determine whether routing or switching fundamentals need to be applied. The following applies to the three layers in a campus design model:

- *Core layer*—Because this is a Layer 3 network that interconnects several remote locations and shared devices across the network, choosing a routing protocol is essential at this layer.
- *Distribution layer*—The distribution block uses a combination of Layer 2 and Layer 3 switching to provide for the appropriate balance of policy and access controls, availability, and flexibility in subnet allocation and VLAN usage. Both routing and switching fundamentals need to be applied.

- *Access layer*—This layer is the demarcation point between network infrastructure and computing devices. This is designed for critical network edge functions to provide intelligent application and device-aware services, to set the trust boundary to distinguish applications, provide identity-based network access to protected data and resources, provide physical infrastructure services to reduce greenhouse emission, and so on. This subsection provides design guidance to enable various types of Layer 1 to Layer 3 intelligent services and to optimize and secure network edge ports.

The recommended routing or switching scheme of each layer is discussed in the following sections.

Designing the Core Layer Network

Because the core layer is a Layer 3 network, routing principles must be applied. Choosing a routing protocol is essential; routing design principles and routing protocol selection criteria are discussed in the following subsections.

Routing Design Principles

Although enabling routing functions in the core is a simple task, the routing blueprint must be well understood and designed before implementation, because it provides the end-to-end reachability path of the enterprise network. For an optimized routing design, the following three routing components must be identified and designed to allow more network growth and provide a stable network, independent of scale:

- *Hierarchical network addressing*—Structured IP network addressing in the borderless network where the LAN and/or WAN design are required to make the network scalable, optimal, and resilient.
- *Routing protocol*—Cisco IOS supports a wide range of Interior Gateway Protocols (IGPs). Cisco recommends deploying a single routing protocol across the borderless network infrastructure.
- *Hierarchical routing domain*—Routing protocols must be designed in a hierarchical model that allows the network to scale and operate with greater stability. Building a routing boundary and summarizing the network minimizes the topology size and synchronization procedure, which improves overall network resource use and re-convergence.

Routing Protocol Selection Criteria

The criteria for choosing the right protocol vary based on the end-to-end network infrastructure. Although all the routing protocols that Cisco IOS and NX-OS currently support can provide a viable solution, network architects must consider all of the following critical design factors when selecting the routing protocol to be implemented throughout the internal network:

- *Network design*—Requires a proven protocol that can scale in full-mesh campus network designs and can optimally function in hub-and-spoke WAN network topologies.

- *Scalability*—The routing protocol function must be network- and system-efficient and operate with a minimal number of updates and re-computation, independent of the number of routes in the network.
- *Rapid convergence*—Link-state versus DUAL re-computation and synchronization. Network re-convergence also varies based on network design, configuration, and a multitude of other factors that may be more than a specific routing protocol can handle. The best convergence time can be achieved from a routing protocol if the network is designed to the strengths of the protocol.
- *Operational*—A simplified routing protocol that can provide ease of configuration, management, and troubleshooting.

Cisco IOS and NX-OS support a wide range of routing protocols, such as Routing Information Protocol (RIP) v1/2, Enhanced Interior Gateway Routing Protocol (EIGRP), Open Shortest Path First (OSPF), and Intermediate System-to-Intermediate System (IS-IS). However, Cisco recommends using EIGRP or OSPF for this network design. EIGRP is a popular version of an Interior Gateway Protocol (IGP) because it has all the capabilities needed for small- to large-scale networks, offers rapid network convergence, and above all is simple to operate and manage. OSPF is a popular link state protocol for large-scale enterprise and service provider networks. OSPF enforces hierarchical routing domains in two tiers by implementing backbone and non-backbone areas. The OSPF area function depends on the network connectivity model and the role of each OSPF router in the domain.

Other technical factors must be considered when implementing OSPF in the network, such as OSPF router type, link type, maximum transmission unit (MTU) considerations, designated router (DR)/backup designated router (BDR) priority, and so on. This document provides design guidance for using simplified EIGRP and OSPF in the borderless network infrastructure.

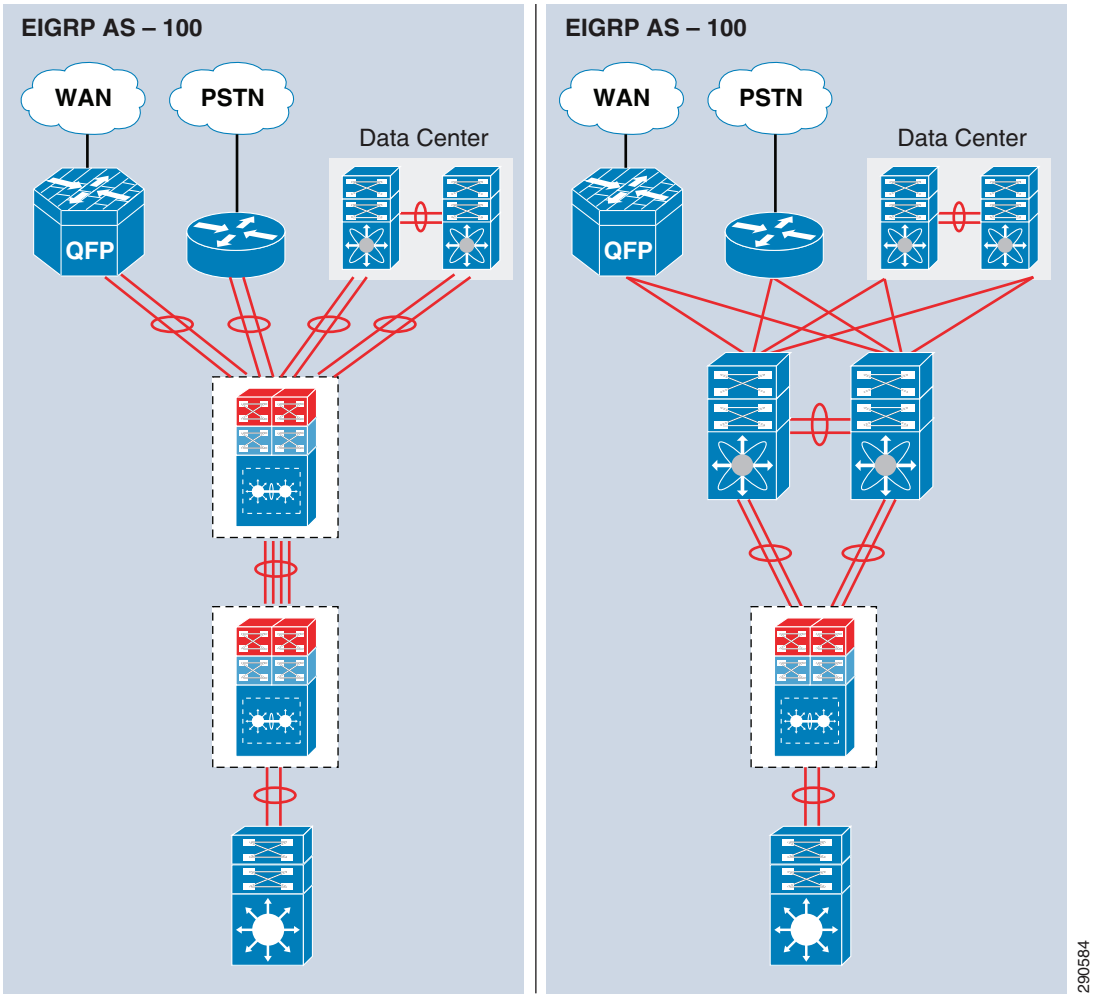
Note For detailed information on EIGRP and OSPF, see: http://www.cisco.com/en/US/docs/solutions/Enterprise/Campus/routed-ex.html.

Designing EIGRP Routing in the Campus Network

EIGRP is a balanced hybrid routing protocol that builds neighbor adjacency and flat routing topology on a per-autonomous system (AS) basis. Cisco recommends considering the following three critical design tasks before implementing EIGRP in the campus core layer network:

- *EIGRP autonomous system*—The Layer 3 LAN and WAN infrastructure of the borderless network must be deployed in a single EIGRP AS, as shown in Figure 29. A single EIGRP AS reduces operational tasks and prevents route redistribution, loops, and other problems that may occur because of misconfiguration. Figure 29 illustrates end-to-end single EIGRP autonomous network design in an enterprise network.

Figure 29 End-to-End EIGRP Routing Design in the Campus Network



Implementing EIGRP Routing Protocol

The following sample configuration provides deployment guidelines for implementing the EIGRP routing protocol on all Cisco IOS Layer 3 network devices and Cisco Nexus 7000 running NX-OS into a single Autonomous System (AS):

Cisco IOS

```
cr23-VSS-Core(config)#router eigrp 100
cr23-VSS-Core(config-router)# network 10.0.0.0
cr23-VSS-Core(config-router)# eigrp router-id 10.125.200.254
cr23-VSS-Core(config-router)# no auto-summary

cr23-VSS-Core#show ip eigrp neighbors
EIGRP-IPv4 neighbors for process 100
H   Address                 Interface       Hold     Uptime   SRTT   RTO  Q  Seq
                                              (sec)          (ms)       Cnt Num
7   10.125.0.13             Po101             12       3d16h       1   200  0  62
0   10.125.0.15             Po102             10      3d16h       1   200  0  503
1   10.125.0.17             Po103             11       3d16h       1   200  0  52
…

cr23-VSS-Core#show ip route eigrp | inc /16|/20|0.0.0.0
     10.0.0.0/8 is variably subnetted, 41 subnets, 5 masks
D       10.126.0.0/16 [90/3072] via 10.125.0.23, 08:33:16, Port-channel106
D       10.125.128.0/20 [90/3072] via 10.125.0.17, 08:33:15, Port-channel103
D       10.125.96.0/20 [90/3072] via 10.125.0.13, 08:33:18, Port-channel101
D       10.125.0.0/16 is a summary, 08:41:12, Null0
…
D*EX 0.0.0.0/0 [170/515072] via 10.125.0.27, 08:33:20, Port-channel108
```

Cisco NX-OS

```
cr35-N7K-Core1(config)#feature eigrp
!Enable EIGRP feature set

cr35-N7K-Core1(config)# router eigrp 100
cr35-N7K-Core1(config-router)# router-id 10.125.100.3
!Configure system-wide EIGRP RID, auto-summarization is by default off in Cisco NX-OS

cr35-N7K-Core1(config)# interface Port-channel 100-103
cr35-N7K-Core1(config-if-range)#ip router eigrp 100
!Associate EIGRP routing process on per L3 Port-channel interface basis

cr35-N7K-Core1(config)# interface Ethernet 1/3-4
cr35-N7K-Core1(config-if-range)#ip router eigrp 100
!Associate EIGRP routing process on per L3 interface basis

cr35-N7K-Core1#show ip eigrp neighbor
IP-EIGRP neighbors for process 100 VRF default
H   Address                 Interface       Hold  Uptime   SRTT   RTO  Q  Seq
                                            (sec)          (ms)       Cnt Num
5   10.125.10.0             Po101           12    00:05:18  2    200   0   83
4   10.125.12.4             Po103           12    00:05:19  3    200   0   89
3   10.125.12.0             Po102           13    00:05:19  1    200   0   113
2   10.125.11.0             Eth1/3          14    00:05:19  1    200   0   75
1   10.125.11.4             Eth1/4          12    00:05:19  1    200   0   72
```

```
0   10.125.21.1             Po100           14   00:05:20  1    200   0   151
```

- *EIGRP adjacency protection*—This increases network infrastructure efficiency and protection by securing the EIGRP adjacencies with internal systems. This task involves two subset implementation tasks on each EIGRP-enabled network device:
  - Increases system efficiency—Blocks EIGRP processing with passive-mode configuration on physical or logical interfaces connected to non-EIGRP devices in the network, such as PCs. The best practice helps reduce CPU utilization and secures the network with unprotected EIGRP adjacencies with untrusted devices. The following sample configuration provides guidelines to enable EIGRP protocol communication on trusted interfaces and block on all system interfaces. This recommended best practice must be enabled on all of the EIGRP Layer 3 systems in the network:

    Cisco IOS

    ```
    cr23-VSS-Core(config)#router eigrp 100
    cr23-VSS-Core(config-router)# passive-interface default
    cr23-VSS-Core(config-router)# no passive-interface Port-channel101
    cr23-VSS-Core(config-router)# no passive-interface Port-channel102
    <snippet>
    ```

    Cisco NX-OS

    ```
    cr35-N7K-Core1(config)#interface Ethernet 1/3
    cr35-N7K-Core1(config-if)#ip passive-interface eigrp 100
    ```

  - Network security—Each EIGRP neighbor in the LAN/WAN network must be trusted by implementing and validating the Message-Digest algorithm 5 (MD5) authentication method on each EIGRP-enabled system in the network. Follow the recommended EIGRP MD5 adjacency authentication configuration on each non-passive EIGRP interface to establish secure communication with remote neighbors. This recommended best practice must be enabled on all of the EIGRP Layer 3 systems in the network:

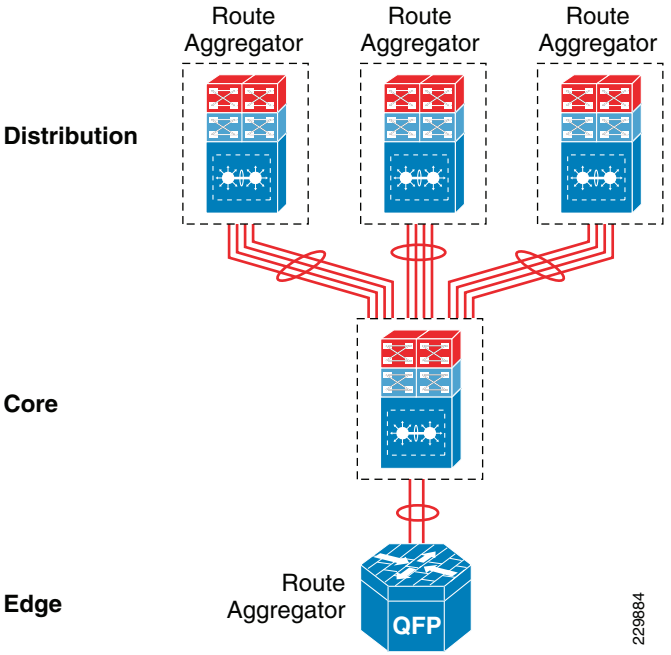
    Cisco IOS and NX-OS

    ```
    cr23-VSS-Core(config)#key chain eigrp-key
    cr23-VSS-Core(config-keychain)# key 1
    cr23-VSS-Core(config-keychain-key)#key-string <password>

    cr23-VSS-Core(config)#interface range Port-Channel 101 - 108
    cr23-VSS-Core(config-if-range)# ip authentication mode eigrp 100 md5
    cr23-VSS-Core(config-if-range)# ip authentication key-chain eigrp 100 eigrp-key
    ```

- *Optimizing EIGRP topology*—EIGRP allows network administrators to summarize multiple individual and contiguous networks into a single summary network before advertising to the neighbor. Route summarization helps improve network performance, stability, and convergence by hiding the fault of an individual network that requires each router in the network to

synchronize the routing topology. Each aggregating device must summarize a large number of networks into a single summary route. Figure 30 shows an example of the EIGRP topology for the campus design.

Figure 30 EIGRP Route Aggregator Design



The following configuration must be applied on each EIGRP route aggregator system as depicted in Figure 30. EIGRP route summarization must be implemented on the upstream logical port channel interface to announce a single prefix from each block.

Distribution

```
cr22-6500-LB(config)#interface Port-channel100
cr22-6500-LB(config-if)# ip summary-address eigrp 100 10.125.96.0 255.255.240.0
```

```
cr22-6500-LB#show ip protocols
…
  Address Summarization:
    10.125.96.0/20 for Port-channel100
```

```
<snippet>

cr22-6500-LB#s ip route | inc Null0
D       10.125.96.0/20 is a summary, 3d16h, Null0
```

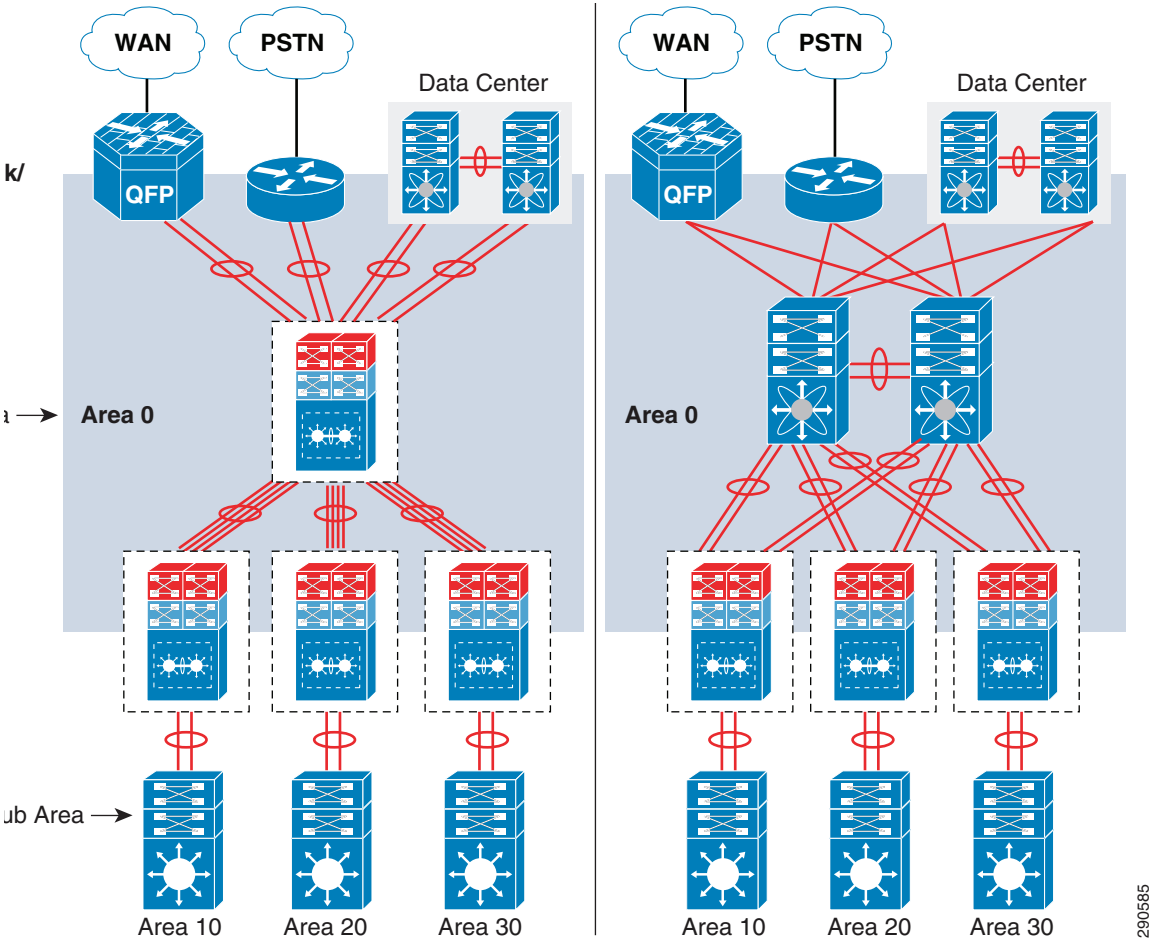
- *EIGRP Timers*—By default, EIGRP speakers running Cisco IOS or NX-OS systems transmit Hello packets every five seconds and terminate EIGRP adjacency if the neighbor fails to receive it within 15 seconds of hold-down time. In this network design, Cisco recommends retaining the default EIGRP Hello and Hold timers on all EIGRP-enabled platforms. Implementing aggressive EIGRP Hello processing and Hold-down timers may create adverse impacts during graceful recovery processes on any of the redundant campus layer systems.

Designing OSPF Routing in the Campus Network

OSPF is a widely-deployed IETF standard link state and adaptive routing protocol for the heterogeneous vendor enterprise network environment. Unlike the EIGRP protocol, the OSPF network builds a structured network boundary into multiple areas, which helps in propagating summarized network topology information and rapidly performs OSPF database computations for intelligent forwarding decisions. OSPF divides the routing boundaries into non-backbone areas that connect to single core backbone area; such design helps simplify administration and optimizes network traffic and resource utilization. The OSPF protocol supports various types of areas; this design guide recommends the following two areas types to implement in an enterprise campus network (see Figure 31):

- Backbone area—The campus core layer is the heart of the network backbone and it must be configured with OSPF backbone area. The campus aggregation layer system must be implemented in the Area Border Router (ABR) role that interconnects to the core backbone area and to the access layer non-backbone area OSPF systems. Cisco recommends that OSPF routing and backbone area design be contiguous.
- Stub/Totally Stub area—The campus access layer network requires concise network topology information and the default route from the distribution layer systems to reach external networks. Hence the non-OSPF backbone area between the distribution and access layer must be configured into stub area or totally stub area mode. Only the non-backbone area can be deployed into stub or totally-stub area mode. For a summarized network topology, these area types do not receive external route information from an area border router. Totally-stub area is a Cisco proprietary area type and an extension to an IETF standard stub area. Cisco's totally-stubby area is the same as stub area, however it does not allow a summarized network and external route to reduce the OSPF database size.

Figure 31 End-to_End Routing Design in the Campus Network



Implementing OSPF Routing Protocol

The following sample configuration provides deployment guidelines for implementing the OSPF routing protocol in the OSPF backbone area:

Core

Cisco IOS

```
cr23-VSS-Core(config)#router ospf 100
```

```
cr23-VSS-Core(config-router)#router-id 10.125.200.254
cr23-VSS-Core(config-router)#network 10.125.0.0 0.0.255.255 area 0
!All connected interfaces configured in OSPF backbone area 0
```

Cisco NX-OS

```
cr35-N7K-Core1(config)# feature ospf
!Enable OSPFv2 feature set

cr35-N7K-Core1(config)#router ospf 100
cr35-N7K-Core1(config-router)#router-id 10.125.100.3
cr35-N7K-Core1(config-router)#log-adjacency-changes detail
!Configure OSPF routing instance and router-id for this system

cr35-N7K-Core1(config)#interface loopback0
cr35-N7K-Core1(config-if)#ip router ospf 100 area 0.0.0.0

cr35-N7K-Core1(config)#interface eth1/1 , et2/1
cr35-N7K-Core1(config-if)#ip router ospf 100 area 0.0.0.0

cr35-N7K-Core1(config)#interface port-channel 100 - 103
cr35-N7K-Core1(config-if)#ip router ospf 100 area 0.0.0.0
!Associate OSPF process and area-id on per-interface basis
```

Distribution

```
cr22-6500-LB(config)#router ospf 100
cr22-6500-LB(config-router)# router-id 10.125.200.6
cr22-6500-LB(config-router)#network 10.125.200.6 0.0.0.0 area 0
cr22-6500-LB(config-router)#network 10.125.0.13 0.0.0.0 area 0
!Loopback and interface connected to Core router configured in OSPF backbone area 0

    cr22-6500-LB#show ip ospf neighbor
    !OSPF adjacency between Distribution and Core successfully established
Neighbor ID     Pri   State           Dead Time   Address         Interface
10.125.200.254  1   FULL/DR       00:00:38    10.125.0.12     Port-channel101
    …..
```

- OSPF adjacency protection—Like EIGRP routing security, these best practices increase network infrastructure efficiency and protection by securing the OSPF adjacencies with internal systems. This task involves two subset implementation tasks on each OSPF-enabled network device:
  - Increases system efficiency—Blocks OSPF processing with passive-mode configuration on physical or logical interfaces connected to non-OSPF devices in the network, such as PCs. The best practice helps reduce CPU utilization and secures the network with unprotected OSPF adjacencies with untrusted neigbors. The following sample configuration provides guidelines to explicitly enable OSPF protocol communication on trusted interfaces and block on all other interfaces. This recommended best practice must be enabled on all of the OSPF Layer 3 systems in the network:

Cisco IOS

```
cr22-6500-LB(config)#router ospf 100
cr22-6500-LB(config-router)# passive-interface default
cr22-6500-LB(config-router)# no passive-interface Port-channel101
```

Cisco NX-OS

```
cr35-N7K-Core1(config)#interface Ethernet 1/3
cr35-N7K-Core1(config-if)#ip ospf passive-interface
```

- Network security—Each OSPF neighbor in the LAN/WAN network must be trusted by implementing and validating the Message-Digest algorithm 5 (MD5) authentication methods on each OSPF-enabled system in the network. The following recommended OSPF MD5 adjacency authentication configuration must be in the OSPF backbone and each non-backbone area to establish secure communication with remote neighbors. This recommended best practice must be enabled on all of the OSPF Layer 3 systems in the network:

Cisco IOS and NX-OS

```
cr22-6500-LB(config)#router ospf 100
cr22-6500-LB (config-router)#area 0 authentication message-digest
cr22-6500-LB (config-router)#area 10 authentication message-digest
!Enables common OSPF MD5 authentication method for all interfaces

    cr22-6500-LB(config)#interface Port-Channel 101
    cr22-6500-LB(config-if-range)# ip ospf message-digest-key 1 <key>

cr22-6500-LB#show ip ospf interface Port-channel101 | inc authen|key
    Message digest authentication enabled
     Youngest key id is 1
```

- Optimizing OSPF topology—Depending on the network design, the OSPF protocol may be required to be fine-tuned in several aspects. Building borderless enterprise campus networks with Cisco VSS and Nexus 7000 with the recommended best practices inherently optimizes several routing components. Leveraging the underlying virtualized campus network benefits, this design guide recommends two fine-tuning parameters to be applied on OSPF-enabled systems:
  - Route Aggregation—OSPF route summarization must be performed at the area border routers (ABR) that connect the OSPF backbone and several aggregated non-backbone; typically ABR routers are the campus distribution or WAN aggregation systems. Route summarization helps network administrators to summarize multiple individual and contiguous networks into a single summary network before advertising into the OSPF backbone area. Route summarization helps improve network performance, stability, and convergence by hiding the fault of an individual network that requires each router in the network to synchronize the routing topology. Refer to Figure 30 for an example of OSPF route aggregation topology in the enterprise campus design.

Distribution

```
!Route Aggregation on distribution layer OSPF ABR router
cr22-6500-LB(config)#router ospf 100
cr22-6500-LB(config-router)# area <non-backbone area> range <subnet> <mask>
```

– Network Type—OSPF supports several network types, each designed to operate optimally in various types of network connectivity and designs. The default network type for the OSPF protocol running over an Ethernet-based network is broadcast. Ethernet is a multi-acccess network that provides the flexibility to interconnect several OSPF neighbors deployed in a single Layer 2 broadcast domain. In a best practice campus network design, two Layer 3 systems interconnect directly to each other, thus forming point-to-point communication. Cisco recommends modifying the default OSPF network type from broadcast to point-to-point on systems running Cisco IOS and NX-OS, which optimizes adjacencies by eliminating DR/BDR processing and reducing routing complexities between all OSPF-enabled systems:

Cisco IOS and NX-OS

```
cr22-6500-LB #show ip ospf interface Port-channel 101 | inc Network
  Process ID 100, Router ID 10.125.100.2, Network Type BROADCAST, Cost: 1

cr22-6500-LB#show ip ospf neighbor
           !OSPF negotiates DR/BDR processing on Broadcast network
Neighbor ID     Pri   State          Dead Time   Address         Interface
10.125.200.254  1   FULL/DR      00:00:38    10.125.0.12     Port-channel101

cr22-6500-LB (config)#interface Port-channel 101
cr22-6500-LB (config-if)#ip ospf network point-to-point

cr22-6500-LB#show ip ospf neighbor
           !OSPF point-to-point network optimizes adjacency processing
Neighbor ID     Pri   State          Dead Time   Address         Interface
10.125.200.254  1   FULL/  -      00:00:32    10.125.0.12     Port-channel101
```

- OSPF Hello Timers—By default, OSPF routers transmit Hello packets every 10 seconds and terminate OSPF adjacency if the neighbor fails to receive it within four intervals or 40 seconds of dead time. In this optimized and best practice network design, Cisco recommends to retain default OSPF Hello and Hold timers on all OSPF-enabled platforms running Cisco IOS or NX-OS operating systems. Implementing aggressive Hello processing timers and Dead Times may adversely impact graceful recovery processes on any of the redundant campus layer systems.
- OSPF Auto-Cost—The metric of an OSPF interface determines the best forwarding path based on lower metric or cost to the destination. By default, the metric or cost of an OSPF interface is automatically derived based on a fixed formula (108/bandwidth in bps) on Cisco Catalyst switches running IOS software. For example, the OSPF cost for 10 Gbps link is computed as 1. In the

EtherChannel/MEC-based network design, bundling multiple 10Gbps links into a logical port channel interface dynamically increases the aggregated bandwidth, however the OSPF cost remains 1 due to the fixed formula:

Cisco IOS

```
Dist-VSS#show interface Port-Channel 3 | inc BW|Member
  MTU 1500 bytes, BW 20000000 Kbit, DLY 10 usec,
  Members in this channel: Te1/1/8 Te2/1/8

Dist-VSS#show ip ospf interface Port-Channel 3 | inc Cost
      Process ID 100, Router ID 10.125.100.2, Network Type POINT_TO_POINT, Cost: 1
```

The default OSPF auto-cost can be adjusted if required, however it is recommended to maintain the default reference cost value on all OSPF switches running Cisco IOS. In a best practice enterprise campus network design, the default OSPF auto cost parameter helps minimize OSPF topology change and maintains network capacity and reliability during individual EtherChannel/MEC member links.

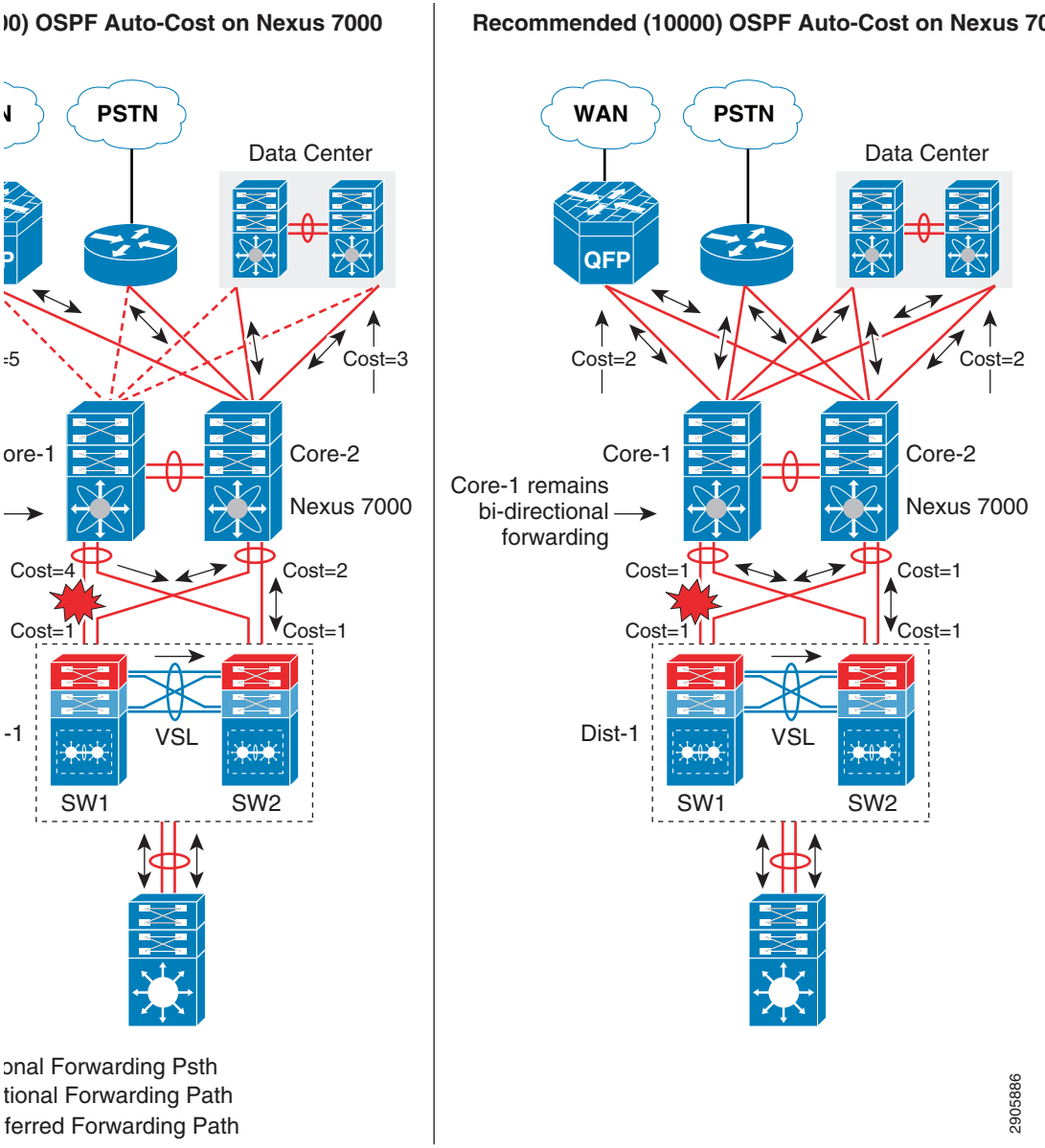
In the Cisco NX-OS, the default OSPF auto-cost reference bandwidth is 4010/bandwidth in bps or 40000 Mbps. Hence by default the OSPF metric of port channel interface can reflect the actual value based on the number of aggregated ports. With up to four 10Gbps interfaces bundled in EtherChannel, the Cisco Nexus 7000 can dynamically adjust the OSPF cost based on aggregated port channel interface bandwidth:

Cisco NX-OS

```
cr35-N7K-Core1#show ip ospf | inc Ref
     Reference Bandwidth is 40000 Mbps
cr35-N7K-Core1#show interface Port-Channel 101 | inc BW|Members
  MTU 1500 bytes, BW 20000000 Kbit, DLY 10 usec
      Members in this channel: Eth1/2, Eth2/2
cr35-N7K-Core1#show ip ospf int Port-Channel 101 | inc cost
    State P2P, Network type P2P, cost 2
```

In the EtherChannel-based campus network design, the default auto-cost reference value on Cisco Nexus 7000 systems should be adjusted to the same as Cisco IOS software. Reverting the default OSPF auto-cost setting to the same as Cisco IOS provides multiple benefits as described earlier. Figure 32 illustrates the network data forwarding impact during individual link loss with default auto-cost setting and with 10000 Mbps setting.

Figure 32 Network with Default and Recommended OSPF Auto-Cost



Designing the Campus Distribution Layer Network

This section provides design guidelines for deploying various types of Layer 2 and Layer 3 technologies in the distribution layer. Independent of which implemented distribution layer design model is deployed, the deployment guidelines remain consistent in all designs.

Because the distribution layer can be deployed with both Layer 2 and Layer 3 technologies, the following two network designs are recommended:

- Multilayer
- Routed access

Designing the Multilayer Network

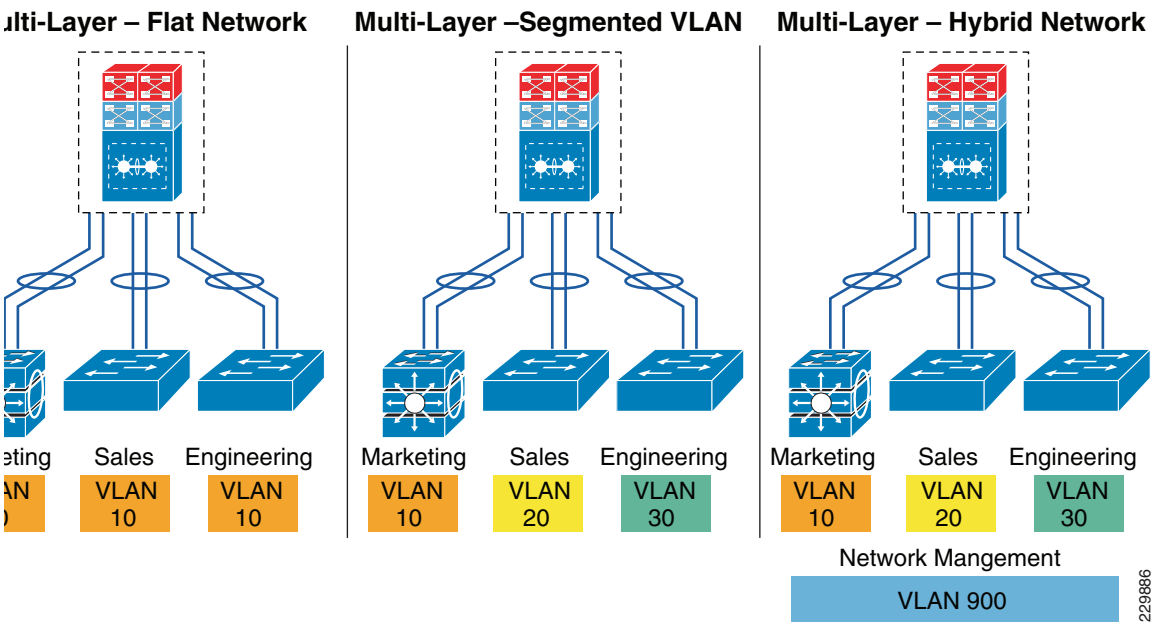
A multilayer network is a traditional, simple, and widely-deployed scenario, regardless of network scale. The access layer switches in the campus network edge interface with various types of endpoints and provide intelligent Layer 1/Layer 2 services. The access layer switches interconnect to distribution switches with the Layer 2 trunk and rely on the distribution layer aggregation switch to perform intelligent Layer 3 forwarding and to set policies and access control.

There are three design variations in building a multilayer network; all variations must be deployed in a V-shaped physical network design and must be built to provide a loop-free topology:

- *Flat*—Certain applications and user access require that the broadcast domain design span more than a single wiring closet switch. The multilayer network design provides the flexibility to build a single large broadcast domain with an extended star topology. Such flexibility introduces scalability, performance, and security challenges and may require extra attention to protect the network against misconfiguration and miswiring that can create spanning-tree loops and de-stabilize the network.
- *Segmented*—Provides a unique VLAN for different organizational divisions and enterprise business functional segments to build a per-department logical network. All network communication between various enterprise and administrative groups passes through the routing and forwarding policies defined at the distribution layer.
- *Hybrid*—A hybrid logical network design segments VLAN workgroups that do not span different access layer switches and allows certain VLANs (for example, that network management VLAN) to span across the access-distribution block. The hybrid network design enables flat Layer 2 communication without impacting the network and also helps reduce the number of subnets used.

Figure 1-33 shows the three design variations for the multilayer network.

Figure 1-33 Multilayer Design Variations



Cisco recommends that the hybrid multilayer access-distribution block design use a loop-free network topology and span a few VLANs that require such flexibility, such as the management VLAN.

The following sample configuration provides guideline to deploy several types of multilayer network components for the hybrid multilayer access-distribution block. All the configurations and best practices remain consistent and can deployed independent of Layer 2 platform type and campus location:

VTP

VLAN Trunking Protocol (VTP) is a Cisco proprietary Layer 2 messaging protocol that manages the addition, deletion, and renaming of VLANs on a network-wide basis. Cisco's VTP simplifies administration in a switched network. VTP can be configured in three modes—server, client, and transparent. It is recommended to deploy VTP in transparent mode; set the VTP domain name and change the mode to the transparent mode as follows:

```
cr22-3750-LB(config)#vtp domain Campus-BN
cr22-3750-LB(config)#vtp mode transparent
cr22-3750-LB(config)#vtp version 2

cr22-3750-LB#show vtp status
VTP Version capable:1 to 3
```

```
VTP version running:2
VTP Domain Name:Campus-BN
```

VLAN

```
cr22-3750-LB(config)#vlan 101
cr22-3750-LB(config-vlan)#name Untrusted_PC_VLAN
cr22-3750-LB(config)#vlan 102
cr22-3750-LB(config-vlan)#name Lobby_IP_Phone_VLAN
cr22-3750-LB(config)#vlan 900
cr22-3750-LB(config-vlan)#name Mgmt_VLAN

cr22-3750-LB#show vlan | inc 101|102|900
101  Untrusted_PC_VLANactive    Gi1/0/1
102  Lobby_IP_Phone_VLANactive    Gi1/0/2
900  Mgmt_VLANactive
```

Implementing Layer 2 Trunk

In a typical campus network design, a single access switch is deployed with more than a single VLAN, such as a data VLAN and a voice VLAN. The Layer 2 network connection between the distribution and access device is a trunk interface. A VLAN tag is added to maintain logical separation between VLANS across the trunk. It is recommended to implement 802.1Q trunk encapsulation in static mode instead of negotiating mode to improve the rapid link bring-up performance.

Enabling the Layer 2 trunk on a port channel automatically enables communication for all of the active VLANs between access and distribution. This may adversely impact the large scale network as the access layer switch may receive traffic flood destined to another access switch. Hence it is important to limit traffic on Layer 2 trunk ports by statically allowing the active VLANS to ensure efficient and secure network performance. Allowing only assigned VLANs on a trunk port automatically filters the rest.

By default on Cisco Catalyst switches, the native VLAN on each Layer 2 trunk port is VLAN 1 and cannot be disabled or removed from the VLAN database. The native VLAN remains active on all access switch Layer 2 ports. The default native VLAN must be properly configured to avoid several security risks— worms, viruses, or data theft. Any malicious traffic originated in VLAN 1 will span across the access layer network. With a VLAN-hopping attack it is possible to attack a system which does not reside in VLAN 1. Best practice to mitigate this security risk is to implement a unused and unique VLAN ID as a native VLAN on the Layer 2 trunk between the access and distribution switch. For example, configure VLAN 801 in the access switch and in the distribution switch. Then change the default native VLAN setting in both the switches. Thereafter, VLAN 801 must not be used anywhere for any purpose in the same access-distribution block.

The following is a configuration example to implement Layer 2 trunk, filter VLAN list, and configure the native VLAN to prevent attacks and optimize port channel interface. When the following configurations are applied on a port-channel interface (i.e., Port-Channel 1), they are automatically inherited on each bundled member link (i.e., Gig1/0/49 and Gig1/0/50):

Access-Layer

```
cr22-3750-LB(config)#vlan 801
cr22-3750-LB(config-vlan)#name Hopping_VLAN

cr22-3750-LB(config)#interface Port-channel1
cr22-3750-LB(config-if)#description Connected to cr22-6500-LB
cr22-3750-LB(config-if)#switchport
cr22-3750-LB(config-if)#switchport trunk encapsulation dot1q
cr22-3750-LB(config-if)#switchport trunk native vlan 801
cr22-3750-LB(config-if)#switchport trunk allowed vlan 101-110,900
cr22-3750-LB(config-if)#switchport mode trunk

cr22-3750-LB#show interface port-channel 1 trunk

Port        Mode             Encapsulation  Status        Native vlan
Po1         on               802.1q               trunking      801

Port        Vlans allowed on trunk
Po1         101-110,900

Port        Vlans allowed and active in management domain
Po1         101-110,900

Port        Vlans in spanning tree forwarding state and not pruned
Po1         101-110,900
```

Spanning-Tree in Multilayer Network

Spanning Tree (STP) is a Layer 2 protocol that prevents logical loops in switched networks with redundant links. The Borderless Campus design uses an Etherchannel or MEC (point-to-point logical Layer 2 bundle) connection between access layer and distribution switches, which inherently simplifies the STP topology and operation. In this point-to-point network design, the STP operation is done on a logical port, therefore it will be assigned automatically in a forwarding state.

Over the years, STP protocols have evolved into the following versions:

- *Per-VLAN Spanning Tree Plus (PVST+)*—Provides a separate 802.1D STP for each active VLAN in the network.
- *IEEE 802.1w-Rapid PVST+*—Provides an instance of RSTP (802.1w) per VLAN. It is easy to implement, proven in large scale networks that support up to 3000 logical ports, and greatly improves network restoration time.

- *IEEE 802.1s-MST*—Provides up to 16 instances of RSTP (802.1w) and combines many VLANs with the same physical and logical topology into a common RSTP instance.

It is recommended to enable the Rapid PSVT+ STP protocol in the multilayer network design. For large scale distribution block, the network administrator can consider IEEE MST as an alternate solution to simplify spanning tree instances. The following is an example configuration for enabling Rapid PVST+ in a multilayer network:

Distribution-Layer

```
cr22-6500-LB(config)#spanning-tree mode rapid-pvst

cr22-6500-LB #show spanning-tree summary | inc mode

!Switch is in rapid-pvst mode
```

Access-Layer

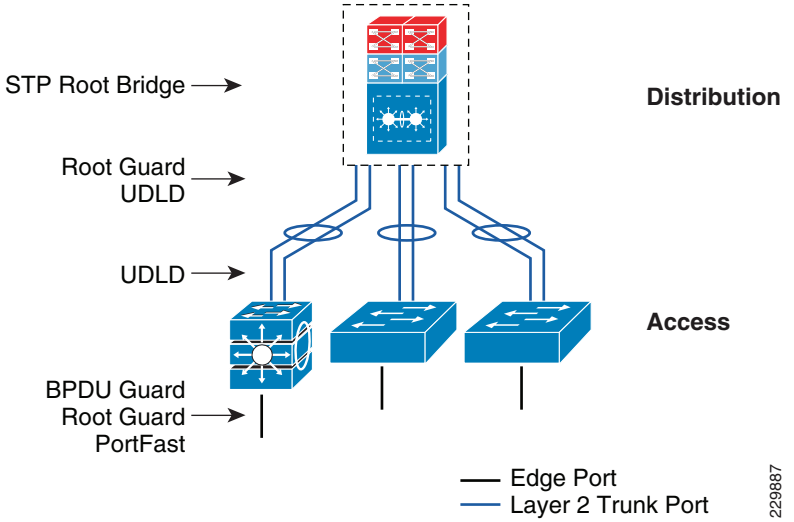
```
cr22-3750-LB(config)#spanning-tree mode rapid-pvst
```

Hardening Spanning-Tree Toolkit

Ensuring a loop-free topology is critical in a multilayer network design. Spanning Tree Protocol (STP) dynamically develops a loop-free multilayer network topology that can compute the best forwarding path and provide redundancy. Although STP behavior is deterministic, it is not optimally designed to mitigate network instability caused by hardware miswiring or software misconfiguration. Cisco has developed several STP extensions to protect against network malfunctions and to increase stability and availability. All Cisco Catalyst LAN switching platforms support the complete STP toolkit suite that must be enabled globally on individual logical and physical ports of the distribution and access layer switches.

Figure 34 shows an example of enabling various STP extensions on distribution and access layer switches in all campus sites.

Figure 34 Protecting Multilayer Network with Cisco STP Toolkit

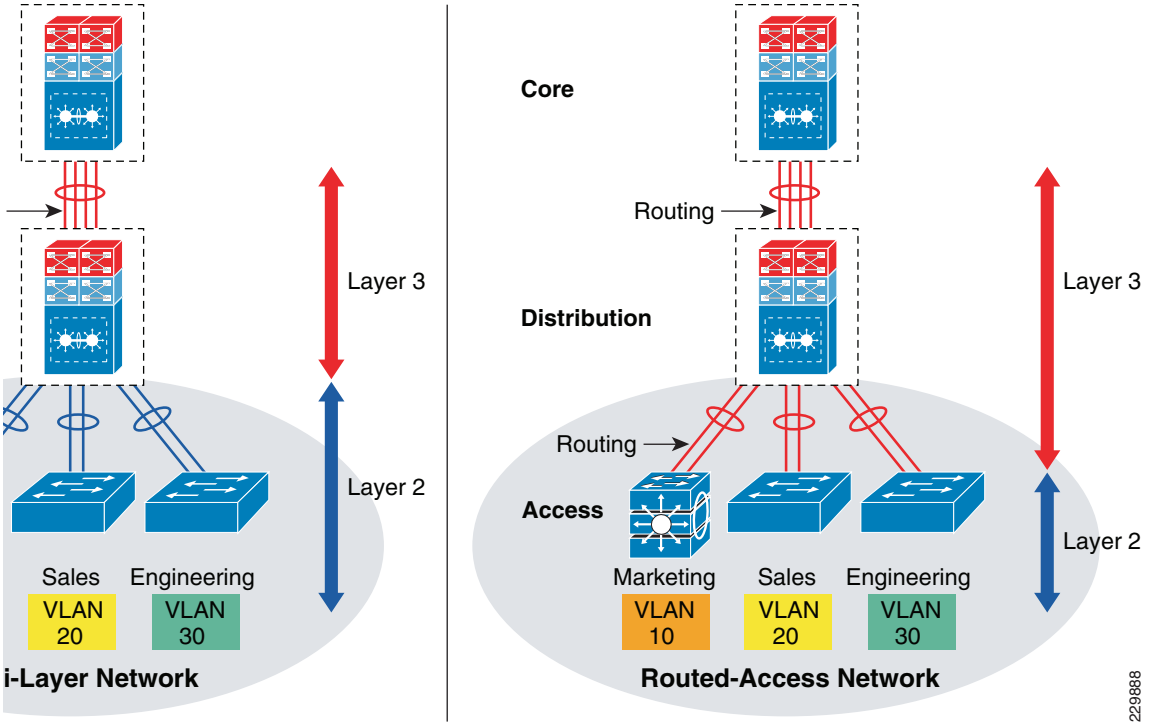


Note For additional STP information, see: http://www.cisco.com/en/US/tech/tk389/tk621/tsd_technology_support_troubleshooting_technotes_list.html.

Designing the Routed Access Network

Routing functions in the access layer network simplify configuration, optimize distribution performance, and provide end-to-end troubleshooting tools. Implementing Layer 3 functions in the access layer replaces the Layer 2 trunk configuration with a single point-to-point Layer 3 interface with a collapsed core system in the aggregation layer. Pushing Layer 3 functions one tier down on Layer 3 access switches changes the traditional multilayer network topology and forwarding development path. Implementing Layer 3 functions in the access switch does not require any physical or logical link reconfiguration; the access-distribution block can be used and is as resilient as in the multilayer network design. Figure 1-35 shows the differences between the multilayer and routed access network designs, as well as where the Layer 2 and Layer 3 boundaries exist in each network design.

Figure 1-35 Layer 2 and Layer 3 Boundaries for Multilayer and Routed Access Network Design



Routed access network design enables Layer 3 access switches to act as a Layer 2 demarcation point and provide Inter-VLAN routing and gateway functions to the endpoints. The Layer 3 access switches make more intelligent, multi-function, and policy-based routing and switching decisions like distribution layer switches.

Although Cisco VSS and a single redundant distribution design are simplified with a single point-to-point EtherChannel, the benefits in implementing the routed access design in enterprises are:

- Eliminates the need for implementing STP and the STP toolkit on the distribution system. As a best practice, the STP toolkit must be hardened at the access layer.
- Shrinks the Layer 2 fault domain, thus minimizing the number of denial-of-service (DoS)/distributed denial-of-service (DDoS) attacks.
- Bandwidth efficiency—Improves Layer 3 uplink network bandwidth efficiency by suppressing Layer 2 broadcasts at the edge port.
- Improves overall collapsed core and distribution resource utilization.

Enabling Layer 3 functions in the access-distribution block must follow the same core network designs as mentioned in previous sections to provide network security as well as optimize the network topology and system resource utilization:

- *EIGRP autonomous system*—Layer 3 access switches must be deployed in the same EIGRP AS as the distribution and core layer systems.
- *EIGRP adjacency protection*—EIGRP processing must be enabled on uplink Layer 3 EtherChannels and must block remaining Layer 3 ports by default in passive mode. Access switches must establish secured EIGRP adjacency using the MD5 hash algorithm with the aggregation system.
- *EIGRP network boundary*—All EIGRP neighbors must be in a single AS to build a common network topology. The Layer 3 access switches must be deployed in EIGRP stub mode for a concise network view.

Implementing Routed Access in Access-Distribution Block

Cisco IOS configuration to implement Layer 3 routing function on the Catalyst access layer switch remains consistent. To implement the routing function in access layer switches, refer to the EIGRP routing configuration and best practices defined in Designing EIGRP Routing in the Campus Network.

Implementing EIGRP

EIGRP creates and maintains a single flat routing topology network between EIGRP peers. Building a single routing domain in a large-scale campus core design allows for complete network visibility and reachability that may interconnect multiple campus components, such as distribution blocks, services blocks, the data center, the WAN edge, and so on.

In the three- or two-tier deployment models, the Layer 3 access switch must always have single physical or logical forwarding to a distribution switch. The Layer 3 access switch dynamically develops the forwarding topology pointing to a single distribution switch as a single Layer 3 next hop. Because the distribution switch provides a gateway function to rest of the network, the routing design on the Layer 3 access switch can be optimized with the following two techniques to improve performance and network reconvergence in the access-distribution block, as shown in Figure 1-36:

- Deploying the Layer 3 access switch in EIGRP stub mode

  An EIGRP stub router in a Layer 3 access switch can announce routes to a distribution layer router with great flexibility.

  The following is an example configuration to enable EIGRP stub routing in the Layer 3 access switch; no configuration changes are required in the distribution system:

Access layer

```
cr22-4507-LB(config)#router eigrp 100
cr22-4507-LB(config-router)# eigrp stub connected
```

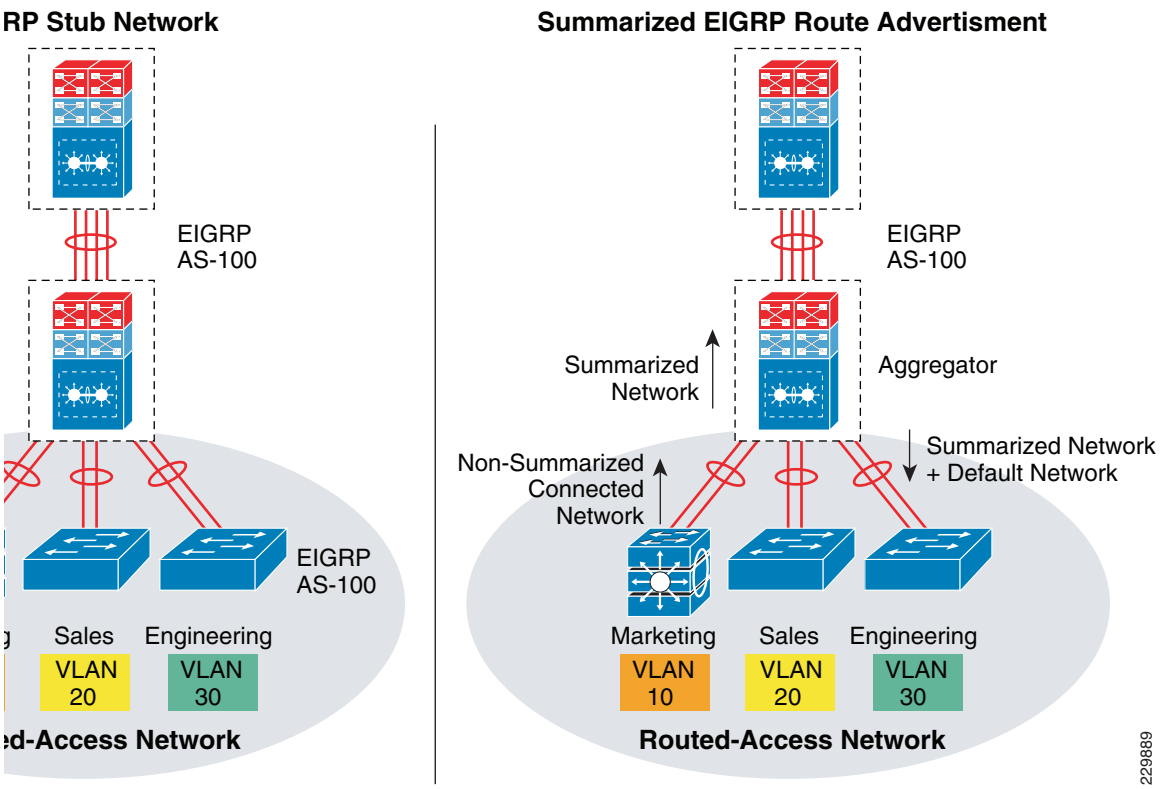
```
cr22-4507-LB#show eigrp protocols detailed

Address Family Protocol EIGRP-IPv4:(100)
  EIGRP metric weight K1=1, K2=0, K3=1, K4=0, K5=0
  EIGRP maximum hopcount 100
  EIGRP maximum metric variance 1
  EIGRP NSF-aware route hold timer is 240
  EIGRP NSF enabled
     NSF signal timer is 20s
     NSF converge timer is 120s
     Time since last restart is 2w2d
  EIGRP stub, connected
  Topologies : 0(base)
```

Distribution layer

```
cr22-6500-LB#show ip eigrp neighbors detail port-channel 101
EIGRP-IPv4 neighbors for process 100
H   Address                 Interface       Hold Uptime   SRTT   RTO  Q  Seq
                                            (sec)         (ms)       Cnt Num
2   10.125.0.1              Po101             13 3d18h       4   2000  98
   Version 4.0/3.0, Retrans: 0, Retries: 0, Prefixes: 6
   Topology-ids from peer - 0
   Stub Peer Advertising ( CONNECTED ) Routes
   Suppressing queries
```

- Summarizing the network view with a default route to the Layer 3 access switch for intelligent routing functions

Figure 1-36 Designing and Optimizing EIGRP Network Boundary for the Access Layer



The following sample configuration demonstrates the procedure to implement route filtering at the distribution layer that allows summarized and default route advertisement. The default route announcement is likely done by the network edge system, i.e., the Internet edge, to reach external networks. To maintain entire network connectivity, the distribution layer must announce the summarized and default route to access layer switches, so that even during loss of default route the internal network operation remains unaffected:

Distribution layer

```
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 5 permit 0.0.0.0/0
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 10 permit 10.122.0.0/16
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 15 permit 10.123.0.0/16
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 20 permit 10.124.0.0/16
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 25 permit 10.125.0.0/16
cr22-6500-LB(config)# ip prefix-list EIGRP_STUB_ROUTES seq 30 permit 10.126.0.0/16
```

```
cr22-6500-LB(config)#router eigrp 100
cr22-6500-LB(config-router)#distribute-list route-map EIGRP_STUB_ROUTES out
Port-channel101
cr22-6500-LB(config-router)#distribute-list route-map EIGRP_STUB_ROUTES out
Port-channel102
cr22-6500-LB(config-router)#distribute-list route-map EIGRP_STUB_ROUTES out
Port-channel103

cr22-6500-LB#show ip protocols
  Outgoing update filter list for all interfaces is not set
    Port-channel101 filtered by
    Port-channel102 filtered by
    Port-channel103 filtered by
```

Access layer

```
cr22-4507-LB#show ip route eigrp
     10.0.0.0/8 is variably subnetted, 12 subnets, 4 masks
D       10.122.0.0/16 [90/3840] via 10.125.0.0, 07:49:11, Port-channel1
D       10.123.0.0/16 [90/3840] via 10.125.0.0, 01:42:22, Port-channel1
D       10.126.0.0/16 [90/3584] via 10.125.0.0, 07:49:11, Port-channel1
D       10.124.0.0/16 [90/64000] via 10.125.0.0, 07:49:11, Port-channel1
D       10.125.0.0/16 [90/768] via 10.125.0.0, 07:49:13, Port-channel1
D *EX 0.0.0.0/0 [170/515584] via 10.125.0.0, 07:49:13, Port-channel1
```

Implementing OSPF

Since OSPF divides the routing function into core backbone and non-backbone area, the Layer 3 access layer systems must be deployed in the same non-backbone area and in totally stub area mode as the distribution layer ABR router to successfully form adjacencies. Deploying Layer 3 access layer switches in totally stub area mode enables complete network reachability and helps optimizie network topology, system, and network resources.

Enabling OSPF routing functions in the access-distribution block must follow the same core network design guidelines mentioned in Designing OSPF Routing in the Campus Network to provide network security as well as optimize the network topology and system resource utilization:

- OSPF area design—Layer 3 access switches must be deployed in the non-backbone area as the distribution and core layer systems.

- OSPF adjacency protection—OSPF processing must be enabled on uplink Layer 3 EtherChannels and must block remaining Layer 3 ports by default in passive mode. Access switches must establish secured OSPF adjacency using the MD5 hash algorithm with the aggregation system.
- OSPF network boundary—All OSPF-enabled access layer switches must be deployed in a single OSPF area to build a common network topology in each distribution block. The OSPF area between Layer 3 access switches and distribution switches must be deployed in totally stub area mode.

The following sample configuration provides deployment guidelines for implementing the OSPF totally stub area routing protocol in the OSPF non-backbone area on the distribution layer system. With the following configuration, the distribution layer system will start announcing all locally-connected networks into the OSPF backbone area:

Access-Layer

```
cr22-4507-LB(config)#router ospf 100
cr22-4507-LB(config-router)#router-id 10.125.200.1
cr22-4507-LB(config-router)#network 10.125.96.0 0.0.3.255 area 10
cr22-4507-LB(config-router)#network 10.125.200.1 0.0.0.0 area 10
cr22-4507-LB(config-router)#area 10 stub no-summary
```

Distributution

```
cr22-4507-LB(config)#router ospf 100
cr22-4507-LB(config-router)#network 10.125.96.0 0.0.15.255 area 10
cr22-4507-LB(config-router)#area 10 stub no-summary

    cr22-4507-LB#show ip ospf neighbor
!OSPF negotiates DR/BDR processing on Broadcast network
Neighbor ID     Pri   State           Dead Time   Address         Interface
10.125.200.6  1   FULL/ -      00:00:34    10.125.0.0     Port-channel101

cr22-4507-LB#show ip route ospf  | inc Port-channel101
O *IA    0.0.0.0/0  [110/2] via 10.125.0.0, 0
```

Multicast for Application Delivery

Because unicast communication is based on the one-to-one forwarding model, it becomes easier in routing and switching decisions to perform destination address lookup, determine the egress path by scanning forwarding tables, and to switch traffic. In the unicast routing and switching technologies discussed in the previous section, the network may need to be made more efficient by allowing certain applications where the same content or application must be replicated to multiple users. IP multicast delivers source traffic to multiple receivers using the least amount of network resources as possible

without placing an additional burden on the source or the receivers. Multicast packet replication in the network is done by Cisco routers and switches enabled with Protocol Independent Multicast (PIM) as well as other multicast routing protocols.

Similar to the unicast methods, multicast requires the following design guidelines:

- Choosing a multicast addressing design
- Choosing a multicast routing protocol
- Providing multicast security regardless of the location within the enterprise design

Multicast Addressing Design

The Internet Assigned Numbers Authority (IANA) controls the assignment of IP multicast addresses. A range of class D address space is assigned to be used for IP multicast applications. All multicast group addresses fall in the range from 224.0.0.0 through 239.255.255.255. Layer 3 addresses in multicast communications operate differently; while the destination address of IP multicast traffic is in the multicast group range, the source IP address is always in the unicast address range. Multicast addresses are assigned in various pools for well-known multicast-based network protocols or inter-domain multicast communications, as listed in Table 4.

Table 4 Multicast Address Range Assignments

Application	Address Range
Reserved—Link local network protocols.	224.0.0.0/24
Global scope—Group communication between an organization and the Internet.	224.0.1.0 – 238.255.255.255
Source Specific Multicast (SSM)—PIM extension for one-to-many unidirectional multicast communication.	232.0.0.0/8
GLOP—Inter-domain multicast group assignment with reserved global AS.	233.0.0.0/8
Limited scope—Administratively scoped address that remains constrained within a local organization or AS. Commonly deployed in enterprise, education, and other organizations.	239.0.0.0/8

During the multicast network design phase, the enterprise network architects must select a range of multicast sources from the limited scope pool (239/8).

Multicast Routing Design

To enable end-to-end dynamic multicast operation in the network, each intermediate system between the multicast receiver and source must support the multicast feature. Multicast develops the forwarding table differently than the unicast routing and switching model. To enable communication, multicast requires specific multicast routing protocols and dynamic group membership.

The enterprise campus design must be able to build packet distribution trees that specify a unique forwarding path between the subnet of the source and each subnet containing members of the multicast group. A primary goal in distribution tree construction is to ensure that no more than one copy of each packet is forwarded on each branch of the tree. The two basic types of multicast distribution trees are:

- *Source trees*—The simplest form of a multicast distribution tree is a source tree, with its root at the source and branches forming a tree through the network to the receivers. Because this tree uses the shortest path through the network, it is also referred to as a shortest path tree (SPT).
- *Shared trees*—Unlike source trees that have their root at the source, shared trees use a single common root placed at a selected point in the network. This shared root is called a rendezvous point (RP).

The PIM protocol is divided into the following two modes to support both types of multicast distribution trees:

- *Dense mode (DM)*—Assumes that almost all routers in the network need to distribute multicast traffic for each multicast group (for example, almost all hosts on the network belong to each multicast group). PIM in DM mode builds distribution trees by initially flooding the entire network and then pruning back the small number of paths without receivers.
- *Sparse mode (SM)*—Assumes that relatively few routers in the network are involved in each multicast. The hosts belonging to the group are widely dispersed, as might be the case for most multicasts over the WAN. Therefore, PIM-SM begins with an empty distribution tree and adds branches only as the result of explicit Internet Group Management Protocol (IGMP) requests to join the distribution. PIM-SM mode is ideal for a network without dense receivers and multicast transport over WAN environments and it adjusts its behavior to match the characteristics of each receiver group.

Selecting the PIM mode depends on the multicast applications that use various mechanisms to build multicast distribution trees. Based on the multicast scale factor and centralized source deployment design for one-to-many multicast communication in Borderless Campus network infrastructures, Cisco recommends deploying PIM-SM because it is efficient and intelligent in building a multicast distribution tree. All the recommended platforms in this design support PIM-SM mode on physical or logical (switched virtual interface [SVI] and EtherChannel) interfaces.

Designing PIM Rendezvous Point

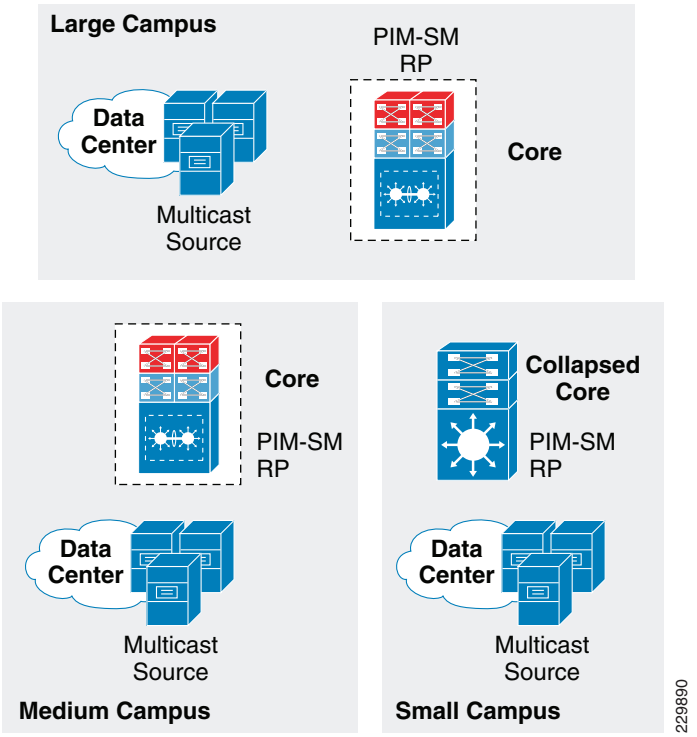
The following sections discuss best practices in designing and deploying the PIM-SM Rendezvous Point.

PIM-SM RP Placement

It is assumed that each borderless network site has a wide range of local multicast sources in the data center for distributed enterprise IT-managed media and employee research and development applications. In such a distributed multicast network design, Cisco recommends deploying PIM RP on each site for wired or wireless multicast receivers and sources to join and register at the closest RP. In a hierarchical campus network design, placing PIM-SM RP at the center point of the campus core layer network builds optimal shortest-path tree (SPT). The PIM operation in a core layer system is in a transit path between multicast receivers residing in local distribution blocks or a remote campus and the multicast sources in data centers.

The Borderless Campus design recommends PIM-SM RP placement on a highly-resilient Cisco VSS and Nexus 7000 core system in the three-tier campus design and on the collapsed core/distribution system in the two-tier campus design model. See Figure 37.

Figure 37 Distributed PIM-SM RP Placement



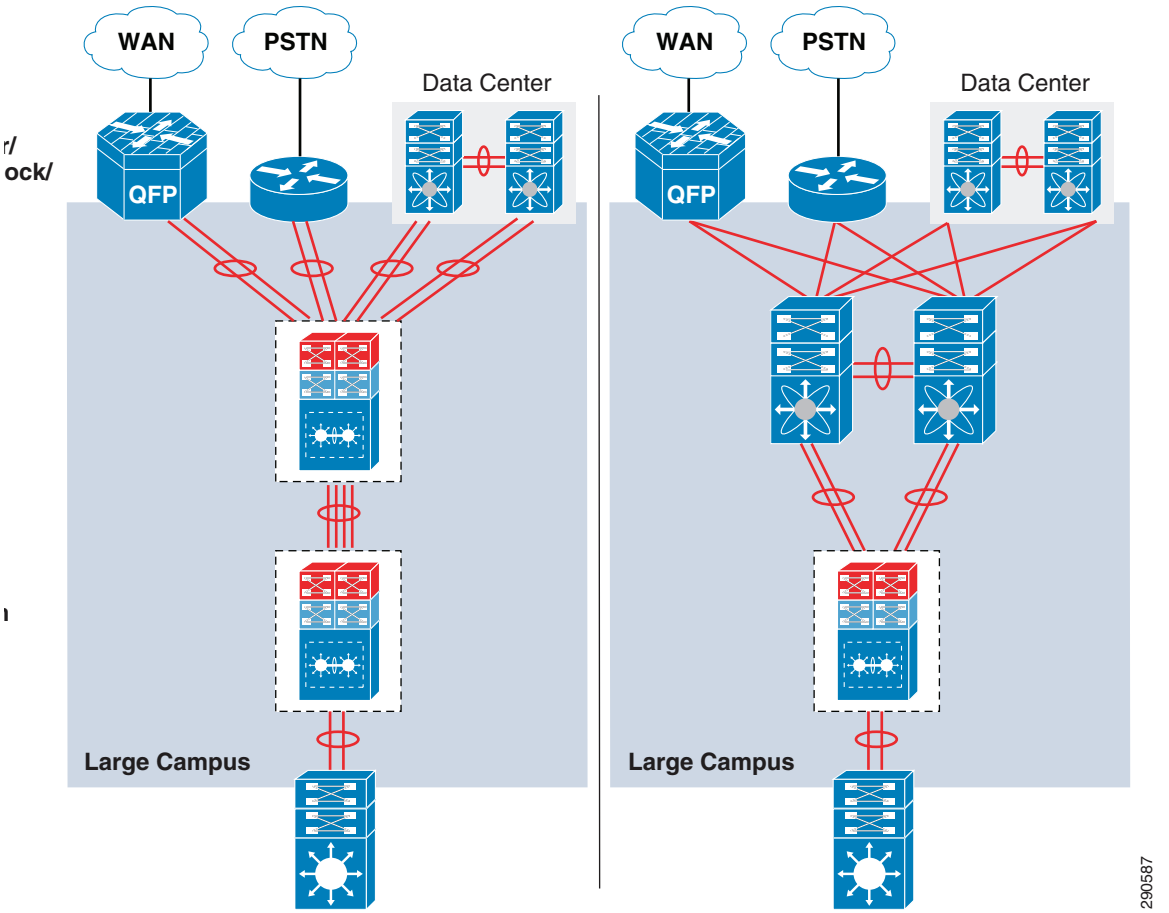
PIM-SM RP Mode

PIM-SM supports RP deployment in the following three modes in the network:

- *Static*—In this mode, RP must be statically identified and configured on each PIM router in the network. RP load balancing and redundancy can be achieved using anycast RP.
- *Auto-RP*—This mode is a dynamic method for discovering and announcing the RP in the network. Auto-RP implementation is beneficial when there are multiple RPs and groups that often change in the network. To prevent network reconfiguration during a change, the RP mapping agent router must be designated in the network to receive RP group announcements and to arbitrate conflicts, as part of the PIM version 1 specification.
- *BootStrap Router (BSR)*—This mode performs the same tasks as Auto-RP but in a different way and is part of the PIM version 2 specification. Auto-RP and BSR cannot co-exist or interoperate in the same network.

In a small- to mid-sized multicast network, static RP configuration is recommended over the other modes. Static RP implementation offers RP redundancy and load sharing and an additional simple access control list (ACL) can be applied to deploy RP without compromising multicast network security. Cisco recommends designing the enterprise campus multicast network using the static PIM-SM mode configuration and MSDP-based Anycast RP to provide RP redundancy in the campus network. See Figure 38.

Figure 38 PIM-SM Network Design in Enterprise Network

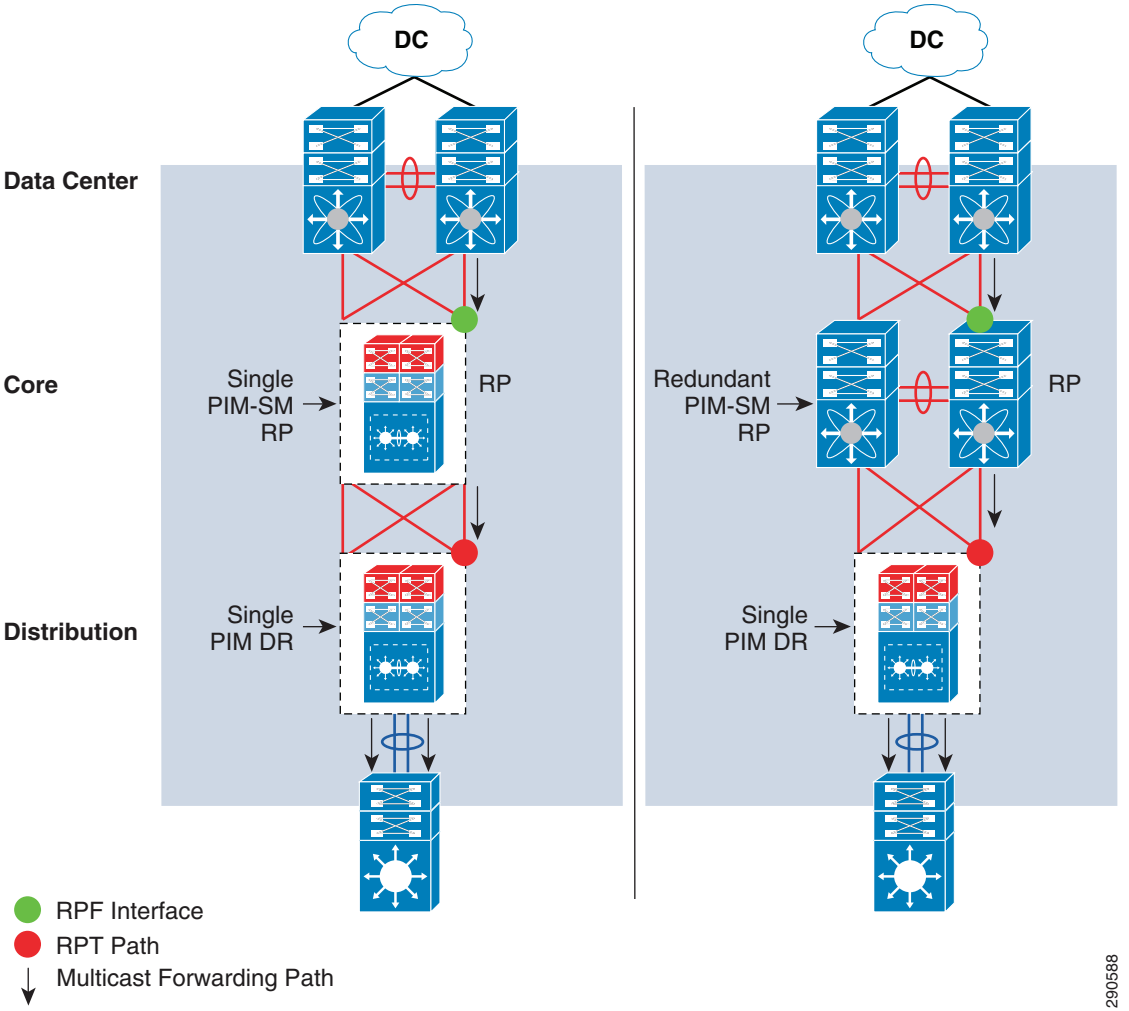


Defining the static role of PIM-SM RP at the campus core simplifies the multicast shortest-path tree (SPT) and RP trees (RPT) development operation before data forwarding begins. Based on an ECMP versus an EtherChannel campus network design, the PIM may not build the optimal forwarding shared-tree to utilize all downstream paths to transmit multicast traffic. This is primarily how the multicast distributed tree gets developed in a full mesh and equal-cost parallel paths are deployed in enterprise campus network environments. Deploying Cisco VSS in the distribution layer simplifies and optimizes the unicast and multicast forwarding planes in the campus distribution access block. With a single unified VSS control plane, it eliminates FHRP for gateway redundancy and represents a single PIM DR on a VLAN. The result of a PIM join request from a distribution layer system to a core layer PIM RP depends on the implemented Layer 3 network design—ECMP versus EtherChannel.

ECMP

Multicast forwarding is a connection-oriented process in the campus network. Multicast sources get registered with the local first-hop PIM-SM router and the multicast receivers dynamically join the PIM DR, which is the last-hop router that is rooted towards PIM RP. Following the ECMP unicast routing table, the last-hop PIM routers send PIM join messages to RP through a single Layer 3 interface. This is determined based on the highest next-hop IP address in the RIB table. The multicast distribution tree in the multicast routing table gets developed based on the interface where the PIM join communication occurred. Upon link failure, the alternate Layer 3 interface with the highest next-hop IP address is dynamically selected for multicast traffic. An ECMP-based campus network design may not build an optimal multicast forwarding network topology as it cannot leverage all physical paths available to switch data traffic.

Figure 39 Campus Multicast Operation in ECMP Design



EtherChannel/MEC

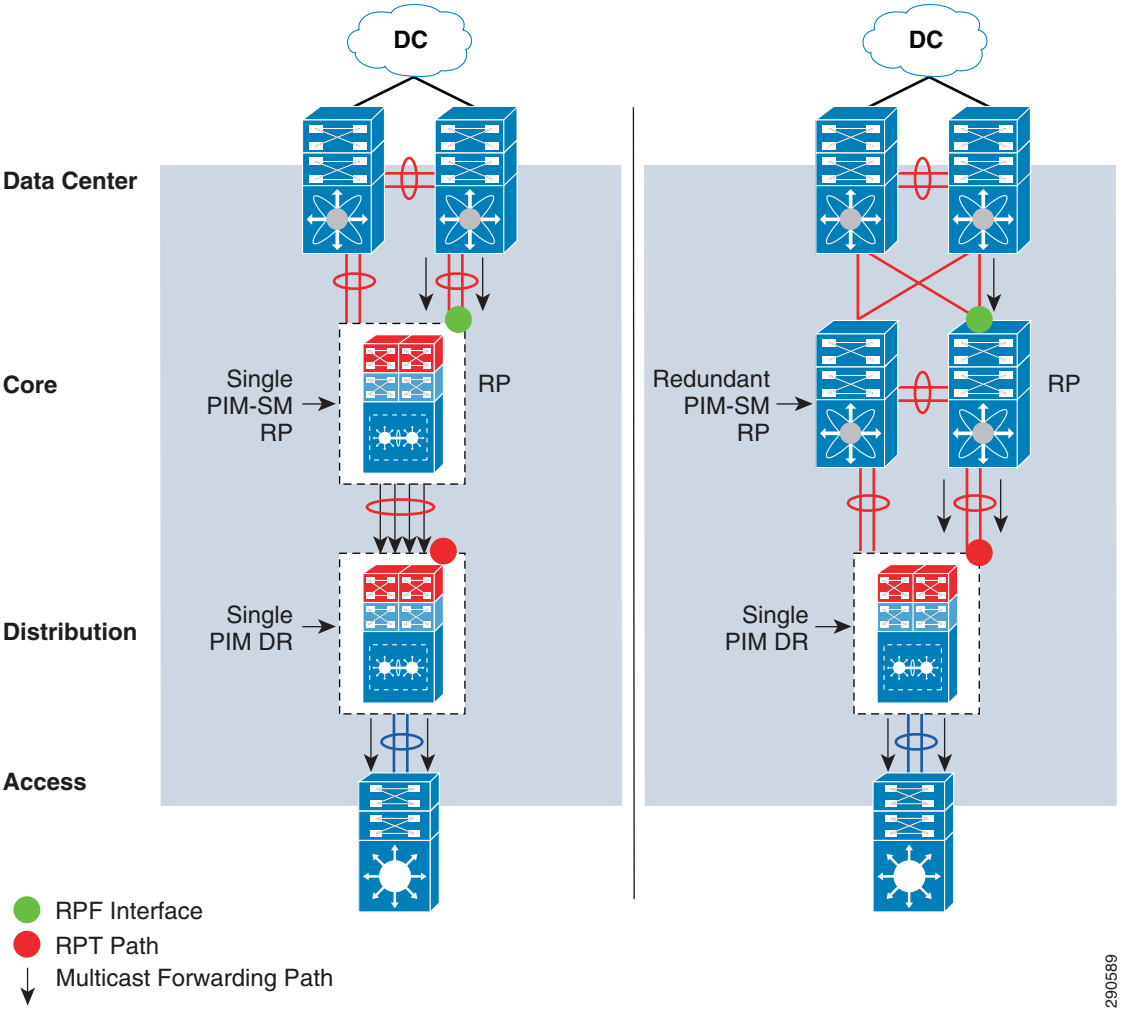
An EtherChannel-based network design offers multiple advantages to multicast protocols over ECMP. Depending on the campus core layer system type, deployment model, and its capability, the network architect can simplify and optimize multicast operation over a logical Layer 3 EtherChannel interface. The number of bundled Etherchannel member links remains transparent for multicast communication

between distribution and core layer systems. During individual member link failures, the EtherChannel provides multicast redundancy as it reduces the unicast and multicast routing topology change that triggers any recomputation or the building of a new distribution tree.

Deploying a full-mesh diverse fiber with a single logical Layer 3 EherChannel/MEC between the distribution layer VSS and core layer VSS builds a single PIM adjacency between two logical systems. Leveraging the EtherChannel Layer 2-Layer 4 load sharing technique, the performance and switching capacity of multicast traffic is increased as it is dynamically load shared across each member link by both virtual switch systems. This high-availability design increases multicast redundancy during individual link failure or even virtual switch failure.

Deploying a Cisco Nexus 7000 system in the same design also offers benefits over ECMP. Since Nexus 7000 is a standalone core system, the distribution layer can be deployed with redundant Layer 3 EtherChannel with each core system. Like ECMP, the PIM join process in this configuration will be the same as it selects the EtherChannel interface with the highest next-hop IP address. The Cisco Nexus 7000 builds the EtherChannel interface as an outgoing-interface-list (OIL) in the multicast routing table and leverages the same load sharing fundamentals to transmit multicast traffic across each bundled member link. Deploying the Cisco Nexus 7000 with redundant hardware components provides constant multicast communication during individual member link or active supervisor switchover.

Figure 40 Campus Multicast Operation in EtherChanne/MEC Design



The following is an example configuration to deploy PIM-SM RP on all PIM-SM running systems. To provide transparent PIM-SM redundancy, static PIM-SM RP configuration must be identical across the campus LAN network and on each of the PIM-SM RP routers.

- Core layer

  Cisco IOS

  ```
  cr23-VSS-Core(config)#ip multicast-routing
  ```

```
cr23-VSS-Core(config)#interface Loopback100
cr23-VSS-Core(config-if)#description Anycast RP Loopback
cr23-VSS-Core(config-if)#ip address 10.100.100.100 255.255.255.255

cr23-VSS-Core(config)#ip pim rp-address 10.100.100.100

cr23-VSS-Core#show ip pim rp

Group: 239.192.51.1, RP: 10.100.100.100, next RP-reachable in 00:00:34
Group: 239.192.51.2, RP: 10.100.100.100, next RP-reachable in 00:00:34
Group: 239.192.51.3, RP: 10.100.100.100, next RP-reachable in 00:00:34

cr23-VSS-Core#show ip pim interface

Address          Interface                      Ver/   Nbr    Query  DR     DR
                                                Mode   Count  Intvl  Prior
10.125.0.12      Port-channel101          v2/S   1            30     1
10.125.0.13
10.125.0.14      Port-channel102          v2/S   1            30     1
10.125.0.15
…

cr23-VSS-Core#show ip mroute sparse
(*, 239.192.51.8), 3d22h/00:03:20, RP 10.100.100.100, flags: S
  Incoming interface: Null, RPF nbr 0.0.0.0
  Outgoing interface list:
    Port-channel105, Forward/Sparse, 00:16:54/00:02:54
    Port-channel101, Forward/Sparse, 00:16:56/00:03:20

(10.125.31.147, 239.192.51.8), 00:16:54/00:02:35, flags: A
  Incoming interface: Port-channel105, RPF nbr 10.125.0.21
  Outgoing interface list:
    Port-channel101, Forward/Sparse, 00:16:54/00:03:20

cr23-VSS-Core#show ip mroute active
Active IP Multicast Sources - sending >= 4 kbps

Group: 239.192.51.1, (?)
   Source: 10.125.31.153 (?)
     Rate: 2500 pps/4240 kbps(1sec), 4239 kbps(last 30 secs), 12 kbps(life avg)
```

Cisco NX-OS

```
cr35-N7K-Core1(config)#feature pim
!Enable PIM feature set

cr35-N7K-Core1(config)#interface Loopback 0, 100
cr35-N7K-Core1(config-if-range)#ip pim sparse-mode
!Enable PIM-SM mode on each Loopback interface
```

```
cr23-VSS-Core(config)#interface Loopback254
cr23-VSS-Core(config-if)#description Anycast RP Loopback
cr23-VSS-Core(config-if)#ip address 10.100.100.100 255.255.255.255

cr35-N7K-Core1(config)#interface Ethernet 1/1 - 8 , Ethernet 2/1 - 8
cr35-N7K-Core1(config-if-range)#ip pim sparse-mode
!Enable PIM-SM mode on each Layer 3 interface

cr35-N7K-Core1(config)#interface Port-Channel 101 - 103
cr35-N7K-Core1(config-if-range)# ip pim sparse-mode


cr35-N7K-Core1(config)#ip pim rp-address 10.100.100.100

cr35-N7K-Core1#show ip pim interface brief
PIM Interface Status for VRF "default"
Interface            IP Address      PIM DR Address  Neighbor  Border
                                                        Count      Interface
Ethernet1/1          10.125.20.0     10.125.20.1     1         no
Ethernet1/2          10.125.10.1     10.125.10.1     1         no
port-channel101     10.125.10.1     10.125.10.1     1         no
port-channel102     10.125.12.1     10.125.12.1     1         no

cr35-N7K-Core1# show ip mroute
IP Multicast Routing Table for VRF "default"

(*, 239.192.101.1/32), uptime: 00:01:18, pim ip
  Incoming interface: loopback254, RPF nbr: 10.125.254.254
  Outgoing interface list: (count: 2)
    port-channel101, uptime: 00:01:17, pim
    port-channel103, uptime: 00:01:18, pim

(10.100.1.9/32, 239.192.101.1/32), uptime: 2d13h, pim ip
  Incoming interface: Ethernet2/1, RPF nbr: 10.125.20.3, internal
  Outgoing interface list: (count: 2)
    port-channel101, uptime: 00:01:17, pim
    port-channel103, uptime: 00:01:18, pim
<snip>
```

- Distribution layer

```
cr22-6500-LB(config)#ip multicast-routing
cr22-6500-LB(config)#ip pim rp-address 10.100.100.100


cr22-6500-LB(config)#interface range Port-channel 100 - 103
cr22-6500-LB(config-if-range)#ip pim sparse-mode

cr22-6500-LB(config)#interface range Vlan 101 - 120
```

```
cr22-6500-LB(config-if-range)#ip pim sparse-mode

cr22-6500-LB#show ip pim rp
Group: 239.192.51.1, RP: 10.100.100.100, uptime 00:10:42, expires never
Group: 239.192.51.2, RP: 10.100.100.100, uptime 00:10:42, expires never
Group: 239.192.51.3, RP: 10.100.100.100, uptime 00:10:41, expires never
Group: 224.0.1.40, RP: 10.100.100.100, uptime 3d22h, expires never

cr22-6500-LB#show ip pim interface

Address          Interface                Ver/   Nbr    QueryDR     DR
                                          Mode   Count  IntvlPrior
10.125.0.13Port-channel100v2/S   1      30     1      10.125.0.13
10.125.0.0Port-channel101v2/S   1      30     1      10.125.0.1
…
10.125.103.129Vlan101v2/S   0      30     1      10.125.103.129
…

cr22-6500-LB#show ip mroute sparse
(*, 239.192.51.1), 00:14:23/00:03:21, RP 10.100.100.100, flags: SC
  Incoming interface: Port-channel100, RPF nbr 10.125.0.12, RPF-MFD
  Outgoing interface list:
    Port-channel102, Forward/Sparse, 00:13:27/00:03:06, H
    Vlan120, Forward/Sparse, 00:14:02/00:02:13, H
    Port-channel101, Forward/Sparse, 00:14:20/00:02:55, H
    Port-channel103, Forward/Sparse, 00:14:23/00:03:10, H
    Vlan110, Forward/Sparse, 00:14:23/00:02:17, H

cr22-6500-LB#show ip mroute active
Active IP Multicast Sources - sending >= 4 kbps

Group: 239.192.51.1, (?)
  RP-tree:
     Rate: 2500 pps/4240 kbps(1sec), 4240 kbps(last 10 secs), 4011 kbps(life avg)
```

- Access layer

```
cr23-3560-LB(config)#ip multicast-routing distributed
cr23-3560-LB(config)#ip pim rp-address 10.100.100.100

cr23-3560-LB(config)#interface range Vlan 101 - 110
cr22-3560-LB(config-if-range)#ip pim sparse-mode

cr22-3560-LB#show ip pim rp
Group: 239.192.51.1, RP: 10.100.100.100, uptime 00:01:36, expires never
Group: 239.192.51.2, RP: 10.100.100.100, uptime 00:01:36, expires never
Group: 239.192.51.3, RP: 10.100.100.100, uptime 00:01:36, expires never
Group: 224.0.1.40, RP: 10.100.100.100, uptime 5w5d, expires never
cr22-3560-LB#show ip pim interface
```

```
Address          Interface                Ver/   Nbr    Query  DR     DR
                      Mode   Count  Intvl  Prior
10.125.0.5       Port-channel1            v2/S   1      30     1      10.125.0.5
10.125.101.1     Vlan101                  v2/S   0      30     1      0.0.0.0
…
10.125.103.65   Vlan110                  v2/S   0      30     1      10.125.103.65

cr22-3560-LB#show ip mroute sparse
(*, 239.192.51.1), 00:06:06/00:02:59, RP 10.100.100.100, flags: SC
  Incoming interface: Port-channel1, RPF nbr 10.125.0.4
  Outgoing interface list:
    Vlan101, Forward/Sparse, 00:06:08/00:02:09
    Vlan110, Forward/Sparse, 00:06:06/00:02:05
```

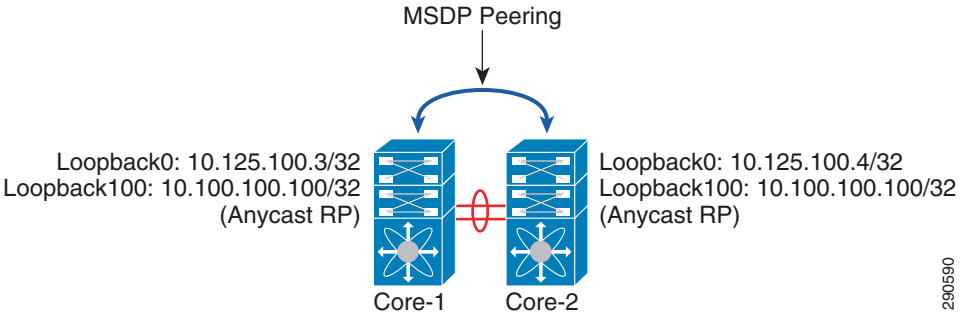
- WAN edge layer

```
cr11-asr-we(config)#ip multicast-routing distributed

cr11-asr-we(config)#ip pim rp-address 10.100.100.100

cr11-asr-we(config)#interface range Port-channel1 , Gig0/2/0 , Gig0/2/1.102
cr11-asr-we(config-if-range)#ip pim sparse-mode
cr11-asr-we(config)#interface Ser0/3/0
cr11-asr-we(config-if)#ip pim sparse-mode

cr11-asr-we#show ip pim rp
Group: 239.192.57.1, RP: 10.100.100.100, uptime 00:23:16, expires never
Group: 239.192.57.2, RP: 10.100.100.100, uptime 00:23:16, expires never
Group: 239.192.57.3, RP: 10.100.100.100, uptime 00:23:16, expires never

cr11-asr-we#show ip mroute sparse

(*, 239.192.57.1), 00:24:08/stopped, RP 10.100.100.100, flags: SP
  Incoming interface: Port-channel1, RPF nbr 10.125.0.22
  Outgoing interface list: Null

(10.125.31.156, 239.192.57.1), 00:24:08/00:03:07, flags: T
  Incoming interface: Port-channel1, RPF nbr 10.125.0.22
  Outgoing interface list:
    Serial0/3/0, Forward/Sparse, 00:24:08/00:02:55

cr11-asr-we#show ip mroute active
Active IP Multicast Sources - sending >= 4 kbps

Group: 239.192.57.1, (?)
   Source: 10.125.31.156 (?)
     Rate: 625 pps/1130 kbps(1sec), 1130 kbps(last 40 secs), 872 kbps(life avg)
```

PIM-SM RP Redundancy

PIM-SM RP redundancy and load sharing becomes imperative in the enterprise campus design because each recommended core layer design model provides resiliency and simplicity. In the Cisco Catalyst 6500 VSS-enabled core layer, the dynamically discovered group-to-RP entries are fully synchronized to the standby switch. Combining NSF/SSO capabilities with IPv4 multicast reduces the network recovery time and retains the user and application performance at an optimal level. In the non-VSS-enabled network design, PIM-SM uses Anycast RP and Multicast Source Discovery Protocol (MSDP) for node failure protection. PIM-SM redundancy and load sharing is simplified with the Cisco VSS-enabled core. Because VSS is logically a single system and provides node protection, there is no need to implement Anycast RP and MSDP on a VSS-enabled PIM-SM RP. If the campus core is deployed with redundant Cisco Nexus 7000 systems, then it is recommended to deploy Anycast-RP with MSDP between both core layer systems to provide multicast RP redundancy in the campus. Figure 41 illustrates enabling Anycast-RP with MSDP protocol between Cisco Nexus 7000 core layer systems in the campus network.

Figure 41 Cisco Nexus 7000 RP Redundancy Design



Implementing MSDP Anycast RP

Core-1/Core-2

```
cr35-N7K-Core1(config)#feature msdp
!Enable MSDP feature set

cr35-N7K-Core1(config)# interface Loopback0, Loopback254
cr35-N7K-Core1(config-if)# ip router eigrp 100
!Enables loopback interface route advertisement
```

Core-1

```
cr35-N7K-Core1(config)# ip msdp peer 10.125.100.4 connect-source loopback0
cr35-N7K-Core1(config)# ip msdp description 10.125.100.4 ANYCAST-PEER-N7K-LrgCampus
!Enables MSDP peering with Core-2
```

Core-2

```
cr35-N7K-Core2(config)# ip msdp peer 10.125.100.3 connect-source loopback0
cr35-N7K-Core2(config)# ip msdp description 10.125.100.3 ANYCAST-PEER-N7K-LrgCampus
!Enables MSDP peering with Core-1

cr35-N7K-Core1# show ip msdp peer 10.125.100.4 | inc "peer|local|status"
MSDP peer 10.125.100.4 for VRF "default"
AS 0, local address: 10.125.100.3 (loopback0)
  Connection status: Established

cr35-N7K-Core2# show ip msdp peer 10.125.100.3 | inc "peer|local|status"
MSDP peer 10.125.100.3 for VRF "default"
AS 0, local address: 10.125.100.4 (loopback0)
  Connection status: Established
```

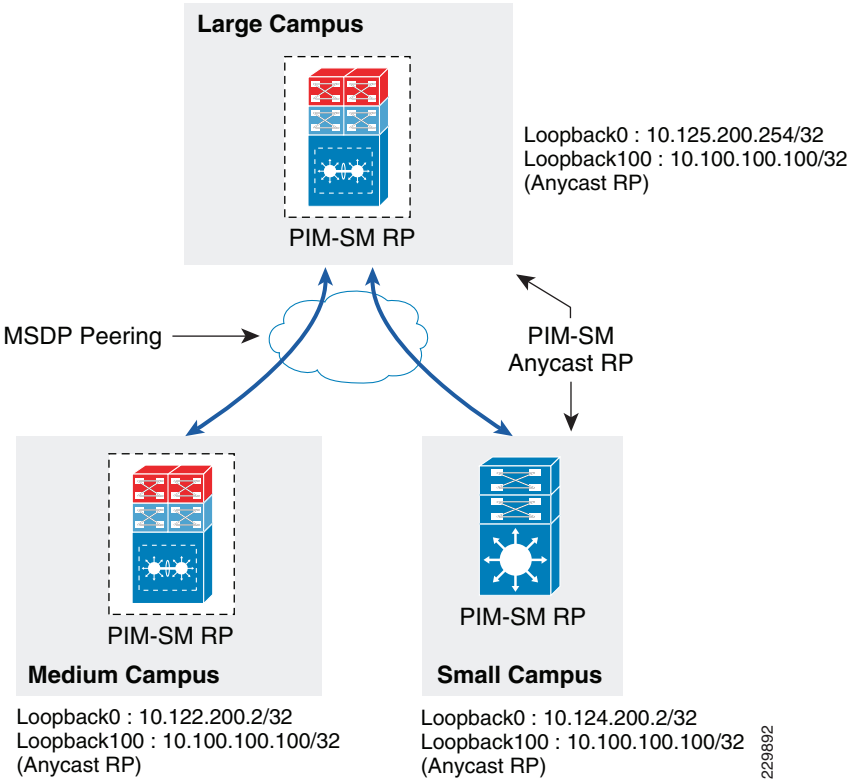
Inter-Site PIM Anycast RP

MSDP allows PIM RPs to share information about the active sources. PIM-SM RPs discover local receivers through PIM join messages, while the multicast source can be in a local or remote network domain. MSDP allows each multicast domain to maintain an independent RP that does not rely on other multicast domains, but does enable RPs to forward traffic between domains. PIM-SM is used to forward the traffic between the multicast domains.

Anycast RP is a useful application of MSDP. Originally developed for inter-domain multicast applications, MSDP used with Anycast RP is an intra-domain feature that provides redundancy and load sharing capabilities. Large networks typically use Anycast RP for configuring a PIM-SM network to meet fault tolerance requirements within a single multicast domain.

The enterprise campus multicast network must be designed with Anycast RP. PIM-SM RP at the main or the centralized core must establish an MSDP session with RP on each remote site to exchange distributed multicast source information and allow RPs to join SPT to active sources as needed. Figure 42 shows an example of an enterprise campus multicast network design.

Figure 42 Inter-Site Multicast Network Design



Implementing Inter-Site MSDP Anycast RP

Large Campus

```
cr23-VSS-Core(config)#ip msdp peer 10.122.200.2 connect-source Loopback0
cr23-VSS-Core(config)#ip msdp description 10.122.200.2 ANYCAST-PEER-6k-RemoteLrgCampus
cr23-VSS-Core(config)#ip msdp peer 10.124.200.2 connect-source Loopback0
cr23-VSS-Core(config)#ip msdp description 10.124.200.2 ANYCAST-PEER-4k-RemoteSmlCampus
cr23-VSS-Core(config)#ip msdp cache-sa-state
cr23-VSS-Core(config)#ip msdp originator-id Loopback0

cr23-VSS-Core#show ip msdp peer | inc MSDP Peer|State
MSDP Peer 10.122.200.2 (?), AS ?
    State: Up, Resets: 0, Connection source: Loopback0 (10.125.200.254)
MSDP Peer 10.123.200.1 (?), AS ?
```

```
    State: Up, Resets: 0, Connection source: Loopback0 (10.125.200.254)
MSDP Peer 10.124.200.2 (?), AS ?
    State: Up, Resets: 0, Connection source: Loopback0 (10.125.200.254)
```

Medium Campus

```
cr14-6500-RLC(config)#ip msdp peer 10.125.200.254 connect-source Loopback0
cr14-6500-RLC(config)#ip msdp description 10.125.200.254 ANYCAST-PEER-6k-MainCampus
cr14-6500-RLC(config)#ip msdp cache-sa-state
cr14-6500-RLC(config)#ip msdp originator-id Loopback0

cr14-6500-RLC#show ip msdp peer | inc MSDP Peer|State|SAs learned
MSDP Peer 10.125.200.254 (?), AS ?
State: Up, Resets: 0, Connection source: Loopback0 (10.122.200.2)
SAs learned from this peer: 94
```

Small Campus

```
cr14-4507-RSC(config)#ip msdp peer 10.125.200.254 connect-source Loopback0
cr14-4507-RSC(config)#ip msdp description 10.125.200.254 ANYCAST-PEER-6k-MainCampus
cr14-4507-RSC(config)#ip msdp cache-sa-state
cr14-4507-RSC(config)#ip msdp originator-id Loopback0

cr14-4507-RSC#show ip msdp peer | inc MSDP Peer|State|SAs learned
MSDP Peer 10.125.200.254 (?), AS ?
    State: Up, Resets: 0, Connection source: Loopback0 (10.124.200.2)
  SAs learned from this peer: 94
```

Dynamic Group Membership

Multicast receiver registration is done via IGMP protocol signaling. IGMP is an integrated component of an IP multicast framework that allows the receiver hosts and transmitting sources to be dynamically added to and removed from the network. Without IGMP, the network is forced to flood rather than multicast the transmissions for each group. IGMP operates between a multicast receiver host in the access layer and the Layer 3 router at the distribution layer.

The multicast system role changes when the access layer is deployed in the multilayer and routed access models. Because multilayer access switches do not run PIM, it becomes complex to make forwarding decisions out of the receiver port. In such a situation, Layer 2 access switches flood the traffic on all ports. This multilayer limitation in access switches is solved by using the IGMP snooping feature, which is enabled by default and is recommended to not be disabled.

IGMP is still required when a Layer 3 access layer switch is deployed in the routed access network design. Because the Layer 3 boundary is pushed down to the access layer, IGMP communication is limited between a receiver host and the Layer 3 access switch. In addition to the unicast routing protocol, PIM-SM must be enabled at the Layer 3 access switch to communicate with RPs in the network.

Implementing IGMP

By default, the Layer 2 access switch dynamically detects IGMP hosts and multicast-capable Layer 3 PIM routers in the network. The IGMP snooping and multicast router detection functions on a per-VLAN basis and is globally enabled by default for all the VLANs.

The multicast routing function changes when the access switch is deployed in routed access mode. PIM operation is performed at the access layer; therefore, the multicast router detection process is eliminated. The following output from a Layer 3 switch verifies that the local multicast ports are in router mode and provide a snooped Layer 2 uplink port channel which is connected to the collapsed core router for multicast routing:

The IGMP configuration can be validated using the following **show** command on the Layer 2 and Layer 3 access switch:

Layer 2 Access

```
cr22-3750-LB#show ip igmp snooping groups
Vlan      Group                Type        Version     Port List
------------------------------------------------------------------
110       239.192.51.1         igmp        v2          Gi1/0/20, Po1
110       239.192.51.2         igmp        v2          Gi1/0/20, Po1
110       239.192.51.3         igmp        v2          Gi1/0/20, Po1

cr22-3750-LB#show ip igmp snooping mrouter
Vlan    ports
-------  -------
110    Po1(dynamic)
```

Layer 3 Access

```
cr22-3560-LB#show ip igmp membership
 Channel/Group                Reporter        Uptime   Exp.  Flags  Interface
*,239.192.51.1               10.125.103.106  00:52:36 02:09 2A     Vl110
 *,239.192.51.2               10.125.103.107  00:52:36 02:12 2A     Vl110
 *,239.192.51.3               10.125.103.109  00:52:35 02:16 2A     Vl110
 *,224.0.1.40                 10.125.0.4      3d22h    02:04 2A     Po1
 *,224.0.1.40                 10.125.101.129  4w4d     02:33 2LA    Vl103

cr22-3560-LB#show ip igmp snooping mrouter
Vlan    ports
```

```
-------  ------
 103     Router
 106     Router
 110     Router
```

Designing Multicast Security

When designing multicast security in the borderless enterprise campus network design, two key concerns are preventing a rogue source and preventing a rogue PIM-RP.

Preventing Rogue Source

In a PIM-SM network, an unwanted traffic source can be controlled with the **pim accept-register** command. When the source traffic hits the first-hop router, the first-hop router (DR) creates the (S,G) state and sends a PIM source register message to the RP. If the source is not listed in the accept-register filter list (configured on the RP), the RP rejects the register and sends back an immediate Register-Stop message to the DR. The drawback with this method of source filtering is that with the **pim accept-register** command on the RP, the PIM-SM (S,G) state is still created on the first-hop router of the source. This can result in traffic reaching receivers local to the source and located between the source and the RP. Furthermore, because the **pim accept-register** command works on the control plane of the RP, this can be used to overload the RP with fake register messages and possibly cause a DoS condition.

The following is the sample configuration with a simple ACL that has been applied to the RP to filter only on the source address. It is also possible to filter the source and the group using an extended ACL on the RP:

Cisco IOS

```
cr23-VSS-Core(config)#ip access-list extended PERMIT-SOURCES
cr23-VSS-Core(config-ext-nacl)# permit ip 10.120.31.0 0.7.0.255 239.192.0.0 0.0.255.255
cr23-VSS-Core(config-ext-nacl)# deny   ip any any

cr23-VSS-Core(config)#ip pim accept-register list PERMIT-SOURCES
```

Cisco NX-OS

```
cr35-N7K-Core1(config)# route-map PERMIT_SOURCES permit 10
cr35-N7K-Core1(config-route-map)# match ip multicast source 10.120.31.0/24 group-range
239.192.0.0 to 239.192.255.255

cr35-N7K-Core1(config)#ip pim register-policy PERMIT_SOURCES
```

Preventing Rogue PIM-RP

Like the multicast source, any router can be misconfigured or can maliciously advertise itself as a multicast RP in the network with the valid multicast group address. With a static RP configuration, each PIM-enabled router in the network can be configured to use static RP for the multicast source and override any other Auto-RP or BSR multicast router announcement from the network.

The following is the sample configuration that must be applied to each PIM-enabled router in the campus network to accept PIM announcements only from the static RP and ignore dynamic multicast group announcement from any other RP:

Cisco IOS

```
cr23-VSS-Core(config)#ip access-list standard Allowed_MCAST_Groups
cr23-VSS-Core(config-std-nacl)# permit 224.0.1.39
cr23-VSS-Core(config-std-nacl)# permit 224.0.1.40
cr23-VSS-Core(config-std-nacl)# permit 239.192.0.0 0.0.255.255
cr23-VSS-Core(config-std-nacl)# deny   any

cr23-VSS-Core(config)#ip pim rp-address 10.100.100.100 Allowed_MCAST_Groups override
```

Cisco NX-OS

```
cr35-N7K-Core1(config)# route-map Allowed_MCAST_Groups permit 10
cr35-N7K-Core1(config-route-map)#   match ip address Allowed_MCAST_Groups
cr35-N7K-Core1(config-route-map)#   match ip multicast group 224.0.1.39/32
cr35-N7K-Core1(config-route-map)# route-map Allowed_MCAST_Groups permit 20
cr35-N7K-Core1(config-route-map)#   match ip multicast group 224.0.1.40/32
cr35-N7K-Core1(config-route-map)# route-map Allowed_MCAST_Groups permit 30
cr35-N7K-Core1(config-route-map)#   match ip multicast group 239.192.0.0/16

cr35-N7K-Core1(config)# ip pim rp-address 10.100.100.100 route-map Allowed_MCAST_Groups
override
```

Designing the LAN network for the enterprise network design establishes the foundation for all other aspects within the service fabric (WAN, security, mobility, and UC) as well as laying the foundation to provide safety and security, operational efficiencies, etc.

4 Summary

This Borderless Campus 1.0 chapter focuses on designing campus hierarchical layers and provides design and deployment guidance. Network architects should leverage Cisco recommended best practices to deploy key network foundation services such as routing, switching, QoS, multicast, and high availability. Best practices are provided for the entire enterprise design.

3

Deploying QoS for Application Performance Optimization

Expectations have evolved significantly over the past few years as users and endpoints use the network in ever-evolving ways and increasingly expect guaranteed low latency bandwith. Application and device awareness have become key tools in providing differentiated service treatment at the campus edge. Media applications, particularly video-oriented media applications, are evolving as the enterprise network conducts business digitally. There are also increased campus network and asset security requirements. Integrating video applications in the enterprise campus network exponentially increases bandwidth utilization and fundamentally shifts traffic patterns. Business drivers behind this media application growth include remote business operations, as well as leveraging the network as a platform to build an energy-efficient network to minimize cost and go “green”. High-definition media is transitioning from the desktop to conference rooms and social networking phenomena are crossing over into enterprise settings. Besides internal and enterprise research applications, media applications are fueling a new wave of IP convergence, requiring the ongoing development of converged network designs.

Converging media applications onto an IP network is much more complex than converging voice over IP (VoIP). Media applications are diverse, generally bandwidth-intensive, and bursty (as compared to VoIP). In addition to IP telephony, applications can include live and on-demand streaming media applications, digital signage applications, high-definition room-based conferencing applications, as well as an infinite array of data-oriented applications. By embracing media applications as the next cycle of convergence, enterprise IT departments can think holistically about their network design and its readiness to support the coming tidal wave of media applications and develop a network-wide strategy to ensure a high quality end user experiences.

The Borderless Campus infrastructure must set administrative policies to provide differentiated forwarding services to network applications, users, and endpoints to prevent contention. The characteristics of network services and applications must be well understood so that policies can be defined that allow network resources to be used for internal applications, to provide best-effort services for external traffic, and to keep the network protected from threats.

Policies for providing network resources to an internal application are further complicated when interactive video and real-time VoIP applications are converged over the same network that is switching mid-to-low priority data traffic. Deploying QoS technologies in the campus allows different types of traffic to contend inequitably for network resources. Real-time applications such as voice, interactive, and physical security video can be given priority or preferential services over generic data applications, but not to the point that data applications are starved for bandwidth.

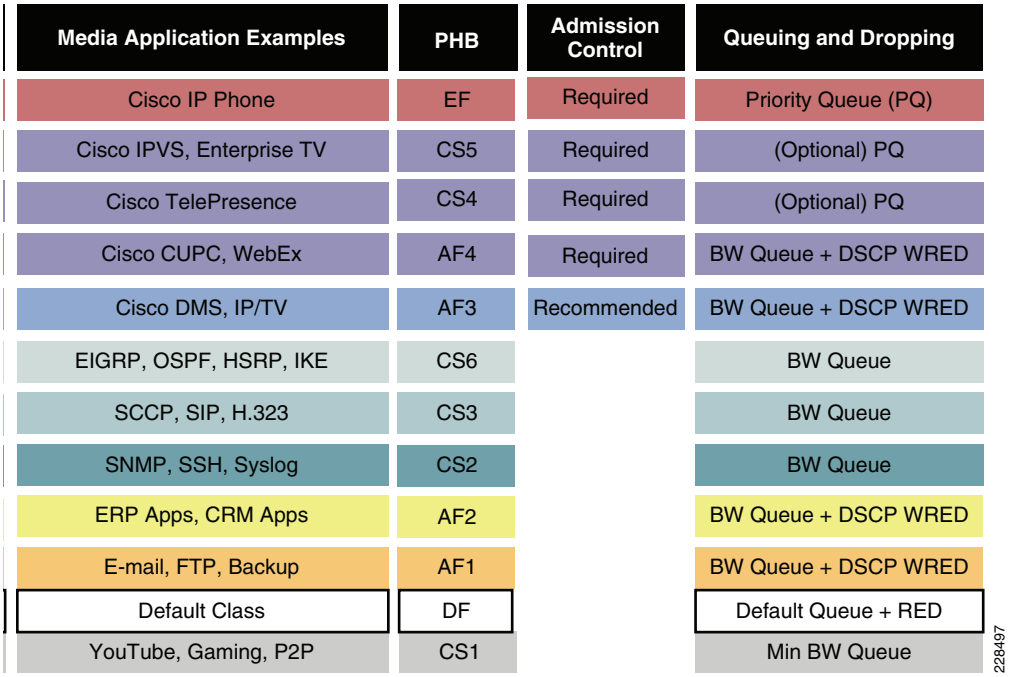
1 Enterprise Campus QoS Framework

Each group of managed and un-managed applications with unique traffic patterns and service level requirements requires a dedicated QoS class to provision and guarantee these service level requirements. The enterprise campus network architect may need to determine the number of classes for various applications, as well as how these individual classes should be implemented to consistently deliver differentiated services in main and remote campus sites. Cisco recommends following the relevant industry standards and guidelines whenever possible to extend the effectiveness of your QoS policies beyond your direct administrative control.

With minor changes, the enterprise campus QoS framework is developed based on RFC4594 that follows industry standard and guidelines to function consistently in heterogeneous network environments. These guidelines are to be viewed as industry best-practice recommendations. Enterprise organizations and service providers are encouraged to adopt these marking and provisioning recommendations with the aim of improving QoS consistency, compatibility, and interoperability. However, because these guidelines are not standards, modifications can be made to these recommendations as specific needs or constraints require. To this end, to meet specific business requirements, Cisco has made a minor modification to its adoption of RFC 4594, namely the switching of call signaling and broadcast video markings (to CS3 and CS5, respectively).

RFC 4594 outlines twelve classes of media applications that have unique service level requirements, as shown in Figure 1-1.

Figure 1-1 Campus 12-Class QoS Policy Recommendation



The twelve classes are:

- *VoIP telephony*—This service class is intended for VoIP telephony (bearer only) traffic (VoIP signaling traffic is assigned to the call signaling class). Traffic assigned to this class should be marked EF. This class is provisioned with expedited forwarding (EF) per-hop behavior (PHB). The EF PHB defined in RFC 3246 is a strict priority queuing service and, as such, admission to this class should be controlled (admission control is discussed in the following section). Examples of this type of traffic include G.711 and G.729a.
- *Broadcast video*—This service class is intended for broadcast TV, live events, video surveillance flows, and similar *inelastic* streaming video flows, which are highly drop sensitive and have no retransmission and/or flow control capabilities. Traffic in this class should be marked class selector 5 (CS5) and may be provisioned with an EF PHB; as such, admission to this class should be controlled. Examples of this traffic include live Cisco Digital Media System (DMS) streams to desktops or to Cisco Digital Media Players (DMPs), live Cisco Enterprise TV (ETV) streams, and Cisco IP Video Surveillance.
- *Real-time interactive*—This service class is intended for (inelastic) room-based, high-definition interactive video applications and is intended primarily for the voice and video components of these applications. Whenever technically possible and administratively feasible, data

sub-components of this class can be separated out and assigned to the transactional data traffic class. Traffic in this class should be marked CS4 and may be provisioned with an EF PHB; as such, admission to this class should be controlled. A sample application is Cisco TelePresence.

- *Multimedia conferencing*—This service class is intended for desktop software multimedia collaboration applications and is intended primarily for the voice and video components of these applications. Whenever technically possible and administratively feasible, data sub-components of this class can be separated out and assigned to the transactional data traffic class. Traffic in this class should be marked assured forwarding (AF) Class 4 (AF41) and should be provisioned with a guaranteed bandwidth queue with Differentiated Services Code Point (DSCP)-based Weighted Random Early Detection (WRED) enabled. Admission to this class should be controlled; additionally, traffic in this class may be subject to policing and re-marking. Sample applications include Cisco Unified Personal Communicator, Cisco Unified Video Advantage, and the Cisco Unified IP Phone 7985G.
- *Multimedia streaming*—This service class is intended for video-on-demand (VoD) streaming video flows, which, in general, are more elastic than broadcast/live streaming flows. Traffic in this class should be marked AF Class 3 (AF31) and should be provisioned with a guaranteed bandwidth queue with DSCP-based WRED enabled. Admission control is recommended on this traffic class (though not strictly required) and this class may be subject to policing and re-marking. Sample applications include Cisco Digital Media System VoD streams.
- *Network control*—This service class is intended for network control plane traffic, which is required for reliable operation of the enterprise network. Traffic in this class should be marked CS6 and provisioned with a (moderate, but dedicated) guaranteed bandwidth queue. WRED should not be enabled on this class, because network control traffic should not be dropped (if this class is experiencing drops, the bandwidth allocated to it should be re-provisioned). Sample traffic includes EIGRP, OSPF, Border Gateway Protocol (BGP), HSRP, Internet Key Exchange (IKE), and so on.
- *Call-signaling*—This service class is intended for signaling traffic that supports IP voice and video telephony. Traffic in this class should be marked CS3 and provisioned with a (moderate, but dedicated) guaranteed bandwidth queue. WRED should not be enabled on this class, because call-signaling traffic should not be dropped (if this class is experiencing drops, the bandwidth allocated to it should be re-provisioned). Sample traffic includes Skinny Call Control Protocol (SCCP), Session Initiation Protocol (SIP), H.323, and so on.
- *Operations/administration/management (OAM)*—This service class is intended for network operations, administration, and management traffic. This class is critical to the ongoing maintenance and support of the network. Traffic in this class should be marked CS2 and provisioned with a (moderate, but dedicated) guaranteed bandwidth queue. WRED should not be enabled on this class, because OAM traffic should not be dropped (if this class is experiencing drops, the bandwidth allocated to it should be re-provisioned). Sample traffic includes Secure Shell (SSH), Simple Network Management Protocol (SNMP), Syslog, and so on.

- *Transactional data (or low-latency data)*—This service class is intended for interactive, "foreground" data applications (foreground refers to applications from which users are expecting a response via the network to continue with their tasks; excessive latency directly impacts user productivity). Traffic in this class should be marked AF Class 2 (AF21) and should be provisioned with a dedicated bandwidth queue with DSCP-WRED enabled. This traffic class may be subject to policing and re-marking. Sample applications include data components of multimedia collaboration applications, Enterprise Resource Planning (ERP) applications, Customer Relationship Management (CRM) applications, database applications, and so on.
- *Bulk data (or high-throughput data)*—This service class is intended for non-interactive "background" data applications (background refers to applications from which users are not awaiting a response via the network to continue with their tasks; excessive latency in response times of background applications does not directly impact user productivity). Traffic in this class should be marked AF Class 1 (AF11) and should be provisioned with a dedicated bandwidth queue with DSCP-WRED enabled. This traffic class may be subject to policing and re-marking. Sample applications include E-mail, backup operations, FTP/SFTP transfers, video and content distribution, and so on.
- *Best effort (or default class)*—This service class is the default class. The vast majority of applications will continue to default to this best-effort service class; as such, this default class should be adequately provisioned. Traffic in this class is marked default forwarding (DF or DSCP 0) and should be provisioned with a dedicated queue. WRED is recommended to be enabled on this class.
- *Scavenger (or low-priority data)*—This service class is intended for non-business-related traffic flows, such as data or video applications that are entertainment and/or gaming-oriented. The approach of a less-than Best-Effort service class for non-business applications (as opposed to shutting these down entirely) has proven to be a popular, political compromise. These applications are permitted on enterprise networks, as long as resources are always available for business-critical voice, video, and data applications. However, as soon as the network experiences congestion, this class is the first to be penalized and aggressively dropped. Traffic in this class should be marked CS1 and should be provisioned with a minimal bandwidth queue that is the first to starve should network congestion occur. Sample traffic includes YouTube, Xbox Live/360 movies, iTunes, BitTorrent, and so on.

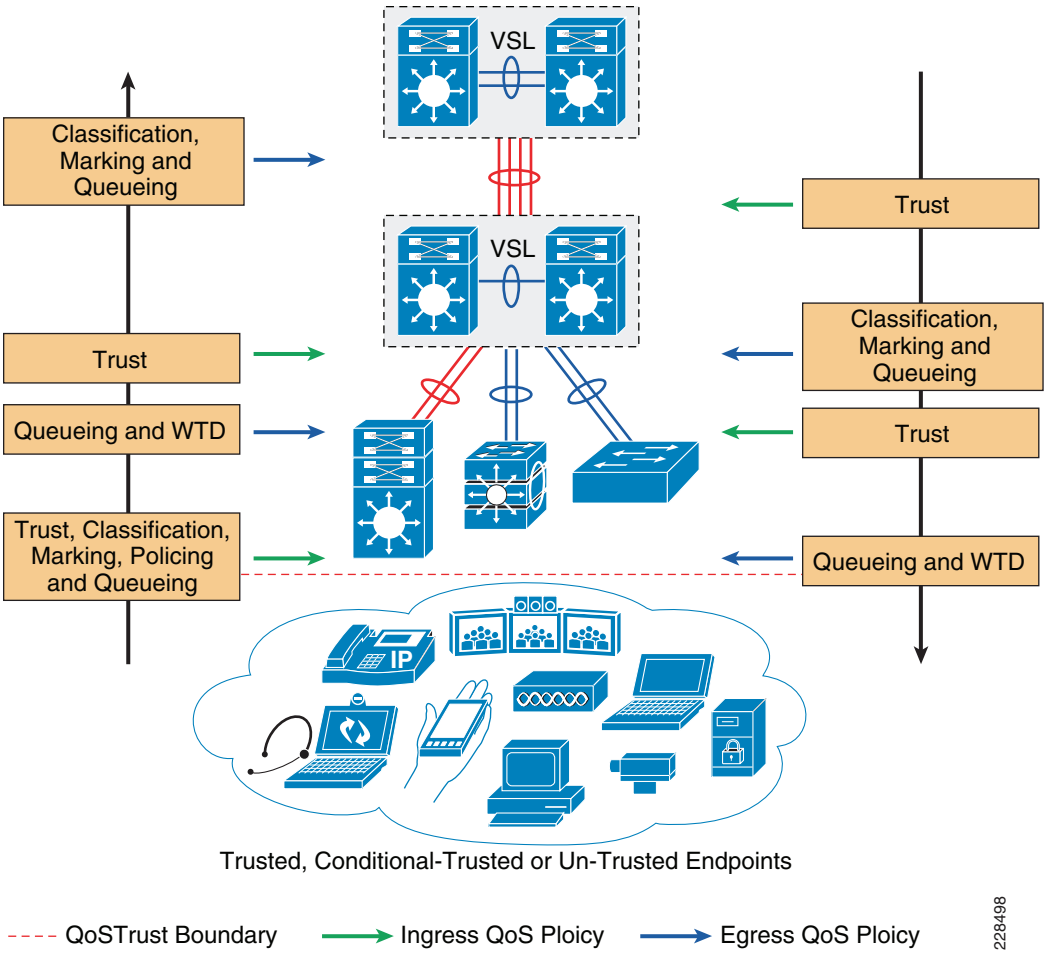
2 Designing Enterprise Campus QoS Trust Boundary and Policies

To build an end-to-end QoS framework that offers transparent and consistent QoS service without compromising performance, it is important to create an blueprint of the network, defining sets of trusted applications, devices, and forwarding paths, and then defining common QoS policy settings independently of how QoS is implemented within the system.

QoS settings applied at the campus network edge set the ingress rule based on deep packet classification and mark the traffic before it is forwarded inside the campus core. To retain the marking set by access layer switches, it is important that other LAN network devices in the campus trust the marking and apply the same policy to retain the QoS settings and offer symmetric treatment. Bi-directional network communication between applications, endpoints, or other network devices requires the same treatment when traffic enters or leaves the network and must be taken into account when designing the trust model between network endpoints and core and edge campus devices.

The trust or un-trust model simplifies the rules for defining bi-directional QoS policy settings. Figure 2 shows the QoS trust model setting that sets the QoS implementation guidelines in Borderless Campus networks.

Figure 2 Borderless Campus QoS Trust and Policies



3 Enterprise Campus QoS Overview

With an overall application strategy in place, end-to-end QoS policies can be designed for each device and interface, as determined by their roles in the network infrastructure. However, because the Cisco QoS toolset provides many QoS design and deployment options, a few succinct design principles can help simplify strategic QoS deployments, as discussed in the following sections.

Hardware versus Software QoS

A fundamental QoS design principle is, whenever possible, to enable QoS policies in hardware rather than software. Cisco IOS routers perform QoS in software, which places incremental loads on the CPU, depending on the complexity and functionality of the policy. Cisco Catalyst switches, on the other hand, perform QoS in dedicated hardware application-specific integrated circuits (ASICs) on Ethernet-based ports and as such do not tax their main CPUs to administer QoS policies. This allows complex policies to be applied at line rates even up to Gigabit or 10-Gigabit speeds.

Classification and Marking

When classifying and marking traffic, a recommended design principle is to classify and mark applications as close to their sources as technically and administratively feasible. This principle promotes end-to-end differentiated services and PHBs.

In general, it is not recommended to trust markings that can be set by users on their PCs or similar devices, because users can easily abuse provisioned QoS policies if permitted to mark their own traffic. For example, if an EF PHB has been provisioned over the network, a PC user can easily configure all their traffic to be marked to EF, thus hijacking network priority queues to service non-realtime traffic. Such abuse can easily ruin the service quality of realtime applications throughout the campus. On the other hand, if enterprise network administrator controls are in place that centrally administer PC QoS markings, it may be possible and advantageous to trust these.

Following this rule, it is recommended to use DSCP markings whenever possible, because these are end-to-end, more granular, and more extensible than Layer 2 markings. Layer 2 markings are lost when the media changes (such as a LAN-to-WAN/VPN edge). There is also less marking granularity at Layer 2. For example, 802.1P supports only three bits (values 0-7), as does Multiprotocol Label Switching Experimental (MPLS EXP). Therefore, only up to eight classes of traffic can be supported at Layer 2 and inter-class relative priority (such as RFC 2597 Assured Forwarding Drop Preference markdown) is not supported. Layer 3-based DSCP markings allow for up to 64 classes of traffic, which provides more flexibility and is adequate in large-scale deployments and for future requirements.

As the network border blurs between the borderless enterprise network and service providers, the need for interoperability and complementary QoS markings is critical. Cisco recommends following the IETF standards-based DSCP PHB markings to ensure interoperability and future expansion. Because enterprise voice, video, and data application marking recommendations are standards-based, as previously discussed, enterprises can easily adopt these markings to interface with service provider classes of service.

Policing and Markdown

There is little reason to forward unwanted traffic that gets policed and dropped by a subsequent tier node, especially when unwanted traffic is the result of DoS or worm attacks in the enterprise network. Excessive volume attack traffic can destabilize network systems, which can result in outages. Cisco recommends policing traffic flows as close to their sources as possible. This principle applies also to legitimate flows, because worm-generated traffic can masquerade under legitimate, well-known TCP/UDP ports and cause extreme amounts of traffic to be poured into the network infrastructure. Such excesses should be monitored at the source and marked down appropriately.

Whenever supported, markdown should be done according to standards-based rules, such as RFC 2597 (AF PHB). For example, excess traffic marked to AFx1 should be marked down to AFx2 (or AFx3 whenever dual-rate policing such as defined in RFC 2698 is supported). Following such markdowns, congestion management policies, such as DSCP-based WRED, should be configured to drop AFx3 more aggressively than AFx2, which in turn should be dropped more aggressively than AFx1.

Queuing and Dropping

Critical media applications require uncompromised performance and service guarantees regardless of network conditions. Enabling outbound queuing in each network tier provides end-to-end service guarantees during potential network congestion. This common principle applies to campus-to-WAN/Internet edges, where speed mismatches are most pronounced, and campus interswitch links, where oversubscription ratios create greater potential for network congestion.

Because each application class has unique service level requirements, each should optimally be assigned a dedicated queue. A wide range of platforms in varying roles exist in enterprise networks, so each must be bounded by a limited number of hardware or service provider queues. No fewer than four queues are required to support QoS policies for various types of applications, specifically:

- Realtime queue (to support a RFC 3246 EF PHB service)
- Guaranteed-bandwidth queue (to support RFC 2597 AF PHB services)
- Default queue (to support a RFC 2474 DF service)
- Bandwidth-constrained queue (to support a RFC 3662 scavenger service)

Additional queuing recommendations for these classes are discussed next.

Strict-Priority Queuing

The realtime or strict priority class corresponds to the RFC 3246 EF PHB. The amount of bandwidth assigned to the realtime queuing class is variable. However, if the majority of bandwidth is provisioned with strict priority queuing (which is effectively a FIFO queue), the overall effect is a dampening of

QoS functionality, both for latency- and jitter-sensitive realtime applications (contending with each other within the FIFO priority queue) and also for non-realtime applications (because these may periodically receive significant bandwidth allocation fluctuations, depending on the instantaneous amount of traffic being serviced by the priority queue). Remember that the goal of convergence is to enable voice, video, and data applications to transparently co-exist on a single enterprise network infrastructure. When realtime applications dominate a link, non-realtime applications fluctuate significantly in their response times, destroying the transparency of the converged network.

For example, consider a 45 Mbps DS3 link configured to support two Cisco TelePresence CTS-3000 calls with an EF PHB service. Assuming that both systems are configured to support full high definition, each such call requires 15 Mbps of strict-priority queuing. Before the TelePresence calls are placed, non-realtime applications have access to 100 percent of the bandwidth on the link; to simplify the example, assume there are no other realtime applications on this link. However, after these TelePresence calls are established, all non-realtime applications are suddenly contending for less than 33 percent of the link. TCP windowing takes effect and many applications hang, timeout, or become stuck in a non-responsive state, which usually translates into users calling the IT help desk to complain about the network (which happens to be functioning properly, albeit in a poorly-configured manner).

Note As previously discussed, Cisco IOS software allows the abstraction (and thus configuration) of multiple strict priority LLQs. In such a multiple LLQ context, this design principle applies to the sum of all LLQs to be within one-third of link capacity.

It is vitally important to understand that this strict priority queuing rule is simply a best practice design recommendation and is not a mandate. There may be cases where specific business objectives cannot be met while holding to this recommendation. In such cases, the enterprise network administrator must provision according to their detailed requirements and constraints. However, it is important to recognize the tradeoffs involved with over-provisioning strict priority traffic and its negative performance impact, both on other realtime flows and also on non-realtime-application response times.

And finally, any traffic assigned to a strict-priority queue should be governed by an admission control mechanism.

Best Effort Queuing

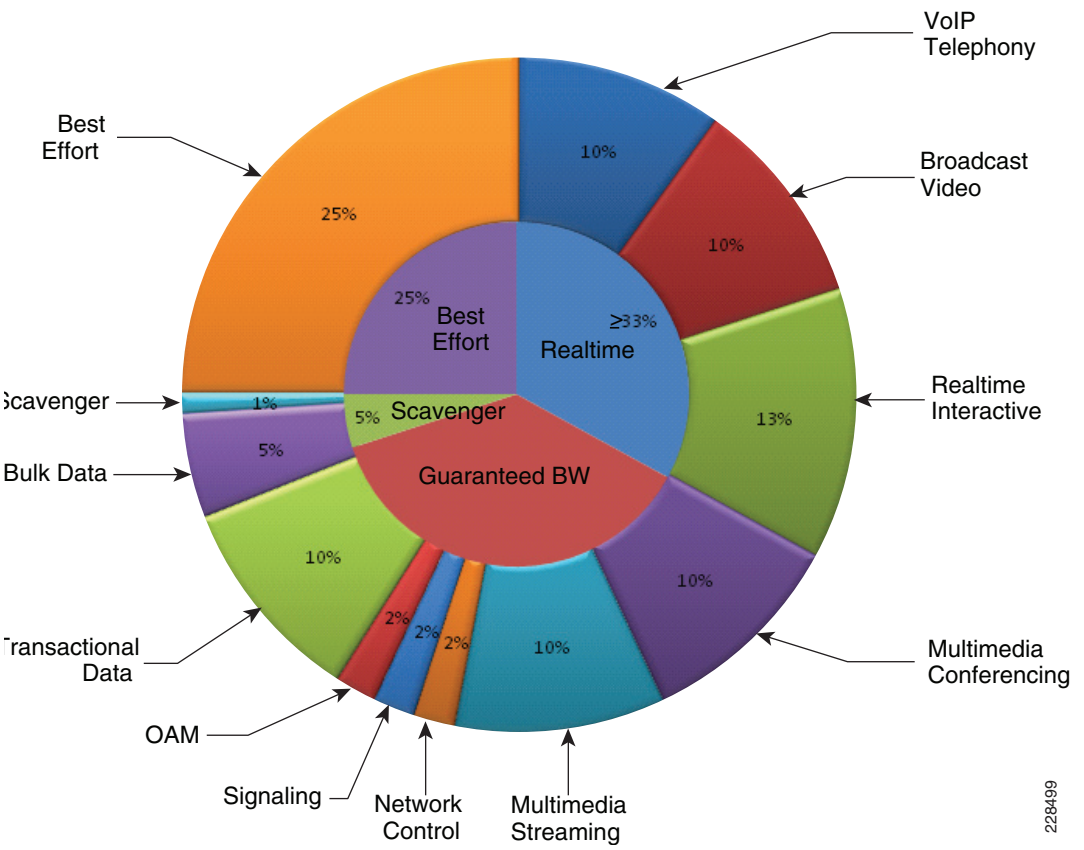
The best effort class is the default class for all traffic that has not been explicitly assigned to another application-class queue. Only if an application has been selected for preferential/deferential treatment is it removed from the default class. Because most enterprises may have several types of applications running in networks, adequate bandwidth must be provisioned for this class as a whole to handle the number and volume of applications that default to it. Therefore, Cisco recommends reserving at least 25 percent of link bandwidth for the default best effort class.

Scavenger Class Queuing

Whenever the scavenger queuing class is enabled, it should be assigned a minimal amount of link bandwidth capacity, such as 1 percent, or whatever minimal bandwidth allocation the platform supports. On some platforms, queuing distinctions between bulk data and scavenger traffic flows cannot be made, either because queuing assignments are determined by class of service (CoS) values (and both of these application classes share the same CoS value of 1) or because only a limited amount of hardware queues exist, precluding the use of separate dedicated queues for each of these two classes. In such cases, the scavenger/bulk queue can be assigned a moderate amount of bandwidth, such as 5 percent.

These queuing rules are summarized in Figure 3, where the inner pie chart represents a hardware or service provider queuing model that is limited to four queues and the outer pie chart represents a corresponding, more granular queuing model that is not bound by such constraints.

Figure 3 Compatible 4-Class and 12-Class Queuing Models



4 Deploying QoS in Borderless Campus Networks

All Layer 2 and Layer 3 systems in IP-based networks forward traffic based on best-effort, providing no differentiated services between different class-of-service network applications. The routing protocol forwards packets over the best low-metric or delay path, but offers no guarantee of delivery. This model works well for TCP-based data applications that adapt gracefully to variations in latency, jitter, and loss. The enterprise campus LAN and WAN is a multi-service network designed to support a wide range of low-latency voice and high bandwidth video with critical and non-critical data traffic over a single network infrastructure. For an optimal user experience, real time applications (such as voice and video) require packets to be delivered within specified loss, delay, and jitter parameters. Cisco QoS is a collection of features and hardware capabilities that allow the network to intelligently

dedicate the network resources for higher priority real-time applications, while reserving sufficient network resources to service medium to lower non-real-time traffic. QoS accomplishes this by creating a more application-aware Layer 2 and Layer 3 network to provide differentiated services to network applications and traffic. For a detailed discussion of QoS, refer to the *Enterprise QoS Design Guide* at:

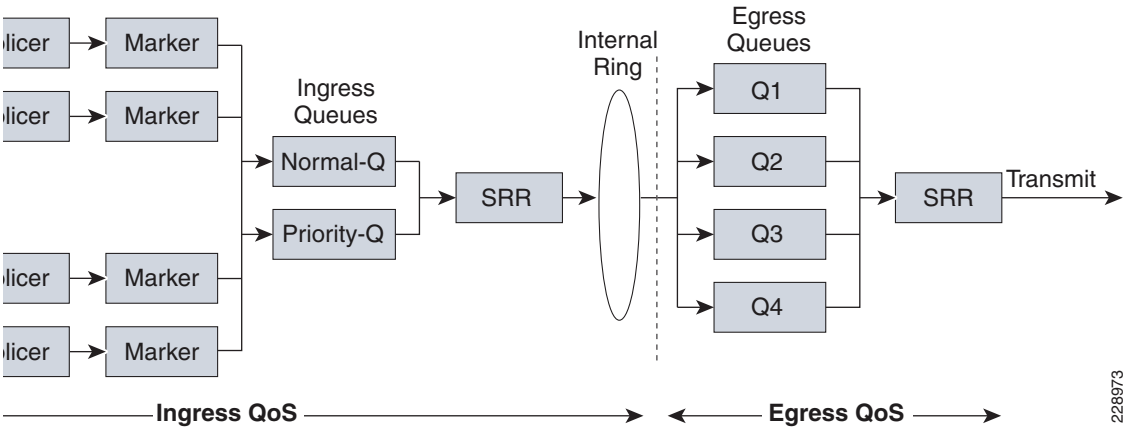
http://www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoS-SRND-Book.html

While the QoS design principles across the network are common, the QoS implementation in hardware- and software-based switching platforms vary due to internal system design. This section discusses the internal switching architecture and the differentiated QoS structure on a per-hop-basis.

QoS in Cisco Catalyst Fixed Configuration Switches

The QoS implementation in Cisco Catalyst 3560-X and 3750-X Series switches are similar. There is no difference in the ingress or egress packet classification, marking, queuing, and scheduling implementation among these Catalyst platforms. Cisco Catalyst switches allow users to create policy-maps by classifying incoming traffic (Layer 2 to Layer 4) and then attaching the policy-map to an individual physical port or to logical interfaces (SVI or port-channel). This creates a common QoS policy that may be used in multiple networks. To prevent switch fabric and egress physical port congestion, the ingress QoS policing structure can strictly filter excessive traffic at the network edge. All ingress traffic from edge ports passes through the switch fabric and moves to the egress ports, where congestion may occur. Congestion in access layer switches can be prevented by tuning queuing scheduler and Weighted Tail Drop (WTD) drop parameters. See Figure 1-4.

Figure 1-4 QoS Implementation in Cisco Catalyst Switches



The main difference between these platforms is the switching capacity, which ranges from 1G to 10G. The switching architecture and some of the internal QoS structure also differ. Some important differences to consider when selecting an access switch include:

- Only the Cisco Catalyst 3560-X and 3750-X support IPv6 QoS.
- Only the Cisco Catalyst 3560-X and 3750-X support policing on 10-Gigabit Ethernet interfaces.
- Only the Cisco Catalyst 3560-X and 3750-X support SRR shaping weights on 10-Gigabit Ethernet interfaces.

QoS in Cisco Catalyst Modular Switches

The Cisco Catalyst 4500E and 6500-E are high-density, resilient switches for large scale networks. The Borderless Campus design uses both platforms across the network; therefore, all the QoS recommendations in this section for these platforms will remain consistent. Both next-generation Catalyst platforms are modular in design; however, there are significant internal hardware architecture differences between the two platforms that impact the QoS implementation model.

Catalyst 4500E QoS

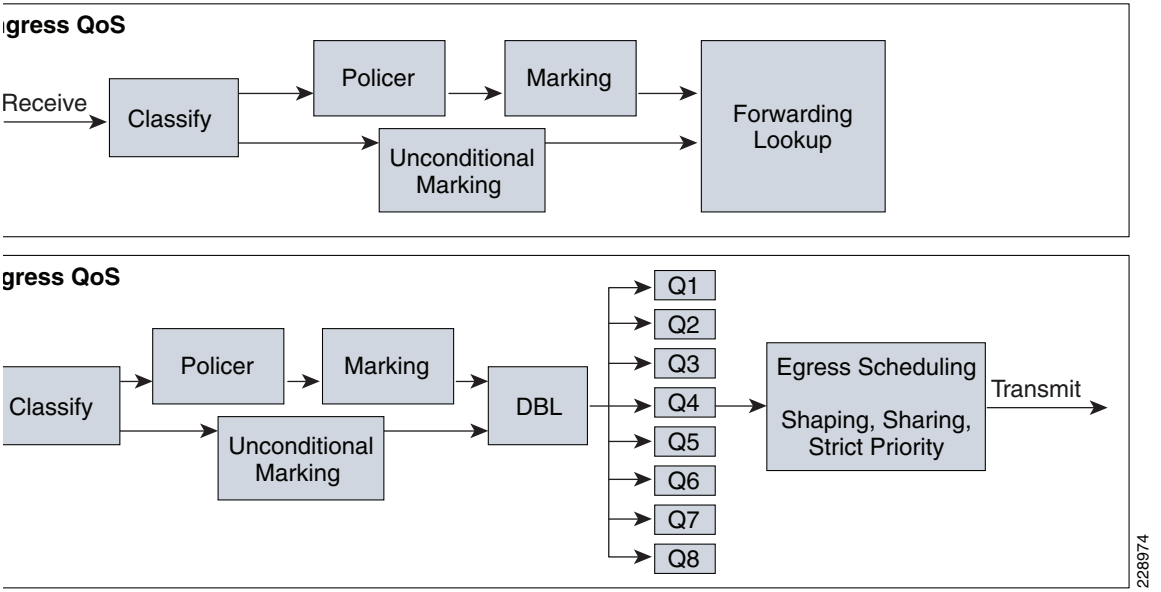
The Cisco Catalyst 4500E Series platforms are widely deployed with classic and next-generation supervisors. This design guide recommends deploying the next-generation supervisor Sup7-E and the current-generation Sup6-E and Sup6L-E that offer a number of technical benefits that are beyond QoS.

The Cisco Catalyst 4500E with Sup7-E, Sup-6E, and Sup6L-E (see Figure 5) is designed to offer better differentiated and preferential QoS services for various class-of-service traffic. New QoS capabilities in the Sup7-E, Sup-6E, and Sup6L-E enable administrators to take advantage of hardware-based intelligent classification and take action to optimize application performance and network availability. The QoS implementation in Sup7-E, Sup-6E, and Sup6L-E supports the Modular QoS CLI (MQC) as implemented in IOS-based routers that enhances QoS capabilities and eases implementation and operations. The following are some of the key QoS features that differentiate the next- and current-generation supervisor modules versus classic supervisors:

- Trust and Table-Map—MQC-based QoS implementation offers a number of implementation and operational benefits over classic supervisors that rely on the Trust model and internal Table-map as a tool to classify and mark ingress traffic.
- Internal DSCP—Unlike the classic supervisors that rely on the internal DSCP value, the queue placement in Sup7-E, Sup-6E, and Sup6L-E is simplified by leveraging the MQC capabilities to explicitly map any class of traffic (e.g., DSCP or CoS traffic) to an egress queue structure. For example, DSCP 46 can be classified with extended ACL and can be matched in PQ class-map of an MQC in Sup7-E, Sup-6E, and Sup6L-E.

- Sequential vs. Parallel Classification—With MQC-based QoS classification, the Sup7-E, Sup-6E, and Sup6L-E provide sequential classification rather than parallel. The sequential classification method allows network administrators to classify traffic at the egress based on the ingress markings.

Figure 5 Catalyst 4500E—Supervisor 7-E, Supervisor 6-E, and Supervisor 6L-E QoS Architecture



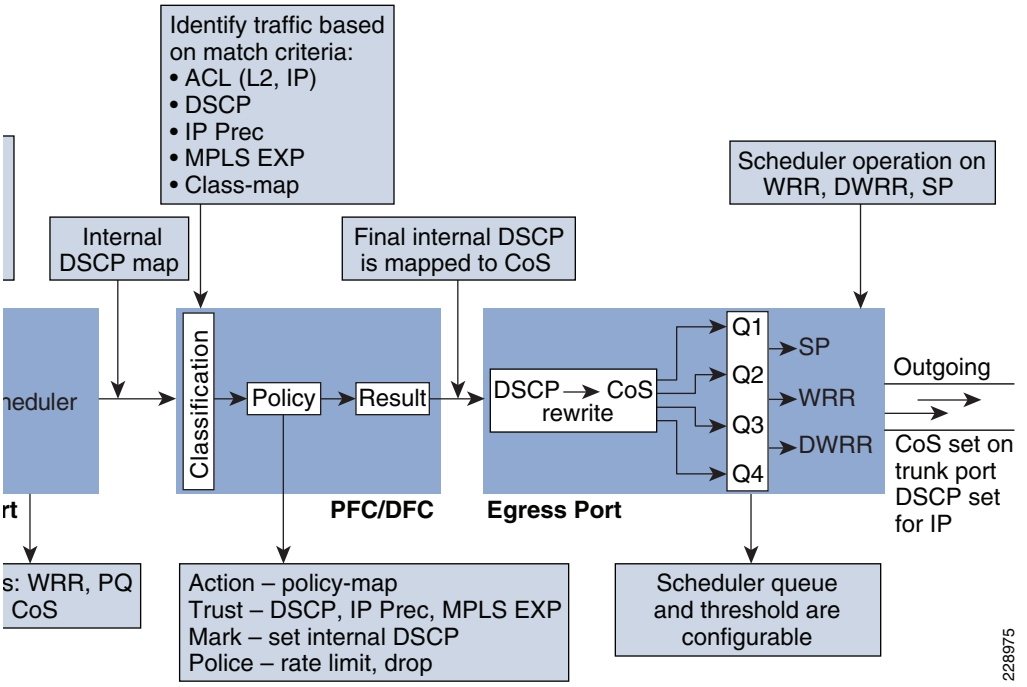
Catalyst 6500-E QoS

The Cisco Catalyst 6500-E Series are enterprise-class switches with next-generation hardware and software capabilities designed to deliver innovative, secure, converged borderless network services regardless of their place in the network. The Cisco Catalyst 6500-E can be deployed as a borderless service node in the campus network to offer a high performance, robust, and intelligent application and network awareness services. The Catalyst 6500-E provides leading-edge Layer 2-Layer 7 services, including rich high availability, manageability, virtualization, security, and QoS feature sets, as well as integrated Power-over-Ethernet (PoE), allowing for maximum flexibility in virtually any role within the campus.

Depending on the network services and application demands of the Cisco Catalyst 6500-E, the platform can be deployed with different types of Supervisor modules—Sup720-10GE, Sup720, and Sup32. This design guide uses the Sup720-10GE supervisor, which is built with next-generation hardware allowing administrators to build virtual network systems in a simplified and

highly-redundant enterprise campus network. These supervisors leverage various featured daughter cards, including the Multilayer Switch Feature Card (MSFC) that serves as the routing engine, the Policy Feature Card (PFC) that serves as the primary QoS engine, as well as various Distributed Feature Cards (DFCs) that serve to scale policies and processing. Specifically relating to QoS, the PFC sends a copy of the QoS policies to the DFC to provide local support for the QoS policies, which enables the DFCs to support the same QoS features that the PFC supports. Since Cisco VSS is designed with a distributed forwarding architecture, the PFC and DFC functions are enabled and active on active and hot-standby virtual switch nodes and on distributed linecard modules. Figure 1-6 illustrates the internal PFC-based QoS architecture.

Figure 1-6 Cisco Catalyst 6500-E PFC QoS Architecture



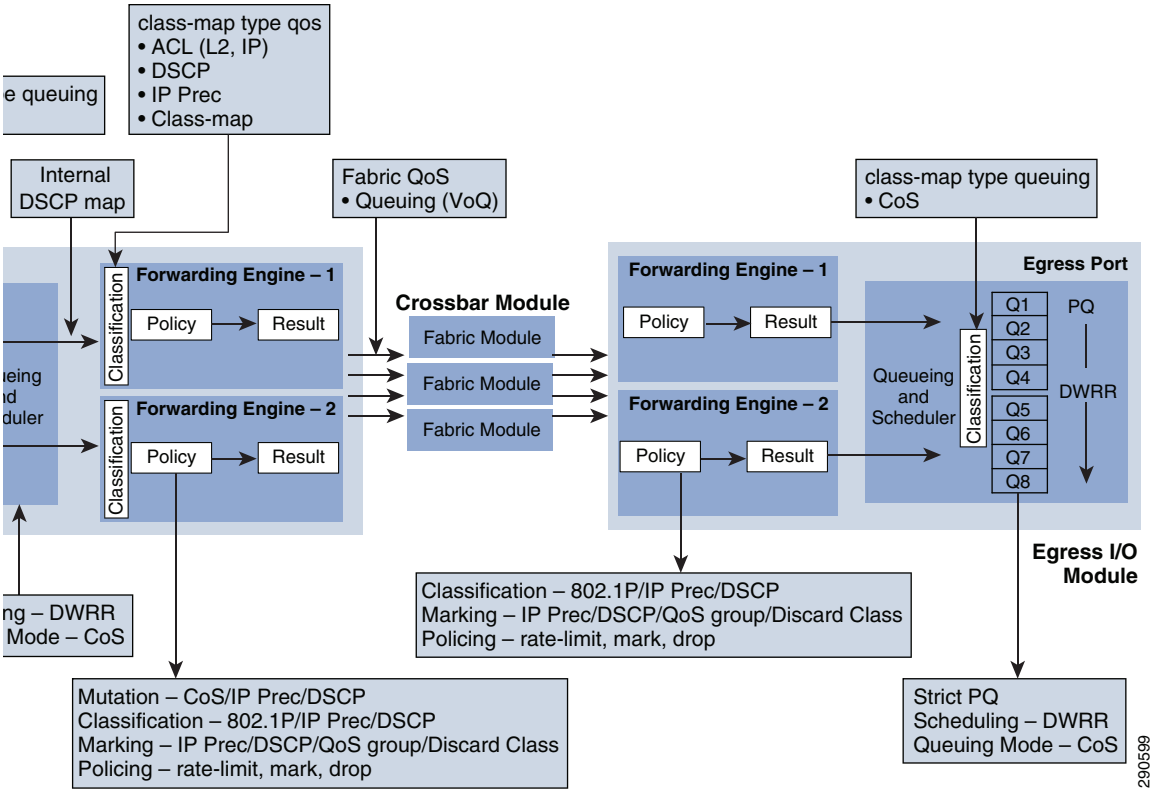
Cisco Nexus 7000 QoS

The internal distributed system architecture of the Cisco Nexus 7000 differs from the modular Cisco Catalyst platforms. For the centralized management plane, users can build the campus core-class QoS policies for ingress and egress data traffic switching through the campus backbone switch. Applying the policy-map to a physical or logical interface programs the distributed forwarding engine on each

I/O module to make the QoS policies effective in the system. The next-generation Cisco Nexus 7000 system supports MQC-based QoS policy configuration to build hierarchical classification and policy-maps to simplify the QoS operation at the core layer. By default QoS is enabled on Nexus 7000 system to perform the classification and queuing function based on ingress data traffic markings and internal mapping tables.

The Cisco Nexus 7000 system increases QoS performance with distributed and multi-stage QoS functions. To prevent congestion and prioritize real-time application traffic, such as VoIP and video, the QoS function is distributed between the port-asic, forwarding engine, and crossbar fabric path on the ingress and egress I/O modules. Each component performs a different level of inbound and outbound QoS policies to make effective switching decisions that minimize congestion for different class-of-service traffic. Each 10Gbps port of the recommended core-layer next-generation M108 I/O module is designed to perform at non-oversubscription and is equipped with a dual forwarding engine to load share the QoS function between two port groups. The M108 I/O module supports the ingress and egress Cos-to-Queue function to enable a 12-class campus QoS model for a broad range of network and data applications. Figure 1-7 illustrates the distributed Nexus 7000 QoS architecture with the recommended M108 10Gbps I/O module.

Figure 1-7 Cisco Nexus 7000 QoS Architecture



The Cisco Nexus 7000 system supports the following QoS class-map and policy configurations. Each method is designed to provide a different set of QoS functions in the system in different network environments:

- Queuing—The queuing class-map and policy-map enable the queuing and scheduling functions on the interface. The queuing class-maps are pre-defined in the Nexus 7000 system and the user cannot create a custom class-map for the ingress or egress queuing function. A user can customize the queuing policy-map and leverage the system-defined common queuing class-map to build the policy-map.
- QoS—Another set of class-maps and MQC policy-maps for the classification, marking, and policing functions. The user can customize the QoS class-map and policy-map to apply inbound and outbound policy-maps on the interfaces. The Nexus 7000 system provides the flexibility to apply two different QoS and queuing policy-maps on physical and logical interfaces.

- Network QoS—Network QoS defines common CoS characteristics across the Data Center Bridging (DCB) network, which is not currently applicable in campus network designs.

Deploying Access Layer QoS

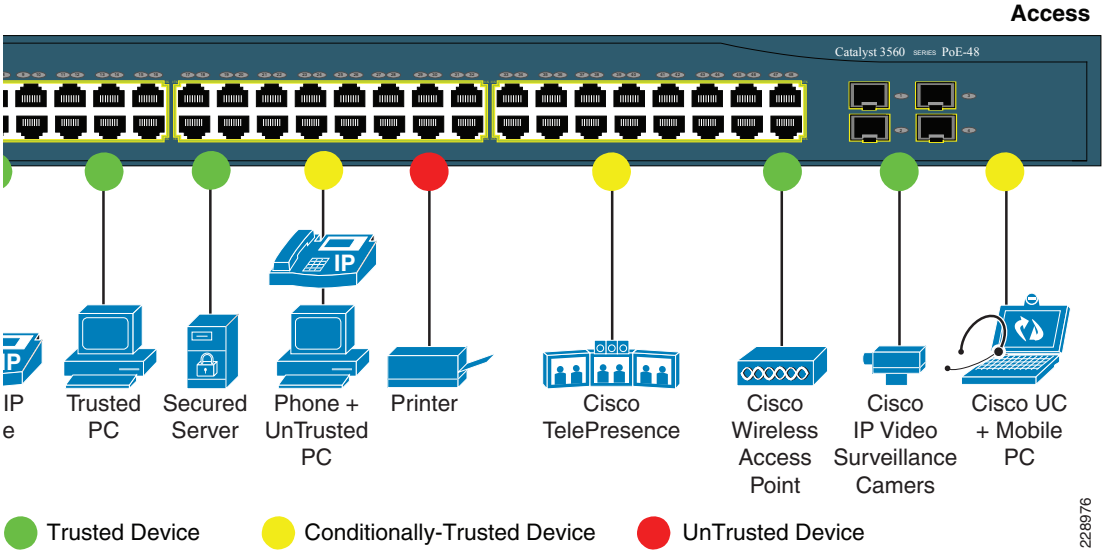
The intelligent Cisco campus access layer switches provide the entry point to the network for various types of end devices managed by the enterprise IT department or employee's personal devices (laptops, etc.). The secured access switch must decide whether to accept the QoS markings from each endpoint or whether to modify them. This is determined by the QoS policies and the trust model with which the endpoint is deployed.

QoS Trust Boundary

Network QoS policies need to be designed and implemented considering the entire borderless network. This includes defining trust points and determining which policies to enforce at each device within the network. Developing the trust model guides policy implementations for each device.

The devices (routers, switches, WLC) within the internal network boundary are managed by the system administrator and hence are classified as trusted devices. Access layer switches communicate with devices that are beyond the network boundary and within the internal network domain. The QoS trust boundary at the access layer communicates with various devices that could be deployed in different trust models (trusted, conditional-trusted, or untrusted). Figure 1-8 illustrates several types of devices in the network edge.

Figure 1-8 Borderless Campus QoS Trust Boundary



The enterprise network administrator must identify and classify each of these device types into one of three different trust models, each with its own unique security and QoS policies to access the network:

- *Untrusted*—An unmanaged device that does not pass through the network security policies, for example, an employee-owned PC or network printer. Packets with 802.1p or DSCP marking set by untrusted endpoints are reset to default by the access layer switch at the edge. Otherwise, it is possible for an unsecured user to consume network bandwidth that may impact network availability and security for other users.
- *Trusted*—Devices that passes through network access security policies and are managed by the network administrator. Even when these devices are maintained and secured by the network administrator, QoS policies must still be enforced to classify traffic and assign it to the appropriate queue to provide bandwidth assurance and proper treatment during network congestion.
- *Conditionally-trusted*—A single physical connection with one trusted endpoint and an indirect untrusted endpoint must be deployed as conditionally-trusted model. The trusted endpoints are still managed by the network administrator, but it is possible that the untrusted user behind the endpoint may or may not be secure (for example, a Cisco Unified IP phone and a PC). These deployment scenarios require a hybrid QoS policy that intelligently distinguishes and applies different QoS policies to the trusted and untrusted endpoints that are connected to the same port.

The ingress QoS policy at the access switches needs to be established, since this is the trust boundary where traffic enters the network. The following ingress QoS techniques are applied to provide appropriate service treatment and prevent network congestion:

- *Trust*—After classifying the endpoint the trust settings must be explicitly set by a network administrator. By default, Catalyst switches set each port in untrusted mode when QoS is enabled.
- *Classification*—An IETF standard has defined a set of application classes and provides recommended DSCP settings. This classification determines the priority the traffic will receive in the network. Using the IETF standard simplifies the classification process and improves application and network performance.
- *Policing*—To prevent network congestion, the access layer switch limits the amount of inbound traffic up to its maximum setting. Additional policing can be applied for known applications to ensure the bandwidth of an egress queue is not completely consumed by one application.
- *Marking*—Based on trust model, classification, and policer settings, the QoS marking is set at the edge before approved traffic enters through the access layer switching fabric. Marking traffic with the appropriate DSCP value is important to ensure traffic is mapped to the appropriate internal queue and treated with the appropriate priority.
- *Queuing*—To provide differentiated services internally in the Catalyst 29xx and 3xxx switching fabric, all approved traffic is queued into a priority or non-priority ingress queue. The ingress queuing architecture ensures real-time applications, like VoIP traffic, are given the appropriate priority (e.g., transmitted before data traffic).

Enabling QoS

By default, QoS is disabled on all Catalyst 3xxx Series switches and must be explicitly enabled in global configuration mode. The QoS configuration is the same for a multilayer or routed access deployment. The following sample QoS configuration must be enabled on all access layer switches deployed in the campus LAN network.

Access Layer 3xxx (Multilayer or Routed Access)

```
cr24-3560X-LB(config)#mls qos
cr24-3560X-LB#show mls qos
QoS is enabled
QoS ip packet dscp rewrite is enabled
```

Note The QoS function on the Catalyst 4500E with Sup7-E, Sup6-E, and Sup6L-E is enabled with the policy-map attached to the port and does not require any additional global configuration.

Upon enabling QoS in the Catalyst switches, all physical ports are assigned untrusted mode. The network administrator must explicitly enable the trust settings on the physical port where trusted or conditionally trusted endpoints are connected. The Catalyst switches can trust the ingress packets based on 802.1P (CoS-based), ToS (ip-prec-based), or DSCP (DSCP-based) values. Best practice is to

deploy DSCP-based trust mode on all the trusted and conditionally-trusted endpoints. This offers a higher level of classification and marking granularity than other methods. The following sample DSCP-based trust configuration must be enabled on the access switch ports connecting to trusted or conditionally-trusted endpoints.

QoS Trust Mode (Multilayer or Routed-Access)

Trusted Port

- Catalyst 3xxx (Multilayer or Routed Access)

```
cr22-3560X-LB(config)#interface GigabitEthernet0/5
cr22-3560X-LB(config-if)# description CONNECTED TO IPVS 2500 - CAMERA
cr22-3560X-LB(config-if)# mls qos trust dscp
cr22-3560X-LB#show mls qos interface Gi0/5
GigabitEthernet0/5
trust state: trust dscp
trust mode: trust dscp
trust enabled flag: ena
COS override: dis
default COS: 0
DSCP Mutation Map: Default DSCP Mutation Map
Trust device: none
qos mode: port-based
```

- 4500E (Multilayer or Routed Access)

  By default all the Sup7-E, Sup6-E, and Sup6L-E ports are in trusted mode; such a configuration leverages internal the DSCP mapping table to automatically classify QoS bit settings from incoming traffic and place it in the appropriate queue based on the mapping table. To set the appropriate network policy, the default settings must be modified by implementing ingress QoS policy-map. Refer to the "Implementing Ingress QoS Policing" section on page 27 for further details.

Conditionally-Trusted Port

At the campus access layer the network edge port can be explicitly implemented to conditionally trust the port QoS setting based on end point, e.g., Cisco IP phone. When Trust Boundary is enabled as shown below, the edge port automatically becomes "untrusted" and the access layer switch marks the 802.1P CoS and DSCP values to 0 until the IP phone is detected on that port. QoS policies are applied according to these modified values.

```
cr22-3560-LB(config)#interface Gi0/4
cr22-3560-LB(config-if)# description CONNECTED TO PHONE+PC
cr22-3560-LB(config-if)# mls qos trust device cisco-phone
cr22-3560-LB(config-if)# mls qos trust dscp
```

```
cr22-3560-LB#show mls qos interface Gi0/4
GigabitEthernet0/4
trust state: not trusted
trust mode: trust dscp
trust enabled flag: dis
COS override: dis
default COS: 0
DSCP Mutation Map: Default DSCP Mutation Map
Trust device: cisco-phone
qos mode: port-based
```

- 4500E (Multilayer or Routed Access)

```
cr22-4507-LB(config)#interface GigabitEthernet3/3
cr22-4507-LB(config-if)# qos trust device cisco-phone

cr22-4507-LB#show qos interface Gig3/3
Operational Port Trust State: Trusted
Trust device: cisco-phone
Default DSCP: 0 Default CoS: 0
Appliance trust: none
```

UnTrusted Port

As described earlier, the default trust mode is untrusted when globally enabling the QoS function. Without explicit trust configuration on the Gi0/1 port, the following **show** command verifies current trust state and mode:

- Catalyst 3xxx (Multilayer or Routed Access)

```
cr22-3560-LB#show mls qos interface Gi0/1
GigabitEthernet0/1
trust state: not trusted
trust mode: not trusted
trust enabled flag: ena
COS override: dis
default COS: 0
DSCP Mutation Map: Default DSCP Mutation Map
Trust device: none
qos mode: port-based
```

- 4500E (Multilayer or Routed Access)

  The QoS trust function on the Cisco Catalyst 4500E with Sup7-E, Sup6-E, and Sup6L-E is enabled by default and must be modified with the policy-map attached to the port.

```
cr22-4507-LB#show qos interface GigabitEthernet3/1
Operational Port Trust State: Trusted
Trust device: none
Default DSCP: 0 Default CoS: 0
```

```
Appliance trust: none
```

Implementing Ingress QoS Classification

When creating QoS classification policies, the network administrator needs to consider what applications are present at the access edge (in the ingress direction) and whether these applications are sourced from trusted or untrusted endpoints. If PC endpoints are secured and centrally administered, then endpoint PCs may be considered trusted endpoints. In most deployments this is not the case, thus PCs are considered untrusted endpoints for the remainder of this document.

Not every application class, as defined in the Cisco-modified RFC 4594-based model, is present in the ingress direction at the access edge; therefore, it is not necessary to provision the following application classes at the access layer:

- *Network Control*—It is assumed that access layer switch will not transmit or receive network control traffic from endpoints; hence this class is not implemented.
- *Broadcast Video*—Broadcast video and a multimedia streaming server can be distributed across the campus network which may be broadcasting live video feed using multicast streams. These live video feed must be originated from the trusted distributed data center servers.
- Operation, Administration and Management—Primarily generated by network devices (routers and switches) and collected by management stations which are typically deployed in the trusted data center network or a network control center.

All applications present at the access edge need to be assigned a classification, as shown in Figure 9. Voice traffic is primarily sourced from Cisco IP telephony devices residing in the voice VLAN (VVLAN). These are trusted devices or conditionally trusted (if users also attach PCs, etc.) to the same port. Voice communication may also be sourced from PCs with soft-phone applications, like Cisco Unified Personal Communicator (CUPC). Since such applications share the same UDP port range as multimedia conferencing traffic (UDP/RTP ports 16384-32767), this soft-phone VoIP traffic is indistinguishable and should be classified with multimedia conferencing streams. See Figure 9.

Figure 9 Ingress QoS Application Model

Application	PHB	Application Examples	Present at Campus Access-Edge (Ingress)?	Trust Boundary
Network Control	CS6	EIGRP, OSPF, HSRP, IKE		
VoIP	EF	Cisco IP Phone	Yes	Trusted
Broadcast Video		Cisco IPVS, Enterprise TV		
Realtime Interactive	CS4	Cisco TelePresence	Yes	Trusted
Multimedia Conferencing	AF4	Cisco CUPC. WebEx	Yes	Untrusted
Multimedia Streaming	AF3	Cisco DMS, IP/TV		
Signaling	CS3	SCCP, SIP, H.323	Yes	Trusted
Transactional Data	AF2	ERP Apps, CRM Apps	Yes	Untrusted
OAM	CS2	SNMP, SSH, Syslog		
Bulk Data	AF1	Email, FTP, Backups	Yes	Untrusted
Best Effort	DF	Default Class	Yes	Untrusted
Scavenger	CS1	YouTube, Gaming, P2P	Yes	Untrusted

228977

Modular QoS MQC offers scalability and flexibility in configuring QoS to classify all 8-application classes by using match statements or an extended access list to match the exact value or range of Layer 4 known ports that each application uses to communicate on the network. The following sample configuration creates an extended access list for each application and then applies it under class-map configuration mode.

- Catalyst 3xxx and 4500E (MultiLayer and Routed Access)

```
cr22-4507-LB(config)#ip access-list extended MULTIMEDIA-CONFERENCING
cr22-4507-LB(config-ext-nacl)# remark RTP
cr22-4507-LB(config-ext-nacl)# permit udp any any range 16384 32767

cr22-4507-LB(config-ext-nacl)#ip access-list extended SIGNALING
cr22-4507-LB(config-ext-nacl)# remark SCCP
cr22-4507-LB(config-ext-nacl)# permit tcp any any range 2000 2002
cr22-4507-LB(config-ext-nacl)# remark SIP
cr22-4507-LB(config-ext-nacl)# permit tcp any any range 5060 5061
cr22-4507-LB(config-ext-nacl)# permit udp any any range 5060 5061

cr22-4507-LB(config-ext-nacl)#ip access-list extended TRANSACTIONAL-DATA
```

```
cr22-4507-LB(config-ext-nacl)# remark HTTPS
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 443
cr22-4507-LB(config-ext-nacl)# remark ORACLE-SQL*NET
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 1521
cr22-4507-LB(config-ext-nacl)# permit udp any any eq 1521
cr22-4507-LB(config-ext-nacl)# remark ORACLE
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 1526
cr22-4507-LB(config-ext-nacl)# permit udp any any eq 1526
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 1575
cr22-4507-LB(config-ext-nacl)# permit udp any any eq 1575
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 1630

cr22-4507-LB(config-ext-nacl)#ip access-list extended BULK-DATA
cr22-4507-LB(config-ext-nacl)# remark FTP
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq ftp
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq ftp-data
cr22-4507-LB(config-ext-nacl)# remark SSH/SFTP
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 22
cr22-4507-LB(config-ext-nacl)# remark SMTP/SECURE SMTP
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq smtp
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 465
cr22-4507-LB(config-ext-nacl)# remark IMAP/SECURE IMAP
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 143
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 993
cr22-4507-LB(config-ext-nacl)# remark POP3/SECURE POP3
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq pop3
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 995
cr22-4507-LB(config-ext-nacl)# remark CONNECTED PC BACKUP
cr22-4507-LB(config-ext-nacl)# permit tcp any eq 1914 any

cr22-4507-LB(config-ext-nacl)#ip access-list extended DEFAULT
cr22-4507-LB(config-ext-nacl)# remark EXPLICIT CLASS-DEFAULT
cr22-4507-LB(config-ext-nacl)# permit ip any any

cr22-4507-LB(config-ext-nacl)#ip access-list extended SCAVENGER
cr22-4507-LB(config-ext-nacl)# remark KAZAA
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 1214
cr22-4507-LB(config-ext-nacl)# permit udp any any eq 1214
cr22-4507-LB(config-ext-nacl)# remark MICROSOFT DIRECT X GAMING
cr22-4507-LB(config-ext-nacl)# permit tcp any any range 2300 2400
cr22-4507-LB(config-ext-nacl)# permit udp any any range 2300 2400
cr22-4507-LB(config-ext-nacl)# remark APPLE ITUNES MUSIC SHARING
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 3689
cr22-4507-LB(config-ext-nacl)# permit udp any any eq 3689
cr22-4507-LB(config-ext-nacl)# remark BITTORRENT
cr22-4507-LB(config-ext-nacl)# permit tcp any any range 6881 6999
cr22-4507-LB(config-ext-nacl)# remark YAHOO GAMES
cr22-4507-LB(config-ext-nacl)# permit tcp any any eq 11999
cr22-4507-LB(config-ext-nacl)# remark MSN GAMING ZONE
```

```
cr22-4507-LB(config-ext-nacl)# permit tcp any any range 28800 29100
```

Creating class-map for each application services and applying match statement:

```
cr22-4507-LB(config)#class-map match-all VVLAN-SIGNALING
cr22-4507-LB(config-cmap)# match ip dscp cs3

cr22-4507-LB(config-cmap)#class-map match-all VVLAN-VOIP
cr22-4507-LB(config-cmap)# match ip dscp ef

cr22-4507-LB(config-cmap)#class-map match-all MULTIMEDIA-CONFERENCING
cr22-4507-LB(config-cmap)# match access-group name MULTIMEDIA-CONFERENCING

cr22-4507-LB(config-cmap)#class-map match-all SIGNALING
cr22-4507-LB(config-cmap)# match access-group name SIGNALING

cr22-4507-LB(config-cmap)#class-map match-all TRANSACTIONAL-DATA
cr22-4507-LB(config-cmap)# match access-group name TRANSACTIONAL-DATA

cr22-4507-LB(config-cmap)#class-map match-all BULK-DATA
cr22-4507-LB(config-cmap)# match access-group name BULK-DATA

cr22-4507-LB(config-cmap)#class-map match-all DEFAULT
cr22-4507-LB(config-cmap)# match access-group name DEFAULT

cr22-4507-LB(config-cmap)#class-map match-all SCAVENGER
cr22-4507-LB(config-cmap)# match access-group name SCAVENGER
```

Implementing Ingress QoS Policing

It is important to limit how much bandwidth each class may use at the ingress to the access layer for two primary reasons:

- *Bandwidth bottleneck*—To prevent network congestion, each physical port at the trust boundary must be rate-limited. The rate-limit value may differ based on several factors—end-to-end network bandwidth capacity, end-station, and application performance capacities, etc.
- *Bandwidth security*—Well-known applications like Cisco IP telephony use a fixed amount of bandwidth per device based on a codec. It is important to police high-priority application traffic which is assigned to the high-priority queue, otherwise it could consume too much overall network bandwidth and impact other application performance.

In addition to policing, the rate-limit function also provides the ability to take different actions on the excess incoming traffic which exceeds the established limits. The exceed-action for each class must be carefully designed based on the nature of the application to provide best-effort services based on network bandwidth availability. Table 1 provides best practice policing guidelines for different classes to be implemented for trusted and conditional-trusted endpoints at the network edge.

Table 1 Access Layer Ingress Policing Guidelines

Application	Policing Rate	Conform-Action	Exceed-Action
VoIP Signaling	<32 kbps	Pass	Drop
VoIP Bearer	<128 kbps	Pass	Drop
Multimedia Conferencing	<5Mbps [1]	Pass	Drop
Signaling	<32 kbps	Pass	Drop
Transactional Data	<10 Mbps [1]	Pass	Remark to CS1
Bulk Data	<10 Mbps [1]	Pass	Remark to CS1
Best Effort	<10 Mbps [1]	Pass	Remark to CS1
Scavenger	<10 Mbps [1]	Pass	Drop

1. The rate varies based on several factors as defined earlier. This table depicts a sample rate-limiting value.

Catalyst 3xxx (Multilayer and Routed-Access)

- Trusted or Conditionally-Trusted Port Policer

```
cr24-3750-LB(config)#policy-map Phone+PC-Policy
cr24-3750-LB(config-pmap)# class VVLAN-VOIP
cr24-3750-LB(config-pmap-c)#  police 128000 8000 exceed-action drop
cr24-3750-LB(config-pmap-c)# class VVLAN-SIGNALING
cr24-3750-LB(config-pmap-c)#  police 32000 8000 exceed-action drop
cr24-3750-LB(config-pmap-c)# class MULTIMEDIA-CONFERENCING
cr24-3750-LB(config-pmap-c)#  police 5000000 8000 exceed-action drop
cr24-3750-LB(config-pmap-c)# class SIGNALING
cr24-3750-LB(config-pmap-c)#  police 32000 8000 exceed-action drop
cr24-3750-LB(config-pmap-c)# class TRANSACTIONAL-DATA
cr24-3750-LB(config-pmap-c)#  police 10000000 8000 exceed-action policed-dscp-transmit
cr24-3750-LB(config-pmap-c)# class BULK-DATA
cr24-3750-LB(config-pmap-c)#  police 10000000 8000 exceed-action policed-dscp-transmit
cr24-3750-LB(config-pmap-c)# class SCAVENGER
cr24-3750-LB(config-pmap-c)#  police 10000000 8000 exceed-action drop
cr24-3750-LB(config-pmap-c)# class DEFAULT
cr24-3750-LB(config-pmap-c)#  police 10000000 8000 exceed-action policed-dscp-transmit
```

Catalyst 4500E (Multilayer and Routed-Access)

```
cr22-4507-LB(config)#policy-map Phone+PC-Policy
cr22-4507-LB(config-pmap)# class VVLAN-VOIP
cr22-4507-LB(config-pmap-c)#  police 128k bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action drop
```

```
cr22-4507-LB(config-pmap-c-police)# class VVLAN-SIGNALING
cr22-4507-LB(config-pmap-c)#  police 32k bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action drop
cr22-4507-LB(config-pmap-c-police)# class MULTIMEDIA-CONFERENCING
cr22-4507-LB(config-pmap-c)#  police 5m bc 8000
cr22-4507-LB(config-pmap-c-police)#  conform-action transmit
cr22-4507-LB(config-pmap-c-police)#  exceed-action drop
cr22-4507-LB(config-pmap-c-police)# class SIGNALING
cr22-4507-LB(config-pmap-c)# police 32k bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action drop
cr22-4507-LB(config-pmap-c-police)# class TRANSACTIONAL-DATA
cr22-4507-LB(config-pmap-c)#  police 10m bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action set-dscp-transmit cs1
cr22-4507-LB(config-pmap-c-police)# class BULK-DATA
cr22-4507-LB(config-pmap-c)#  police 10m bc 8000
cr22-4507-LB(config-pmap-c-police)#  conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action set-dscp-transmit cs1
cr22-4507-LB(config-pmap-c-police)# class SCAVENGER
cr22-4507-LB(config-pmap-c)#  police 10m bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#  exceed-action drop
cr22-4507-LB(config-pmap-c-police)# class class-default
cr22-4507-LB(config-pmap-c)#  police 10m bc 8000
cr22-4507-LB(config-pmap-c-police)#   conform-action transmit
cr22-4507-LB(config-pmap-c-police)#   exceed-action set-dscp-transmit cs1
```

- UnTrusted Port Policer

All ingress traffic (default class) from untrusted endpoints must be policed without explicit classification that requires differentiated services. The following sample configuration shows how to deploy policing on untrusted ingress ports in access layer switches:

```
cr22-3560X-LB(config)#policy-map UnTrusted-PC-Policy
cr22-3560X-LB(config-pmap)# class class-default
cr22-3560X-LB(config-pmap-c)#  police 10000000 8000 exceed-action drop
```

Implementing Ingress Marking

Accurate DSCP marking of ingress traffic at the access layer switch is critical to ensure proper QoS service treatment as traffic traverses through the network. All classified and policed traffic must be explicitly marked using the policy-map configuration based on an 8-class QoS model as shown in Figure 14.

Best practice is to use a explicit marking command (**set dscp**) even for trusted application classes (like VVLAN-VOIP and VVLAN-SIGNALING), rather than a trust policy-map action. A trust statement in a policy map requires multiple hardware entries with the use of an explicit (seemingly redundant) marking command and improves hardware efficiency.

The following sample configuration shows how to implement explicit marking for multiple classes on trusted and conditionally-trusted ingress ports in access layer switches:

Trusted or Conditionally-Trusted Port

- Catalyst 3xxx and 4500E (Multilayer and Routed-Access)

```
cr22-3750-LB(config)#policy-map Phone+PC-Policy
cr22-3750-LB(config-pmap)# class VVLAN-VOIP
cr22-3750-LB(config-pmap-c)#  set dscp ef
cr22-3750-LB(config-pmap-c)# class VVLAN-SIGNALING
cr22-3750-LB(config-pmap-c)#  set dscp cs3
cr22-3750-LB(config-pmap-c)# class MULTIMEDIA-CONFERENCING
cr22-3750-LB(config-pmap-c)#  set dscp af41
cr22-3750-LB(config-pmap-c)# class SIGNALING
cr22-3750-LB(config-pmap-c)#  set dscp cs3
cr22-3750-LB(config-pmap-c)# class TRANSACTIONAL-DATA
cr22-3750-LB(config-pmap-c)#  set dscp af21
cr22-3750-LB(config-pmap-c)# class BULK-DATA
cr22-3750-LB(config-pmap-c)#  set dscp af11
cr22-3750-LB(config-pmap-c)# class SCAVENGER
cr22-3750-LB(config-pmap-c)#  set dscp cs1
cr22-3750-LB(config-pmap-c)# class DEFAULT
cr22-3750-LB(config-pmap-c)#  set dscp default
```

All ingress traffic (default class) from an untrusted endpoint must be marked without a explicit classification. The following sample configuration shows how to implement explicit DSCP marking:

Untrusted Port

- Catalyst 3xxx and 4500E (Multilayer and Routed-Access)

```
cr22-3750-LB(config)#policy-map UnTrusted-PC-Policy
cr22-3750-LB(config-pmap)# class class-default
cr22-3750-LB(config-pmap-c)#  set dscp default
```

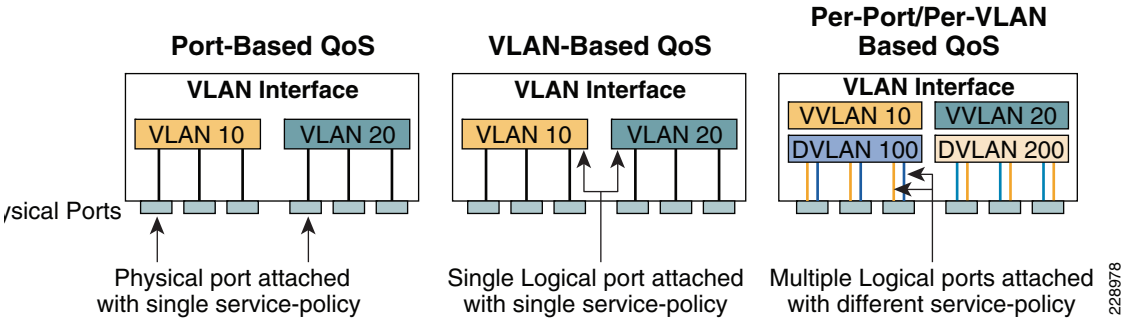
Applying Ingress Policies

After creating a policy-map on all the Layer 2 and Layer 3 access switches with QoS policies defined, the service-policy must be applied on the edge interface of the access layer to enforce the QoS configuration. Cisco Catalyst switches offer three simplified methods to apply service-policies; depending on the deployment model any of the methods can be implemented:

- *Port-Based QoS*—Applying the service-policy on a per-physical-port basis forces traffic to pass through the QoS policies before entering into the campus network. Port-Based QoS discretely functions on a per-physical port basis even if it is associated with a logical VLAN which is applied on multiple physical ports.
- *VLAN-Based QoS*—Applying the service-policy on a per-VLAN basis requires the policy-map to be attached to a VLAN. Every physical port associated to a VLAN that requires bandwidth guarantee or traffic shaping needs extra configuration at the interface level.
- *Per-Port / Per-VLAN-Based QoS*—This is not supported on all Catalyst platforms and the configuration commands are platform-specific. Per-port/per-VLAN-based QoS allows policy-map to operate on a trunk interface. A different policy-map can be applied to specific VLANs within a trunk interface.

See Figure 10.

Figure 10 QoS Policies Implementation Methods



The following sample configuration provides guidelines to deploy port-based QoS on the access layer switches in the campus network:

- Catalyst 3xxx and 4500E (Multilayer and Routed-Access)

```
cr22-3560X-LB(config)#interface GigabitEthernet0/1
cr22-3560X-LB(config-if)# service-policy input UnTrusted-PC-Policy

cr22-3560X-LB#show mls qos interface GigabitEthernet0/1
GigabitEthernet0/1
Attached policy-map for Ingress: UnTrusted-PC-Policy
trust state: not trusted
trust mode: not trusted
trust enabled flag: ena
COS override: dis
default COS: 0
DSCP Mutation Map: Default DSCP Mutation Map
Trust device: none
qos mode: port-based
```

Applying Ingress Queuing

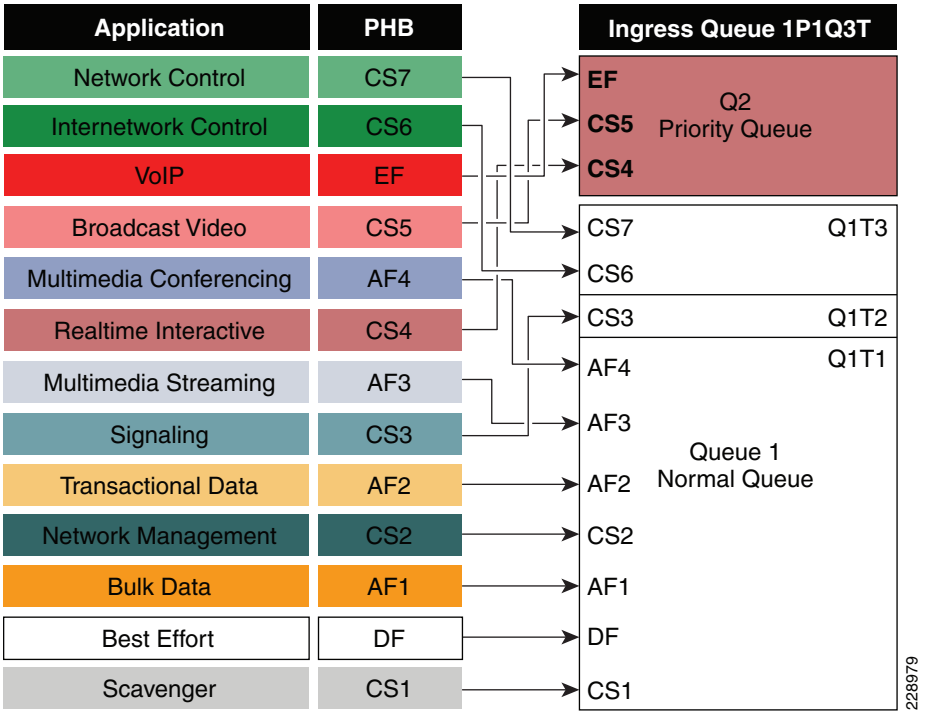
Fixed configuration Cisco Catalyst switches not only offer differentiated services on the network ports, but also internally on the switching fabric. After enabling QoS and attaching inbound policies on the physical ports, all the packets that meet the specified policy are forwarded to the switching fabric for egress switching. The aggregate bandwidth from all edge ports may exceed the switching fabric bandwidth and cause internal congestion.

Cisco Catalyst 3xxx platforms support two internal ingress queues—normal queue and priority queue. The ingress queue inspects the DSCP value on each incoming frame and assigns it to either the normal or priority queue. High priority traffic, like DSCP EF marked packets, are placed in the priority queue and switched before processing the normal queue.

The Catalyst 3750-X family of switches supports the weighted tail drop (WTD) congestion avoidance mechanism. WTD is implemented on queues to manage the queue length. WTD drops packets from the queue based on DSCP value and the associated threshold. If the threshold is exceeded for a given internal DSCP value, the switch drops the packet. Each queue has three threshold values. The internal DSCP determines which of the three threshold values is applied to the frame. Two of the three thresholds are configurable (explicit) and one is not (implicit). This last threshold corresponds to the tail of the queue (100 percent limit).

Figure 11 depicts how different class-of-service applications are mapped to the ingress queue structure (1P1Q3T) and how each queue is assigned a different WTD threshold.

Figure 11 Catalyst 3xxx DSCP-based Ingress QoS Model



- **Catalyst 3xxx (Multilayer and Routed-Access)**

```
cr22-3750-LB(config)#mls qos srr-queue input priority-queue 2 bandwidth 30
 ! Q2 is enabled as a strict-priority ingress queue with 30% BW

cr22-3750-LB (config)#mls qos srr-queue input bandwidth 70 30
 ! Q1 is assigned 70% BW via SRR shared weights
 ! Q1 SRR shared weight is ignored (as it has been configured as a PQ)

cr22-3750-LB (config)#mls qos srr-queue input threshold 1 80 90
 ! Q1 thresholds are configured at 80% (Q1T1) and 90% (Q1T2)
 ! Q1T3 is implicitly set at 100% (the tail of the queue)
 ! Q2 thresholds are all set (by default) to 100% (the tail of Q2)

! This section configures ingress DSCP-to-Queue Mappings
cr22-3750-LB (config)# mls qos srr-queue input dscp-map queue 1 threshold 1 0 8 10 12
14
 ! DSCP DF, CS1 and AF1 are mapped to ingress Q1T1
cr22-3750-LB (config)# mls qos srr-queue input dscp-map queue 1 threshold 1 16 18 20
22
```

```
 ! DSCP CS2 and AF2 are mapped to ingress Q1T1
cr22-3750-LB (config)# mls qos srr-queue input dscp-map queue 1 threshold 1 26 28 30
34 36 38
 ! DSCP AF3 and AF4 are mapped to ingress Q1T1
cr22-3750-LB (config)#mls qos srr-queue input dscp-map queue 1 threshold 2 24
 ! DSCP CS3 is mapped to ingress Q1T2

cr22-3750-LB(config)#mls qos srr-queue input dscp-map queue 1 threshold 3 48 56
 ! DSCP CS6 and CS7 are mapped to ingress Q1T3 (the tail of Q1)
cr22-3750-LB(config)#mls qos srr-queue input dscp-map queue 2 threshold 3 32 40 46
 ! DSCP CS4, CS5 and EF are mapped to ingress Q2T3 (the tail of the PQ)

cr22-3750-LB#show mls qos input-queue
Queue:       12
----------------------------------------
buffers   :9010
bandwidth :7030
priority  :030
threshold1:80100
threshold2:90100

cr22-3750-LB#show mls qos maps dscp-input-q
   Dscp-inputq-threshold map:
     d1 :d2    0        1        2        3         4         5        6         7
8         9

-----------------------------------------------------------------------------------
      0 :    01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01
      1 :    01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01 01-01
      2 :    01-01 01-01 01-01 01-01 01-02 01-01 01-01 01-01 01-01 01-01
      3 :    01-01 01-01 02-03 01-01 01-01 01-01 01-01 01-01 01-01 01-01
      4 :    02-03 02-01 02-01 02-01 02-01 02-01 02-03 02-01 01-03 01-01
      5 :    01-01 01-01 01-01 01-01 01-01 01-01 01-03 01-01 01-01 01-01
      6 :    01-01 01-01 01-01 01-01
```

Note The ingress queuing function on the Catalyst 4500E Sup7-E, Sup6-E, and Sup6L-E is not supported as described in Figure 5.

Implementing Access Layer Egress QoS

The QoS implementation of egress traffic towards network edge devices on access layer switches is much simplified compared to ingress traffic which requires stringent QoS policies to provide differentiated services and network bandwidth protection. Unlike the ingress QoS model, the egress

QoS model must provide optimal queuing policies for each class and set the drop thresholds to prevent network congestion and degraded application performance. With egress queuing in DSCP mode, the Cisco Catalyst switching platforms are bounded by a limited number of hardware queues.

Catalyst 3xxx Egress QoS

The Cisco Catalyst 3xxx Series platform supports four egress queues that are required to support the variable class QoS policies for the enterprise campus network; specifically, the following queues would be considered a minimum:

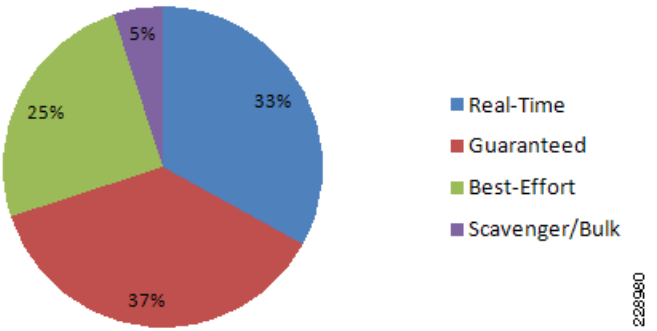
- Realtime queue (to support a RFC 3246 EF PHB service)
- Guaranteed bandwidth queue (to support RFC 2597 AF PHB services)
- Default queue (to support a RFC 2474 DF service)
- Bandwidth constrained queue (to support a RFC 3662 scavenger service)

As a best practice, each physical or logical interface must be deployed with IETF recommended bandwidth allocations for different class-of-service applications:

- The real-time queue should not exceed 33 percent of the link's bandwidth.
- The default queue should be at least 25 percent of the link's bandwidth.
- The bulk/scavenger queue should not exceed 5 percent of the link's bandwidth.

Figure 12 illustrates the egress bandwidth allocation best practices design for different classes.

Figure 12 Class-of-Service Egress Bandwidth Allocations



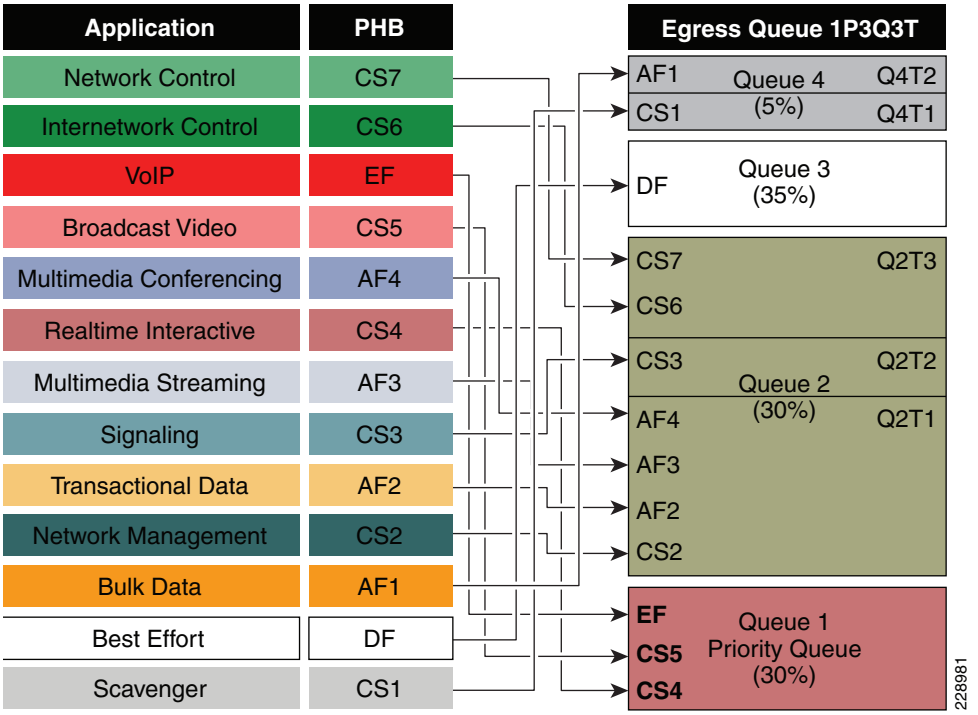
Given these minimum queuing requirements and bandwidth allocation recommendations, the following application classes can be mapped to the respective queues:

- *Realtime Queue*—Voice, broadcast video, and realtime interactive may be mapped to the realtime queue (per RFC 4594).

- *Guaranteed Queue*—Network/internetwork control, signaling, network management, multimedia conferencing, multimedia streaming, and transactional data can be mapped to the guaranteed bandwidth queue. Congestion avoidance mechanisms (i.e., selective dropping tools), such as WRED, can be enabled on this class; furthermore, if configurable drop thresholds are supported on the platform, these may be enabled to provide intra-queue QoS to these application classes in the respective order they are listed (such that control plane protocols receive the highest level of QoS within a given queue).
- *Scavenger/Bulk Queue*—Bulk data and scavenger traffic can be mapped to the bandwidth-constrained queue and congestion avoidance mechanisms can be enabled on this class. If configurable drop thresholds are supported on the platform, these may be enabled to provide inter-queue QoS to drop scavenger traffic ahead of bulk data.
- *Default Queue*—Best-effort traffic can be mapped to the default queue; congestion avoidance mechanisms can be enabled on this class.

Like the ingress queuing structure that maps various applications based on DSCP value into two ingress queues, the egress queuing must be similarly designed to map with four egress queues. The DSCP-to-queue mapping for egress queuing must be mapped to each egress queue as stated above, which allows better queuing-policy granularity. A campus egress QoS model example for a platform that supports DSCP-to-queue mapping with a 1P3Q3T queuing structure is depicted in Figure 13.

Figure 13 Catalyst 3xxx DSCP-based 1P3Q3T Egress QoS Model



DSCP marked packets are assigned to the appropriate queue and each queue is configured with appropriate WTD threshold as defined in Figure 13. Egress queuing settings are common between all the trust-independent network edge ports as well as on the Layer 2 or Layer 3 uplinks connected to the internal network. The following egress queue configuration entered in global configuration mode must be enabled on every access layer switch in the network.

Catalyst 3xxx (Multilayer and Routed-Access)

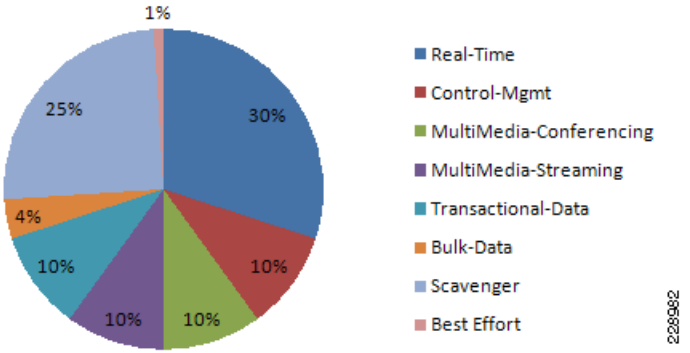
```
cr22-3750-LB(config)#mls qos queue-set output 1 buffers 15 30 35 20
! Queue buffers are allocated
cr22-3750-LB (config)#mls qos queue-set output 1 threshold 1 100 100 100 100
! All Q1 (PQ) Thresholds are set to 100%
cr22-3750-LB (config)#mls qos queue-set output 1 threshold 2 80 90 100 400
! Q2T1 is set to 80%; Q2T2 is set to 90%;
 ! Q2 Reserve Threshold is set to 100%;
 ! Q2 Maximum (Overflow) Threshold is set to 400%
cr22-3750-LB (config)#mls qos queue-set output 1 threshold 3 100 100 100 400
! Q3T1 is set to 100%, as all packets are marked the same weight in Q3
```

```
 ! Q3 Reserve Threshold is set to 100%;
 ! Q3 Maximum (Overflow) Threshold is set to 400%
cr22-3750-LB (config)#mls qos queue-set output 1 threshold 4 60 100 100 400
! Q4T1 is set to 60%; Q4T2 is set to 100%
 ! Q4 Reserve Threshold is set to 100%;
 ! Q4 Maximum (Overflow) Threshold is set to 400%

cr22-3750-LB(config)# mls qos srr-queue output dscp-map queue 1 threshold 3 32 40 46
! DSCP CS4, CS5 and EF are mapped to egress Q1T3 (tail of the PQ)
cr22-3750-LB(config)# mls qos srr-queue output dscp-map queue 2 threshold 1 16 18 20 22
! DSCP CS2 and AF2 are mapped to egress Q2T1
cr22-3750-LB(config)# mls qos srr-queue output dscp-map queue 2 threshold 1 26 28 30 34 36
38
! DSCP AF3 and AF4 are mapped to egress Q2T1
cr22-3750-LB(config)#mls qos srr-queue output dscp-map queue 2 threshold 2 24
! DSCP CS3 is mapped to egress Q2T2
cr22-3750-LB(config)#mls qos srr-queue output dscp-map queue 2 threshold 3 48 56
! DSCP CS6 and CS7 are mapped to egress Q2T3
cr22-3750-LB(config)#mls qos srr-queue output dscp-map queue 3 threshold 3 0
! DSCP DF is mapped to egress Q3T3 (tail of the best effort queue)
cr22-3750-LB(config)#mls qos srr-queue output dscp-map queue 4 threshold 1 8
! DSCP CS1 is mapped to egress Q4T1
cr22-3750-LB(config)# mls qos srr-queue output dscp-map queue 4 threshold 2 10 12 14
! DSCP AF1 is mapped to Q4T2 (tail of the less-than-best-effort queue)

 ! This section configures edge and uplink port interface with common egress queuing
parameters
cr22-3750-LB(config)#interface range GigabitEthernet1/0/1-48
cr22-3750-LB(config-if-range)# queue-set 1
! The interface(s) is assigned to queue-set 1
cr22-3750-LB(config-if-range)# srr-queue bandwidth share 1 30 35 5
! The SRR sharing weights are set to allocate 30% BW to Q2
 ! 35% BW to Q3 and 5% BW to Q4
 ! Q1 SRR sharing weight is ignored, as it will be configured as a PQ
cr22-3750-LB(config-if-range)# priority-queue out
! Q1 is enabled as a strict priority queue

cr22-3750-LB#show mls qos interface GigabitEthernet1/0/27 queueing
GigabitEthernet1/0/27
Egress Priority Queue : enabled
Shaped queue weights (absolute) :  25 0 0 0
Shared queue weights  :  1 30 35 5
The port bandwidth limit : 100  (Operational Bandwidth:100.0)
The port is mapped to qset : 1
```

Catalyst 4500E Sup7-E, Sup6-E, and Sup6L-E Egress QoS

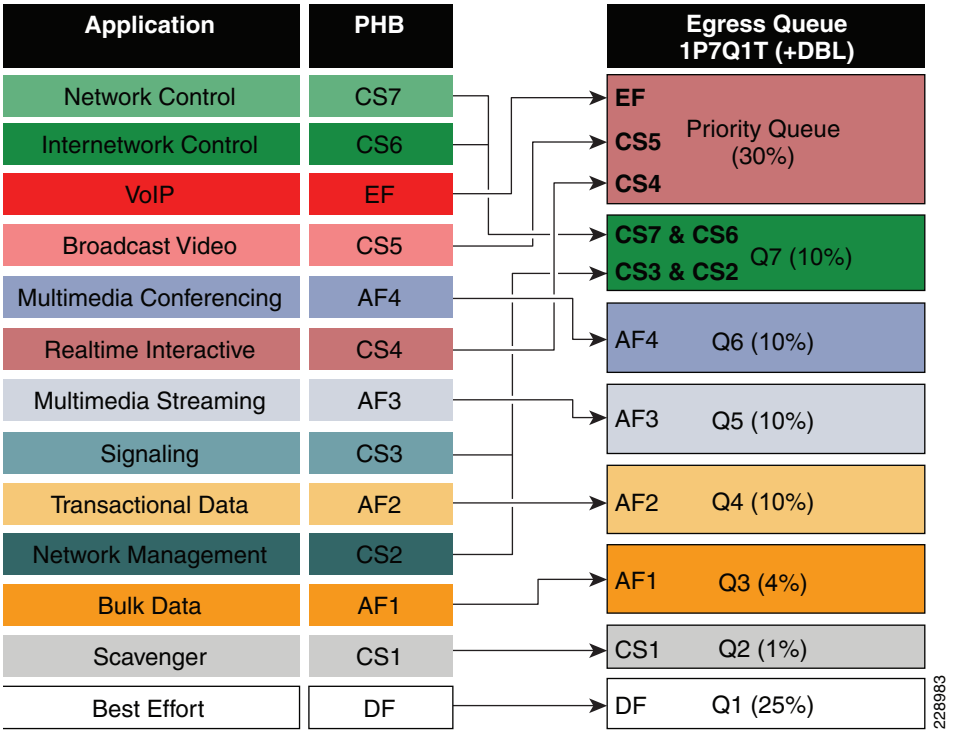
The enterprise-class 4500E switch with next-generation supervisor hardware architecture is designed to offer better egress QoS techniques, capabilities, and flexibility to provide for a well-diversified queuing structure for multiple class-of-service traffic types. Deploying the next-generation Sup7-E, Sup6-E, and Sup6L-E in the campus network provides more QoS granularity to map the 8-class traffic types to hardware-based egress-queues as illustrated in Figure 14.

Figure 14 Eight Class-of-Service Egress Bandwidth Allocations



The Cisco Catalyst 4500E Sup7-E, Sup6-E, and Sup6L-E supervisors support platform-specific congestion avoidance algorithms to provide Active Queue Management (AQM), namely Dynamic Buffer Limiting (DBL). DBL is flow-based and uses logical flow tables per port/per queue for each port. It operates by tracking the amount of buffering and credits for each flow currently active in that queue. When the queue length of a flow exceeds its limit, DBL drops packets or sets the Explicit Congestion Notification (ECN) bits in the TCP packet headers. With 8 egress (1P7Q1T) queues and DBL capability in the Sup7-E- and Sup6-E-based supervisors, the bandwidth distribution for different classes change. Figure 15 provides the new recommended bandwidth allocation.

Figure 15 Catalyst 4500E DSCP-based 1P7Q1T Egress QoS Model



The QoS architecture and implementation procedures are identical among the Sup7-E, Sup6-E, and Sup6L-E modules. Implementing QoS policies on a Sup7-E-based Catalyst 4500E platform follows the IOS (MQC)-based configuration model instead of the Catalyst OS-based QoS model. To take advantage of hardware-based QoS egress, the queuing function using MQC must be applied per member-link of the EtherChannel interface. Therefore, load-sharing egress per-flow traffic across EtherChannel links offers the advantage of optimally using distributed hardware resources.

The recommended DSCP markings for each traffic class can be classified in a different class-map for egress QoS functions. Based on Figure 15, the following configuration uses the new egress policy-map with queuing and the DBL function implemented on the Catalyst 4500E deployed with a Sup7-E, Sup6-E, and Sup6L-E supervisor module. All network edge port and core-facing uplink ports must use a common egress policy-map.

- Catalyst 4500E Sup7-E, Sup6-E, and Sup6L-E (MultiLayer and Routed-Access)

```
! Creating class-map for each classes using match dscp statement as marked by edge
systems
cr22-4507-LB(config)#class-map match-all PRIORITY-QUEUE
cr22-4507-LB(config-cmap)# match  dscp ef
```

```
cr22-4507-LB(config-cmap)# match  dscp cs5
cr22-4507-LB(config-cmap)# match  dscp cs4
cr22-4507-LB(config-cmap)#class-map match-all CONTROL-MGMT-QUEUE
cr22-4507-LB(config-cmap)# match  dscp cs7

cr24-4507-LB(config-cmap)# match  dscp cs6
cr24-4507-LB(config-cmap)# match  dscp cs3
cr24-4507-LB(config-cmap)# match  dscp cs2
cr24-4507-LB(config-cmap)#class-map match-all MULTIMEDIA-CONFERENCING-QUEUE
cr24-4507-LB(config-cmap)# match  dscp af41  af42  af43
cr24-4507-LB(config-cmap)#class-map match-all MULTIMEDIA-STREAMING-QUEUE
cr24-4507-LB(config-cmap)# match  dscp af31  af32  af33
cr24-4507-LB(config-cmap)#class-map match-all TRANSACTIONAL-DATA-QUEUE
cr24-4507-LB(config-cmap)# match  dscp af21  af22  af23
cr24-4507-LB(config-cmap)#class-map match-all BULK-DATA-QUEUE
cr24-4507-LB(config-cmap)# match  dscp af11  af12  af13
cr24-4507-LB(config-cmap)#class-map match-all SCAVENGER-QUEUE
cr24-4507-LB(config-cmap)# match  dscp cs1

! Creating policy-map and configure queueing for class-of-service
cr22-4507-LB(config)#policy-map EGRESS-POLICY
cr22-4507-LB(config-pmap)# class PRIORITY-QUEUE
cr22-4507-LB(config-pmap-c)#    priority
cr22-4507-LB(config-pmap-c)# class CONTROL-MGMT-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 10
cr22-4507-LB(config-pmap-c)# class MULTIMEDIA-CONFERENCING-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 10
cr22-4507-LB(config-pmap-c)# class MULTIMEDIA-STREAMING-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 10
cr22-4507-LB(config-pmap-c)# class TRANSACTIONAL-DATA-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 10
cr22-4507-LB(config-pmap-c)#    dbl
cr22-4507-LB(config-pmap-c)# class BULK-DATA-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 4
cr22-4507-LB(config-pmap-c)#    dbl
cr22-4507-LB(config-pmap-c)# class SCAVENGER-QUEUE
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 1
cr22-4507-LB(config-pmap-c)# class class-default
cr22-4507-LB(config-pmap-c)#    bandwidth remaining percent 25
cr22-4507-LB(config-pmap-c)#    dbl

! Attaching egress service-policy on all physical member-link ports
cr22-4507-LB(config)#int range Ten3/1 , Te4/1 , Ten5/1 , Ten5/4, Ten Gi1/1 - 6
cr22-4507-LB(config-if-range)# service-policy output EGRESS-POLICY
```

Policing Priority-Queue

EtherChannel is an aggregated logical bundle of interfaces that do not perform queuing and rely on individual member links to queue egress traffic by using hardware-based queuing. The hardware-based priority-queue implementation on the Catalyst 4500E does not support a built-in policer to restrict traffic during network congestion. To mitigate this challenge, it is recommended to implement an additional policy-map to rate-limit the priority class traffic and the policy-map must be attached on the EtherChannel to govern the aggregated egress traffic limits. The following additional policy-map must be created to classify priority-queue class traffic and rate-limit up to 30 percent egress link capacity:

```
cr22-4507-LB(config)#class-map match-any PRIORITY-QUEUE
cr22-4507-LB (config-cmap)# match  dscp ef
cr22-4507-LB (config-cmap)# match  dscp cs5
cr22-4507-LB (config-cmap)# match  dscp cs4

cr22-4507-LB (config)#policy-map PQ-POLICER
cr22-4507-LB (config-pmap)# class PRIORITY-QUEUE
cr22-4507-LB (config-pmap-c)# police cir 300 m conform-action transmit exceed-action drop

cr22-4507-LB (config)#interface range Port-Channel 1
cr22-4507-LB (config-if-range)#service-policy output PQ-POLICER
```

Table 1-2 Summarized Access Layer Ingress QoS Deployment Guidelines

End-Point	Trust Model	DSCP Trust	Classification	Marking	Policing	In Qu
Unmanaged devices, printers etc	UnTrusted	Don't Trust. Default.	None	None	Yes	Ye
Managed secured devices, Servers etc	Trusted	Trust	8 Class Model	Yes	Yes	Ye
Phone	Trusted	Trust	Yes	Yes	Yes	Ye
Phone + Mobile PC	Conditionally-Trusted	Trust	Yes	Yes	Yes	Ye
IP Video surveillance Camera	Trusted	Trust	No	No	No	Ye
Digital Media Player	Trusted	Trust	No	No	No	Ye
Core facing Uplinks	Trusted	Trust	No	No	No	Ye

1. Catalyst 3xxx only.

Table 1-3 Summarized Access Layer Egress QoS Deployment Guidelines

End-Point	Trust Model	Classification / Marking / Policing	Egress Queuing	Ba Sh
Unmanaged devices, printers etc	UnTrusted	None	Yes	Ye
Managed secured devices, Servers etc	Trusted	None	Yes	Ye
Phone	Trusted	None	Yes	Ye
Phone + Mobile PC	Conditionally-Trusted	None	Yes	Ye
IP Video surveillance Camera	Trusted	None	Yes	Ye
Digital Media Player	Trusted	None	Yes	Ye
Core facing Uplinks	Trusted	Yes (PQ Policer)	Yes	Ye

Deploying Network-Layer QoS

Borderless Campus network systems at the large campus and remote medium and small campus are managed and maintained by the enterprise IT administration to provide key network foundation services such as routing, switching, QoS, and virtualization. In a best practice network environment, these systems must be implemented with the recommended configurations to provide differentiated borderless network services on a per-hop basis. To allow for consistent application delivery throughout the network, it is recommended to implement bidirectional QoS policies on distribution and core layer systems.

QoS Trust Boundary

All enterprise IT managed campus LAN and WAN network systems can be classified as trusted devices and must follow the same QoS best practices recommended in a previous subsection. It is recommended to avoid deploying trusted or untrusted endpoints directly to the campus distribution and core layer systems.

Based on the global network QoS policy, each class-of-service application receives the same treatment. Independent of the enterprise network tier—LAN/WAN, platform type, and their capabilities— each device in the network protects service quality and enables communication across the network without degrading application performance.

Implementing Network-Layer Ingress QoS

As described earlier, the internal campus core network must be considered to be trusted. The next-generation Cisco Catalyst access layer platform must be deployed with more application-aware and intelligence at the network edge. The campus core and distribution network devices should rely on the access layer switches to implement QoS classification and marking based on a wide-range of applications and IP-based devices deployed at the network edge.

To provide consistent and differentiated QoS services on per-hop basis across the network, the distribution and core network must be deployed to trust incoming pre-marked DSCP traffic from the downstream Layer 2 or Layer 3 network devices. This Borderless Campus network design recommends deploying a broad range of Layer 3 Catalyst switching platforms in the campus distribution layer and Catalyst 6500-E VSS and Nexus 7000 in the core layer. As mentioned in the previous section, the hardware architecture of each switching platform is different, based on the platform capabilities and resources. This changes how the different class-of-service traffic types are handled in different directions—ingress, switching fabric, and egress.

Cisco Catalyst access layer switches must classify the application and device type to mark DSCP value based on the trust model with deep packet inspection using access lists (ACL) or protocol-based device discovery. Therefore there is no need to reclassify the same class-of-service at the campus distribution and core layers. The campus distribution and core layers can trust DSCP markings from the access layer and provide QoS transparency without modifying the original parameters unless the network is congested.

Based on the simplified internal network trust model, the ingress QoS configuration also becomes more simplified and manageable. This subsection provides common ingress QoS deployment guidelines for the campus distribution and core for all locations:

QoS Trust Mode

As described earlier, the QoS function in the Cisco Nexus 7000 system is trusted and enabled by default. The Nexus 7000 system automatically performs the ingress and egress classification and queuing QoS function with default Cos-to-Queue map settings. The network data traffic is automatically placed in ingress and egress queues based on marking done at the campus access layer to appropriately utilize port bandwidth resources. In the default configuration setting, the Cisco Nexus 7000 protects the original DSCP markings performed by the end-point or an access layer switch.

The Catalyst 4500E deployed with either a Sup7-E, Sup6-E, or Sup6L-E supervisor module in the distribution or in the collapsed core layer automatically sets the physical ports in the trust mode. The Catalyst 4500E by default performs DSCP-CoS or CoS-DSCP mappings to transmit traffic transparently without any QoS bits rewrites. However the default QoS function on campus distribution or core platforms like the 6500-E Series switches is disabled.

When QoS trust is disabled by default, the network administrator must manually enable QoS globally on the switch and explicitly enable DSCP trust mode on each logical EtherChannel and each member link interface connected to upstream and downstream devices. The distribution layer QoS trust

configuration is the same for a multilayer or routed-access deployment. The following sample QoS configuration must be enabled on all the distribution and core layer switches deployed in a campus LAN network.

Distribution and Core Layer 6500-E

```
cr22-6500-LB(config)#mls qos
cr22-6500-LB#show mls qos
  QoS is enabled globally
…
```

Core Layer Nexus 7000

```
cr35-N7K-Core1# show queuing int et1/1 summary | begin Ingress
Ingress Queuing for Ethernet1/1 [Interface]
------------------------------------------------
Template: 8Q2T
Trust: Trusted
```

Implement DSCP Trust Mode

- Catalyst 6500-E

```
cr22-6500-LB(config)#interface Port-channel100
cr22-6500-LB(config-if)# description Connected to cr22-4507-LB
cr22-6500-LB(config-if)# mls qos trust dscp

Catalyst 6500-E will automatically replicate "mls qos trust dscp" command from
port-channel interface to each bundled member-links.

cr22-6500-LB#show queueing interface Ten1/1/2 | inc QoS|Trust
  Port QoS is enabled
  Trust boundary disabled
  Trust state: trust DSCP
```

- Cisco Nexus 7000

  This document characterized the Cisco Nexus 7000 running the Cisco NX-OS 5.1.3 software version which currently supports a CoS-based trust model and does not support a DSCP-based model for ingress or egress network traffic. The incoming pre-marked DSCP data traffic is automatically classified and appropriately queued based on the internal system CoS-to-DSCP mapping table.

```
cr35-N7K-Core1#show queuing interface Ethernet1/1 | inc Mode
    Queuing Mode in TX direction: mode-cos
    Queuing Mode in RX direction: mode-cos
```

Applying Ingress Queuing

The Catalyst 6500-E and Nexus 7000 system support the ingress queuing function to classify ingress traffic and place it in the ingress queue for ingress scheduling and data prioritization prior to forwarding data traffic to the switch fabric. Implementing the ingress queuing function is effective when the high-speed ports on the network module operate at an over-subscription rate. In a non-oversubscription network module, the ingress queuing provides data prioritization and protection if internal fabric backplane bandwidth is reduced, creating internal forwarding congestion.

The Cisco Catalyst 4500E deployed with a Sup7-E, Sup6-E, or a Sup6L-E supervisor module does not support ingress queuing.

Implementing Catalyst 6500-E Ingress Queuing

When 6500-E switching platforms receive various class-of-service requests from different physical ports, then depending on the DSCP and CoS markings, they can queue the traffic prior sending it to the switching fabric in a FIFO manner. There are two main considerations relevant to ingress queuing design on the Catalyst 6500-E:

- The degree of oversubscription (if any) of the linecard
- Whether the linecard requires trust-CoS to be enabled to engage ingress queuing

Some linecards may be designed to support a degree of oversubscription that theoretically offers more traffic to the linecard than the sum of all GE/10GE switch ports than can collectively access the switching backplane at once. Since such a scenario is extremely unlikely, it is often more cost-effective to use linecards that have a degree of oversubscription within the campus network. However, if this design choice has been made, it is important for network administrators to recognize the potential for drops due to oversubscribed linecard architectures. To manage application class service levels during such extreme scenarios, ingress queuing models may be enabled.

While the presence of oversubscribed linecard architectures may be viewed as the sole consideration in enabling or not enabling ingress queuing, a second important consideration is that many Catalyst 6500-E linecards only support CoS-based ingress queuing models that reduce classification and marking granularity—limiting the administrator to an 8-class 802.1Q/p model. Once CoS is trusted, DSCP values are overwritten (via the CoS-to-DSCP mapping table) and application classes sharing the same CoS values are longer distinguishable from one another. Therefore, given this classification and marking limitation and the fact that the value of enabling ingress queuing is only achieved in extremely rare scenarios, it is not recommended to enable CoS-based ingress queuing on the Catalyst 6500-E. Rather, limit such linecards and deploy either non-oversubscribed linecards and/or linecards supporting DSCP-based queuing at the distribution and core layers of the campus network.

Table 1 summarizes recommended linecards by listing and oversubscription ratios and whether the ingress queuing models are CoS- or DSCP-based.

Table 1-4 Catalyst 6500-E Switch Module Ingress Queuing Architecture

Switch Module	Maximum Input	Maximum Output (To Backplane)	Oversubscription Ratio	Ingress Queuing Structure	CoS / DSCP Based	Ingress Recomm
WS-6724-SFP	24 Gbps (24 x GE ports)	40 Gbps (2 x 20 Gbps)	-	1P3Q8T	CoS based	Not Rec
WS-6704-10 GE	40 Gbps (4 x 10GE ports)		-	8Q8T	CoS or DSCP based	Not Rec
WS-6708-10 GE	80 Gbps (8 x 10GE ports)		2:1	8Q4T	CoS or DSCP based	Use DSC 8Q4T in queuing
WS-6716-10 GE	160 Gbps (16 x 10GE ports)		4:1	8Q4T / 1P7Q2T*	CoS or DSCP based	Use DSC 1P7Q2T queuing

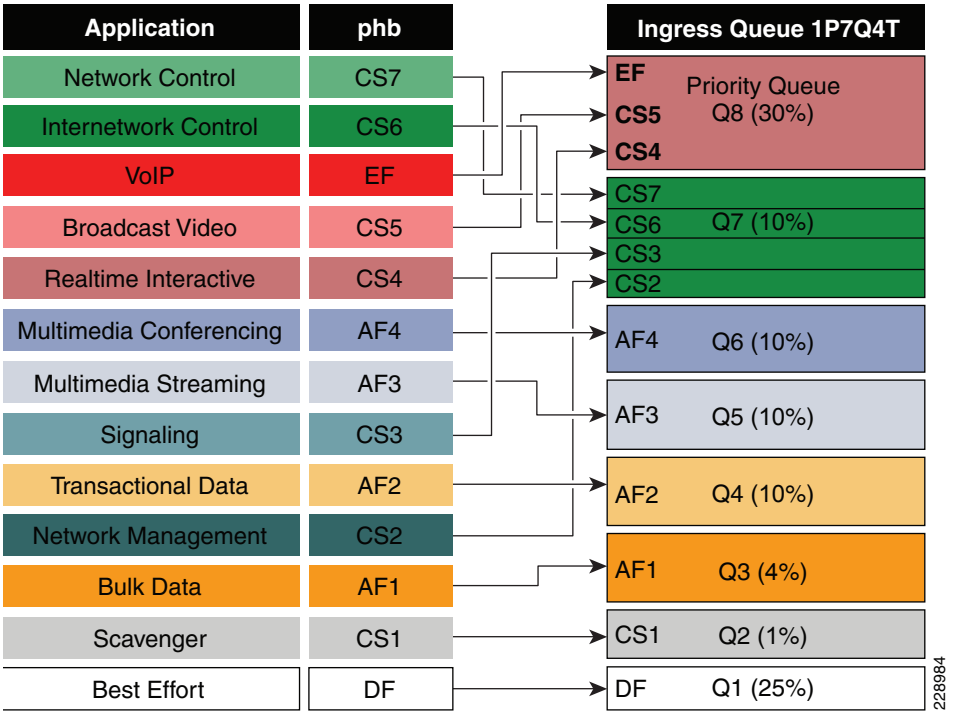
Note The Catalyst WS-X6716-10GE can be configured to operate in Performance Mode (with an 8Q4T ingress queuing structure) or in Oversubscription Mode (with a 1P7Q2T ingress queuing structure). In Performance mode, only one port in every group of four is operational (while the rest are administratively shut down), which eliminates any oversubscription on this linecard and as such ingress queuing is not required (as only 4 x 10GE ports are active in this mode and the backplane access rate is also at 40 Gbps). In Oversubscription Mode (the default mode), all ports are operational and the maximum oversubscription ratio is 4:1. Therefore, it is recommended to enable 1P7Q2T DSCP-based ingress queuing on this linecard in Oversubscription Mode.

Additional details on these WS-X6716-10GE operational modes can be found at: http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/qa_cisco_catalyst_6500_series_16port_10gigabit_ethernet_module.html

If 6708 and 6716 linecards (with the latter operating in oversubscription mode) are used in the distribution and core layers of the campus network, then 8Q4T DSCP-based ingress queuing and 1P7Q2T DSCP-based ingress queuing (respectively) are recommended to be enabled. These queuing models are detailed in the following sections.

Figure 16 depicts how different class-of-service applications are mapped to the ingress queue structure (8Q4T) and how each queue is assigned a different WTD threshold.

Figure 16 Catalyst 6500-E Ingress Queuing Model



The corresponding configuration for 8Q8T (DSCP-to-Queue) ingress queuing on a Catalyst 6500-E VSS in the distribution and core layers is shown below. The PFC function is active on active and hot-standby virtual-switch nodes; therefore, ingress queuing must be configured on each distributed member link of Layer 2 or Layer 3 MEC.

- Distribution and Core-Layer Catalyst 6500-E in VSS mode

```
! This section configures the port for DSCP-based Ingress queuing
cr22-vss-core(config)#interface range TenGigabitEthernet 1/1/2 - 8 , 2/1/2-8
cr22-vss-core(config-if-range)# mls qos queue-mode mode-dscp
 ! Enables DSCP-to-Queue mapping

! This section configures the receive queues BW and limits
cr22-vss-core(config-if-range)# rcv-queue queue-limit 10 25 10 10 10 10 10 15
 ! Allocates 10% to Q1, 25% to Q2, 10% to Q3, 10% to Q4,
 ! Allocates 10% to Q5, 10% to Q6, 10% to Q7 and 15% to Q8
cr22-vss-core(config-if-range)# rcv-queue bandwidth 1 25 4 10 10 10 10 30
 ! Allocates 1% BW to Q1, 25% BW to Q2, 4% BW to Q3, 10% BW to Q4,
 ! Allocates 10% BW to Q5, 10% BW to Q6, 10% BW to Q7 & 30% BW to Q8
```

```
! This section enables WRED on all queues except Q8
cr22-vss-core(config-if-range)# rcv-queue random-detect 1
 ! Enables WRED on Q1
cr22-vss-core(config-if-range)# rcv-queue random-detect 2
 ! Enables WRED on Q2
cr22-vss-core(config-if-range)# rcv-queue random-detect 3
 ! Enables WRED on Q3
cr22-vss-core(config-if-range)# rcv-queue random-detect 4
 ! Enables WRED on Q4
cr22-vss-core(config-if-range)# rcv-queue random-detect 5
 ! Enables WRED on Q5
cr22-vss-core(config-if-range)# rcv-queue random-detect 6
 ! Enables WRED on Q6
cr22-vss-core(config-if-range)# rcv-queue random-detect 7
 ! Enables WRED on Q7
cr22-vss-core(config-if-range)# no rcv-queue random-detect 8
 ! Disables WRED on Q8

! This section configures WRED thresholds for Queues 1 through 7
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 1 100 100 100 100
 ! Sets all WRED max thresholds on Q1 to 100%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 1 80 100 100 100
 ! Sets Q1T1 min WRED threshold to 80%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 2 80 100 100 100
 ! Sets Q2T1 min WRED threshold to 80%
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 2 100 100 100 100
 ! Sets all WRED max thresholds on Q2 to 100%

cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 3 70 80 90 100
 ! Sets WRED min thresholds for Q3T1, Q3T2, Q3T3 to 70 %, 80% and 90%
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 3 80 90 100 100
 ! Sets WRED max thresholds for Q3T1, Q3T2, Q3T3 to 80%, 90% and 100%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 4 70 80 90 100
 ! Sets WRED min thresholds for Q4T1, Q4T2, Q4T3 to 70 %, 80% and 90%
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 4 80 90 100 100
 ! Sets WRED max thresholds for Q4T1, Q4T2, Q4T3 to 80%, 90% and 100%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 5 70 80 90 100
 ! Sets WRED min thresholds for Q5T1, Q5T2, Q5T3 to 70 %, 80% and 90%
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 5 80 90 100 100
! Sets WRED max thresholds for Q5T1, Q5T2, Q5T3 to 80%, 90% and 100%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 6 70 80 90 100
 ! Sets WRED min thresholds for Q6T1, Q6T2, Q6T3 to 70 %, 80% and 90%
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 6 80 90 100 100
 ! Sets WRED max thresholds for Q6T1, Q6T2, Q6T3 to 80%, 90% and 100%
cr22-vss-core(config-if-range)# rcv-queue random-detect min-threshold 7 60 70 80 90
 ! Sets WRED min thresholds for Q7T1, Q7T2, Q7T3 and Q7T4
 ! to 60%, 70%, 80% and 90%, respectively
cr22-vss-core(config-if-range)# rcv-queue random-detect max-threshold 7 70 80 90 100
 ! Sets WRED max thresholds for Q7T1, Q7T2, Q7T3 and Q7T4
```

```
 ! to 70%, 80%, 90% and 100%, respectively

! This section configures the DSCP-to-Receive-Queue mappings
cr22-vss-core(config-if-range)# rcv-queue dscp-map 1 1 8
 ! Maps CS1 (Scavenger) to Q1T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 2 1 0
 ! Maps DF (Best Effort) to Q2T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 3 1 14
! Maps AF13 (Bulk Data-Drop Precedence 3) to Q3T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 3 2 12
 ! Maps AF12 (Bulk Data-Drop Precedence 2) to Q3T2
cr22-vss-core(config-if-range)# rcv-queue dscp-map 3 3 10
 ! Maps AF11 (Bulk Data-Drop Precedence 1) to Q3T3
cr22-vss-core(config-if-range)# rcv-queue dscp-map 4 1 22
 ! Maps AF23 (Transactional Data-Drop Precedence 3) to Q4T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 4 2 20
 ! Maps AF22 (Transactional Data-Drop Precedence 2) to Q4T2
cr22-vss-core(config-if-range)# rcv-queue dscp-map 4 3 18
 ! Maps AF21 (Transactional Data-Drop Precedence 1) to Q4T3
cr22-vss-core(config-if-range)# rcv-queue dscp-map 5 1 30
 ! Maps AF33 (Multimedia Streaming-Drop Precedence 3) to Q5T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 5 2 28
 ! Maps AF32 (Multimedia Streaming-Drop Precedence 2) to Q5T2
cr22-vss-core(config-if-range)# rcv-queue dscp-map 5 3 26
 ! Maps AF31 (Multimedia Streaming-Drop Precedence 1) to Q5T3
cr22-vss-core(config-if-range)# rcv-queue dscp-map 6 1 38
 ! Maps AF43 (Multimedia Conferencing-Drop Precedence 3) to Q6T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 6 2 36
 ! Maps AF42 (Multimedia Conferencing-Drop Precedence 2) to Q6T2
cr22-vss-core(config-if-range)# rcv-queue dscp-map 6 3 34
 ! Maps AF41 (Multimedia Conferencing-Drop Precedence 1) to Q6T3
cr22-vss-core(config-if-range)# rcv-queue dscp-map 7 1 16
 ! Maps CS2 (Network Management) to Q7T1
cr22-vss-core(config-if-range)# rcv-queue dscp-map 7 2 24
 ! Maps CS3 (Signaling) to Q7T2
cr22-vss-core(config-if-range)# rcv-queue dscp-map 7 3 48
 ! Maps CS6 (Internetwork Control) to Q7T3
cr22-vss-core(config-if-range)# rcv-queue dscp-map 7 4 56
 ! Maps CS7 (Network Control) to Q7T4
cr22-vss-core(config-if-range)# rcv-queue dscp-map 8 4 32 40 46
 ! Maps CS4 (Realtime Interactive), CS5 (Broadcast Video),
 ! and EF (VoIP) to Q8

cr23-VSS-Core#show queueing interface Ten1/1/2 | begin Rx
    Queueing Mode In Rx direction: mode-dscp
    Receive queues [type = 8q4t]:
    Queue Id    Scheduling  Num of thresholds
    -----------------------------------------
       01         WRR                 04
```

```
        02          WRR                 04
        03          WRR                 04
        04          WRR                 04
        05          WRR                 04
        06          WRR                 04
        07          WRR                 04
        08          WRR                 04

    WRR bandwidth ratios:    1[queue 1]  25[queue 2]   4[queue 3]  10[queue 4]  10[queue
5]  10[queue 6]  10[queue 7]  30[queue 8]
    queue-limit ratios:     10[queue 1]  25[queue 2]  10[queue 3]  10[queue 4]  10[queue
5]  10[queue 6]  10[queue 7]  15[queue 8]

    queue tail-drop-thresholds
    --------------------------
    1     70[1] 80[2] 90[3] 100[4]
    2     100[1] 100[2] 100[3] 100[4]
    3     100[1] 100[2] 100[3] 100[4]
    4     100[1] 100[2] 100[3] 100[4]
    5     100[1] 100[2] 100[3] 100[4]
    6     100[1] 100[2] 100[3] 100[4]
    7     100[1] 100[2] 100[3] 100[4]
    8     100[1] 100[2] 100[3] 100[4]

    queue random-detect-min-thresholds
    ----------------------------------
      1    80[1] 100[2] 100[3] 100[4]
      2    80[1] 100[2] 100[3] 100[4]
      3    70[1] 80[2] 90[3] 100[4]
      4    70[1] 80[2] 90[3] 100[4]
      5    70[1] 80[2] 90[3] 100[4]
      6    70[1] 80[2] 90[3] 100[4]
      7    60[1] 70[2] 80[3] 90[4]
      8    100[1] 100[2] 100[3] 100[4]

    queue random-detect-max-thresholds
    ----------------------------------
      1    100[1] 100[2] 100[3] 100[4]
      2    100[1] 100[2] 100[3] 100[4]
      3    80[1] 90[2] 100[3] 100[4]
      4    80[1] 90[2] 100[3] 100[4]
      5    80[1] 90[2] 100[3] 100[4]
      6    80[1] 90[2] 100[3] 100[4]
      7    70[1] 80[2] 90[3] 100[4]
      8    100[1] 100[2] 100[3] 100[4]

    WRED disabled queues:      8
…
    queue thresh dscp-map
```

```
    ----------------------------------------
    1      1       1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 39 41 42 43 44 45
47
    1      2
    1      3
    1      4
    2      1       0
    2      2
    2      3
    2      4
    3      1       14
    3      2       12
    3      3       10
    3      4
    4      1       22
    4      2       20
    4      3       18
    4      4
    5      1       30 35 37
    5      2       28
    5      3       26
    5      4
    6      1       38 49 50 51 52 53 54 55 57 58 59 60 61 62 63
    6      2       36
    6      3       34
    6      4
    7      1       16
    7      2       24
    7      3       48
    7      4       56
    8      1
    8      2
    8      3
    8      4       32 40 46
…
  Packets dropped on Receive:
    BPDU packets:  0

    queue               dropped  [dscp-map]
    ----------------------------------------------
    1                         0  [1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 39
41 42 43 44 45 47 ]
    2                         0  [0 ]
    3                         0  [14 12 10 ]
    4                         0  [22 20 18 ]
    5                         0  [30 35 37 28 26 ]
    6                         0  [38 49 50 51 52 53 54 55 57 58 59 60 61 62 63 36 34 ]
    7                         0  [16 24 48 56 ]
    8                         0  [32 40 46 ]
```

Implementing Cisco Nexus 7000 Ingress Queuing

The Nexus 7000 system supports ingress queuing on all M1 series I/O modules, however implementing the ingress queuing policy is different in the Nexus 7000 system. The NX-OS supports pre-defined multiple ingress queuing class-maps for different models—8Q2T and 2Q4T. Based on the ingress queue capability of the I/O module, the network administrator can create an ingress queue policy-map and leverage system-defined ingress queuing class-maps to classifying CoS value for each queue. The ingress queue policies are by default attached to every physical port of the Nexus 7000 system and are always in effect until new user-defined QoS policies are applied to the port to override the default configuration. The default ingress class-map names cannot be modified or removed, however the NX-OS provides the flexibility to remap default CoS-to-Queue as needed.

The Cisco Nexus 7000 supports 8 port 10Gbps M108 and 32 port 10Gbps M132 series I/O modules with advanced Layer 3 campus core layer technologies. Both modules operate at 10Gbps and provide up to 80Gbps backplane switching capacity. However these modules are designed to be deployed in a specific network layer to provide the appropriate level port scalability and switching performance. The M108 is a high-performance I/O module designed to be deployed in the high-speed campus and data center core layer. This module provides wire-speed throughput and distributed services, such as QoS and ACL, through a dual-forwarding engine attached to a single M108 module. The M132 I/O module is commonly deployed in a high-density aggregation layer to interconnect access layer switches through high-speed 10Gbps links.

The default port bandwidth and resource allocation settings for each physical port differ on both module types. Table 1 summarizes the ingress QoS comparison between both Nexus 7000 M1 series I/O modules.

Table 1-5 Cisco Nexus 7000 M1 Ingress QoS Architecture

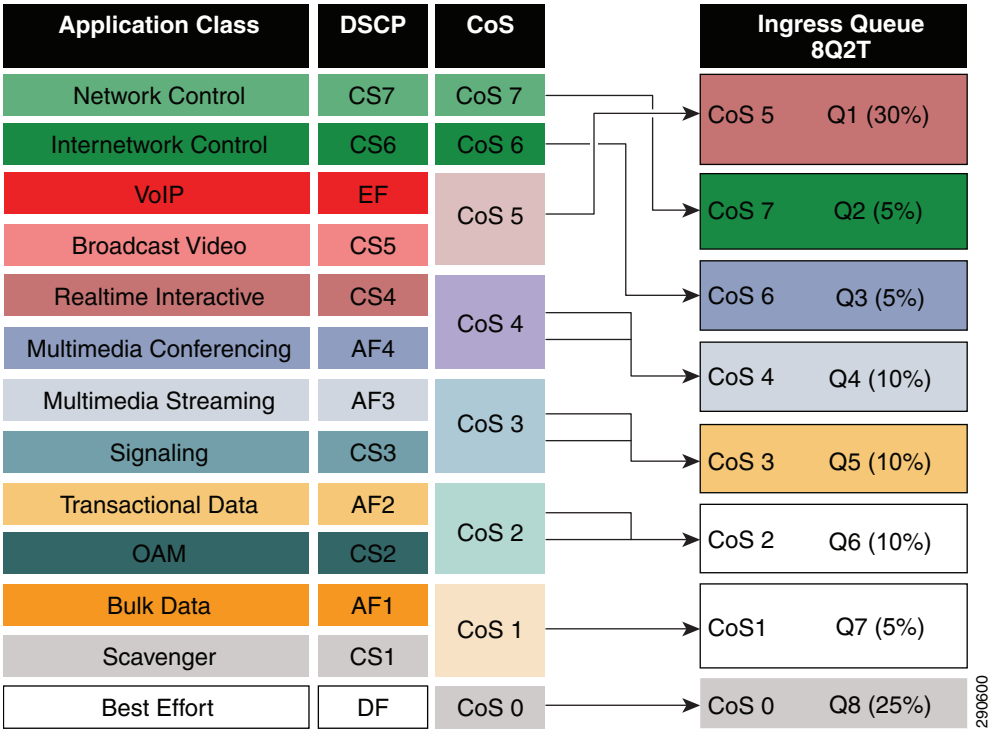
I/O Module	Maximum Input	Maximum Output (To Backplane)	Oversubscription Ratio	Ingress Queue Structure	Trust Mode	Ingress Queuing Recommendatio
M108	80 Gbps (8 x 10 GE ports)	80 Gbps (2 x Crossbar Fabric)[1]	-	8Q2T	CoS Based	Use 8Q2T CoS ingress queueing
M132	80 Gbps (32 x 10GE ports)	80 Gbps (2 x Crossbar Fabric)[1]	4:1 (Shared Mode)	2Q4T	CoS Based	Use 2Q4T CoS ingress queueing
M132	80 Gbps (8 x 10GE ports)[2]	80 Gbps (2 x Crossbar Fabric)[1]	- (Dedicated Mode)	8Q2T	CoS Based	Use 8Q2T CoS ingress queueing

1. Requires at least dual crossbar fabric module for a maximum 80 Gbps backplane throughput from each M1 series I/O modul

2. 8 ports in operational state and the remaining 24 x 10GE port must be disable for dedicated mode.

Figure 17 illustrates how different class-of-service applications are mapped to the ingress queue structure (8Q2T) and how each queue is assigned a different bandwidth and WTD threshold.

Figure 17 Nexus 7000 CoS-Based 8Q2T Ingress QoS Model



For each ingress queue model, the NX-OS currently supports, system-wide, a single set of ingress queue class-maps (for example, a single CoS classification rule for each ingress queue). In a best practice campus network design, the network administrator should follow the recommendation to implement ingress policy based on a 12-class QoS model. Such a QoS design provides consistent QoS treatment on a hop-by-hop basis in a campus network.

As described previously, the Cisco Nexus 7000 system supports a default system-defined ingress queuing and the default policy-map is attached by default to each physical port of the system. The default ingress queuing policy by default uses the first queue and the default or last queue. The default settings can be verified with the **show policy-map** command. as illustrated below:

```
cr35-N7K-Core1# show policy-map interface Ethernet 1/1 input type queuing
```

```
Global statistics status :   enabled
Ethernet1/1
  Service-policy (queuing) input:   default-in-policy
    policy statistics status:   enabled (current status: enabled)
    Class-map (queuing):   in-q1 (match-any)
      queue-limit percent 50
      bandwidth percent 80
      queue dropped pkts : 0
    Class-map (queuing):   in-q-default (match-any)
      queue-limit percent 50
      bandwidth percent 20
      queue dropped pkts : 0
```

The default class-map may vary based on the type of linecard used. The system default in-q1 class-map is general class-map for the module that supports 8Q2T or 2Q4T. Since the M108 module supports a 8Q2T ingress queuing model, the network administrator may verify default CoS-to-Queue settings based on 8q2t-in.

```
cr35-n7k-Core1#show class-map type queuing in-q1
Class-maps referred to in default policies can be printed using
appropriate show commands of that particular linecard type.
Prefix the linecard type in front to get the actual cos2q map. e.g.
        To print in-q1, use 2q4t-in-q1 or 8q2t-in-q1

cr35-n7k-Core1#show class-map type queuing 8q2t-in-q1
  Type queuing class-maps
  ========================
    class-map type queuing match-any 8q2t-in-q1
      Description: Classifier for ingress queue 1 of type 8q2t
      match cos 5-7

cr35-n7k-Core1# show class-map type queuing 8q2t-in-q-default
  Type queuing class-maps
  ========================
    class-map type queuing match-any 8q2t-in-q-default
      Description: Classifier for ingress default queue of type 8q2t
      match cos 0-4
```

The system-defined default ingress bandwidth allocation, queue-limit, and drop threshold parameters are also different than the recommended ingress QoS model as illustrated in Figure 17.

```
cr35-n7k-Core1# show queuing int et1/1 summary | be Ingress
Ingress Queuing for Ethernet1/1 [Interface]
------------------------------------------------
Template: 8Q2T
Trust: Trusted
-----------------------------------
Que# Group Qlimit% IVL      CoSMap
-----------------------------------
   0     -      50   -         0-4
```

```
   1      -        0    -            -
   2      -        0    -            -
   3      -        0    -            -
   4      -        0    -            -
   5      -        0    -            -
   6      -        0    -            -
   7      -       50    -          5-7

cr35-n7k-Core1# show queuing int et1/1 | be "RX Queuing"
  Interface Ethernet1/1 RX Queuing strategy: Weighted Round-Robin
    Queuing Mode in RX direction: mode-cos
    Receive queues [type = 8q2t]
    Port Cos not configured

    <snip>

   Configured WRR
      WRR bandwidth ratios:   20[8q2t-in-q-default] 0[8q2t-in-q2] 0[8q2t-in-q3]
0[8q2t-in-q4] 0[8q2t-in-q5] 0[8q2t-in-q6] 0[8q2t-in-q7] 80[8q2t-in-q1]

   <snip>
```

As described previously, the system-defined ingress queue QoS policy is active and operational with bandwidth and thresholds for ingress queue 1 and ingress queue 8 class-maps. By default no CoS bits are mapped or bandwidth assigned to the remaining ingress queues, hence these queues are not utilized and remain un-used. If the network administrator modifies the system-defined default queuing policy-maps without following implementation best practices, the campus core layer network may experience instability that can cause severe service disruption.

Modifying the default class-map configuration by re-mapping CoS values to un-used class-maps may result in insufficient bandwidth for internal processing. Due to such mis-configuration and lack of internal bandwidth, network control traffic may be impacted and de-stabilize the entire network topology.

Note It is highly recommended to follow best practices to seamlessly enable ingress queuing policy on the Cisco Nexus 7000 system. The network administrator must follow the exact steps described below to implement ingress queuing policy.

Step 1 Create the ingress queue policy-map. To utilize all ingress queues for a 12-class QoS model, the policy-map should have all 8Q2T class-maps assigned in this policy. On a per-class basis, apply the recommended bandwidth, queue-limit, and WRED parameters as required for ingress class-map:

```
!Create ingress queuing policy-map
cr35-n7k-Core1(config)# policy-map type queuing INGRESS-POLICY
```

```
!Assign ingress bandwidth and queue-limit for Q1 for CoS=5 traffic
cr35-n7k-Core1(config-pmap-que)#   class type queuing 8q2t-in-q1
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 30
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 15

!Assign ingress bandwidth and queue-limit for Q2 for CoS=7 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q2
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 5
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10

!Assign ingress bandwidth and queue-limit for Q3 for CoS=6 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q3
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 5
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10

!Assign ingress bandwidth, queue-limit and WRED for Q4 for CoS=4 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q4
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 10
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos-based
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos 4 minimum-threshold percent 80
maximum-threshold percent 100

!Assign ingress bandwidth, queue-limit and WRED for Q5 for CoS=3 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q5
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 10
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos-based
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos 3 minimum-threshold percent 80
maximum-threshold percent 100

!Assign ingress bandwidth, queue-limit and WRED for Q6 for CoS=2 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q6
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 10
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos-based
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos 2 minimum-threshold percent 80
maximum-threshold percent 100

!Assign ingress bandwidth, queue-limit and WRED for Q7 for CoS=1 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q7
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 5
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 10
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos-based
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos 1 minimum-threshold percent 80
maximum-threshold percent 100

!Assign ingress bandwidth, queue-limit and WRED for Q8 for CoS=0 traffic
cr35-n7k-Core1(config-pmap-c-que)# class type queuing 8q2t-in-q-default
```

```
cr35-n7k-Core1(config-pmap-c-que)# queue-limit percent 25
cr35-n7k-Core1(config-pmap-c-que)# bandwidth percent 25
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos-based
cr35-n7k-Core1(config-pmap-c-que)# random-detect cos 0 minimum-threshold percent 80
maximum-threshold percent 100
```

Step 2 Attach the ingress queuing policy-map to a physical Layer 3 interface or to a logical port channel interface in an EtherChannel-based network design. The QoS policy automatically becomes effective on each member link of EtherChannel once attached to a port-channel interface:

```
cr35-N7K-Core1(config)# int Ethernet 1/1 , Ethernet 2/1
cr35-N7K-Core1(config-if-range)# service-policy type queuing input INGRESS-POLICY

cr35-N7K-Core1(config)# int Port-Channel 100 - 103
cr35-N7K-Core1(config-if-range)# service-policy type queuing input INGRESS-POLICY
cr35-N7K-Core1(config-if-range)#

cr35-N7K-Core1# show policy-map interface brief

 Interface/VLAN [Status]:INP QOS       OUT QOS       INP QUE       OUT QUE
================================================================================
port-channel100 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel101 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel102 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel103 [Active]:                           INGRESS-POLIC EGRESS-POLICY
Ethernet1/1      [Active]:                           INGRESS-POLIC EGRESS-POLICY
<snip>

cr35-N7K-Core1# show queuing interface Ethernet 1/1 summary | be Ingress

Ingress Queuing for Ethernet1/1 [Interface]
------------------------------------------------
Template: 8Q2T
Trust: Trusted
----------------------------------
Que# Group Qlimit% IVL      CoSMap
----------------------------------
   0     -      25   -           0
   1     -      10   -           7
   2     -      10   -           6
   3     -      10   -           4
   4     -      10   -           3
   5     -      10   -           2
   6     -      10   -           1
   7     -      15   -           5
```

Step 3 Once the ingress queue policy-map is created and bandwidth and queue-limit are allocated to each class, the default CoS-to-Queue can be safely re-mapped across each ingress queue class-map that was created in Step 1. To utilize each ingress queue, the network administrator must assign a single CoS value to enable one queue per class configuration:

```
cr35-N7K-Core1(config)# class-map type queuing match-any 8q2t-in-q1
cr35-N7K-Core1(config-cmap-que)#   match cos 5
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q2
cr35-N7K-Core1(config-cmap-que)#   match cos 7
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q3
cr35-N7K-Core1(config-cmap-que)#   match cos 6
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q4
cr35-N7K-Core1(config-cmap-que)#   match cos 4
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q5
cr35-N7K-Core1(config-cmap-que)#   match cos 3
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q6
cr35-N7K-Core1(config-cmap-que)#   match cos 2
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q7
cr35-N7K-Core1(config-cmap-que)#   match cos 1
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 8q2t-in-q-default
cr35-N7K-Core1(config-cmap-que)#   match cos 0
```

Implementing Network Core Egress QoS

The QoS implementation of egress traffic towards network edge devices on access layer switches are much simplified compared to ingress traffic, which requires stringent QoS policies to provide differentiated services and network bandwidth protection. Unlike the ingress QoS model, the egress QoS model must provide optimal queuing policies for each class and sets the drop thresholds to prevent network congestion and an impact to application performance. With egress queuing in DSCP mode, the Cisco Catalyst switching platforms and linecards are bounded by a limited number of egress hardware queues.

Catalyst 4500E

The configuration and implementation guidelines for egress QoS on the Catalyst 4500E with Sup7-E, Sup6-E, or Sup6L-E in distribution and access layer roles remains consistent. All conformed traffic marked with DSCP values must be manually assigned to each egress queue based on a four class-of-service QoS model. Refer to the "Implementing Access Layer Egress QoS" section on page 34 for the deployment details.

Catalyst 6500-E—VSS

The Cisco Catalyst 6500-E in VSS mode operates in a centralized management mode, but uses a distributed forwarding architecture. The Policy Feature Card (PFC) on active and hot-standby is functional on both nodes and is independent of the virtual switch role. Like ingress queuing, the network administrator must implement egress queuing on each of the member links of the Layer 2 or Layer 3 MEC. The egress queuing model on the Catalyst 6500-E is based on linecard type and its capabilities; when deploying Catalyst 6500-E in VSS mode only, the WS-67xx series 1G/10G linecard with daughter card CFC or DFC3C/DFC3CXL is supported.

Table 1 describes the deployment guidelines for the Catalyst 6500-E Series linecard module in the campus distribution and core layer network. In the solutions lab, WS-6724-SFP and WS-6708-10GE were validated in the campus distribution and core layers. As both modules support different egress queuing models, this subsection provides deployment guidelines for both module types.

Table 1-6 Catalyst 6500-E Switch Module Egress Queuing Architecture

Switch Module	Daughter Card	Egress Queue and Drop Thresholds	Egress Queue Scheduler	Total Buffer Size	Egres Size
WS-6724-SFP	CFC or DFC3	1P3Q8T	DWRR	1.3 MB	1.2 M
WS-6704-10GE	CFC	1P7Q8T	DWRR	16 MB	14 M
	DFC3				
WS-6708-10GE	DFC3	1P7Q4T	DWRR SRR	198 MB	90 M
WS-6716-10GE	DFC3	1P7Q8T (Oversubscription and Perf. Mode)		198 MB[1] 91 MB[2]	90 M 1 MB

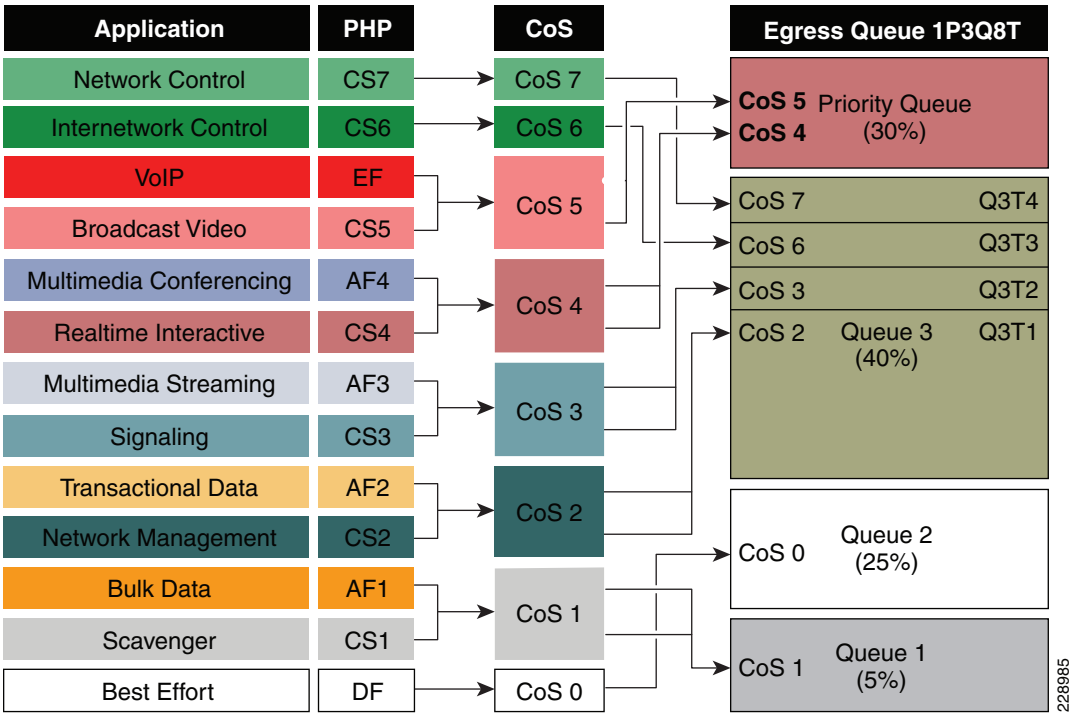
1. Per Port Capacity in Performance Mode
2. Per Port Capacity in Oversubscription Mode

WS-6724-SFP—1P3Q8T Egress Queuing Model

On the WS-6724-SFP module the egress queuing functions on a per-physical-port basis and independent of link layer and above protocols settings; these functions remain consistent when the physical port is deployed in standalone or bundled into an EtherChannel. Each 1G physical port supports four egress queues with the default CoS based on the transmit side. This module is a cost-effective, 1G, non-blocking, high-speed network module, but does not provide deep application granularity based on different DSCP markings. It does not have the flexibility to use various class-of-service egress queues for applications. Campus LAN QoS consolidation to a four class model occurs on the physical paths that connect to the WAN or Internet edge routers, which forward traffic

across a private WAN or the Internet. Deploying the WS-6724-SFP module in a four class model would be recommended in that design. Figure 18 illustrates 1P3Q8T egress queuing model to be applied on the Catalyst 6500-E – WS-6724-SF module.

Figure 18 Catalyst 6500-E CoS-based 1P3Q8T Egress QoS Model



The following corresponding 1P3Q8T egress queuing configuration must be applied on each member link of MEC.

- Catalyst 6500-E VSS (Distribution and Core)

```
cr23-vss-core(config)#interface range GigabitEthernet 1/2/1-24 , Gi2/2/1 - 24
cr23-vss-core(config-if-range)# wrr-queue queue-limit 20 25 40
! Allocates 20% of the buffers to Q1, 25% to Q2 and 40% to Q3
cr23-vss-core(config-if-range)# priority-queue queue-limit 15
 ! Allocates 15% of the buffers to the PQ
cr23-vss-core(config-if-range)# wrr-queue bandwidth 5 25 40
 ! Allocates 5% BW to Q1, 25% BW to Q2 and 30% BW to Q3

 ! This section enables WRED on Queues 1 through 3
cr23-vss-core(config-if-range)# wrr-queue random-detect 1
```

```
 ! Enables WRED on Q1
cr23-vss-core(config-if-range)# wrr-queue random-detect 2
 ! Enables WRED on Q2
cr23-vss-core(config-if-range)# wrr-queue random-detect 3
 ! Enables WRED on Q3

 ! This section configures WRED thresholds for Queues 1 through 3
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 1 100 100 100 100
100 100 100 100
 ! Sets all WRED max thresholds on Q1 to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 1 80 100 100 100 100
100 100 100
 ! Sets Q1T1 min WRED threshold to 80%; all others set to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 2 100 100 100 100
100 100 100 100
 ! Sets all WRED max thresholds on Q2 to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 2 80 100 100 100 100
100 100 100
 ! Sets Q2T1 min WRED threshold to 80%; all others set to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 3 70 80 90 100 100
100 100 100
 ! Sets Q3T1 max WRED threshold to 70%; Q3T2 max WRED threshold to 80%;
 ! Sets Q3T3 max WRED threshold to 90%; Q3T4 max WRED threshold to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 3 60 70 80 90 100
100 100 100
 ! Sets Q3T1 min WRED threshold to 60%; Q3T2 min WRED threshold to 70%;
 ! Sets Q3T3 min WRED threshold to 80%; Q3T4 min WRED threshold to 90%

 ! This section configures the CoS-to-Queue/Threshold mappings
cr23-vss-core(config-if-range)# wrr-queue cos-map 1 1 1
 ! Maps CoS 1 (Scavenger and Bulk Data) to Q1T1
cr23-vss-core(config-if-range)# wrr-queue cos-map 2 1 0
 ! Maps CoS 0 (Best Effort) to Q2T1
cr23-vss-core(config-if-range)# wrr-queue cos-map 3 1 2
 ! Maps CoS 2 (Network Management and Transactional Data) to Q3T1
cr23-vss-core(config-if-range)# wrr-queue cos-map 3 2 3
 ! Maps CoS 3 (Signaling and Multimedia Streaming) to Q3T2
cr23-vss-core(config-if-range)# wrr-queue cos-map 3 3 6
 ! Maps CoS 6 (Internetwork Control) to Q3T3
cr23-vss-core(config-if-range)# wrr-queue cos-map 3 4 7
 ! Maps CoS 7 (Network Control) to Q3T4
cr23-vss-core(config-if-range)# priority-queue cos-map 1 4 5
 ! Maps CoS 4 (Realtime Interactive and Multimedia Conferencing) to PQ
 ! Maps CoS 5 (VoIP and Broadcast Video) to the PQ

cr23-VSS-Core#show queueing interface GigabitEthernet 1/2/1
Interface GigabitEthernet1/2/1 queueing strategy:  Weighted Round-Robin
  Port QoS is enabled
Trust boundary disabled
```

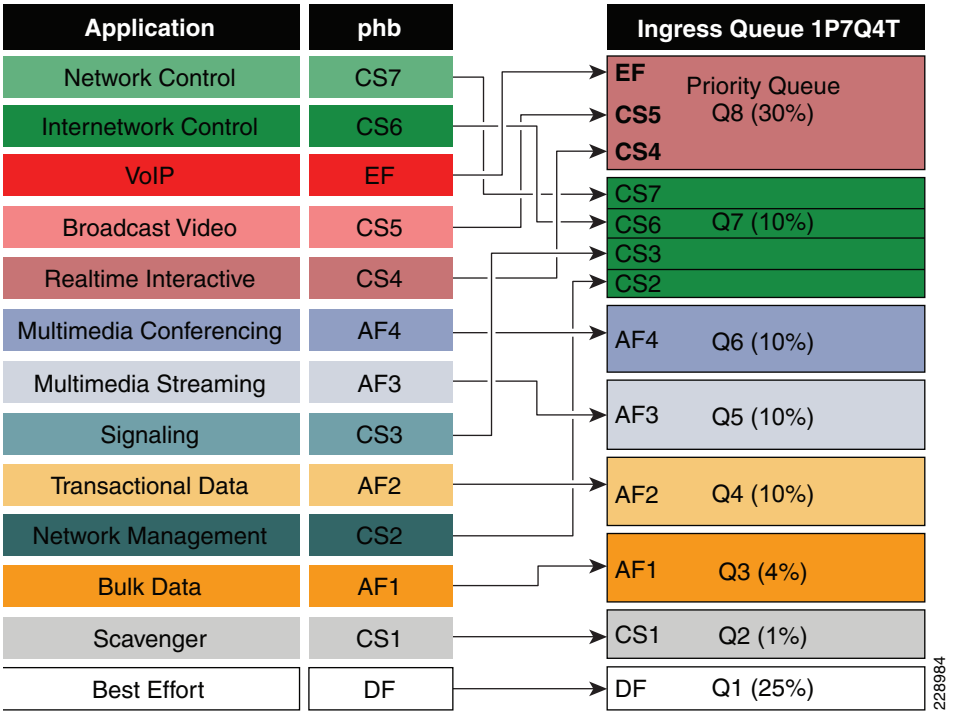
```
Trust state: trust DSCP
Extend trust state: not trusted [COS = 0]
Default COS is 0
  Queueing Mode In Tx direction: mode-cos
  Transmit queues [type = 1p3q8t]:
  Queue Id    Scheduling  Num of thresholds
  -----------------------------------------
     01         WRR                 08
     02         WRR                 08
     03         WRR                 08
     04         Priority            01

  WRR bandwidth ratios:    5[queue 1]  25[queue 2]  40[queue 3]
  queue-limit ratios:     20[queue 1]  25[queue 2]  40[queue 3]  15[Pri Queue]

  queue tail-drop-thresholds
  --------------------------
  1     70[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
  2     70[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
  3     100[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]

  queue random-detect-min-thresholds
  ----------------------------------
    1    80[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
    2    80[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
    3    60[1] 70[2] 80[3] 90[4] 100[5] 100[6] 100[7] 100[8]

  queue random-detect-max-thresholds
  ----------------------------------
    1    100[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
    2    100[1] 100[2] 100[3] 100[4] 100[5] 100[6] 100[7] 100[8]
    3    70[1] 80[2] 90[3] 100[4] 100[5] 100[6] 100[7] 100[8]

 WRED disabled queues:
  queue thresh cos-map
  ---------------------------------------
  1     1      1
  1     2
  1     3
  1    4
  1     5
  1     6
  1     7
  1     8
  2     1      0
  2     2
  2     3
  2     4
```

```
    2       5
    2       6
    2       7
    2       8
    3       1       2
    3       2       3
    3       3       6
    3       4       7
    3       5
    3       6
    3       7
    3       8
    4       1       4 5
…
```

WS-6708-10GE and WS-6716-10GE—1P7Q4T Egress Queuing Model

The hardware design of the next-generation 10G linecards are designed with advanced ASICs and higher capacity to ensure the campus backbone of large enterprise networks are ready for the future. Both modules support DSCP based on the eight queue model to deploy flexible and scalable QoS in the campus core. With 8-egress queue support the WS-6708-10G and WS-6716-10G modules increased application granularity based on various DSCP markings done at the network edge. Figure 19 illustrates the DSCP-based 1P7Q4T egress queuing model.

Figure 19 Catalyst 6500-E DSCP-based P7Q4T Egress QoS Model



The following corresponding 1P7Q4T egress queuing configuration must be applied on each member link of MEC.

- Catalyst 6500-E VSS (Distribution and Core)

```
cr23-vss-core(config)#interface range TenGigabitEthernet 1/1/2 - 8 , 2/1/2 - 8
cr23-vss-core(config-if-range)# wrr-queue queue-limit 10 25 10 10 10 10 10
 ! Allocates 10% of the buffers to Q1, 25% to Q2, 10% to Q3, 10% to Q4,
 ! Allocates 10% to Q5, 10% to Q6 and 10% to Q7
cr23-vss-core(config-if-range)# wrr-queue bandwidth 1 25 4 10 10 10 10
 ! Allocates 1% BW to Q1, 25% BW to Q2, 4% BW to Q3, 10% BW to Q4,
 ! Allocates 10% BW to Q5, 10% BW to Q6 and 10% BW to Q7
cr23-vss-core(config-if-range)# priority-queue queue-limit 15
 ! Allocates 15% of the buffers to the PQ

 ! This section enables WRED on Queues 1 through 7
cr23-vss-core(config-if-range)# wrr-queue random-detect 1
 ! Enables WRED on Q1
cr23-vss-core(config-if-range)# wrr-queue random-detect 2
 ! Enables WRED on Q2
```

```
cr23-vss-core(config-if-range)# wrr-queue random-detect 3
 ! Enables WRED on Q3
cr23-vss-core(config-if-range)# wrr-queue random-detect 4
 ! Enables WRED on Q4
cr23-vss-core(config-if-range)# wrr-queue random-detect 5
 ! Enables WRED on Q5
cr23-vss-core(config-if-range)# wrr-queue random-detect 6
 ! Enables WRED on Q6
cr23-vss-core(config-if-range)# wrr-queue random-detect 7
 ! Enables WRED on Q7

 ! This section configures WRED thresholds for Queues 1 through 7
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 1 100 100 100 100
 ! Sets all WRED max thresholds on Q1 to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 1 80 100 100 100
  ! Sets Q1T1 min WRED threshold to 80%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 2 100 100 100 100
 ! Sets all WRED max thresholds on Q2 to 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 2 80 100 100 100
  ! Sets Q2T1 min WRED threshold to 80%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 3 80 90 100 100
 ! Sets WRED max thresholds for Q3T1, Q3T2, Q3T3 to 80%, 90% and 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 3 70 80 90 100
 ! Sets WRED min thresholds for Q3T1, Q3T2, Q3T3 to 70 %, 80% and 90%

cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 4 70 80 90 100
 ! Sets WRED min thresholds for Q4T1, Q4T2, Q4T3 to 70 %, 80% and 90%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 4 80 90 100 100
 ! Sets WRED max thresholds for Q4T1, Q4T2, Q4T3 to 80%, 90% and 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 5 70 80 90 100
 ! Sets WRED min thresholds for Q5T1, Q5T2, Q5T3 to 70 %, 80% and 90%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 5 80 90 100 100
 ! Sets WRED max thresholds for Q5T1, Q5T2, Q5T3 to 80%, 90% and 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 6 70 80 90 100
 ! Sets WRED min thresholds for Q6T1, Q6T2, Q6T3 to 70 %, 80% and 90%
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 6 80 90 100 100
 ! Sets WRED max thresholds for Q6T1, Q6T2, Q6T3 to 80%, 90% and 100%
cr23-vss-core(config-if-range)# wrr-queue random-detect min-threshold 7 60 70 80 90
 ! Sets WRED min thresholds for Q7T1, Q7T2, Q7T3 and Q7T4
 ! to 60%, 70%, 80% and 90%, respectively
cr23-vss-core(config-if-range)# wrr-queue random-detect max-threshold 7 70 80 90 100
 ! Sets WRED max thresholds for Q7T1, Q7T2, Q7T3 and Q7T4
 ! to 70%, 80%, 90% and 100%, respectively


 ! This section configures the DSCP-to-Queue/Threshold mappings
cr23-vss-core(config-if-range)# wrr-queue dscp-map 1 1 8
 ! Maps CS1 (Scavenger) to Q1T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 2 1 0
```

```
 ! Maps DF (Best Effort) to Q2T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 3 1 14
 ! Maps AF13 (Bulk Data-Drop Precedence 3) to Q3T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 3 2 12
 ! Maps AF12 (Bulk Data-Drop Precedence 2) to Q3T2
cr23-vss-core(config-if-range)# wrr-queue dscp-map 3 3 10
 ! Maps AF11 (Bulk Data-Drop Precedence 1) to Q3T3
cr23-vss-core(config-if-range)# wrr-queue dscp-map 4 1 22
 ! Maps AF23 (Transactional Data-Drop Precedence 3) to Q4T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 4 2 20
 ! Maps AF22 (Transactional Data-Drop Precedence 2) to Q4T2
cr23-vss-core(config-if-range)# wrr-queue dscp-map 4 3 18
 ! Maps AF21 (Transactional Data-Drop Precedence 1) to Q4T3
cr23-vss-core(config-if-range)# wrr-queue dscp-map 5 1 30
 ! Maps AF33 (Multimedia Streaming-Drop Precedence 3) to Q5T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 5 2 28
 ! Maps AF32 (Multimedia Streaming-Drop Precedence 2) to Q5T2
cr23-vss-core(config-if-range)# wrr-queue dscp-map 5 3 26
 ! Maps AF31 (Multimedia Streaming-Drop Precedence 1) to Q5T3
cr23-vss-core(config-if-range)# wrr-queue dscp-map 6 1 38
 ! Maps AF43 (Multimedia Conferencing-Drop Precedence 3) to Q6T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 6 2 36
 ! Maps AF42 (Multimedia Conferencing-Drop Precedence 2) to Q6T2
cr23-vss-core(config-if-range)# wrr-queue dscp-map 6 3 34
 ! Maps AF41 (Multimedia Conferencing-Drop Precedence 1) to Q6T3
cr23-vss-core(config-if-range)# wrr-queue dscp-map 7 1 16
 ! Maps CS2 (Network Management) to Q7T1
cr23-vss-core(config-if-range)# wrr-queue dscp-map 7 2 24
 ! Maps CS3 (Signaling) to Q7T2
cr23-vss-core(config-if-range)# wrr-queue dscp-map 7 3 48
 ! Maps CS6 (Internetwork Control) to Q7T3
cr23-vss-core(config-if-range)# wrr-queue dscp-map 7 4 56
 ! Maps CS7 (Network Control) to Q7T4
cr23-vss-core(config-if-range)# priority-queue dscp-map 1 32 40 46
 ! Maps CS4 (Realtime Interactive), CS5 (Broadcast Video),
 ! and EF (VoIP) to the PQ
```

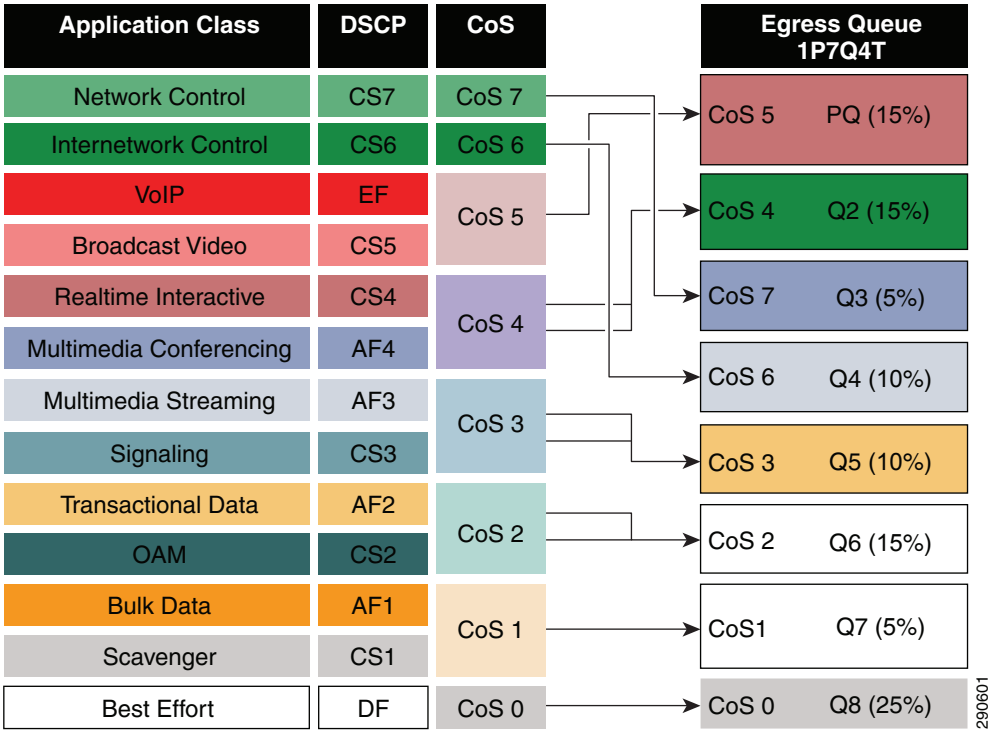
Note Due to the default WRED threshold settings, at times the maximum threshold needs to be configured before the minimum (as is the case on queues one through three in the example above); at other times, the minimum threshold needs to be configured before the maximum (as is the case on queues four through seven in the example above).

Implementing Cisco Nexus 7000 Egress Queuing

The egress queuing function on the Cisco Nexus 7000 is performed on the forwarding engine and port-asic of an egress I/O module. The forwarding engine performs the CoS classification, remarking, and policing whereas the port-asic provides performs the egress scheduling and queuing function. The recommended M108 I/O module supports 1P7Q4T queue support on a per-port level.

Like ingress queuing, the egress queuing function is enabled by default in Cisco Nexus 7000 switches. The system-defined egress class-map and policy-map are by default attached to each physical interface of the system to automatically perform outbound QoS functions. The system-defined default egress queuing policy-map is applied on the low-speed I/O module that supports reduced 1P3Q4T queue capacity. Hence the default egress queue configuration must be fine tuned for the M108 I/O module that supports 1P7Q4T queuing model. The CoS-based egress queue mode is enabled by default, however to align with the recommended 12-class QoS model in a campus design, the network administrator must modify the default CoS-to-Queue mapping table to use a single queue for each class-of-service application traffic. Figure 20 illustrates the recommended 1P7Q4T egress queuing model on the Nexus 7000 system.

Figure 20 Nexus 7000 CoS-Based 1P7Q4T Egress QoS Model



To enable egress QoS on a Cisco Nexus 7000 platform, perform the following steps:

Step 1 The network administrator must first create a custom queuing type policy-map. To utilize each egress queue, the system-defined egress queuing class-map must be used to assign bandwidth, queue limit, and WRED threshold for different class-of-service applications and network data traffic.

```
!Create custom egress queuing policy-map
cr35-N7K-Core1(config)# policy-map type queuing EGRESS-POLICY

!Assign egress bandwidth and queue-limit for PQ-Q1 for CoS=5 traffic
cr35-N7K-Core1(config-pmap-que)#   class type queuing 1p7q4t-out-pq1
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 15
cr35-N7K-Core1(config-pmap-c-que)# priority level 1

!Assign egress bandwidth and queue-limit for Q2 for CoS=4 traffic
    cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q2
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 15
```

```
    cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 20

!Assign egress bandwidth and queue-limit for Q3 for CoS=7 traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q3
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 5
    cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 5

!Assign egress bandwidth and queue-limit for Q4 for CoS=6 traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q4
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 5
    cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 5

!Assign egress bandwidth, queue-limit and WRED for Q5 for CoS=3 traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q5
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 15
cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 20
cr35-N7K-Core1(config-pmap-c-que)# random-detect cos-based
    cr35-N7K-Core1(config-pmap-c-que)# random-detect cos 3 minimum-threshold percent 80
maximum-threshold percent 100

!Assign egress bandwidth, queue-limit and WRED for Q6 for CoS=2 traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q6
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 15
cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 20
cr35-N7K-Core1(config-pmap-c-que)# random-detect cos-based
    cr35-N7K-Core1(config-pmap-c-que)# random-detect cos 2 minimum-threshold percent 80
maximum-threshold percent 100

!Assign egress bandwidth, queue-limit and WRED for Q7 for CoS=1 traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q7
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 5
cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 5
cr35-N7K-Core1(config-pmap-c-que)# random-detect cos-based
    cr35-N7K-Core1(config-pmap-c-que)# random-detect cos 1 minimum-threshold percent 80
maximum-threshold percent 100

!Assign egress bandwidth, queue-limit and WRED for Q8 for CoS=0 Default-class traffic
cr35-N7K-Core1(config-pmap-c-que)# class type queuing 1p7q4t-out-q-default
cr35-N7K-Core1(config-pmap-c-que)# queue-limit percent 25
cr35-N7K-Core1(config-pmap-c-que)# bandwidth remaining percent 25
cr35-N7K-Core1(config-pmap-c-que)# random-detect cos-based
cr35-N7K-Core1(config-pmap-c-que)# random-detect cos 0 minimum-threshold percent 80
maximum-threshold percent 100
```

Step 2 Modify the default egress CoS-to-Queue mapping to align with the recommended CoS setting as illustrated in Figure 20:

```
cr35-N7K-Core1(config)# class-map type queuing match-any 1p7q4t-out-pq1
cr35-N7K-Core1(config-cmap-que)#   match cos 5
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q2
```

```
cr35-N7K-Core1(config-cmap-que)#   match cos 4
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q3
cr35-N7K-Core1(config-cmap-que)#   match cos 7
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q4
cr35-N7K-Core1(config-cmap-que)#   match cos 6
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q5
cr35-N7K-Core1(config-cmap-que)#   match cos 3
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q6
cr35-N7K-Core1(config-cmap-que)#   match cos 2
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q7
cr35-N7K-Core1(config-cmap-que)#   match cos 1
cr35-N7K-Core1(config-cmap-que)# class-map type queuing match-any 1p7q4t-out-q-default
cr35-N7K-Core1(config-cmap-que)#   match cos 0
```

Once the policy-map is configured on the system, the network administrator can apply the egress policy-map on each physical and logical port-channel interface. The policy-map applied on a logical port-channel is automatically applied on each member link interface.

```
cr35-N7K-Core1(config)# interface Ethernet 1/1 , Ethernet 2/1
cr35-N7K-Core1(config-if-range)# service-policy type queuing output EGRESS-POLICY

cr35-N7K-Core1(config)# interface Port-Channel 100 - 103
cr35-N7K-Core1(config-if-range)# service-policy type queuing output EGRESS-POLICY

cr35-N7K-Core1# show policy-map interface brief

 Interface/VLAN [Status]:INP QOS       OUT QOS       INP QUE       OUT QUE
================================================================================
port-channel100 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel101 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel102 [Active]:                           INGRESS-POLIC EGRESS-POLICY
port-channel103 [Active]:                           INGRESS-POLIC EGRESS-POLICY
Ethernet1/1      [Active]:                           INGRESS-POLIC EGRESS-POLICY
<snip>

cr35-N7K-Core1# show queuing interface ethernet 1/1 summary
Egress Queuing for Ethernet1/1 [Interface]
------------------------------------------------
Template: 1P7Q4T
------------------------------------------------
Que# Group Bandwidth% PrioLevel Shape%     CoSMap
------------------------------------------------
   0     -         25         -      -          0
   1     -         20         -      -          4
   2     -          5         -      -          7
   3     -          5         -      -          6
   4     -         20         -      -          3
   5     -         20         -      -          2
   6     -          5         -      -          1
```

```
7       -               -       High        -           5
```

5 Summary

As enterprise customers expand their applications onto their campus networks, ensuring the right amount of bandwidth and prioritization of traffic is essential. This chapter provides an overview of QoS and recommendations and best practices for all switches in the Borderless Campus 1.0 architecture.

4

Deploying High Availability in Campus

The requirement for network reliability and availability is not a new demand, but one that must be well planned for during the early network design phase. To prevent catastrophic network failures and network outages, it is important to identify network fault domains and define rapid recovery plans to minimize application impact during minor and major network outages.

Because every tier of the LAN network design can be classified as a fault domain, deploying a strong campus network foundation with redundant system components and a resilient network design becomes highly effective for non-stop borderless services operation and business communication. However this introduces a new set of challenges, such as higher cost and the added complexity of managing more systems. Network reliability and availability can be simplified using several Cisco high availability technologies that offer complete failure transparency to end users and applications during planned or unplanned network outages.

Cisco high availability technologies can be deployed based on whether platforms have a critical or non-critical role in the network. Some of the high availability techniques can be achieved with the campus network design inherent within the borderless enterprise

network design, without making major network changes. However the critical network systems that are deployed in the main campus that provide global connectivity may require additional hardware and software components to provide uninterrupted communication. The following three major resiliency requirements encompass most of the common types of failure conditions; depending on the LAN design tier, the resiliency option appropriate to the role and network service type must be deployed:

- *Network resiliency*—Provides redundancy during physical link failures, such as fiber cut, bad transceivers, incorrect cabling, and so on.
- *Device resiliency*—Protects the network during abnormal node failure triggered by hardware or software, such as software crashes, a non-responsive supervisor, and so on.
- *Operational resiliency*—Enables resiliency capabilities to the next level, providing complete network availability even during planned network outages using In Service Software Upgrade (ISSU) features.

1 Borderless Campus High-Availability Framework

Independent of the business function, the goal of the network architect should always be to build a strong, scalable, and resilient next-generation IP network. Networks that are built on these three fundamentals provide the high availability necessary to use the network as a core platform that allows you to overlay advanced and emerging technologies as well as provision non-stop network communications. The Borderless Campus network must be built on the same fundamentals, providing highly available network services for uninterrupted business operation, campus security, and the protection of campus physical assets.

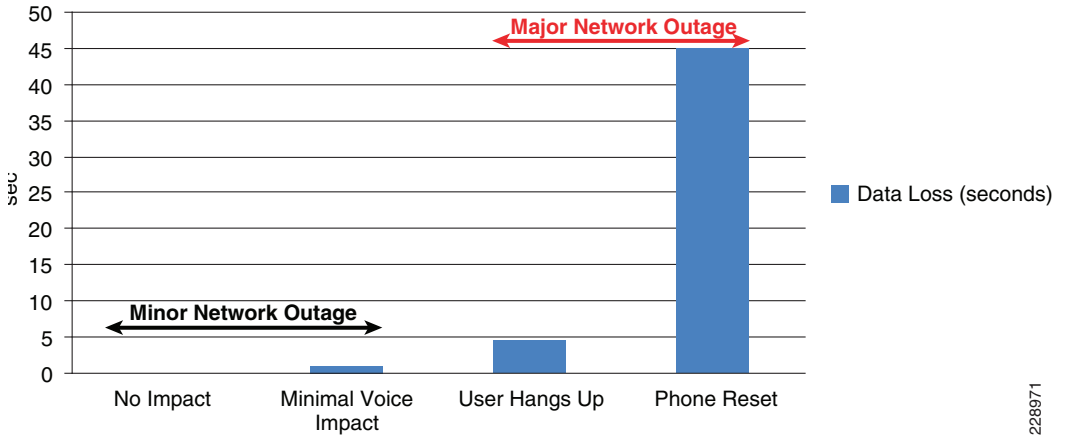
Network fault domains in this reference architecture are identifiable, but the failure conditions within the domains are unpredictable. Improper network design or non-resilient network systems can lead to more faults that not only degrade the user experience, but may severely impact application performance, such as the failure to capture critical physical security video information. The fault levels can range from network interruption to disaster, which can be triggered by the system, humans, or even by nature. Network failures can be classified in two ways:

- *Planned Failure*—A planned network outage occurs when any network system is administratively planned to be “down” for a for scheduled event (software upgrade, etc.).
- *Unplanned Failure*—Any unforeseen failures of network elements can be considered as unplanned failures. Such failures can include internal faults in the network device caused by hardware or software malfunctions, which includes software crashes, linecard or link transceiver failures, etc.

Campus High Availability Baseline

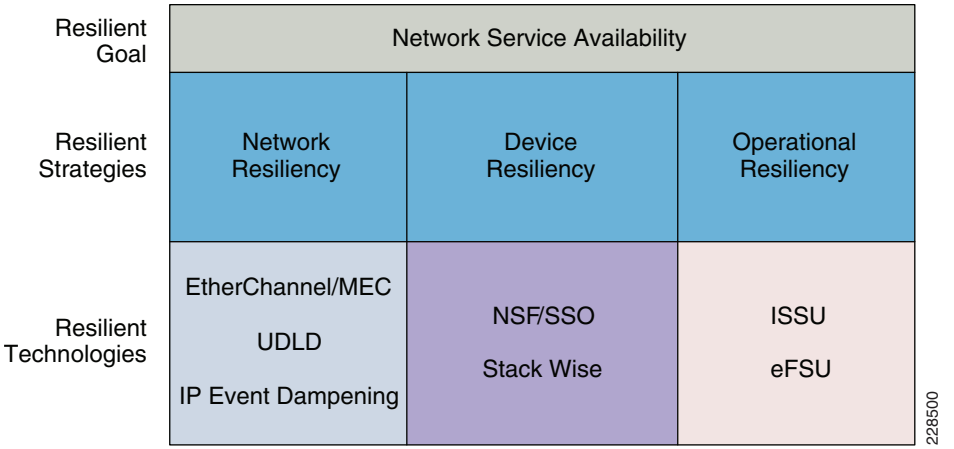
Typical application response time is measured in milliseconds when the campus network is built with high-speed backbone connections and in a fully-operational state. In deterministic network environments, users typically accomplish their work very rapidly. However, during network failures, abnormal traffic loss, congestion, and application retries impact performance and alert the user to a network problem. During major network faults, users determine network connection problem based on routine experience even before an application's protocol mechanism does (e.g,. slow Internet browsing). Protocol-based failure detection and recovery is intentional and is designed to minimize overall productivity impact and allow the network to gracefully adjust and recover during minor failures. While retries for non-critical data traffic may be acceptable, the same level of retries for applications running in real-time may not. Figure 1 illustrates some of the more typical user reactions to varying levels of real-time VoIP application outage, from minor network outages that have no user impact at all to major outages requiring a full device reset.

Figure 1 VoIP Impact During Minor and Major Network Outage



This high availability framework is based on the three major resiliency strategies to effectively mitigate a wide-range of planned and unplanned network outages. Several high availability technologies must be deployed at each layer to provide high network availability and rapid recovery during failure conditions and to prevent communication failure or degraded network-wide application performance (see Figure 2).

Figure 2 High-Availability Goals, Strategy, and Technologies



2 Network Resiliency Overview

The most common network fault occurrence in the LAN network is a link failure between two systems. Link failures can be caused by issues such as a fiber cut, miswiring, linecard module failure, and so on. In the modular platform design, the redundant parallel physical links between distributed modules in two systems reduces fault probabilities and can increase network availability. It is important to remember how multiple parallel paths between two systems also affect how higher layer protocols construct adjacencies and loop-free forwarding topologies.

Deploying redundant parallel paths in the recommended Borderless Campus design by default develops a non-optimal topology that keeps the network underutilized and requires protocol-based network recovery. In the same network design, the routed access model eliminates such limitations and enables full load balancing capabilities to increase bandwidth capacity and minimize application impact during a single path failure. To develop a consistent network resiliency service in the centralized main and remote campus sites, the following basic principles apply:

- Deploying redundant parallel paths is a basic requirement for network resiliency at any tier. It is critical to simplify the control plane and forwarding plane operation by bundling all physical paths into a single logical bundled interface (EtherChannel).
- Implement a defense-in-depth approach to failure detection and recovery. An example of this is configuring the UniDirectional Link Detection (UDLD) protocol, which uses a Layer 2 keep-alive to test that the switch-to-switch links are connected and operating correctly and acts as a backup to the native Layer 1 unidirectional link detection capabilities provided by 802.3z and 802.3ae standards. UDLD is not an EtherChannel function; it operates independently over each individual physical port at Layer 2 and remains transparent to the rest of the port configuration.

- Ensure that the network design is self-stabilizing. Hardware or software errors may cause ports to flap, which creates false alarms and destabilizes the network topology. Implementing route summarization advertises a concise topology view to the network, which prevents core network instability. However, within the summarized boundary, the network may not be protected from flooding. Deploy IP event dampening as an tool to prevent control and forwarding plane impact caused by physical topology instability.

These principles are intended to be a complementary part of the overall structured modular campus design approach and serve primarily to reinforce good resilient design practices.

3 Device Resiliency Overview

Another major component of an overall campus high availability framework is providing device- or node-level protection that can be triggered during any type of abnormal internal hardware or software process within the system. Some of the common internal failures are a software-triggered crash, power outages, line card failures, and so on. LAN network devices can be considered as a single-point-of-failure and are considered to be major failure conditions because recovery may require a network administrator to mitigate the failure and recover the system. The network recovery time can remain undeterministic, causing complete or partial network outage, depending on the network design.

Redundant hardware components for device resiliency vary between fixed configuration and modular Cisco Catalyst switches. To protect against common network faults or resets, all critical Borderless Campus network devices must be deployed with a similar device resiliency configuration. This section provides basic redundant hardware deployment guidelines at the access layer and collapsed core switching platforms in the campus network.

Redundant Power System

Redundant power supplies for network systems protect against power outages, power supply failures, and so on. It is important not only to protect the internal network system, but also the endpoints that rely on power delivery over the Ethernet network. Redundant power systems can be deployed in the following two configuration modes:

- *Modular switch*—Dual power supplies can be deployed in modular switching platforms such as the Cisco Catalyst 6500-E and 4500E Series platforms. Depending on the Cisco Nexus 7000 chassis model, it can be deployed with multiple redundant power supplies, each designed to include two isolated power units. By default, the power supply operates in redundant mode, offering the 1+1 redundant option. In modular Catalyst and Nexus switching systems, the network administrator must perform overall power capacity planning to allow for dynamic network growth with new linecard modules while maintaining power redundancy. Smaller power supplies can be combined to allocate power to all internal and external resources, but may not be able to offer power redundancy.

- *Fixed configuration switch*—Depending on the Catalyst switch, fixed configuration switches offer a wide range of power redundancy options, including the latest innovation, Cisco StackPower, in the Catalyst 3750-X series platform. To prevent network outages on fixed configuration Catalyst switches, they must be deployed with power redundancy:
  - Cisco StackPower technology on 3750-X switches
  - Internal and external redundant power supplies on Catalyst 3560-X switches

  A single Cisco RPS 2300 power supply uses a modular power supply and fan for flexibility and can deliver power to multiple switches. Deploying an internal and external power supply solution protects critical access layer switches during power outages and provides complete fault transparency and constant network availability.

Redundant Control Plane

Device or node resiliency in modular Cisco Catalyst 6500-E, Cisco Nexus 7000, 4500E, and Cisco StackWise Plus platforms provides 1+1 redundancy with enterprise-class high availability and deterministic network recovery time. The following subsections provide high availability design details, as well as graceful network recovery techniques that do not impact the control plane and provide constant forwarding capabilities during failure events.

Stateful Switchover

The stateful switchover (SSO) capability in modular switching platforms such as the Cisco Catalyst 6500-E, Nexus 7000, and 4500E provide complete enterprise-class high availability in the campus network. Cisco recommends the distribution and core layer design model to be the center point of high-availability in the enterprise network. Deploying redundant supervisors in the mission-critical distribution and core system provides non-stop communication throughout the network.

Core/Distribution Layer Redundancy

Increase network- and device-level resiliency by designing the enterprise campus to operate in a deterministic capacity, with network resiliency and the availability of rich, integrated services. The Catalyst 6500-E system running VSS mode must be deployed with a redundant supervisor module in each virtual switch chassis in the aggregation layer and backbone network. In the Cisco best practice campus design, the Cisco 6500-E system provides constant network availability and deterministic recovery with minimal application impact during supervisor switchover.

The system architecture of the Cisco Nexus 7000 system is built to deliver a lossless networking solution in large-scale enterprise campus and data center networks. Decoupling the control plane from the forwarding plane, the supervisor switchover process becomes graceful and hitless in the Cisco

Nexus 7000 system. The resilient hardware and software in the Nexus 7000 architecture is designed to protect campus network capacity and services availability using redundant components—supervisor, I/O, and crossbar fabric modules.

Access Layer Redundancy

Depending on the redundancy capabilities of the access layer system, the campus access layer may become a single-point of failure. To provide 99.999 percent service availability in the access layer, the Catalyst 4500E must be equipped with redundant supervisors to critical endpoints, such as Cisco TelePresence.

Cisco StackWise Plus is a low-cost solution to provide device-level high availability. Cisco StackWise Plus is designed with unique hardware and software capabilities that distribute, synchronize, and protect common forwarding information across all member switches in a stack ring. During master switch failure, the new master switch re-election remains transparent to the network devices and endpoints. Deploying Cisco StackWise Plus according to the recommended guidelines protects against network interruption and recovers the network in less than one second during master switch re-election.

Bundling SSO with NSF capability and the awareness function allows the network to operate without errors during a primary supervisor module failure. Users of realtime applications such as VoIP do not hang up the phone and IP video surveillance cameras do not freeze.

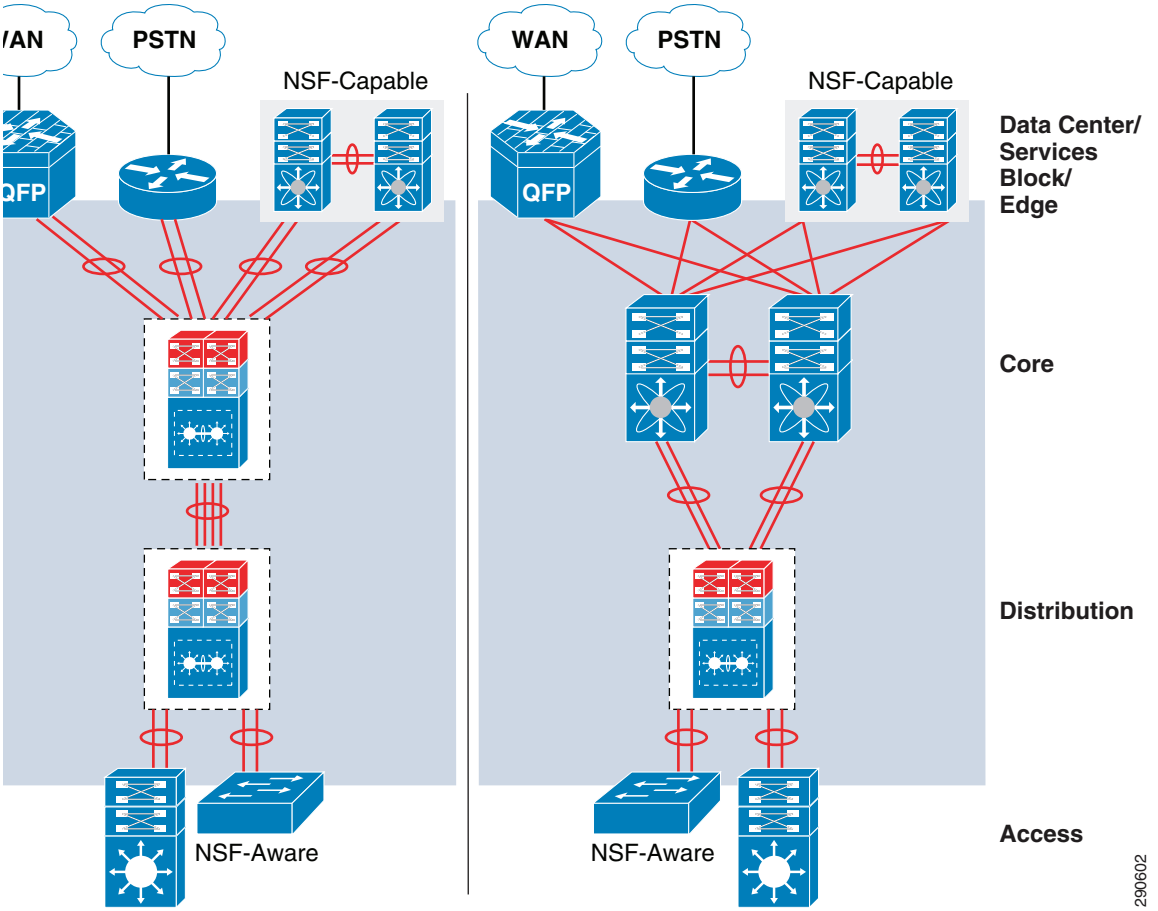
Non-Stop Forwarding

Every borderless campus recommended system deployed in redundant SSO configuration mode provides graceful protocol and network recovery during active supervisor or switch resets. The systems deployed with dual supervisor or route processors are NSF-capable systems that have the capability to initialize graceful protocol recovery with neighbors during the active supervisor or route processor reset. The neighbor system must have the NSF-Aware capability—to support the NSF-capable system to gracefully recover—by protecting routing adjacencies and topology.

It is important to enable the NSF capability for Layer 3 protocols running in a campus network. During the graceful switchover process, the new active supervisor or switch sends graceful recovery signals to neighbors to protect adjacencies and topology reset. Combining SSO with protocol intelligence using NSF technology enables graceful control plane recovery to maintain a bi-directional non-stop forwarding plane for continuous network communication.

As device redundancy is critically important in each campus network tier, the modular Cisco Catalyst and Nexus 7000 systems are designed to support NSF capability for Layer 3 unicast and multicast routing protocols. The non-modular systems, such as the Catalyst 3560-X and Cisco ISR routers, provide network-level redundancy while a SSO-capable neighbor switch is going through the recovery process. (See Figure 1-3.)

Figure 1-3 Borderless Campus NSF/SSO Capable and Aware Systems

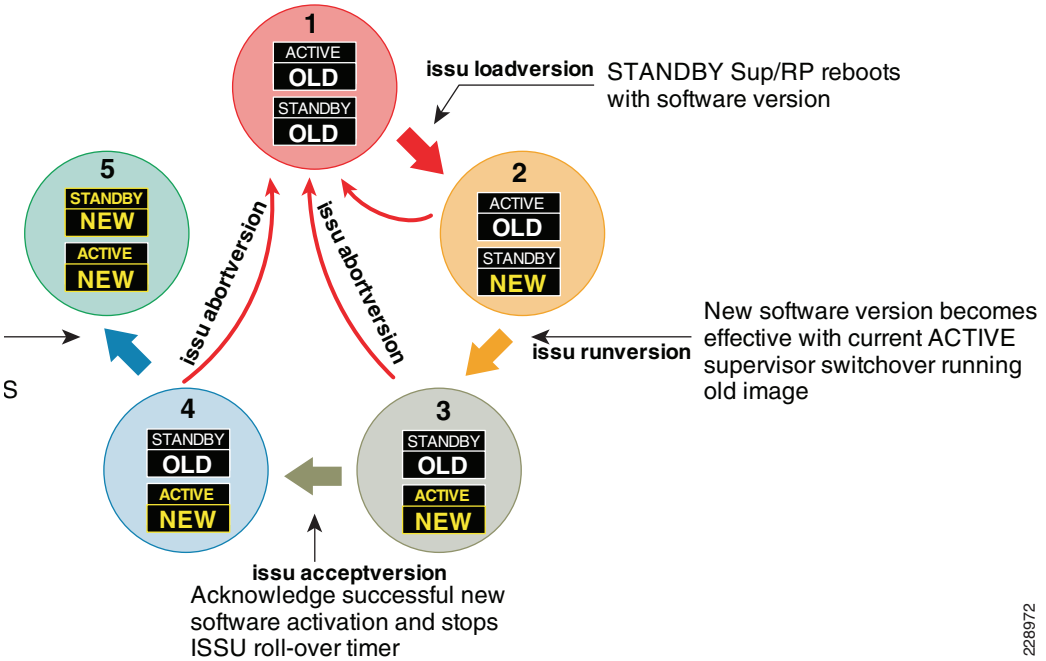


4 Operational Resiliency Overview

Designing the network to recover from failure events is only one aspect of the overall campus non-stop design. Converged network environments are continuing to move toward requiring true 7x24x365 availability. The Borderless Campus network is part of the backbone of the enterprise network and must be designed to enable standard operational processes, configuration changes, and software and hardware upgrades without disrupting network services.

The ability to make changes, upgrade software, and replace or upgrade hardware becomes challenging without a redundant system in the campus core. Upgrading individual devices without taking them out of service is similarly based on having internal component redundancy (such as with power supplies and supervisors) complemented with the system software capabilities. The Cisco Catalyst 6500-E, Nexus 7000, 4507R+E, and ASR 1000 series platforms support real-time software upgrades in the campus without introducing network downtime or impacting network availability. The Cisco In-Service Software Upgrade (ISSU) and Enhanced Fast Software Upgrade (eFSU) leverage NSF/SSO technology to provide continuous network availability while upgrading critical systems. This helps to greatly reduce the need for planned service downtime and maintenance. Figure 1-4 demonstrates the platform-independent Cisco IOS software upgrade flow process using ISSU technology.

Figure 1-4 Cisco IOS ISSU Software Process Cycle



Catalyst 4500E—ISSU

Full-image ISSU on the Cisco Catalyst 4500E leverages dual redundant supervisors to allow for a full, in-service Cisco IOS upgrade, such as moving from IOS Release 12.2(53)SG to 12.2(53)SG1. This leverages the NSF/SSO capabilities and unique uplink port capability to keep ports in an operational

and forwarding state even when supervisor module is reset. This design helps retain bandwidth capacity while upgrading both supervisor (Sup7-E, Sup6-E, or Sup6L-E) modules at the cost of less than a sub-second of traffic loss during a full Cisco IOS upgrade.

Having the ability to operate the campus as a non-stop system depends on the appropriate capabilities being designed into the network from the start. Network and device level redundancy, along with the necessary software control mechanisms, guarantee controlled and fast recovery of all data flows following any network failure, while concurrently providing the ability to proactively manage the infrastructure.

The Catalyst 4500E can perform ISSU with the following two methods:

- Manual—Follow each ISSU process as illustrated in Figure 1-4. The manual IOS upgrade mode is more attentive and requires users to upgrade IOS by manually going through each upgrade cycle. Executing each ISSU upgrade step provides flexibility for users to verify the stability of network operation and services by introducing new IOS software individually, as well as providing an option to abort the upgrade process and roll back to an older IOS version if any abnormal behavior is observed.
- Automatic—Follows the same ISSU upgrade process as illustrated in Figure 1-4. However the automatic upgrade process is the new, single-step automatic IOS-XE upgrade process that automates each ISSU step on the active and standby Sup7-E supervisor modules without user intervention. This simplified upgrade process helps network administrators of large Catalyst 4500E-based campus networks to roll out new Cisco IOS software in the network. The Catalyst 4500E Sup6-E and Sup6L-E supervisor modules currently do not support the automatic upgrade process.

Cisco recommends using both ISSU methods when upgrading the IOS software process on the Cisco Catalyst 4500E Sup7-E module in order to minimize the disruptive impact to network operation and services and upgrade the network rapidly. It is recommended that network administrators first upgrade the Catalyst 4500E Sup7-E system using manual procedures that allow verification of stability at each upgrade step. They should then identify the reliability of the new Cisco IOS version and verify that it is ready to be deployed across the campus network. Later the remainder of the Sup-7E-based systems can be upgraded using the single-step automatic ISSU upgrade procedure.

Catalyst 6500 VSS—eFSU

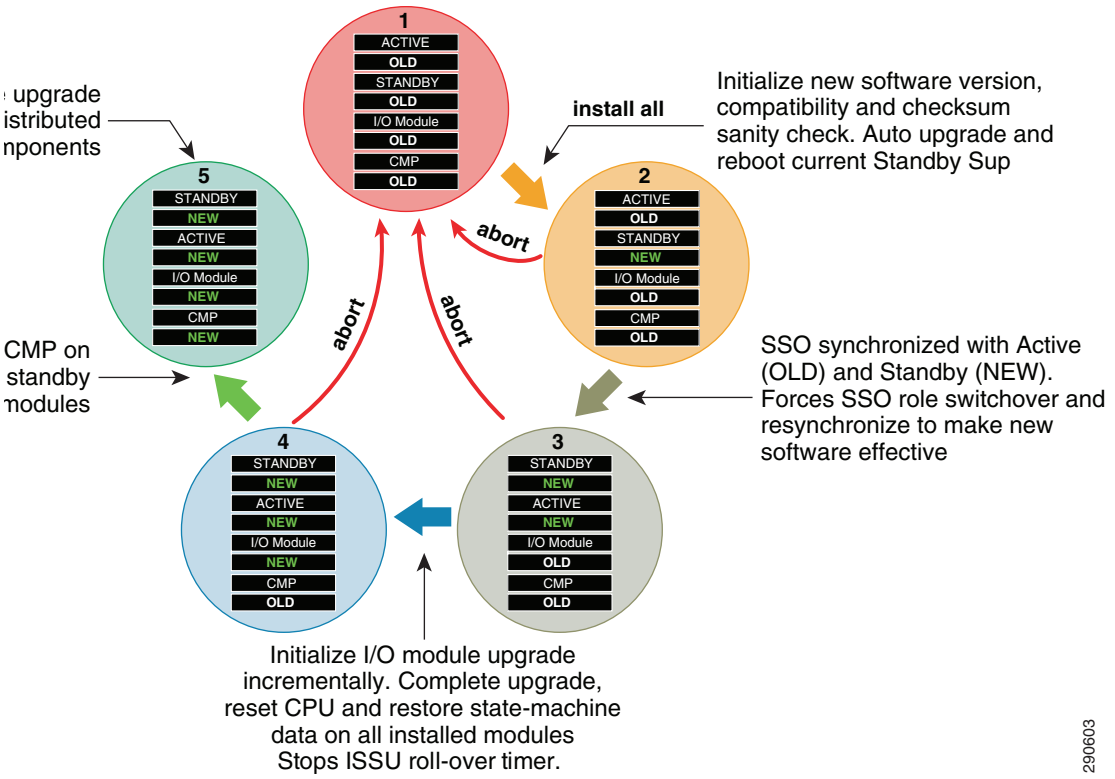
A network upgrade requires planned network and system downtime. VSS offers unmatched network availability to the core. With the Enhanced Fast Software Upgrade (eFSU) feature, VSS can continue to provide network services during an upgrade. With the eFSU feature, the VSS network upgrade remains transparent to applications and end users. Because eFSU works in conjunction with NSF/SSO technology, network devices can gracefully restore control and forwarding information during the upgrade process, while the bandwidth capacity operates at 50 percent and the data plane converges in less than one second.

For a transparent software update, the ISSU process requires three sequential upgrade events on both virtual switch systems. Each upgrade event causes traffic to be re-routed to a redundant MEC path, causing sub-second traffic loss that does not impact realtime network applications, such as VoIP.

Cisco Nexus 7000—ISSU

To provide non-disruptive network services in the campus core, the Nexus 7000 provides a simplified and resilient upgrade procedure. The distributed hardware components require software upgrades with the latest Cisco NX-OS software and protect against control plane disruption, maintaining campus backbone network availability and capacity. During a graceful software upgrade process on a dual-supervisor module, all I/O module and CMP complexes go through a five-step automatic upgrade procedure initiated by a single user step. Each step performs several non-disruptive checks to ensure the Cisco NX-OS upgrade procedure will not introduce any network instabilities. Combined with the resilient Nexus 7000 system architecture, best practice campus network design, and NSF/SSO capability, the Cisco NX-OS software upgrade process results in zero packet loss. Figure 1-5 illustrates the Cisco NX-OS ISSU-based software upgrade process.

Figure 1-5 Cisco NX-OS ISSU Software Process Cycle



5 Design Strategies for Network Survivability

Each network tier can be classified as a fault domain, with the deployment of redundant components and systems increasing redundancy and load sharing capabilities. However, this introduces a new set of challenges— namely, higher costs and increased complexity in managing a greater number of systems. Network reliability and availability can be simplified using several Cisco high-availability and virtual system technologies such as VSS, which offers complete failure transparency to end users and applications during planned or un-planned network outages. In this sense, minor and major network failures are considered broad terms that includes several types of network faults which must be taken into consideration in order to implement a rapid recovery solution.

Cisco high availability technologies can be deployed based on whether platforms have a critical or non-critical role in the network. Some of the high-availability techniques can be achieved in the campus network design without making major network changes; however, the critical network systems that are deployed in the center of the network to provide global connectivity may require additional hardware and software components to offer non-stop communication.

The network survivability strategy can be categorized using three major resiliency requirements that can encompass most of the common types of failure conditions. Depending on the network system tier, role, and network service types, the appropriate resiliency option must be deployed (see).

Table 1-1 Borderless Campus Network High Availability Strategy

Platform	Role	Network Resiliency	Device Resiliency	Operational Efficienc
Catalyst 3560-X	Access	EtherChannel UDLD Dampening	RPS 2300	None. Standalone Sy
Catalyst 3750-X				
Catalyst 3750-X StackWise Plus			Cisco StackPower NSF-Capable and Aware	StackWise Plus
Catalyst 4500E	Access		Redundant Power Supplies Redundant Linecard Modules Redundant Supervisor Modules SSO/NSF Capable and Aware[1]	ISSU
	Distribution			
	Core			
Catalyst 6500-E	Distribution			VSS eFSU
	Core			
Nexus 7000	Core	EtherChannel UDLD	Redundant Power Supplies Redundant Linecard modules Redundant Crossbar Fabric Module Redundant Supervisor modules SSO/NSF Capable and Aware	ISSU

Table 1-1 Borderless Campus Network High Availability Strategy

ASR 1006	WAN Edge	EtherChannel Dampening	Redundant Power Supplies Redundant ESP modules Redundant Route Processors SSO/NSF Capable and Aware	ISSU
ASR 1004	Internet Edge		Red. Power Supplies SSO/NSF Capable and Aware[2]	ISSU

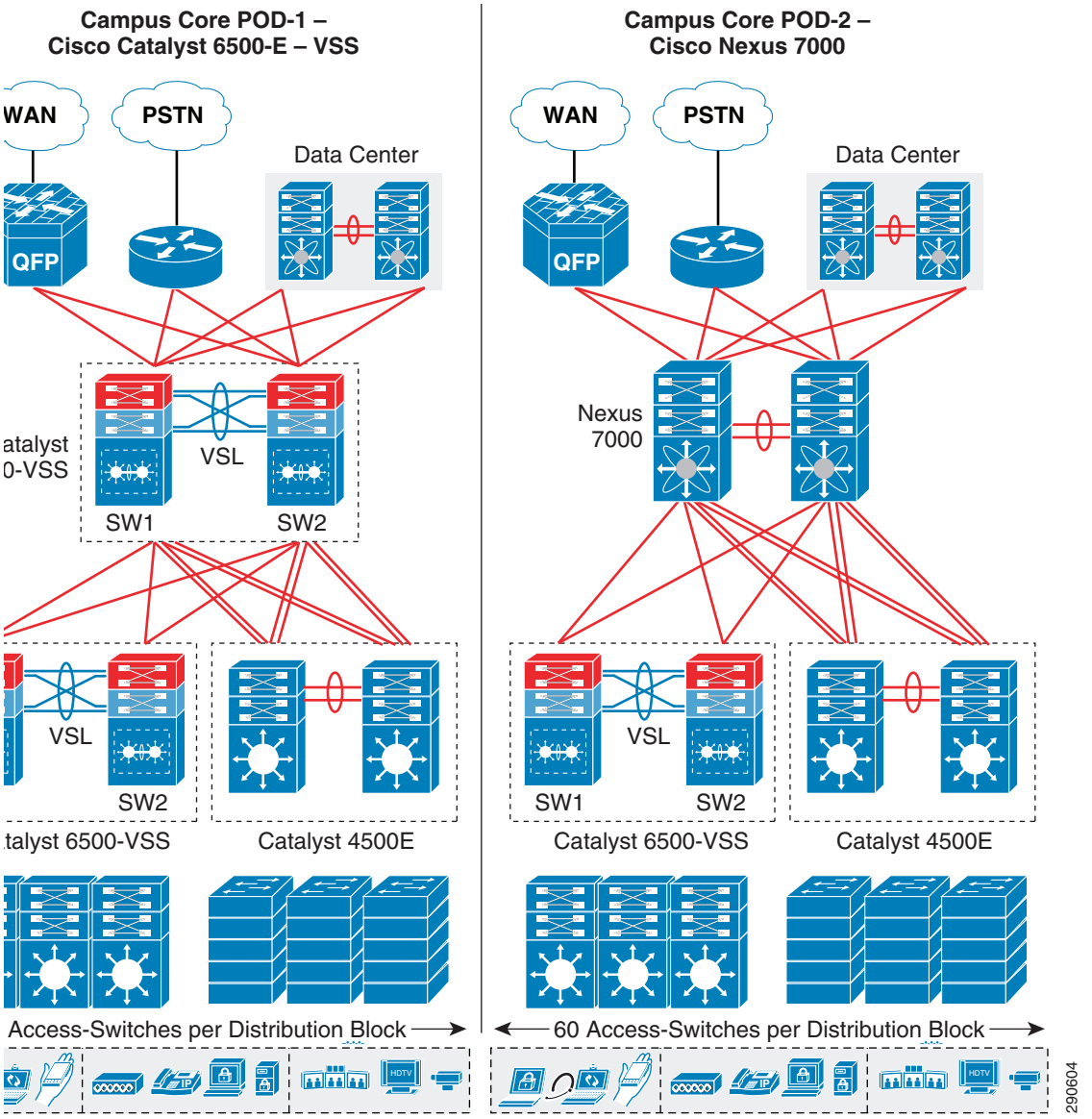
1. Redundant quad supervisor per VSS Domain (two per virtual switch node basis)and dual supervisor module on Catalyst 4500 chassis.
2. Software-based SSO Redundancy.

6 Borderless Campus Design Validation

This design guide validates the operation, integration, and performance of an end-to-end reference borderless campus network architecture. Evaluating network and application performance through thorough and rigorous testing in a Cisco solution lab derives the recommended network design, systems, technologies, and best practices. To align with real-world large enterprise networks and understand the impact to end points and application in an end-to-end, large scale environment, the Cisco solution lab is equipped with a large number of real-time and non-real time devices, such as IP phones, TelePresence units, PCs, laptops, etc.

Previous chapters provided guidance on deploying key foundational technologies and optimizing application performance with QoS techniques. This chapter provides strategies and guidance on building a resilient campus network design. To meet the campus high-availability baseline—enabling real-time applications such as unified and video communication—this document provides validated results obtained by inducing faults on system components and measuring the impact on the network and applications. Figure 1-6 illustrates a sample solution lab network topology for a large campus network design based on a reference architecture.

Figure 1-6 Sample Large Campus Solution Lab Network Topology



To characterize end-to-end application impact during system or network failure, the solution architect collects bi-directional test data to analyze overall application-level impact and the recovery mechanisms in the network. Unicast and multicast data, voice, and video traffic directions are divided into the following categories:

- Unicast upstream—Traffic flows in unique patterns (point-to-point, client-to-one server, client-to-many servers) originated by end points from the access layer and routed towards the data center or a remote medium and small campus over the WAN infrastructure.
- Unicast downstream—Traffic flows originated from data centers by a single or many servers destined to many clients connected at the access layers.
- Multicast downstream—Traffic flows of multicast data and video originated by multicast sources in data centers and sent to many multicast receivers connected at the access layers.

All results described in subsequent sections are validated with the bi-directional traffic patterns described above.

7 Implementing Network Resiliency

The Borderless Campus design guide recommends deploying a mix of hardware and software resiliency designed to address the most common campus LAN network faults and instabilities. It is important to analyze network and application impact using a top-down approach and implement the appropriate high availability solution to create a resilient network. Implementing a resilient hardware and software design maintains the availability of all upper layer network services that are deployed in a Borderless Campus design. This section provides Layer 2 and Layer 3 network design recommendations to build a simplified, flexible, scalable, and resilient multi-tier enterprise campus network. Cisco recommendations and best practices are consistent across different campus network sizes and designs (two-tier versus three-tier).

Campus network stability and reliability are challenged during most common path failures caused by fiber cuts, faulty hardware, or Layer 1 link errors. Such fault conditions de-stabilize the network and result in service disruptions and degraded application performance. Network-level resiliency can be stabilized and service disruptions minimized by suppressing link faults and dampening un-stable network paths by implementing Cisco recommended network resiliency techniques.

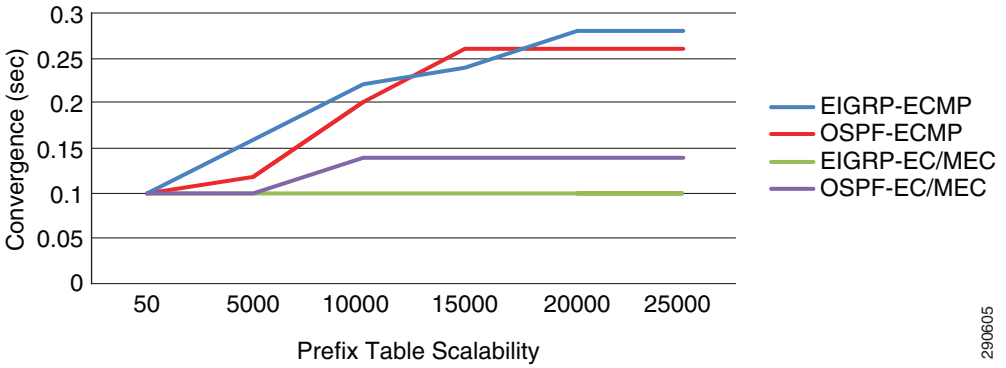
ECMP versus EtherChannel

Chapter 1, "Deploying Network Foundation Services," describes the functional and operational impact of several Layer 2 and Layer 3 network foundation technologies based on ECMP and EtherChannel. The key point to consider in a network design with multiple parallel paths between two system is to simplify the operation with a single logical EtherChannel that builds a concise network routing and switching topology and lets intelligent hardware perform network optimization with parallel forwarding paths that increase network capacity and resiliency.

In the Multilayer campus network design, depending on the aggregation system configuration mode—VSS versus Standalone—the network administrator must deploy EtherChannel/MEC when there are multiple parallel Layer 2 paths between the logical distribution and access layer systems. Bundling Layer 2 paths between two systems offers several architectural and operational benefits (see Chapter 1, "Deploying Network Foundation Services," for more details). If the distribution layer system is deployed in standalone configuration mode, then it may operate in a sub-optimal configuration with two distributed Layer 2 uplinks from the access layer network. Depending on the Layer 2 VLAN design—Flat versus Segmented VLANs in the distribution block—the forwarding path may become asymmetric. Alternatively, the Layer 3 routing boundary can be extended to the wiring closet with a subset routing function to build active/active Layer 3 forwarding paths between the distribution and access layer systems. From a network resiliency perspective, both recommended Layer 2 MEC and Layer 3 routed access designs deliver deterministic sub-second network recovery during link faults.

As next-generation campus systems are evolving with high-performance systems and network virtualization, the redundant and mission-critical enterprise campus distribution and core systems must be simplified to scale, enable borderless network services, improve application quality, and increase user satisfaction. If the distribution and core layer are deployed with the Catalyst 6500-E in VSS mode, then it is highly recommended to build a single unified point-to-point Layer 3 MEC between both campus layer systems. A full-mesh, diversified, and distributed fiber between both virtual switch systems helps increase hardware-driven data load sharing and builds a prefix scale independent campus backbone network. Figure 7 provides evidence of how a well-designed network simplifies and future-proofs network operation and resiliency and delivers consistent, deterministic enterprise-class network recovery independent of prefix scale size in the campus backbone.

Figure 7 6500-E VSS—ECMP versus EC/MEC Link Loss Analysis



EtherChannel technology should be leveraged when the Cisco Nexus 7000 is deployed in the campus core layer. The Nexus 7000 system should be deployed with a single Layer 3 EtherChannel when there are multiple parallel Layer 3 paths with a standalone neighbor device or with distributed Layer 3 paths between logical switches, i.e., VSS or 4500E in redundant mode. Deploying Layer 3 EtherChannel in

the high-scale campus backbone, the Nexus 7000 system is specifically designed to offer the same consistent application performance and user experience as Catalyst 6500-E VSS mode. In a recommended EtherChannel-based campus network design, the Nexus 7000 performs as consistently as the Catalyst 6500-E and delivers network stability and resiliency during path failures.

EtherChannel/Multi-Chassis EtherChannel

In a non-EtherChannel network environment, the network protocol requires fault detection, topology synchronization, and best path recomputation in order to reroute traffic requiring variable timing and to restart the forwarding of traffic. Conversely, EtherChannel or MEC network environments provide significant benefits in such conditions, as the network protocol remains unaware of the topology changes and allows the hardware to self-recover from faults. Re-routing traffic over an alternate member link of EtherChannel or MEC is based on minor internal system EtherChannel hash re-computations instead of an entire network topology re-computation. Hence an EtherChannel and MEC-based network provides deterministic sub-second network recovery of minor to major network faults.

The design and implementation considerations for deploying diverse physical connectivity across redundant standalone systems and virtual systems to create a single point-to-point logical EtherChannel is explained in the Designing the Campus LAN Network in Chapter 1, "Deploying Network Foundation Services."

EtherChannel/MEC Network Recovery Analysis

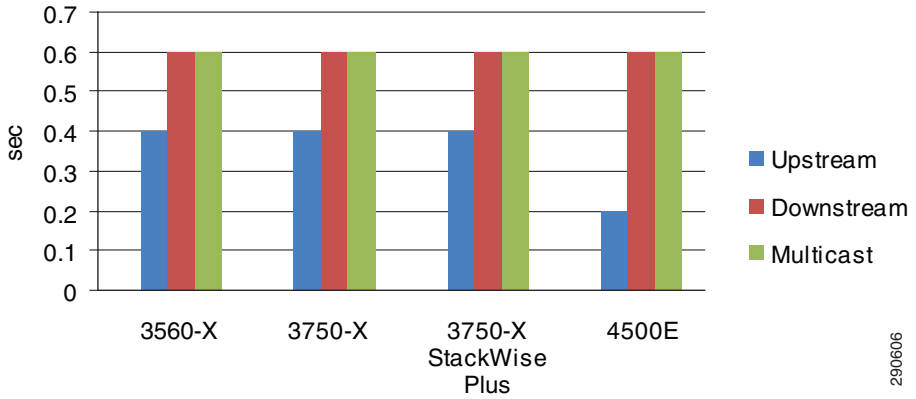
Network recovery with EtherChannel and MEC is platform- and diverse-physical-path-dependent instead of Layer 2 or Layer 3 network protocol dependent. The Borderless Campus design deploys EtherChannel and MEC throughout the network in order to to develop a simplified single point-to-point network topology which does not build any parallel routing paths between any devices at any network tiers.

During individual member link failures, the Layer 2 and Layer 3 protocols dynamically adjust the metrics of the aggregated port channel interfaces. Spanning-Tree updates the port costs and Layer 3 routing protocols like EIGRP update the composite metrics (note that OSPF may change the interface cost). In such events, the metric change will require the generation of minor update messages in the network and will not require end-to-end topology recomputations that impact the overall network recovery process. Since the network topology remains intact during individual link failures, the re-computation to select alternate member links in EtherChannel and MEC becomes locally significant to each impacted EtherChannel neighbor on either end. EtherChannel re-computation requires recreating a new logical hash table and re-programming the hardware to re-route the traffic over the remaining available paths in the bundled interface. The Layer 2 or Layer 3 EtherChannel and MEC re-computation is rapid and independent of network scale.

Catalyst 6500-E VSS MEC Link Recovery Analysis

Several types of network faults can trigger link failures in the network (e.g., fiber pullout, GBIC failure, etc.). Network recovery remains consistent and deterministic in all network fault conditions. In standalone or non-virtual systems using switches such as the Catalyst 3560-X or 4500E, the EtherChannel recomputation is fairly easy as the alternate member link resides within the system. However, with the distributed forwarding architecture in virtual systems like Catalyst 6500-E VSS and Catalyst 3750-X StackWise Plus, extra computation may be required to select alternate member link paths through its inter-chassis backplane interface—VSL or StackRing. Such designs still provide deterministic recovery, but with an additional delay to recompute a new forwarding path through the remote virtual switch node. The link failure analysis chart with inter-chassis reroute in Figure 8 summarizes several types of faults induced in large scale EIGRP and OSPF campus topologies during the development of this Cisco Validated Design guide.

Figure 8 Catalyst 6500-E VSS Inter-Chassis MEC Link Recovery Analysis



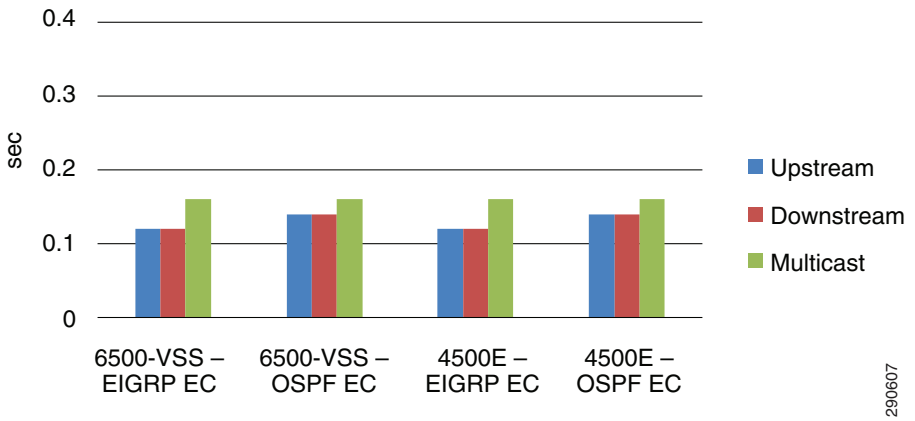
The Borderless Campus network can be designed optimally for deterministic and bidirectional symmetric network recovery for unicast and multicast traffic. For an intra-chassis recovery analysis with the same network faults tested in inter-chassis scenarios, refer to Redundant Linecard Modules.

Nexus 7000 EtherChannel Link Recovery Analysis

As described earlier, designing an EtherChannel-based campus network minimizes routing topology recomputation during individual member link failures. Member link failure in an EtherChannel-based network design suppresses notification to upper layer protocols such as EIGRP, OSPF, and multicast PIM, while the same link fault in an ECMP network design may force a network-wide topology change and could cause forwarding path switchover due to metric adjustments. Based on the best practices in this design guide, the Nexus 7000 maintains next-hop Layer 3 OSPF or EIGRP paths in the URIB/FIB table or the multicast OIF interface in the MRIB/MFIB table during individual member link failures.

With fully-synchronized forwarding information across all system-wide installed I/O modules, the hardware rapidly re-computes the EtherChannel hash and performs data switching based on the new lookup. Even if the new forwarding egress path is within the same I/O module or another I/O module, data plane re-routing within the system across crossbar fabric module remains deterministic and within the campus HA baseline. Figure 9 summarizes several types of campus core layer link faults induced in large-scale EIGRP and OSPF campus core network topologies during the development of this Cisco Validated Design guide.

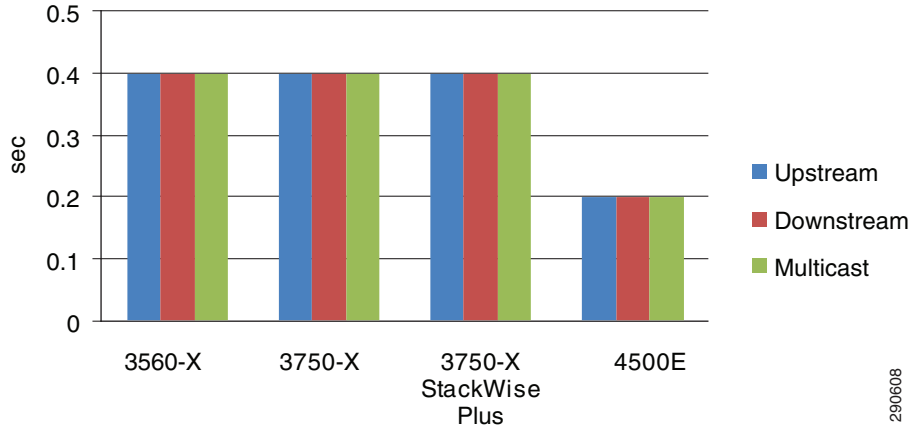
Figure 9 Cisco Nexus 7000 EtherChannel Link Recovery Analysis



Catalyst 4500E EtherChannel Link Recovery Analysis

In the Borderless Campus design, a Catalyst 4500E with redundant hardware components is deployed in the different campus LAN network tiers. A Cisco Catalyst 4500E can only be deployed in standalone mode with in-chassis supervisor and module redundancy. The traffic load balancing and rerouting across different EtherChannel member links occurs within the local chassis. The centralized forwarding architecture in the Catalyst 4500E can rapidly detect link failures and reprogram the hardware with new EtherChannel hash results. The test results in Figure 10 confirm the deterministic and consistent network recovery in large-scale campus topologies running EIGRP and OSPF during individual Layer 2/Layer 3 EtherChannel member link failures.

Figure 10 Catalyst 4500E EtherChannel Link Recovery Analysis



Unidirectional Link Detection (UDLD)

UDLD is a Layer 2 protocol that works with Layer 1 features to determine the physical status of a link. At Layer 1, auto-negotiation takes care of physical signaling and fault detection. UDLD performs tasks that auto-negotiation cannot perform, such as detecting the identity of neighbors and shutting down misconnected ports. When auto-negotiation and UDLD are both enabled, the Layer 1 and Layer 2 detection methods work together to prevent physical and logical unidirectional connections and protocol malfunctions. The UDLD protocol functions transparently on Layer 2 or Layer 3 physical ports. The protocol level, uni-directional communication between two systems should be deployed based on these recommendations:

- Layer 2 Network—In the multilayer standalone or EtherChannel-based network design, the UDLD protocol can be enabled on a per-trunk port level between the access and distribution switches.
- Layer 3 ECMP—In the Layer 3 ECMP-based campus core or in a routed access network design, the uni-directional communication between two systems can be detected by Layer 3 routing protocols as it operates on per-physical interface basis.
- Layer 3 EtherChannel—In a recommended EtherChannel-based network design, the UDLD should be implemented between two Layer 3 systems. Enabling UDLD on each member link of the Layer 3 EtherChannel provides uni-directional path detection at the Layer 2 level.

Copper media ports use Ethernet link pulses as a link monitoring tool and are not susceptible to unidirectional link problems. However, because one-way communication is possible in fiber optic environments, mismatched transmit/receive pairs can cause a link up/up condition even though bidirectional upper layer protocol communication has not been established. When such physical connection errors occur, it can cause loops or traffic black holes. UDLD operates in one of two modes:

- *Normal mode (Recommended)*—If bidirectional UDLD protocol state information times out, it is assumed there is no fault in the network and no further action is taken. The port state for UDLD is marked as undetermined and the port behaves according to its STP state.
- *Aggressive mode*—If bidirectional UDLD protocol state information times out, UDLD attempts to reestablish the state of the port provided it detects that the link on the port is operational. Failure to reestablish communication with UDLD neighbor forces the port into the err-disable state, which either must be manually recovered by the user or the switch if it is be configured for auto-recovery within a specified time interval.

Unidirectional fiber cable anomalies can trigger asymmetric communication and may cause network instability, e.g.., STP loops. Normal UDLD operation detects such faults and prevents network instability by disabling the physical port. The default time to detect the unidirectional links and take action in normal or aggressive mode UDLD may still involve a delay of several seconds in a mission critical campus network. To address this, Cisco has introduced fast UDLD technology that can provide sub-second detection of the fault , thus helping to minimize network impact. Currently fast UDLD is supported on Cisco Catalyst 4500 switches running 12.2(54)SG and 6500 12.2(33)SXI4. Cisco Catalyst 4500E Sup7-E running IOS-XE 3.1.0 SG does not support fast UDLD.

While fast UDLD solves the unidirectional link condition with acceptable delay, it introduces the following challenges for large, redundant campus network designs:

- CPU Utilization—Since the fast UDLD hello packets are processed in milliseconds, it requires heavy CPU interruption. Depending on the number of fast UDLD-enabled links and other software processing network applications, fast UDLD may introduce network and system instability challenges for the network administrator.
- SSO Switchover—This design guide recommends deploying Cisco Catalyst modular platforms with dual supervisor modules on each chassis to provide redundancy. Any Layer 2 or Layer 3 protocols implemented with a sub-second timer may trigger a session timeout and create a false positive alarm, which may result in an entire network outage during a supervisor switchover event. The new active supervisor module in the recovery system cannot restart software processing until several seconds have elapsed. Hence, the peer device initiates a session reset due to not receiving the required keepalives within the time period specified by the timeout parameter.

Note It is recommended to avoid implementing UDLD in aggressive mode as well as fast UDLD on the Cisco Catalyst switches deployed with redundant supervisor modules.

The following illustrates a configuration example to implement the UDLD protocol in normal mode:

Cisco IOS

```
cr22-6500-VSS#config t
cr22-6500-VSS(config)#interface range Ten1/1/8 , Ten2/1/8
cr22-6500-VSS(config-if-range)#udld port

cr22-6500-VSS#show udld neighbors
Te1/1/8    TBM14364802     1            Ethernet1/2     Bidirectional
Te2/1/8    TBM14364802     1            Ethernet2/2     Bidirectional
```

Cisco NX-OS

```
cr35-N7K-Core2(config)# feature udld
!Enable UDLD feature set

cr35-N7K-Core2#show udld neighbors
Port             Device Name    Device ID   Port ID         Neighbor State
-------------------------------------------------------------------------
Ethernet1/2      08E3FFFC4      1           Te1/1/8         bidirectional
Ethernet2/2      08E3FFFC4      1           Te2/1/8         bidirectional
```

IP Event Dampening

Unstable physical network connectivity with poor signaling or loose connections may cause continuous port flaps. When the Borderless Campus network is not deployed using best practice guidelines to summarize the network boundaries at the aggregation layer, a single interface flap can severely impact the stability and availability of the entire campus network. Route summarization is one technique used to isolate the fault domain and contain local network faults within the domain.

To ensure local network domain stability during port flaps, all Layer 3 interfaces can be implemented with IP Event Dampening, which uses the same fundamental principles as BGP dampening. Each time the Layer 3 interface flaps, IP dampening tracks and records the flap event. On multiple flaps, a logical penalty is assigned to the port and it suppresses link status notifications to IP routing until the port becomes stable.

IP Event Dampening is a local specific function and does not have any signaling mechanism to communicate with remote systems. It can be implemented on each individual physical or logical Layer 3 interface—physical ports, SVI, or port-channels:

- Layer 3 Port-Channel

```
cr24-4507e-MB(config)#interface Port-Channel 1
cr24-4507e-MB(config-if)#no switchport
cr24-4507e-MB(config-if)#dampening
```

- Layer 2 Port-Channel

```
cr24-4507e-MB(config)#interface Port-Channel 15
```

```
cr24-4507e-MB(config-if)#switchport
cr24-4507e-MB(config-if)#dampening
```

- SVI Interface

```
cr24-4507e-MB(config)#interface range Vlan101 - 120
cr24-4507e-MB(config-if-range)#dampening

cr24-4507e-MB#show interface dampening
Vlan101
  Flaps Penalty    Supp ReuseTm   HalfL  ReuseV   SuppV  MaxSTm    MaxP Restart
      3       0   FALSE       0       5    1000    2000      20   16000       0
…
TenGigabitEthernet3/1 Connected to cr23-VSS-Core
  Flaps Penalty    Supp ReuseTm   HalfL  ReuseV   SuppV  MaxSTm    MaxP Restart
     10       0   FALSE       0       5    1000    2000      20   16000       0
…
Port-channel1 Connected to cr23-VSS-Core
  Flaps Penalty    Supp ReuseTm   HalfL  ReuseV   SuppV  MaxSTm    MaxP Restart
      3       0   FALSE       0       5    1000    2000      20   16000       0
Port-channel15 Connected to cr24-3560X-MB
  Flaps Penalty    Supp ReuseTm   HalfL  ReuseV   SuppV  MaxSTm    MaxP Restart
      3       0   FALSE       0       5    1000    2000      20   16000       0
```

8 Implementing Device Resiliency

Each device in the borderless enterprise LAN and WAN network design is connected to a critical system or end-point to provide network connectivity and services for business operations. Like network resiliency, device resiliency integrates redundant hardware components and software-based solutions into a single standalone or virtual systems. Depending on the platform architecture of the Cisco router or switch deployed in the campus network design, device redundancy is divided into four major categories—Redundant Power Supplies, Redundant Line cards, Redundant Supervisor/RP, and Non-Stop Forwarding (NSF) with Stateful Switchover (SSO).

Redundant Power

To provide non-stop network communication during power outages, critical network devices must be deployed with redundant power supplies. To maintain network services operation and prevent disruption in any campus network tier, the Cisco Catalyst and Nexus 7000 systems are designed to provide power redundancy during power outages or hardware failure. Deploying redundant power supplies offers 1+1 or N+1 power redundancy against power supply unit or power source failure that helps reduce mean-time-to-repair (MTTR) in the mission critical campus system. In the recommended

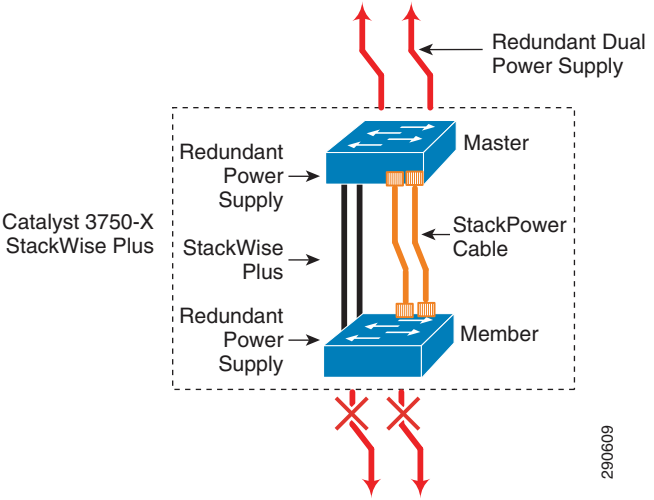
system power redundancy design, campus network communication remains transparent and uninterrupted during power failure, with graceful switchover to redundant power supply units or power input sources.

At the campus access layer, the network administrator must identify the network systems that provide network connectivity and services to mission critical servers. This would also include Layer 1 services such as PoE to boot IP phones and IP video surveillance cameras for campus physical security and communications.

Catalyst 3750-X—Cisco StackPower Redundancy

The next-generation Catalyst 3750-X Series platform introduces innovative Cisco StackPower technology to provide power redundancy solutions for fixed configuration switches. Cisco StackPower unifies the individual power supplies installed in the switches and creates a pool of power, directing that power where it is needed. Up to four switches can be configured in a StackPower stack with the special Cisco proprietary StackPower cable. The StackPower cable is different than the StackWise data cables and is available on all Cisco Catalyst 3750-X models. See Figure 11.

Figure 11 Cisco StackPower Redundancy



A stack member switch experiencing a power fault with its own power supply can derive power from the global power pool so as to provide seamless, continued operation in the network. With the modular power supply design in Catalyst 3750-X Series platform, the defective power supply can be swapped out without disrupting network operation. Cisco StackPower technology can be deployed in two modes:

- *Sharing mode*—All input power is available to be used for power loads. The total aggregated available power in all switches in the power stack (up to four) is treated as a single large power supply. All switches in the stack can provide this shared power to all powered devices connected to PoE ports. In this mode, the total available power is used for power budgeting decisions without any power reserved to accommodate power supply failures. If a power supply fails, powered devices and switches could be shut down. This is the default mode of operation.
- *Redundant mode*—The power from the largest power supply in the system is subtracted from the power budget and held in reserve. This reduces the total power available to PoE devices, but provides backup power in case of a power supply failure. Although there is less available power in the pool for switches and powered devices to draw upon, the possibility of having to shut down switches or powered devices in case of a power failure or extreme power load is reduced. It is recommended to budget the required power and deploy each Catalyst 3750-X switch in the stack with dual power supplies to meet demand. Enabling redundant mode offers power redundancy as a backup should one of the power supply units fail.

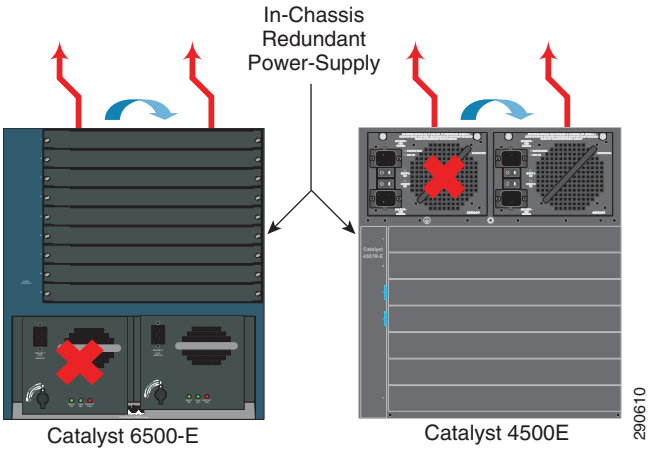
Since Cisco StackWise Plus can group up to nine 3750-X Series switches in the stack ring, Cisco StackPower must be deployed with two power stack groups in order to accommodate up to four switches. The following sample configuration demonstrates deploying Cisco StackPower in redundancy mode and grouping the stack members into power stack groups. To make the new power configuration effective, it is important that network administrator plan for network downtime as all the switches in the stack ring must be reloaded:

```
cr36-3750X-xSB(config)#stack-power stack PowerStack
cr36-3750X-xSB(config-stackpower)#mode redundant

cr36-3750X-xSB(config)#stack-power switch 1
cr36-3750X-xSB(config-switch-stackpower)#stack-id PowerStack
%The change may not take effect until the entire data stack is reloaded

cr36-3750X-xSB(config)#stack-power switch 2
cr36-3750X-xSB(config-switch-stackpower)#stack-id PowerStack
%The change may not take effect until the entire data stack is reloaded
```

Catalyst 4500E and 6500-E (In-Chassis Power Redundancy)

The Cisco Catalyst 4500E and 6500-E Series modular platforms allocate power to several internal hardware components, such as linecards, fans, etc., and externally powered devices, such as IP phones, wireless access points, etc. All of the power is allocated from the internal power supply. With a dual power supply unit hardware design, the Catalyst 6500-E and 4500E systems provide the flexibility to expand the use of power supplies as the network grows. Like linecard module hardware design, power supplies are hot-swappable and implementing 1+1 power redundancy provides network services resiliency while replacing the faulty unit.

Figure 12 Catalyst 4500E and 6500-E Power Redundancy



Dual power supplies in these systems can operate in two different modes:

- *Redundant Mode* (**Recommended**)—By default, power supplies operate in redundant mode offering a 1+1 redundant option. The system determines power capacity and the number of power supplies required based on the allocated power to all internal and external power components. Both power supplies must have sufficient power to provide power to all the installed modules in order to operate in 1+1 redundant mode.

```
cr24-4507e-LB(config)#power redundancy-mode redundant

cr24-4507e-LB#show power supplies
Power supplies needed by system    :1
Power supplies currently available :2

cr22-vss-core(config)#power redundancy-mode redundant switch 1
cr22-vss-core(config)#power redundancy-mode redundant switch 2

cr2-6500-vss#show power switch 1 | inc Switch|mode
Switch Number: 1
system power redundancy mode = redundant

cr2-6500-vss#show power switch 2 | inc Switch|mode
Switch Number: 2
system power redundancy mode = redundant
```

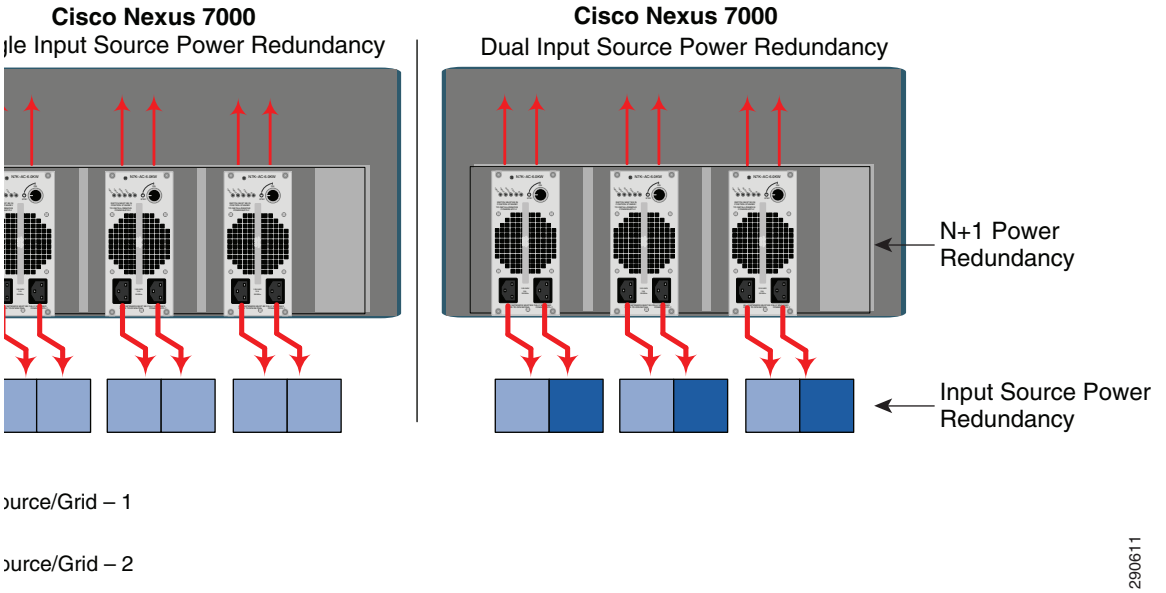
- *Combined mode*—If the system power requirement exceeds the capacity of a single power supply, then the network administrator can utilize both power supplies in combined mode to increase overall capacity. However it may not offer 1+1 power redundancy during a primary power supply failure. The following global configuration enables power redundancy operation in combined mode:

```
cr24-4507e-LB(config)#power redundancy-mode combined

cr24-4507-LB#show power supplies
Power supplies needed by system:2
Power supplies currently available:2
```

Cisco Nexus 7000 (In-Chassis Power Redundancy)

The Cisco Nexus 7000 system can be protected by three internally redundant power supplies with two internal isolated power units that provide up to six active power paths in a fully redundant configuration. Several hardware components, such as supervisor, I/O modules, fan, and crossbar fabric module, consume power from the total aggregated power wattage. All active power supplies use a proportional load sharing method for power distribution to each hardware component that allows efficient use of dissimilar capacity power supplies in the same system. The Cisco Nexus 7000 offers power redundancy to the system in two power source environments—Single Input and Dual Input. The single input source power provides N+1 power unit redundancy, while the dual input source power provides system power protection in multi-failure conditions—power source or grid and power unit failure.

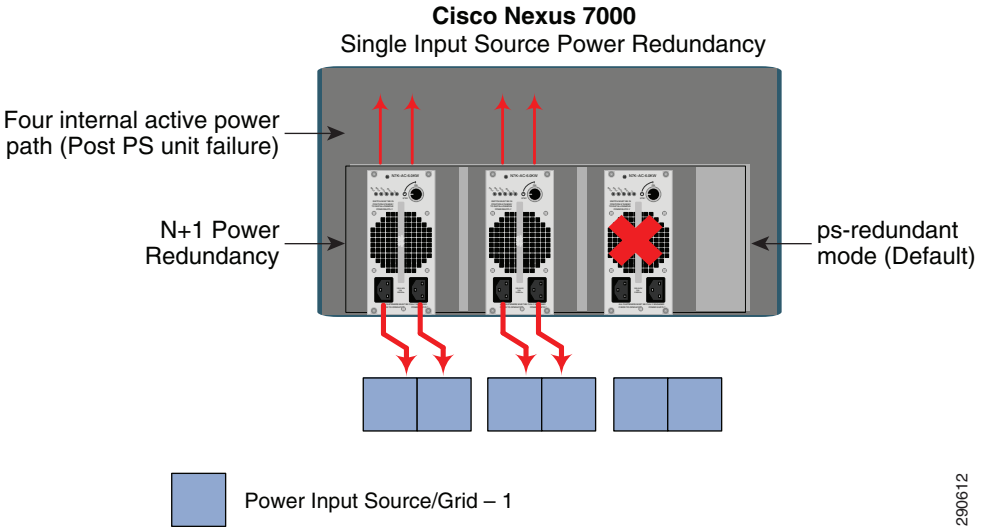
Figure 1-13 Cisco Nexus 7000 Power Redundancy



Implementing redundant power subsystem allows all three units to be configured in the following redundancy modes.

- *PS Redundant mode* (**Recommended**)—By default, deploying redundant power supply units provides N+1 power supply unit redundancy. This redundant mode provides protection against a single power supply unit failure where all power sources are distributed through a single power grid. The cumulative available power to distribute between components is the sum of all installed power supplies minus that of the largest (for redundancy). During single power supply failure, loads are redistributed using the available capacity across the remaining functional power supply units. N+1 power redundancy becomes available with two or three power supplies installed in the system. In a single power circuit/grid environment, the default PS-redundant mode is recommended for N+1 power supply unit redundancy.

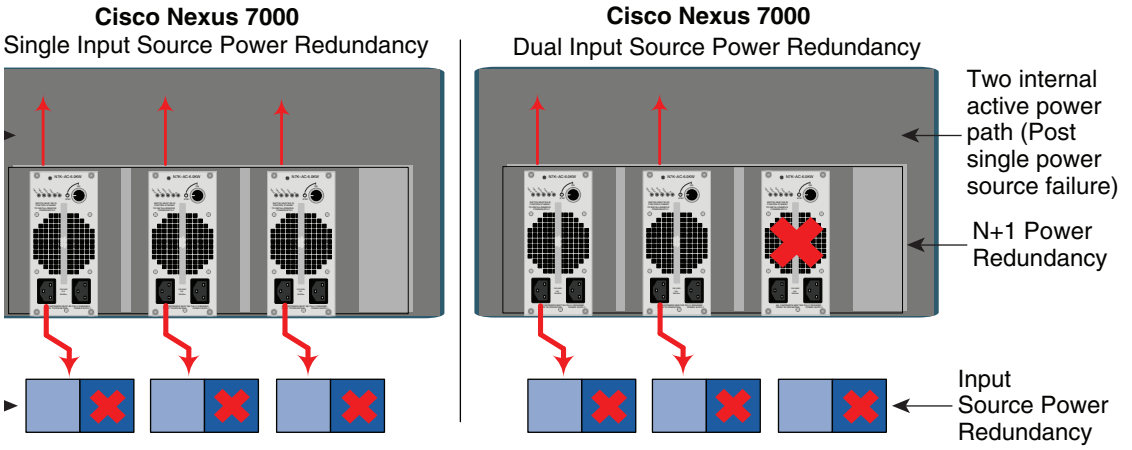
Figure 14 Recommended Single Input Source Power Redundancy




```
cr35-N7K-Core2#show environment power detail | inc Ok|redundancy
1         N7K-AC-6.0KW                515 W       6000 W      Ok
2         N7K-AC-6.0KW                443 W       6000 W      Ok
3         N7K-AC-6.0KW                525 W       6000 W      Ok
Power Supply redundancy mode (configured)                     PS-Redundant
Power Supply redundancy mode (operational)                    PS-Redundant
```

- *Input Source Redundant mode*—In dual power grid network designs, the Nexus 7000 provides the ability to increase power protection against input source failures. To implement input grid power redundancy in the system, each power supply unit must be connected in a distributed model between two independent power sources (grids). During single power grid failure, the cumulative power capacity reduces to half, however with an alternate power source, the remaining half internal power paths remain operational. This mode does not provide power redundancy during individual power unit failure.
- *Redundant mode* (**Recommended**)—The redundant power mode provides multi-failure power protection. Implementing redundant mode provides power protection to the system during power input source (grid) failure and power supply unit failure. This mode provides increased level power redundancy to the Nexus 7000 system by logically combining the N+1 (PS-redundancy) and input grid (*Input Source Redundant*) modes. Each of the power supply redundancy modes imposes different power budgeting and allocation models, which in turn deliver varying usable power yields and capacities. In a dual power input source environment, it is recommended to implement redundancy mode in the Nexus 7000 system.

Figure 1-15 Recommended Dual Input Source Power Redundancy




```
cr35-N7K-Core2(config)# power redundancy-mode redundant

cr35-N7K-Core2# show environment power detail | inc Ok|redundancy
1         N7K-AC-6.0KW                519 W       6000 W      Ok
2         N7K-AC-6.0KW                438 W       6000 W      Ok
3         N7K-AC-6.0KW                521 W       6000 W      Ok
Power Supply redundancy mode (configured)                   Redundant
Power Supply redundancy mode (operational)                  Redundant
```

- *Combined mode*—The cumulative available power watts can be combined with all installed power supplies to provide the sum of all available power to the usable power budget. The combined mode does not provide power redundancy. In this mode the power failure or the unit failure degrades available power to the system. Based on the number of installed hardware components, if power draw is exceeded after failure, it may cause I/O module power down, which may severely impact network services availability and campus backbone capacity. This mode may become an un-reliable power design for power protection during source or unit failure and may introduce network instability or complete outage.

Network Recovery Analysis with Power Redundancy

Each campus LAN router and switch providing critical network services must be powered with either an in-chassis or external redundant power supply system. This best practice is also applicable to the standalone or virtual system devices. Each physical Catalyst 6500-E chassis in VSS mode at the campus distribution and core layer must be deployed with a redundant in-chassis power supply. The Nexus 7000 system at the mission critical core layer must be deployed with three redundant power supply units. Depending on the number of power input sources, the network administrate must implement Cisco recommended power redundancy techniques. The Catalyst 3750-X StackWise Plus must be deployed following the same rule, with the master and member switches in the stack ring deployed using the external redundant power system. Powering virtual systems with redundant power supplies prevents a reduction in network bandwidth capacity, topology changes, and poor application performance in the event of a power failure event.

Several power failures on redundant power systems were conducted during the production of this Cisco Validated Eesign in order to characterize overall network and application impact. Several test cases performed on all redundant power campus systems confirms zero-packet loss during individual power supply failures. Note that the network administrator must analyze the required power capacity that will be drawn by different hardware components (e.g., network modules, PoE+ etc.).

Redundant Linecard Modules

Modular Catalyst platforms support a wide range of linecards for connectivity to the network core and edge. The high-speed core linecards are equipped with special hardware components to build the campus backbone, whereas the network edge linecards are developed with more intelligence and application awareness. Using internal system protocols, each line card communicates with the centralized control plane processing supervisor module through the internal backplane. Any type of internal communication failure or protocol malfunction may disrupt communication between the linecard and the supervisor, which may lead to the linecard and all the physical ports associated with it forcibly resetting to resynchronize with the supervisor.

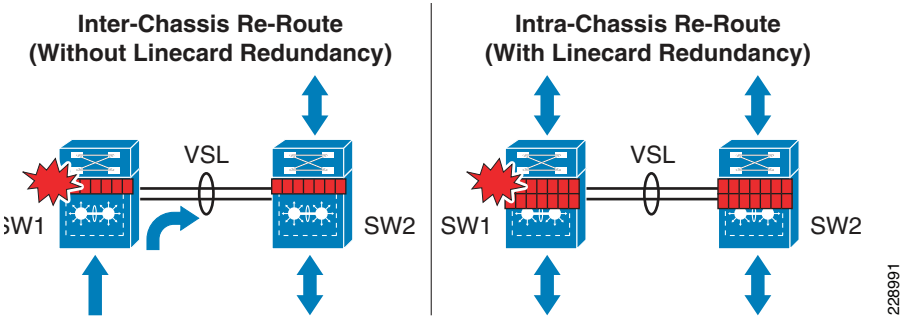
Catalyst 6500-E Linecard Module Recovery Analysis

When the distribution and core layer 6500-E systems are deployed with multiple redundant line cards, the network administrator must design the network by diversifying the physical cables across multiple linecard modules. A full-mesh, diversified fiber design between two virtual switching systems and linecard modules minimizes service disruption and prevents network congestion. The distributed forwarding architecture in hardware is fully synchronized on each DFC-based linecard deployed in the virtual switch. In a steady network state, this software design minimizes data routing across system critical VSL paths. Data traffic traverses the VSL links as a “last-resort” in hardware if either of the virtual switch chassis loses a local member link from the MEC link due to a fiber cut or a major fault

condition like a linecard failure. The impact on traffic could be in the sub-second to seconds range and it may create congestion on the VSL Etherchannel link if the rerouted traffic exceeds overall VSL bandwidth capacity.

Deploying redundant linecards and diversifying paths across the modules prevents inter-chassis re-route, which may cause network congestion if there is not sufficient VSL bandwidth to accommodate the rerouted traffic. Figure 16 demonstrates inter-chassis reroute (without linecard redundancy) and intra-chassis re-route (with linecard redundancy).

Figure 16 Intra-Chassis versus Inter-Chassis Traffic Re-route

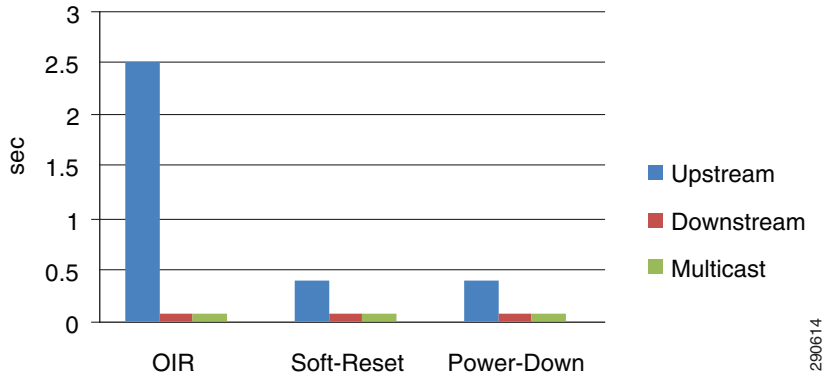


Implementing distributed and diversified fibers between modules mitigates VSL congestion problems. To minimize the service disruption and increase network recovery, the network administrator can follow Cisco recommended best practices to swap or replace module in production network. Removing linecard modules from the modular system while it is in service requires several system internal checks to detect the removal and update distributed forwarding information across all operational modules in the 6500E chassis. This process may take seconds to restore traffic through alternative forwarding paths. To minimize the downtime and restore the service within sub-seconds, Cisco recommends to first disable the linecard from the service and then remove it from the system. The linecard can be put out-of-service in two recommended ways:

- Soft-Reset—Issuing **hw-module switch <1|2> module** <#> from exec mode is a graceful module reset from a software and hardware forwarding perspective, which helps minimize traffic losses bi-directionally. With MEC it also helps minimize control plane changes that trigger topology computation or re-routing. The traffic remains operational through alternate modules and distributed without going through an inter-switch VSL path.
- Power-Down—Disabling power allocation to the network module produces the same impact to the system and network as a soft-reset. The key difference in this procedure is that the module in the specified slot will remain powered down until a new module is installed or power is re-allocated. Power allocation to a module can be disabled using the **no power enable switch <1|2> module** <#> command from global configuration mode.

Both recommended procedures provide graceful network recovery during the linecard removal process. Figure 17 provides an analysis of linecard OIR, soft-reset, and power-down.

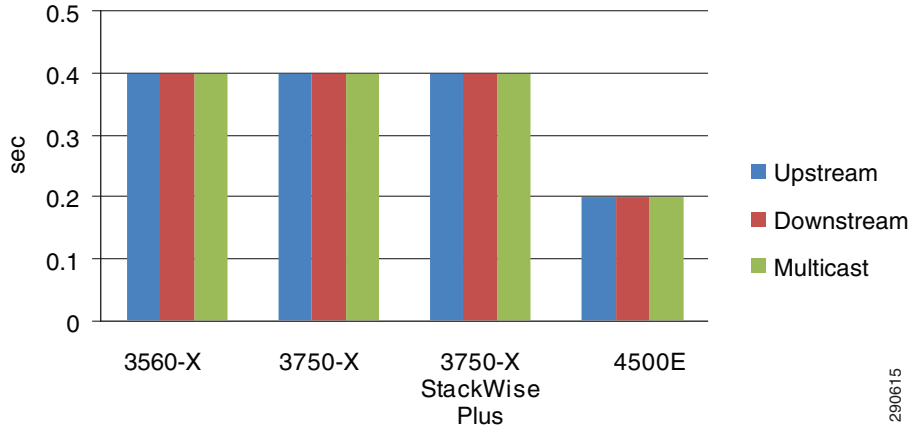
Figure 17 6500E VSS Linecard Recovery Analysis



Catalyst 4500E Linecard Module Recovery Analysis

The centralized forwarding architecture in a Catalyst 4500E programs all the forwarding information on the active and standby supervisor Sup7-E, Sup6-E, or Sup6L-E modules. All the redundant linecards in the chassis are stub and maintain low-level information to handle ingress and egress forwarding information. During a link or linecard module failure, new forwarding information gets rapidly reprogrammed on both supervisors in the chassis. However, deploying EtherChannel utilizing diversified fibers across different linecard modules provides consistent sub-second network recovery during abnormal failure or the removal of a linecard from the Catalyst 4500E chassis. The chart in Figure 18 provides test results associated with removing a linecard from the Catalyst 4500E chassis deployed in various campus network roles.

Figure 18 Catalyst 4500E Distribution Layer Linecard Recovery Analysis

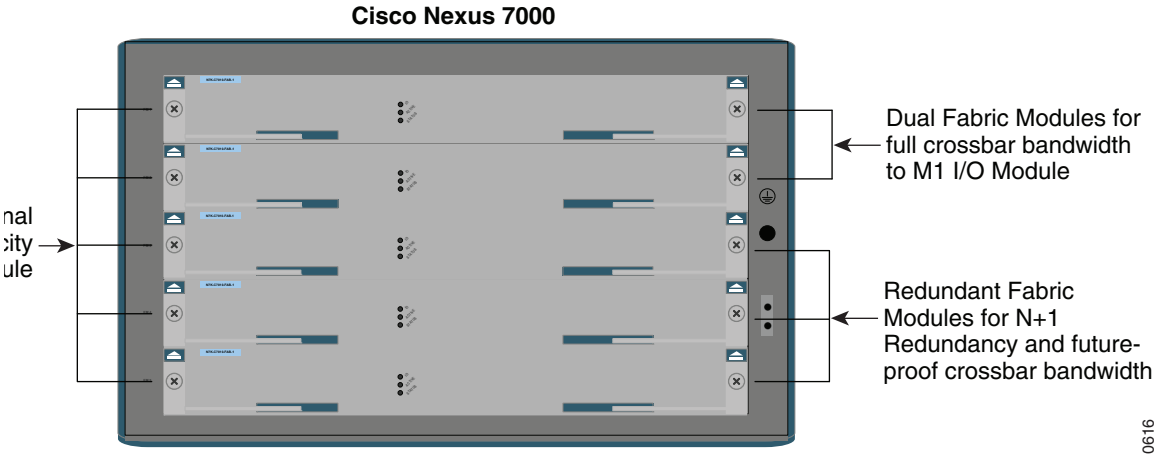


Redundant Nexus 7000 Crossbar Fabric Module

The distributed forwarding architecture in the Nexus 7000 system is built upon an intelligent hardware and software design that decouples the centralized control plane operation from the supervisor module. Each Nexus 7000 I/O module is designed with a distributed forwarding engine that builds and maintains hardware-based forwarding information based on global unicast and multicast RIB. The ingress and egress data switching between ports performs local switching with the I/O module without a centralized lookup procedure or backplane bandwidth involvement to (**??with??**) local switch traffic.

Data traffic switching between different I/O modules is performed through high-speed crossbar modules. The switch fabric capacity per I/O module is determined based on the I/O module's internal throughput capacity and the number of crossbar fabric modules installed in the system. Deploying at least two crossbar fabric modules enables the campus core recommended M108 I/O module to operate at its full 80 Gbps capacity. However during abnormal fabric module failure, the system may introduce backplane congestion due to a lack of sufficient switching capacity from a single crossbar module. It is highly recommended to deploy at least three fabric modules to provide module and backplane capacity redundancy. Cisco recommends deploying additional crossbar fabric modules in the system to future proof the switch fabric bandwidth and increase N+1 fabric module redundancy during abnormal module failure.

Figure 1-19 Cisco Nexus 7000 Crossbar Fabric Bandwidth and Module Redundancy



Crossbar Fabric Bandwidth and Module Redundancy

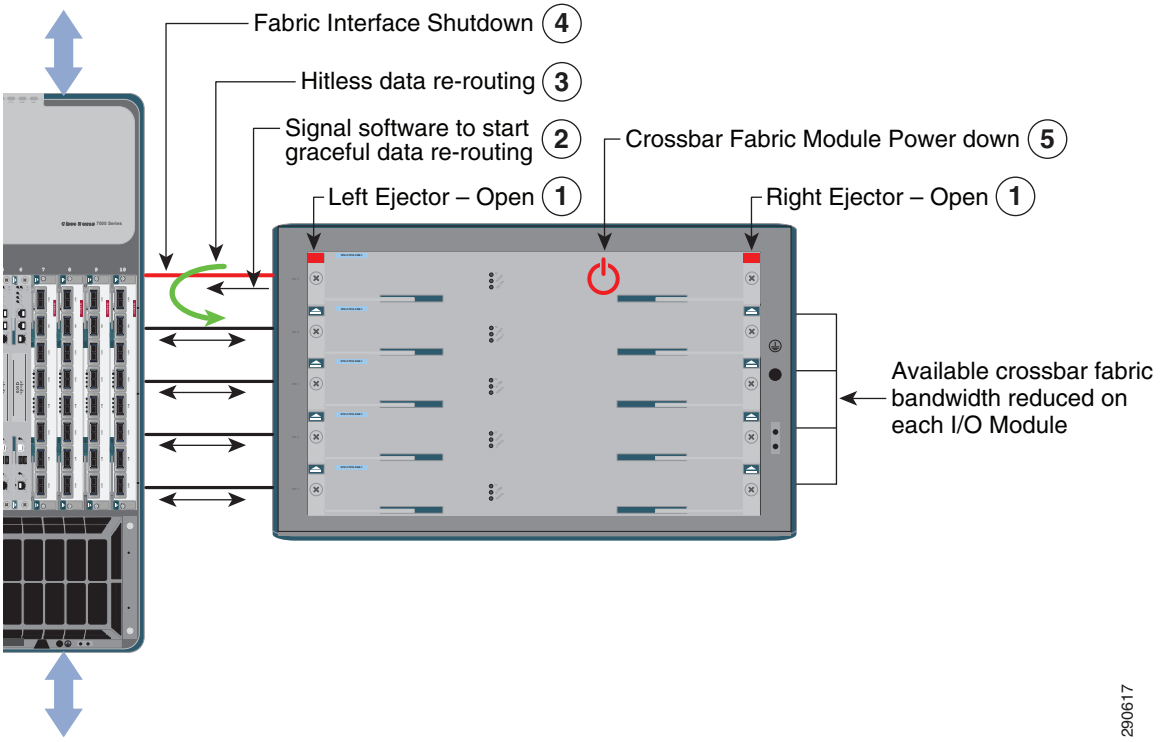
The crossbar fabric modules are hot-swappable. The new fabric module can be inserted in the system without any service disruption or introducing any maintenance window in a production campus network. To swap the fabric module in the Nexus 7000 system, the network administrator must press dual ejector buttons to open and release the internal lock prior to removing the module from operation. The crossbar fabric module get internally shutdown when both ejector buttons are opened; the module remains in an operational state even when one button is open and the other is closed.

```
!Fabric Module remains in operational state with single ejector button in OPEN state
%PLATFORM-3-EJECTOR_STAT_CHANGED: Ejectors' status in slot 13 has changed, Left Ejector is
CLOSE, Right Ejector is OPEN

cr35-N7K-Core2#show module xbar 3 | inc Fabric|Left
3    0      Fabric Module 1                  N7K-C7010-FAB-1    ok
Left ejector OPEN, Right ejector CLOSE, Module HW does support ejector based shutdown.

!Fabric Module status with both ejector button open
%PLATFORM-3-EJECTOR_STAT_CHANGED: Ejectors' status in slot 13 has changed, Left Ejector is
OPEN, Right Ejector is OPEN
%PLATFORM-2-XBAR_REMOVE: Xbar 3 removed (Serial number JAF1442AHKB)

cr35-N7K-Core2# show module | inc Fabric
1    0      Fabric Module 1                  N7K-C7010-FAB-1    ok
2    0      Fabric Module 1                  N7K-C7010-FAB-1    ok
```

To provide hitless fabric module switchover, the hardware sensors are capable of transmitting a signal to the software system for graceful fabric module shutdown when both ejector buttons are opened. This intelligent and highly-available design first gracefully re-routes the data plane to an alternate fabric module prior to internally powering down the fabric module. This graceful OIR proces remains

transparent to the centralized control plane running on a supervisor module and it does not trigger any change in network operation. Figure 1-20 illustrates the step-by-step internal graceful data plane recovery procedure with crossbar fabric module redundancy.

Figure 1-20 Hitless Crossbar Fabric Module Redundancy




```
!System state prior crossbar fabric module failure
cr35-N7K-Core2# show module xbar | inc Fabric
1    0       Fabric Module 1                  N7K-C7010-FAB-1     ok
2    0       Fabric Module 1                  N7K-C7010-FAB-1     ok
3    0       Fabric Module 1                  N7K-C7010-FAB-1     ok
cr35-N7K-Core2# show hardware fabric-utilization
-------------------------------------------------
Slot        Total Fabric        Utilization
            Bandwidth      Ingress % Egress %
-------------------------------------------------
1             138 Gbps           0.0       0.0
2             138 Gbps           0.0       0.0
<snip>
```

```
!System state post crossbar fabric module failure
%PLATFORM-3-EJECTOR_STAT_CHANGED: Ejectors' status in slot 13 has changed, Left Ejector is
OPEN, Right Ejector is OPEN
%PLATFORM-2-XBAR_PWRFAIL_EJECTORS_OPEN: Both ejectors open, Xbar 3 will not be powered up
cr35-N7K-Core2#show module | inc Fabric
1    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
2    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
3    0      Fabric Module                   N/A                powered-dn

!46Gbps Fabric Bandwidth reduced on each I/O module
cr35-N7K-Core2#show hardware fabric-utilization
-----------------------------------------------
Slot        Total Fabric        Utilization
            Bandwidth      Ingress % Egress %
-----------------------------------------------
1              92 Gbps          0.0       0.0
2              92 Gbps          0.0       0.0
<snip>
```

Insufficient Fabric Bandwidth

The high-speed I/O modules may operate under capacity with a non-redundant, single operational crossbar fabric module in a Nexus 7000 system. The 10Gbps M108 I/O module operates at 80 Gbps per slot. With a single operational crossbar fabric, the I/O module remains in an operational state, however backplane switching gets reduced to 46 Gbps. Due to insufficient backplane bandwidth, it may not handle wire-speed campus backbone traffic and may create backplane congestion in the critical campus core.

```
!Steady system state with redundant crossbar fabric modules
cr35-N7K-Core2# show module | inc Fabric
1    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
2    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
3    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
cr35-N7K-Core2# show hardware fabric-utilization
-----------------------------------------------
Slot        Total Fabric        Utilization
            Bandwidth      Ingress % Egress %
-----------------------------------------------
1             138 Gbps          0.0       0.0
2             138 Gbps          0.0       0.0
<snip>

!System state post two crossbar fabric module failure
cr35-N7K-Core2# show module | inc Fabric
1    0      Fabric Module 1                 N7K-C7010-FAB-1    ok
2    0      Fabric Module                   N/A                powered-dn
3    0      Fabric Module                   N/A                powered-dn
```

```
%XBAR-2-XBAR_INSUFFICIENT_XBAR_BANDWIDTH: Module in slot 1 has insufficient
xbar-bandwidth.
%XBAR-2-XBAR_INSUFFICIENT_XBAR_BANDWIDTH: Module in slot 2 has insufficient
xbar-bandwidth.

! Insufficient 46Gbps Fabric Bandwidth for 80Gbps per slot I/O module
cr35-N7K-Core2# show hardware fabric-utilization
-----------------------------------------------
Slot        Total Fabric        Utilization
            Bandwidth      Ingress % Egress %
-----------------------------------------------
1               46 Gbps           0.0       0.0
2               46 Gbps           0.0       0.0
<snip>
```

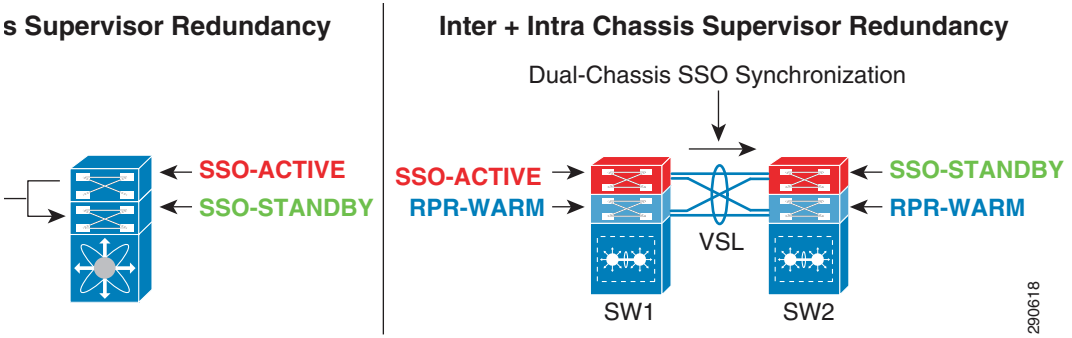
Redundant Supervisor

The enterprise-class modular Cisco Catalyst and Nexus 7000 system support dual-redundant supervisor modules to prevent borderless services disruption due to network control plane and topology resets in the event of supervisor module failure or a forced reset. Deploying redundant supervisor modules in mission critical campus access, distribution, and core layer systems protects network availability and bandwidth capacity during an active supervisor switchover process. Based on the system architecture, the primary supervisor synchronizes all required hardware and software state machines, forwarding information to a secondary supervisor module for seamless operation. The graceful SSO redundancy provides transparent and graceful network recovery that leverages the NSF capability to protect the forwarding plane with completely hitless network recovery. The supervisor redundancy architecture in the recommended modular systems depends on the system hardware design and implemented mode.

- Intra-Chassis Supervisor Redundancy—This mode provides redundancy between two supervisor modules deployed within a single chassis. Depending on the campus system and deployed mode, intra-chassis supervisor redundancy can be in two redundancy modes:
  - SSO—The standalone Nexus 7000 and Catalyst 4500E borderless campus systems provide intra-chassis SSO redundancy. This mode provides single chassis supervisor protection by synchronizing state machines from the active supervisor module to the standby supervisor deployed within the same chassis. The Catalyst 6500-E deployed in standalone mode provides the same intra-chassis SSO redundancy as this system.
  - RPR-WARM—The Catalyst 6500-E deployed in VSS mode is designed to provide inter-chassis redundancy. Deploying each virtual switch with a redundant supervisor module leverages the same set of hardware and supervisor modules to provide quadrupled supervisor redundancy. The intra-chassis supervisor provides virtual switch redundancy if the primary supervisor self-recovery fails.

- Inter-Chassis Supervisor Redundancy—The Cisco VSS innovation with the next-generation Sup720-10GE supervisor module extends the single-chassis SSO supervisor redundancy capability between two separate physical chassis deployed in the same campus network layer. By extending internal backplane communication between two supervisors modules over VSL links, the VSS becomes a single, large, logical, and redundant system to build a unified campus network system. The centralized control plane running on the active supervisor module deployed in the virtual switch, i.e., Switch-1, performs the same SSO synchronization task with the standby supervisor deployed in the remote virtual-switch, i.e., Switch-2.

Figure 1-21 Intra-Chassis versus Inter-Chassis Supervisor Redundancy



Intra-Chassis Supervisor Redundancy

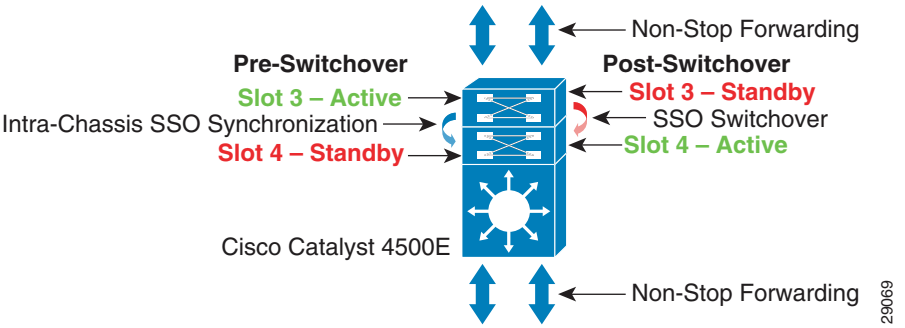
The intra- or single-chassis provides 1+1 supervisor redundancy in the Nexus 7000 system and the Catalyst 4500E switch provides continuous network availability across all the installed modules while the supervisor module is going through the graceful recovery process. Even these systems are modular and provide intra-chassis supervisor redundancy. The hardware and software operation is different when the system is in a steady operational state or going through the switchover process.

Catalyst 4500E

The Catalyst 4500E series platform is modular with a simple hardware and software design. The Catalyst 4500E system is designed for a high-density access layer with end-point-aware intelligence to enable several rich borderless network services at the edge. The active supervisor module holds the ownership of the control and management plane to build the centralized forwarding plane by communicating with end points and upstream network devices. The high-speed linecards are non-distributed and rely on the supervisor for all intelligent forwarding decisions and applying network policies, such as QoS, ACL, etc.

The Catalyst 4500E deployed with a redundant supervisor in SSO configuration dynamically synchronizes the system configuration, network protocol state machines, forwarding information, and more in real-time from the active to the standby supervisor module. During an administrator or software forced supervisor switchover, the Layer 3 network protocols gracefully recover with neighboring systems, however the system maintains its overall network capacity. In the event of supervisor switchover, the uplink ports from both supervisors and linecard modules remains fully operational and in a forwarding state to protect switching capacity and provide continuous non-disruptive borderless services.

Figure 22 Cisco Catalyst 4500E Supervisor Redundancy



Nexus 7000

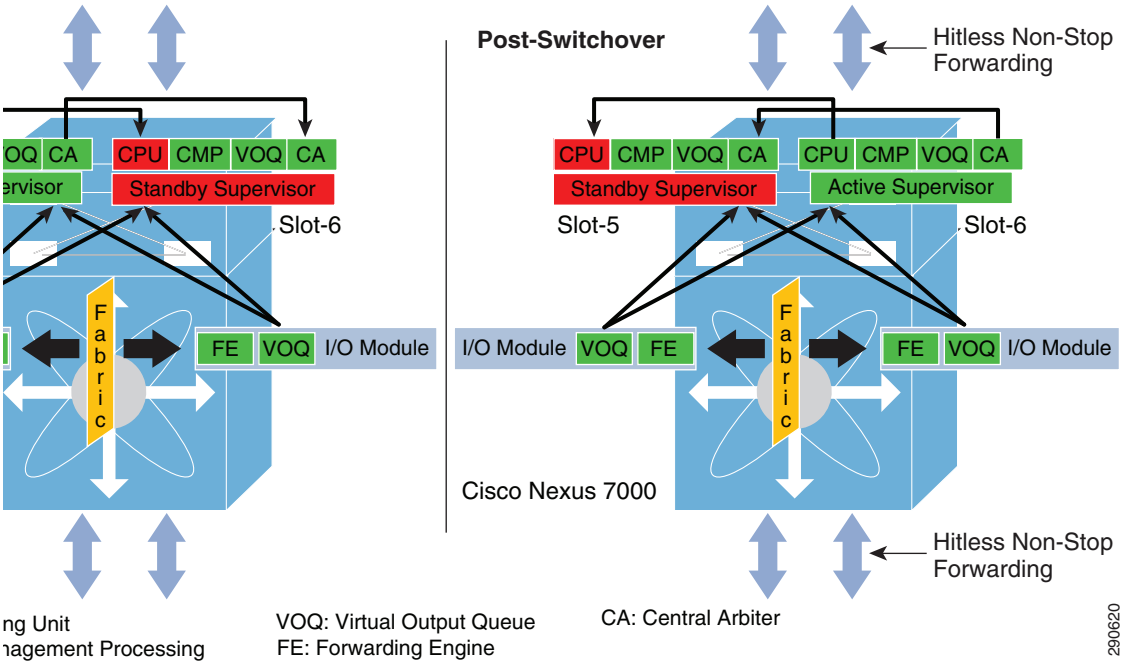
With the increased number of borderless network services and applications running in today's enterprise, the campus network foundation is becoming a core platform to enable a broad ange of digital communication in various emerging forms. As the demands continue to expand at a rapid pace, the backbone network of a next-generation campus network demands a new multi-terabit carrier-class system architecture that scales network capacity several fold. The Nexus 7000 system is a core-class system specifically designed with hardware and software to enable high performance in next-generation campus and data center networks.

As described previously, the Nexus 7000 system decouples the control, management, and data plane within the system. The active supervisor builds the routing adjacencies and forwarding information that gets dynamically updated on each I/O module designed with a distributed forwarding architecture. The system configuration, network protocol state machines, and active supervisor are constantly synchronized to the standby supervisor module for graceful switchover. The distributed forwarding information from the supervisor is stored in a forwarding engine on an I/O module to maintain a local copy of unicast and multicast forwarding information for rapid egress port or module lookup without supervisor involvement. The forwarding engine also provides distributed services like QoS, ACL, Netflow, etc. to optimize throughput and improve application performance with a rapid lookup and forwarding decision process. The multi-stage crossbar fabric module enables backplane

communication between I/O modules. The I/O modules access to the switch fabric is based on VoQ buffer requests and a granting process that involves a central arbiter operating in an active/active state on both supervisor modules.

With a fully-distributed forwarding information and decoupled crossbar switch fabric module design, the active supervisor module switchover remains completely hitless and transparent to other hardware components in the Nexus 7000 system. To provide hitless forwarding, the crossbar fabric module remains operational and the distributed I/O module maintains local forwarding information to seamlessly switch data traffic while the standby supervisor goes through the recovery process.

Figure 1-23 Cisco Nexus 7000 Supervisor Redundancy



Inter- and Intra-Chassis Supervisor Redundancy

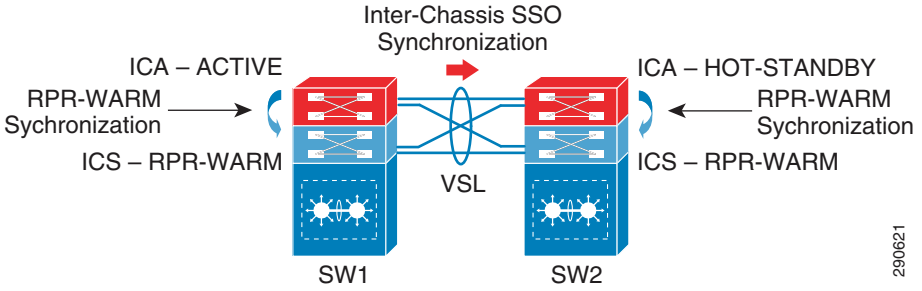
The Cisco VSS solution extends supervisor redundancy by synchronizing SSO and all system internal communication over the special VSL EtherChannel interface between the paired virtual systems. Note that VSS does not currently support stateful intra-chassis supervisor redundancy on each individual virtual node. The virtual switch node running in the active supervisor mode is forced to reset during the switchover. This may disrupt the network topology if it is not deployed with the best practices defined in this design guide. The “triangle”-shaped, distributed, full-mesh fiber paths combined with

single point-to-point EtherChannel or MEC links play a vital role during such network events. During the failure, the new active virtual switch node performs a Layer 3 protocol graceful recovery with its neighbors in order to provide constant network availability over the local interfaces.

6500-E VSS Intra-Chassis RPR-WARM Redundancy

As described earlier the Cisco Catalyst 6500-E introduces innovations aimed at providing intra-chassis stateless supervisor redundancy with the quad-supervisor in the VSS domain. These new innovations allow each redundant supervisor module in each virtual switch chassis to operate in a hybrid role—Supervisor in RPR mode and Distributed Line card. With this hybrid redundancy role, Cisco VSS deployed in quad-sup design operates in a dual redundancy mode—Inter-Chassis SSO with remote virtual-switch chassis and Intra-Chassis RPR within the virtual switch chassis as illustrated in Figure 24.

Figure 24 VSS Quad-Sup Synchronization Process



The ICA supervisor from each virtual-switch chassis synchronizes all critical configurations to the local ICS supervisor module to provide transparent switchover. Even with the stateless intra-chassis redundancy implementation, Cisco VSS offers the ability to maintain full system virtualization, up to date network configuration and protocol state information between both virtual switch chassis. The SSO communication and the synchronization process between ICA supervisors in each virtual switch chassis remains transparent and independent of RPR-WARM. Cisco VSS RPR-WARM provides intra-chassis or local redundancy options, hence it synchronizes the following set of system critical parameters between ICA and ICS supervisor modules:

- Startup-Configuration—Saving the configuration in NVRAM forces the ICA supervisor modules to synchronize their startup configuration with the local in-chassis ICS supervisor module. As part of the SSO synchronization process, the running configuration is synchronized with the remote STANDBY supervisor module in order to maintain an up-to-date configuration.
- BOOT Variables—The boot parameters and registers defined by network administrators are stored as boot parameters in the ROMMON on all four supervisor modules. This synchronization process helps all supervisor modules have consistent bootup information in order to maintain quad-sup redundancy.

- VSS Switch ID—The ICA supervisor module automatically synchronizes the virtual switch ID from its own ROMMON setting to the local ICS ROMMON. Automatically synchronizing the virtual switch ID provides these benefits:
  - Ease of deployment of the in-chassis redundant supervisor module without any additional configuration to synchronize with existing ICA supervisor in the virtual switch.
  - Ability to quickly swap Sup720-10GE module with previous VSS configuration. The ICA supervisor module rewrites old the VSS switch ID to align with its own ID.
- VLAN Database—All VLAN database information is fully synchronized between the ICA and ICS supervisor module.

Deploying Cisco VSS with quad-sup is a plug-n-play operation and no extra configuration is required to enable RPR-WARM. All the intra-chassis ICA and ICS role negotiation and configuration synchronization occurs without any additional settings. The following sample **show** command depicts the SSO synchronized state between the ICA supervisors of SW1 and SW2, which are also running full Cisco IOS software. The in-chassis redundant supervisor modules have been initialized with special Sup720-LC IOS software that enables the hybrid role capability to synchronize RPR-WARM with the local ICA module:

```
cr22-6500-LB#show switch virtual redundancy | inc Switch|Software

My Switch Id = 1
Peer Switch Id = 2

Switch 1 Slot 5 Processor Information :
Current Software state =
Image Version = Cisco IOS Software, s72033_rp Software (s72033_rp-ADVENTERPRISEK9_WAN-M),
Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)

Switch 1 Slot 6 Processor Information :
Current Software state = RPR-Warm
Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M), Version
12.2(33)SXI4, RELEASE SOFTWARE (fc3)

Switch 2 Slot 5 Processor Information :
Current Software state = STANDBY HOT (switchover target)
Image Version = Cisco IOS Software, s72033_rp Software (s72033_rp-ADVENTERPRISEK9_WAN-M),
Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)

Switch 2 Slot 6 Processor Information :
Current Software state = RPR-Warm
Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M), Version
12.2(33)SXI4, RELEASE SOFTWARE (fc3)
```

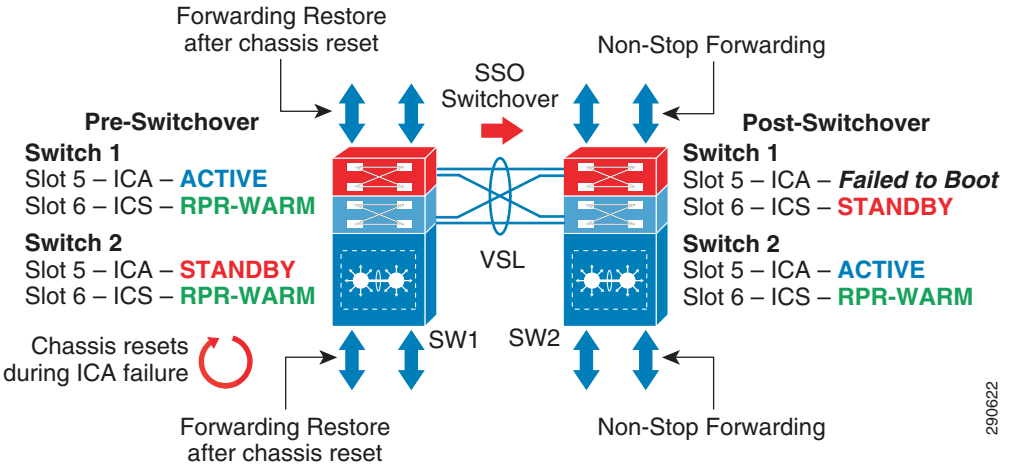
Note Users must use the **show switch virtual redundancy** or **show module** commands to verify the current role of the quad-sup and their status. The **show redundancy** command continues to provide dual-sup role and status information, however it doesnot provide any quad-sup specific information.

6500-E VSS Intra-Chassis Supervisor Switchover

The design using the quad-sup stateless intra-chassis redundancy option is same as the dual-sup VSS design. During an ICA supervisor (ACTIVE or STANDBY) failure, the entire chassis and all modules in the impacted chassis are reset. If the original ICA supervisor fails to reboot, the redundant ICS supervisor module takes over chassis ownership and bootup in the ICA role in order to restore the original network capacity and reliability in the virtual switch system. Administrative reset or failure of a redundant ICS supervisor module does not cause virtual switch chassis reset, since it also acts as a distributed linecard and is not actively handling any of the control plane or switch fabric ownership in the chassis.

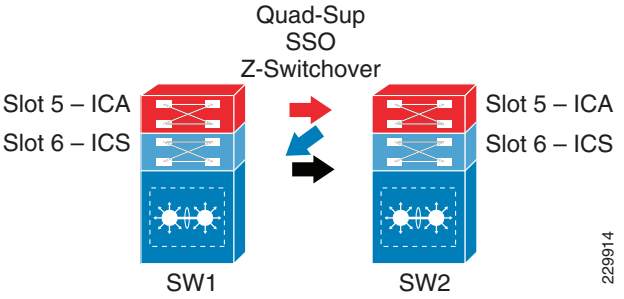
Deploying Catalyst 6500-E in VSS mode with quad-sup capability continues to provide the same level of inter-chassis SSO redundancy as the dual-sup design. The SSO ACTIVE supervisor synchronizes all the run time and stateful information from the SSO HOT-STANDBY supervisor module that resides on the peer virtual switch chassis. Hence during ACTIVE supervisor failure, the operation of the network remains transparent, as the remote virtual switch gracefully takes over software control plane ownership (as illustrated in Figure 25).

Figure 25 VSS Quad-Sup Switchover



Since the VSS domain is now equipped with a quad-sup, the logic utilized to synchronize information and identify the redundancy role of each supervisor module is different than that utilized in the dual-sup deployment. Depending on the reason supervisor reset, Cisco VSS internally sets a bootup parameter that modifies ICA or ICS role preference in the next bootup process. Such software design provides built-in system reliability in order to detect the fault, take over ICA ownership, and stabilize the overall virtual switch and network operation. This integrated quad-sup switchover capability is known as "Quad-Sup SSO Z-switchover" and is transparent to the user. It does not require any manual user intervention for optimization. Figure 26 illustrates the deterministic supervisor roles that occur during multiple switchover events.

Figure 26 VSS Quad-Sup Z-Switchover Process



Network administrators can verify the target supervisor module of the next SSO switchover event using the **show switch virtual redundancy exec** command.

Implementing SSO Redundancy

To deploy SSO supervisor redundancy, it is important to remember that both supervisor modules must be identical in hardware type, software version, and all the internal hardware components—memory and bootflash must be the same to provide complete operational transparency during failure. The default redundancy mode on all modular Catalyst and Nexus 7000 series platforms is SSO. Hence it does not require any additional configuration to enable SSO redundancy. The SSO redundant status can be verified using the following command on each recommended system:

Cisco IOS—Catalyst 6500-E VSS Mode

```
cr23-VSS-Core#show switch virtual redundancy
My Switch Id = 1
Peer Switch Id = 2
Configured Redundancy Mode = sso
Operating Redundancy Mode = sso

Switch 1 Slot 5 Processor Information :
```

```
-------------------------------------------
Current Software state = ACTIVE
<snippet>
Fabric State = ACTIVE
Control Plane State = ACTIVE

Switch 2 Slot 5 Processor Information :
-------------------------------------------
Current Software state = STANDBY HOT (switchover target)
<snippet>
Fabric State = ACTIVE
Control Plane State = STANDBY
```

Cisco IOS—Catalyst 4500-E

```
cr40-4507-1#show redundancy states
       my state = 13 -ACTIVE
     peer state = 8  -STANDBY HOT
       <snip>
Redundancy Mode (Operational) = Stateful Switchover
Redundancy Mode (Configured)  = Stateful Switchover
Redundancy State              = Stateful Switchover
    Manual Swact = enabled
 Communications = Up
<snip>
```

Cisco NX-OS—Cisco Nexus 7000

```
cr35-N7K-Core1# show redundancy status
Redundancy mode
---------------
      administrative:   HA
         operational:   HA
This supervisor (sup-5)
-----------------------
    Redundancy state:   Active
    Supervisor state:   Active
      Internal state:   Active with HA standby
Other supervisor (sup-6)
------------------------
    Redundancy state:   Standby
    Supervisor state:   HA standby
      Internal state:   HA standby
```

Non-Stop Forwarding (NSF)

When implementing NSF technology in systems using SSO redundancy mode, network disruptions are transparent to campus users and applications and high availability is provided even during periods where the control plane processing module (Supervisor/Route-Processor) is reset. During a failure, the underlying Layer 3 NSF-capable protocols perform graceful network topology re-synchronization. The preset forwarding information on the redundant processor or distributed linecard hardware remains intact and continues to switch network packets. This service availability significantly lowers the Mean Time To Repair (MTTR) and increases the Mean Time Between Failure (MTBF) to achieve the highest level of network availability.

NSF is an integral part of a routing protocol and depends on the following fundamental principles of Layer 3 packet forwarding:

- *Cisco Express Forwarding (CEF)*—CEF is the primary mechanism used to program the network path into the hardware for packet forwarding. NSF relies on the separation of the control plane update and the forwarding plane information. The control plane provides the routing protocol with a graceful restart and the forwarding plane switches packets using hardware acceleration where available. CEF enables this separation by programming hardware with FIB entries in all Catalyst switches. This ability plays a critical role in NSF/SSO failover.
- *Routing protocol*—The motivation behind NSF is route convergence avoidance. From a protocol operation perspective, this requires the adjacent routers to support a routing protocol with special intelligence that allows a neighbor to be aware that NSF-capable routers can undergo switchover so that its peer can continue to forward packets. This may bring its adjacency to a hold-down state (NSF recovery mode) for a brief period and request that routing protocol information be resynchronized.

A router that has the capability for continuous forwarding during a switchover is *NSF-capable*. Devices that support the routing protocol extensions such that they continue to forward traffic *to* a restarting router are *NSF-aware*. A Cisco device that is NSF-capable is also NSF-aware. The NSF capability must be manually enabled on each redundant system on a per-routing-protocol basis. The NSF-aware function is enabled by default on all Layer 3 platforms. describes the Layer 3 NSF-capable and NSF-aware platforms deployed in the campus network environment.

Implementing EIGRP NSF Capability

The following sample configuration illustrates how to enable the NSF capability within EIGRP (the same procedure applies to OSPF) on each Layer 3 campus LAN/WAN system deployed with redundant supervisors and route-processors or in virtual-switching modes (i.e., Cisco VSS, Catalyst 4500E, and StackWise Plus). EIGRP NSF capability is enabled by default on the Cisco Nexus 7000 system:

Cisco IOS—Catalyst Platforms

```
cr23-vss-core(config)#router eigrp 100
cr23-vss-core (config-router)#nsf
```

```
cr23-vss-core #show ip protocols | inc NSF
*** IP Routing is NSF aware ***
  EIGRP NSF-aware route hold timer is 240
  EIGRP NSF enabled
     NSF signal timer is 20s
     NSF converge timer is 120s
```

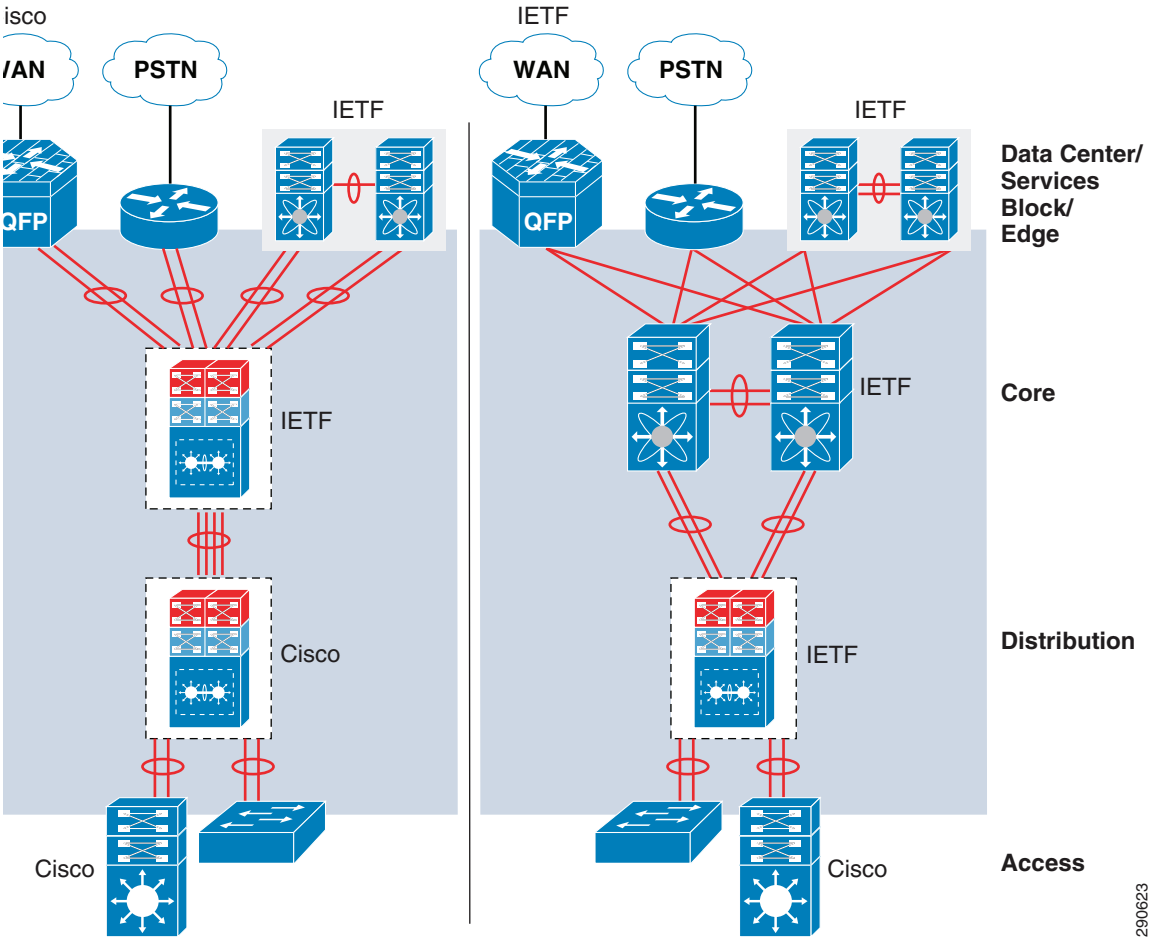
Cisco NX-OX—Nexus 7000

```
cr35-N7K-Core1#show ip eigrp | inc Grace
  Graceful-Restart: Enabled
```

Implementing OSPF NSF Capability

The OSPF NSF capability and helper function in Cisco IOS-based systems is supported in two modes—Cisco proprietary and IETF standard-based. The NX-OS running on the Nexus 7000 system supports OSPF NSF capability and helper function based on the IETF standard. Depending on the campus network design, the network administrator must implement the correct OSPF NSF capability between two adjacent Layer 3 campus systems to recognize and respond to the graceful restart capability in an OSPF TLV packet during supervisor switchover. By default, enabling OSPF NSF capability on Cisco IOS routers and switches enables the Cisco proprietary NSF function, whereas the IETF NSF capability is by default enabled on the Nexus 7000 system. Figure 27 illustrates the recommended OSPF NSF capability in each campus network design.

Figure 27 Recommended OSPF NSF Capability in Campus



Cisco IOS—Cisco NSF Capability

```
cr23-vss-core(config)#router ospf 100
cr23-vss-core (config-router)#nsf
cr23-vss-core# show ip ospf | inc Non-Stop|helper
 Non-Stop Forwarding enabled
 IETF NSF helper support enabled
 Cisco NSF helper support enabled
```

Cisco IOS—IETF NSF Capability

```
cr23-vss-core(config)#router ospf 100
cr23-vss-core(config-router)#nsf ietf
cr23-vss-core#show ip ospf | inc Non-Stop|helper
 IETF Non-Stop Forwarding enabled
 IETF NSF helper support enabled
 Cisco NSF helper support enabled
```

Cisco NX-OS—IETF NSF Capability

```
!IETF OSPF NSF capability is enabled by default
cr35-N7K-Core1#show ip ospf | inc Stateful|Graceful
 Stateful High Availability enabled
 Graceful-restart is configured
```

Graceful Restart Example

The following example demonstrates how the EIGRP protocol gracefully recovers when active supervisor/chassis switchover on a Cisco VSS and Nexus 7000 core system is forced by a reset:

- Cisco IOS

```
cr23-VSS-Core#redundancy force-switchover
This will reload the active unit and force switchover to standby[confirm]y

! VSS active system reset will force all linecards and ports to go down
!the  following logs confirms connectivity loss to core system
%LINK-3-UPDOWN: Interface TenGigabitEthernet2/1/2, changed state to down
%LINK-3-UPDOWN: Interface TenGigabitEthernet2/1/4, changed state to down

! Downed interfaces are automatically removed from EtherChannel/MEC,
! however additional interface to new active chassis retains port-channel in up/up
state
%EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet2/1/2 left the port-channel
Port-channel100
%EC-SW1_SP-5-UNBUNDLE: Interface TenGigabitEthernet2/1/4 left the port-channel
Port-channel100

! EIGRP protocol completes graceful recovery with new active virtual-switch.
%DUAL-5-NBRCHANGE: EIGRP-IPv4:(613) 100: Neighbor 10.125.0.12 (Port-channel100) is
resync: peer graceful-restart
```

- Cisco NX-OS

```
cr35-N7K-Core1#system switchover

! EIGRP protocol completes graceful recovery with new active supervisor
%DUAL-5-NBRCHANGE: EIGRP-IPv4 100: Neighbor 10.125.10.1 (Port-channel3) is resync:
peer graceful-restart
```

NSF Timers

The OSPF routing information stalls the routes and forwarding information for several seconds to gracefully recover the OSPF adjacencies and re-synchronize the database. By default the OSPF NSF timer on Cisco Catalyst switches is 120 seconds and the Nexus 7000 system can hold routing information for up to 60 seconds. Lowering the timer values may abruptly terminate graceful recovery, which can cause network instability. The default timer setting is tuned for a well-structured and concise campus LAN network topology. It is recommended to retain the default route hold timers in the network unless it is observed that NSF recovery takes more than the default values.

NSF/SSO Recovery Analysis

As described in a previous section, the NSF/SSO implementation and its recovery process differ on the Nexus 7000, Catalyst 4500E (Intra-Chassis), and Catalyst 6500-E VSS (Inter-Chassis) in the Borderless Campus LAN design. In each deployment scenario, the Cisco enterprise solution architecture validated the network recovery and application performance by inducing several types of active supervisor faults that trigger Layer 3 protocol graceful recovery. During each test, the switches continued to provide network accessibility during the recovery stage.

The Nexus 7000 and Catalyst 4500E systems retain the operational and forwarding state of the linecard and fabric modules for non-stop forwarding while the new active supervisor module goes through the graceful recovery process.

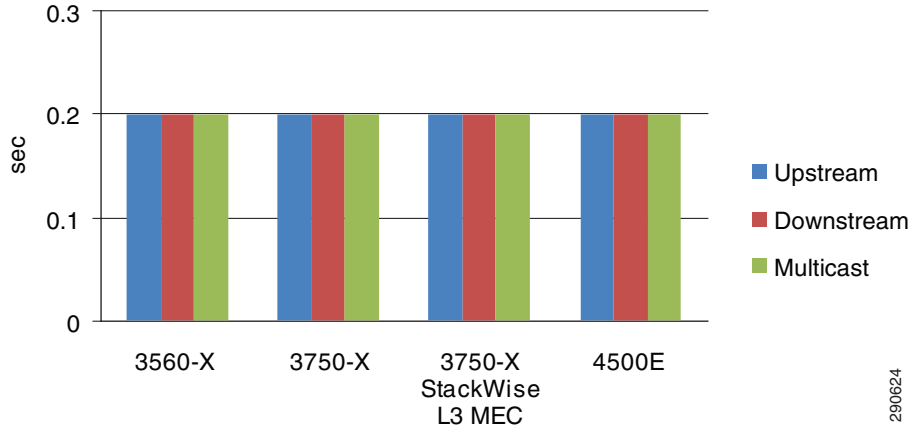
The inter-chassis SSO implementation in Catalyst 6500-E VSS differs from the single-chassis redundant implementation in that during active virtual switch node failure the entire chassis and all the linecards installed reset. However, with Layer 2/3 MEC links, the network protocols and forwarding information remain protected via the remote virtual switch node that can provide seamless network availability.

Catalyst 6500-E VSS NSF/SSO Recovery Analysis

As described earlier, in dual-sup or quad-sup Cisco VSS designs, the entire Catalyst 6500-E chassis and all installed linecard modules are reset during an in-chassis active (SSO ACTIVE or HOT-STANDBY) virtual switch switchover event. With a diverse full-mesh fiber network design, the Layer 2/Layer 3 remote device perceives this event as a loss of a member link since the alternate link to the standby switch is in an operational and forwarding state. The standby virtual switch detects the loss of the VSL Etherchannel and transitions into the active role and initializes Layer 3 protocol graceful recovery with the remote devices. Since there are no major network topology changes and member links are still in an operational state, the NSF/SSO recovery in Catalyst 6500-E VSS system is identical to the scenario where individual links are lost.

Additionally, the Cisco Catalyst 6500-E supports Multicast Multilayer Switching (MMLS) NSF with SSO, thereby enabling the system to maintain the multicast forwarding state in PFC3- and DFC3-based hardware during an active virtual switch reset. The new active virtual switch reestablishes PIM adjacency while continuing to switch multicast traffic based on pre-switchover programmed information (see Figure 28).

Figure 28 Catalyst 6500-E Dual-Sup and Quad-Sup VSS NSF/SSO Recovery Analysis



Nexus 7000 NSF/SSO Recovery Analysis

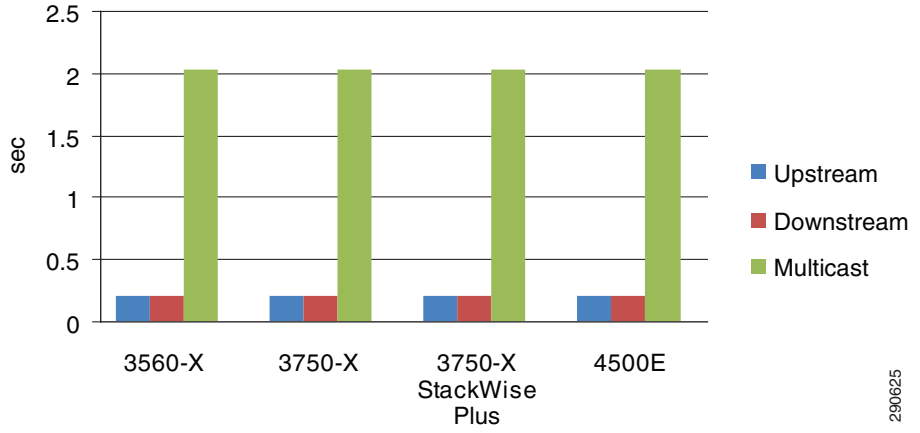
Since the Nexus 7000 system is designed with a distributed architecture to decouple the centralized control plane from the distributed data plane, the supervisor switchover process remains transparent and hitless to the network. During the supervisor switchover process, the distributed I/O and crossbar modules remain intact with synchronized forwarding information across the system. The egress forwarding information lookup and network services, such as QoS and ACL, are perfomed at the I/O module and the campus backbone network remains hitless with zero packet loss during an active or standby supervisor switchover event. The campus core remains hitless when the Cisco Nexus 7000 system is in various supervisor fault conditions, such as administrative forced switchover, manual OIR, or a hardware or software crash.

Catalyst 4500E NSF/SSO Recovery Analysis

Figure 29 illustrates an intra-chassis NSF/SSO recovery analysis for the Catalyst 4500E chassis deployed with Sup7-E, Sup6-E, or Sup6L-E in redundant mode. With EIGRP NSF/SSO capability, the unicast traffic consistently recovers within 200 msec. or less. However, the Catalyst 4500E does not currently support redundancy for Layer 3 multicast routing and forwarding information. Therefore,

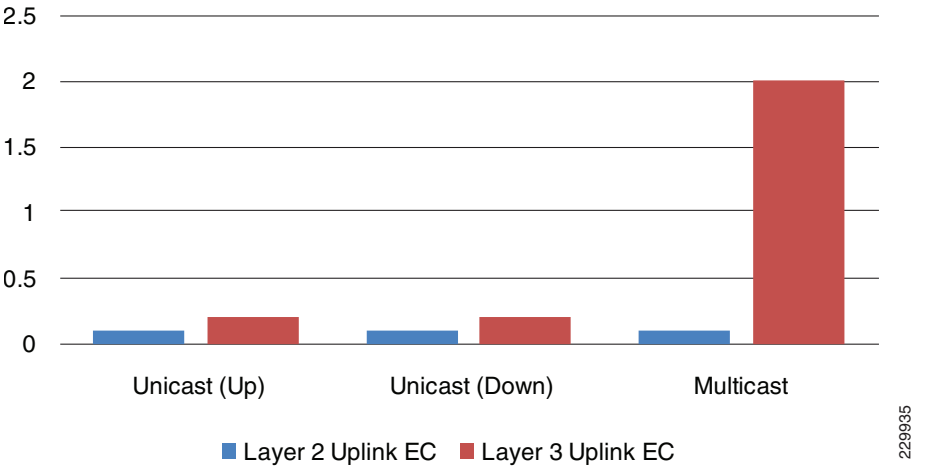
there may be an approximately two second loss of multicast traffic, since the switch has to reestablish all the multicast routing and forwarding information during the switchover event associated with the Sup7-E, Sup6-E, or Sup6L-E.

Figure 29 Catalyst 4500E Distribution Layer NSF/SSO Recovery Analysis



At the campus access layer, the Cisco Catalyst 4500E series platform provides unparalleled high availability to network applications with its unique forwarding architecture. During supervisor switchover, all synchronized Layer 2 forwarding information remains on the standby supervisor module that gracefully takes over control plane ownership; with the uplink port active on the failed supervisor module, the uplink capacity and Layer 2 adjacency are unaffected. Due to the highly-resilient platform design, network recovery is low sub-seconds for unicast and multicast traffic, as illustrated in Figure 30, when the Cisco Catalyst 4500E in the access layer is deployed in Multilayer and Routed-Access mode.

Figure 30 Catalyst 4500E Access Layer NSF/SSO Recovery Analysis



Catalyst 4500E Standby Supervisor Failure and Recovery Analysis

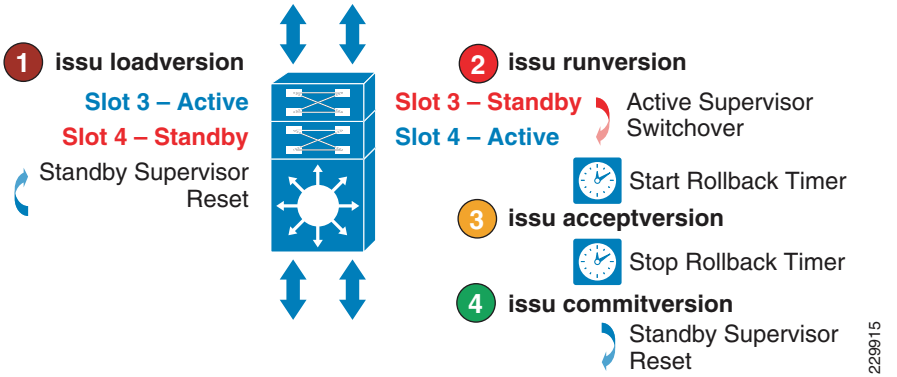
The standby Sup7-E, Sup6-E, or Sup6L-E supervisor remains in redundant mode while the active supervisor is in an operational state. If the standby supervisor gets reset or re-inserted, protocol graceful recovery is not triggered, nor are there any changes in network topology. Hence the standby supervisor remains completely transparent to the system and to rest of the network. The uplink port of the standby supervisor remains in an operational and forwarding state and the network bandwidth capacity remains intact during a standby supervisor soft switchover event.

9 Implementing Operational Resiliency

Path redundancy is often used to facilitate access during maintenance activity. However, single standalone systems are single points of failure and this type of network design simply does not provide user access if a critical node is taken out of service. Leveraging enterprise-class high availability features like NSF/SSO in the distribution and core layer Catalyst 4500E and 6500-E Series platforms enables support for ISSU and real-time network upgrade capability. Using ISSU and eFSU technology, the network administrator can upgrade the Cisco IOS software to implement new features, software bug fixes, or critical security fixes in real time.

Catalyst 4500E ISSU Software Design and Upgrade Process

Figure 31 Catalyst 4500E Manual ISSU Software Upgrade Process



ISSU Software Upgrade Pre-Requisite

ISSU Compatibility Matrix

When a redundant Catalyst 4500E system is brought up with a different Cisco IOS software version, the ISSU stored compatibility matrix information is analyzed internally to determine interoperability between the software running on the active and standby supervisors. ISSU provides SSO compatibility between several versions of software releases shipped during a 18 month period. Prior to upgrading the software, the network administrator must verify ISSU software compatibility with the following **show** command. Incompatible software may cause the standby supervisor to boot in RPR mode, which may result in a network outage:

```
cr24-4507e-MB#show issu comp-matrix stored
Number of Matrices in Table = 1
                My Image ver:  12.2(53)SG
                Peer Version    Compatibility
                ----------------    -------------
                12.2(44)SGBase(2)
                12.2(46)SG              Base(2)
                12.2(44)SG1             Base(2)
    …
```

Software

Prior to starting the software upgrade process, it is recommended to copy the old and new Cisco IOS software on the Catalyst 4500E active and standby supervisor into local file systems—Bootflash or Compact Flash.

```
cr24-4507e-MB#dir slot0:
Directory of slot0:/
 1  -rw- 25442405 Nov 23 2009 17:53:48 -05:00  cat4500e-entservicesk9-mz.122-53.SG1 ← new image
 2  -rw- 25443451 Aug 22 2009 13:26:52 -04:00  cat4500e-entservicesk9-mz.122-53.SG  ← old image

cr24-4507e-MB#dir slaveslot0:
Directory of slaveslot0:/

  1  -rw- 25443451 Aug 22 2009 13:22:00 -04:00  cat4500e-entservicesk9-mz.122-53.SG ← old image
  2  -rw- 25442405 Nov 23 2009 17:56:46 -05:00  cat4500e-entservicesk9-mz.122-53.SG1 ← new image
```

Configuration

It is recommended to save the running configuration to NVRAM and other local or remote locations such as bootflash or TFTP server prior to upgrading IOS software.

Boot Variable and String

The system default boot variable is defined to boot from the local file system. Make sure the default setting is not changed and the configuration register is set to 0x2102.

Modify the boot string to point to the new image to boot from a new IOS software version after the next reset triggered during the ISSU upgrade process. Refer to following URL for additional ISSU pre-requisites:

http://www.cisco.com/en/US/partner/docs/switches/lan/catalyst4500/12.2/53SG/configuration/issu.html#wp1072849

Catalyst 4500E Manual ISSU Software Upgrade Procedure

This subsection provides the manual software upgrade procedure for a Catalyst 4500E deployed in the enterprise campus LAN network design in several different roles—access, distribution, core, collapsed core, and Metro Ethernet WAN edge. The manual ISSU upgrade capability is supported on Catalyst 4500E Sup7-E, Sup6-E, and Sup6L-E supervisors running the Cisco IOS Enterprise feature set. However the automatic ISSU upgrade capability is only supported on the next generation Catalyst 4500E Sup7-E supervisor module.

In the following sample output, the Sup6-E supervisor is installed in Slot3 and Slot4 respectively. The Slot3 supervisor is in the SSO Active role and the Slot4 supervisor is in Standby role. Both supervisors are running identical 12.2(53)SG Cisco IOS software versions and are fully synchronized with SSO.

```
cr24-4507e-MB#show module | inc Chassis|Sup|12.2
Chassis Type : WS-C4507R-E
!Common Supervisor Module Type
 3     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXQ3
 4     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXRQ
!Common operating system version
 3     0021.d8f5.45c0 to 0021.d8f5.45c5 0.4 12.2(33r)SG     ( 12.2(53)SG       Ok
 4     0021.d8f5.45c6 to 0021.d8f5.45cb 0.4 12.2(33r)SG     ( 12.2(53)SG       Ok
!SSO Synchronized
 3     Active Supervisor       SSO Active
 4     Standby Supervisor       SSO Standby hot
```

The following provides the step-by-step procedure to upgrade the Cisco IOS Release 12.2(53)SG to 12.2(53)SG1 Cisco IOS release without causing network topology and forwarding disruption. Prior to issuing the **issu commitversion** command, the ISSU software upgrade can be aborted at any stage by issuing the **issu abortversion** command if any failure is detected.

1. *ISSU loadversion*—This first step will direct the active supervisor to initialize the ISSU software upgrade process.

   ```
   cr24-4507e-MB#issu loadversion 3 slot0:cat4500e-entservicesk9-mz.122-53.SG1 4
   slaveslot0: cat4500e-entservicesk9-mz.122-53.SG1
   ```

   After issuing the above command, the active supervisor ensures the new IOS software is downloaded on both supervisors' file systems and performs several additional checks on the standby supervisor for the graceful software upgrade process. ISSU changes the boot variable with the new IOS software version if no errors are found and resets the standby supervisor module.

   ```
   %RF-5-RF_RELOAD: Peer reload. Reason: ISSU Loadversion
   ```

Note Resetting the standby supervisor will not trigger a network protocol graceful recovery and all standby supervisor uplink ports will remain in operational and forwarding state for the transparent upgrade process.

With the broad range of ISSU version compatibility used in conducting SSO communication, the standby supervisor will then successfully bootup again in its original standby state:

```
cr24-4507e-MB#show module | inc Chassis|Sup|12.2
Chassis Type : WS-C4507R-E
! Common Supervisor Module Type
 3     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXQ3
 4     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXRQ
! Mismatch operating system version
 3     0021.d8f5.45c0 to 0021.d8f5.45c5 0.4 12.2(33r)SG( 12.2(53)SG       Ok
 4     0021.d8f5.45c6 to 0021.d8f5.45cb 0.4 12.2(33r)SG( 12.2(53)SG1      Ok
!SSO Synchronized
 3     Active Supervisor       SSO Active
 4     Standby Supervisor       SSO Standby hot
```

This bootup process will force the active supervisor to re-synchronize all SSO redundancy and checkpoints, VLAN database, and forwarding information with the standby supervisor and will notify the user to proceed with the next ISSU step.

```
%C4K_REDUNDANCY-5-CONFIGSYNC: The config-reg has been successfully synchronized to the
standby supervisor
%C4K_REDUNDANCY-5-CONFIGSYNC: The startup-config has been successfully synchronized to
the standby supervisor
%C4K_REDUNDANCY-5-CONFIGSYNC: The private-config has been successfully synchronized to
the standby supervisor
%C4K_REDUNDANCY-5-CONFIGSYNC_RATELIMIT: The vlan database has been successfully
synchronized to the standby supervisor

%ISSU_PROCESS-7-DEBUG: Peer state is [ STANDBY HOT ]; Please issue the runversion
command
```

2. *ISSU runversion*—After ensuring that the newly-loaded software is stable on the standby supervisor, the network administrator must proceed to the second step:

```
cr24-4507e-MB#issu runversion 4
This command will reload the Active unit.  Proceed ? [confirm]y
%RF-5-RF_RELOAD: Self reload. Reason: Admin ISSU runversion CLI
%SYS-5-RELOAD: Reload requested by console. Reload reason: Admin ISSU runversion
```

This step will force the current active supervisor to reset itself, thereby triggering network protocol graceful recovery with peer devices. However the uplink ports of the active supervisor remain intact and the data plane is not impacted during the switchover process. From an overall network perspective, the active supervisor reset caused by the **issu runversion** command will be no different than in similar switchover procedures (e.g., administrator-forced switchover or supervisor online insertion and removal). During the entire software upgrade procedure, this is the only step that performs SSO-based network graceful recovery. The following syslog on various Layer 3 systems confirm stable and EIGRP graceful recovery with the new supervisor running the new Cisco IOS software version.

- NSF-Aware Core

```
cr23-VSS-Core#
%DUAL-5-NBRCHANGE: EIGRP-IPv4:(415) 100: Neighbor 10.125.0.15 (Port-channel102) is
resync: peer graceful-restart
```

- NSF-Aware Layer 3 Access

```
cr24-3560-MB#
%DUAL-5-NBRCHANGE: EIGRP-IPv4:(100) 100: Neighbor 10.125.0.10 (Port-channel1) is
resync: peer graceful-restart
```

The previously active supervisor module will boot up in the standby role with the older IOS software version instead of the new IOS software version.

```
cr24-4507e-MB#show module | inc Chassis|Sup|12.2
```

```
Chassis Type : WS-C4507R-E
! Common Supervisor Module Type
 3      6   Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXQ3
 4      6   Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXRQ
! Mismatch operating system version
 3      0021.d8f5.45c0 to 0021.d8f5.45c5 0.4 12.2(33r)SG( 12.2(53)SG        Ok
 4      0021.d8f5.45c6 to 0021.d8f5.45cb 0.4 12.2(33r)SG( 12.2(53)SG1       Ok
!SSO Synchronized
 3     Active Supervisor       SSO Standby hot
 4     Standby Supervisor       SSO Active
```

This safeguarded software design provides an opportunity to roll back to the previous IOS software if the system upgrade causes any network abnormalities. At this stage, ISSU automatically starts internal rollback timers to re-install the old IOS image. The default rollback timer is up to 45 minutes, which provides a network administrator with an opportunity to perform several sanity checks. In small to mid-size network designs, the default timer may be sufficient. However, for large networks, network administrators may want to increase the timer up to two hours:

```
cr24-4507e-MB#show issu rollback-timer
       Rollback Process State = In progress
      Configured Rollback Time = 45:00
       Automatic Rollback Time = 19:51
```

The system will notify the network administrator with the following, instructing them to move to the next ISSU upgrade step if no stability issues are observed and all the network services are operating as expected.

```
%ISSU_PROCESS-7-DEBUG: Peer state is [ STANDBY HOT ]; Please issue the acceptversion
command
```

3. *ISSU acceptversion* (Optional)—This step provides confirmation from the network administrator that the system and network is stable after the IOS install and they are ready to accept the new IOS software on the standby supervisor. This step stops the rollback timer and instructs the network administrator to issue the final commit command. The network admnistrator can optionally skip this upgrade step and issue the final commit within the rollback timer window:

```
cr24-4507e-MB#issu acceptversion 4
% Rollback timer stopped. Please issue the commitversion command.

cr24-4507e-MB#show issu rollback-timer
     Rollback Process State = Not in progress
   Configured Rollback Time = 45:00


cr24-4507e-MB#show module | inc Chassis|Sup|12.2
Chassis Type : WS-C4507R-E
! Common Supervisor Module Type
 3      6   Sup 6-E 10GE (X2), 1000BaseX (SFP)      WS-X45-SUP6-E      JAE1132SXQ3
```

```
 4     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)     WS-X45-SUP6-E     JAE1132SXRQ
! Mismatch operating system version
 3     0021.d8f5.45c0 to 0021.d8f5.45c5 0.4 12.2(33r)SG( 12.2(53)SG       Ok
 4     0021.d8f5.45c6 to 0021.d8f5.45cb 0.4 12.2(33r)SG( 12.2(53)SG1      Ok
!SSO Synchronized
 3    Active Supervisor      SSO Standby hot
 4    Standby Supervisor      SSO Active
```

4. *ISSU commitversion*—This final ISSU step forces the active supervisor to synchronize its configuration with the standby supervisor and forces it to reboot with the new IOS software. This stage concludes the ISSU upgrade procedure and the new IOS version is permanently committed on both supervisor modules. If for some reason the network administrator wants to rollback to the older image, it is recommended to perform an ISSU-based downgrade procedure to retain the network operational state without any downtime.

```
cr24-4507e-MB#issu commitversion 3
Building configuration...
Compressed configuration from 24970 bytes to 10848 bytes[OK]
%C4K_REDUNDANCY-5-CONFIGSYNC: The private-config has been successfully synchronized to
the standby supervisor
%RF-5-RF_RELOAD: Peer reload. Reason: ISSU Commitversion

cr24-4507e-MB#show module | inc Chassis|Sup|12.2
Chassis Type : WS-C4507R-E
! Common Supervisor Module Type
 3     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)     WS-X45-SUP6-E     JAE1132SXQ3
 4     6  Sup 6-E 10GE (X2), 1000BaseX (SFP)     WS-X45-SUP6-E     JAE1132SXRQ
! Common new operating system version
 3     0021.d8f5.45c0 to 0021.d8f5.45c5 0.4 12.2(33r)SG( 12.2(53)SG1      Ok
 4     0021.d8f5.45c6 to 0021.d8f5.45cb 0.4 12.2(33r)SG( 12.2(53)SG1      Ok

!SSO Synchronized
 3    Active Supervisor      SSO Standby hot
 4    Standby Supervisor      SSO Active
```

Catalyst 4500E Automatic ISSU Software Upgrade Procedure

The network administrator can use the automatic ISSU upgrade method for large Catalyst 4500E Sup7-E-based campus networks once the manual ISSU upgrade procedure is successfully performed. It is recommended that the status of all network communication, operation, and manageability components on the Catalyst 4500E system now running with the new Cisco IOS-XE software be verified. Once the stability of the new IOS-XE software is confirmed, the network administrator can start a single-step automatic ISSU upgrade procedure on the remaining systems. Cisco IOS-XE also provides the flexibility to program the system for an automatic ISSU upgrade based on a user-defined future time.

The new **issu changeversion** command automates the upgrade of all four ISSU upgrade procedures into a single step that does not require manual intervention from the network administrator. The syntax must include the new targeted Cisco IOS-XE software to be installed on both supervisor modules. This gets set in the BOOT variable and the rest of the upgrade process becomes fully automated. Even with the automatic ISSU upgrade procedure, the standby supervisor module still gets reset when **issu loadversion**, **issu runversion**, and **issu commitversion** are executed by the software.

The following provides the single-step procedure to upgrade the Cisco IOS-XE Release from 3.1.0SG to pre- release Cisco IOS-XE software without causing network topology and forwarding disruption. Like the manual upgrade steps, the automatic ISSU upgrade can also be aborted at any stage by issuing the **issu abortversion** command:

1. ISSU changeversion—The only manual step to initialize automatic ISSU upgrade procedure on Cisco Catalyst 4500E system with Sup7-E supervisor module. The Catalyst 4500E system ensures the correct location and Cisco IOS-XE software information to initialize the automated ISSU software upgrade procedure. Both supervisors perform file system and other checks on the standby supervisor in order to ensure a graceful software upgrade process. The automatic ISSU procedure will modify the boot variable with the new IOS-XE software version if no errors are found. The rest of the ISSU upgrade procedure will automate starting with the standby supervisor module being force reset with a new software version. The network administrator can monitor the automated upgrade status and has flexibility to abort the entire process if any abnormal condition occurs during the new software installation process:

```
cr19-4507-MB#show issu state detail | exclude Pre|Post
    Slot = 3
    RP State = Active
    ISSU State = Init
    Operating Mode = Stateful Switchover
    Current Image = bootflash:cat4500e-universalk9.SPA.03.01.00.SG.150-1.XO.bin

    Slot = 4
    RP State = Standby
    ISSU State = Init
    Operating Mode = Stateful Switchover
    Current Image = bootflash: cat4500e-universalk9.SPA.03.01.00.SG.150-1.XO.bin

cr19-4507-MB#dir bootflash:
59009  -rw-  <truncated>  cat4500e-universalk9.SPA.03.01.00.SG.150-1.XO.bin <- old
image
29513  -rw-  <truncated>  cat4500e-universalk9.SSA.03.01.01.0.74.150.2.SG.bin <- new
image

cr19-4507-MB#dir slavebootflash:
14769  -rw-  <truncated>  cat4500e-universalk9.SPA.03.01.00.SG.150-1.XO.bin <- old
image
14758  -rw-  <truncated>  cat4500e-universalk9.SSA.03.01.01.0.74.150.2.SG.bin <- new
image
```

```
cr19-4507-LB#issu changeversion
bootflash:cat4500e-universalk9.SSA.03.01.01.0.74.150.2.SG.bin

!Automatically triggers issu loadversion that resets current standby supervisor module
% 'issu changeversion' is now executing 'issu loadversion'
% issu loadversion executed successfully, Standby is being reloaded
% changeversion finished executing loadversion, waiting for standby to reload and
reach SSO ...

cr19-4507-MB#show issu state detail | exclu Pre|Post
Slot = 3
RP State = Active
ISSU State = Load Version
Changeversion = TRUE
Operating Mode = not reached
Current Image = bootflash:cat4500e-universalk9.SPA.03.01.00.SG.150-1.XO.bin
Standby information is not available because it is in 'DISABLED' state

!Automatically triggers issu runversion that resets current active supervisor module.
This step will force SSO switchover, the new supervisor module gracefully recovers
protocol with neighbosr. Automatically starts ISSU roll-back timer

%INSTALLER-7-ISSU_OP_SUCC:  issu changeversion is now executing 'issu runversion'
%INSTALLER-7-ISSU_OP_SUCC:  issu changeversion successfully executed 'issu runversion'
Please stand by while rebooting the system...
Restarting system.

%INSTALLER-7-ISSU_OP_SUCC: Rollback timer started with timer value (2700)

cr19-4507-MB#show issu rollback-timer
     Rollback Process State = In progress
   Configured Rollback Time = 00:45:00
    Automatic Rollback Time = 00:43:18
```

Layer 3 neighbors gracefully resynchronize routing information with the new supervisor while maintaining and forwarding traffic across all EtherChannel member links, including on uplink ports of the old active supervisor module.

```
%DUAL-5-NBRCHANGE: EIGRP-IPv4:(100) 100: Neighbor 10.125.0.7 (Port-channel11) is
resync: peer graceful-restart

!Automatically triggers issu commitversion that stops roll-back timer and resets
current standby supervisor module to bootup with new targeted Cisco IOS-XE software

%INSTALLER-7-ISSU_OP_SUCC:  issu changeversion is now executing 'issu commitversion'
%HA_CONFIG_SYNC-6-BULK_CFGSYNC_SUCCEED: Bulk Sync succeeded
%RF-5-RF_TERMINAL_STATE: Terminal state reached for (SSO)

cr19-4507-MB#show issu rollback-timer
```

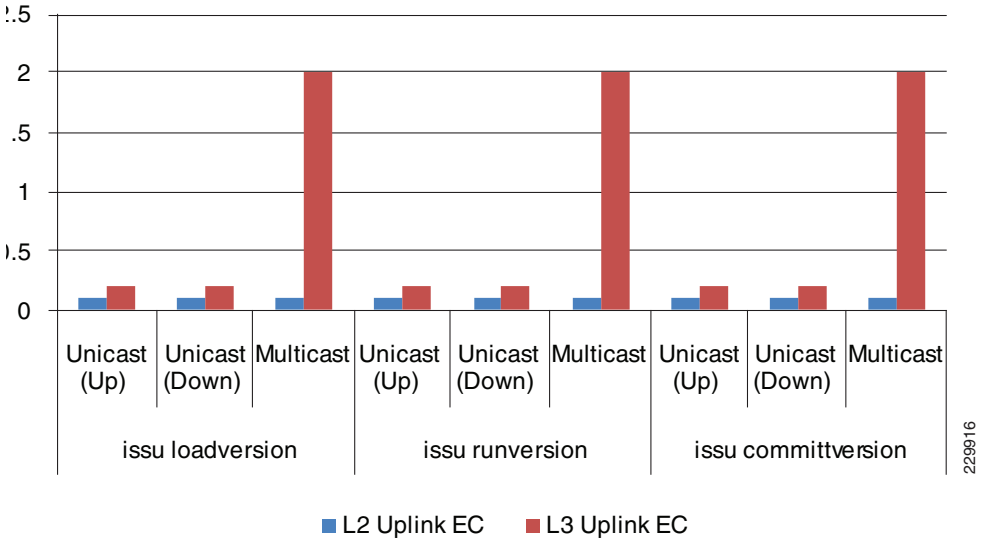
```
      Rollback Process State = Not in progress
    Configured Rollback Time = 00:45:00

cr19-4507-MB#show issu state detail | exc Pre|Post
        Slot = 4
        RP State = Active
        ISSU State = Init
        Operating Mode = Stateful Switchover
        Current Image = bootflash:cat4500e-universalk9.SSA.03.01.01.0.74.150.2.SG.bin

        Slot = 3
        RP State = Standby
        ISSU State = Init
        Operating Mode = Stateful Switchover
        Current Image = bootflash:cat4500e-universalk9.SSA.03.01.01.0.74.150.2.SG.bin
```

Catalyst 4500E Network Recovery with ISSU Software Upgrade

As described in the previous section, the Cisco Catalyst 4500E chassis in redundant supervisor mode gracefully resets the supervisor module without impacting any of its uplink ports. Hence even during the software upgrade procdure, the Cisco Catalyst 4500E chassis maintains its original network capacity and gracefully synchronizes with peer network devices for continuous forwarding of network traffic. This highly resilient architecture provides the network administrator with the flexibility to upgrade the Catalyst 4500E chassis with new Cisco IOS software without downtime or disruption in the operation of the network.

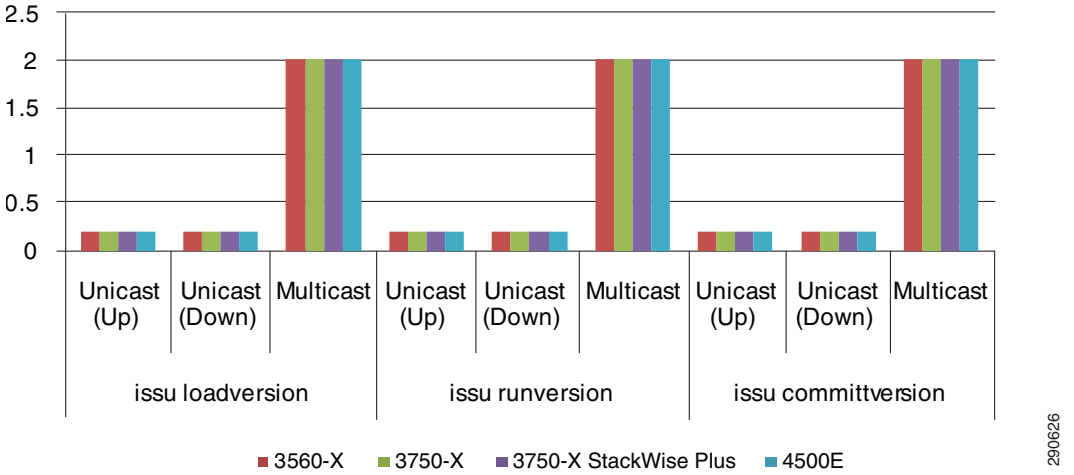
The ISSU software upgrade procedure is even more graceful and transparent with EtherChannel-based network topologies, which offer entire system upgrades with deterministic traffic loss information. The following two charts provide characterized ISSU test results during a supervisor reset that is triggered by **issu loadversion**, **issu runversion**, and i**ssu committversion** via a manual CLI or an automatic ISSU upgrade procedure on the Cisco Catalyst 4500E. These results are collected from a Catalyst 4500E chassis deployed in the campus access layer and at the distribution layer deployed with Layer 2 and Layer 3 EtherChannel:

Figure 32 Catalyst 4500E Access Layer Network Recovery with ISSU Software Upgrade



The Catalyst 4500E can be deployed in Layer 2 or Layer 3 mode at the campus access layer. During each ISSU software upgrade step, the network impact to the unicast or multicast traffic flow is at or below 200 msec range when the Catalyst 4500E system is deployed in Layer 2 mode. However when the routing boundary is extended to the access layer, the current Catalyst 4500E chassis does not fully support Layer 3 multicast high-availability. This means that the multicast traffic loss can be higher than the unicast flows.

Figure 33 Catalyst 4500E Distribution Layer Network Recovery with ISSU Software Upgrade



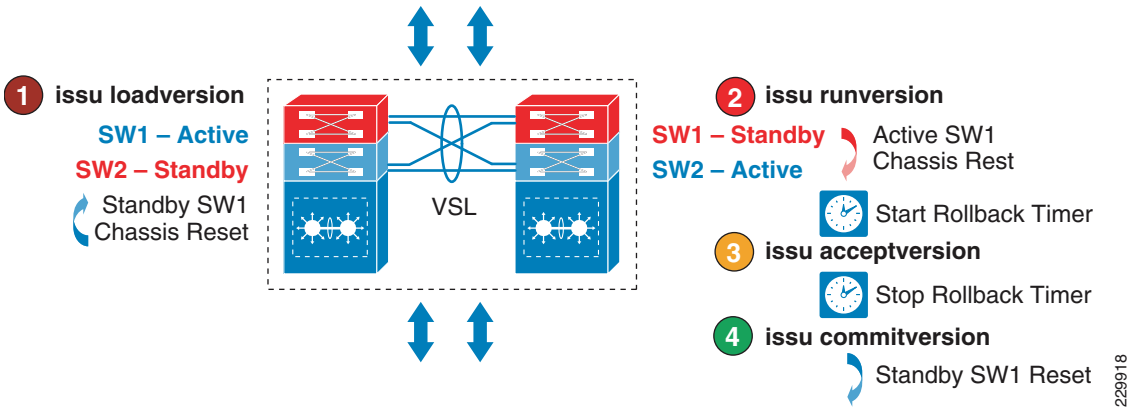
The Catalyst 4507R+E deployed in the distribution layer typically represents a demarcation point between Layer 2 and Layer 3 network boundaries. As described in the previous section, the current software architecture of the Catalyst 4507R+E series platform does not support Layer 3 multicast high-availability. Thus the multicast PIM neighbor adjacency and forwarding information are reset during the supervisor switchover process. This reset causes about a two second multicast traffic loss, but with consistent unicast traffic loss at or below 200 msec baseline range.

Catalyst 6500-E VSS eFSU Software Design and Upgrade Process

Cisco Catalyst VSS was introduced in the initial IOS Release 12.2(33)SXH that supported Fast Software Upgrade (FSU). In the initial introduction, it had limited high-availability capabilities to upgrade the IOS software release. The ISSU mismatched software version compatibility was not supported by the FSU infrastructure, which could cause network down time. This may not be a desirable solution when deploying the Catalyst 6500-E in the critical aggregation or core network tier.

Starting with IOS Release 12.2(33)SXI, the Catalyst 6500-E supports true transparent IOS software upgrade in standalone and virtual switch network designs. Enhanced Fast Software Upgrade (eFSU) made it completely ISSU infrastructure compliant and enhances the software and hardware design to retain its functional state during the graceful upgrade process.

Figure 34 Catalyst 6500-E VSS eFSU Software Upgrade Process

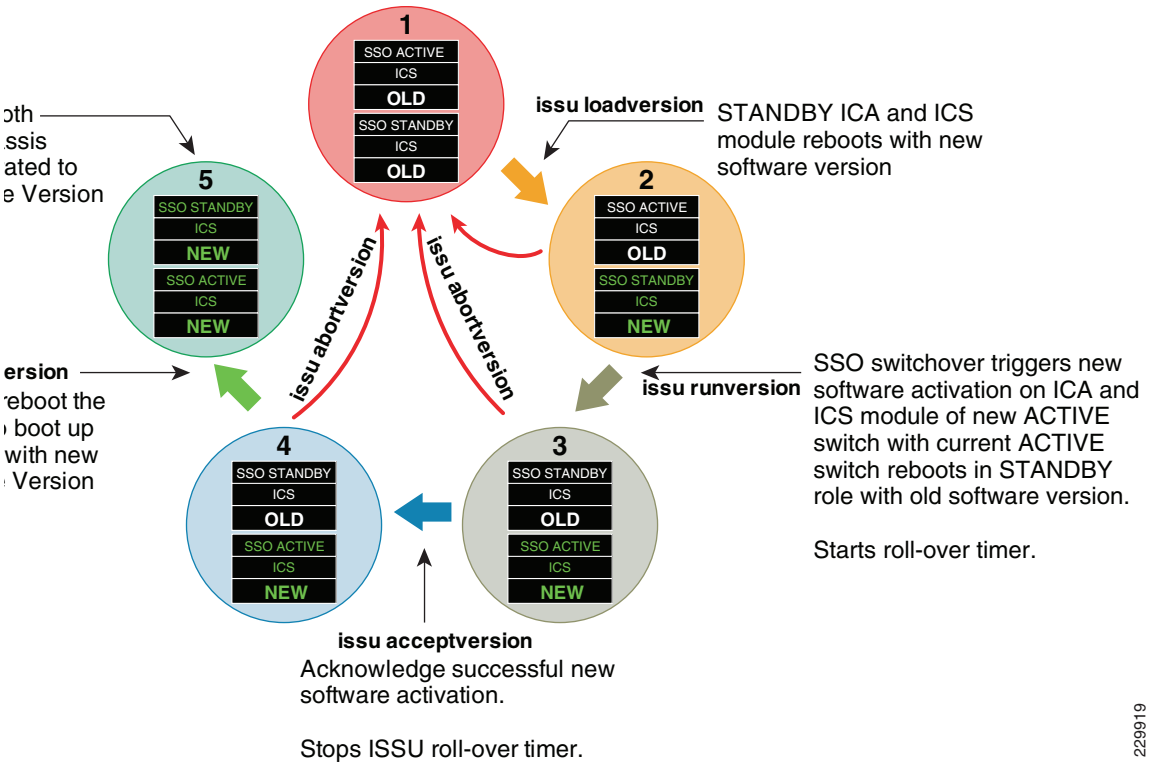


Since eFSU in the Catalyst 6500-E system is built on the ISSU infrastructure, most of the eFSU pre-requisites for Cisco VSS dual-sup design and IOS upgrade procedures remain consistent as explained in a previous sub-section. As described earlier, the Cisco VSS technology enables inter-chassis SSO communication between two virtual switch nodes. However, while the software upgrade procedure for inter-chassis eFSU upgrades is similar, the network operation slightly differs compared to ISSU implemented on intra-chassis based SSO design.

Catalyst 6500-E VSS Quad-Sup eFSU Software Upgrade Process

The eFSU software upgrade process remained simplified even when VSS is deployed with an increased number of supervisor modules in the domain. Upgrading four operational supervisor modules with minimal upgrade time and with reduced complexities, Cisco VSS software is designed to perform parallel ICA and ICS module upgrades by leveraging the existing eFSU cycle. As described earlier, the ICA supervisor module updates the BOOT parameters in ROMMON of the ICS supervisor module in order to to boot up with the new targeted Cisco IOS software. The Cisco VSS quad-sup follows the SSO Z-switchover mechanism as it goes through the entire eFSU upgrade cycle, as illustrated in Figure 1-35.

Figure 1-35 VSS Quad-Sup eFSU Process Cycle



Before going through the eFSU upgrade procedure on Cisco VSS deployed with quad-sup, the network administrator must make sure that the following prerequisites and guidelines for graceful system upgrade are followed:

- To gracefully upgrade all four supervisor modules, Cisco highly recommends that the ICA and ICS supervisor on both virtual switch chassis meet the following requirements:
  - Run common software version and license type.
  - The new IOS software version must be copied to local storage (e.g., disk0, bootdisk) of the ICA and ICS supervisor module.
  - All four supervisor modules are in a fully operational state (SSO ACTIVE/HOT_STANDBY or RPR-WARM mode)
- Do not insert a new ICS supervisor or swap ICS supervisors during any step of the eFSU upgrade procedure.

- During the eFSU software upgrade cycle, intra-chassis role switchover may occur when ISSU triggers a chassis reset. Hence it is strongly recommended to design the network as per the recommendations made in this design document.
- Save the configuration to startup-config, local storage, or at a remote location (e.g.. TFTP/FTP)
- Do not manually modify any BOOT variables and strings.

Catalyst 6500-E eFSU Software Upgrade Procedure

This subsection provides the software upgrade procedure for the Catalyst 6500-Es deployed in VSS mode with quad-sup in the Borderless Campus design. eFSU is supported on the Catalyst 6500-E Sup720-10GE supervisor module running Cisco IOS release with the Enterprise feature set.

In the following sample output, VSS capable Sup720-10G supervisor modules are installed in Slot5 and Slot6 of virtual switch SW1 and SW2, respectively. The virtual switch SW1 supervisor is in the SSO Active role and the SW2 supervisor is in the Standby hot role. In addition, with MEC and the distributed forwarding architecture, the forwarding plane is in an active state on both virtual switch nodes. Both ICA supervisors are running the Cisco IOS Release 12.2(33)SXI4 software version and are fully synchronized with SSO. ICS supervisor modules are running the same Sup720-LC software version and operating in RPR-WARM mode.

```
cr23-VSS-Core#show switch virtual redundancy | inc Mode|Switch|Image|Control
                  My Switch Id = 1
                Peer Switch Id = 2
    Configured Redundancy Mode = sso
     Operating Redundancy Mode = sso
Switch 1 Slot 5 Processor Information :
                  Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State =
Switch 1 Slot 6 Processor Information :
                  Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M),
Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State = RPR-Warm
Switch 2 Slot 5 Processor Information :
                  Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State = STANDBY
Switch 2 Slot 6 Processor Information :
                  Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M),
Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State = RPR-Warm
```

The following provides a step-by-step procedure to upgrade from Cisco IOS Release 12.2(33)SXI4 to 12.2(33)SXI4a without causing network topology and forwarding disruption. Each upgrade step can be aborted at any stage by issuing the **issu abortversion** command if any failures are detected.

1. *ISSU loadversion*—This first step will direct the active virtual switch node to initialize the ISSU software upgrade process.

```
cr22-6500-LB#show issu state
                           Slot = 1/5
                     RP State =
                    ISSU State = Init
                  Boot Variable =
bootdisk:s72033-adventerprisek9_wan-mz.122-33.SXI4.bin,12;
                           Slot = 2/5
                     RP State = Standby
                    ISSU State = Init
                  Boot Variable =
bootdisk:s72033-adventerprisek9_wan-mz.122-33.SXI4.bin,12;

cr23-VSS-Core#issu loadversion 1/5 disk0:s72033-adventerprisek9_wan-mz.122-33.SXI4a
2/4 slavedisk0: s72033-adventerprisek9_wan-mz.122-33.SXI4a
```

After issuing the above command, the active virtual switch ensures the new IOS software is downloaded on both supervisors' file systems and performs several additional checks on the ICA and ICS supervisor modules on the remote virtual switch for the graceful software upgrade process. ISSU changes the boot variable to the new IOS software version if no error is found and resets the standby virtual switch and installed modules.

```
%RF-SW1_SP-5-RF_RELOAD: Peer reload. Reason: ISSU Loadversion
%SYS-SW2_SPSTBY-5-RELOAD: Reload requested - From Active Switch (Reload peer unit).
%issu loadversion executed successfully, Standby is being reloaded
```

Note Resetting the standby virtual switch node will not trigger the network protocol graceful recovery process and will not reset the ICS supervisor module or the linecards on the active virtual switch. It will remain in an operational and forwarding state for the transparent upgrade process.

With the broad range of ISSU version compatibility for SSO communication, the standby supervisor will successfully bootup again in its original standby state:

```
cr23-VSS-Core#show switch virtual redundancy | inc Mode|Switch|Image|Control
                  My Switch Id = 1
                Peer Switch Id = 2
    Configured Redundancy Mode = sso
     Operating Redundancy Mode = sso
Switch 1 Slot 5 Processor Information :
```

```
                Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State =
Switch 1 Slot 6 Processor Information :
                Image Version = Cisco IOS Software, s72033_lc Software
(s72033_lc-SP-M), Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State = RPR-Warm
Switch 2 Slot 5 Processor Information :
                Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = STANDBY
Switch 2 Slot 6 Processor Information :
                Image Version = Cisco IOS Software, s72033_lc Software
(s72033_lc-SP-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = RPR-Warm
```

To rejoin the virtual switch domain, both chassis will reestablish distributed VSL EtherChannel communication and force the active supervisor to resynchronize all SSO redundancy and checkpoints, VLAN database, and forwarding information with the standby virtual switch. The network administrator is notified to proceed with the next ISSU step.

```
%HA_CONFIG_SYNC-6-BULK_CFGSYNC_SUCCEED: Bulk Sync succeeded
%PFREDUN-SW2_SPSTBY-6-STANDBY: Ready for SSO mode

%ISSU_PROCESS-SW1_SP-7-DEBUG: Peer state is [ STANDBY HOT ]; Please issue the
runversion command
```

2. *ISSU runversion*—After performing several steps to ensure the new loaded software is stable on the standby virtual switch, the network administrator is now ready to proceed to the runversion step.

```
cr23-VSS-Core#show issu state
                        Slot = 1/5
                  RP State =
                ISSU State = Load Version
              Boot Variable =
disk0:s72033-adventerprisek9_wan-mz.122-33.SXI4.bin,12

                        Slot = 2/5
                  RP State = Standby
                ISSU State = Load Version
              Boot Variable =
disk0:s72033-adventerprisek9_wan-mz.122-33.SXI4a,12;disk0:s72033-adventerprisek9_wan-m
z.122-33.SXI4.bin,12

cr23-VSS-Core#issu runversion 2/5
This command will reload the Active unit.  Proceed ? [confirm]y
%issu runversion initiated successfully
```

```
%RF-SW1_SP-5-RF_RELOAD: Self reload. Reason: Admin ISSU runversion CLI
```

This step will force the current active virtual switch (SW1) to reset itself along with its ICS supervisor module and the linecards, which triggers network protocol graceful recovery with peer devices. However the linecard and the ICS supervisor module on the current standby virtual switch (SW2) will remain intact and the data plane traffic will continue to be switched during the switchover process. From a network perspective, the effects of the active supervisor resetting during the ISSU runversion step will be no different than the normal switchover procedure (i.e., administration-forced switchover or supervisor online insertion and removal). In the entire eFSU software upgrade procedure, this is the only time that the systems will perform an SSO-based network graceful recovery. The following sample syslogs confirm stable and EIGRP graceful recovery on the virtual-switch running the new Cisco IOS software version.

NSF-Aware Distribution

```
cr24-4507e-MB#
%DUAL-5-NBRCHANGE: EIGRP-IPv4:(100) 100: Neighbor 10.125.0.14 (Port-channel1) is
resync: peer graceful-restart
```

After re-negotiating and establishing the VSL EtherChannel link and going through the VSLP protocol negotiation process, the rebooted virtual switch module boots up in the standby role with the older IOS software version instead of the new IOS software version. The ISSU runversion procedure starts the SSO Z-Switchover process in the VSS quad-sup design where the ICS can take over the ICA ownership during the next boot up process. Hence all the network configuration and the VSL connectivity must be deployed as recommended in this document for transparent network operation:

```
Dist-VSS#show switch virtual redundancy | inc Mode|Switch|Image|Control
                  My Switch Id = 2
                Peer Switch Id = 1
    Configured Redundancy Mode = sso
     Operating Redundancy Mode = sso
Switch 2 Slot 5 Processor Information :
Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State =
Switch 2 Slot 6 Processor Information :
Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M), Version
12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = RPR-Warm
Switch 1 Slot 6 Processor Information :
Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4, RELEASE SOFTWARE (fc3)
           Control Plane State = STANDBY
Switch 1 Slot 5 Processor Information :
Image Version = Cisco IOS Software, s72033_lc Software (s72033_lc-SP-M), Version
12.2(33)SXI4, RELEASE SOFTWARE (fc3)
```

```
            Control Plane State = RPR-Warm
```

Like intra-chassis ISSU implementation, eFSU also provides a safeguarded software design for additional network stability and the opportunity to roll back to the previous IOS software if the system upgrade causes any type of network abnormality. At this stage, ISSU automatically starts a set of internal rollback timers to re-install the old IOS image if there are any problems. The default rollback timer is up to 45 minutes, which provides the network administrator with an opportunity to perform several sanity checks. In small- to mid-size network designs, the default timer may be sufficient. However for large networks, the network administrator may want to adjust the timer up to two hours:

```
%ISSU_PROCESS-SP-7-DEBUG: rollback timer process has been started
cr23-VSS-Core#show issu rollback-timer
        Rollback Process State = In progress
      Configured Rollback Time = 00:45:00
       Automatic Rollback Time = 00:36:08
```

The system will notify the network administrator with the following syslog to continue to the next ISSU upgrade step if no stability issues occur and all the network services are operating as expected.

```
%ISSU_PROCESS-SW2_SP-7-DEBUG: Peer state is [ STANDBY HOT ]; Please issue the
acceptversion command
```

3. *ISSU acceptversion*—This eFSU step provides confirmation from the network administrator regarding the system and network stability after installing the new software. It confirms readiness to accept the new IOS software on the standby supervisor. This step stops the rollback timer and instructs the network administrator to continue to the final commit state. However, it does not perform any additional steps to install the new software on the standby supervisor.

```
cr23-VSS-Core#show issu state
                          Slot = 2/5
                     RP State =
                    ISSU State = Run Version
                  Boot Variable =
disk0:s72033-adventerprisek9_wan-mz.122-33.SXI4a,12;disk0:s72033-adventerprisek9_wan-m
z.122-33.SXI4.bin,12

                          Slot = 1/6
                     RP State = Standby
                    ISSU State = Run Version
                  Boot Variable =
disk0:s72033-adventerprisek9_wan-mz.122-33.SXI4.bin,12


cr23-VSS-Core#issu acceptversion 2/5
% Rollback timer stopped. Please issue the commitversion command.
cr23-VSS-Core#show issu rollback-timer
```

```
Rollback Process State = Not in progress
Configured Rollback Time = 00:45:00
```

4. *ISSU commitversion*—The final eFSU step forces the active virtual switch to synchronize the configuration with the standby supervisor and force it to reboot with the new IOS software. This stage concludes the eFSU upgrade procedure and the new IOS version is permanently committed on the ICA and ICS supervisor modules of both virtual switches. If for some reason the network administrator needs to rollback to the older image, it is recommended to perform the eFSU-based downgrade procedure to maintain the network operational state without any downtime.

```
cr23-VSS-Core#issu commitversion 1/6
Building configuration...
[OK]
%RF-SW2_SP-5-RF_RELOAD: Peer reload. Reason: Proxy request to reload peer
%SYS-SW1_SPSTBY-5-RELOAD: Reload requested - From Active Switch (Reload peer unit).
%issu commitversion executed successfully

cr23-VSS-Core#show switch virtual redundancy | inc Mode|Switch|Image|Control
                  My Switch Id = 2
                Peer Switch Id = 1
    Configured Redundancy Mode = sso
     Operating Redundancy Mode = sso
Switch 2 Slot 5 Processor Information :
                  Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State =
Switch 2 Slot 6 Processor Information :
                  Image Version = Cisco IOS Software, s72033_lc Software
(s72033_lc-SP-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = RPR-Warm
Switch 1 Slot 5 Processor Information :
                  Image Version = Cisco IOS Software, s72033_rp Software
(s72033_rp-ADVENTERPRISEK9_WAN-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = STANDBY
Switch 1 Slot 6 Processor Information :
                  Image Version = Cisco IOS Software, s72033_lc Software
(s72033_lc-SP-M), Version 12.2(33)SXI4a, RELEASE SOFTWARE (fc2)
           Control Plane State = RPR-Warm
```

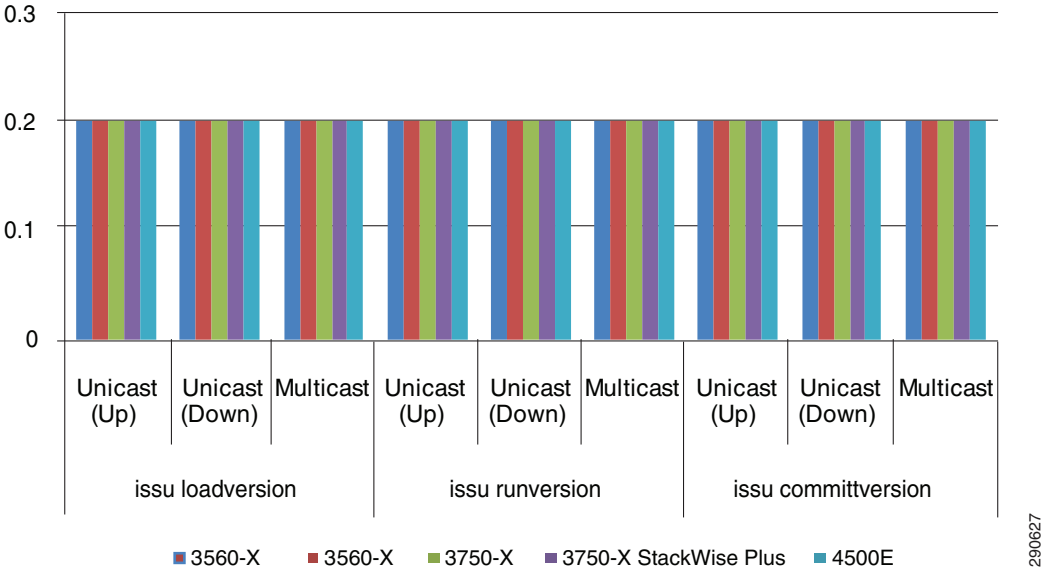
Catalyst 6500-E Network Recovery with eFSU Software Upgrade

As described in a previous section, the Cisco Catalyst 6500-E chassis in virtual switch mode will be reset to gracefully install the new Cisco IOS software version on all supervisor modules. When the campus network is deployed in an environment consistent with the best practices recommended in this design guide, network recovery is transparent to end users and applications, as the other chassis in the virtual switch domain continues to provide constant network availability. Bundling the parallel

physical paths simplifies topology and forwarding paths. Thus, even during virtual switch chassis reset, there is no major topology re-computation or re-programming required on the Cisco VSS or on the peer multihomed network devices.

The eFSU software upgrade procedure becomes more graceful and transparent in Cisco VSS and the MEC-based network topologies, which offer the ability to upgrade the system within a consistent window of time. Figure 36 illustrates the amount of traffic loss during the **issu loadversion**, **issu runversion**, and **issu commitversion** process.

Figure 36 Network Recovery with eFSU Software Upgrade



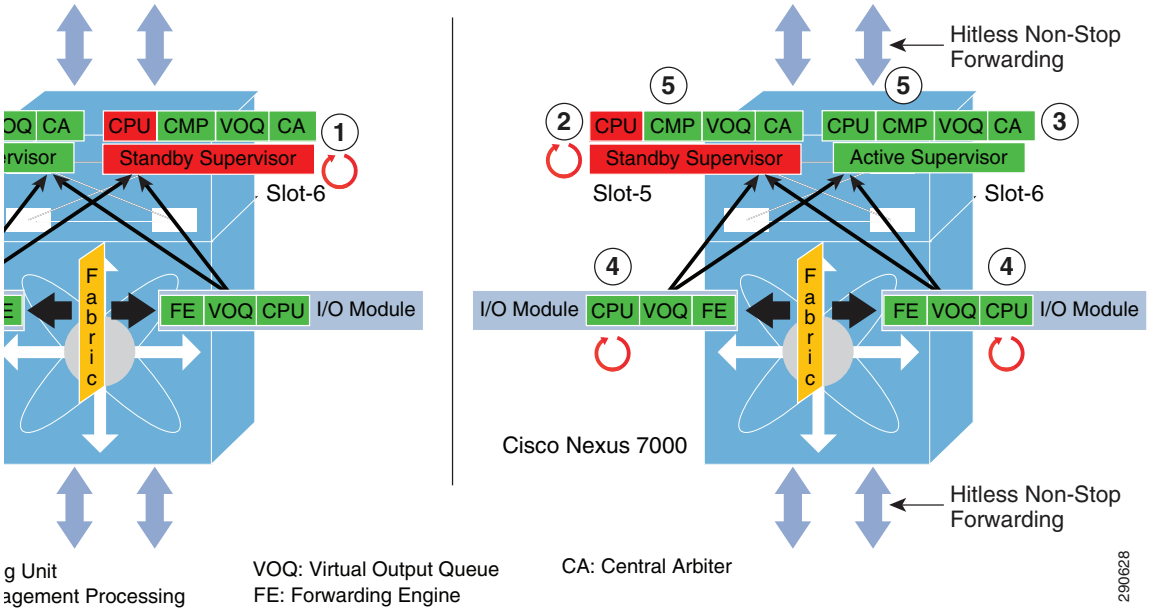
Nexus 7000 ISSU Software Design and Upgrade Process

The ISSU software upgrade process on the Cisco Nexus 7000 running the NX-OS operating system is different than on IOS. Each Nexus 7000 system component is intelligent in design with local hardware and software resources, such as CPU, memory for internal communication, and synchronization. For a common distributed software capability across all distributed hardware, the Nexus 7000 installs a consistent and common software version across the system. The NX-OS software image is bundled and compressed with the system software image, the linecard image to install on each I/O module, and the Connectivity Management Processor (CMP) BIOS and software image. The NX-OS ISSU software provides an option to manually upgrade each component separately or all at once.

The ISSU NX-OS software upgrade process in the Nexus 7000 system offers many benefits:

- Simplified—The network administrator can upgrade redundant supervisor modules, CMP, and all installed I/O modules with a single command or manually upgrade each component incrementally.
- Automated—Once initiated by the user, the Nexus 7000 system goes through five automatic NX-OS upgrade steps without user involvement. Each upgrade step ensures that the preceding step was successful and automatically proceeds through the following steps to upgrade all system components.
- Reliable—Assures the user that upgrading the NX-OS software will be non-disruptive. The Nexus 7000 performs hardware and software compatibility, integrity, and capability checks to prevent an upgrade failure. The system generates an impact report that provides detailed information about the software upgrade process on each distributed module. The Nexus 7000 aborts the upgrade process and automatically reverts to the previous software version if it detects any upgrade failures.
- Hitless software upgrade—Leveraging the distributed forwarding architecture, NSF/SSO technology, and graceful software upgrade design, the entire five-step ISSU upgrade process in the Nexus 7000 is non-disruptive to borderless services and hitless in switching campus core data traffic without dropping a single packet. The software upgrade procedure on each I/O module is also graceful, while data forwarding remains non-disruptive and the active supervisor and I/O CPU protect and restore distributed software applications and run-time data information.

Figure 1-37 Nexus 7000 Hitless ISSU Upgrade Process



The following steps briefly describe the system and network states during a hitless ISSU software upgrade on a Nexus 7000 system as shown in Figure 1-37:

1. The user initiates the NX-OS software process with a single command. The Nexus 7000 system generates a system impact report to verify if the user wants to accept or abort the upgrade procedure. If the user confirms the upgrade, the system synchronizes the kickstart and system image on the standby supervisor module (slot-6), extracts the image from the bundled new software, and upgrades the bios/bootloader/bootrom on each hardware component. On completion, the standby supervisor module is reset to install the new NX-OS software version.
2. The new boot variable gets updated on the current active supervisor module (slot-5) and forces a self-reset to install the new NX-OS software version.
3. This automatic upgrade step performs two key tasks in the system—SSO role switchover and making the new NX-OS software in effect on the new active supervisor module (slot-6). The Layer 3 network protocols perform graceful recovery with neighbor systems and maintain forwarding plane information for continuous data forwarding. This step remains graceful and non-disruptive with complete hitless switchover.
4. The system initializes a graceful software upgrade on the linecard CPU of each module. The I/O module upgrade occurs incrementally, one module at a time. This software design ensures that the new installed linecard software version is non-disruptive, reliable, and successful prior to proceeding to the next I/O module.

5. The final ISSU upgrade step is upgrading CMP on both supervisor modules. The CMP runs a subset of the operating system that operates independently from the NX-OS running on the supervisor module. The CMP software gets installed if the bundled version is the most recent compared to the currently installed version. This step may not proceed if the current CMP software version is the same or older. Since CMP software operates independently of the system, the version may not be the same as NX-OS.

Nexus 7000 ISSU Software Upgrade Procedure

This subsection provides guidelines to gracefully upgrade software and best practice to consider when upgrading NX-OS on the Cisco Nexus 7000 system.

Prior to starting the NX-OS software upgrade procedure on the Cisco Nexus 7000, the network administrator must make sure that the following prerequisites and guidelines for a graceful system upgrade are followed:

- To gracefully upgrade both supervisor modules, Cisco highly recommends that both Sup-1 supervisor modules in the Nexus 7000 chassis meet the following requirements:
  - Install a common new software version on both supervisor modules.
  - The new NX-OS software version must be copied to local storage (e.g., slot0, bootflash) of the supervisor module.
  - Both supervisor modules are in a a full operational and SSO redundant state.
- Inspect the impact report generated as part of the upgrade process to ensure a non-disruptive and hitless software upgrade process.
- Do not insert, swap, or remove the supervisor, crossbar fabric, or I/O module during the ISSU upgrade procedure.
- Use a direct console or CMP connection and login with network-admin privileges on both the active and standby supervisors during the entire ISSU upgrade process.
- During the ISSU software upgrade cycle, active supervisor role switchover will occur. By default NSF capability is enabled for Layer 3 protocol on the Nexus 7000. It is recommended to ensure that the neighbor system is NSF-aware and supports a compatible NSF protocol capability in the network as per the recommendations in this document.
- Save the system configuration to startup-config, local storage, or in a remote location (e.g., TFTP/FTP).
- Do not manually modify any BOOT variables or strings.

The following subsection demonstrate the entire NX-OS software upgrade procedure and sample output that follows the recommended best practices to install a new software version on the Cisco Nexus 7000 system running version 5.0.5 and upgrading to version 5.1.1a.

Preparing for NX-OS Software Upgrade

Prior to initializing the ISSU software upgrade procedure, the network administrator must prepare the Nexus 7000 system with the proper installation and validation to prevent services disruption and upgrade failures. In the sample output below, the Nexus 7000 system is equipped with dual redundant supervisor modules and M108 I/O network modules. The supervisor module is operating in SSO redundant mode and running system image 5.0(5); the I/O network module is running the same version of the linecard image as the supervisor module.

```
cr35-N7K-Core1#show module
Mod  Ports  Module-Type                        Model              Status
---  -----  --------------------------------- ------------------ ------------
1    8      10 Gbps Ethernet XL Module      N7K-M108X2-12L     ok
2    8      10 Gbps Ethernet XL Module      N7K-M108X2-12L     ok
5    0      Supervisor module-1X             N7K-SUP1                 *
6    0      Supervisor module-1X             N7K-SUP1           ha-standby

Mod  Sw              Hw
---  ------------  ------
1    5.0(5)          1.1
2    5.0(5)          1.1
5    5.0(5)          1.8
6    5.0(5)          1.6

cr35-N7K-Core1#show version | inc "CMP version"
  CMP Image version:       5.0(2) [build 5.0(0.66)]
  CMP Image version:       5.0(2) [build 5.0(0.66)]
```

Copy the new NX-OS system and kickstart software images on the local storage of both supervisor modules. If the new software is copied on compact flash, then it is recommended to not remove or swap until the system is successfully upgraded with the new NX-OS software version.

```
cr35-N7K-Core1# dir bootflash://sup-1 | inc 5.1.1a.bin
  145433028    Mar 03 21:52:15 2011  n7000-s1-dk9.5.1.1a.bin
   30484992    Dec 16 20:02:47 2010  n7000-s1-kickstart.5.1.1a.bin
cr35-N7K-Core1# dir bootflash://sup-2 | inc 5.1.1a.bin
  145433028    Mar 05 09:53:31 2011  n7000-s1-dk9.5.1.1a.bin
   30484992    Dec 16 20:08:23 2010  n7000-s1-kickstart.5.1.1a.bin
```

Cisco recommends generating the upgrade impact report to assess if migrating to new targeted NX-OS software will be graceful and non-disruptive or if it will fail due to any particular hardware or software failure. This pre-determination step performs multiple hardware and software integrity checks on each installed system component. The report indicates which system component will be upgraded from the current software version and how gracefully the new software version becomes effective in the system.

The user must enter the following syntax to generate an impact report. This step does **not** initialize the upgrade process; it performs an internal hardware and software verification procedure with the new software version and generates a detailed impact report:

```
cr35-N7K-Core1#show install all impact system bootflash:/n7000-s1-dk9.5.1.1a.bin kickstart
bootflash:/n7000-s1-kickstart.5.1.1a.bin
!Step 1 - Verify the boot variable for kickstart and system image.
Verifying image bootflash:/n7000-s1-kickstart.5.1.1a.bin for boot variable "kickstart".
[####################] 100% -- SUCCESS
<snip>

!Step 2 - Decompress all bundled software from new version (kickstart, system, cmp,
linecard & bios)
Extracting "lc1n7k" version from image bootflash:/n7000-s1-dk9.5.1.1a.bin.
[####################] 100% -- SUCCESS
<snip>

!Step 3 - Verify the new software compatibility with all installed I/O modules
Performing module support checks.
[####################] 100% -- SUCCESS

!Step 4 - Active supervisor sends new software upgrade signal to each distributed system
components.
Notifying services about system upgrade.
[####################] 100% -- SUCCESS

!Step 5 - Generate new NX-OS system impact report with ISSU upgrade. The first following
table briefly describes if new targeted software will be disruptive or non-disruptive. It
also describes the software installation process - supervisor reset (graceful switchover)
and I/O module rolling (incremental I/O upgrade procedure). The second table describes
which hardware components will be upgraded and provides detail information about current
and new software version information.

Compatibility check is done:
Module  bootable     Impact  I          Install-type  Reason
------  --------  --------------      ------------  ------
     1       yes  non-disruptive        rolling
     2       yes  non-disruptive        rolling
     5       yes  non-disruptive        reset
     6       yes  non-disruptive        reset

Module    Image     Running-Version(pri:alt)                New-Version       Upg-Required
------  ----------  ------------------------                ------------      ------------
1       lc1n7k      5.0(5)                                  5.1(1a)               yes
1        bios       v1.10.14(04/02/10):v1.10.14(04/02/10)   v1.10.14(04/02/10)    no
2       lc1n7k      5.0(5)                                  5.1(1a)               yes
2        bios       v1.10.14(04/02/10):v1.10.14(04/02/10)   v1.10.14(04/02/10)    no
5       system      5.0(5)                                  5.1(1a)               yes
5       kickstart   5.0(5)                                  5.1(1a)               yes
5        bios       v3.22.0(02/20/10):v3.22.0(02/20/10)     v3.22.0(02/20/10      no
5       cmp         5.0(2)                                  5.1(1)                yes
5       cmp-bios    02.01.05                                02.01.05              no
6       system      5.0(5)                                  5.1(1a)               yes
```

```
6       kickstart   5.0(5)                                  5.1(1a)             yes
6        bios       v3.22.0(02/20/10):v3.22.0(02/20/10)     v3.22.0(02/20/10    no
6       cmp         5.0(2)                                  5.1(1)              yes
6       cmp-bios    02.01.05                                02.01.05            no
```

Initiating NX-OS Software Upgrade

After confirming a non-disruptive and graceful upgrade based on the new software report, the network administrator can proceed to initiate the simplified NX-OS upgrade procedure on the Cisco Nexus 7000 system. As described earlier, the ISSU upgrade process on the Nexus 7000 is performed with a single user-initiated command. If the system detects any hardware or software failure during the ISSU upgrade process, the Nexus 7000 automatically aborts the upgrade procedure. Even though the upgrade process is completely automated, it is recommended to be vigilent during the process. Any upgrade failures or abnormal errors must be recorded during the process to ensure the new NX-OS software is installed successfully on each hardware component.

Once the user initiates the upgrade procedure, the Nexus 7000 system will go through five installation steps to upgrade each distributed hardware component, as illustrated in Figure 1-37:

1. Initiate the single-step NX-OS software installation process with the following:

```
cr35-N7K-Core1#install all system bootflash:///n7000-s1-dk9.5.1.1a.bin kickstart
bootflash:///n7000-s1-kickstart.5.1.1a.bin
```

This step repeats the same software initialization, extraction, and cross-system verification process as performed in generating the impact table. Once the report is generated, the network administrator must carefully review the upgrade impact and confirm to proceed with installation or abort.

```
<snip>
6    cmp-bios                                  02.01.05             02.01.05
no
Do you want to continue with the installation (y/n)?  [n] Y
```

The first step starts upgrading the bios, bootloader, and bootrom on each I/O and supervisor module. Once the basic software upgrade completes, the boot variable is modified and the standby supervisor module is force reset to boot with the new NX-OS software image.

```
Install is in progress, please wait.

Module 1: Refreshing compact flash and upgrading bios/loader/bootrom.
Warning: please do not remove or power off the module at this time.
[####################] 100% -- SUCCESS

<snip>

%$ VDC-1 %$ %PLATFORM-2-MOD_REMOVE: Module 6 removed (Serial number JAF1432AERD)
```

```
%$ VDC-1 %$ %CMPPROXY-STANDBY-2-LOG_CMP_WENT_DOWN: Connectivity Management processor
(on module 6) went DOWN


Module 6: Waiting for module online.
 -- SUCCESS        <- Standby supervisor online and successfully upgrade

%CMPPROXY-STANDBY-2-LOG_CMP_UP: Connectivity Management processor(on module 6) is now
UP
```

2. The active supervisor gets ready to gracefully reset. It notifies system components about its active role switchover and resets to boot up with the new NX-OS software. After reboot, it changes its role to standby and re-synchronizes with the new active supervisor module.

```
Notifying services about the switchover.
[####################] 100% -- SUCCESS

%MODULE-5-STANDBY_SUP_OK: Supervisor 5 is standby <- The supervisor is successfully
upgraded and reboots in Standby mode
%CMPPROXY-STANDBY-2-LOG_CMP_UP: Connectivity Management processor(on module 5) is now
UP
```

3. This step triggers two important changes in the Nexus 7000 system:
   - The standby supervisor takes over active ownership and performs graceful protocol recovery with neighbors.
   - The newly-installed NX-OS software becomes effective in the Nexus 7000 system.

```
%MODULE-5-STANDBY_SUP_OK: Supervisor 6 is standby   <- Pre-Switchover slot-6
Supervisor role

  %SYSMGR-2-HASWITCHOVER_PRE_START: This supervisor is becoming active (pre-start
phase). <- Slot-5 supervisor got reset and slot-6 supervisor taking over active role.
  %SYSMGR-2-HASWITCHOVER_START: Supervisor 6 is becoming active.
  %SYSMGR-2-SWITCHOVER_OVER: Switchover completed. <- SSO Switchover successfully
complete

  IP-EIGRP(0) 100: Neighbor 10.125.21.1 (port-channel100) is up: new adjacency <-
Gracefully EIGRP adjacency recovered

!Following graceful EIGRP adjacency synch message from EIGRP neighbor system
IP-EIGRP(0) 100: Neighbor 10.125.21.0 (port-channel100) is resync: peer
graceful-restart
```

4. After upgrading and activating the system image on both supervisor modules, in the next step the Nexus 7000 automatically initializes the I/O module upgrade process. Each I/O modules runs a linecard image on its local CPU for various different types of software applications. For consistent internal system communication, each I/O module must be upgraded to the same software version

as the system image running on the supervisor. All internal and external state machines and dynamic forwarding information remains intact on the distributed forwarding engine and other components of the I/O modules to provide non-stop communication. While the linecard CPU is gracefully installing and resetting itself to make the new software version effective, the I/O module remains fully operational and in service to provide a hitless software upgrade.

```
 Module 1: Non-disruptive upgrading.
  -- SUCCESS   <- The CPU on I/O Module in slot-1 successfully upgraded
 Module 2: Non-disruptive upgrading.
  -- SUCCESS   <- The CPU on I/O Module in slot-2 successfully upgraded
```

5. The final automated upgrade step is the CMP software upgrade process. The Nexus 7000 automatically upgrades CMP complexes on each supervisor module. This step may become optional if CMP is running a software version that is more current than the bundle version. The new CMP software version does not become effective after installation. It becomes effective on the next supervisor or system reset. The user may manually reboot the CMP complex to immediately put the new software into effect.

```
Module 6: Upgrading CMP image.
 Warning: please do not reload or power cycle CMP module at this time.
  -- SUCCESS
 Module 5: Upgrading CMP image.
 Warning: please do not reload or power cycle CMP module at this time.
  -- SUCCESS

 Recommended action::
 "Please reload CMP(s) manually to have it run in the newer version." <- Reload CMP
manually to immediately run new software version.

 Install has been successful. <- ISSU Software Upgrade finish

!Reload CMP complex on both supervisor module
cr35-N7K-Core1# reload cmp module 5
This command will reload the CMP on the supervisor in slot 5.  Continue (y/n)?  [no] Y

cr35-N7K-Core1# reload cmp module 6
This command will reload the CMP on the supervisor in slot 6.  Continue (y/n)?  [no] Y

!Verify new upgrades software status in Nexus 7000 system
cr35-N7K-Core1# show version | inc version
  BIOS:      version 3.22.0
  kickstart: version 5.1(1a)
  system:    version 5.1(1a)
  System version: 5.0(5)
  CMP BIOS version:        02.01.05
  CMP Image version:       5.1(1) [build 5.0(0.66)]
  CMP BIOS version:        02.01.05
  CMP Image version:       5.1(1) [build 5.0(0.66)]
```

Nexus 7000 Network Recovery with ISSU Software Upgrade

The distributed forwarding design and resilient software architecture of the Cisco Nexus 7000 system provides a hitless upgrade, thereby reducing the need for a maintenance window to install new software versions. In a hitless software upgrade design, campus backbone availability and switching capacity remain non-disruptive and intact while the Nexus 7000 is rolled out with a new NX-OS software image. The Cisco Nexus 7000 system had zero packet loss in several successful software upgrades in various network designs. Upgrading the Cisco Nexus 7000 system with the recommended procedure and best practices helps ensure a successful software upgrade and minimizes impact to network services.

10 Summary

Cisco Borderless Networks, a Cisco next-generation architecture, delivers a new workspace experience, securely, reliably, and smoothly connecting anyone, anywhere, using any device, to any resource. This borderless experience is only possible with a strong and resilient intelligent network that is designed to meet the needs of a global workspace. The Cisco-enabled network platform is the primary component of this architecture, providing borderless services such as mobility, security, medianet, location, and EnergyWise, to deliver an optimal user experience. Building network designs with intelligence at the edge provides mobility and secure collaboration, as well as the overall infrastructure backbone to provide network-wide differentiated services for a consistent, highly-available, and reliable user experience. Cisco Borderless Networks enable innovative business models, creating new user experiences that lead to increased customer satisfaction and loyalty.