



Cisco Data Center Infrastructure 2.1 Design Guide

Corporate Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
<http://www.cisco.com>
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

Text Part Number: OL-11565-01



THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS IN THIS MANUAL ARE SUBJECT TO CHANGE WITHOUT NOTICE. ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS MANUAL ARE BELIEVED TO BE ACCURATE BUT ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. USERS MUST TAKE FULL RESPONSIBILITY FOR THEIR APPLICATION OF ANY PRODUCTS.

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

The Cisco implementation of TCP header compression is an adaptation of a program developed by the University of California, Berkeley (UCB) as part of UCB's public domain version of the UNIX operating system. All rights reserved. Copyright © 1981, Regents of the University of California.

NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED "AS IS" WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

CCIP, CCSP, the Cisco Arrow logo, the Cisco *Powered* Network mark, the Cisco Systems Verified logo, Cisco Unity, Follow Me Browsing, FormShare, iQ Net Readiness Scorecard, Networking Academy, and ScriptShare are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn, The Fastest Way to Increase Your Internet Quotient, and iQuick Study are service marks of Cisco Systems, Inc.; and Aironet, ASIST, BPX, Catalyst, CCDA, CCDP, CCIE, CCNA, CCNP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, the Cisco IOS logo, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Empowering the Internet Generation, Enterprise/Solver, EtherChannel, EtherSwitch, Fast Step, GigaStack, Internet Quotient, IOS, IP/TV, iQ Expertise, the iQ logo, LightStream, MGX, MICA, the Networkers logo, Network Registrar, *Packet*, PIX, Post-Routing, Pre-Routing, RateMUX, Registrar, SlideCast, SMARTnet, StrataView Plus, Stratm, SwitchProbe, TeleRouter, TransPath, and VCO are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0303R)



CONTENTS

CHAPTER 1

Data Center Architecture Overview 1-1

Data Center Architecture Overview 1-1

Data Center Design Models 1-3

Multi-Tier Model 1-3

Server Cluster Model 1-5

HPC Cluster Types and Interconnects 1-6

Logical Overview 1-8

Physical Overview 1-9

CHAPTER 2

Data Center Multi-Tier Model Design 2-1

Data Center Multi-Tier Design Overview 2-2

Data Center Core Layer 2-3

Recommended Platform and Modules 2-3

Distributed Forwarding 2-3

Traffic Flow in the Data Center Core 2-4

Data Center Aggregation Layer 2-6

Recommended Platforms and Modules 2-6

Distributed Forwarding 2-8

Traffic Flow in the Data Center Aggregation Layer 2-8

Path Selection in the Presence of Service Modules 2-8

Server Farm Traffic Flow with Service Modules 2-10

Server Farm Traffic Flow without Service Modules 2-10

Scaling the Aggregation Layer 2-11

Layer 2 Fault Domain Size 2-12

Spanning Tree Scalability 2-13

10 GigE Density 2-13

Default Gateway Redundancy with HSRP 2-14

Data Center Access Layer 2-14

Recommended Platforms and Modules 2-17

Distributed Forwarding 2-18

Resiliency 2-18

Sharing Services at the Aggregation Layer 2-19

Data Center Services Layer 2-20

Recommended Platforms and Modules 2-20

Performance Implications	2-21
Traffic Flow through the Service Layer	2-22
Resiliency	2-24

CHAPTER 3

Server Cluster Designs with Ethernet 3-1

Technical Objectives	3-1
Distributed Forwarding and Latency	3-2
Catalyst 6500 System Bandwidth	3-3
Equal Cost Multi-Path Routing	3-4
Redundancy in the Server Cluster Design	3-5
Server Cluster Design—Two-Tier Model	3-6
4- and 8-Way ECMP Designs with Modular Access	3-7
2-Way ECMP Design with 1RU Access	3-9
Server Cluster Design—Three-Tier Model	3-10
Calculating Oversubscription	3-12
Recommended Hardware and Modules	3-12

CHAPTER 4

Data Center Design Considerations 4-1

Factors that Influence Scalability	4-1
Why Implement a Data Center Core Layer?	4-1
Why Use the Three-Tier Data Center Design?	4-2
Determining Maximum Servers	4-2
Determining Maximum Number of VLANs	4-3
Server Clustering	4-4
NIC Teaming	4-7
Pervasive 10GigE	4-8
Server Consolidation	4-9
Top of Rack Switching	4-10
Blade Servers	4-13
Importance of Team Planning	4-13

CHAPTER 5

Spanning Tree Scalability 5-1

Extending VLANs in the Data Center	5-1
STP Active Logical Ports and Virtual Ports per Line Card	5-2
Calculating the Active Logical Ports	5-4
Calculating Virtual Ports per Line Card	5-5
Steps to Resolve Logical Port Count Implications	5-6

CHAPTER 6**Data Center Access Layer Design 6-1**

- Overview of Access Layer Design Options 6-1
 - Service Module Influence on Design 6-3
 - Service Module/Appliance and Path Preferences 6-4
 - General Recommendations 6-5
- Layer 2 Looped Access Layer Model 6-6
 - Layer 2 Looped Access Topologies 6-6
 - Triangle Looped Topology 6-8
 - Spanning Tree, HSRP, and Service Module Design 6-8
 - Failure Scenarios 6-9
 - Square Looped Topology 6-12
 - Spanning Tree, HSRP, and Service Module Design 6-13
 - Failure Scenarios 6-14
 - Layer 2 Loop-Free Access Layer Model 6-17
 - Layer 2 Loop-Free Access Topologies 6-17
 - Layer 2 Loop-Free U Topology 6-19
 - Spanning Tree, HSRP, and Service Module Design 6-20
 - Failure Scenarios 6-20
 - Layer 2 Loop-Free Inverted U Topology 6-23
 - Spanning Tree, HSRP, and Service Module Design 6-24
 - Failure Scenarios 6-24
 - FlexLinks Access Model 6-28
 - Spanning Tree, HSRP, and Service Module Design 6-31
 - Implications Related to Possible Loop Conditions 6-31
 - Failure Scenarios 6-33
 - Using EtherChannel Min-Links 6-37

CHAPTER 7**Increasing HA in the Data Center 7-1**

- Establishing Path Preference with RHI 7-1
 - Aggregation 1 CSM Configuration 7-3
 - Aggregation 1 OSPF and Route Map Configurations 7-4
 - Aggregation Inter-switch Link Configuration 7-4
 - Aggregation 2 Route Map Configuration 7-5
- Service Module FT Paths 7-5
- NSF-SSO in the Data Center 7-6
 - Possible Implications 7-8
 - HSRP 7-8
 - IGP Timers 7-8

Slot Usage versus Improved HA	7-9
Recommendations	7-9

CHAPTER 8

Configuration Reference	8-1
Core Switch 1	8-2
Aggregation Switch 1	8-6
Core Switch 2	8-13
Aggregation Switch 2	8-16
Access Switch 4948-7	8-22
Access Switch 4948-8	8-24
Access Switch 6500-1	8-26
FWSM 1-Aggregation Switch 1 and 2	8-28
Additional References	8-32



Data Center Architecture Overview



Note

The README file posted with this guide contains details about the technologies, hardware, and software that were used in producing this document. The README file also contains a revision history that details updates made to each chapter.

This chapter is an overview of proven Cisco solutions for providing architecture designs in the enterprise data center, and includes the following topics:

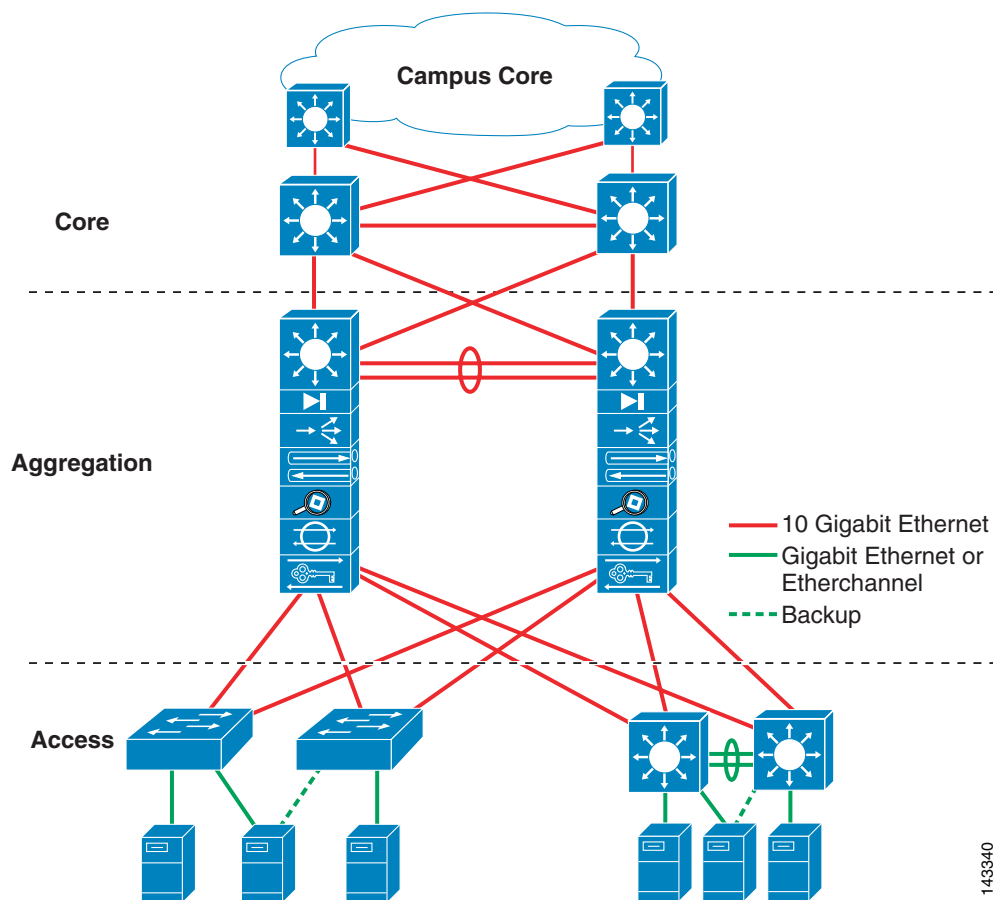
- [Data Center Architecture Overview](#)
- [Data Center Design Models](#)

Data Center Architecture Overview

The data center is home to the computational power, storage, and applications necessary to support an enterprise business. The data center infrastructure is central to the IT architecture, from which all content is sourced or passes through. Proper planning of the data center infrastructure design is critical, and performance, resiliency, and scalability need to be carefully considered.

Another important aspect of the data center design is flexibility in quickly deploying and supporting new services. Designing a flexible architecture that has the ability to support new applications in a short time frame can result in a significant competitive advantage. Such a design requires solid initial planning and thoughtful consideration in the areas of port density, access layer uplink bandwidth, true server capacity, and oversubscription, to name just a few.

The data center network design is based on a proven *layered* approach, which has been tested and improved over the past several years in some of the largest data center implementations in the world. The layered approach is the basic foundation of the data center design that seeks to improve scalability, performance, flexibility, resiliency, and maintenance. [Figure 1-1](#) shows the basic layered design.

Figure 1-1 Basic Layered Design

The layers of the data center design are the *core*, *aggregation*, and *access* layers. These layers are referred to extensively throughout this guide and are briefly described as follows:

- **Core layer**—Provides the high-speed packet switching backplane for all flows going in and out of the data center. The core layer provides connectivity to multiple aggregation modules and provides a resilient Layer 3 routed fabric with no single point of failure. The core layer runs an interior routing protocol, such as OSPF or EIGRP, and load balances traffic between the campus core and aggregation layers using Cisco Express Forwarding-based hashing algorithms.
- **Aggregation layer modules**—Provide important functions, such as service module integration, Layer 2 domain definitions, spanning tree processing, and default gateway redundancy. Server-to-server multi-tier traffic flows through the aggregation layer and can use services, such as firewall and server load balancing, to optimize and secure applications. The smaller icons within the aggregation layer switch in [Figure 1-1](#) represent the integrated service modules. These modules provide services, such as content switching, firewall, SSL offload, intrusion detection, network analysis, and more.
- **Access layer**—Where the servers physically attach to the network. The server components consist of 1RU servers, blade servers with integral switches, blade servers with pass-through cabling, clustered servers, and mainframes with OSA adapters. The access layer network infrastructure consists of modular switches, fixed configuration 1 or 2RU switches, and integral blade server switches. Switches provide both Layer 2 and Layer 3 topologies, fulfilling the various server broadcast domain or administrative requirements.

This chapter defines the framework on which the recommended data center architecture is based and introduces the primary data center design models: the *multi-tier* and *server cluster* models.

Data Center Design Models

The *multi-tier* model is the most common design in the enterprise. It is based on the web, application, and database layered design supporting commerce and enterprise business ERP and CRM solutions. This type of design supports many web service architectures, such as those based on Microsoft .NET or Java 2 Enterprise Edition. These web service application environments are used by ERP and CRM solutions from Siebel and Oracle, to name a few. The multi-tier model relies on security and application optimization services to be provided in the network.

The *server cluster* model has grown out of the university and scientific community to emerge across enterprise business verticals including financial, manufacturing, and entertainment. The server cluster model is most commonly associated with high-performance computing (HPC), parallel computing, and high-throughput computing (HTC) environments, but can also be associated with grid/utility computing. These designs are typically based on customized, and sometimes proprietary, application architectures that are built to serve particular business objectives.

Although server clusters are dominated today by the Linux operating system, a Microsoft Windows-based HPC product (Windows Server 2003 Compute Cluster Edition) is planned for shipment in the 2005-2006 timeframe. Microsoft reports that approximately 20 percent of enterprises are now experimenting with HPC applications. It is expected that enterprise businesses will continue to embrace HPC technology, and server cluster implementations will become more mainstream.

[Chapter 2, “Data Center Multi-Tier Model Design,”](#) provides an overview of the multi-tier model, and [Chapter 3, “Server Cluster Designs with Ethernet,”](#) provides an overview of the server cluster model. Later chapters of this guide address the design aspects of these models in greater detail.

Multi-Tier Model

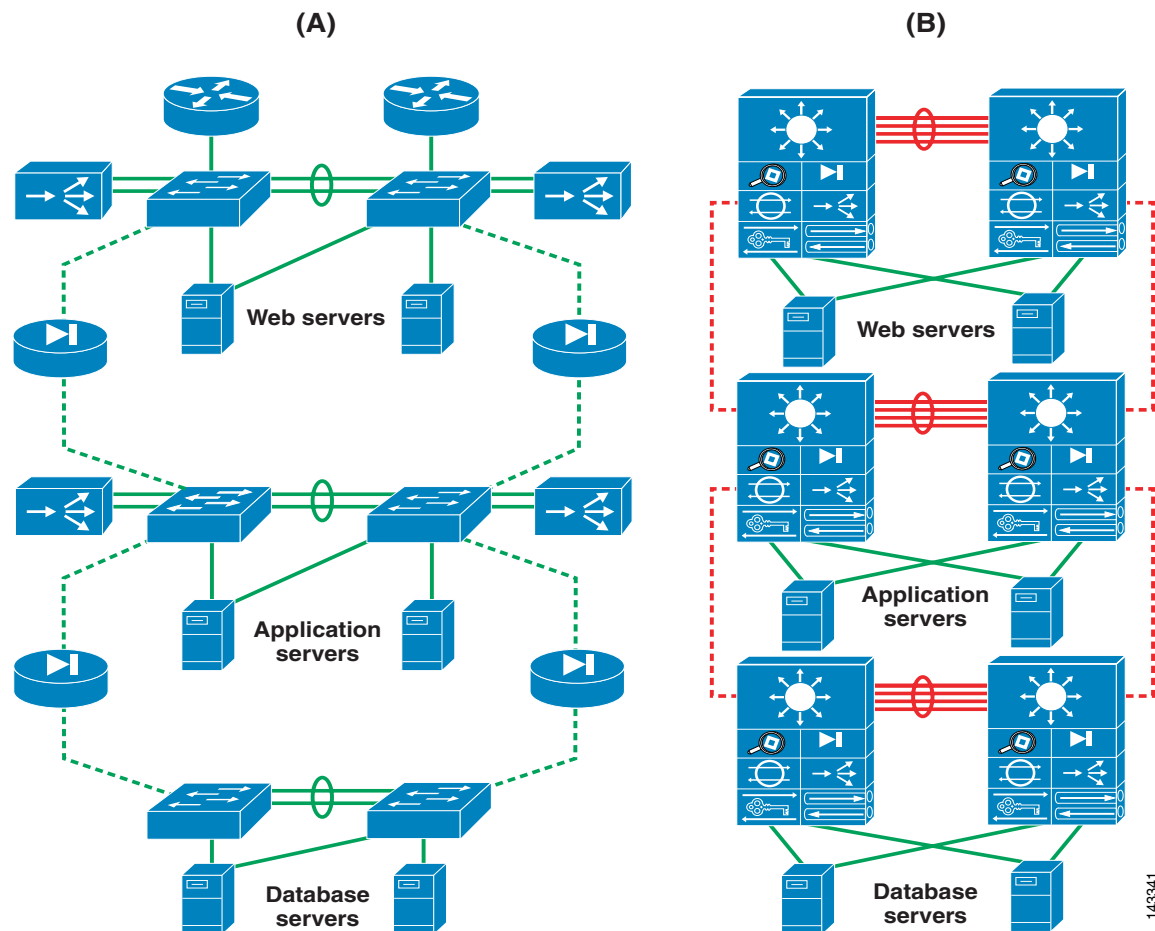
The multi-tier data center model is dominated by HTTP-based applications in a multi-tier approach. The multi-tier approach includes web, application, and database tiers of servers. Today, most web-based applications are built as multi-tier applications. The multi-tier model uses software that runs as separate processes on the same machine using interprocess communication (IPC), or on different machines with communications over the network. Typically, the following three tiers are used:

- Web-server
- Application
- Database

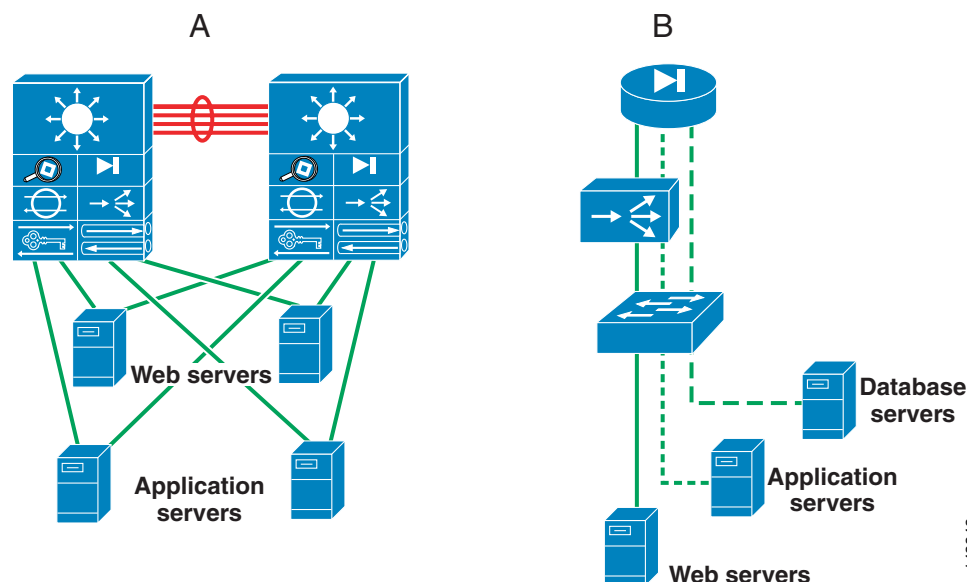
Multi-tier server farms built with processes running on separate machines can provide improved resiliency and security. Resiliency is improved because a server can be taken out of service while the same function is still provided by another server belonging to the same application tier. Security is improved because an attacker can compromise a web server without gaining access to the application or database servers. Web and application servers can coexist on a common physical server; the database typically remains separate.

Resiliency is achieved by load balancing the network traffic between the tiers, and security is achieved by placing firewalls between the tiers. You can achieve segregation between the tiers by deploying a separate infrastructure composed of aggregation and access switches, or by using VLANs (see [Figure 1-2](#)).

Figure 1-2 Physical Segregation in a Server Farm with Appliances (A) and Service Modules (B)



The design shown in [Figure 1-3](#) uses VLANs to segregate the server farms. The left side of the illustration (A) shows the physical topology, and the right side (B) shows the VLAN allocation across the service modules, firewall, load balancer, and switch. The firewall and load balancer, which are VLAN-aware, enforce the VLAN segregation between the server farms. Note that not all of the VLANs require load balancing. For example, the database in the example sends traffic directly to the firewall.

Figure 1-3 Logical Segregation in a Server Farm with VLANs

Physical segregation improves performance because each tier of servers is connected to dedicated hardware. The advantage of using logical segregation with VLANs is the reduced complexity of the server farm. The choice of physical segregation or logical segregation depends on your specific network performance requirements and traffic patterns.

Business security and performance requirements can influence the security design and mechanisms used. For example, the use of wire-speed ACLs might be preferred over the use of physical firewalls. Non-intrusive security devices that provide detection and correlation, such as the Cisco Monitoring, Analysis, and Response System (MARS) combined with Route Triggered Black Holes (RTBH) and Cisco Intrusion Protection System (IPS) might meet security requirements. Cisco Guard can also be deployed as a primary defense against distributed denial of service (DDoS) attacks. For more details on security design in the data center, refer to the *Server Farm Security SRND* at the following URL: http://www.cisco.com/en/US/netsol/ns656/networking_solutions_design_guidances_list.html#anchor3.

Server Cluster Model

In the modern data center environment, clusters of servers are used for many purposes, including high availability, load balancing, and increased computational power. This guide focuses on the high performance form of clusters, which includes many forms. All clusters have the common goal of combining multiple CPUs to appear as a unified high performance system using special software and high-speed network interconnects. Server clusters have historically been associated with university research, scientific laboratories, and military research for unique applications, such as the following:

- Meteorology (weather simulation)
- Seismology (seismic analysis)
- Military research (weapons, warfare)

Server clusters are now in the enterprise because the benefits of clustering technology are now being applied to a broader range of applications. The following applications in the enterprise are driving this requirement:

- Financial trending analysis—Real-time bond price analysis and historical trending
- Film animation—Rendering of artist multi-gigabyte files
- Manufacturing—Automotive design modeling and aerodynamics
- Search engines—Quick parallel lookup plus content insertion

In the enterprise, developers are increasingly requesting higher bandwidth and lower latency for a growing number of applications. The time-to-market implications related to these applications can result in a tremendous competitive advantage. For example, the cluster performance can directly affect getting a film to market for the holiday season or providing financial management customers with historical trending information during a market shift.

HPC Cluster Types and Interconnects

In the high performance computing landscape, various HPC cluster types exist and various interconnect technologies are used. The top 500 supercomputer list at www.top500.org provides a fairly comprehensive view of this landscape. The majority of interconnect technologies used today are based on Fast Ethernet and Gigabit Ethernet, but a growing number of specialty interconnects exist, for example including Infiniband and Myrinet. Specialty interconnects such as Infiniband have very low latency and high bandwidth switching characteristics when compared to traditional Ethernet, and leverage built-in support for Remote Direct Memory Access (RDMA). 10GE NICs have also recently emerged that introduce TCP/IP offload engines that provide similar performance to Infiniband.

The Cisco SFS line of Infiniband switches and Host Channel Adapters (HCAs) provide high performance computing solutions that meet the highest demands. For more information on Infiniband and High Performance Computing, refer to the following URL:
<http://www.cisco.com/en/US/products/ps6418/index.html>.

The remainder of this chapter and the information in [Chapter 3, “Server Cluster Designs with Ethernet”](#) focus on large cluster designs that use Ethernet as the interconnect technology.

Although high performance clusters (HPCs) come in various types and sizes, the following categorizes three main types that exist in the enterprise environment:

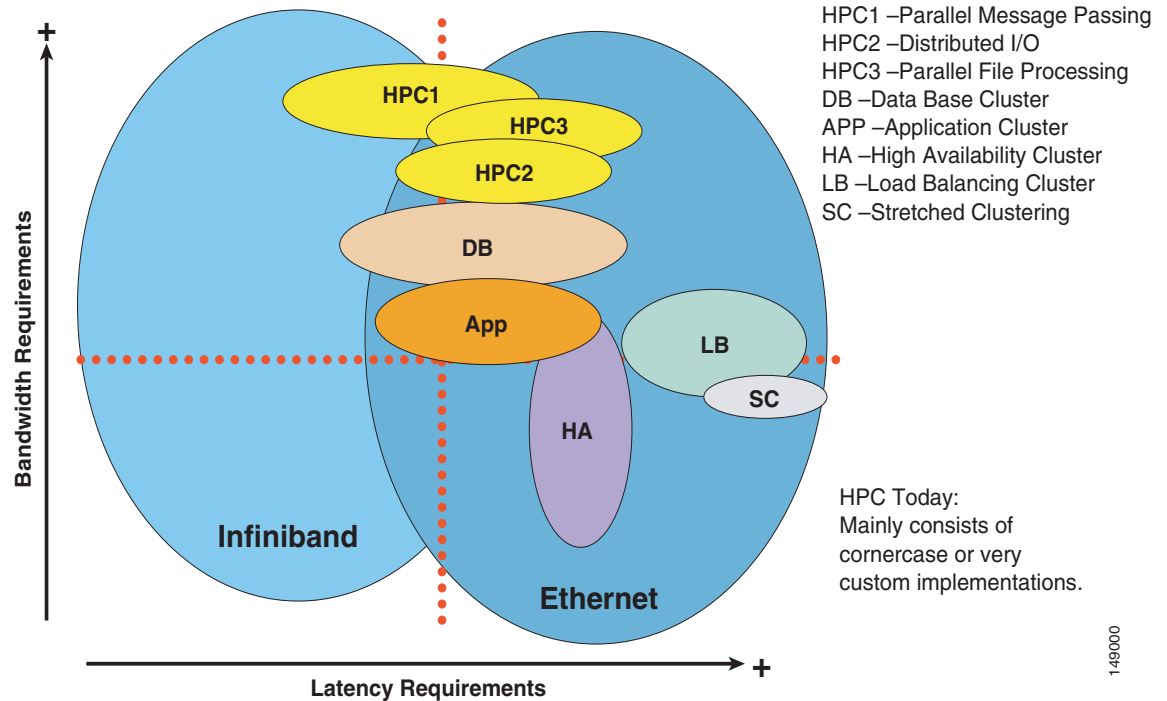
- HPC type 1—Parallel message passing (also known as tightly coupled)
 - Applications run on all compute nodes simultaneously in parallel.
 - A master node determines input processing for each compute node.
 - Can be a large or small cluster, broken down into hives (for example, 1000 servers over 20 hives) with IPC communication between compute nodes/hives.
- HPC type 2—Distributed I/O processing (for example, search engines)
 - The client request is balanced across master nodes, then sprayed to compute nodes for parallel processing (typically unicast at present, with a move towards multicast).
 - This type obtains the quickest response, applies content insertion (advertising), and sends to the client.
- HPC Type 3—Parallel file processing (also known as loosely coupled)
 - The source data file is divided up and distributed across the compute pool for manipulation in parallel. Processed components are rejoined after completion and written to storage.

- Middleware controls the job management process (for example, platform linear file system [LFS]).

The traditional high performance computing cluster that emerged out of the university and military environments was based on the type 1 cluster. The new enterprise HPC applications are more aligned with HPC types 2 and 3, supporting the entertainment, financial, and a growing number of other vertical industries.

Figure 1-4 shows the current server cluster landscape.

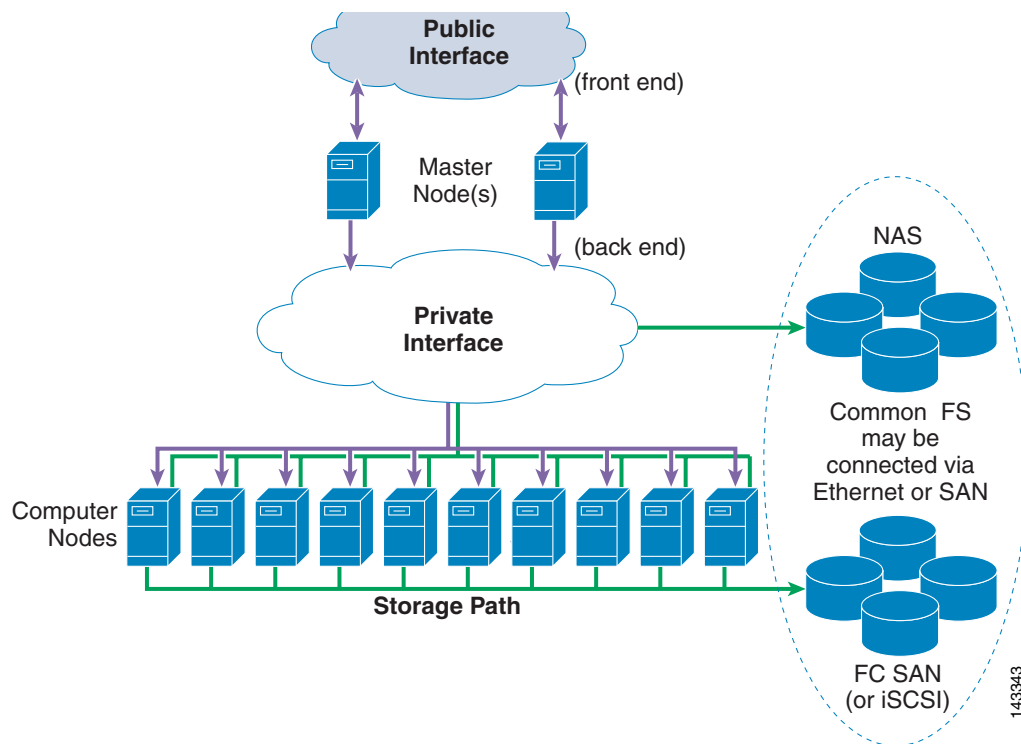
Figure 1-4 Server Cluster Landscape



The following section provides a general overview of the server cluster components and their purpose, which helps in understanding the design objectives described in [Chapter 3, “Server Cluster Designs with Ethernet.”](#)

Figure 1-5 shows a logical view of a server cluster.

Figure 1-5 Logical View of a Server Cluster



Logical Overview

The components of the server cluster are as follows:

- **Front end**—These interfaces are used for external access to the cluster, which can be accessed by application servers or users that are submitting jobs or retrieving job results from the cluster. An example is an artist who is submitting a file for rendering or retrieving an already rendered result. This is typically an Ethernet IP interface connected into the access layer of the existing server farm infrastructure.
- **Master nodes (also known as head node)**—The master nodes are responsible for managing the compute nodes in the cluster and optimizing the overall compute capacity. Usually, the master node is the only node that communicates with the outside world. Clustering middleware running on the master nodes provides the tools for resource management, job scheduling, and node state monitoring of the computer nodes in the cluster. Master nodes are typically deployed in a redundant fashion and are usually a higher performing server than the compute nodes.
- **Back-end high-speed fabric**—This high-speed fabric is the primary medium for master node to compute node and inter-compute node communications. Typical requirements include low latency and high bandwidth and can also include jumbo frame and 10 GigE support. Gigabit Ethernet is the most popular fabric technology in use today for server cluster implementations, but other technologies show promise, particularly Infiniband.

- **Compute nodes**—The compute node runs an optimized or full OS kernel and is primarily responsible for CPU-intense operations such as number crunching, rendering, compiling, or other file manipulation.
- **Storage path**—The storage path can use Ethernet or Fibre Channel interfaces. Fibre Channel interfaces consist of 1/2/4G interfaces and usually connect into a SAN switch such as a Cisco MDS platform. The back-end high-speed fabric and storage path can also be a common transport medium when IP over Ethernet is used to access storage. Typically, this is for NFS or iSCSI protocols to a NAS or SAN gateway, such as the IPS module on a Cisco MDS platform.
- **Common file system**—The server cluster uses a common parallel file system that allows high performance access to all compute nodes. The file system types vary by operating system (for example, PVFS or Lustre).

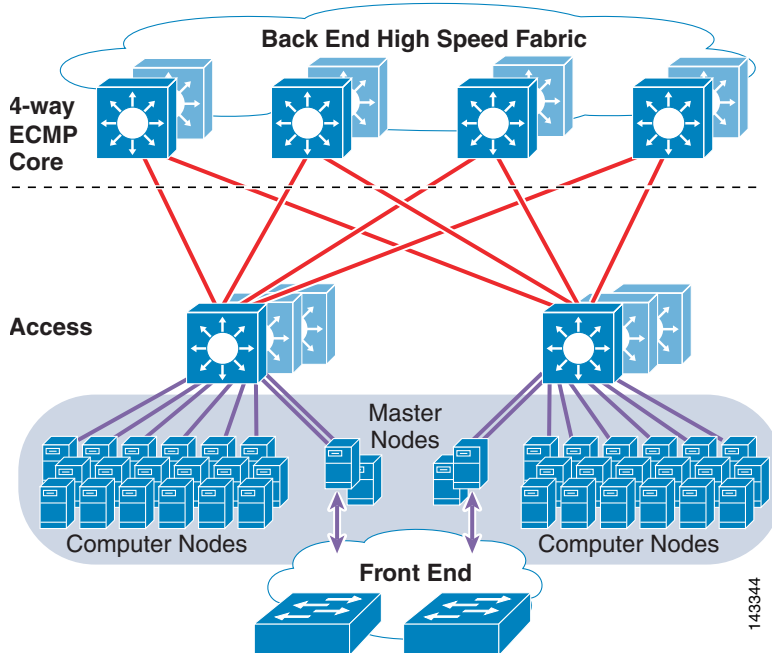
Physical Overview

Server cluster designs can vary significantly from one to another, but certain items are common, such as the following:

- **Commodity off the Shelf (CotS) server hardware**—The majority of server cluster implementations are based on 1RU Intel- or AMD-based servers with single/dual processors. The spiraling cost of these high performing 32/64-bit low density servers has contributed to the recent enterprise adoption of cluster technology.
- **GigE or 10 GigE NIC cards**—The applications in a server cluster can be bandwidth intensive and have the capability to burst at a high rate when necessary. The PCI-X or PCI-Express NIC cards provide a high-speed transfer bus speed and use large amounts of memory. TCP/IP offload and RDMA technologies are also used to increase performance while reducing CPU utilization.
- **Low latency hardware**—Usually a primary concern of developers is related to the message-passing interface delay affecting the overall cluster/application performance. This is not always the case because some clusters are more focused on high throughput, and latency does not significantly impact the applications. The Cisco Catalyst 6500 with distributed forwarding and the Catalyst 4948-10G provide consistent latency values necessary for server cluster environments.
- **Non-blocking or low-over-subscribed switch fabric**—Many HPC applications are bandwidth-intensive with large quantities of data transfer and interprocess communications between compute nodes. GE attached server oversubscription ratios of 2.5:1 (500 Mbps) up to 8:1 (125 Mbps) are common in large server cluster designs.
- **Mesh/partial mesh connectivity**—Server cluster designs usually require a mesh or partial mesh fabric to permit communication between all nodes in the cluster. This mesh fabric is used to share state, data, and other information between master-to-compute and compute-to-compute servers in the cluster.
- **Jumbo frame support**—Many HPC applications use large frame sizes that exceed the 1500 byte Ethernet standard. The ability to send large frames (called jumbos) that are up to 9K in size, provides advantages in the areas of server CPU overhead, transmission overhead, and file transfer time.

Figure 1-6 takes the logical cluster view and places it in a physical topology that focuses on addressing the preceding items.

Figure 1-6 Physical View of a Server Cluster Model Using ECMP



The recommended server cluster design leverages the following technical aspects or features:

- Equal cost multi-path—ECMP support for IP permits a highly effective load distribution of traffic across multiple uplinks between servers across the access layer. Although Figure 1-6 demonstrates a four-way ECMP design, this can scale to eight-way by adding additional paths.
- Distributed forwarding—By using distributed forwarding cards on interface modules, the design takes advantage of improved switching performance and lower latency.
- L3 plus L4 hashing algorithms—Distributed Cisco Express Forwarding-based load balancing permits ECMP hashing algorithms based on Layer 3 IP source-destination plus Layer 4 source-destination port, allowing a highly granular level of load distribution.
- Scalable server density—The ability to add access layer switches in a modular fashion permits a cluster to start out small and easily increase as required.
- Scalable fabric bandwidth—ECMP permits additional links to be added between the core and access layer as required, providing a flexible method of adjusting oversubscription and bandwidth per server.

In the preceding design, master nodes are distributed across multiple access layer switches to provide redundancy as well as to distribute load.

Further details on multiple server cluster topologies, hardware recommendations, and oversubscription calculations are covered in Chapter 3, “Server Cluster Designs with Ethernet.”



Data Center Multi-Tier Model Design



Note

The README file posted with this guide contains details about the technologies, hardware, and software that were used in producing this document. The README file also contains a revision history that details updates made to each chapter.

This chapter provides details about the multi-tier design that Cisco recommends for data centers. The multi-tier design model supports many web service architectures, including those based on Microsoft .NET and Java 2 Enterprise Edition. These web service application environments are used for common ERP solutions, such as those from PeopleSoft, Oracle, SAP, BAAN, and JD Edwards; and CRM solutions from vendors such as Siebel and Oracle.

The multi-tier model relies on a multi-layer network architecture consisting of *core*, *aggregation*, and *access* layers, as shown in [Figure 2-1](#). This chapter describes the hardware and design recommendations for each of these layers in greater detail. The following major topics are included:

- [Data Center Multi-Tier Design Overview](#)
- [Data Center Core Layer](#)
- [Data Center Aggregation Layer](#)
- [Data Center Access Layer](#)
- [Data Center Services Layer](#)



Note

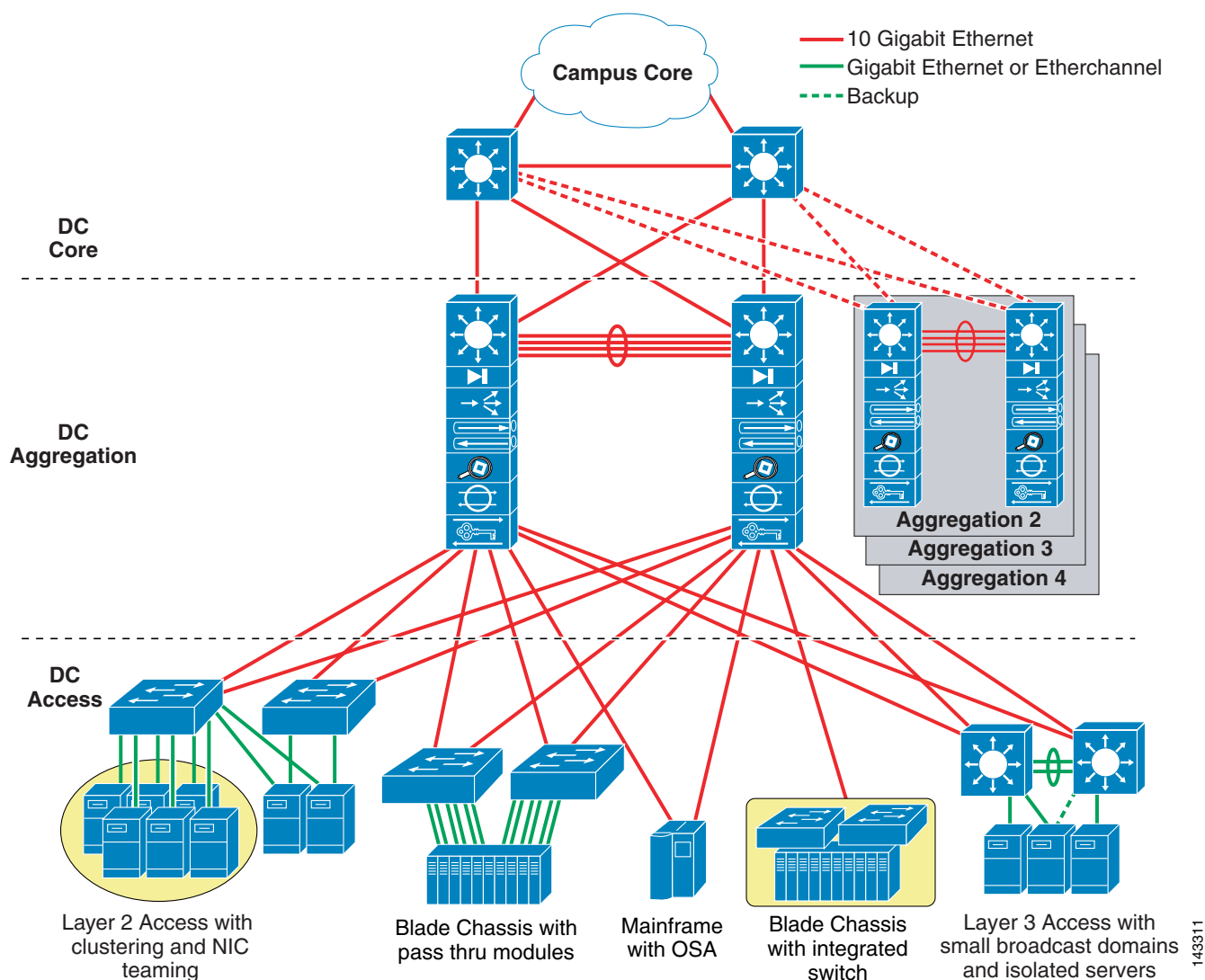
For a high-level overview of the multi-tier model, refer to [Chapter 1, “Data Center Architecture Overview.”](#)

Data Center Multi-Tier Design Overview

The multi-tier model is the most common model used in the enterprise today. This design consists primarily of web, application, and database server tiers running on various platforms including blade servers, one rack unit (1RU) servers, and mainframes.

Figure 2-1 shows the data center multi-tier model topology. Familiarize yourself with this diagram before reading the subsequent sections, which provide details on each layer of this recommended architecture.

Figure 2-1 Data Center Multi-Tier Model Topology



Data Center Core Layer

The data center core layer provides a fabric for high-speed packet switching between multiple aggregation modules. This layer serves as the gateway to the campus core where other modules connect, including, for example, the extranet, WAN, and Internet edge. All links connecting the data center core are terminated at Layer 3 and typically use 10 GigE interfaces for supporting a high level of throughput, performance, and to meet oversubscription levels.

The data center core is distinct from the campus core layer, with a different purpose and responsibilities. A data center core is not necessarily required, but is recommended when multiple aggregation modules are used for scalability. Even when a small number of aggregation modules are used, it might be appropriate to use the campus core for connecting the data center fabric.

When determining whether to implement a data center core, consider the following:

- Administrative domains and policies—Separate cores help isolate campus distribution layers and data center aggregation layers in terms of administration and policies, such as QoS, access lists, troubleshooting, and maintenance.
- 10 GigE port density—A single pair of core switches might not support the number of 10 GigE ports required to connect the campus distribution layer as well as the data center aggregation layer switches.
- Future anticipation—The business impact of implementing a separate data center core layer at a later date might make it worthwhile to implement it during the initial implementation stage.

Recommended Platform and Modules

In a large data center, a single pair of data center core switches typically interconnect multiple aggregation modules using 10 GigE Layer 3 interfaces.

The recommended platform for the enterprise data center core layer is the Cisco Catalyst 6509 with the Sup720 processor module. The high switching rate, large switch fabric, and 10 GigE density make the Catalyst 6509 ideal for this layer. Providing a large number of 10 GigE ports is required to support multiple aggregation modules. The Catalyst 6509 can support 10 GigE modules in all positions because each slot supports dual channels to the switch fabric (the Catalyst 6513 cannot support this). We do not recommend using non-fabric-attached (classic) modules in the core layer.

**Note**

By using all fabric-attached CEF720 modules, the global switching mode is *compact*, which allows the system to operate at its highest performance level.

The data center core is interconnected with both the campus core and aggregation layer in a redundant fashion with Layer 3 10 GigE links. This provides for a fully redundant architecture and eliminates a single core node from being a single point of failure. This also permits the core nodes to be deployed with only a single supervisor module.

Distributed Forwarding

The Cisco 6700 Series line cards support an optional daughter card module called a Distributed Forwarding Card (DFC). The DFC permits local routing decisions to occur on each line card via a local Forwarding Information Base (FIB). The FIB table on the Sup720 policy feature card (PFC) maintains synchronization with each DFC FIB table on the line cards to ensure accurate routing integrity across the system. Without a DFC card, a compact header lookup must be sent to the PFC on the Sup720 to

determine where on the switch fabric to forward each packet to reach its destination. This occurs for both Layer 2 and Layer 3 switched packets. When a DFC is present, the line card can switch a packet directly across the switch fabric to the destination line card without consulting the Sup720 FIB table on the PFC. The difference in performance can range from 30 Mpps system-wide to 48 Mpps *per slot* with DFCs.

With or without DFCs, the available system bandwidth is the same as determined by the Sup720 switch fabric. [Table 2-1](#) summarizes the throughput and bandwidth performance for modules that support DFCs and the older CEF256, in addition to classic bus modules for comparison.

Table 2-1 Performance Comparison with Distributed Forwarding

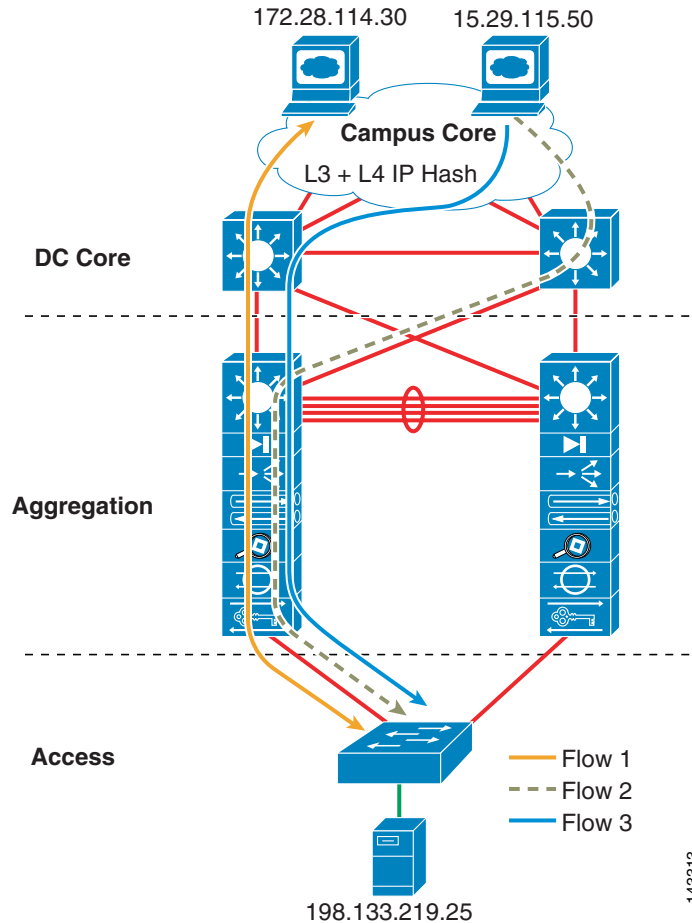
System Config with Sup720	Throughput in Mpps	Bandwidth in Gbps
CEF720 Series Modules (6748, 6704, 6724)	Up to 30 Mpps per system	2 x 20 Gbps (dedicated per slot) (6724=1 x 20 Gbps)
CEF720 Series Modules with DFC3 (6704 with DFC3, 6708 with DFC3, 6724 with DFC3)	Sustain up to 48 Mpps (per slot)	2x 20 Gbps (dedicated per slot) (6724=1 x 20 Gbps)
CEF256 Series Modules (FWSM, SSLSM, NAM-2, IDSM-2, 6516)	Up to 30 Mpps (per system)	1x 8 Gbps (dedicated per slot)
Classic Series Modules (CSM, 61xx-64xx)	Up to 15 Mpps (per system)	16 Gbps shared bus (classic bus)

Using DFCs in the core layer of the multi-tier model is optional. An analysis of application session flows that can transit the core helps to determine the maximum bandwidth requirements and whether DFCs would be beneficial. If multiple aggregation modules are used, there is a good chance that a large number of session flows will propagate between server tiers. Generally speaking, the core layer benefits with lower latency and higher overall forwarding rates when including DFCs on the line cards.

Traffic Flow in the Data Center Core

The core layer connects to the campus and aggregation layers using Layer 3-terminated 10 GigE links. Layer 3 links are required to achieve bandwidth scalability, quick convergence, and to avoid path blocking or the risk of uncontrollable broadcast issues related to extending Layer 2 domains.

The traffic flow in the core consists primarily of sessions traveling between the campus core and the aggregation modules. The core aggregates the aggregation module traffic flows onto optimal paths to the campus core, as shown in [Figure 2-2](#). Server-to-server traffic typically remains within an aggregation module, but backup and replication traffic can travel between aggregation modules by way of the core.

Figure 2-2 Traffic Flow through the Core Layer

As shown in [Figure 2-2](#), the path selection can be influenced by the presence of service modules and the access layer topology being used. Routing from core to aggregation layer can be tuned for bringing all traffic into a particular aggregation node where primary service modules are located. This is described in more detail in [Chapter 7, “Increasing HA in the Data Center.”](#)

From a campus core perspective, there are at least two equal cost routes to the server subnets, which permits the core to load balance flows to each aggregation switch in a particular module. By default, this is performed using CEF-based load balancing on Layer 3 source/destination IP address hashing. An option is to use Layer 3 IP plus Layer 4 port-based CEF load balance hashing algorithms. This usually improves load distribution because it presents more unique values to the hashing algorithm.

To globally enable the Layer 3- plus Layer 4-based CEF hashing algorithm, use the following command at the global level:

```
CORE1(config)#mls ip cef load full
```

**Note**

Most IP stacks use automatic source port number randomization, which contributes to improved load distribution. Sometimes, for policy or other reasons, port numbers are translated by firewalls, load balancers, or other devices. We recommend that you always test a particular hash algorithm before implementing it in a production network.

Data Center Aggregation Layer

The aggregation layer, with many access layer uplinks connected to it, has the primary responsibility of aggregating the thousands of sessions leaving and entering the data center. The aggregation switches must be capable of supporting many 10 GigE and GigE interconnects while providing a high-speed switching fabric with a high forwarding rate. The aggregation layer also provides value-added services, such as server load balancing, firewalling, and SSL offloading to the servers across the access layer switches.

The aggregation layer switches carry the workload of spanning tree processing and default gateway redundancy protocol processing. The aggregation layer might be the most critical layer in the data center because port density, over-subscription values, CPU processing, and service modules introduce unique implications into the overall design.

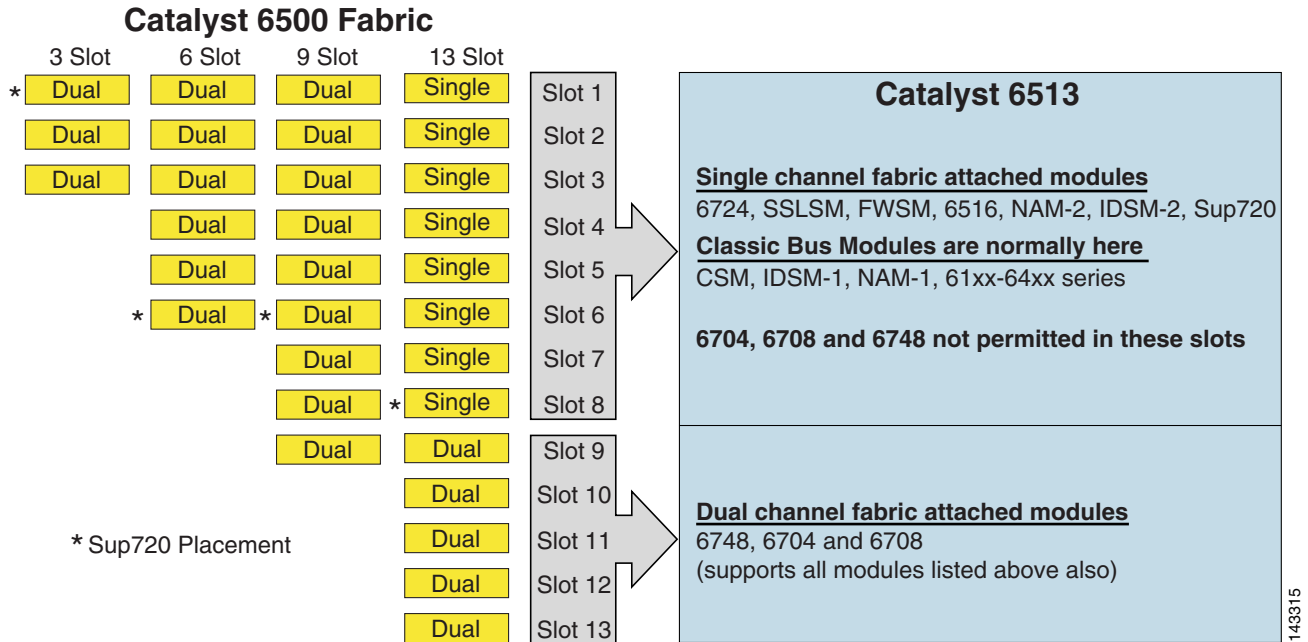
Recommended Platforms and Modules

The enterprise data center contains at least one aggregation module that consists of two aggregation layer switches. The aggregation switch pairs work together to provide redundancy and to maintain session state while providing valuable services to the access layer.

The recommended platforms for the aggregation layer include the Cisco Catalyst 6509 and Catalyst 6513 switches equipped with Sup720 processor modules. The high switching rate, large switch fabric, and ability to support a large number of 10 GigE ports are important requirements in the aggregation layer. The aggregation layer must also support security and application devices and services, including the following:

- Firewall Services Modules (FWSM)
- Secure Sockets Layer Services Modules (SSLSM)
- Content Switching Module (CSM)
- Intrusion Detection
- Network Analysis Module (NAM)
- Distributed denial-of-service attack protection (Guard)

Although the Cisco Catalyst 6513 might appear to be a good fit for the aggregation layer because of the high number of slots, note that it supports a mixture of single and dual channel slots. Slots 1 to 8 are single channel and slots 9 to 13 are dual-channel (see [Figure 2-3](#)).

Figure 2-3 Catalyst 6500 Fabric Channels by Chassis and Slot

Dual-channel line cards, such as the 6704-10 GigE, 6708-10G, or the 6748-SFP (TX) can be placed in slots 9–13. Single-channel line cards such as the 6724-SFP, as well as older single-channel or classic bus line cards can be used and are best suited in slots 1–8, but can also be used in slots 9–13. In contrast to the Catalyst 6513, the Catalyst 6509 has fewer available slots, but it can support dual-channel modules in every slot.

**Note**

A dual-channel slot can support all module types (CEF720, CEF256, and classic bus). A single-channel slot can support all modules with the exception of dual-channel cards, which currently include the 6704, 6708, and 6748 line cards.

The choice between a Cisco Catalyst 6509 or 6513 can best be determined by reviewing the following requirements:

- Cisco Catalyst 6509—When the aggregation layer requires many 10 GigE links with few or no service modules and very high performance.
- Cisco Catalyst 6513—When the aggregation layer requires a small number of 10 GigE links with many service modules.

If a large number of service modules are required at the aggregation layer, a service layer switch can help optimize the aggregation layer slot usage and performance. The service layer switch is covered in more detail in [Traffic Flow through the Service Layer, page 2-22](#).

Other considerations are related to air cooling and cabinet space usage. The Catalyst 6509 can be ordered in a NEBS-compliant chassis that provides front-to-back air ventilation that might be required in certain data center configurations. The Cisco Catalyst 6509 NEBS version can also be stacked two units high in a single data center cabinet, thereby using space more efficiently.

Distributed Forwarding

Using DFCs in the aggregation layer of the multi-tier model is optional. An analysis of application session flows that can transit the aggregation layer helps to determine the maximum forwarding requirements and whether DFCs would be beneficial. For example, if server tiers across access layer switches result in a large amount of inter-process communication (IPC) between them, the aggregation layer could benefit by using DFCs. Generally speaking, the aggregation layer benefits with lower latency and higher overall forwarding rates when including DFCs on the line cards.

**Note**

For more information on DFC operations, refer to [Distributed Forwarding, page 2-3](#).

**Note**

Refer to the Caveats section of the Release Notes for more detailed information regarding the use of DFCs when service modules are present or when distributed Etherchannels are used in the aggregation layer.

Traffic Flow in the Data Center Aggregation Layer

The aggregation layer connects to the core layer using Layer 3-terminated 10 GigE links. Layer 3 links are required to achieve bandwidth scalability, quick convergence, and to avoid path blocking or the risk of uncontrollable broadcast issues related to trunking Layer 2 domains.

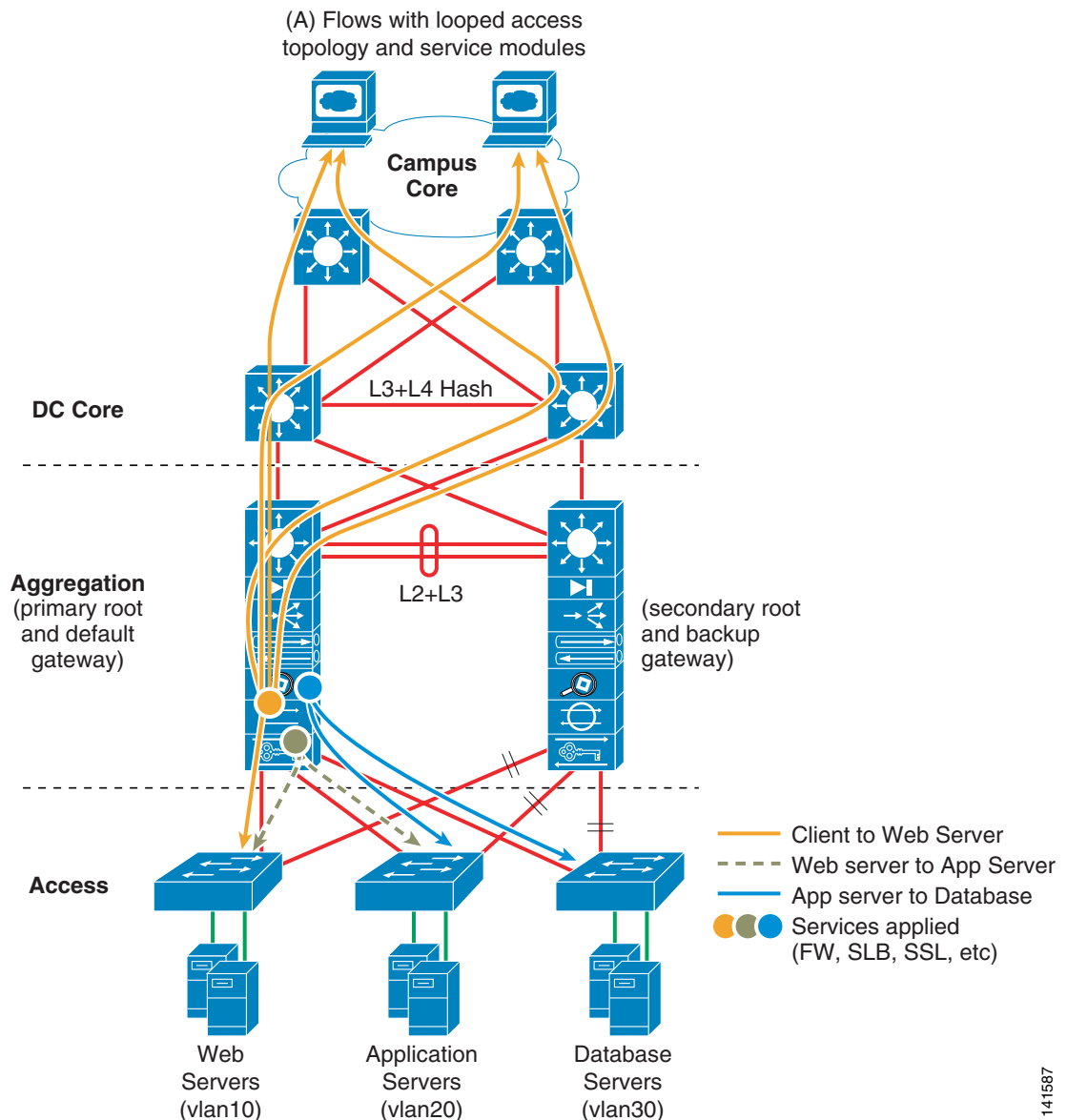
The traffic in the aggregation layer primarily consists of the following flows:

- Core layer to access layer—The core-to-access traffic flows are usually associated with client HTTP-based requests to the web server farm. At least two equal cost routes exist to the web server subnets. The CEF-based L3 plus L4 hashing algorithm determines how sessions balance across the equal cost paths. The web sessions might initially be directed to a VIP address that resides on a load balancer in the aggregation layer, or sent directly to the server farm. After the client request goes through the load balancer, it might then be directed to an SSL offload module or a transparent firewall before continuing to the actual server residing in the access layer.
- Access layer to access layer—The aggregation module is the primary transport for server-to-server traffic across the access layer. This includes server-to-server, multi-tier traffic types (web-to-application or application-to-database) and other traffic types, including backup or replication traffic. Service modules in the aggregation layer permit server-to-server traffic to use load balancers, SSL offloaders, and firewall services to improve the scalability and security of the server farm.

The path selection used for the various flows varies, based on different design requirements. These differences are based primarily on the presence of *service modules* and by the *access layer topology* used.

Path Selection in the Presence of Service Modules

When service modules are used in an active-standby arrangement, they are placed in both aggregation layer switches in a redundant fashion, with the primary active service modules in the Aggregation 1 switch and the secondary standby service modules is in the Aggregation 2 switch, as shown in [Figure 2-4](#).

Figure 2-4 Traffic Flow with Service Modules in a Looped Access Topology

In a service module-enabled design, you might want to tune the routing protocol configuration so that a primary traffic path is established towards the active service modules in the Aggregation 1 switch and, in a failure condition, a secondary path is established to the standby service modules in the Aggregation 2 switch. This provides a design with predictable behavior and traffic patterns, which facilitates troubleshooting. Also, by aligning all active service modules in the same switch, flows between service modules stay on the local switching bus without traversing the trunk between aggregation switches.

**Note**

More detail on path preference design is provided in [Chapter 7, “Increasing HA in the Data Center.”](#)

Without route tuning, the core has two equal cost routes to the server farm subnet; therefore, sessions are distributed across links to both aggregation layer switches. Because Aggregation 1 contains the active service modules, 50 percent of the sessions unnecessarily traverse the inter-switch link between

Aggregation 1 and Aggregation 2. By tuning the routing configuration, the sessions can remain on symmetrical paths in a predictable manner. Route tuning also helps in certain failure scenarios that create active-active service module scenarios.

Server Farm Traffic Flow with Service Modules

Traffic flows in the server farm consist mainly of multi-tier communications, including client-to-web, web-to-application, and application-to-database. Other traffic types that might exist include storage access (NAS or iSCSI), backup, and replication.

As described in the previous section of this chapter, we recommend that you align active services in a common aggregation layer switch. This keeps session flows on the same high speed bus, providing predictable behavior, while simplifying troubleshooting. A looped access layer topology, as shown in [Figure 2-4](#), provides a proven model in support of the active/standby service module implementation. By aligning spanning tree primary root and HSRP primary default gateway services on the Aggregation 1 switch, a symmetrical traffic path is established.

If multiple pairs of service modules are used in an aggregation switch pair, it is possible to distribute active services, which permits both access layer uplinks to be used. However, this is not usually a viable solution because of the additional service modules that are required. Future active-active abilities should permit this distribution without the need for additional service modules.

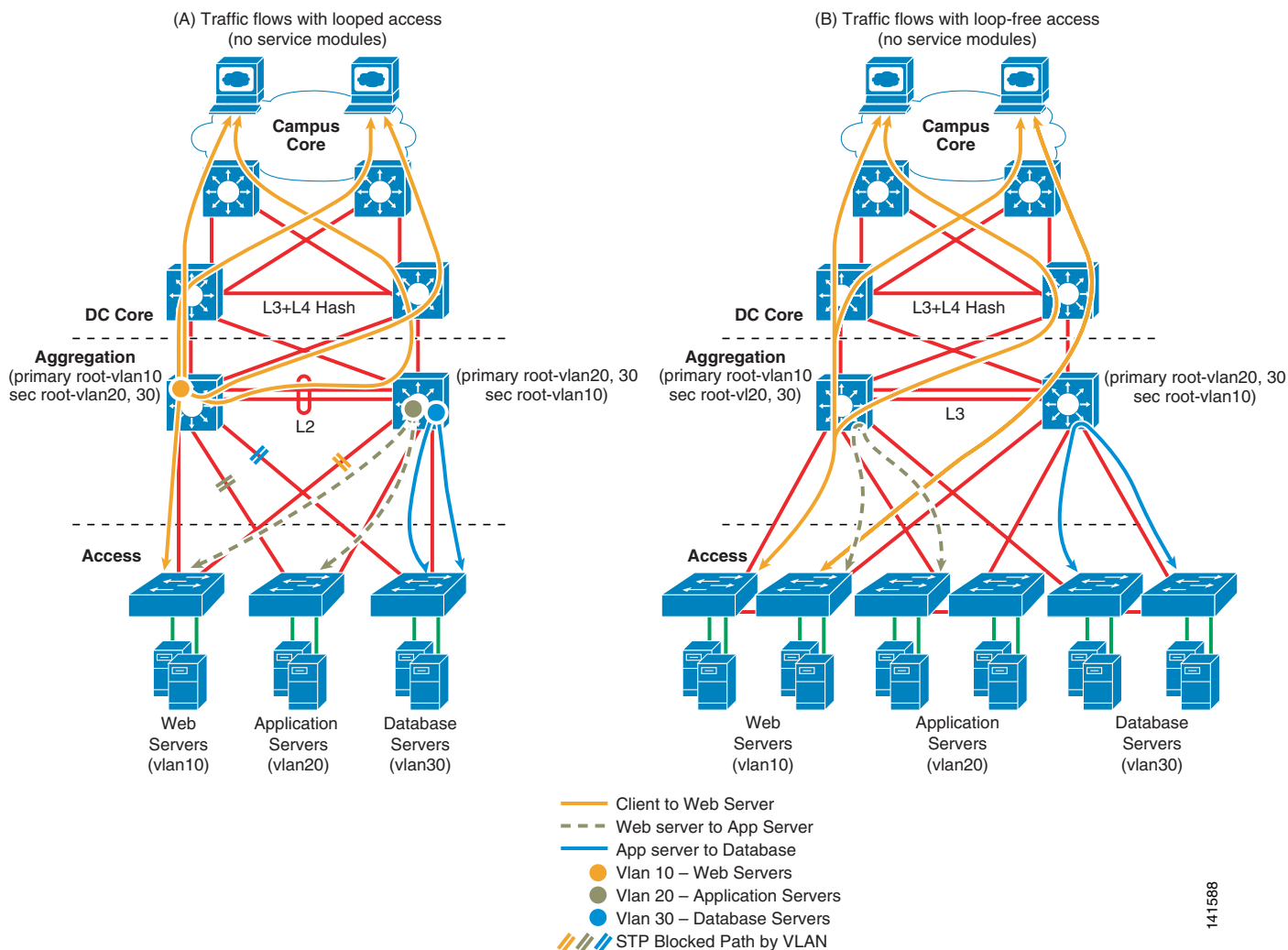


Note

The CSM and FWSM-2.x service modules currently operate in active/standby modes. These module pairs require identical configurations. The access layer design must ensure that connection paths remain symmetrical to the active service module. For more information on access layer designs, refer to [Chapter 6, “Data Center Access Layer Design.”](#) The Application Control Engine (ACE) is a new module that introduces several enhancements with respect to load balancing and security services. A key difference between the CSM, FWSM release 2.x, and ACE is the ability to support active-active contexts across the aggregation module with per context failover. Because the ACE module is not released at the time of this writing, it is not covered.

Server Farm Traffic Flow without Service Modules

When service modules are not used in the aggregation layer switches, multiple access layer topologies can be used. [Figure 2-5](#) shows the traffic flows with both looped and loop-free topologies.

Figure 2-5 Traffic Flow without Service Modules

When service modules are not present, it is possible to distribute the root and HSRP default gateway between aggregation switches as shown in Figure 2-5. This permits traffic to be balanced across both the aggregation switches and the access layer uplinks.

Scaling the Aggregation Layer

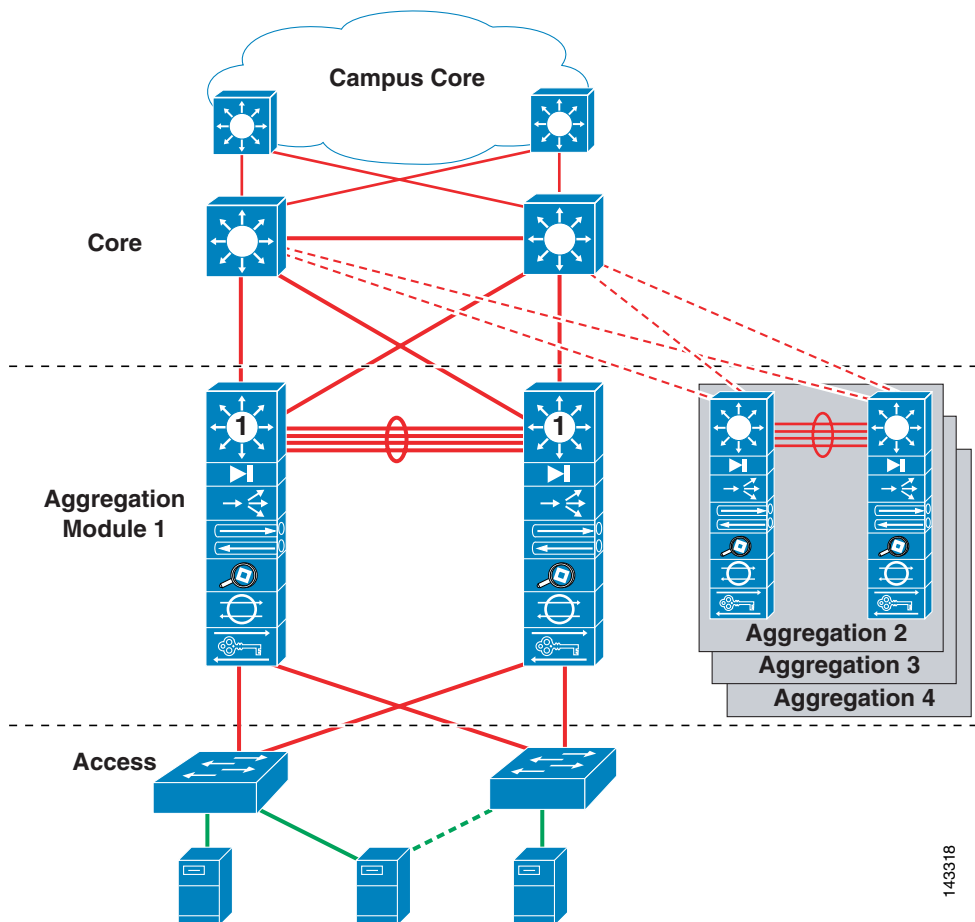
The aggregation layer design is critical to the stability and scalability of the overall data center architecture. All traffic in and out of the data center not only passes through the aggregation layer but also relies on the services, path selection, and redundant architecture built in to the aggregation layer design. This section describes the following four areas of critical importance that influence the aggregation layer design:

- Layer 2 fault domain size
- Spanning tree scalability
- 10 GigE density
- Default gateway redundancy scaling (HSRP)

The aggregation layer consists of pairs of interconnected aggregation switches referred to as modules.

Figure 2-6 shows a multiple aggregation module design using a common core.

Figure 2-6 Multiple Aggregation Modules



The use of aggregation modules helps solve the scalability challenges related to the four areas listed previously. These areas are covered in the following subsections.

Layer 2 Fault Domain Size

As Layer 2 domains continue to increase in size because of clustering, NIC teaming, and other application requirements, Layer 2 diameters are being pushed to scale further than ever before. The aggregation layer carries the largest burden in this regard because it establishes the Layer 2 domain size and manages it with a spanning tree protocol such as Rapid-PVST+ or MST.

The first area of concern related to large Layer 2 diameters is the *fault domain size*. Although features continue to improve the robustness and stability of Layer 2 domains, a level of exposure still remains regarding broadcast storms that can be caused by malfunctioning hardware or human error. Because a loop is present, all links cannot be in a forwarding state at all times because broadcasts/multicast packets would travel in an endless loop, completely saturating the VLAN, and would adversely affect network performance. A spanning tree protocol such as Rapid PVST+ or MST is required to automatically block a particular link and break the loop condition.

**Note**

Details on spanning tree protocol types and comparisons are covered in version 1.1 of this guide that can be found at the following URL:
http://www.cisco.com/en/US/netsol/ns656/networking_solutions_design_guidances_list.html#anchor3

Large data centers should consider establishing a maximum Layer 2 domain size to determine their maximum exposure level to this issue. By using multiple aggregation modules, the Layer 2 domain size can be limited; thus, the failure exposure can be pre-determined. Many customers use a “maximum number of servers” value to determine their maximum Layer 2 fault domain.

Spanning Tree Scalability

Extending VLANs across the data center is not only necessary to meet application requirements such as Layer 2 adjacency, but to permit a high level of flexibility in administering the servers. Many customers require the ability to group and maintain departmental servers together in a common VLAN or IP subnet address space. This makes management of the data center environment easier with respect to additions, moves, and changes.

When using a Layer 2 looped topology, a loop protection mechanism such as the Spanning Tree Protocol is required. Spanning tree automatically breaks loops, preventing broadcast packets from continuously circulating and melting down the network. The spanning tree protocols recommended in the data center design are 802.1w-Rapid PVST+ and 802.1s-MST. Both 802.1w and 802.1s have the same quick convergence characteristics but differ in flexibility and operation.

The aggregation layer carries the workload as it pertains to spanning tree processing. The quantity of VLANs and to what limits they are extended directly affect spanning tree in terms of scalability and convergence. The implementation of aggregation modules helps to distribute and scale spanning tree processing.

**Note**

More details on spanning tree scaling are provided in [Chapter 5, “Spanning Tree Scalability.”](#)

10 GigE Density

As the access layer demands increase in terms of bandwidth and server interface requirements, the uplinks to the aggregation layer are migrating beyond GigE or Gigabit EtherChannel speeds and moving to 10 GigE. This trend is expected to increase and could create a density challenge in existing or new aggregation layer designs. Although the long term answer might be higher density 10 GigE line cards and larger switch fabrics, a current proven solution is the use of multiple aggregation modules.

Currently, the maximum number of 10 GigE ports that can be placed in the aggregation layer switch is 64 when using the WS-X6708-10G-3C line card in the Catalyst 6509. However, after considering firewall, load balancer, network analysis, and other service-related modules, this is typically a lower number. Using a data center core layer and implementing multiple aggregation modules provides a higher level of 10 GigE density.

**Note**

It is also important to understand traffic flow in the data center when deploying these higher density 10 GigE modules, due to their oversubscribed nature.

The access layer design can also influence the 10 GigE density used at the aggregation layer. For example, a square loop topology permits twice the number of access layer switches when compared to a triangle loop topology. For more details on access layer design, refer to [Chapter 6, “Data Center Access Layer Design.”](#)

Default Gateway Redundancy with HSRP

The aggregation layer provides a primary and secondary router “default gateway” address for all servers across the entire access layer using HSRP, VRRP, or GLBP default gateway redundancy protocols. This is applicable only with servers on a Layer 2 access topology. The CPU on the Sup720 modules in both aggregation switches carries the processing burden to support this necessary feature. The overhead on the CPU is linear to the update timer configuration and the number of VLANs that are extended across the entire access layer supported by that aggregation module because the state of each active default gateway is maintained between them. In the event of an aggregation hardware or medium failure, one CPU must take over as the primary default gateway for each VLAN configured.

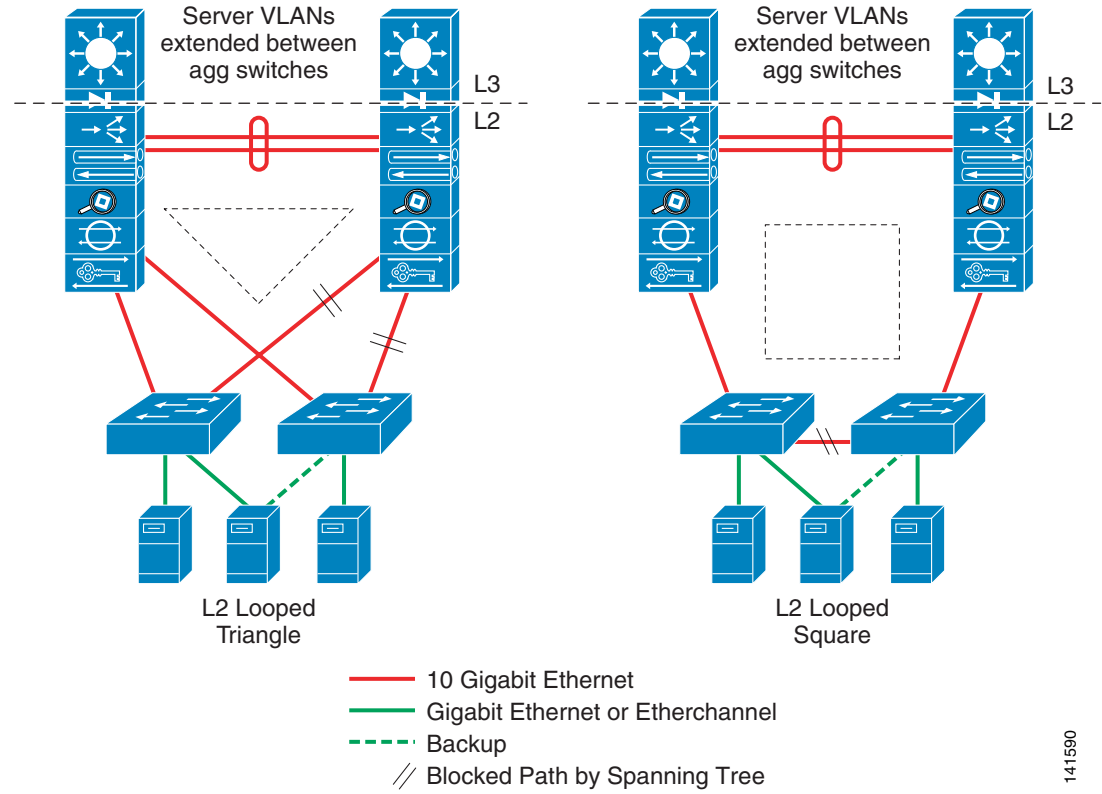
HSRP is the most widely used protocol for default gateway redundancy. HSRP provides the richest feature set and flexibility to support multiple groups, adjustable timers, tracking, and a large number of instances. Current testing results recommend the maximum number of HSRP instances in an aggregation module to be limited to ~ 500, with recommended timers of a one second hello and a three second hold time. Consideration of other CPU interrupt-driven processes that could be running on the aggregation layer switch (such as tunneling and SNMP polling) should be taken into account as they could reduce this value further downward. If more HSRP instances are required, we recommend distributing this load across multiple aggregation module switches. More detail on HSRP design and scalability is provided in [Chapter 4, “Data Center Design Considerations.”](#)

Data Center Access Layer

The access layer provides the physical level attachment to the server resources, and operates in Layer 2 or Layer 3 modes. The mode plays a critical role in meeting particular server requirements such as NIC teaming, clustering, and broadcast containment. The access layer is the first oversubscription point in the data center because it aggregates the server traffic onto Gigabit EtherChannel or 10 GigE/10 Gigabit EtherChannel uplinks to the aggregation layer. Spanning tree or Layer 3 routing protocols are extended from the aggregation layer into the access layer, depending on which access layer model is used. Cisco recommends implementing access layer switches logically paired in groups of two to support server redundant connections or to support diverse connections for production, backup, and management Ethernet interfaces.

The access layer consists mainly of three models: Layer 2 looped, Layer 2 loop-free, and Layer 3.

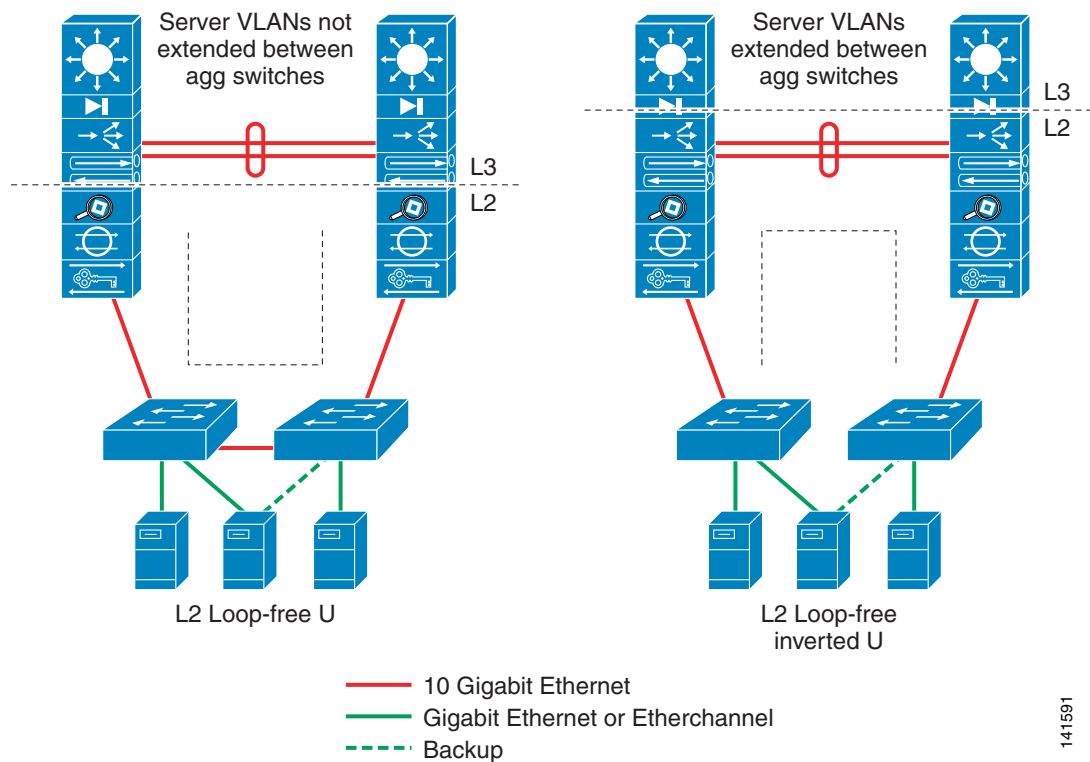
[Figure 2-7](#) illustrates the access layer using the Layer 2 looped model in triangle and square loop topologies.

Figure 2-7 Access Layer Looped Topologies

The triangle-based looped topology is the most widely used today. Looped topologies are the most desirable in the data center access layer for the following reasons:

- VLAN extension—The ability to add servers into a specific VLAN across the entire access layer is a key requirement in most data centers.
- Resiliency—Looped topologies are inherently redundant.
- Service module interoperability—Service modules operating in active-standby modes require Layer 2 adjacency between their interfaces.
- Server requirements for Layer 2 adjacency in support of NIC teaming and high availability clustering.

Figure 2-8 illustrates the access layer using the Layer 2 loop-free model, in loop-free U and loop-free inverted U topologies.

Figure 2-8 Access Layer Loop-free Topologies

The loop-free Layer 2 model is typically used when looped topology characteristics are undesirable. This could be due to inexperience with Layer 2 spanning tree protocols, a need for all uplinks to be active, or bad experiences related to STP implementations. The following are the main differences between a looped and loop-free topology:

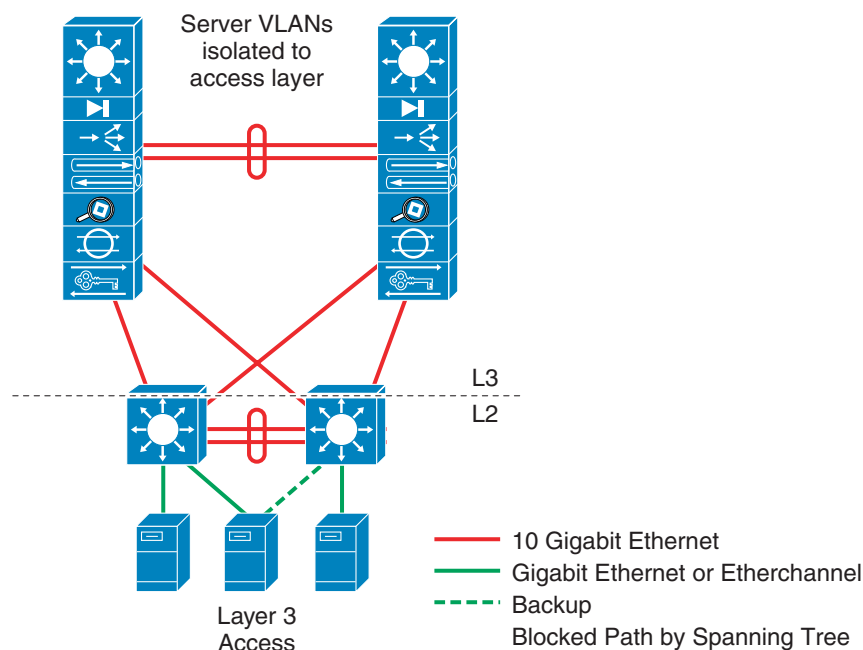
- No blocking on uplinks, all links are active in a loop-free topology
- Layer 2 adjacency for servers is limited to a single pair of access switches in a loop-free topology
- VLAN extension across the data center is not supported in a loop-free U topology but is supported in the inverted U topology.

When using a loop-free model, it is still necessary to run STP as a loop prevention tool.

**Note**

Service modules might not operate as expected during certain failure conditions when using a loop-free U topology. More detail on access layer design is provided in [Chapter 6, “Data Center Access Layer Design.”](#)

[Figure 2-9](#) illustrates the access layer using the Layer 3 model.

Figure 2-9 Access Layer 3 Topology

141592

The Layer 3 access model is typically used to limit or contain broadcast domains to a particular size. This can be used to reduce exposure to broadcast domain issues or to shelter particular servers that could be adversely affected by a particular broadcast level. Layer 3 access has the following characteristics:

- All uplinks are active and use CEF load balancing up to the ECMP maximum (currently 8)
- Layer 2 adjacency for servers is limited to a single pair of access switches in the Layer 3 topology
- VLAN extension across the data center is not possible

When using a Layer 3 access model, Cisco still recommends running STP as a loop prevention tool. STP protocol would be active only on the inter-switch trunk and server ports.

**Note**

Because active-standby service modules require Layer 2 adjacency between their interfaces, the Layer 3 access design does not permit service modules to reside at the aggregation layer and requires placement in each access switch pair. A Layer 3 access design that leverages the use of VRF-Lite might provide an aggregation layer service module solution, but this has not been tested for inclusion in this guide.

Recommended Platforms and Modules

The recommended platforms for the access layer include all Cisco Catalyst 6500 Series switches that are equipped with Sup720 processor modules for modular implementations, and the Catalyst 4948-10GE for top of rack implementations.

The Catalyst 6500 modular access switch provides a high GE port density, 10GE(C) uplinks, redundant components and security features while also providing a high bandwidth switching fabric that the server farm requires. The Catalyst 4948-10GE provides dual 10GE uplinks, redundant power, plus 48 GE server ports in a 1RU form factor that makes it ideal for top of rack solutions. Both the Catalyst 6500 Series switch and the Catalyst 4948-10GE use the IOS image to provide the same configuration look and feel, simplifying server farm deployments.

The following are some of the most common considerations in choosing access layer platforms:

- **Density**—The density of servers together with the maximum number of interfaces used per rack/row can help determine whether a modular or a 1RU solution is a better fit. If a high number of ports per rack are used, it might take many 1RU switches in each rack to support them. Modular switches that are spaced out in the row might reduce the complexity in terms of the number of switches, and permit more flexibility in supporting varying numbers of server interfaces.
- **10 GigE/10 Gigabit EtherChannel uplink support**—It is important to determine what the oversubscription ratio is per application. When this value is known, it can be used to determine the correct amount of uplink bandwidth that is required on the access layer switch. Choosing a switch that can support 10 GigE and 10 Gigabit EtherChannel might be an important option when considering current or future oversubscription ratios.
- **Resiliency features**—When servers are connected with a single NIC interface card at the access layer, the access switch becomes a single point of failure. This makes features such as redundant power and redundant processor much more important in the access layer switch.
- **Production compared to development use**—A development network might not require the redundancy or the software-rich features that are required by the production environment.
- **Cabling design/cooling requirements**—Cable density in the server cabinet and under the floor can be difficult to manage and support. Cable bulk can also create cooling challenges if air passages are blocked. The use of 1RU access switches can improve the cabling design.

The recommended access layer platforms include the following Cisco Catalyst models:

- All Catalyst 6500 Series platforms with the Sup720 processor
- Catalyst 4948-10G

Distributed Forwarding

Using DFCs in the access layer of the multi-tier model is optional. The performance requirements for the majority of enterprise data center access switches are met without the need for DFCs, and in many cases they are not necessary.

If heavy server-to-server traffic on the same modular chassis is expected, such as in HPC designs, DFCs can certainly improve performance. [Table 2-1](#) provides a performance comparison.



Note

The forwarding performance attained when using DFCs apply to both Layer 2 and Layer 3 packet switching.

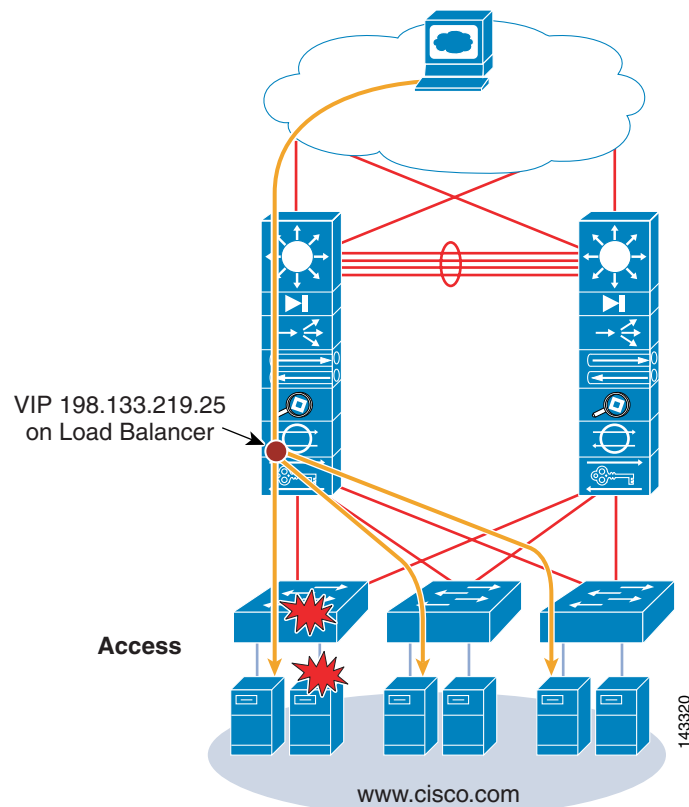
Resiliency

The servers connecting at the access layer can be single-homed or dual-homed for redundancy. A single-homed server has no protection from a NIC card or access switch-related failure and represents a single point of failure. CPU and power redundancy are critical at the access layer when single-attached servers are used because an access switch failure can have a major impact on network availability. If single attached servers create a large exposure point, consideration should be given to platforms that provide full load-redundant power supplies, CPU redundancy, and stateful switchover.

Applications that are running on single-attached servers can use server load balancers, such as the CSM, to achieve redundancy. In this case, servers in a particular application group (VIP) are distributed across two or more access layer switches, which eliminates the access switch as a single point of failure.

Figure 2-10 shows how to use load balancers to achieve redundancy with single-attached servers in a web server farm.

Figure 2-10 Server Redundancy with Load Balancers



In this example, a server NIC failure or an access switch failure causes the servers to be automatically taken out of service by the load balancer, and sessions to be directed to the remaining servers. This is accomplished by leveraging the health monitoring features of the CSM.

Sharing Services at the Aggregation Layer

A Layer 2-looped access topology has the unique advantage of being able to use services provided by service modules or appliances located at the aggregation layer. The integrated service modules in the aggregation layer optimize rack space and cabling, simplify configuration management, and improve the overall flexibility and scalability.

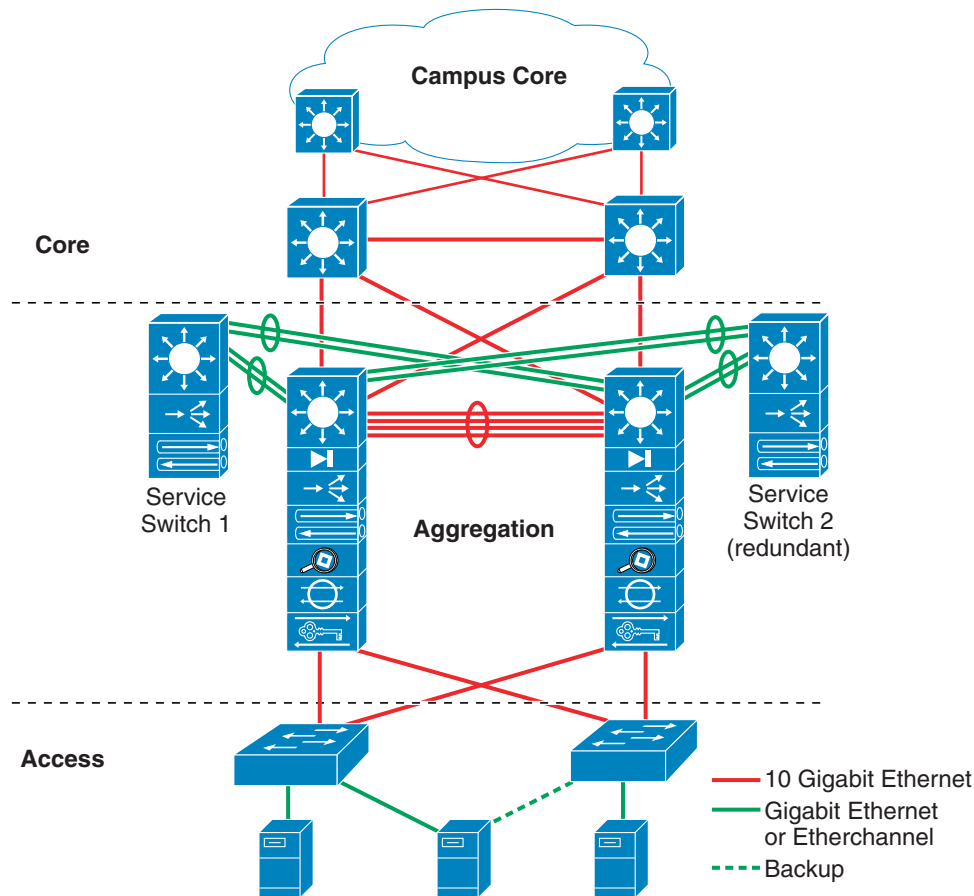
The services in the aggregation layer that can be used by the access layer servers include the following:

- Load balancing
- Firewalling
- SSL offloading (session encryption/decryption)
- Network monitoring
- Intrusion detection and prevention
- Cache engines

Data Center Services Layer

The service layer switch provides a method of scaling up services using service modules without using slots in the aggregation layer switch. Moving certain service modules out of the aggregation layer switch increases the number of available slots and improves aggregation layer performance. For example, this is useful when a farm of CSMs or SSL offload modules are required. Figure 2-11 shows a topology with a service layer switch design.

Figure 2-11 Data Center Service Layer Switch



Recommended Platforms and Modules

Typically, the CSM and SSL service modules are located in the aggregation switch to provide services to servers across the access layer switches. By locating these service modules in a separate standalone switch connected using 802.1Q trunks, the aggregation layer can support a higher access layer uplink density and continue to operate in compact switching mode to optimize performance. This is particularly useful when 10 GigE port density requirements increase at the aggregation layer.

Performance Implications

Mixing older line card or service modules with the Sup720 integrated switch fabric architecture can limit the overall switching capacity and might not meet the performance requirements of the aggregation layer switches. This section examines the implications related to placing classic bus line cards in the aggregation layer switch.

The CSM module connects to the Catalyst 6500 bus using a 4 Gbps EtherChannel connection on the backplane. This interface can be viewed by examining the reserved EtherChannel address of port 259 as shown below:

```
AGG1#sh etherchannel 259 port-channel
Port-channel: Po259
-----
Age of the Port-channel    = 4d:00h:33m:39s
Logical slot/port         = 14/8           Number of ports = 4
GC                        = 0x00000000     HotStandBy port = null
Port state                = Port-channel Ag-Inuse
Protocol                  = -
Ports in the Port-channel:
Index  Load  Port      EC state  No of bits
-----+-----+-----+-----+-----
0      11    Gi3/1     On/FEC    2
1      22    Gi3/2     On/FEC    2
2      44    Gi3/3     On/FEC    2
3      88    Gi3/4     On/FEC    2
```

This 4 Gbps EtherChannel interface is used for all traffic entering and exiting the load balancer and uses hashing algorithms to distribute session load across it just as would an external physical EtherChannel connection. The CSM is also based on the classic bus architecture and depends on the Sup720 to switch packets in and out of its EtherChannel interface because it does not have a direct interface to the Sup720 integrated switch fabric.

If a single 4 Gbps EtherChannel does not provide enough bandwidth to meet design requirements, then multiple CSM modules can be used to scale, as required. Supporting multiple CSM modules does not create any particular problem but it does increase the number of slots used in the aggregation layer switch, which might not be desirable.

Because the CSM is classic bus-based, it must send truncated packet headers to the Sup720 PFC3 to determine packet destination on the backplane. When a single classic bus module exists in the switch, all non-DFC enabled line cards must perform truncated header lookup, which limits the overall system performance. [Table 2-2](#) provides an overview of switching performance by module type.

Table 2-2 Performance with Classic Bus Modules

System Config with Sup720	Throughput in Mpps	Bandwidth in Gbps
Classic Series Modules (CSM, 61XX-64XX)	Up to 15 Mpps (per system)	16 Gbps shared bus (classic bus)
CEF256 Series Modules (FWSM, SSLSM, NAM-2, IDSM-2, 6516)	Up to 30 Mpps (per system)	1 x 8 Gbps (dedicated per slot)

Table 2-2 Performance with Classic Bus Modules (continued)

CEF720 Series Modules (6748, 6704, 6724)	Up to 30 Mpps (per system)	2 x 20 Gbps (dedicated per slot) (6724=1 x 20 Gbps)
CEF720 Series Modules with DFC3 (6704 with DFC3, 6708 with DFC3, 6748 with DFC3, 6724 with DFC3)	Sustain up to 48 Mpps (per slot)	2 x 20 Gbps (dedicated per slot) (6724=1 x 20 Gbps)

The service switch can be any of the Catalyst 6500 Series platforms that use a Sup2 or Sup720 CPU module. The supervisor engine choice should consider sparing requirements, future migration to next generation modules, performance requirements, and uplink requirements to the aggregation module. For example, if 10 GigE uplinks are planned, you must use the sup720 to support the 6704 10 GigE module. A Sup32-10 GigE can also be used if only classic bus-enabled modules are being used, such as the CSM, but this has not been tested as part of this guide.

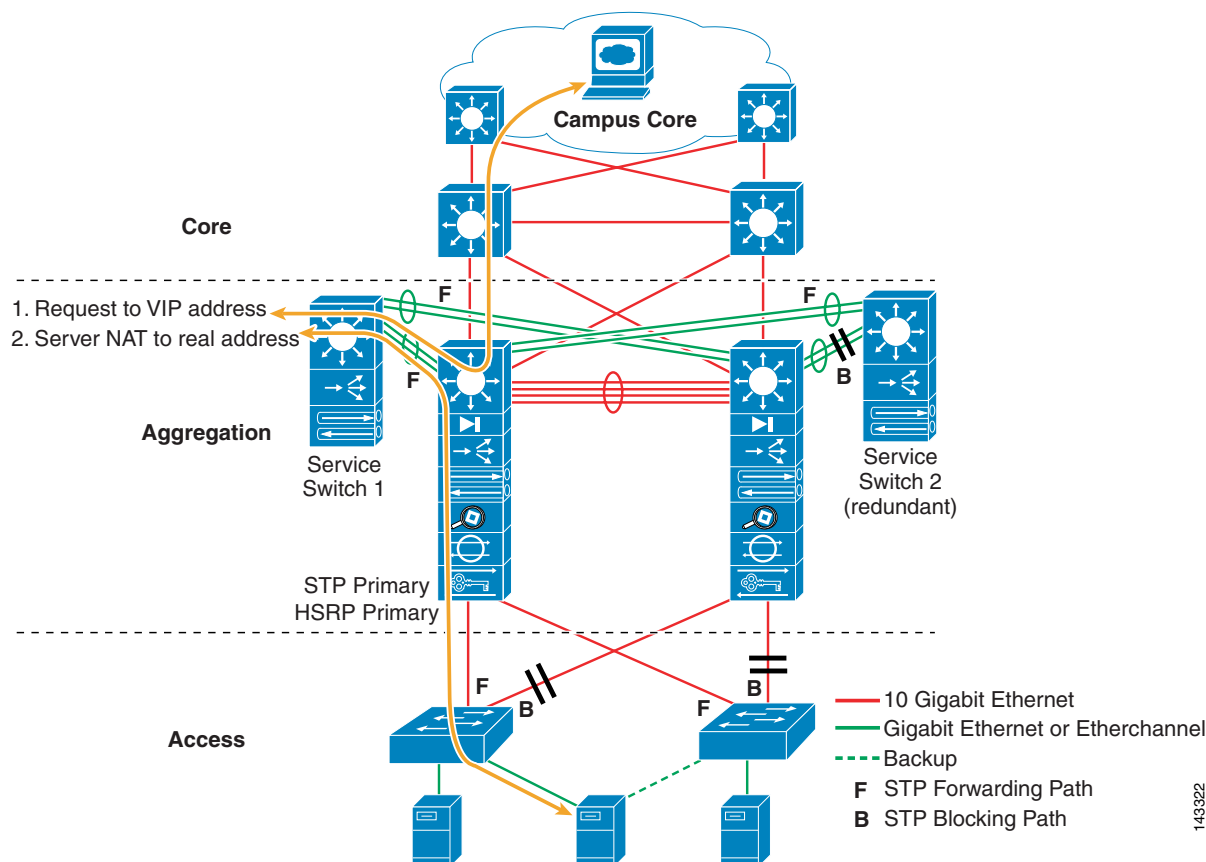
**Note**

If a CSM in a service switch is configured for Route Health Injection (RHI), a Layer 3 configuration to the aggregation layer switch is necessary, because RHI knows how to insert a host route into only the routing table of the local MSFC. A Layer 3 link permits a routing protocol to redistribute the host route to the aggregation layer.

Traffic Flow through the Service Layer

The service switch is connected to both aggregation switches with Gigabit EtherChannel (GigEC) or 10 GigE links configured as 802.1Q trunks. From a logical perspective, this can be viewed as extending the service module VLANs across an 802.1Q trunk. [Figure 2-12](#) shows the flow of a session using a CSM in one-arm mode on a service layer switch.

Figure 2-12 Service Layer Switch Traffic Flow

**Note**

Configuration examples are provided in [Chapter 8, “Configuration Reference.”](#)

The VLANs used in supporting the CSM configuration are extended across 802.1Q trunks (GEC or 10GigE) from the aggregation layer switch to each service switch. The Fault Tolerant (FT) VLANs are also extended across these trunks. Spanning tree blocks only a single trunk from the secondary service switch to the Aggregation 2 switch. This configuration provides a forwarding path to the primary service switch from both aggregation switches.

**Note**

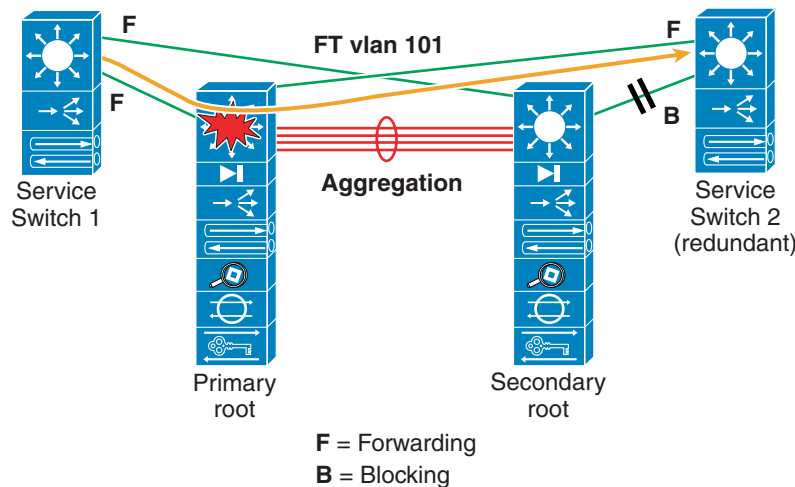
The bandwidth used to connect service switches should be carefully considered. The CSM is capable of supporting up to 4 Gbps of server and FT (replication and state) traffic on its bus interface. The switch uplink should be enough to avoid congestion on the FT path to ensure that CSMs do not go into an active-active state. A separate link containing the FT VLANs can be used to isolate the data and FT paths.

Resiliency

The service switch layer should be deployed in pairs to support a fully redundant configuration. This permits primary/active service modules to be located in one chassis and backup/standby in a second chassis. Both service switches should be redundantly connected to each aggregation switch to eliminate any single point of failure.

In [Figure 2-13](#), Service Switch 1 is configured with the active service modules while Service Switch 2 is used for standby service modules. If Service Switch 1 fails, Service Switch 2 becomes active and provides stateful failover for the existing connections.

Figure 2-13 Redundancy with the Service Layer Switch Design



The FT VLAN is used to maintain session state between service modules. The FT VLAN must be extended between service switches across each 802.1Q trunk that connects to the aggregation layer. The FT VLAN is not passed across the link connected between the aggregation switch pair.

The service switch is provisioned to participate in spanning tree and automatically elects the paths to the primary root on Aggregation 1 for the server VLAN. If Service Switch 1 fails, service module FT communication times out and Service Switch 2 becomes primary.

Service layer switches should not be used across more than one aggregation module. This is required to keep the spanning tree domains isolated between aggregation modules; otherwise, root switches in one aggregation module could become root for another, creating undesirable path selection and link blocking.



Server Cluster Designs with Ethernet



Note

The README file posted with this guide contains details about the technologies, hardware, and software that were used in producing this document. The README file also contains a revision history that details updates made to each chapter.

A high-level overview of the servers and network components used in the server cluster model is provided in [Chapter 1, “Data Center Architecture Overview.”](#) This chapter describes the purpose and function of each layer of the server cluster model in greater detail. The following sections are included:

- [Technical Objectives](#)
- [Distributed Forwarding and Latency](#)
- [Equal Cost Multi-Path Routing](#)
- [Server Cluster Design—Two-Tier Model](#)
- [Server Cluster Design—Three-Tier Model](#)
- [Recommended Hardware and Modules](#)



Note

The design models covered in this chapter have not been fully verified in Cisco lab testing because of the size and scope of testing that would be required. The two-tier models that are covered are similar designs that have been implemented in customer production networks.

Technical Objectives

When designing a large enterprise cluster network, it is critical to consider specific objectives. No two clusters are exactly alike; each has its own specific requirements and must be examined from an application perspective to determine the particular design requirements. Take into account the following technical considerations:

- **Latency**—In the network transport, latency can adversely affect the overall cluster performance. Using switching platforms that employ a low-latency switching architecture helps to ensure optimal performance. The main source of latency is the protocol stack and NIC hardware implementation used on the server. Driver optimization and CPU offload techniques, such as TCP Offload Engine (TOE) and Remote Direct Memory Access (RDMA), can help decrease latency and reduce processing overhead on the server.

Latency might not always be a critical factor in the cluster design. For example, some clusters might require high bandwidth between servers because of a large amount of bulk file transfer, but might not rely heavily on server-to-server Inter-Process Communication (IPC) messaging, which can be impacted by high latency.

- **Mesh/partial mesh connectivity**—Server cluster designs usually require a mesh or partial mesh fabric to permit communication between all nodes in the cluster. This mesh fabric is used to share state, data, and other information between master-to-compute and compute-to-compute servers in the cluster. Mesh or partial mesh connectivity is also application-dependent.
- **High throughput**—The ability to send a large file in a specific amount of time can be critical to cluster operation and performance. Server clusters typically require a minimum amount of available non-blocking bandwidth, which translates into a low oversubscription model between the access and core layers.
- **Oversubscription ratio**—The oversubscription ratio must be examined at multiple aggregation points in the design, including the line card to switch fabric bandwidth and the switch fabric input to uplink bandwidth.
- **Jumbo frame support**—Although jumbo frames might not be used in the initial implementation of a server cluster, it is a very important feature that is necessary for additional flexibility or for possible future requirements. The TCP/IP packet construction places additional overhead on the server CPU. The use of jumbo frames can reduce the number of packets, thereby reducing this overhead.
- **Port density**—Server clusters might need to scale to tens of thousands of ports. As such, they require platforms with a high level of packet switching performance, a large amount of switch fabric bandwidth, and a high level of port density.

Distributed Forwarding and Latency

The Cisco Catalyst 6500 Series switch has the unique ability to support a central packet forwarding or optional distributed forwarding architecture, while the Cisco Catalyst 4948-10GE is a single central ASIC design with fixed line rate forwarding performance. The Cisco 6700 line card modules support an optional daughter card module called a Distributed Forwarding Card (DFC). The DFC permits local routing decisions to occur on each line card by implementing a local Forwarding Information Base (FIB). The FIB table on the Sup720 PFC maintains synchronization with each DFC FIB table on the line cards to ensure routing integrity across the system.

When the optional DFC card is not present, a compact header lookup is sent to the PFC3 on the Sup720 to determine where on the switch fabric to forward each packet. When a DFC is present, the line card can switch a packet directly across the switch fabric to the destination line card without consulting the Sup720. The difference in performance can be from 30 Mpps system-wide without DFCs to 48 Mpps per slot with DFCs. The fixed configuration Catalyst 4948-10GE switch has a wire rate, non-blocking architecture supporting up to 101.18 Mpps performance, providing superior access layer performance for top of rack designs.

Latency performance can vary significantly when comparing the distributed and central forwarding models. [Table 3-1](#) provides an example of latencies measured across a 6704 line card with and without DFCs.

Table 3-1 Cisco Catalyst 6500 Latency Measurements based on RFC1242-LIFO (L2 and L3)

6704 with DFC (Port-to-Port in Microseconds through Switch Fabric)										
Packet size (B)	64	128	256	512	1024	1280	1518	4096	6000	9018
Latency (ms)	8.03	8.03	8.20	8.71	9.59	9.99	10.31	14.20	17.08	21.57
6704 without DFC (Port-to-Port in Microseconds through Switch Fabric)										
Packet size (B)	64	128	256	512	1024	1280	1518	4096	6000	9018
Latency (ms)	9.52	9.23	9.43	9.83	10.66	11.05	11.37	15.26	18.09	22.68

The difference in latency between a DFC-enabled and non-DFC-enabled line card might not appear significant. However, in a 6500 central forwarding architecture, latency can increase as traffic rates increase because of the contention for the shared lookup on the central bus. With a DFC, the lookup path is dedicated to each line card and the latency is constant.

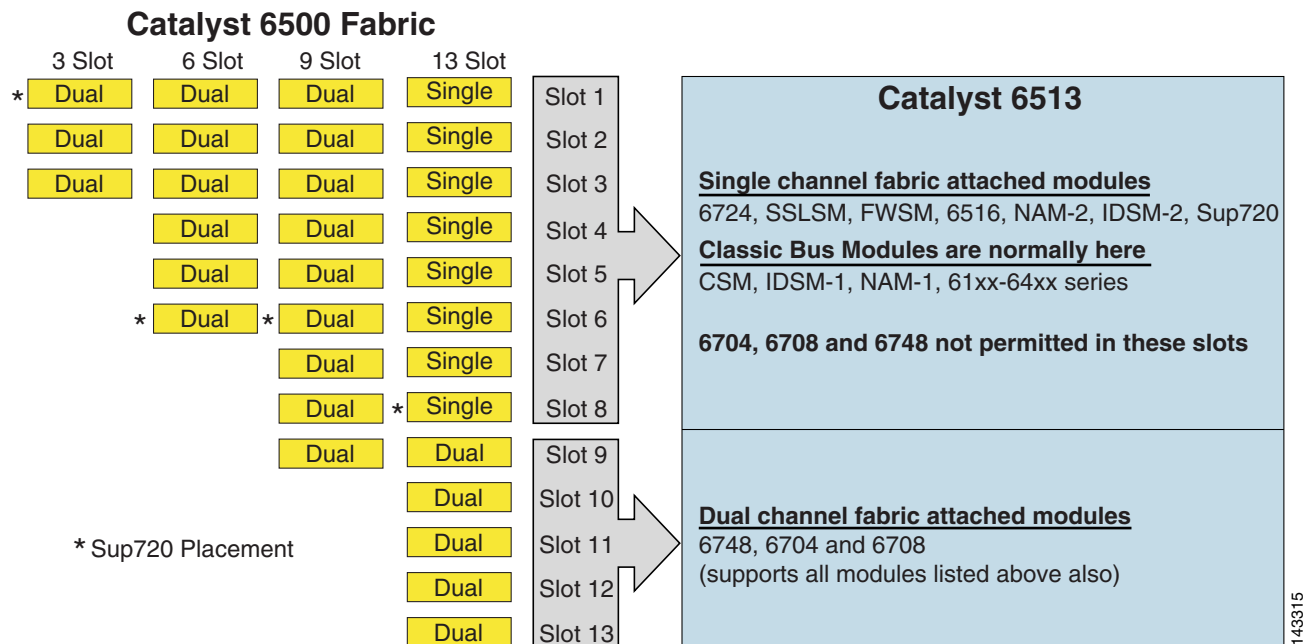
Catalyst 6500 System Bandwidth

The available system bandwidth does not change when DFCs are used. The DFCs improve the packets per second (pps) processing of the overall system. [Table 3-2](#) summarizes the throughput and bandwidth performance for modules that support DFCs, in addition to the older CEF256 and classic bus modules.

Table 3-2 Performance Comparison with Distributed Forwarding

System Configuration with Sup720	Throughput in Mpps	Bandwidth in Gbps
Classic series modules (CSM, 61xx–64xx)	Up to 15 Mpps (per system)	16 G shared bus (classic bus)
CEF256 Series modules (FWSM, SSLSM, NAM-2, IDSM-2, 6516)	Up to 30 Mpps (per system)	1x 8 G (dedicated per slot)
Mix of classic with CEF256 or CEF720 Series modules	Up to 15 Mpps (per system)	Card dependent
CEF720 Series modules (6748, 6704, 6724)	Up to 30 Mpps (per system)	2x 20 G (dedicated per slot) (6724=1x20G)
CEF720 Series modules with DFC3 (6704 with DFC3, 6708 with DFC3, 6748 with DFC3 6724+DFC3)	Sustain up to 48 Mpps (per slot)	2x 20 G (dedicated per slot) (6724=1x20 G)

Although the 6513 might be a valid solution for the access layer of the large cluster model, note that there is a mixture of single and dual channel slots in this chassis. Slots 1 to 8 are single channel and slots 9 to 13 are dual channel, as shown in [Figure 3-1](#).

Figure 3-1 Catalyst 6500 Fabric Channels by Chassis and Slot (6513 Focus)

When a Cisco Catalyst 6513 is used, the dual channel cards, such as the 6704-4 port 10GigE, the 6708-8 port 10GigE, and the 6748-48 port SFP/copper line cards can be placed only in slots 9 to 13. The single channel line cards such as the 6724-24 port SFP/copper line cards can be used in slots 1 to 8. The Sup720 uses slots 7 and 8, which are single channel 20G fabric attached. In contrast to the 6513, the 6509 has fewer available slots but can support dual channel modules in all slots because each slot has dual channels to the switch fabric.

**Note**

Because the server cluster environment usually requires high bandwidth with low latency characteristics, we recommend using DFCs in these types of designs.

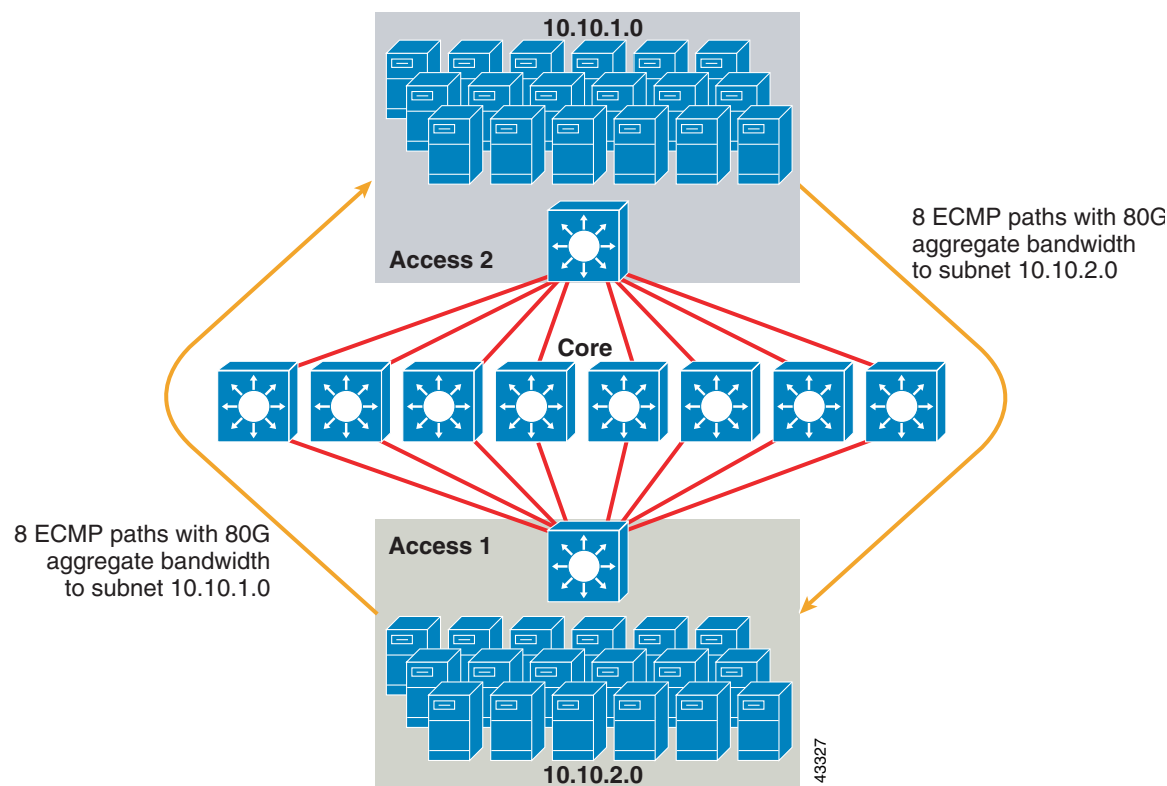
Equal Cost Multi-Path Routing

Equal cost multi-path (ECMP) routing is a load balancing technology that optimizes flows across multiple IP paths between any two subnets in a Cisco Express Forwarding-enabled environment. ECMP applies load balancing for TCP and UDP packets on a per-flow basis. Non-TCP/UDP packets, such as ICMP, are distributed on a packet-by-packet basis. ECMP is based on RFC 2991 and is leveraged on other Cisco platforms, such as the PIX and Cisco Content Services Switch (CSS) products. ECMP is supported on both the 6500 and 4948-10GE platforms recommended in the server cluster design.

The dramatic changes resulting from Layer 3 switching hardware ASICs and Cisco Express Forwarding hashing algorithms helps to distinguish ECMP from its predecessor technologies. The main benefit in an ECMP design for server cluster implementations is the hashing algorithm combined with little to no CPU overhead in Layer 3 switching. The Cisco Express Forwarding hashing algorithm is capable of distributing granular flows across multiple line cards at line rate in hardware. The hashing algorithm default setting is to hash flows based on Layer 3 source-destination IP addresses, and optionally adding Layer 4 port numbers for an additional layer of differentiation. The maximum number of ECMP paths allowed is eight.

Figure 3-2 illustrates an 8-way ECMP server cluster design. To simplify the illustration, only two access layer switches are shown, but up to 32 can be supported (64 10GigEs per core node).

Figure 3-2 8-Way ECMP Server Cluster Design



In Figure 3-2, each access layer switch can support one or more subnets of attached servers. Each switch has a single 10GigE connection to each of the eight core switches using two 6704 line cards. This configuration provides eight paths of 10GigE for a total of 80 G Cisco Express Forwarding-enabled bandwidth to any other subnet in the server cluster fabric. A **show ip route** query to another subnet on another switch shows eight equal-cost entries.

The core is populated with 10GigE line cards with DFCs to enable a fully-distributed high-speed switching fabric with very low port-to-port latency. A **show ip route** query to an access layer switch shows a single route entry on each of the eight core switches.



Note

Although it has not been tested for this guide, there is a new 8-port 10 Gigabit Ethernet module (WS-X6708-10G-3C) that has recently been introduced for the Catalyst 6500 Series switch. This line card will be tested for inclusion in this guide at a later date. For questions about the 8-port 10GigE card, refer to the product data sheet.

Redundancy in the Server Cluster Design

The server cluster design is typically not implemented with redundant CPU or switch fabric processors. Resiliency is typically achieved inherently in the design and by the method the cluster functions as a whole. As described in Chapter 1, “Data Center Architecture Overview,” the compute nodes in the

cluster are managed by master nodes that are responsible for assigning specific jobs to each compute node and monitoring their performance. If a compute node drops out of the cluster, it reassigns to an available node and continues to operate, although with less processing power, until the node is available. Although it is important to diversify master node connections in the cluster across different access switches, it is not critical for the compute nodes.

Although redundant CPUs are certainly optional, it is important to consider port density, particularly with respect to 10GE ports, where an extra slot is available in place of a redundant Sup720 module.

**Note**

The examples in this chapter use non-redundant CPU designs, which permit a maximum of 64 10GE ports per 6509 core node available for access node uplink connections based on using a 6708 8-port 10GigE line card.

Server Cluster Design—Two-Tier Model

This section describes the various approaches of a server cluster design that leverages ECMP and distributed CEF. Each design demonstrates how different configurations can achieve various oversubscription levels and can scale in a flexible manner, starting with a few nodes and growing to many that support thousands of servers.

The server cluster design typically follows a two-tier model consisting of core and access layers. Because the design objectives require the use of Layer 3 ECMP and distributed forwarding to achieve a highly deterministic bandwidth and latency per server, a three-tier model that introduces another point of oversubscription is usually not desirable. The advantages with a three-tier model are described in [Server Cluster Design—Three-Tier Model](#).

The three main calculations to consider when designing a server cluster solution are maximum server connections, bandwidth per server, and oversubscription ratio. Cluster designers can determine these values based on application performance, server hardware, and other factors, including the following:

- Maximum number of server GigE connections at scale—Cluster designers typically have an idea of the maximum scale required at initial concept. A benefit of the way ECMP designs function is that they can start with a minimum number of switches and servers that meet a particular bandwidth, latency, and oversubscription requirement, and flexibly grow in a low/non-disruptive manner to maximum scale while maintaining the same bandwidth, latency, and oversubscription values.
- Approximate bandwidth per server—This value can be determined by simply dividing the total aggregated uplink bandwidth by the total server GigE connections on the access layer switch. For example, an access layer Cisco 6509 with four 10GigE ECMP uplinks with 336 server access ports can be calculated as follows:

$$4 \times 10\text{GigE Uplinks with } 336 \text{ servers} = 120 \text{ Mbps per server}$$

Adjusting either side of the equation decreases or increases the amount of bandwidth per server.

**Note**

This is only an approximate value and serves only as a guideline. Various factors influence the actual amount of bandwidth that each server has available. The ECMP load-distribution hash algorithm divides load based on Layer 3 plus Layer 4 values and varies based on traffic patterns. Also, configuration parameters such as rate limiting, queuing, and QoS values can influence the actual achieved bandwidth per server.

- Oversubscription ratio per server—This value can be determined by simply dividing the total number of server GigE connections by the total aggregated uplink bandwidth on the access layer switch. For example, an access layer 6509 with four 10GigE ECMP uplinks with 336 server access ports can be calculated as follows:

336 GigE server connections with 40G uplink bandwidth = 8.4:1 oversubscription ratio

The following sections demonstrate how these values vary, based on different hardware and interconnection configurations, and serve as a guideline when designing large cluster configurations.

**Note**

For calculation purposes, it is assumed there is no line card to switch fabric oversubscription on the Catalyst 6500 Series switch. The dual channel slot provides 40G maximum bandwidth to the switch fabric. A 4-port 10GigE card with all ports at line rate using maximum size packets is considered the best possible condition with little or no oversubscription. The actual amount of switch fabric bandwidth available varies, based on average packet sizes. These calculations would need to be recomputed if you were to use the WS-X6708 8-port 10GigE card which is oversubscribed at 2:1.

4- and 8-Way ECMP Designs with Modular Access

The following four design examples demonstrate various methods of building and scaling the two-tier server cluster model using 4-way and 8-way ECMP. The main issues to consider are the number of core nodes and the maximum number of uplinks, because these directly influence the maximum scale, bandwidth per server, and oversubscription values.

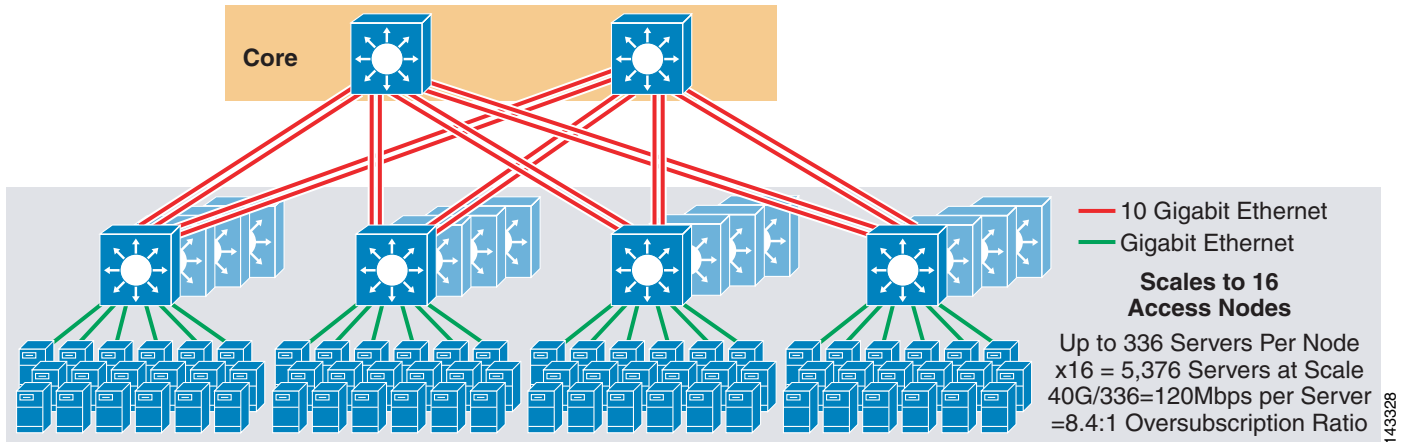
**Note**

Although it has not been tested for this guide, there is a new 8-port 10 Gigabit Ethernet Module (WS-X6708-10G-3C) that has recently been introduced for the Catalyst 6500 Series switch. This line card will be tested for inclusion in the guide at a later date. For questions about the 8-port 10GigE card, refer to the product data sheet.

**Note**

The links necessary to connect the server cluster to an outside campus or metro network are not shown in these design examples but should be considered.

[Figure 3-3](#) provides an example in which two core nodes are used to provide a 4-way ECMP solution.

Figure 3-3 4-Way ECMP using Two Core Nodes

An advantage of this approach is that a smaller number of core switches can support a large number of servers. The possible disadvantage is a high oversubscription-low bandwidth per server value and large exposure to a core node failure. Note that the uplinks are individual L3 uplinks and are not EtherChannels.

Figure 3-4 demonstrates how adding two core nodes to the previous design can dramatically increase the maximum scale while maintaining the same oversubscription and bandwidth per-server values.

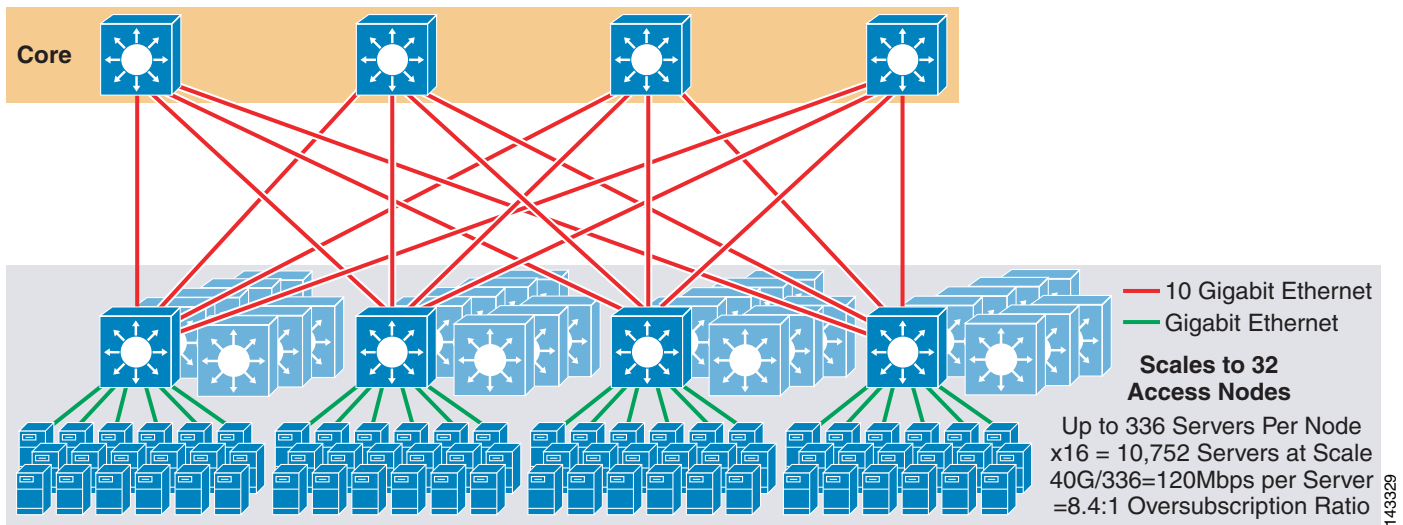
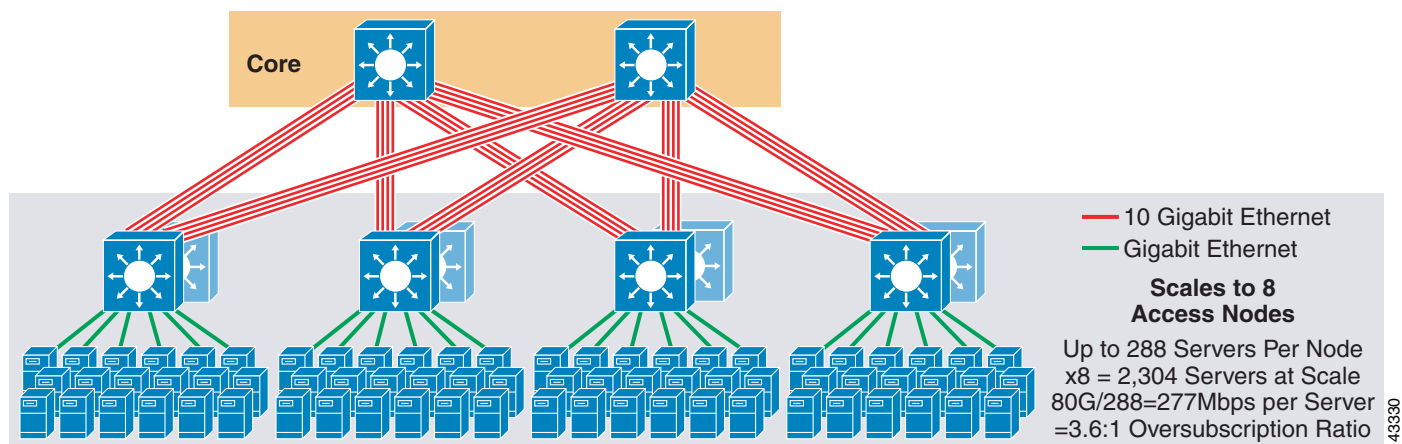
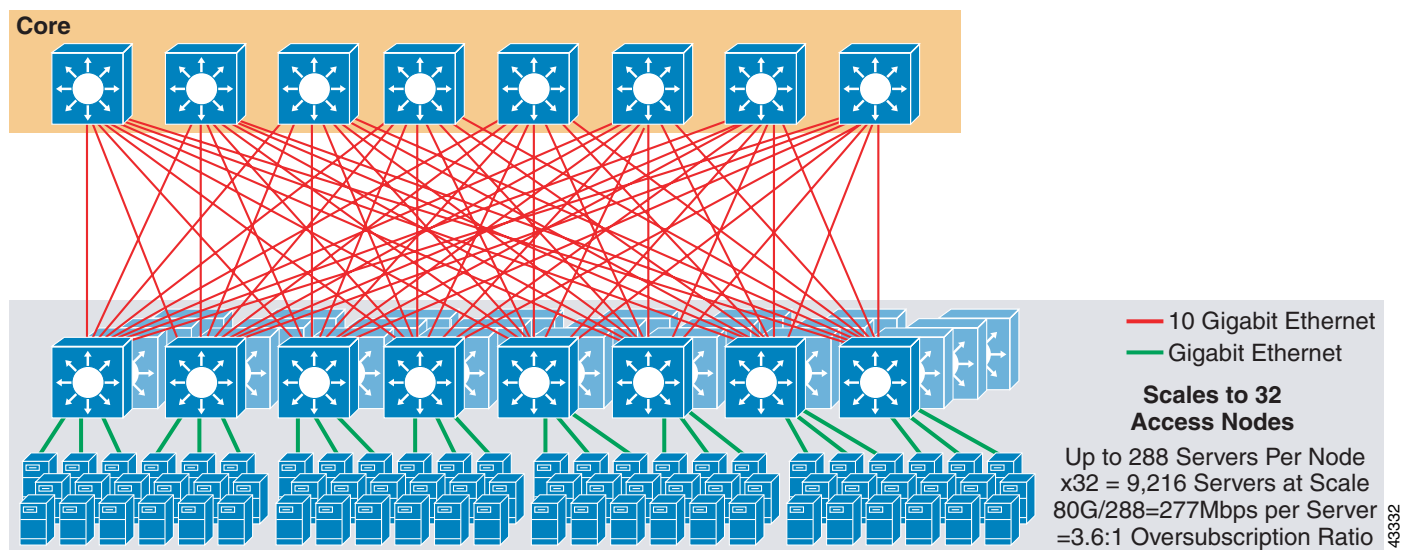
Figure 3-4 4-Way ECMP using Four Core Nodes

Figure 3-5 shows an 8-way ECMP design using two core nodes.

Figure 3-5 8-Way ECMP using Two Core Nodes

As expected, the additional uplink bandwidth dramatically increases the bandwidth per server and reduces the oversubscription ratio per server. Note how the additional slots taken in each access layer switch to support the 8-way uplinks reduces the maximum scale as the number of servers per-switch is reduced to 288. Note that the uplinks are individual L3 uplinks and are not EtherChannels.

Figure 3-6 shows an 8-way ECMP design with eight core nodes.

Figure 3-6 8-Way ECMP using Eight Core Nodes

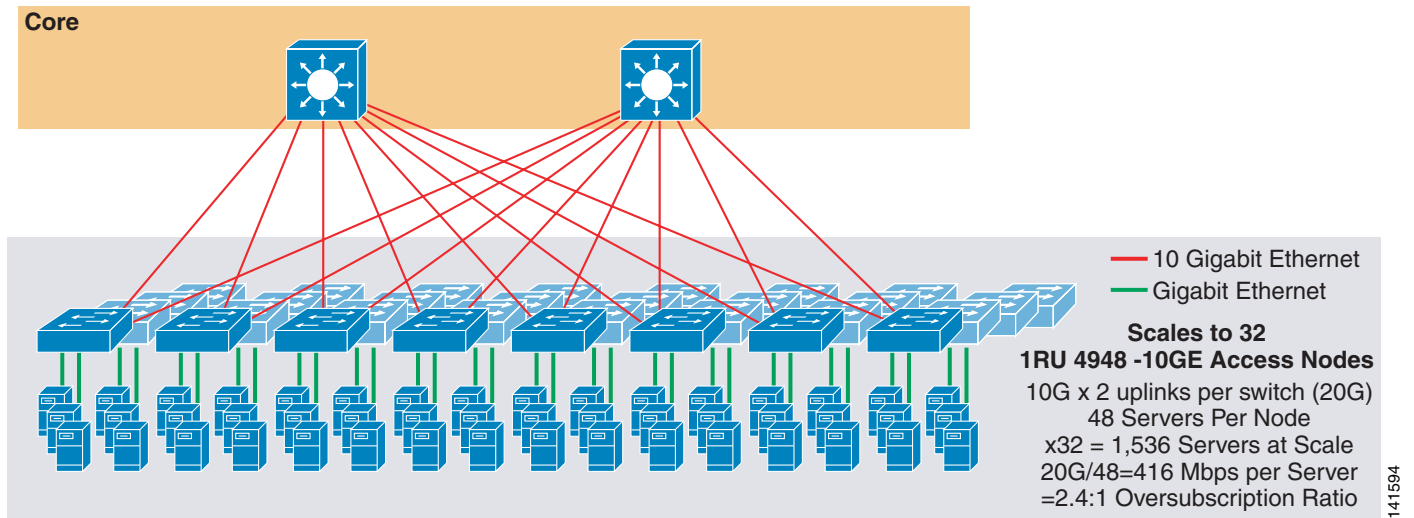
This demonstrates how adding four core nodes to the same previous design can dramatically increase the maximum scale while maintaining the same oversubscription and bandwidth per server values.

2-Way ECMP Design with 1RU Access

In many cluster environments, rack-based server switching using small switches at the top of each server rack is desired or required because of cabling, administrative, real estate issues, or to meet particular deployment model objectives.

Figure 3-7 shows an example in which two core nodes are used to provide a 2-way ECMP solution with 1RU 4948-10GE access switches.

Figure 3-7 2-Way ECMP using Two Core Nodes and 1RU Access



The maximum scale is limited to 1536 servers but provides over 400 Mbps of bandwidth with a low oversubscription ratio. Because the 4948 has only two 10GigE uplinks, this design cannot scale beyond these values.



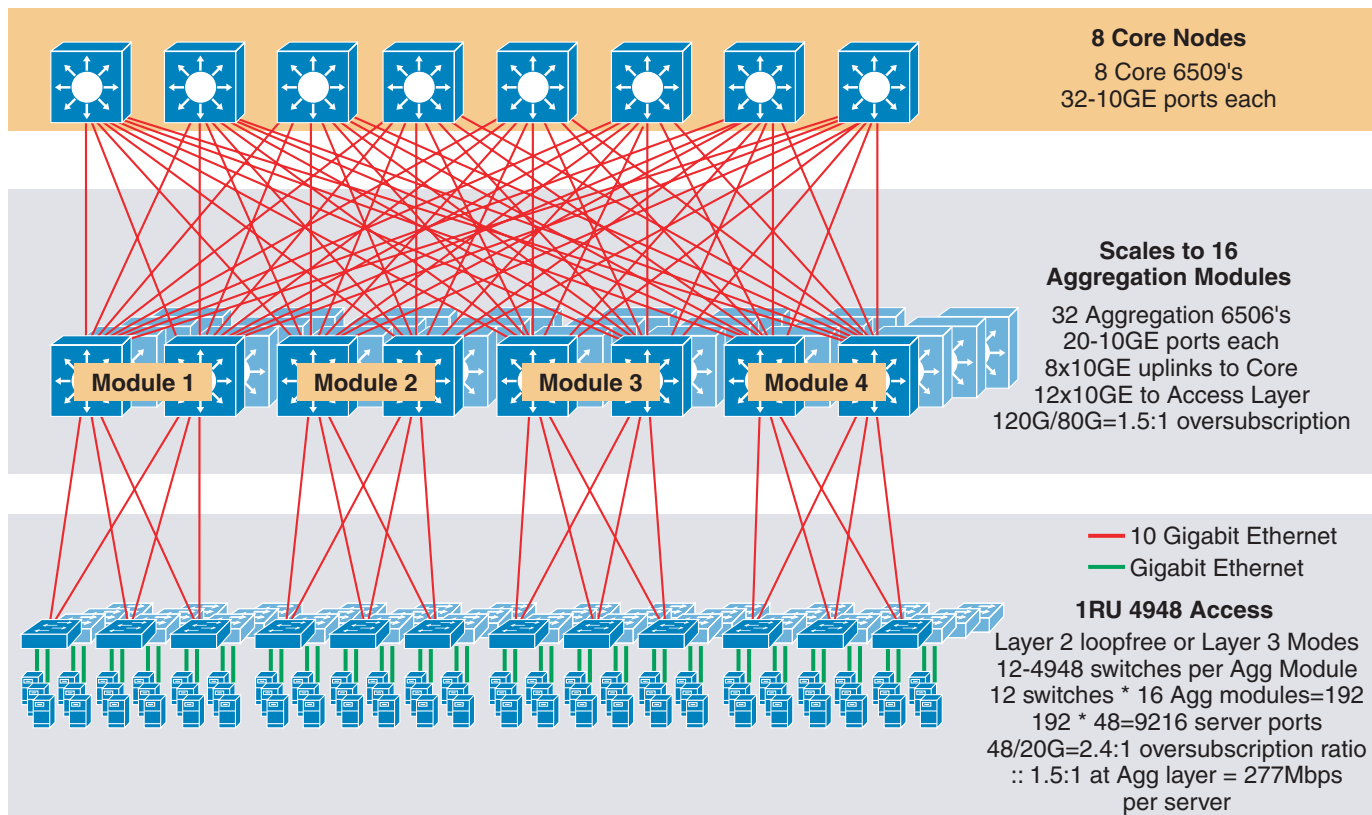
Note

More information on rack-based server switching is provided in [Chapter 3, “Server Cluster Designs with Ethernet.”](#)

Server Cluster Design—Three-Tier Model

Although a two-tier model is most common in large cluster designs, a three-tier model can also be used. The three-tier model is typically used to support large server cluster implementations using 1RU or modular access layer switches.

Figure 3-8 shows a large scale example leveraging 8-way ECMP with 6500 core and aggregation switches and 1RU 4948-10GE access layer switches.

Figure 3-8 Three-Tier Model with 8-Way ECMP

141595

The maximum scale is over 9200 servers with 277 Mbps of bandwidth with a low oversubscription ratio. Benefits of the three-tier approach using 1RU access switches include the following:

- **1RU deployment models**—As mentioned previously, many large cluster model deployments require a 1RU approach for simplified installation. For example, an ASP rolls out racks of servers at a time as they scale large cluster applications. The server rack is pre-assembled and staged offsite such that it can quickly be installed and added to the running cluster. This usually involves a third party that builds the racks, pre-configures the servers, and pre-cables them with power and Ethernet to a 1RU switch. The rack rolls into the data center and is simply plugged in and added to the cluster after connecting the uplinks.

Without an aggregation layer, the maximum size of the 1RU access model is limited to just over 1500 servers. Adding an aggregation layer allows the 1RU access model to scale to a much larger size while still leveraging the ECMP model.

- **Centralization of core and aggregation switches**—With 1RU switches deployed in the racks, it is possible to centralize the larger core and aggregation modular switches. This can simplify power and cabling infrastructure and improve rack real estate usage.
- **Permits Layer 2 loop-free topology**—A large cluster network using Layer 3 ECMP access can use a lot of address space on the uplinks and can add complexity to the design. This is particularly important if public address space is used. The three-tier model approach lends itself well to a Layer 2 loop-free access topology that reduces the number of subnets required.

When a Layer 2 loop-free model is used, it is important to use a redundant default gateway protocol such as HSRP or GLBP to eliminate a single point of failure if an aggregation node fails. In this design, the aggregation modules are not interconnected, permitting a loop-free Layer 2 design that can leverage GLBP

for automatic server default gateway load balancing. GLBP automatically distributes the servers default gateway assignment between the two nodes in the aggregation module. After a packet arrives at the aggregation layer, it is balanced across the core using the 8-way ECMP fabric. Although GLBP does not provide a Layer 3/Layer 4 load distribution hash similar to CEF, it is an alternative that can be used with a Layer 2 access topology.

Calculating Oversubscription

The three-tier model introduces two points of oversubscription at the access and aggregation layers, as compared to the two-tier model that has only a single point of oversubscription at the access layer. To properly calculate the approximate bandwidth per server and the oversubscription ratio, perform the following two steps, which use [Figure 3-8](#) as an example:

Step 1 Calculate the oversubscription ratio and bandwidth per server for both the aggregation and access layers independently.

- Access layer
 - Oversubscription—48GE attached servers/20G uplinks to aggregation = 2.4:1
 - Bandwidth per server—20G uplinks to aggregation/48GigE attached servers = 416Mbps
- Aggregation layer
 - Oversubscription—120G downlinks to access/80G uplinks to core = 1.5:1

Step 2 Calculate the combined oversubscription ratio and bandwidth per server.

The actual oversubscription ratio is the sum of the two points of oversubscription at the access and aggregation layers.

$$1.5 + 2.4 = 3.9:1$$

To determine the true bandwidth per server value, use the algebraic formula for proportions:

$$a:b = c:d$$

The bandwidth per server at the access layer has been determined to be 416 Mbps per server. Because the aggregation layer oversubscription ratio is 1.5:1, you can apply the above formula as follows:

$$416:1 = x:1.5$$

$$x \approx 277 \text{ Mbps per server}$$

Recommended Hardware and Modules

The recommended platforms for the server cluster model design consist of the Cisco Catalyst 6500 family with the Sup720 processor module and the Catalyst 4948-10GE 1RU switch. The high switching rate, large switch fabric, low latency, distributed forwarding, and 10GigE density makes the Catalyst 6500 Series switch ideal for all layers of this model. The 1RU form factor combined with wire rate forwarding, 10GE uplinks, and very low constant latency makes the 4948-10GE an excellent top of rack solution for the access layer.

The following are recommended:

- Sup720—The Sup720 can consist of both PFC3A (default) or the newer PFC3B type daughter cards.

- Line cards—All line cards should be 6700 Series and should all be enabled for distributed forwarding with the DFC3A or DFC3B daughter cards.

**Note**

By using all fabric-attached CEF720 series modules, the global switching mode is *compact*, which allows the system to operate at its highest performance level. The Catalyst 6509 can support 10 GigE modules in all positions because each slot supports dual channels to the switch fabric (the Cisco Catalyst 6513 does not support this).

- Cisco Catalyst 4948-10GE—The 4948-10GE provides a high performance access layer solution that can leverage ECMP and 10GigE uplinks. No special requirements are necessary. The 4948-10GE can use a Layer 2 Cisco IOS image or a Layer 2/3 Cisco IOS image, permitting an optimal fit in either environment.



Data Center Design Considerations



Note

The README file posted with this guide contains details about the technologies, hardware, and software that were used in producing this document. The README file also contains a revision history that details updates made to each chapter.

This chapter describes factors that influence the enterprise data center design. The following topics are included:

- [Factors that Influence Scalability](#)
- [Server Clustering](#)
- [NIC Teaming](#)
- [Pervasive 10GigE](#)
- [Server Consolidation](#)
- [Top of Rack Switching](#)
- [Blade Servers](#)
- [Importance of Team Planning](#)

Factors that Influence Scalability

Determining scalability is never an easy task because there are always unknown factors and inter-dependencies. This section examines some of the most common scalability-related questions that arise when designing a data center network.

Why Implement a Data Center Core Layer?

Do I need a separate core layer for the data center? Can I use my existing campus core?

The campus core can be used as the data center core. The recommendation is to consider the long-term requirements so that a data center core does not have to be introduced at a later date. Advance planning helps avoid disruption to the data center environment. Consider the following items when determining the right core solution:

- 10GigE density—Will there be enough 10GigE ports on the core switch pair to support both the campus distribution as well as the data center aggregation modules?

- Administrative domains and policies—Separate cores help to isolate campus distribution layers from data center aggregation layers in terms of troubleshooting, administration, and policies (QoS, ACLs, troubleshooting, and maintenance).
- Future anticipation—The impact that can result from implementing a separate data center core layer at a later date might make it worthwhile to install it at the beginning.

Why Use the Three-Tier Data Center Design?

Why not connect servers directly to a distribution layer and avoid installing an access layer?

The three-tier approach consisting of the access, aggregation, and core layers permit flexibility in the following areas:

- Layer 2 domain sizing— When there is a requirement to extend a VLAN from one switch to another, the domain size is determined at the distribution layer. If the access layer is absent, the Layer 2 domain must be configured across the core for extension to occur. Extending Layer 2 through a core causes path blocking by spanning tree and has the risk of uncontrollable broadcast issues related to extending Layer 2 domains, and therefore should be avoided.
- Service modules—An aggregation plus access layer solution enables services to be shared across the entire access layer of switches. This lowers TCO and lowers complexity by reducing the number of components to configure and manage. Consider future service capabilities that include Application-Oriented Networking (AON), ACE, and others.
- Mix of access layer models—The three-tier approach permits a mix of both Layer 2 and Layer 3 access models with 1RU and modular platforms, permitting a more flexible solution and allowing application environments to be optimally positioned.
- NIC teaming and HA clustering support—Supporting NIC teaming with switch fault tolerance and high availability clustering requires Layer 2 adjacency between NIC cards, resulting in Layer 2 VLAN extension between switches. This would also require extending the Layer 2 domain through the core, which is not recommended.

Determining Maximum Servers

What is the maximum number of servers that should be on an access layer switch? What is the maximum number of servers to an aggregation module?

The answer is usually based on considering a combination of oversubscription, failure domain sizing, and port density. No two data centers are alike when these aspects are combined. The right answer for a particular data center design can be determined by examining the following areas:

- Oversubscription—Applications require varying oversubscription levels. For example, the web servers in a multi-tier design can be optimized at a 15:1 ratio, application servers at 6:1, and database servers at 4:1. An oversubscription ratio model helps to determine the maximum number of servers that should be placed on a particular access switch and whether the uplink should be Gigabit EtherChannel or 10GE. It is important for the customer to determine what the oversubscription ratio should be for each application environment. The following are some of the many variables that must be considered when determining oversubscription:
 - NIC—Interface speed, bus interface (PCI, PCI-X, PCI-E)
 - Server platform—Single or dual processors, offload engines
 - Application characteristics—Traffic flows, inter-process communications

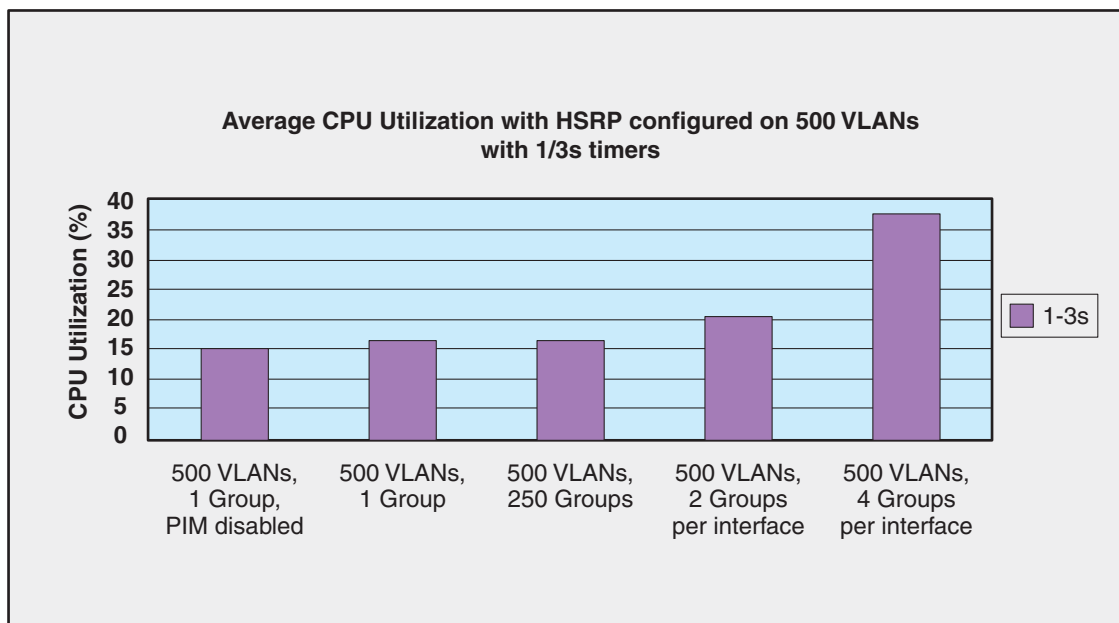
- Usage characteristics—Number of clients, transaction rate, load balancing
- Failure domain sizing—This is a business decision and should be determined regardless of the level of resiliency that is designed into the network. This value is not determined based on MTBF/MTTR values and is not meant to be a reflection of the robustness of a particular solution. No network design should be considered immune to failure because there are many uncontrollable circumstances to consider, including human error and natural events. The following areas of failure domain sizing should be considered:
 - Maximum number of servers per Layer 2 broadcast domain
 - Maximum number of servers per access switch (if single-homed)
 - Maximum number of servers per aggregation module
 - Maximum number of access switches per aggregation module
- Port density—The aggregation layer has a finite number of 10GigE ports that can be supported, which limits the quantity of access switches that can be supported. When a Catalyst 6500 modular access layer is used, thousands of servers can be supported on a single aggregation module pair. In contrast, if a 1RU Catalyst 4948 is used at the access layer, the number of servers supported is less. Cisco recommends leaving space in the aggregation layer for growth or changes in design.

The data center, unlike other network areas, should be designed to have flexibility in terms of emerging services such as firewalls, SSL offload, server load balancing, AON, and future possibilities. These services will most likely require slots in the aggregation layer, which would limit the amount of 10GigE port density available.

Determining Maximum Number of VLANs

What is the maximum number of VLANs that can be supported in an aggregation module?

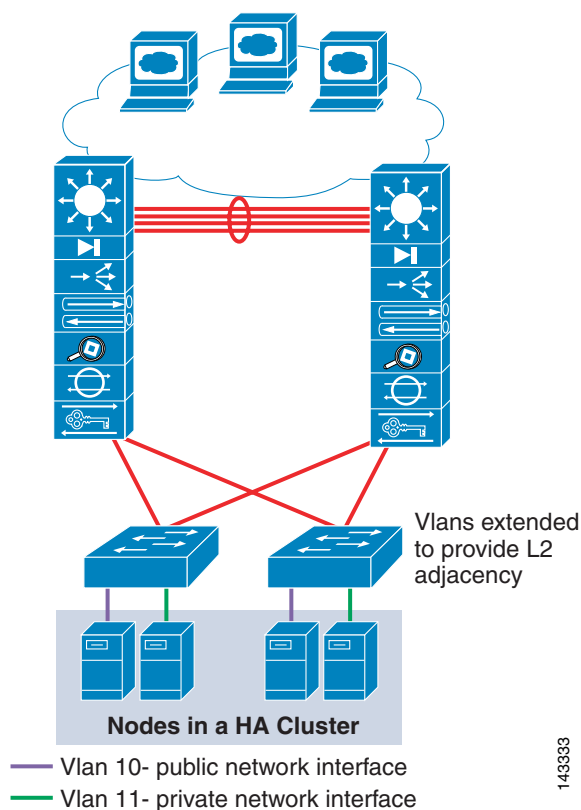
- Spanning tree processing—When a Layer 2 looped access topology is used, which is the most common, the amount of spanning tree processing at the aggregation layer needs to be considered. There are specific watermarks related to the maximum number of system-wide active logical instances and virtual port instances per line card that, if reached, can adversely affect convergence and system stability. These values are mostly influenced by the total number of access layer uplinks and the total number of VLANs. If a data center-wide VLAN approach is used (no manual pruning on links), the watermark maximum values can be reached fairly quickly. More details and recommendations are provided in [Chapter 5, “Spanning Tree Scalability.”](#)
- Default Gateway Redundancy Protocol— The quantity of HSRP instances configured at the aggregation layer is usually equal to the number of VLANs. As Layer 2 adjacency requirements continue to gain importance in data center design, proper consideration for the maximum HSRP instances combined with other CPU-driven features (such as GRE, SNMP, and others) have to be considered. Lab testing has shown that up to 500 HSRP instances can be supported in an aggregation module, but close attention to other CPU driven features must be considered. The graph in [Figure 4-1](#) shows test results when using 500 VLANs with one or multiple groups with the hello and holddown timer configuration at 1/3 seconds.

Figure 4-1 Graphed Average CPU Utilization with HSRP

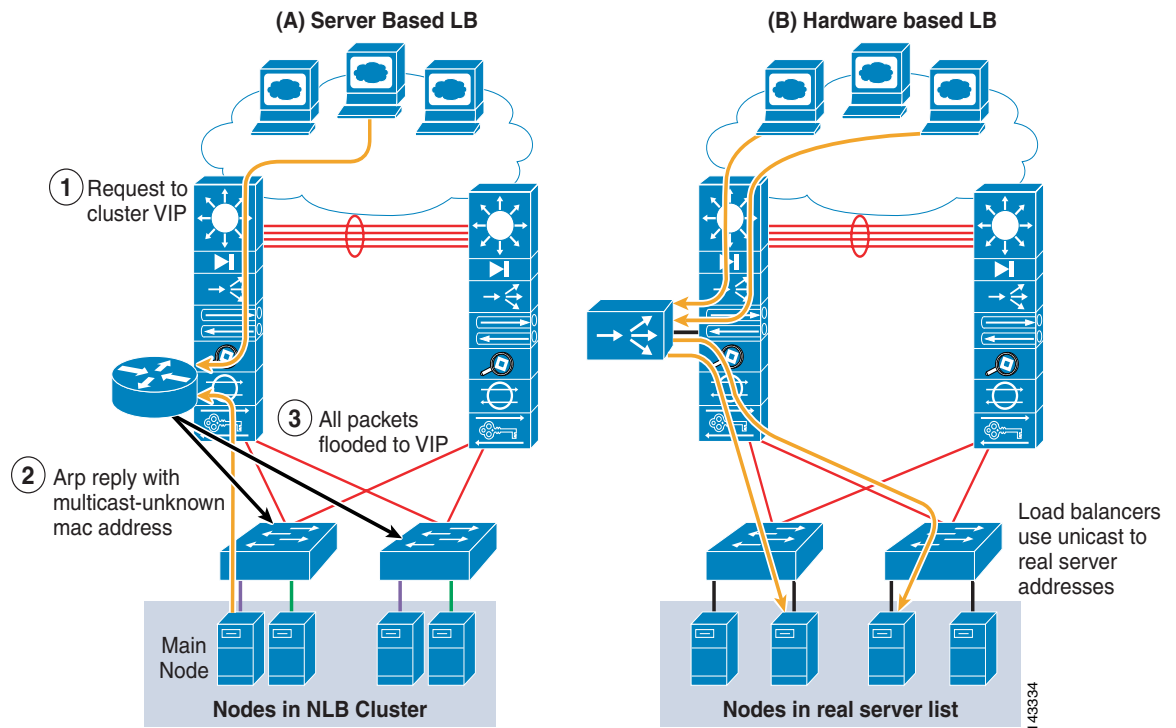
Server Clustering

The goal of server clustering is to combine multiple servers so that they appear as a single unified system through special software and network interconnects. “Clusters” were initially used with the Digital Equipment Corporation VAX VMS Clusters in the late 1980s. Today, “clustering” is a more general term that is used to describe a particular type of grouped server arrangement that falls into the following four main categories:

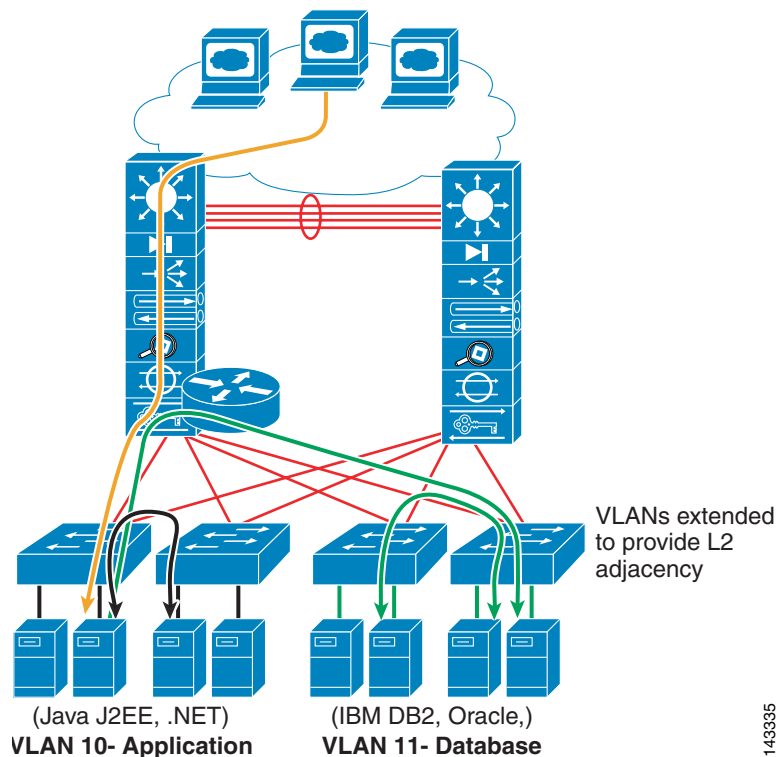
- High availability clusters—This type of cluster uses two or more servers and provides redundancy in the case of a server failure (see [Figure 4-2](#)). If one node fails, another node in the cluster takes over with little or no disruption. This type of cluster is usually up to a maximum of eight nodes and requires Layer 2 adjacency between their public and private interfaces. High availability clusters are common in the data center multi-tier model design.

Figure 4-2 High Availability Cluster

- Network load balanced clusters (NLBs)—This type of cluster typically supports up to a maximum of 32 servers that work together to load balance HTTP sessions on a website. It uses a broadcast mechanism in which the ARP reply from the main server to the gateway router is an unknown MAC address, so that all packets to the destination web site address are essentially broadcast to all servers in the VLAN. This type of implementation is usually much less robust than hardware-based load balancing solutions, and requires Layer 2 adjacency. Hardware-based server load balancers such as the Cisco CSM provide a unicast-based solution, scale beyond 32 servers, and provide many value-added features. NLB clusters are common in the data center multi-tier model design. [Figure 4-3](#) illustrates both a server based load balancing solution and a hardware-based load balancing solution.

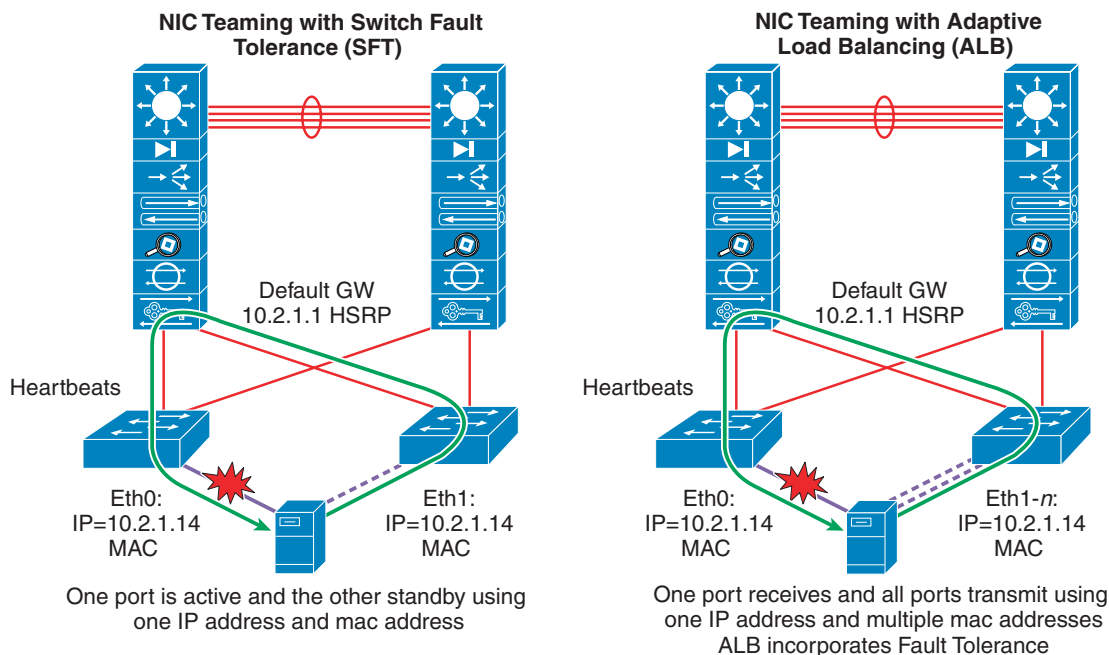
Figure 4-3 Network Load Balanced Clusters

- Database clusters—As databases become larger, the ability to search the database becomes more complex and time sensitive. Database clusters provide a way to enable efficient parallel scans and improve database lock times. Some examples of parallel database implementations are Oracle RAC and IBM DB2. These implementations also require Layer 2 adjacency between the servers. Database clusters are typically two to eight nodes in size and are common in the data center multi-tier model design. Figure 4-4 illustrates how the application and database layers communicate across the aggregation layer router and how the interfaces require Layer 2 adjacency within each layer, resulting in VLANs being extended across multiple access layer switches.

Figure 4-4 Database Clusters

NIC Teaming

Servers with a single Network Interface Card (NIC) interface can have many single points of failure. The NIC card, the cable, and the switch to which it connects are all single points of failure. NIC teaming is a solution developed by NIC card vendors to eliminate this single point of failure by providing special drivers that allow two NIC cards to be connected to two different access switches or different line cards on the same access switch. If one NIC card fails, the secondary NIC card assumes the IP address of the server and takes over operation without disruption. The various types of NIC teaming solutions include active/standby and active/active. All solutions require the NIC cards to have Layer 2 adjacency with each other. NIC teaming solutions are common in the data center multi-tier model design and are shown in [Figure 4-5](#).

Figure 4-5 NIC Teaming Configurations

143336

Note the following:

- Switch fault tolerance (SFT)—With SFT designs, one port is active and the other standby using one common IP address and MAC address.
- Adaptive load balancing (ALB) —With ALB designs, one port receives and all ports transmit using one IP address and multiple MAC addresses.

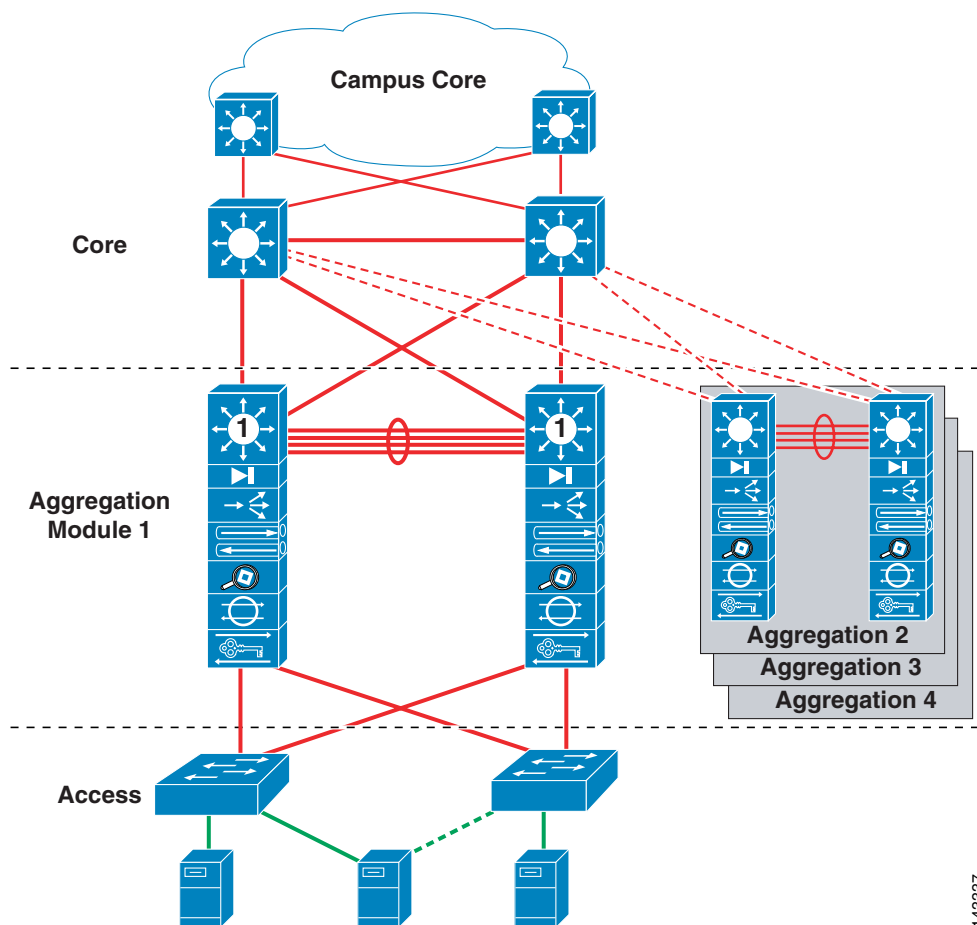
Pervasive 10GigE

Customers are seeing the benefit of moving beyond Gigabit or Gigabit EtherChannel implementations to 10GigE, which includes benefits such as the following:

- Improving IP-based storage access (iSCSI)
- Improving network use of SMP-based servers including virtual machine implementations
- Improving access layer uplink use because of Gigabit EtherChannel hashing algorithm barriers
- Improving server backup and recovery times
- Improving NAS performance

Many customers are also moving to 10GigE access layer uplinks in anticipation of future requirements. The implications relative to this trend are usually related to density in the aggregation layer.

A proven method of increasing aggregation layer 10GigE ports is to use multiple aggregation modules, as shown in [Figure 4-6](#).

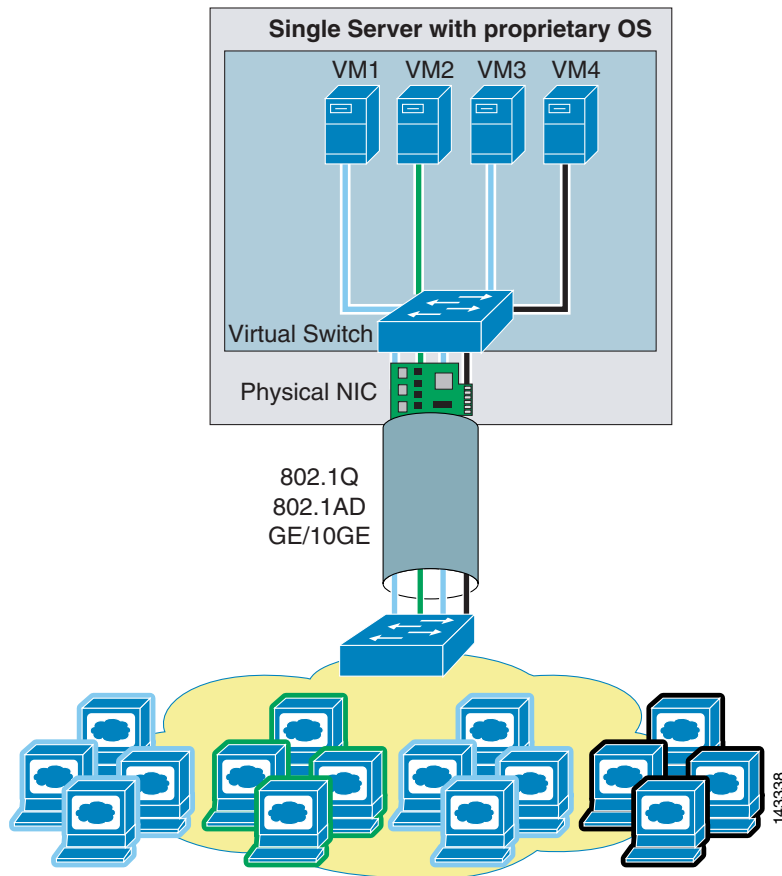
Figure 4-6 Multiple Aggregation Modules

Aggregation modules provide a way to scale 10GigE port requirements while also distributing CPU processing for spanning tree and HSRP. Other methods to increase port density include using the service layer switch, which moves service modules out of the aggregation layer and into a standalone chassis, making slots available for 10GigE ports. Other methods of improving 10GE density are in the access layer design topology used such as a looped square topology. More on this subject is covered in [Chapter 2, “Data Center Multi-Tier Model Design,”](#) and [Chapter 6, “Data Center Access Layer Design.”](#)

Server Consolidation

The majority of servers in the data center are underutilized in terms of CPU and memory; particularly the web server tier and development environment server resources. Virtual machine solutions are being used to solve this deficiency and to improve the use of server resources.

- The virtual machine solution is a vendor software product that can install multiple server images on a single hardware server platform to make it appear the same as multiple, separate physical servers. This was initially seen in development or lab environments but is now a production solution in the enterprise data center. The virtual machine solution is supported on small single processor server platforms to large SMP platforms with greater memory support and multi-processors. This software-based solution allows over 32 virtual machines to coexist on the same physical server. [Figure 4-7](#) shows a server with multiple virtual machine instances running on it.

Figure 4-7 *Multiple Virtual Machines*

Virtual machine solutions can be attached to the network with multiple GE network interfaces, one for each virtual host implementation. Other implementations include the use of 802.1Q on GE or 10GE interfaces to connect virtual hosts directly to a specific VLAN over a single interface. Although 10GE interfaces are not supported on all virtual machine solutions available today, it is expected that this will change in the near future. This requirement could have implications on access layer platform selection, NIC teaming support, and in determining uplink oversubscription values.

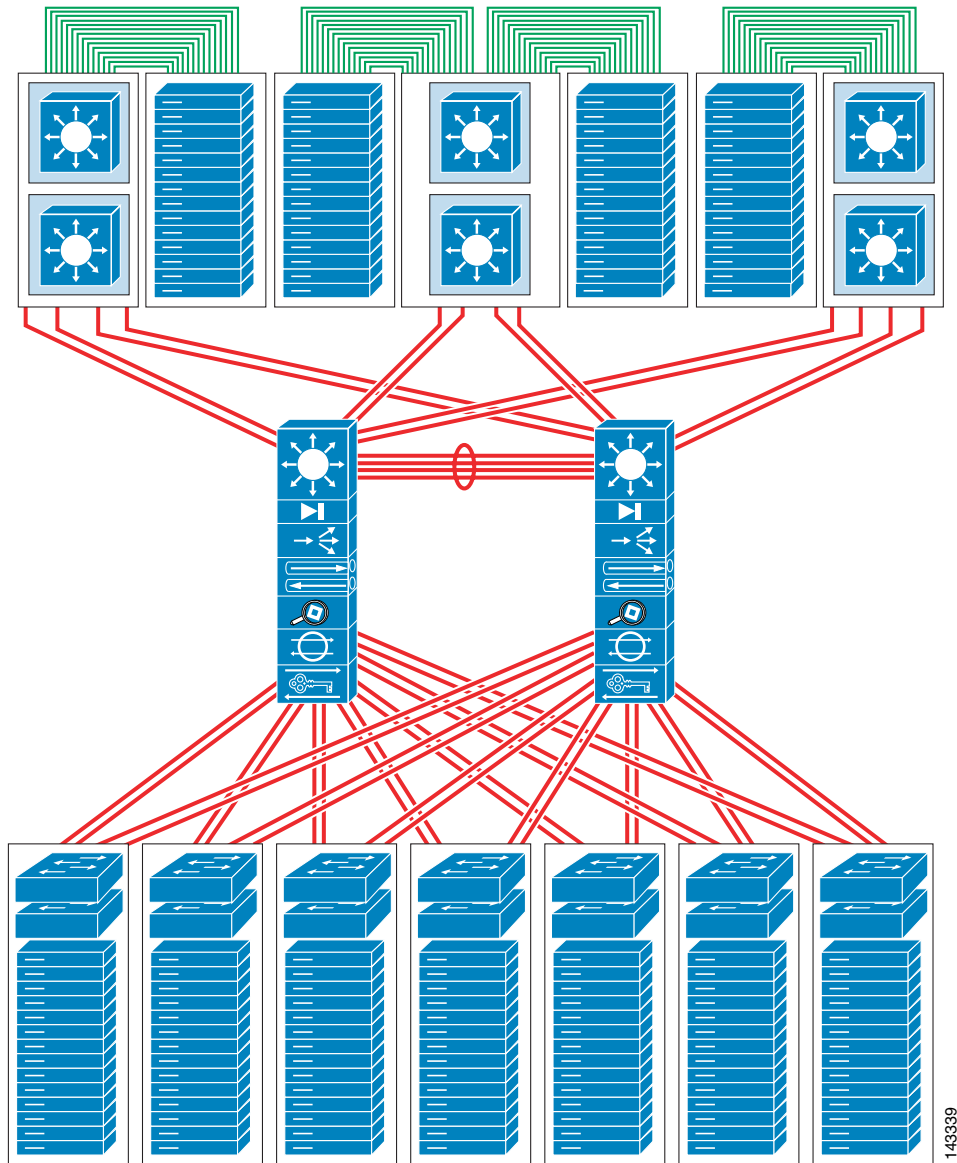
Top of Rack Switching

The most common access layer topology in the enterprise today is based on the modular chassis Catalyst 6500 or 4500 Series platforms. This method has proven to be a very scalable method of building out server farms that are providing high density, high speed uplinks, and redundant power and processors. Although this approach has been very successful, it has certain challenges related to the environments of data centers. The enterprise data center is experiencing a large amount of growth in the sheer number of servers while at the same time server density has been improved with 1RU and blade server solutions. Three particular challenges that result from this trend are related to the following:

- **Cable bulk**—There are typically three to four interfaces connected on a server. With a higher density of servers per rack, cable routing and management can become quite difficult to manage and maintain.

- **Power**—The increased density of components in the rack is driving a need for a larger power feed to the rack. Many data centers do not have the power capacity at the server rows to support this increase.
- **Cooling**—The amount of cables laying under the raised floor and the cable bulk at the cabinet base entry is blocking necessary airflow required to cool equipment in the racks. At the same time, the servers in the rack are requiring more cooling volume because of their higher density.

These challenges have forced customers to find alternative solutions by spacing out cabinets, modifying cable routes, or other means, including to not deploy high density server solutions. Another way that customers are seeking to solve some of these problems is by using a rack-based switching solution. By using 1RU top of rack switches, the server interface cables are kept in the cabinet, reducing the amount of cabling in the floor and thus reducing the cabling and cooling issues. [Figure 4-8](#) shows both a modular (top) and rack-based (bottom) access layer approach.

Figure 4-8 *Modular and 1RU Access Layers*

The upper half of [Figure 4-8](#) has the following characteristics:

- Less devices to manage
- Increased cabling and cooling challenges
- Lower spanning tree processing

The lower half of [Figure 4-8](#) has the following characteristics:

- More devices to manage
- Less cabling and cooling challenges
- Higher spanning tree processing
- Uplink density challenge at the aggregation layer

When considering a 1RU top of rack switch implementation, the following should be considered:

- 10GigE density—The increase in uplinks requires higher 10GigE density at the aggregation layer or additional aggregation modules.
- Spanning tree virtual and logical ports—For Layer 2 looped access layer topologies, the increase in uplinks increases the STP active logical and virtual port per line card instances at the aggregation layer, creating more overhead and processing requirements.
- How many 1RU switches per rack?—The maximum number of ports that might need to be connected in a worst case scenario could create a need for three, four, or more 1RU switches in the rack. This has obvious cost issues and further impacts 10GigE density and STP overhead.
- Management—More switches mean more elements to manage, adding complexity.

Blade Servers

Blade-server chassis have become very popular in the enterprise data center, driven mostly by the IBM BladeCenter, HP BladeServer, and Dell Blade Server products. Although the blade server seeks to reduce equipment footprint, improve integration, and improve management, it has the following specific challenges related to designing and supporting the data center network:

- Administrative domains—Blade server products can support either integrated switches or pass-through modules for connecting its servers to the network. Who is responsible for configuring and managing these integral switches? Usually the system administration team is responsible for the components inside of a server product. So, who configures spanning tree? How should the trunks be configured? How are change control and troubleshooting supported? It is important for customers to address these questions before implementation.
- Interoperability—Blade servers support many different vendor-integral switches, including Cisco, Nortel, and D-Link, to name a few. Although many of the technologies in use are expected to meet interoperability standards such as spanning tree 802.1w, they must be verified and tested to ensure proper operation.
- Spanning tree scaling—The integral switch on the blade server is logically similar to the external rack-based server switching design. The same challenges apply relative to the increase in spanning tree logical/virtual ports.
- Pass-through cabling—The pass-through module option on blade servers permits customers to use their existing external access switches for connecting the servers in the blade server chassis and to avoid the integral switch option. Customers should examine the pass-through cabling system to make sure it can properly be supported in their cabinets.
- Topologies—Each vendor blade server implementation has unique internal and external switch trunk connectivity options. Careful consideration should be taken in determining the proper access layer topology that meets the requirements such as VLAN extension and NIC teaming while staying within the watermark values of spanning tree design.

Importance of Team Planning

Considering the roles of different personnel in an IT organization shows that there is a growing need for team planning with data center design efforts. The following topics demonstrate some of the challenges that the various groups in an IT organization have related to supporting a “business ready” data center environment:

- System administrators usually do not consider physical server placement or cabling to be an issue in providing application solutions. When the need arises for one server to be connected into the same VLAN as other servers, it is usually expected to simply happen without thought or concern about possible implications. The system administrators are faced with the challenge of being business-ready and must be able to deploy new or to scale existing applications in a timely fashion.
- Network administrators have traditionally complied with these requests by extending the VLAN across the Layer 2 looped topology and supporting the server deployment request. This is the flexibility of having a Layer 2 looped access layer topology, but is becoming more of a challenge now than it was in the past. The Layer 2 domain diameters are getting larger, and now the network administrator is concerned with maintaining spanning tree virtual/logical port counts, manageability, and the failure exposure that exists with a large Layer 2 broadcast domain. Network designers are faced with imposing restrictions on server geography in an effort to maintain spanning tree processing, as well as changing design methods to include consideration for Layer 2 domain sizing and maximum failure domain sizing.
- Facilities administrators are very busy trying to keep all this new dense hardware from literally burning up. They also see the additional cabling as very difficult if not impossible to install and support with current design methods. The blocked air passages from the cable bulk can create serious cooling issues, and they are trying to find ways to route cool air into hot areas. This is driving the facilities administrators to look for solutions to keep cables minimized, such as when using 1RU switches. They are also looking at ways to locate equipment so that it can be cooled properly.

These are all distinct but related issues that are growing in the enterprise data center and are creating the need for a more integrated team planning approach. If communication takes place at the start, many of the issues are addressed, expectations are set, and the requirements are understood across all groups.



Spanning Tree Scalability



Note

The README file posted with this guide contains details relative to technologies, hardware, and software that were used in producing this document. There is also a revision history that details updates made to each chapter.

This chapter provides details about spanning tree design in the data center.

Extending VLANs in the Data Center

The ability to extend VLANs across the data center is not only necessary to meet application requirements such as Layer 2 adjacency, but to permit a high level of flexibility in administering the servers. Many customers require the ability to group and maintain departmental servers together in a common VLAN or IP subnet address space. This makes management of the data center environment easier with respect to additions, moves, and changes.

For example, consider a medium-to-large data center environment where rows of servers cover a fairly large area. The HR department might have a small cluster of web and application servers physically located in rack 4 of row A in the main data center. They need to add another server that requires Layer 2 adjacency with the existing application cluster. All available space in row A is currently filled. The new server is physically located in row F, which is currently served by a different access layer switch. Because the data center design uses a Layer 2 loop-based topology, it is quite simple to add the new server port in row F by extending the VLAN to the associated access layer switch. Without this type of data center topology, the new server would have to be physically moved or located closer to the existing access layer switch, which can be disruptive and does not provide a very flexible or scalable data center design. This example demonstrates the type of flexibility that is becoming increasingly important in the data center.

Efforts to consolidate many data centers into a few is also driving the need for larger Layer 2 domains. As data centers become larger, the geographical distance between servers is increasing, creating more opportunities for the need to extend VLANs. Customers want to meet server installation requirements without the need to place them in the same physical proximity. There are many factors associated with having to physically rearrange servers in the data center, including such things as labor and administrative costs, and possible server downtime, driving up the overall total cost of ownership.

1RU access layer designs places a smaller number of servers on an access switch. This in itself creates the likelihood that the use of VLAN extension will also increase. 1RU switches also increase the number of uplinks required on the aggregation layer switch, which increases the number of STP logical and virtual ports on the aggregation layer switches. [Table 5-1](#) provides an example of a 4000 server data center design.

Table 5-1 4000 Server Data Center Design

Modular Access w/4000 servers	1RU Access w/4000 servers
200 servers per 6509=20 access layer 6509s	48 servers per 4948=84 access layer 4948s
Uplink Ports Required (port-channel or 10GE)	Uplink Ports Required (port-channel or 10GE)
20x2 with triangle loop access	84x2 with triangle loop access
10x2 with square loop access	42x2 with square loop access

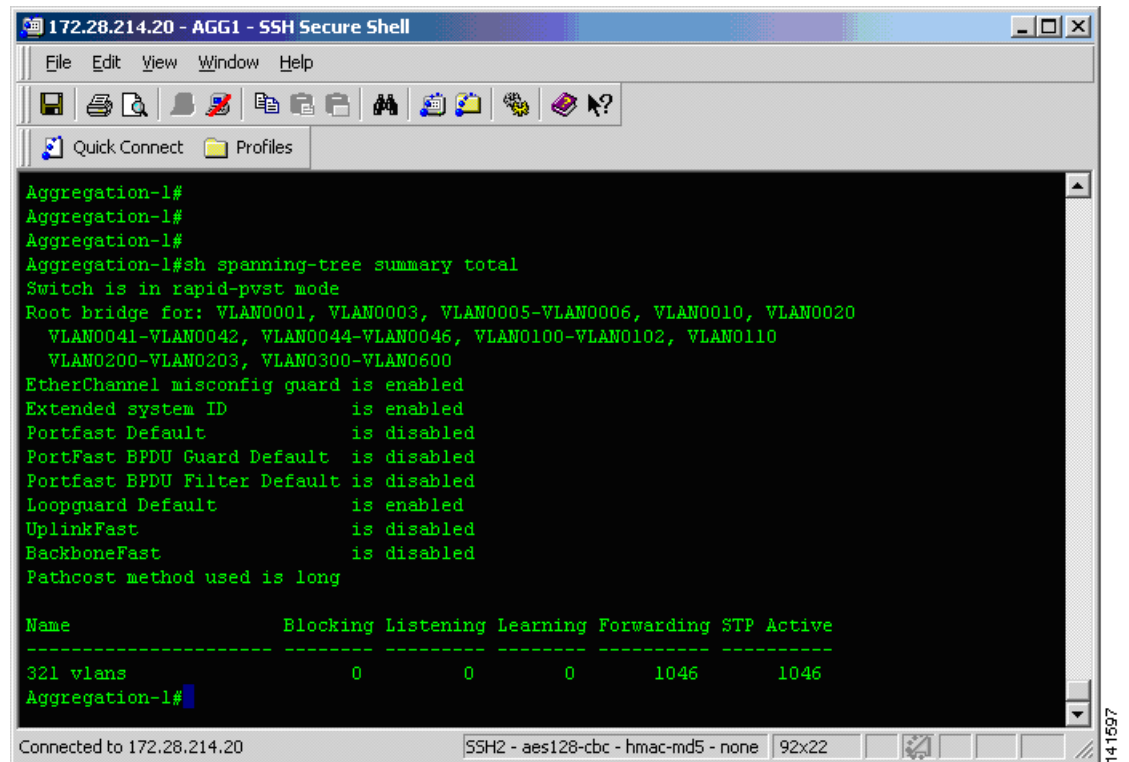
With a triangle loop access topology, the modular access requires a total of 40 uplinks (20 per agg switch) while the 1RU requires 168 for the same number of servers. This can have a large impact on the spanning tree processing at the aggregation layer. A subsequent section of this chapter covers how uplinks affect the total spanning tree logical and virtual port counts, which can impact performance and stability of a Layer 2 looped design.

When using a Layer 2 looped topology, a loop protection mechanism such as a Spanning Tree Protocol (STP) type is required. STPs automatically break loops, preventing broadcast packets from continuously circulating and melting down the network. The STPs recommended in the data center design consist of 802.1w-Rapid PVST+ and 802.1s-MST. Both 802.1w and 802.1s have the same quick convergence characteristics but differ in flexibility and operation.

STP Active Logical Ports and Virtual Ports per Line Card

In a Layer 2 looped topology design, spanning tree processing instances are created on each interface for each active VLAN. These logical instances are used by the spanning tree process in processing the spanning tree-related packets for each VLAN. These instances are referred to as active logical ports and virtual ports. Both active logical ports and virtual ports are important values to consider in spanning tree designs because they affect STP convergence time and stability. These values are usually only of concern on the aggregation layer switches because they typically have a larger number of trunks and VLANs configured than other layers in the data center topology. The rest of this section is focused on these values in the aggregation layer for this reason.

Active logical ports are a system-wide value that reflects the total number of spanning tree processing instances used in the whole system. This value can be determined by entering a **show spantree summary total** command on the console, as shown in [Figure 5-1](#).

Figure 5-1 Total Active Logical Ports Used


```

172.28.214.20 - AGG1 - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

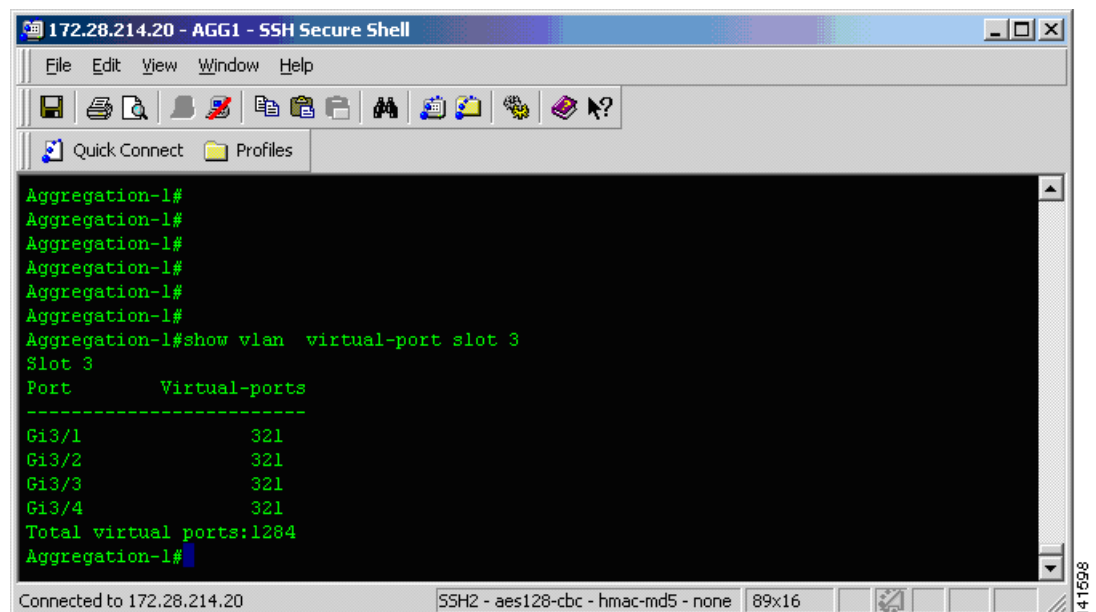
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#sh spanning-tree summary total
Switch is in rapid-pvst mode
Root bridge for: VLAN0001, VLAN0003, VLAN0005-VLAN0006, VLAN0010, VLAN0020
VLAN0041-VLAN0042, VLAN0044-VLAN0046, VLAN0100-VLAN0102, VLAN0110
VLAN0200-VLAN0203, VLAN0300-VLAN0600
EtherChannel misconfig guard is enabled
Extended system ID is enabled
Portfast Default is disabled
PortFast BPDU Guard Default is disabled
Portfast BPDU Filter Default is disabled
Loopguard Default is enabled
UplinkFast is disabled
BackboneFast is disabled
Pathcost method used is long

Name Blocking Listening Learning Forwarding STP Active
-----
321 vlans 0 0 0 1046 1046
Aggregation-1#

```

Connected to 172.28.214.20 SSH2 - aes128-cbc - hmac-md5 - none 92x22 141697

Virtual ports are a per-line card value that reflects the total number of spanning tree processing instances used on a particular line card. This value can be determined by entering a **show vlan virtual-port slot X** command on the console, as shown in Figure 5-2.

Figure 5-2 Total Virtual Ports used Per Line Card


```

172.28.214.20 - AGG1 - SSH Secure Shell
File Edit View Window Help
Quick Connect Profiles

Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#show vlan virtual-port slot 3
Slot 3
Port Virtual-ports
-----
Gi3/1 321
Gi3/2 321
Gi3/3 321
Gi3/4 321
Total virtual ports:1284
Aggregation-1#

```

Connected to 172.28.214.20 SSH2 - aes128-cbc - hmac-md5 - none 89x16 141698

Table 5-2 **Total Virtual Ports used Per Line Card**

1. CSCed33864 is resolved in Release 12.2(17d)SXB and later releases
2. 10 Mbps, 10/100 Mbps, and 100 Mbps switching modules support a maximum of 1200 logical interfaces per module

When designing a large data center using extended Layer 2 VLAN topologies, it is necessary to calculate the spanning tree logical and virtual ports in advance to ensure that spanning tree operates with optimal convergence and stability characteristics.

Figure 5-3 Calculating STP Logical Ports



Figure 5-3 shows a pair of aggregation switches (single aggregation module) connecting to 45 access layer switches in a looped access topology using four Gig-EtherChannel 802.1Q trunks. This might seem like a large number of access switches to have on a single aggregation pair, but with 1RU switch implementations, this amount or more is not uncommon. The following formula is used to determine the total active logical interfaces on the system:

trunks on the switch * active VLANs on trunks + number of non-trunking interfaces on the switch

Using Figure 5-3 for example, the calculation for aggregation 1 and 2 is as follows:

46 trunks * 120 VLANs + 2 = 5,522 active logical interfaces

This value is below the maximum values of both 802.1s MST and 802.1w Rapid-PVST+.

**Note**

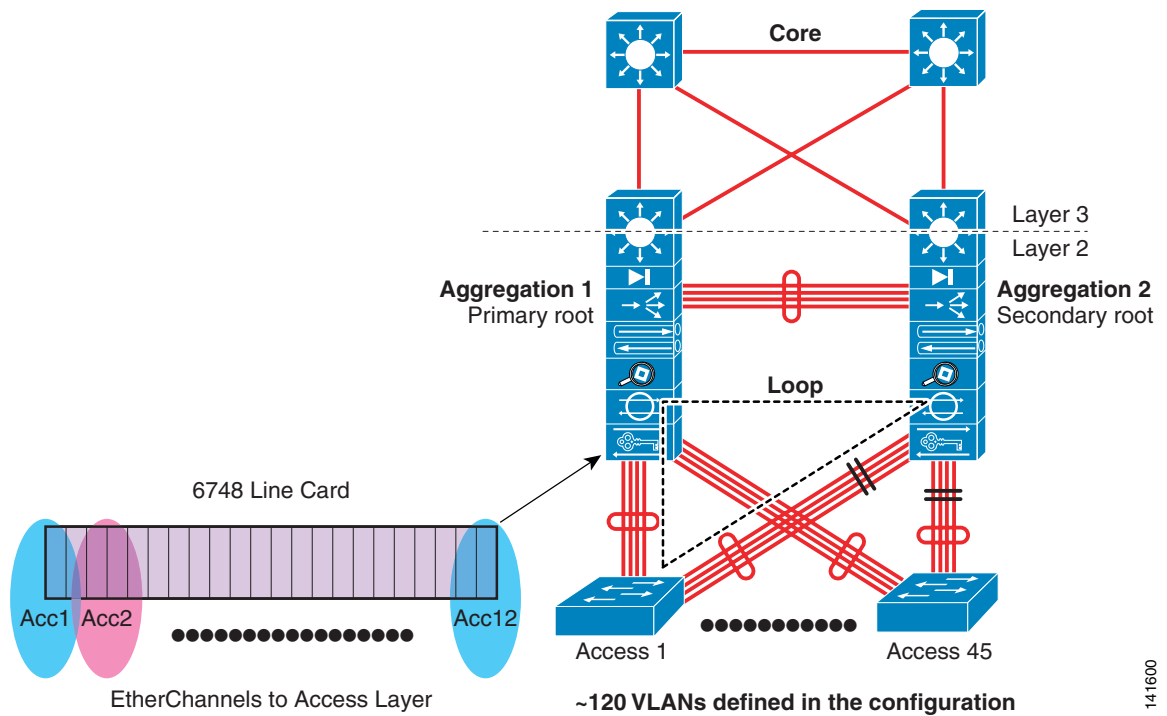
An STP instance for all 120 VLANs defined in the system configuration is present on each trunk unless manual VLAN pruning is performed. For example, on each trunk configuration the **switchport trunk allowed vlan X,Y** command must be performed to reduce the number of spanning tree logical interfaces being used on that port. The VTP Pruning feature does not remove STP logical instances from the port.

Calculating Virtual Ports per Line Card

Virtual ports are instances allocated to each trunk port on a line card. These ports are used to communicate the spanning tree-related state to the switch processor on the Sup720. A maximum number can be supported on each particular line card, as shown in Table 5-2. The following formula is used to determine the number of spanning tree virtual instances used per line card:

sum of all ports used as trunks or part of a port-channel in a trunk * active VLANs on trunks

Figure 5-4 shows a single line card on Aggregation 1 switch connecting to 12 access layer switches in a looped access topology using four Gig-EtherChannel 802.1Q trunks.

Figure 5-4 Calculating STP Virtual Ports per Line Card

A similar example could show two 6748 line cards connecting to 24 access switches using distributed EtherChannel. The formula below applies equally to both scenarios.

Using [Figure 5-4](#) for example, the calculation for the 6748 line card is as follows:

12 trunks with 4 ports each=48 ports * 120 vlans = 5,760 virtual ports

**Note**

Virtual ports are allocated for each VLAN for every port participating in a trunk.

This value is well over the maximum values of 802.1w Rapid-PVST+ and very close to the maximum for 802.1s MST. This scenario experiences various issues such as long convergence times and possibly degraded system level stability.

Steps to Resolve Logical Port Count Implications

The following steps can be taken to reduce the total number of logical ports or to resolve the issues related to a large amount of logical ports being used in a system:

- Implementing multiple aggregation modules—As covered in [Chapter 2, “Data Center Multi-Tier Model Design,”](#) using multiple aggregation modules permits the spanning tree domain to be distributed, thus reducing total port count implications.
- Performing manual pruning on switchport trunk configurations—Although this can be somewhat cumbersome, it dramatically reduces the total number of both active logical and virtual port instances used.

- Using 801.1s MST—MST supports a very large number of logical port instances and is used in some of the largest data centers in the world. The drawbacks of using MST are that it does not have as much flexibility as other STP protocols, such as Rapid-PVST+, and it might not be supported in certain service module configurations.



Note Refer to the Release Notes for information on MST interoperability with service modules.

- Distributing trunks and non-trunk ports across line cards—This can reduce the number of virtual ports used on a particular line card.
- Remove unused VLANs going to Content Switching Modules (CSMs)—The CSM automatically has all available VLANs defined in the system configuration extended to it via the internal 4GEC bus connection. This is essentially the same as any other trunk configured in the system. Although there is no officially documented method of removing unnecessary VLANs, it can be performed. [Figure 5-5](#) provides an example of the steps to remove VLANs from the CSM configuration. Note that the “command rejected” output is not valid.

Figure 5-5 Removing VLANs from CSM Configuration

```

172.28.214.5-port14_AGG1-sup6.r2w - WRQ Reflection for UNIX and Digital
File Edit Connection Setup Macro Window Help

Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#
Aggregation-1#show spannin
Aggregation-1#show spanning-tree int port-ch 259

Vlan          Role Sts Cost      Prio.Nbr Type
-----
VLAN0005      Desg FWD 5000      128.1674 Edge P2p
VLAN0044      Desg FWD 5000      128.1674 Edge P2p
VLAN0102      Desg FWD 5000      128.1674 Edge P2p
Aggregation-1#conf t
Enter configuration commands, one per line. End with CNTL/Z.
Aggregation-1(config)#int range port-ch 255 - 259
Aggregation-1(config-if-range)#sw tr all vl rem 5
Command rejected: Po255 not a switching port.
Aggregation-1(config-if-range)#exit
Aggregation-1(config)#exit
Aggregation-1#
Oct 20 13:40:09.156: %SYS-5-CONFIG_I: Configured from console by mnce onconsole
Aggregation-1#show spanning-tree int port-ch 259

Vlan          Role Sts Cost      Prio.Nbr Type
-----
VLAN0044      Desg FWD 5000      128.1674 Edge P2p
VLAN0102      Desg FWD 5000      128.1674 Edge P2p
Aggregation-1#
Aggregation-1#
  
```



Note There is no way to view which VLANs are attached to the CSM module via the configuration on the CLI. The only way to determine which VLANs are present is with the **show spanning-tree interface port-channel 259** command.



Data Center Access Layer Design



Note

The README file posted with this guide contains details relative to technologies, hardware, and software that were used in producing this document. There is also a revision history that details updates made to each chapter.

This chapter provides details of Cisco tested access layer solutions in the enterprise data center. It includes the following topics:

- [Overview of Access Layer Design Options](#)
- [Layer 2 Looped Access Layer Model](#)
- [Layer 2 Loop-Free Access Layer Model](#)
- [FlexLinks Access Model](#)

Overview of Access Layer Design Options

Access layer switches are primarily deployed in Layer 2 mode in the data center. A Layer 2 access topology provides the following unique capabilities required in the data center:

- **VLAN extension**—The Layer 2 access topology provides the flexibility to extend VLANs between switches that are connected to a common aggregation module. This makes provisioning of servers to a particular subnet/VLAN simple, and without the worry of physical placement of the server in a particular rack or row.
- **Layer 2 adjacency requirements**—NIC teaming, high availability clusters, and database clusters are application examples that typically require NIC cards to be in the same broadcast domain (VLAN). The list of applications used in a clustered environment is growing, and Layer 2 adjacency is a common requirement.
- **Custom applications**—Many developers write custom applications without considering the Layer 3 network environment, either because of lack of skills or available tools. This can create challenges in a Layer 3 IP access topology. These servers usually depend on Layer 2 adjacency with other servers and could require rewriting code when changing IP addresses.
- **Service modules**—A Layer 2 access permits services provided by service modules or appliances to be shared across the entire access layer. Examples of this are when using the FWSM, CSM, and SSLSM. The active-standby modes of operation used by service modules require Layer 2 adjacency with the servers that use them.






- Administrative reasons—Large enterprise customers commonly consist of multiple business units and departments, often with their own individual set of IT personnel, which might be the result of acquisitions or scaling of a business. IP address space is often divided and used by these business units with specific boundaries defined, or it might be completely overlapping. As data center consolidations occur, these business units/departments begin to share common floor and rack space. The ability to group these departments with Layer 2 VLANs across multiple access switches could be a critical requirement in these environments.

The table in [Figure 6-1](#) outlines the available access layer design models and provides a comparison of various factors to consider with each. Each access layer design model is covered in more detail in the remainder of this chapter.

**Note**

It might be more valuable to institute a point system in place of the plus-minus rating to determine which access layer model would be more appropriate for a particular design.

Figure 6-1 Comparison Chart of Access Layer Designs

		Uplinks on Agg Switch in Blocking or Standby State	VLAN Extension Supported Across Access	Service Module Black-Holing on Uplink Failure (5)	Single Attached Server Black- Holing on Uplink Failure	Access Switch Density per Agg Module	Must Consider Inter-Switch Link Scaling
	Looped Triangle	-	+	+	+	-	(3) +
	Looped Square	+	+	+	+	+	-
	Loop-free U	+	-	(4) -	+	+	+
	Loop-free Inverted U	+	+	+	(1, 2) +/-	+	-
	FlexLinks	-	+	+	+	-	+

1. Use of Distributed EtherChannel Greatly Reduces Chances of Black Holing Condition

2. NIC Teaming Can Eliminate Black Holing Condition

3. When Service Modules Are Used and Active Service Modules Are Aligned to Agg1

4. ACE Module Permits L2 Loopfree Access with per Context Switchover on Uplink failure

5. Applies to when using CSM or FWSM in active/standby arrangement

153046

The table in [Figure 6-1](#) contains the following column headings:

- Uplinks in blocking or standby state—Some access layer designs can use both uplinks (active-active), while others have one link active and the other blocked on a per-VLAN basis by spanning tree, or completely unused in a backup mode only. A plus is given to those models that have both uplinks active.
- VLAN extension across the access layer—A plus is given to those access design models that permit a VLAN to be extended to all access switches that are connected to a common aggregation module.

- Service module black holing—An uplink failure on the access layer switch could break connectivity between the servers and the service modules being used.
- Single attached server black holing—If an access switch has a single uplink, it could be a large failure exposure point. Uplinks that use Distributed EtherChannel can reduce the chances of black holing. Server load balancing to a VIP that includes servers physically connected across multiple access switches is another technique that can be used, as well as server NIC teaming.
- Access switch density per aggregation module—When 10GE uplinks are used, port density at the aggregation layer can be a challenge. Some access layer designs permit a larger number of access layer switches per aggregation module than others.
- Inter-switch link bandwidth scaling—Some access layer designs send all traffic towards the primary root aggregation switch, while other designs send traffic towards both aggregation switches. When sending to both aggregation switches, 50 percent of the traffic typically passes over the inter-switch link to reach the active HSRP default gateway and active service module pair. The amount of bandwidth used for the inter-switch links becomes very important in these designs and can create scaling challenges.

Service Module Influence on Design

This section contains recommendations for service module implementations for each of the access layer design models described. Because service modules can be implemented in many different ways or none at all, the focus is on a single service module design that is commonly implemented using the FWSM and CSM modules (see [Figure 6-2](#)).



Note

The Application Control Engine (ACE) is a new module that introduces several enhancements with respect to load balancing and security services. A key difference between the CSM, FWSM release 2.x, and ACE is the ability to support active-active contexts across the aggregation module with per context failover. The ACE module is not released at the time of this writing, so it is not covered.

The diagram illustrates a multi-context switch configuration for Policy Based Routing. It shows two aggregation blocks, Aggregation 1 and Aggregation 2, connected via L3 links and a 802.1Q Trunk. Each aggregation block contains a CSM, an MSFC (labeled 44), and an FWSM. Core 1 and Core 2 are connected to the MSFCs. The FWSMs are configured with three contexts: Context 1 (VLAN 23), Context 2 (VLAN 22), and Context 3 (VLAN 21). The diagram shows the flow of traffic from the CSMs through the MSFCs and FWSMs to the end hosts.

Aggregation 1

Core 1

Core 2

Aggregation 2

CSM

MSFC

CSM

L3 Links

802.1Q Trunk

Policy Based Routing

FWSM

FWSM

STP Primary Root
HSRP Active
Service Module Active

STP Secondary Root
HSRP Standby
Service Module Standby

Context 1

Context 2

Context 3

VLAN 23

VLAN 22

VLAN 21

153047

The FWSM can be virtualized when operating in transparent mode. This allows individual contexts of firewall instances to be created, configured, and managed independently of each other. This allows a single FWSM module to be used across different lines of business or operations as if multiple physical firewalls existed.

To achieve redundancy, service modules are deployed in pairs. One module in the pair acts as the primary/active service module while the other module acts as the secondary/standby. Although service module pairs can be deployed in the same aggregation chassis, they are typically placed in separate chassis to achieve the highest level of redundancy. Service modules are required to be Layer 2 adjacent on their configured VLAN interfaces to permit session state and monitoring to occur. For example, in [Figure 6-2](#), vlan 11 on the FWSM in aggregation 1 must be extended to vlan 11 on the FWSM in aggregation 2 via the 802.1Q trunk inter-switch link. This is also true for vlans 12, 13, 21, 22, and 23. This also applies to the server vlan 44 used by the CSM module in [Figure 6-2](#).

Because only one service module in one aggregation switch can be active at any one time, Cisco recommends aligning traffic flow towards the primary service module(s). The active default gateway and spanning tree root bridge are two components that influence path selection in a Layer 2 network. If primary service modules are located in the aggregation 1 switch, it is desirable to define the HSRP primary default gateway and spanning tree root bridge to also be on the aggregation 1 switch. This prevents session flow from hopping back and forth between aggregation switches, optimizing inter-switch link usage and providing a more deterministic environment.

**Note**

It is possible to double up on service modules and create a design such that active service modules are in each aggregation switch. This permits load balancing of access layer VLANs across uplinks to each aggregation switch without the need for flows to cross the inter-switch link between them. The disadvantage of this type of design is that there are twice the number of devices, with a corresponding increase in management and complexity.

When service modules/appliances are not used, access layer VLANs can be distributed across uplinks without concern for traffic flow issues. This can be achieved by alternating the HSRP active default gateway and spanning tree root configurations for each VLAN between the aggregation 1 and aggregation 2 switch, or by using Gateway Load Balancing Protocol (GLBP) in place of HSRP.

Because most data center implementations use service modules or appliances, the remainder of this chapter focuses on access layer topologies using service modules.

General Recommendations

The remainder of this chapter covers the details of the various access layer design models. Although each meets their own specific requirements, the following general recommendations apply to all:

- Spanning tree pathcost—Cisco recommends optimizing the spanning tree design by implementing the “spanning-tree pathcost method long” global feature. The pathcost method long option causes spanning tree to use a 32 bit-based value in determining port path costs compared to the default 16 bit, which improves the root path selection when various EtherChannel configurations exist.
- EtherChannel protocol—Cisco also recommends using Link Aggregation Control Protocol (LACP) as the link aggregation protocol for EtherChannel configurations. LACP considers the total available bandwidth of an EtherChannel path to STP in determining the path cost. This is advantageous in situations where only a portion of an EtherChannel link fails and the blocked alternate link can provide a higher bandwidth path.
- Failover tracking with service modules and HSRP—HSRP tracking of an interface can be used to control switchover of the primary default gateway between aggregation switches. Service modules can also track interface state and be configured to failover based on various up/down criteria. Unfortunately, the service modules and HSRP do not work together and have different mechanisms to determine failover, which can create situations where active and standby components are misaligned across the aggregation layer.

There are specific situations where tracking can be of benefit, but for the most part Cisco recommends not using the various failover tracking mechanisms and relying instead on using the inter-switch aggregation links to reach active default gateway and service module(s) during failure conditions. For this reason, it is important to consider failure scenarios when determining the proper inter-switch link bandwidth to be used.

- Service module timers—The convergence characteristics of various failure scenarios are influenced by the service module(s) failover timer configurations. Test lab results show that average service module failover times with these values are under ~6 seconds. The recommended service module failover timer configurations are as follows:
 - CSM


```
module ContentSwitchingModule 3
  ft group 1 vlan 102
  priority 20
  heartbeat-time 1
  failover 3
  preempt
```
 - FWSM


```
Unit Poll frequency 500 milliseconds, holdtime 3 seconds
Interface Poll frequency 3 seconds
```
- Using Distributed EtherChannel (DEC)—Cisco generally recommends that the inter-switch link between aggregation switches be implemented with a DEC connection to provide the highest level of resiliency. There are known caveats in certain Cisco IOS releases related to using DEC when service modules are present. For more details, refer to the Release Notes for this guide.

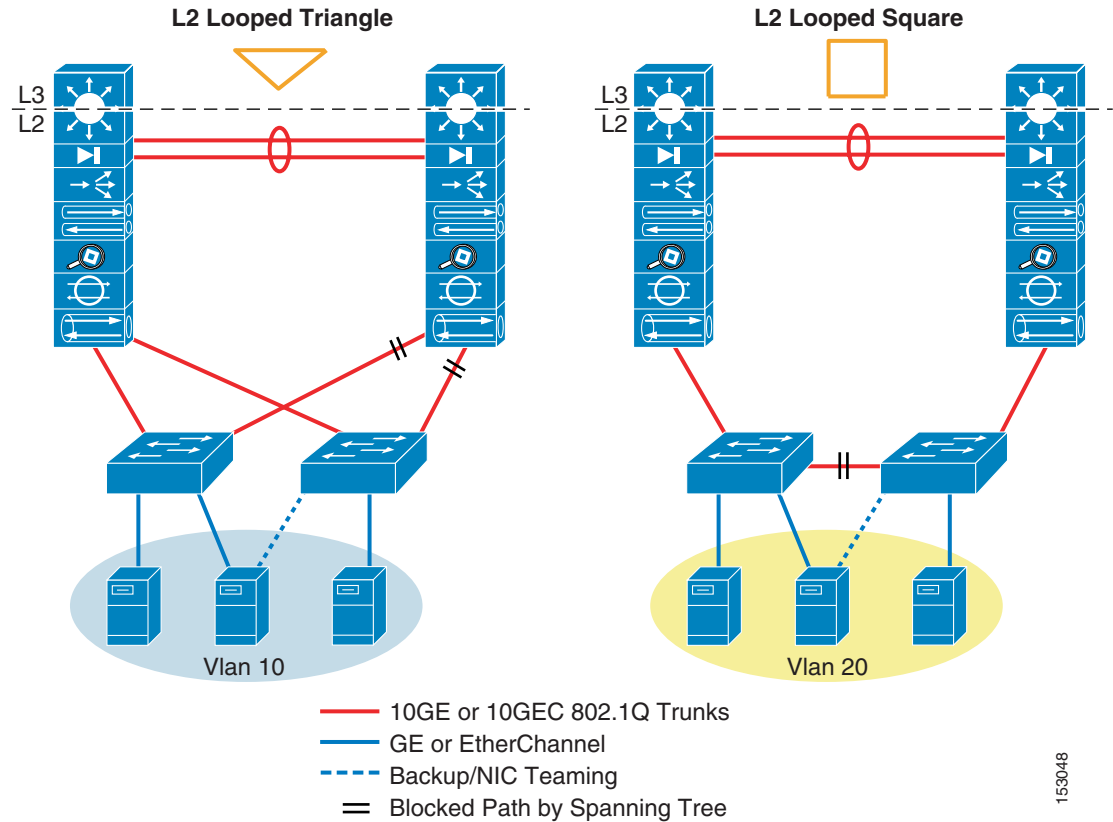
Layer 2 Looped Access Layer Model

This section covers Layer 2 looped access topologies, and includes of the following topics:

- [Layer 2 Looped Access Topologies](#)
- [Triangle Looped Topology](#)
- [Square Looped Topology](#)

Layer 2 Looped Access Topologies

In a Layer 2 looped access topology, a pair of access layer switches are connected to the aggregation layer using 802.1Q trunks. Looped access topologies consist of a triangle and square design, as shown in [Figure 6-3](#).

Figure 6-3 Triangle and Square Looped Access Topologies

In [Figure 6-3](#), a VLAN is configured on each access switch on the corresponding 802.1Q uplink, and is also extended between aggregation switches, forming a looped topology for that VLAN. The left side of the diagram shows an access layer with a triangle looped topology, and the right side shows a square looped topology. In the triangle looped topology, the access switch is dual homed to each aggregation switch. In the square looped topology, a pair of access switches are interconnected together, with each connected to a single aggregation switch.

Because a loop is present, all links cannot be in a forwarding state at all times. Because broadcasts/multicast packets and unknown unicast MAC address packets must be flooded, they would travel in an endless loop, completely saturating the VLAN and adversely affecting network performance. A spanning tree protocol such as Rapid PVST+ or MST is required to automatically block a particular link and break this loop condition.

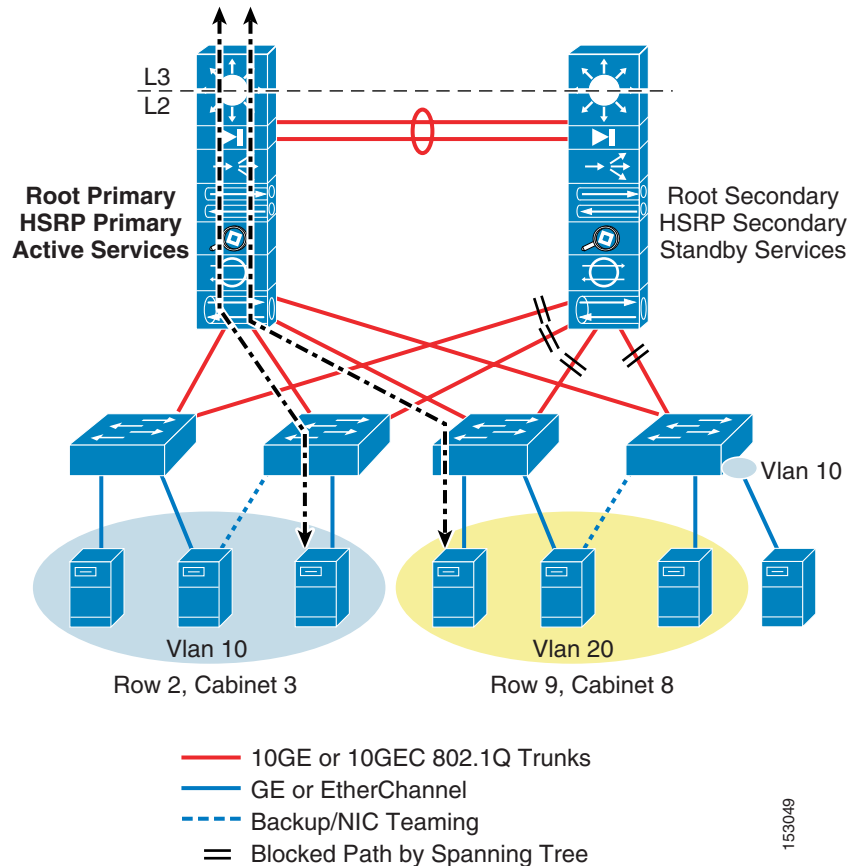
The dashed black lines on the aggregation layer switches represent the demarcation between Layer 2 and Layer 3 for the VLANs that are extended to the access layer switches. All packets processed in the VLAN beneath this line are in the same Layer 2 broadcast domain and are Layer 3 routed above the line. As denoted by the double solid lines, spanning tree automatically blocks one path to break the loop condition.

In both looped topologies, the service module fault-tolerant VLANs are extended between aggregation switches over the 802.1Q inter-switch link. This permits active-standby hellos and session state communications to take place to support redundancy.

Triangle Looped Topology

The triangle looped topology is currently the most widely implemented in the enterprise data center. This topology provides a deterministic design that makes it easy to troubleshoot while providing a high level of flexibility (see [Figure 6-4](#)).

Figure 6-4 Triangle Looped Access Topology



Spanning Tree, HSRP, and Service Module Design

In a triangle looped access layer design, it is desirable to align the spanning tree root, HSRP default gateway, and active service modules on the same aggregation switch, as shown in [Figure 6-4](#). Aligning the access layer switch uplink that is in the forwarding state directly to the same switch that is the primary default gateway and active service module/appliance optimizes the traffic flows. Otherwise, traffic flows can hop back and forth between aggregation switches, creating undesirable conditions and difficulty in troubleshooting.

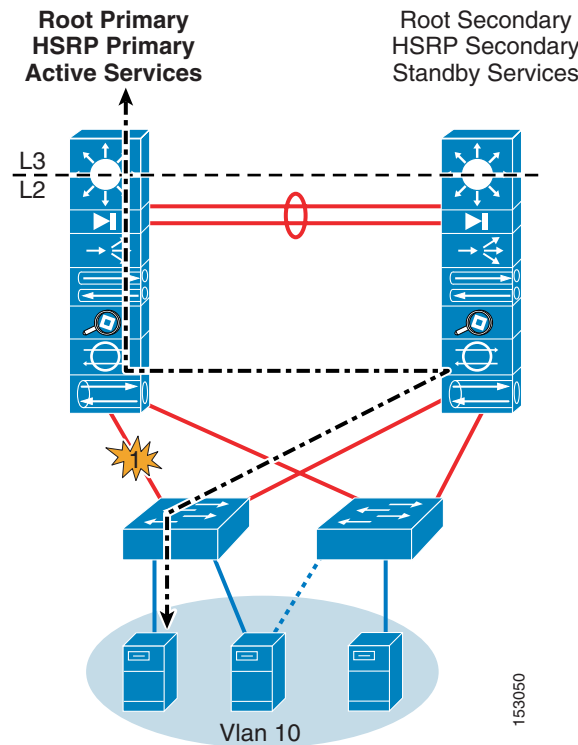
Failure Scenarios

The level of resiliency that is incorporated into the access layer design can vary based on the model used. Other features such as route health injection and route tuning can influence this. This section describes the four main failure scenarios that can occur in a looped access design. Understanding the amount of exposure a customer faces in these scenarios helps in selecting the best access layer design.

Failure 1—Access Layer Uplink Failure

In this failure scenario, spanning tree unblocks the uplink to aggregation 2 because no loop exists (see [Figure 6-5](#)).

Figure 6-5 Triangle Looped Failure Scenario 1—Uplink Down

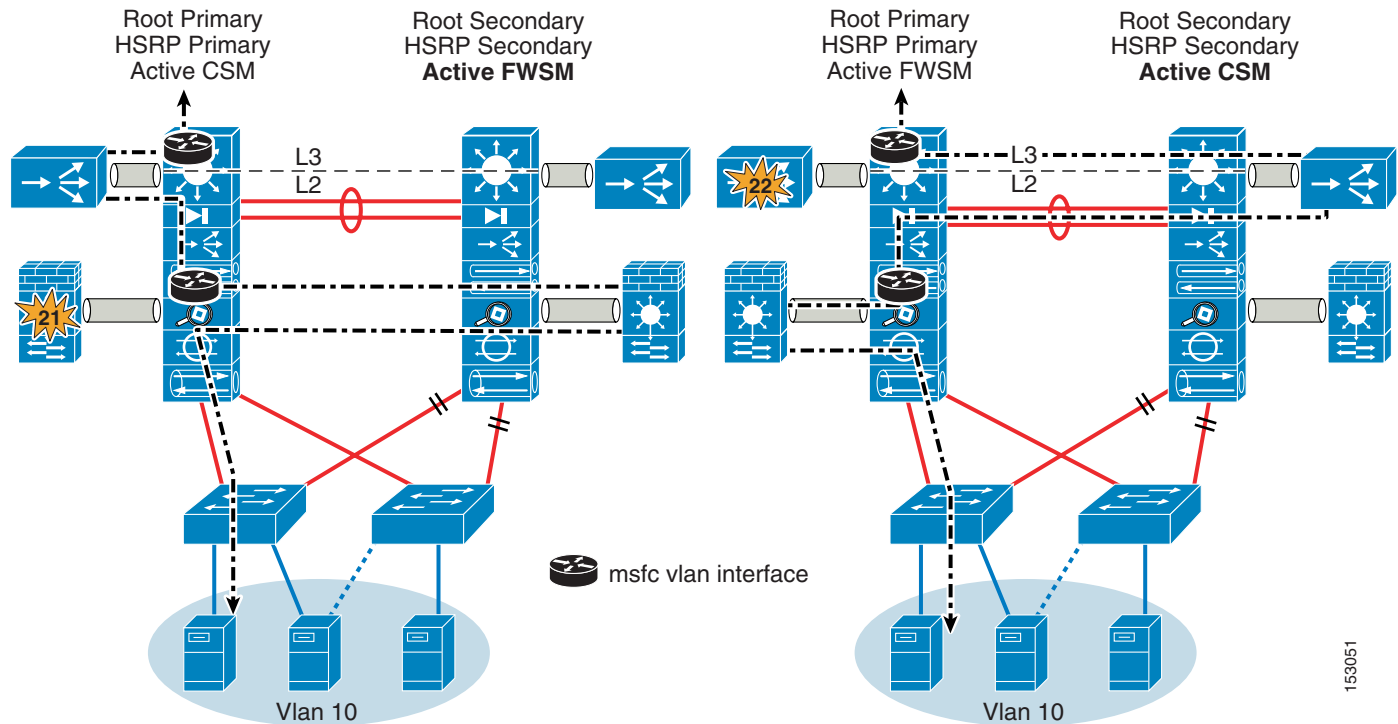


Default gateway and active service modules remain on aggregation 1 unless tracking mechanisms are configured and triggered. Traffic flow goes through aggregation 2 and uses the inter-switch link to aggregation 1 to reach the active HSRP default gateway and active service module.

The convergence characteristics of this failure scenario depend on spanning tree. Test lab results show that with Rapid-PVST+ implementations, this value should be under ~1.5 seconds, but can vary based on the number of spanning tree logical and virtual ports per line card values used.

Failure 2—Service Module Failure (Using CSM One-arm and FWSM Transparent Mode)

In this failure scenario, there is no spanning tree convergence, and the primary default gateway remains active on the aggregation 1 switch (see [Figure 6-6](#)).

Figure 6-6 Triangle Looped Failure Scenario 2—Service Modules

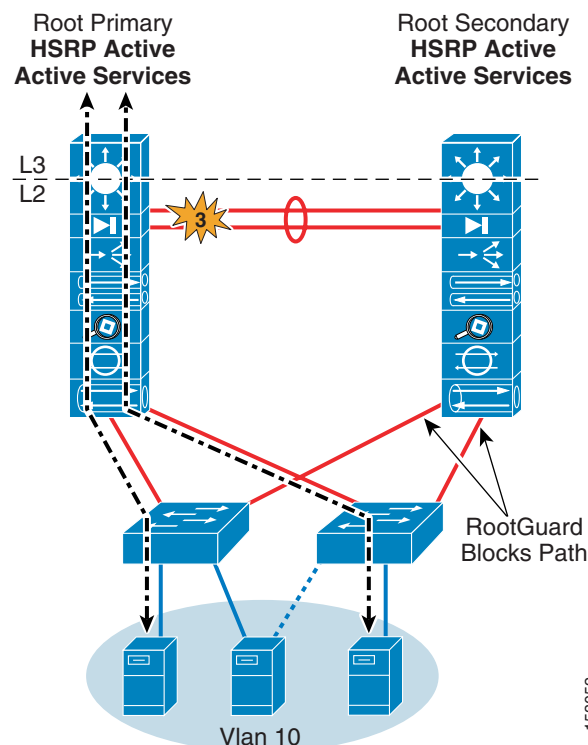
The backup service module moves to the active state on aggregation 2 because it no longer receives hello packets from the active service module, and times out.

Figure 6-6 shows the two following failure instances:

- 2.1 (FWSM failure)—Traffic flow goes through aggregation 1 and across the inter-switch link to aggregation 2, through the now active FWSM module context, and back across the inter-switch link to the active HSRP default gateway on the aggregation 1 MSFC. Because the CSM is still active in aggregation 1, return traffic flow is directed to the CSM based on the PBR configuration on the MSFC VLAN interface, and on to the client via the core.
- 2.2 (CSM failure)—Traffic flow goes through aggregation 1, through the active FWSM module context in aggregation 1, and to the MSFC VLAN interface. The MSFC VLAN interface PBR configuration forces the return CSM traffic to travel across the inter-switch link to aggregation 2 and through the now active CSM module. Because the active default gateway of the CSM server VLAN is still active on aggregation 1, the traffic must flow back across the inter-switch link to the MSFC on aggregation 1 and then on to the client via the core.

Failure 3—Inter-Switch Link Failure

Figure 6-7 shows failure scenario 3.

Figure 6-7 Triangle Looped Failure Scenario 3—Inter-Switch Link Failure

This failure scenario has many side effects to consider. First, spanning tree unblocks the uplink to aggregation 2 because no loop exists. RootGuard on the aggregation switch then automatically disables the link to access 2 because it sees root BPDUs on the now-unblocked path to Aggregation 1 via the access layer switch.

With the inter-switch link down and RootGuard disabling the path to Aggregation 1 via access2, HSRP multicast hello messages no longer have a path between Aggregation 1 and 2, so HSRP goes into an active state on both switches for all VLANs.

Because the service module failover VLANs are configured across the inter-switch link only, service modules in both aggregation switches determine that the other has failed and become active (this is referred to as a split-brain effect).

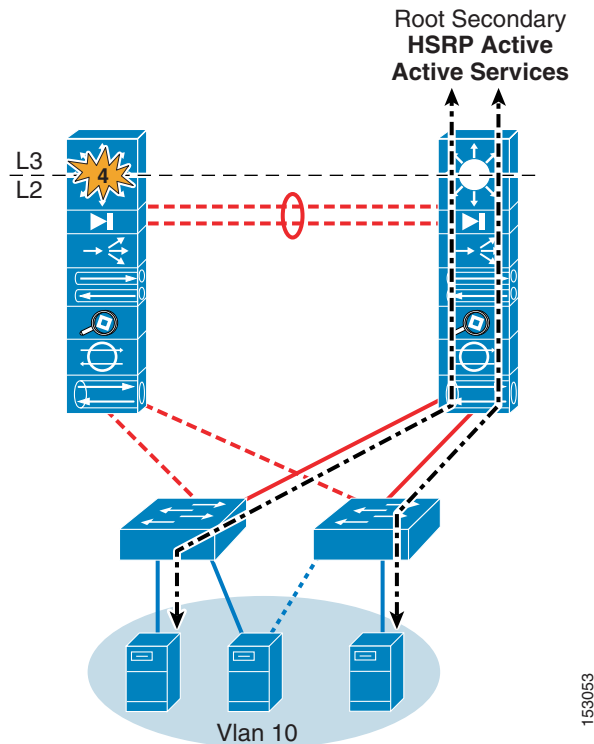
If inbound traffic from the core flows into the aggregation 2 switch during this failure scenario, it attempts to flow through the now-active service modules and stop, because RootGuard has the path to the servers blocked. If for some reason RootGuard is not configured, this still results in asymmetrical flows and breaks connectivity. It is for these reasons that Cisco recommends tuning the aggregation-core routing configuration such that the aggregation 1 switch is the primary route advertised to the core for the primary service module-related VLANs.

Route tuning plus RootGuard prevents asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions. More detail on route tuning can be found in [Establishing Path Preference with RHI](#), page 7-1.

Failure 4—Switch Power or Sup720 Failure (Non-redundant)

Figure 6-8 shows failure scenario 4.

Figure 6-8 Triangle Looped Failure Scenario 4—Single Sup720 or Power Failure

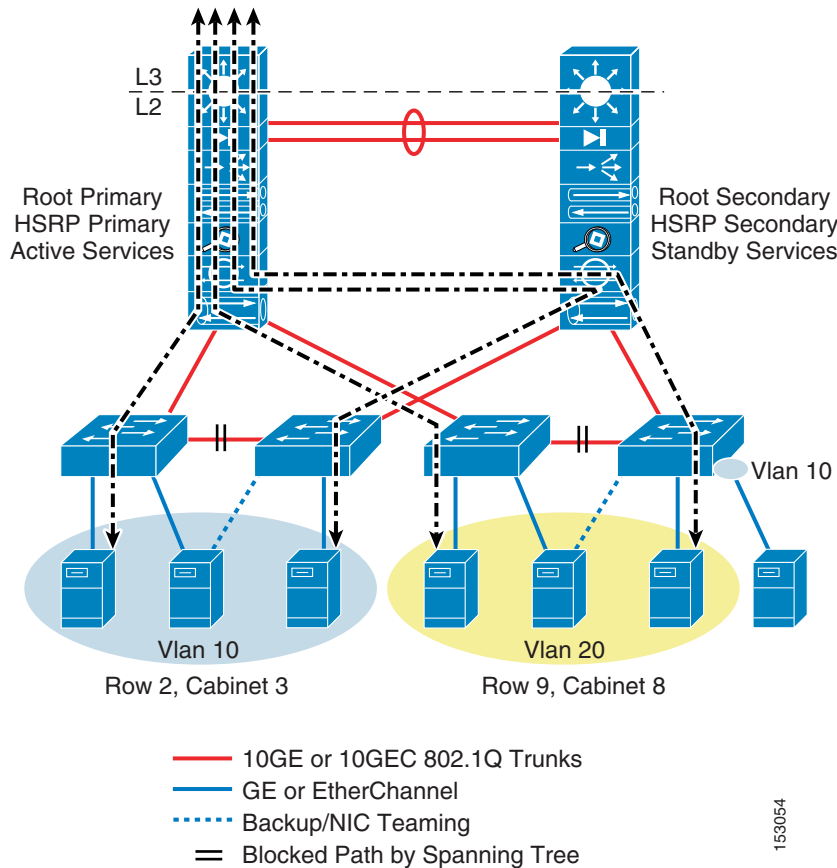


In this failure scenario, the spanning tree root, primary default gateway, and active service modules transition to the aggregation 2 switch.

The convergence characteristics of this failure scenario depend on spanning tree, HSRP, and service module failover times. Because the spanning tree and HSRP failover times are expected to be under that of service modules, the actual convergence time depends on service module timer configurations.

Square Looped Topology

The square-based looped topology is not as common today in the enterprise data center but has recently gained more interest. The square looped topology increases the access layer switch density when compared to a triangle loop topology while retaining the same loop topology characteristics. This becomes particularly important when 10GE uplinks are used. This topology is very similar to the triangle loop topology, with differences in where spanning tree blocking occurs (see Figure 6-9).

Figure 6-9 Square Looped Access Topology

Spanning tree blocks the link between the access layer switches, with the lowest cost path to root being via the uplinks to the aggregation switches, as shown in Figure 6-9. This allows both uplinks to be active to the aggregation layer switches while providing a backup path in the event of an uplink failure. The backup path can also be a lower bandwidth path because it is used only in a backup situation. This might also permit configurations such as 10GE uplinks with GEC backup.

The possible disadvantages of the square loop design relate to inter-switch link use, because 50 percent of access layer traffic might cross the inter-switch link to reach the default gateway/active service module. There can also be degradation in performance in the event of an uplink failure because, in this case, the oversubscription ratio doubles.

Figure 6-9 shows the spanning tree blocking point on the link between the access switch pair. This is ideal if active services are deployed in each aggregation switch because it permits the uplinks to be load balanced without traversing the aggregation layer inter-switch trunk. If active services are only on Agg1, it might be desirable to adjust the STP cost such that the uplink to Agg2 is blocking instead of the link between the access pair. This forces all traffic to the Agg1 switch without having to traverse the aggregation layer inter-switch trunk.

Spanning Tree, HSRP, and Service Module Design

Similar to a triangle design, it is desirable in a square looped access layer design to align the spanning tree root, HSRP default gateway, and active service modules on the same aggregation switch, as shown in Figure 6-9. By aligning the access layer switch uplink that is in the forwarding state directly to the

same switch that is the primary default gateway and active service module/appliance, traffic flows are optimized. Otherwise, traffic flows can hop back and forth between aggregation switches, creating undesirable conditions that are unpredictable and difficult to troubleshoot.

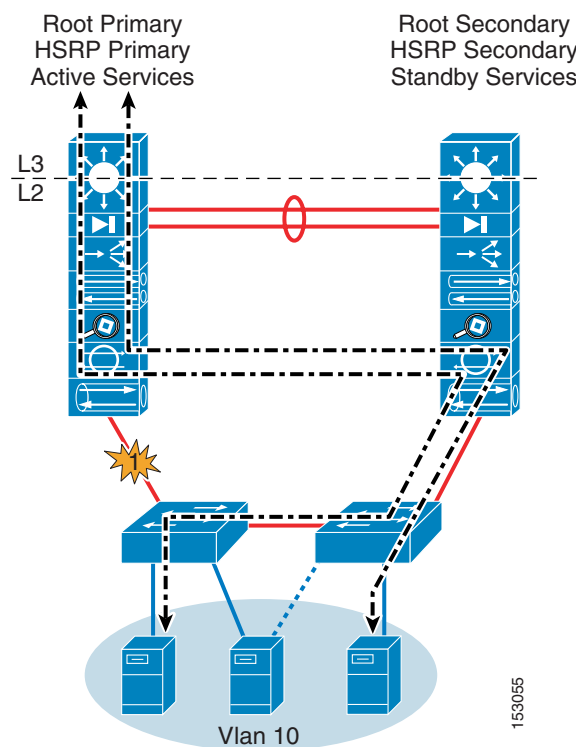
Failure Scenarios

This section examines the square loop design in various failure scenarios.

Failure 1—Access Layer Uplink Failure

Figure 6-10 shows failure scenario 1.

Figure 6-10 Square Looped Failure Scenario 1—Uplink Down

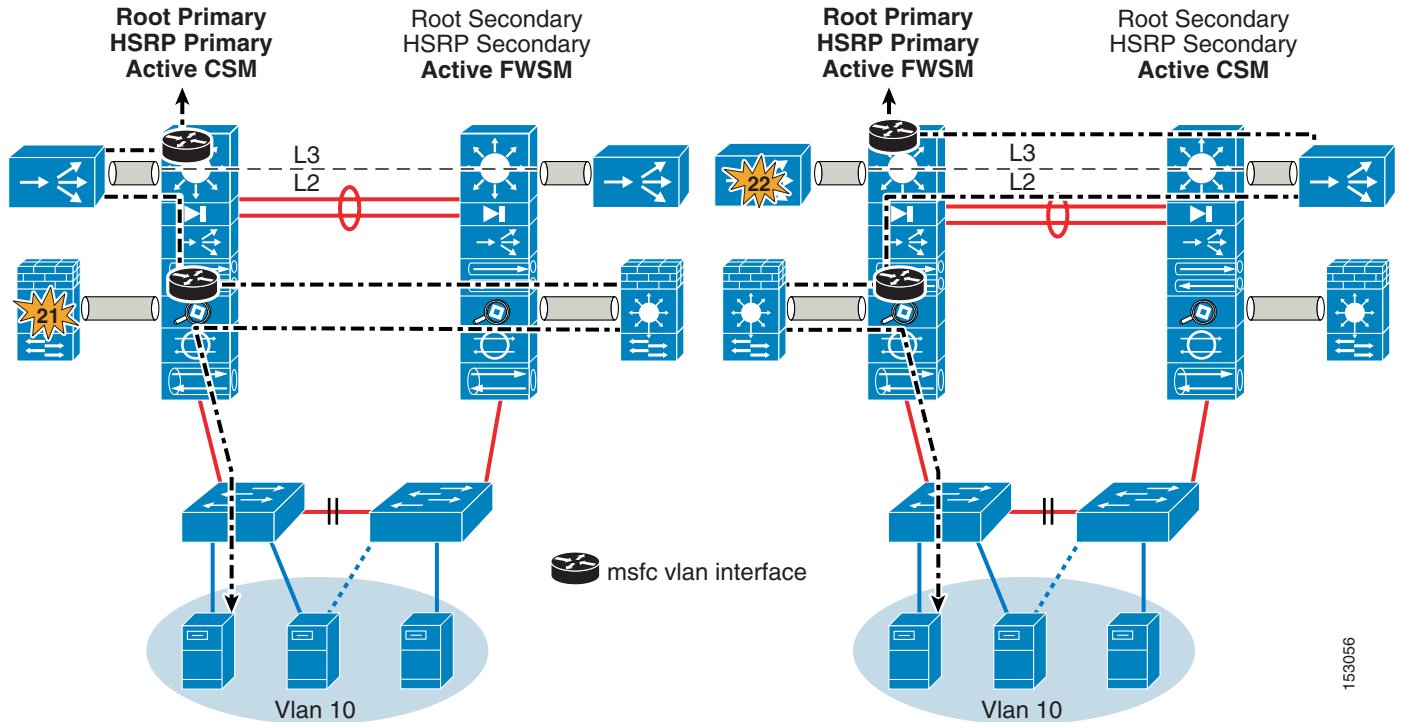


In this failure scenario, spanning tree unblocks the link between access switches because a loop no longer exists. The default gateway and active service modules remain on aggregation 1 unless tracking mechanisms are configured and triggered. Traffic flows go through aggregation 2 and use the inter-switch link to aggregation 1 to reach the active HSRP default gateway and active service module.

The convergence characteristics of this failure scenario depend on spanning tree. Test lab results show that with Rapid-PVST+ implementations, this value should be under ~1.5 seconds, but can vary based on the number of spanning tree logical and virtual ports per line card values present.

Failure 2—Service Module Failure (using CSM One-arm and FWSM Transparent Mode)

In the failure scenario shown in Figure 6-11, there is no spanning tree convergence, and the primary default gateway remains active on the aggregation 1 switch.

Figure 6-11 Square Looped Failure Scenario 2—Service Modules

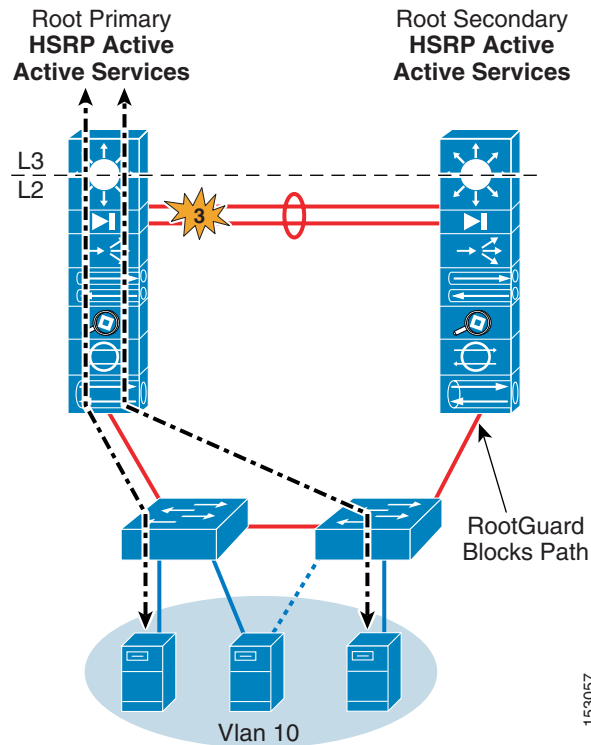
The backup service module moves to the active state on aggregation 2 because it no longer receives hello packets from the active service module, and times out.

The following failure scenarios are shown:

- 2.1 (FWSM failure)—Traffic flow goes through aggregation 1 and across the inter-switch link to aggregation 2, through the now-active FWSM module context, and back across the inter-switch link to the active HSRP default gateway on the aggregation 1 MSFC. Because the CSM is still active in aggregation 1, return traffic flow is directed to the CSM based on the PBR configuration on the MSFC VLAN interface, and then on to the client via the core.
- 2.2 (CSM failure)—Traffic flow goes through aggregation 1, through the active FWSM module context in aggregation 1, and to the MSFC VLAN interface. The MSFC VLAN interface PBR configuration forces return CSM traffic to travel across the inter-switch link to aggregation 2 and through the now-active CSM module. Because the active default gateway of the CSM server VLAN is still active on aggregation 1, the traffic must flow back across the inter-switch link to the MSFC on aggregation 1, and then on to the client via the core.

Failure 3—Inter-Switch Link Failure

Figure 6-12 shows failure scenario 3.

Figure 6-12 Square Looped Failure Scenario 3—Inter-Switch Link Failure

This failure scenario has many side effects to consider. First, spanning tree unblocks the access layer inter-switch link because a loop no longer exists. RootGuard on the aggregation switch then automatically disables the link to access 2 because it sees root BPDUs via the now-unblocked path to aggregation 1.

With the inter-switch link down and RootGuard disabling the path to aggregation 1 via access 2, HSRP multicast hello messages no longer have a path between aggregation 1 and 2, so HSRP goes into an active state on both switches for all VLANs.

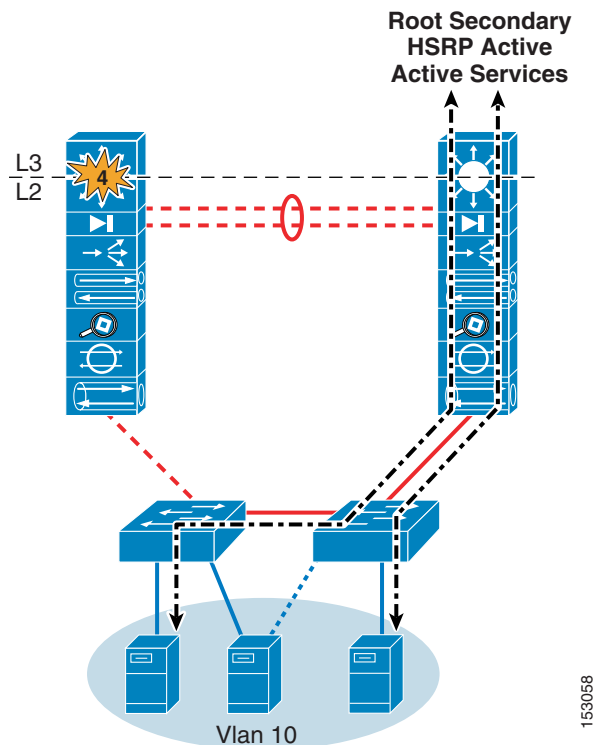
Because the service module failover VLANs are configured across the inter-switch link only, service modules in both aggregation switches determine that the other has failed. This results in service modules in aggregation 1 remaining in the active state, and service modules in aggregation 2 moving from standby to the active state as well. This is commonly referred to as a split-brain effect and is very undesirable.

If inbound traffic from the core flows into the aggregation 2 switch during this failure scenario, it attempts to flow through the now-active service modules and stops, because RootGuard has the path to the servers blocked. If for some reason RootGuard is not configured, this still results in asymmetrical flows and breaks connectivity. For these reasons, Cisco recommends tuning the aggregation-core routing configuration such that the aggregation 1 switch is the primary route advertised to the core for the primary service module-related VLANs.

Route tuning plus RootGuard prevents asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions.

Failure 4—Switch Power or Sup720 Failure (Non-redundant)

Figure 6-13 shows failure scenario 4.

Figure 6-13 Square Looped Failure Scenario 3—Switch Power or Sup720 Failure

In this failure scenario, the spanning tree root, primary default gateway, and active service modules transition to the aggregation 2 switch.

The convergence characteristics of this failure scenario depend on spanning tree, HSRP, and service module failover times. Because the spanning tree and HSRP failover times are expected to be under that of service modules, the actual convergence time depends on service module timer configurations.

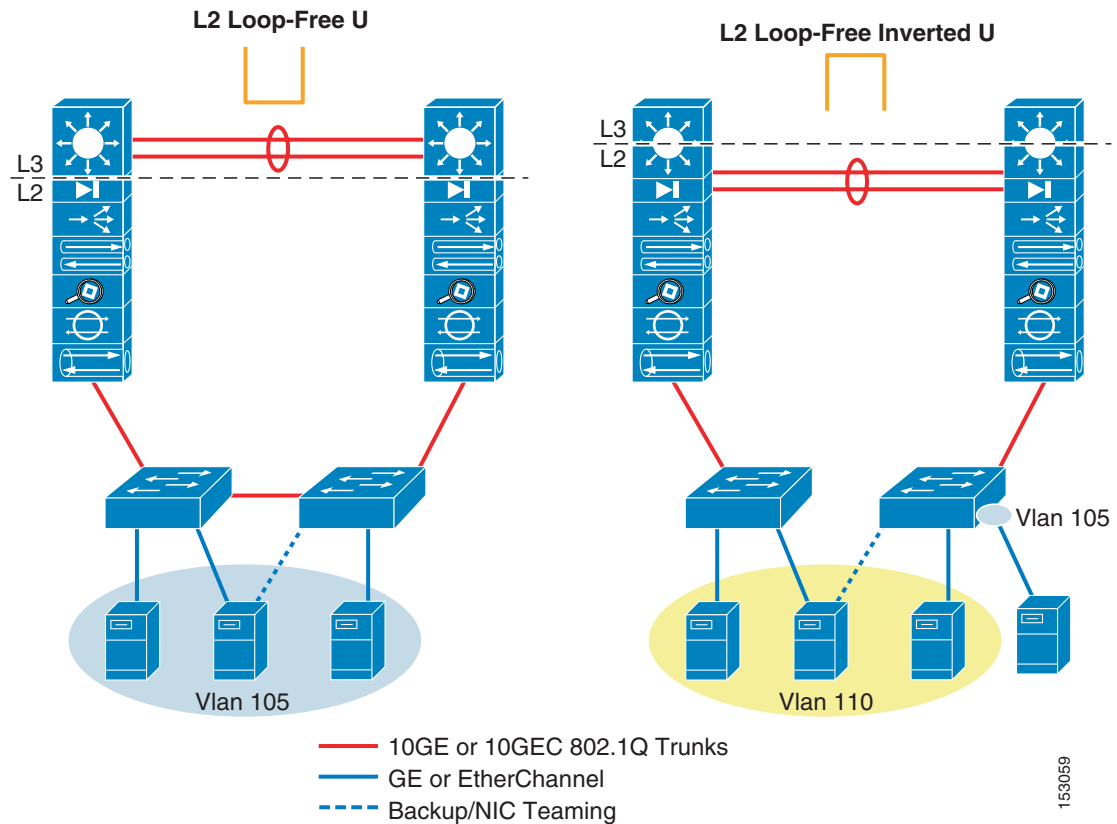
Layer 2 Loop-Free Access Layer Model

This section covers Layer 2 looped access topologies and includes the following topics:

- [Layer 2 Loop-Free Access Topologies](#)
- [Layer 2 Loop-Free U Topology](#)
- [Layer 2 Loop-Free Inverted U Topology](#)

Layer 2 Loop-Free Access Topologies

Figure 6-14 illustrates the access layer using the Layer 2 loop-free model, with loop-free U and loop-free inverted U topologies.

Figure 6-14 Access Layer Loop-Free Topologies

Note that the Layer 2/Layer 3 line of demarcation is different in each design. In a loop-free U, a VLAN is configured on each access switch, and on the 802.1Q inter-switch link between access switches and its corresponding 802.1Q uplink, but it is not extended between aggregation switches; thereby avoiding a looped topology.

In a loop-free inverted U design, a VLAN is configured on each access switch and its corresponding 802.1Q uplink, and is also extended between aggregation switches, but is not extended between access switches, avoiding a looped topology.

Although no loop is present in either loop-free design topology, it is still necessary to run STP as a loop prevention tool. In the event that a cabling or configuration error that creates a loop is encountered, STP prevents the loop from possibly bringing down the network.

**Note**

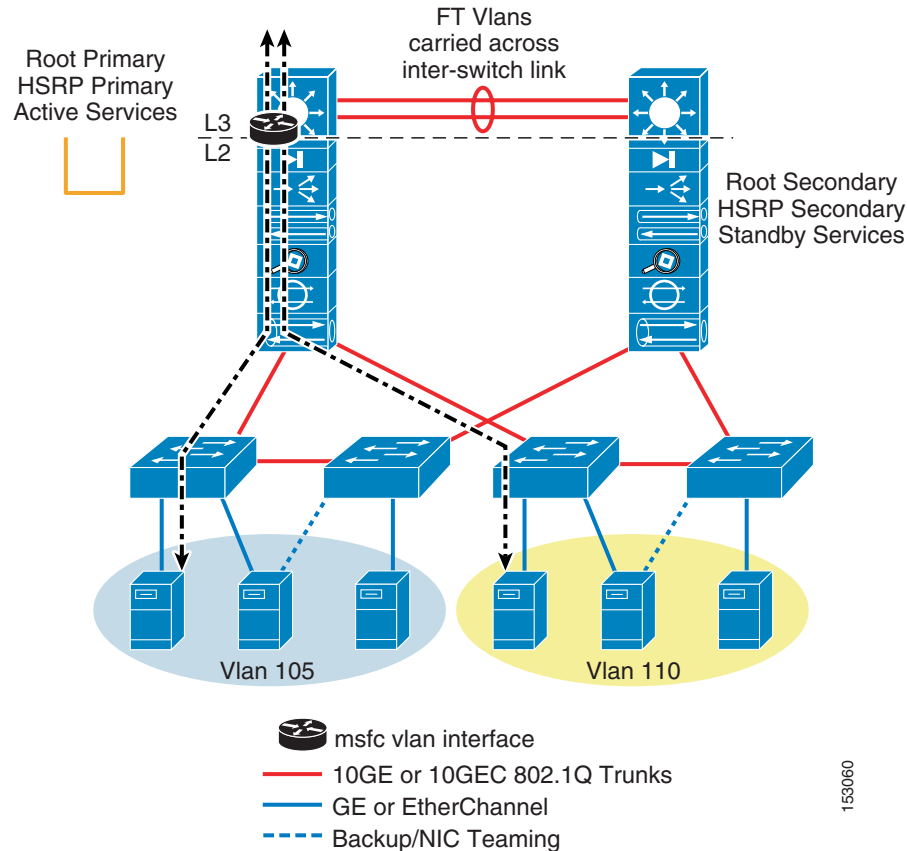
In the loop-free U design, you cannot use RootGuard on the aggregation to access layer links because the aggregation 2 switch would automatically disable these links because root BPDUs would be seen. Details on spanning tree protocol types and comparisons are covered in the version 1.1 of this guide.

In both loop-free topologies, the service module fault-tolerant VLANs are extended between aggregation switches over the 802.1Q inter-switch link. This permits active-standby hellos and session state communications to take place to support redundancy.

Layer 2 Loop-Free U Topology

The loop-free U topology design provides a Layer 2 access solution with active uplinks and redundancy via an inter-switch link between the access layer switches. The chance of a loop condition is reduced but spanning tree is still configured in the event of cabling or configuration errors occur (see [Figure 6-15](#)).

Figure 6-15 Loop-Free U Access Topology



With a loop-free U topology, there are no blocked paths by spanning tree because a loop does not exist. The VLANs are configured on the access layer uplink 802.1Q trunks and access layer inter-switch 802.1Q trunks but are not extended between the aggregation layer switches (note the dashed line designating the Layer 2 and Layer 3 boundaries). The service module fault tolerant VLANs are carried across the 802.1Q trunk for redundancy operations.

This topology allows both uplinks to be active for all VLANs to the aggregation layer switches while providing a backup path in the event of an uplink failure. This also permits a higher density of access switches to be supported on the aggregation module.

The main disadvantages of the loop-free U design is the inability to extend VLANs outside of an access pair, and failure conditions that can create black holes in the event of an uplink failure when service modules are used. Extending VLANs outside of a single access pair creates a loop through the aggregation layer, essentially creating a four-node looped topology with blocked links. The black holes condition is covered in the failure scenarios later in this section.

Spanning Tree, HSRP, and Service Module Design

Because a loop does not exist in the topology, it does not actually require a spanning tree protocol to be running. However, it is very wise to maintain spanning tree in case an error creates a loop condition. It is also still recommended to maintain spanning tree primary root and secondary root configurations just as in the triangle and square looped topology designs. This way, if a loop error condition does exist, the service module and default gateway still operate optimally.



Note

Cisco does not recommend using the loop-free U design in the presence of service modules because of black holing in the event of an uplink failure. More detail is covered in the failure scenarios part of this section. Service modules can be used with a loop-free inverted U topology when the design permits server black holing conditions and uses other mechanisms, such as when load balancers are combined with server distribution across the access layer.

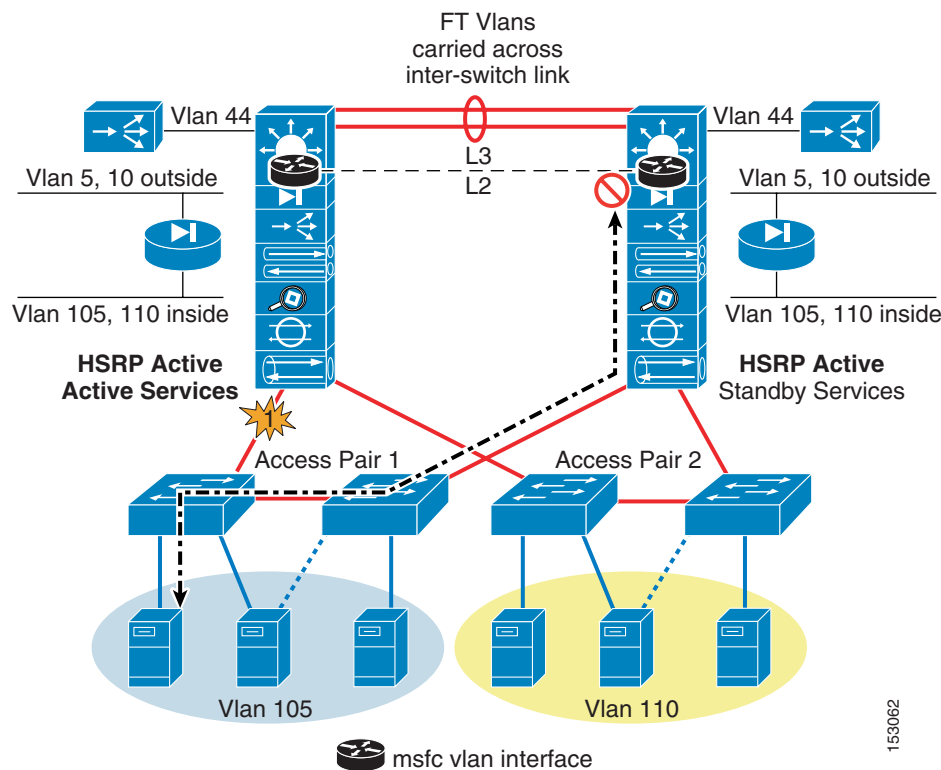
Failure Scenarios

This section describes the loop-free U design in various failure scenarios.

Failure 1—Access Layer Uplink Failure

Figure 6-16 shows failure scenario 1.

Figure 6-16 Loop-Free U Failure Scenario 1—Uplink Failure



In this failure scenario, HSRP multicast hellos are no longer exchanged between the aggregation switches, which creates an active-active HSRP state for the vlan 5 and 10 MSFC interfaces on both aggregation switches.

The servers on access pair 1 are not able to reach the active FWSM context on aggregation 1 because there is no Layer 2 path for vlan 105 across the aggregation layer inter-switch links. Although the FWSM can be configured to switchover the active-standby roles by using the interface monitoring features, this requires the entire module to switchover (all contexts) on a single uplink failure. This is not a desirable condition and is further complicated if there are multiple uplink failures, or when maintenance requires taking down an access layer switch/uplink.

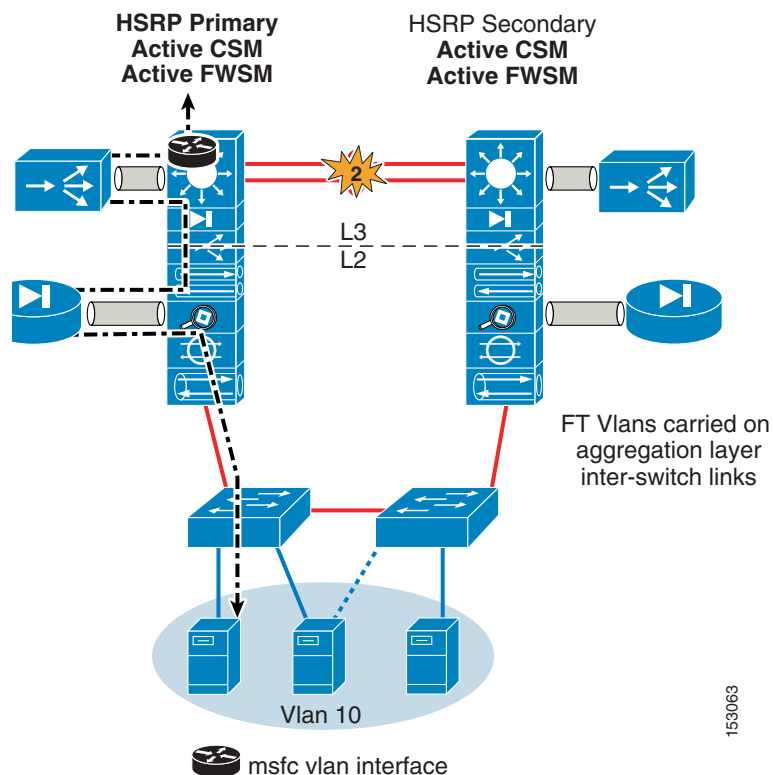
**Note**

Because of the lack of single context failover, improved tracking and mis-aligned components, Cisco does not recommend using service modules with the loop-free U topology.

Failure 2—Inter-Switch Link Failure

Figure 6-17 shows failure scenario 2.

Figure 6-17 Loop-Free U Failure Scenario 2—Inter-Switch Link Failure



This failure scenario has many side effects to consider. Because the service module failover VLANs are configured across the inter-switch link only, service modules in both aggregation switches determine that the other has failed. This results in service modules in aggregation 1 remaining in the active state, and service modules in aggregation 2 moving from standby to the active state as well. This is commonly referred to as a split-brain effect, and is very undesirable because the opportunity for asymmetrical connection failure exists.

The HSRP heartbeats travel along the access layer path, so HSRP remains in the same state with primary on aggregation 1 and standby on aggregation 2.

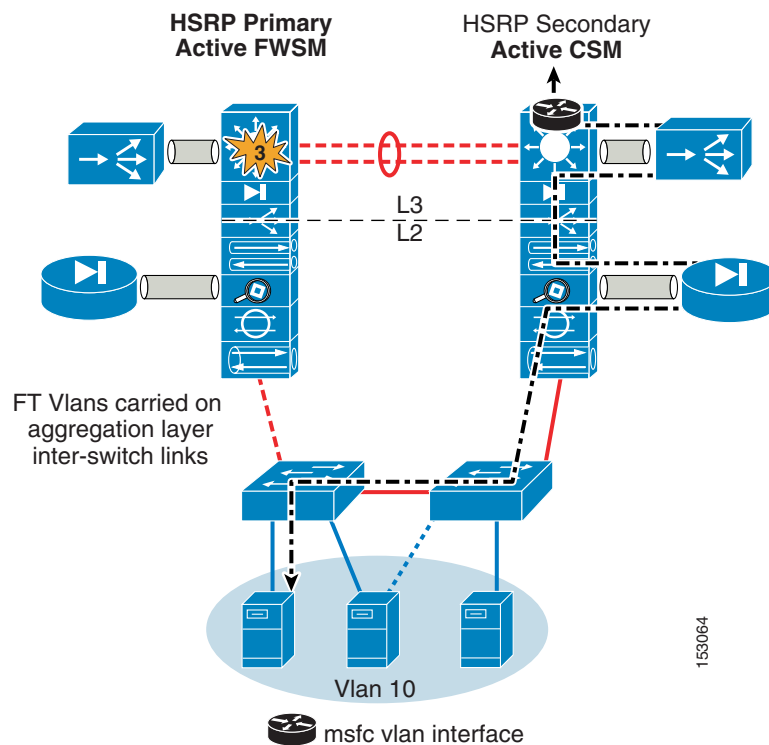
If inbound traffic from the core flows into the aggregation 2 switch during this failure scenario, it reaches the MSFC and then attempts to flow through the now-active service modules. By default, the core switches are performing CEF-based load balancing, thereby distributing sessions to both aggregation 1 and 2. Because state is maintained on the service modules, it is possible that asymmetrical connection failures can occur. For these reasons, Cisco recommends tuning the aggregation-core routing configuration such that the aggregation 1 switch is the primary route from the core for the primary service module-related VLANs.

Route tuning prevents asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions. More information on route tuning can be found in [Establishing Path Preference with RHI](#), page 7-1.

Failure 3—Switch Power or Sup720 Failure (Non-redundant)

Figure 6-18 shows failure scenario 3.

Figure 6-18 Loop-Free U Failure Scenario 3—Single Sup720 or Power Failure



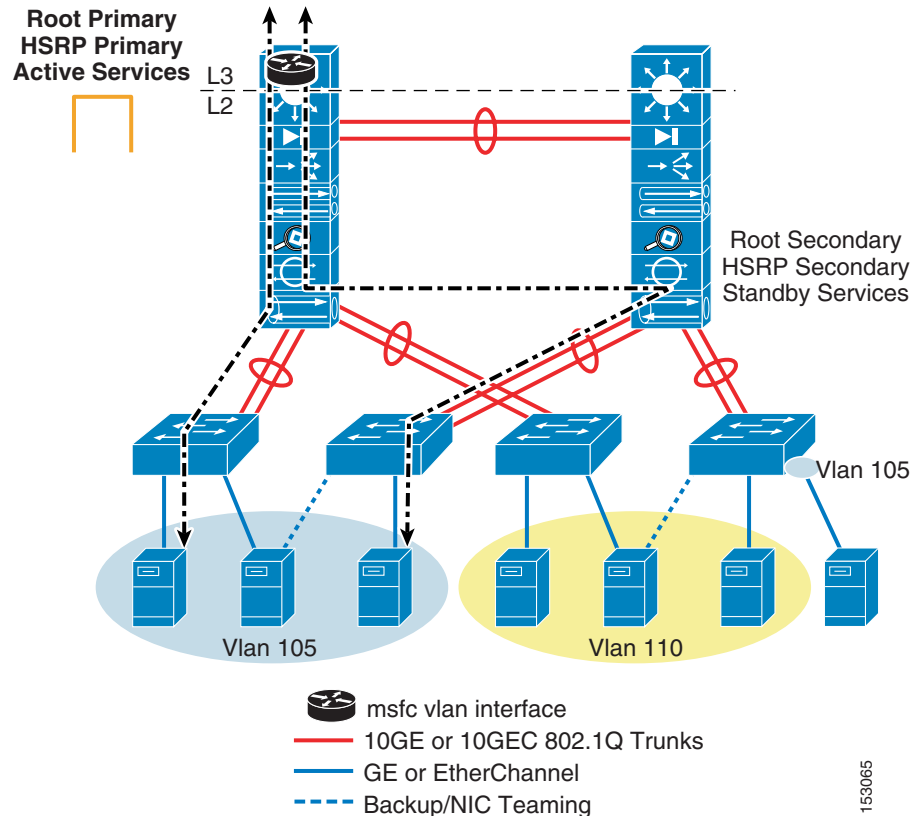
In this failure scenario, the spanning tree root, primary default gateway, and active service modules transition to the aggregation 2 switch.

The convergence characteristics of this failure scenario depend on spanning tree, HSRP, and service module failover times. Because the spanning tree and HSRP failover times are expected to be under that of service modules, the actual convergence time depends on service module timer configurations. Test lab results show this convergence time to be ~ 6 seconds.

Layer 2 Loop-Free Inverted U Topology

The loop-free inverted-U topology design provides a Layer 2 access solution with a single active access layer uplink to a single aggregation switch, as shown in Figure 6-19.

Figure 6-19 Loop-Free Inverted-U Access Topology



With a loop-free inverted-U topology, there are no blocked paths by spanning tree because a loop does not exist. The VLANs are configured on the access layer uplink 802.1Q trunks and are extended between the aggregation layer switches (note the dashed line designating the Layer 2 and Layer 3 boundaries). The service module fault tolerant VLANs are carried across the aggregation inter-switch 802.1Q trunk for redundancy operations. This topology allows both uplinks to be active for all VLANs to the aggregation layer switches and permits VLAN extension across the access layer. The loop-free inverted-U design does not provide a backup link at the access layer, but resiliency can be improved by the use of distributed EtherChannel (DEC), as shown in Figure 6-19.

The main disadvantage of the loop-free inverted-U design can be attributed to an aggregation switch failure or access switch uplink failure that black holes servers because there is no alternate path available. The following improvements to the design can offset the effects of these failures and improve overall resiliency:

- Aggregation nodes with redundant Sup720s using NSF/SSO
- Distributed EtherChannel uplinks
- NIC teaming
- Server load balancing with REALS spread across access switches

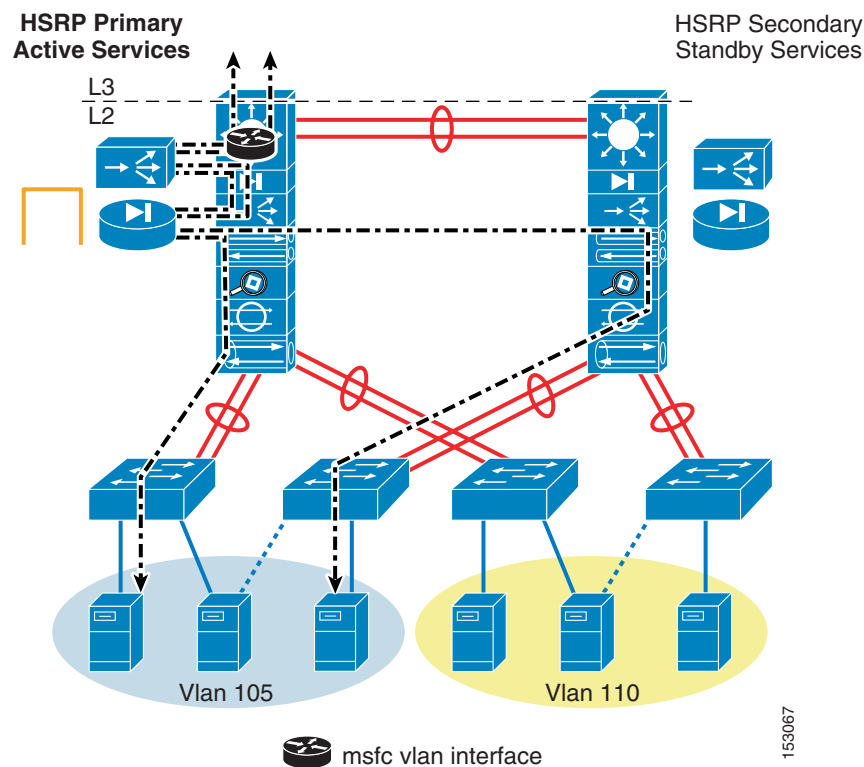
Spanning Tree, HSRP, and Service Module Design

Because a loop does not exist in the topology, it does not require a spanning tree protocol to be running. However, Cisco recommends maintaining spanning tree in case an error creates a loop condition. Cisco also still recommends maintaining spanning tree primary root and secondary root configurations just as in the triangle and square looped topology designs. This way if a loop error condition does exist, the service module and default gateway still operate optimally.

As in all other access layer designs that use service modules, Cisco recommends aligning the HSRP default gateway, STP root, and active service modules on the same aggregation switch. If the primary default gateway and active service modules are not aligned, it creates session flows that travel across the inter-switch links unnecessarily.

When HSRP, STP root, and primary service modules are aligned, the session flows are more optimal, easier to troubleshoot, and deterministic, as shown in [Figure 6-20](#). Note that in a loop-free inverted-U topology, 50 percent of the session flows use the aggregation layer inter-switch link to reach the active HSRP default gateway and active service modules.

Figure 6-20 Loop-Free Inverted-U with HSRP and Service Modules Aligned –Recommended

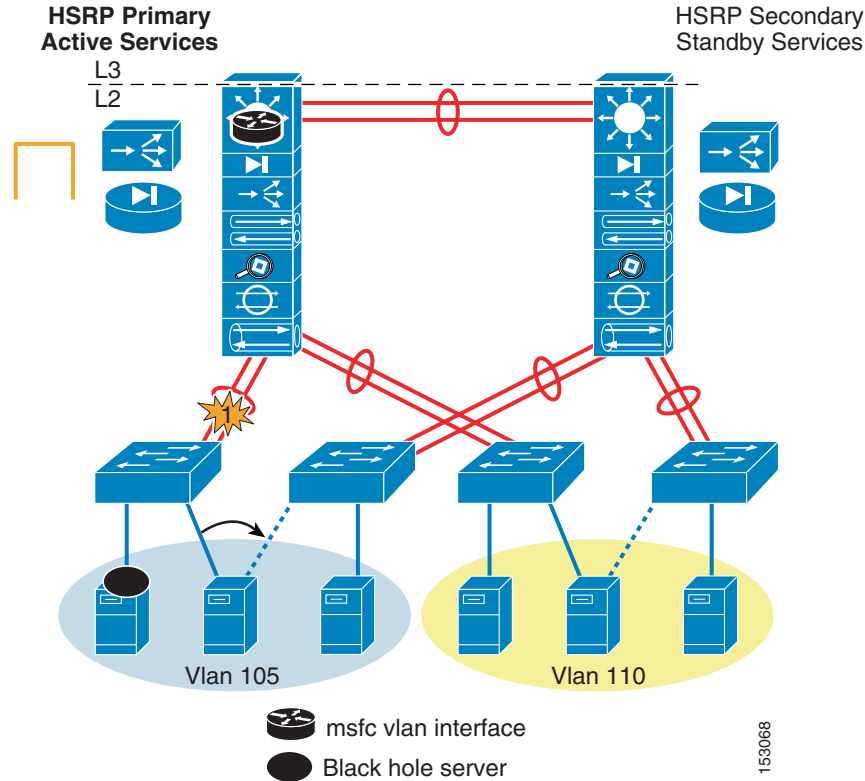


Failure Scenarios

This section describes the loop-free inverted-U design in various failure scenarios.

Failure 1—Access Layer Uplink Failure

[Figure 6-21](#) shows failure scenario 1.

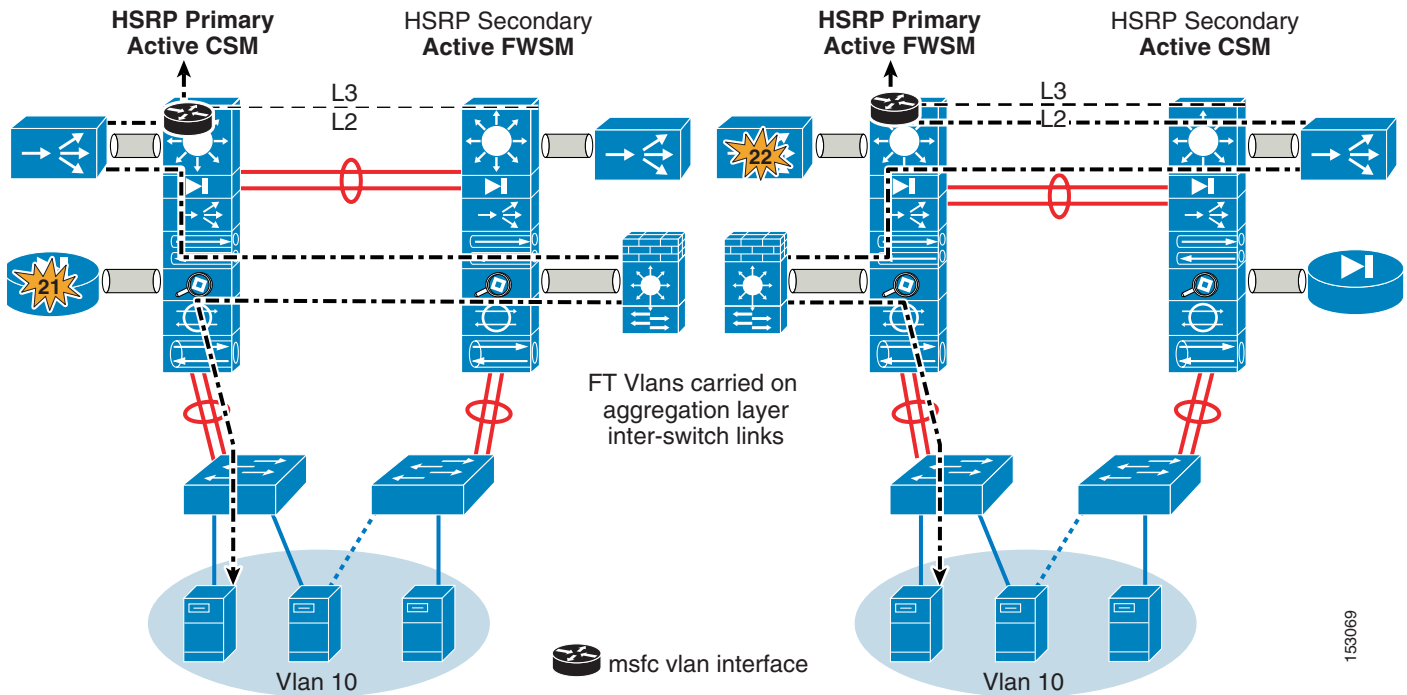
Figure 6-21 Loop-Free Inverted-U Failure Scenario 1—Uplink Failure

This failure is fairly obvious and straightforward. If servers are single attached, this results in a black hole condition. If servers use NIC teaming, they should experience a fairly short outage as they transition to the backup NIC and access switch.

As mentioned earlier, the use of DEC is recommended to reduce the chances of this failure scenario. Convergence times with a single link failure within a port channel group are under one second. The use of redundant supervisors in the access layer can also increase the resiliency of this design.

Failure 2—Service Module Failure (using CSM One-arm and FWSM Transparent Mode)

Figure 6-22 shows failure scenario 2.

Figure 6-22 Failure Scenario 2—Service Module Failure with Loop-Free Inverted U Topology

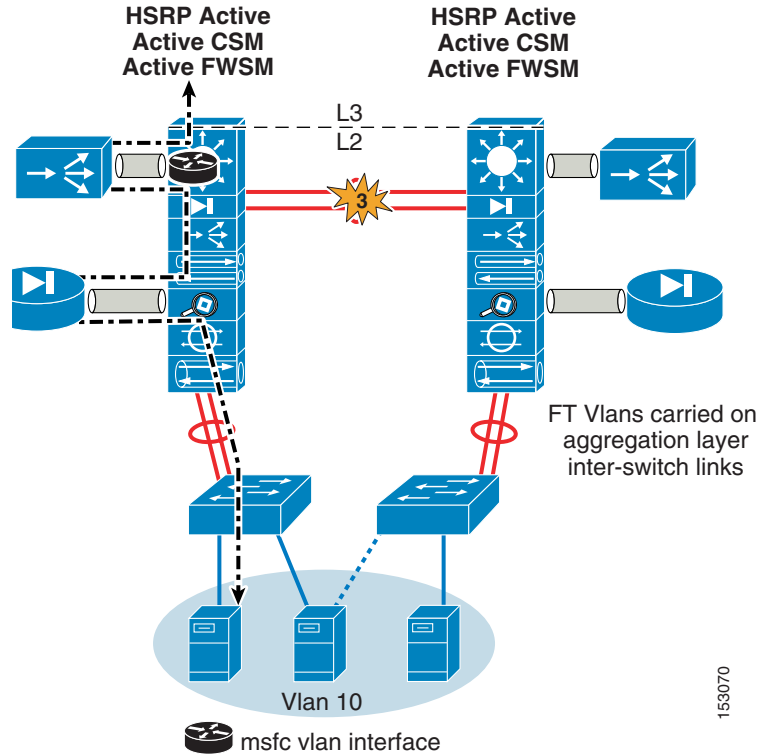
In this failure scenario, the backup service module moves to the active state on aggregation 2 because it no longer receives hello packets from the active service module, and times out.

Figure 6-22 shows the following two scenarios:

- 2.1 (FWSM failure)—Sessions cross the inter-switch link to aggregation 2 through the now-active FWSM module context and return back through the inter-switch link to the active HSRP default gateway on the aggregation 1 MSFC. Because the CSM is still active in aggregation 1, return traffic flow is directed to the CSM based on the PBR configuration on the MSFC VLAN interface, and on to the client via the core.
- 2.2 (CSM failure)—Sessions flow through the active FWSM module context in aggregation 1 and to the MSFC VLAN interface. The MSFC VLAN interface PBR configuration forces return CSM traffic to travel across the inter-switch link to aggregation 2 and through the now-active CSM module. Because the active default gateway of the CSM server VLAN is still active on aggregation 1, the traffic must return back across the aggregation layer inter-switch link to the MSFC on aggregation 1, and then on to the client via the core.

Failure 3—Inter-Switch Link Failure

Figure 6-23 shows failure scenario 3.

Figure 6-23 Loop-Free Inverted-U Failure Scenario 3—Inter-Switch Link Failure

This failure scenario has many side effects to consider. Because the service module fault tolerant (failover) VLANs are configured across the inter-switch link only, service modules in both aggregation switches determine that the other has failed. This results in service modules in aggregation 1 remaining in the active state and service modules in aggregation 2 moving from standby to the active state as well. This is commonly referred to as a split-brain effect, and is very undesirable because the opportunity for asymmetrical connection failure exists.

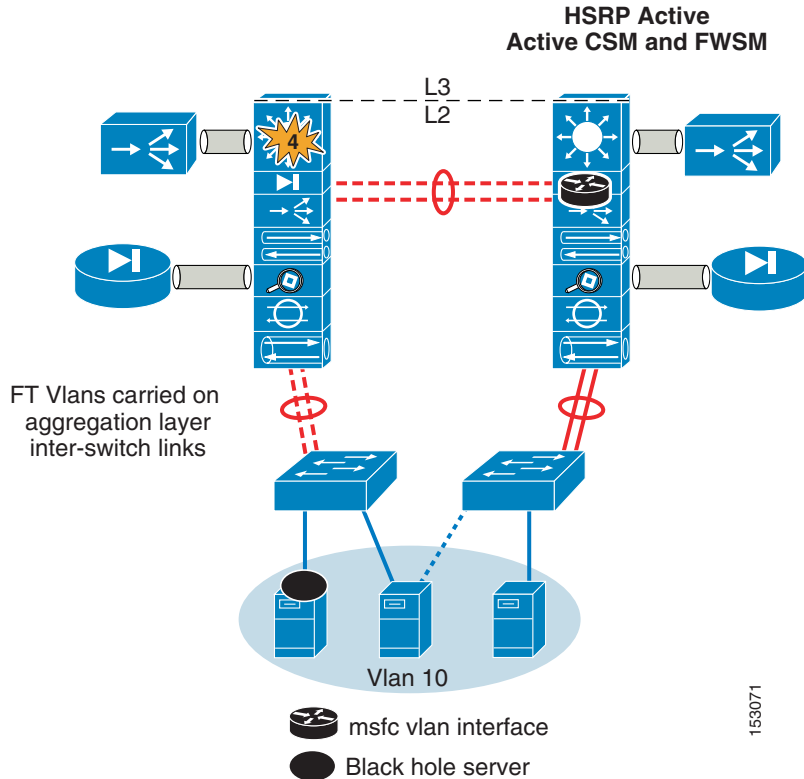
The path for HSRP heartbeats between MSFCs is also broken so both MSFC VLAN interfaces go into the HSRP active state without a standby.

If inbound traffic from the core flows into the aggregation 2 switch during this failure scenario, it reaches the MSFC and then attempts to flow through the now-active service modules. The core switches are performing CEF-based load balancing, thereby distributing sessions to both aggregation 1 and 2. Because state is maintained on the service modules, it is possible that asymmetrical connection failures can occur. For these reasons, Cisco recommends tuning the aggregation-core routing configuration such that the aggregation 1 switch is the primary route from the core for the primary service module-related VLANs.

Route tuning prevents asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combinations, thus preventing asymmetrical conditions. More information on route tuning is provided in [Establishing Path Preference with RHI, page 7-1](#).

Failure 4—Switch Power or Sup720 Failure (Non-redundant)

Figure 6-24 shows failure scenario 4.

Figure 6-24 Loop-Free Inverted-U Failure Scenario 4—Single Sup720 or Power Failure

In this failure scenario, the primary HSRP default gateway and active service modules transition to the aggregation 2 switch. Servers that are single attached to an access layer switch are black holed. NIC teaming can be used to prevent this failure.

The convergence characteristics of this failure scenario depend on HSRP and service module failover times. Because the HSRP failover time is expected to be under that of service modules, the actual convergence time depends on service module timer configurations. Test lab results show this convergence time to be ~5–6 seconds.

FlexLinks Access Model

FlexLinks are an alternative to the looped access layer topology. FlexLinks provide an active-standby pair of uplinks defined on a common access layer switch. After an interface is configured to be a part of an active-standby FlexLink pair, spanning tree is turned off on both links and the secondary link is placed in a standby state, which prevents it from being available for packet forwarding. FlexLinks operate in single pairs only, participate only in a single pair at a time, and can consist of mixed interface types with mixed bandwidth. FlexLinks are configured with local significance only because the opposite end of a FlexLink is not aware of its configuration or operation. FlexLinks also has no support for preempt, or an ability to return to the primary state automatically after a failure condition is restored.

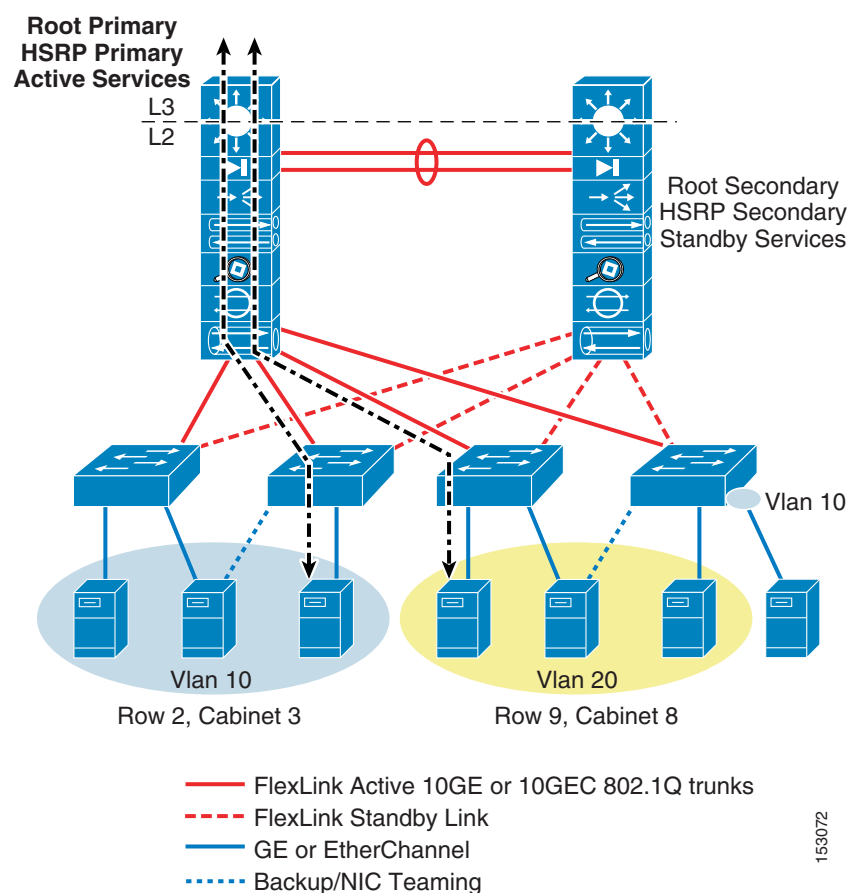
The main advantage of using FlexLinks is that there is no loop in the design and spanning tree is not enabled. Although this can have advantages in reducing complexity and reliance on STP, there is the drawback of possible loop conditions that can exist, which is covered in more detail later in this chapter. Other disadvantages are a slightly longer convergence time than R-PVST+, and the inability to balance traffic across both uplinks. Failover times measured using FlexLinks were usually under two seconds.

**Note**

When FlexLinks are enabled on the access layer switch, it is locally significant only. The aggregation switch ports to which FlexLinks are connected do not have any knowledge of this state, and the link state appears as up and active on both the active and standby links. CDP and UDLD packets still traverse and operate as normal. Spanning tree is disabled (no BPDUs flow) on the access layer ports configured for FlexLink operation, but spanning tree logical and virtual ports are still allocated on the aggregation switch line card. VLANs are in the forwarding state as type P2P on the aggregation switch ports.

Figure 6-25 shows the FlexLinks access topology.

Figure 6-25 FlexLinks Access Topology



The configuration steps for FlexLinks are as follows. FlexLinks are configured only on the primary interface.

```
ACCESS1#conf t
ACCESS1(config-if)#interface tenGigabitEthernet 1/1
ACCESS1(config-if)#switchport backup interface tenGigabitEthernet 1/2
ACCESS1(config-if)#
May 2 09:04:14: %SPANTREE-SP-6-PORTDEL_ALL_VLANS: TenGigabitEthernet1/2 deleted from all
Vlans
May 2 09:04:14: %SPANTREE-SP-6-PORTDEL_ALL_VLANS: TenGigabitEthernet1/1 deleted from all
Vlans
ACCESS1(config-if)#end
```

To view the current status of interfaces configured as FlexLinks:

```
ACCESS1#show interfaces switchport backup
```

Switch Backup Interface Pairs:

Active Interface	Backup Interface	State
TenGigabitEthernet1/1	TenGigabitEthernet1/2	Active Up/Backup Standby

Note that both the active and backup interface are in up/up state when doing a “show interface” command:

```
ACCESS1#sh interfaces tenGigabitEthernet 1/1
TenGigabitEthernet1/1 is up, line protocol is up (connected)
  Hardware is C6k 10000Mb 802.3, address is 000e.83ea.b0e8 (bia 000e.83ea.b0e8)
  Description: to_AGG1
  MTU 1500 bytes, BW 10000000 Kbit, DLY 10 usec,
    reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA, loopback not set
  Keepalive set (10 sec)
  Full-duplex, 10Gb/s
  input flow-control is off, output flow-control is off
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input 00:00:09, output 00:00:09, output hang never
  Last clearing of "show interface" counters 00:00:30
  Input queue: 0/2000/0/0 (size/max/drops/flushes); Total output drops: 0
  Queueing strategy: fifo
  Output queue: 0/40 (size/max)
  5 minute input rate 32000 bits/sec, 56 packets/sec
  5 minute output rate 0 bits/sec, 0 packets/sec
    1150 packets input, 83152 bytes, 0 no buffer
    Received 1137 broadcasts (1133 multicasts)
    0 runs, 0 giants, 0 throttles
    0 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored
    0 watchdog, 0 multicast, 0 pause input
    0 input packets with dribble condition detected
    26 packets output, 2405 bytes, 0 underruns
    0 output errors, 0 collisions, 0 interface resets
    0 babbles, 0 late collision, 0 deferred
    0 lost carrier, 0 no carrier, 0 PAUSE output
    0 output buffer failures, 0 output buffers swapped out
```

```
ACCESS1#sh interfaces tenGigabitEthernet 1/2
TenGigabitEthernet1/2 is up, line protocol is up (connected)
  Hardware is C6k 10000Mb 802.3, address is 000e.83ea.b0e9 (bia 000e.83ea.b0e9)
  Description: to_AGG2
  MTU 1500 bytes, BW 10000000 Kbit, DLY 10 usec,
    reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA, loopback not set
  Keepalive set (10 sec)
  Full-duplex, 10Gb/s
  input flow-control is off, output flow-control is off
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input 00:00:51, output 00:00:03, output hang never
  Last clearing of "show interface" counters 00:00:33
  Input queue: 0/2000/0/0 (size/max/drops/flushes); Total output drops: 0
  Queueing strategy: fifo
  Output queue: 0/40 (size/max)
  5 minute input rate 32000 bits/sec, 55 packets/sec
  5 minute output rate 0 bits/sec, 0 packets/sec
    1719 packets input, 123791 bytes, 0 no buffer
    Received 1704 broadcasts (1696 multicasts)
```

```

0 runs, 0 giants, 0 throttles
0 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored
0 watchdog, 0 multicast, 0 pause input
0 input packets with dribble condition detected
7 packets output, 1171 bytes, 0 underruns
0 output errors, 0 collisions, 0 interface resets
0 babbles, 0 late collision, 0 deferred
0 lost carrier, 0 no carrier, 0 PAUSE output
0 output buffer failures, 0 output buffers swapped out
ACCESS1#

```

Note that both the spanning-tree is no longer sending BPDUs in an effort to detect loops on interfaces in the FlexLink pair:

```

ACCESS1#sh spanning-tree interface tenGigabitEthernet 1/1
no spanning tree info available for TenGigabitEthernet1/1
ACCESS1#sh spanning-tree interface tenGigabitEthernet 1/2
no spanning tree info available for TenGigabitEthernet1/2

```

CDP and UDLD packets are still transmitted across Flexlinks as shown below:

```

ACCESS1#show cdp neighbor
Capability Codes: R - Router, T - Trans Bridge, B - Source Route Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater, P - Phone

Device ID          Local Intrfce    Holdtme    Capability  Platform  Port ID
Aggregation-1.cisco.com
                  Ten 1/1          156        R S         WS-C6509  Ten 7/4
Aggregation-2.cisco.com
                  Ten 1/2          178        R S         WS-C6509  Ten 7/4

ACCESS1#show udld neighbor
Port      Device Name      Device ID      Port ID      Neighbor State
-----
Te1/1     TBM06108988      1              Te7/4        Bidirectional
Te1/2     SCA0332000T      1              Te7/4        Bidirectional
ACCESS1#

```

Spanning Tree, HSRP, and Service Module Design

FlexLinks automatically disable spanning tree BPDUs on both the active and standby links, as noted in the preceding section. Cisco still recommends enabling spanning tree on the aggregation switches that are connected to FlexLink-enabled access switches. It is also desirable to align the spanning tree root, HSRP default gateway, and active service modules on the same aggregation switch just as recommended in looped access layer designs. This is shown in [Figure 6-25](#). By aligning the primary access layer switch uplink directly to the same switch that is the primary default gateway and active service module/appliance, traffic flows are optimized. Otherwise, traffic flows can hop back and forth between aggregation switches, creating undesirable conditions and difficulty in troubleshooting.

Implications Related to Possible Loop Conditions

Because spanning tree is disabled on FlexLinks, there is the possibility that a loop condition can exist in particular scenarios, such as a patch cable that is mistakenly connected between access layer switches that are configured for FlexLinks. This is shown in [Figure 6-26](#).

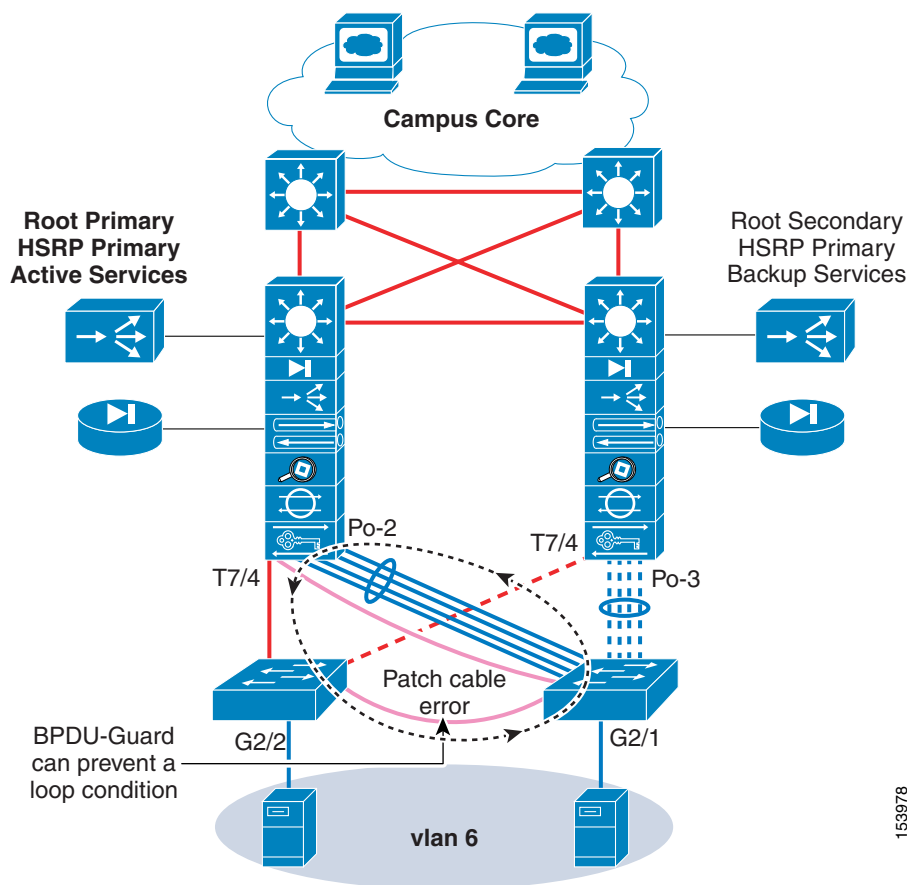
Figure 6-26 Possible Loop Condition

Figure 6-26 demonstrates two possible loop conditions that can be introduced by configuration error or patch cable error. The first example demonstrates a connection between the aggregation switch and an access switch. This can be the result of an incorrect patch/uplink cable or simply the configuration of a separate link that is not part of the FlexLink channel group. Because STP BPDUs are not passed along the FlexLink path, a loop in the topology cannot be detected, and an endless replication of broadcast/multicast frames occurs that can have a very negative impact on the whole aggregation module. Note that RootGuard is ineffective in this scenario because Agg1 does not see a path to the root (Agg2) through the access switch with FlexLinks enabled.

The second example demonstrates a patch cable connection error between access switches.

If BPDU Guard is supported and enabled on access layer server ports, the port is automatically disabled when BPDUs are detected, as shown in the following console message:

```
ACCESS1#
Apr 13 16:07:33: %SPANTREE-SP-2-BLOCK_BPDUGUARD: Received BPDU on port GigabitEthernet2/2
with BPDU Guard enabled. Disabling port.
Apr 13 16:07:33: %PM-SP-4-ERR_DISABLE: bpduguard error detected on Gi2/2, putting Gi2/2 in
err-disable state
ACCESS1#sh int g2/2
GigabitEthernet2/2 is administratively down, line protocol is down (disabled)
```

If BPDU Guard is not supported or is not enabled on access layer server ports, a loop condition occurs. This loop condition endlessly forwards multicast and broadcast packets through the aggregation 1 switch and back through the access switches via the patch cable link that now extends between them. This could create negative conditions that affect all servers connected to this aggregation module.

**Note**

Because spanning tree BPDUs are not passed from the access layer switch when using FlexLinks, cabling or configuration errors can create a loop condition with negative implications. Although cabling mistakes such as these might be considered rare, the degree of change control in your data center environment can be the barometer in determining whether Flexlinks are a proper solution.

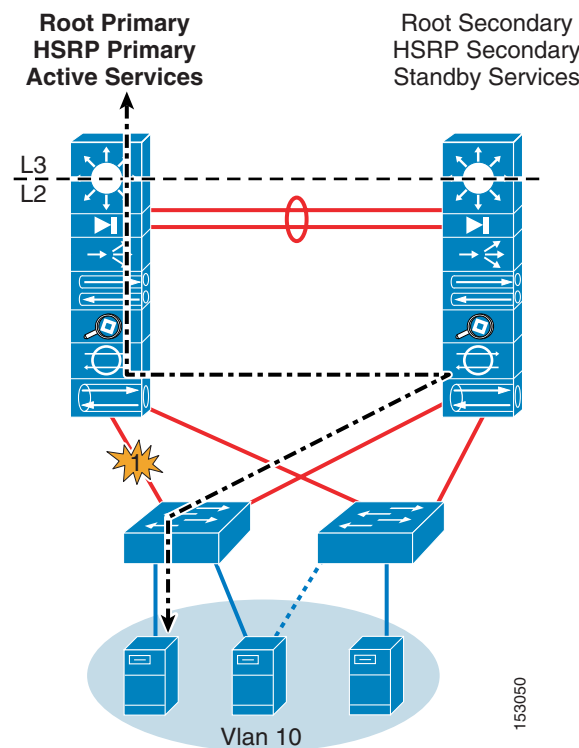
Failure Scenarios

The level of resiliency that is incorporated into the access layer design can vary based on the model used. Other features such as route health injection and route tuning can influence this. The four main failure scenarios that can occur in a looped access design are covered in this section. Understanding the amount of exposure in these scenarios helps to determine the best access layer design selection.

Failure 1—Access Layer Uplink Failure

Figure 6-27 shows failure scenario 1.

Figure 6-27 FlexLinks Failure Scenario 1—Uplink Down



153050

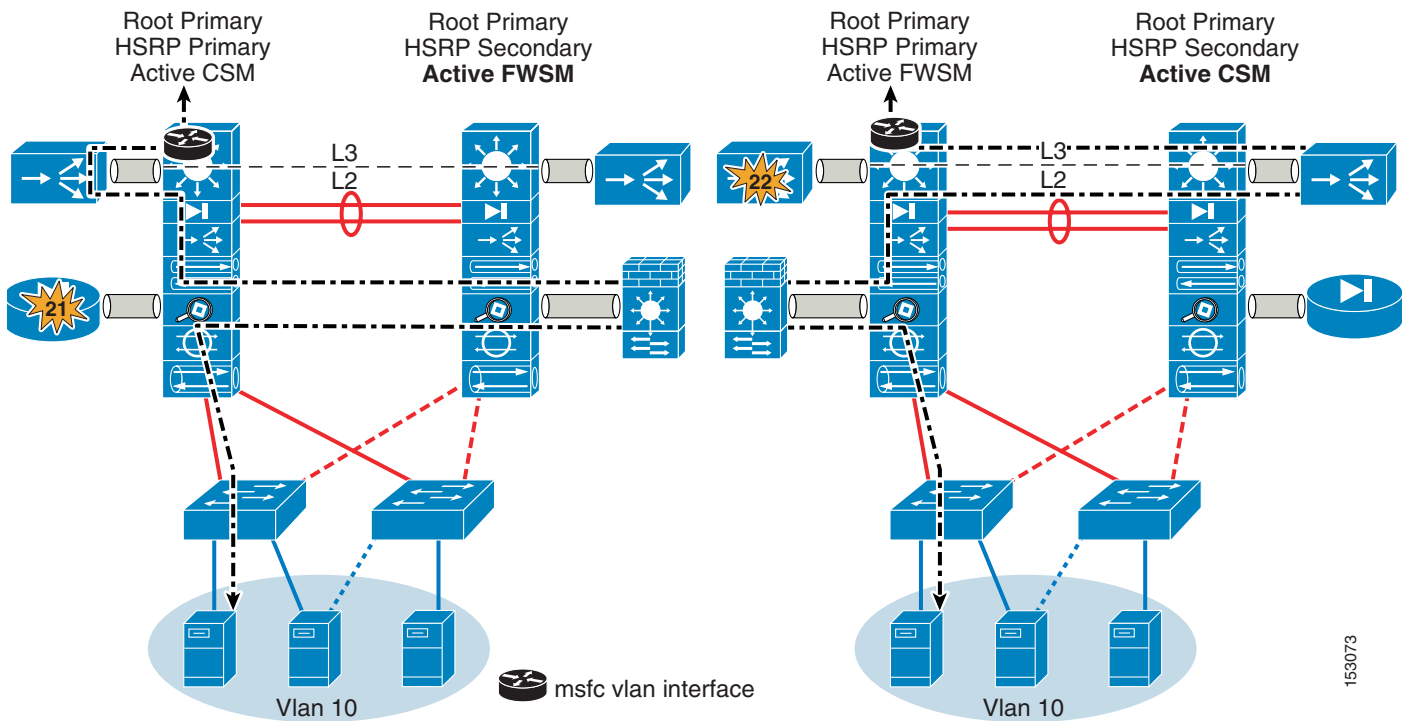
In this failure scenario, the backup FlexLink goes active and immediately begins to pass packets over its interface. Default gateway and active service modules remain on aggregation 1 unless tracking mechanisms are configured and triggered. Traffic flow goes through aggregation 2 and uses the inter-switch link to aggregation 1 to reach the active HSRP default gateway and active service modules.

The convergence characteristics of this failure scenario are typically less than 2 seconds.

Failure 2—Service Module Failure (using CSM One-arm and FWSM Transparent Mode)

Figure 6-28 shows failure scenario 2.

Figure 6-28 FlexLinks Failure Scenario 2—Service Modules



In this failure scenario, there is no FlexLink convergence and the primary default gateway remains active on the aggregation 1 switch. The backup service module moves to the active state on aggregation 2 because it no longer receives hello packets from the failed active service module and times out.

Figure 6-28 shows the following two failure instances:

- 2.1 (FWSM failure)—Traffic flow goes through aggregation 1 and across the inter-switch link to aggregation 2, through the now-active FWSM module context, and back across the inter-switch link to the active HSRP default gateway on the aggregation 1 MSFC. Because the CSM is still active in aggregation 1, return traffic flow is directed to the CSM based on the PBR configuration on the MSFC VLAN interface and on to the client via the core.
- 2.2 (CSM failure)—Traffic flow goes through aggregation 1, through the active FWSM module context in aggregation 1, and to the MSFC VLAN interface. The MSFC VLAN interface PBR configuration forces return CSM traffic to travel across the inter-switch link to aggregation 2 and through the now-active CSM module. Because the active default gateway of the CSM server VLAN is still active on aggregation 1, the traffic must flow back across the inter-switch link to the MSFC on aggregation 1 and then on to the client via the core.

The convergence characteristics of these failure scenarios depend on the service module(s) failover time. The recommended service module failover timer configurations are as follows:

- CSM:

```
module ContentSwitchingModule 3
  ft group 1 vlan 102
  priority 20
  heartbeat-time 1
  failover 3
  preempt
```

- FWSM:

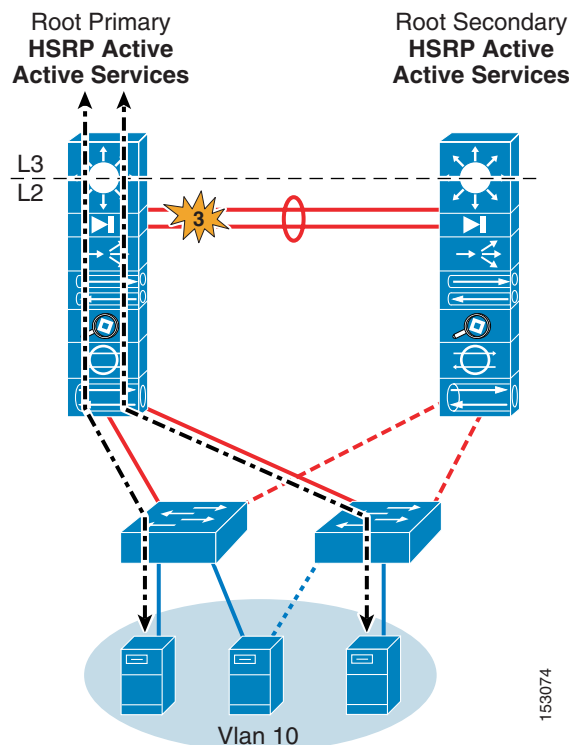
```
Unit Poll frequency 500 milliseconds, holdtime 3 seconds
Interface Poll frequency 3 seconds
```

Test lab results show that average service module failover times with these values is under ~5 seconds.

Failure 3—Inter-Switch Link Failure

Figure 6-29 shows failure scenario 3.

Figure 6-29 FlexLinks Failure Scenario 3—Inter-Switch Link Failure



FlexLinks do not converge in this failure scenario.

With the inter-switch link down, HSRP multicast hello messages no longer have a path between aggregation 1 and 2, so HSRP goes into an active state on both switches for all VLANs.

Service modules in both aggregation switches determine that the other has failed and become active (this is referred to as a split-brain effect).

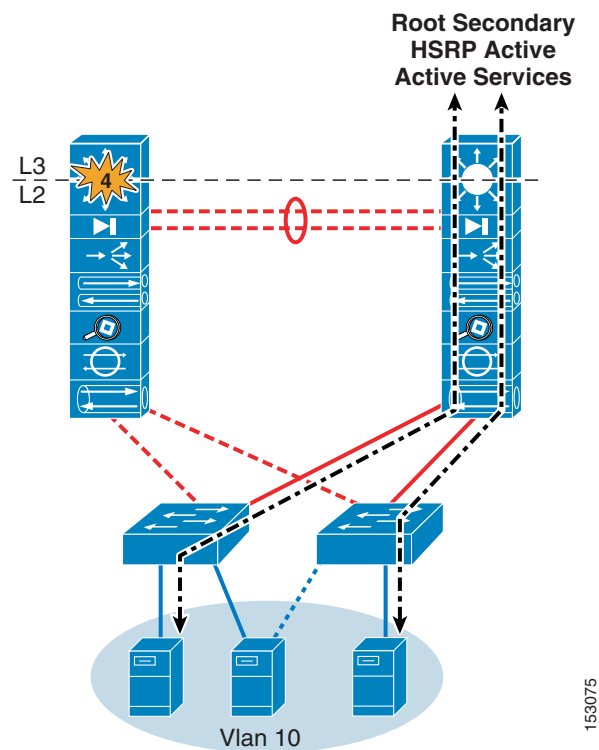
If inbound traffic from the core flows into the aggregation 2 switch during this failure scenario, it attempts to flow through the now-active service modules and stops because the path to the servers is blocked by a standby FlexLink. For these reasons, Cisco recommends tuning the aggregation-core routing configuration such that the aggregation 1 switch is the primary route advertised to the core for the primary service module-related VLANs.

Route tuning helps to prevent asymmetrical connections and black holing in a split-brain scenario because traffic flows are aligned with the same default gateway and service module combination, preventing asymmetrical conditions.

Failure 4—Switch Power or Sup720 Failure (Non-redundant)

Figure 6-30 shows failure scenario 4.

Figure 6-30 FlexLinks Failure Scenario 4—Switch Power or Sup720 Failure



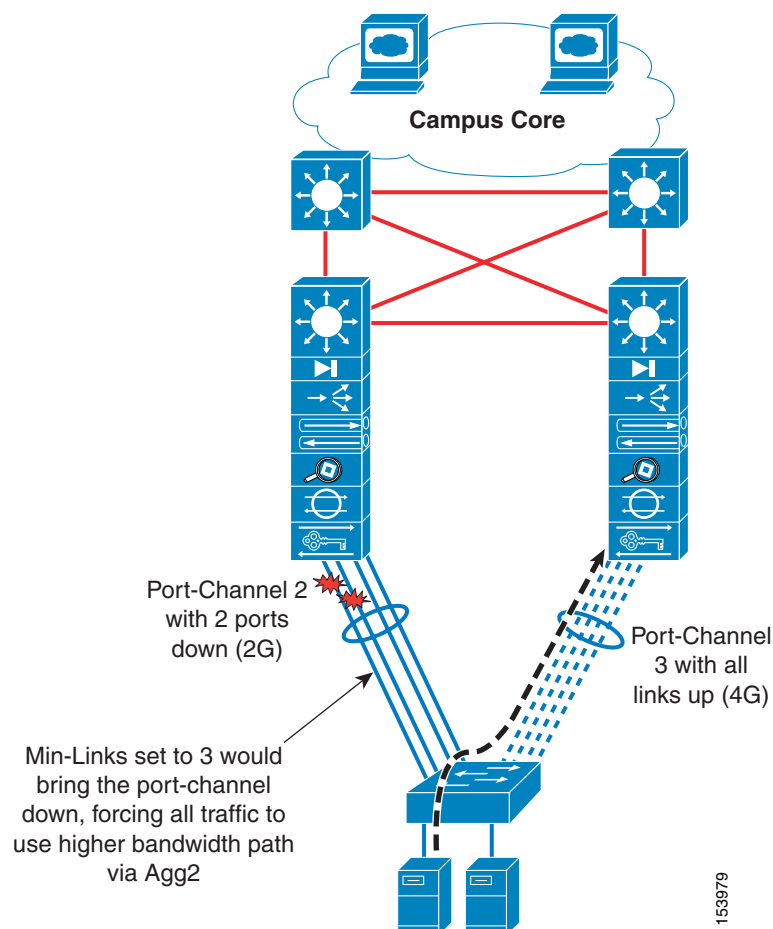
In this failure scenario, the active FlexLinks, primary default gateway, and active service modules transition to the aggregation 2 switch.

The convergence characteristics of this failure scenario depend on FlexLink failure detection, HSRP failover, and service module failover times. Because the FlexLink and HSRP failover times are expected to be under that of service modules, the actual convergence time depends on service module timer configurations.

Using EtherChannel Min-Links

EtherChannel Min-Links is a new feature as of the Cisco 12.2.18 SXF IOS Release. EtherChannel Min-Links permit you to designate the minimum number of member ports that must be in the link-up state and bundled in an LACP EtherChannel for a port channel interface to be in a link-up state. In the data center access layer, this can be useful in making sure that a higher bandwidth uplink path is chosen as the active path. For example, consider the diagram in [Figure 6-31](#).

Figure 6-31 Using EtherChannel Min-Links



In the above example, 4G EtherChannels connect the access layer switch to both aggregation layer switches. A failure has occurred that has taken down two of the port members on the EtherChannel to the aggregation 1 switch. Because two members of the port channel are still up, the port channel itself remains up and server traffic uses this path as normal, although it is a path with less available bandwidth. With EtherChannel Min-Links, you can designate a minimum number of required ports that must be active or the port channel is taken down. In this example, if EtherChannel Min-Links are set to 3, the port channel is taken down and server traffic is forced to use the higher 4G bandwidth path towards the aggregation 2 switch.

The EtherChannel Min-Links feature requires the LACP EtherChannel protocol to be used. The access layer topology can consist of looped, loop-free, or FlexLink models. The Min-Links feature works at the physical interface level and is independent of spanning tree path selection. Consider the following when deciding whether Min-Links should be used:

- Active/standby service modules are used—If active services are primarily on the aggregation 1 switch, a failure that forces Min-Links to use the path to aggregation 2 will likely cause all traffic to also traverse the inter-switch link between the aggregation switches.
- Looped topologies with spanning tree—If a looped access topology is used, it is possible to provide a similar capability by using the **spanning-tree pathcost method long** global option. This permits spanning tree to use larger cost values when comparing the cost of different paths to root, which in turn can differentiate the cost value of various paths when a port member fails.
- Dual failures—With Min-Links, it is possible to have a situation where if both EtherChannels do not have the minimum required port members, both uplinks would be forced down, which would black-hole all connected servers.

The configuration steps for EtherChannel Min-Links are as follows:

```
ACCESS2#conf t
Enter configuration commands, one per line. End with CNTL/Z.
ACCESS2(config)#interface port-channel 2
ACCESS2(config-if)#port-channel ?
    min-links  Minimum number of bundled ports needed to bring up this port
                channel
ACCESS2(config-if)#port-channel min-links ?
    <2-8>  The minimum number of bundled ports needed before this port channel
          can come up.
ACCESS2(config-if)#port-channel min-links 3
ACCESS2(config-if)#end
```



Increasing HA in the Data Center



Note

The README file posted with this guide contains details relative to technologies, hardware, and software that were used in producing this document. There is also a revision history that details updates made to each chapter.

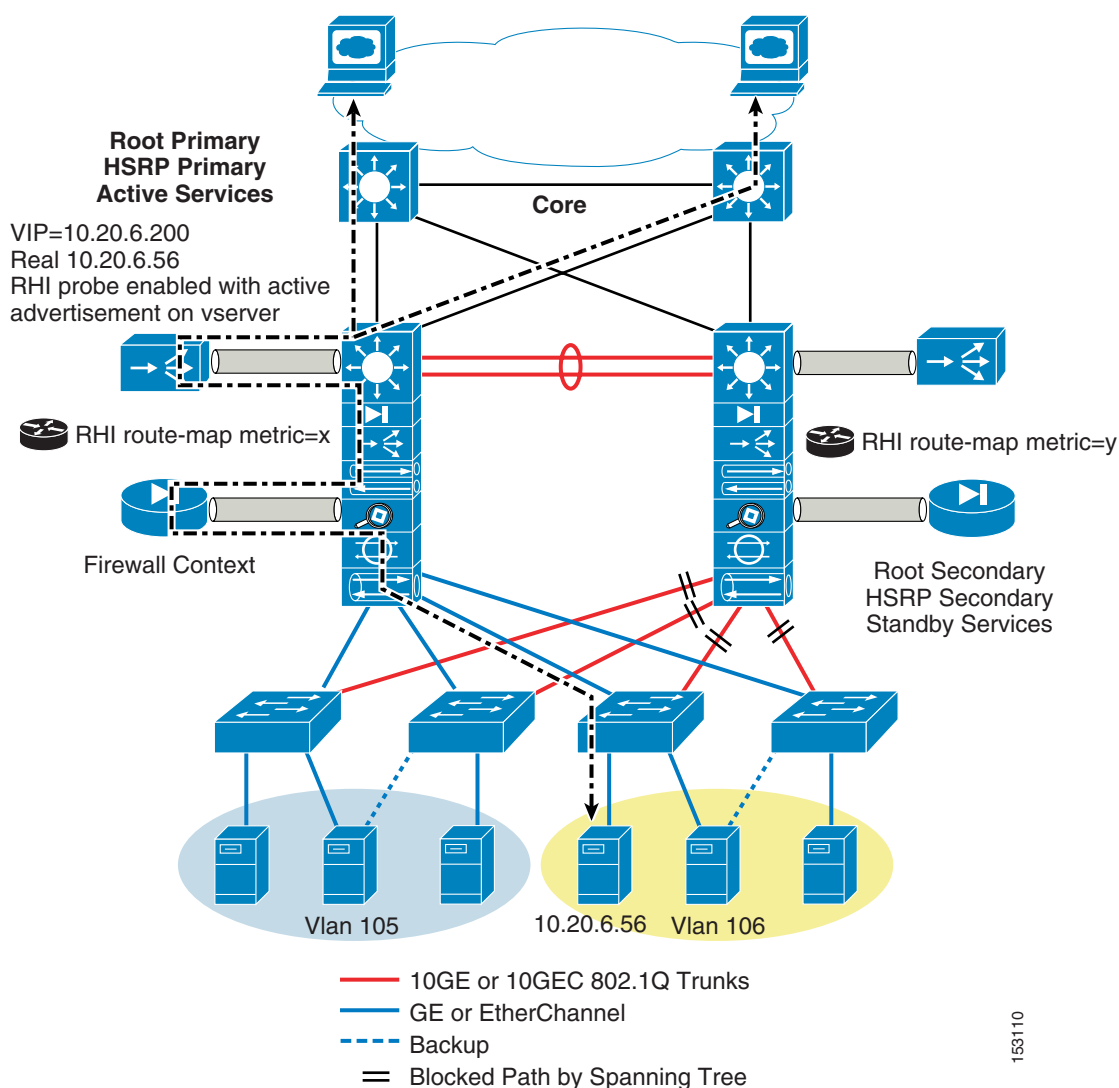
This chapter provides details of Cisco tested high availability solutions in the enterprise data center. It includes the following topics:

- [Establishing Path Preference with RHI](#)
- [Service Module FT Paths](#)
- [NSF-SSO in the Data Center](#)

Establishing Path Preference with RHI

When active/standby service module pairs are used, it becomes important to align traffic flows such that the active/primary service modules are the preferred path to a particular server application. This is desirable because it creates a design that is more deterministic and easier to troubleshoot but it also becomes particularly important in failure scenarios such as the inter-switch trunk failure described previously.

[Figure 7-1](#) shows an aggregation layer with route preference established toward the primary service modules in an active/standby pair.

Figure 7-1 Route Preference toward Primary Service Modules in an Active/Standby Pair

By using Route Health Injection (RHI) combined with specific route map attributes, a path preference is established with the core so that all sessions to a particular VIP go to agg1 where the primary service modules are located.

The RHI configuration is accomplished by defining a probe type and related values in the Cisco Content Switching Module (CSM) portion of the Cisco IOS configuration. The following probe types are supported:

```

dns          slb dns probe
ftp          slb ftp probe
http         slb http probe
icmp         slb icmp probe
kal-ap-tcp   KAL-AP TCP probe
kal-ap-udp   KAL-AP UDP probe
name         probe with this name
real         SLB probe real suspects information
script       slb script probe
smtp         slb smtp probe
tcp          slb tcp probe

```

```
telnet      slb telnet probe
udp         slb udp probe
```

In the following configuration, a simple ICMP probe is defined, after which it is attached to the server farm configuration. This initiates the probe packets and the monitoring of each real server defined in the server farm.

The last step in configuring RHI is indicating that you want the VIP address to be advertised based on the probe status being operational. If the probe determines that the servers are active and healthy, it inserts a /32 static route for the VIP address into the MSFC configuration.

Aggregation 1 CSM Configuration

```
module ContentSwitchingModule 3
  ft group 1 vlan 102
    priority 20
    heartbeat-time 1
    failover 3
    preempt
  !
  vlan 44 server
    ip address 10.20.44.42 255.255.255.0
    gateway 10.20.44.1
    alias 10.20.44.44 255.255.255.0
  !
  probe RHI icmp
    interval 3
    failed 10
  !
  serverfarm SERVER200
    nat server
    no nat client
    real 10.20.6.56
    inservice
    probe RHI
  !
  vserver SERVER200
    virtual 10.20.6.200 any
    vlan 44
    serverfarm SERVER200
    advertise active
    sticky 10
    replicate csrp sticky
    replicate csrp connection
    persistent rebalance
    inservice
```

With the static host route installed by the CSM into the MSFC based on the health of the server farm, you can now advertise the host route to the core with a route path preference to the active VIP in aggregation 1. This is accomplished with the **redistribute** command in the router process that points to a specific route map. The route map points to an access list that identifies the VIP addresses to match against, and also permits metric attributes to be set to establish path preference. In the following OSPF configuration, **set metric-type type-1** is used to set the host route advertisement to an OSPF external type-1. Alternative methods can include setting the actual OSPF metric value. By setting the host route to an OSPF external type-1, the route appears in the core with the actual accumulated cost of the path used. This approach could prove to be more desirable than attempting to set specific values because it better reflects actual path costs in the case of link failures.

**Note**

The VIP server subnet itself should not be included in the router network statements. If 10.20.6.0 were to be advertised, it would be the next most exact route if the VIP host route, 10.20.6.200, were to not be advertised in an actual failure situation. This would defeat the purpose of advertising the VIP as healthy, allowing sessions to continue to be directed to the agg1 switch.

Aggregation 1 OSPF and Route Map Configurations

```
router ospf 10
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 nsf
 area 10 authentication message-digest
 area 10 nssa
 timers throttle spf 1000 1000 1000
 redistribute static subnets route-map rhi
 passive-interface default
 no passive-interface Vlan3
 no passive-interface TenGigabitEthernet7/2
 no passive-interface TenGigabitEthernet7/3
 network 10.10.20.0 0.0.0.255 area 10
 network 10.10.40.0 0.0.0.255 area 10
 network 10.10.110.0 0.0.0.255 area 10
 (note: server subnet 10.20.6.0 is not advertised)
 access-list 44 permit 10.20.6.200 log
 route-map rhi permit 10
 match ip address 44
 set metric-type type-1
 set metric +(value) (on Agg2)
```

Aggregation Inter-switch Link Configuration

The design in [Figure 7-1](#) uses VLAN 3 between the aggregation switches to establish a Layer 3 OSPF peering between them. This VLAN traverses a 10GE-802.1Q trunk. With VLANs that cross a 10GE trunk, OSPF sets the bandwidth value equal to a GE interface, not a 10GE interface as one might expect. If the bandwidth on the VLAN configuration is not adjusted to reflect an equal value to the aggregation-core 10GE links, the route from agg2 to the active VIP appears better via the core instead of via VLAN 3 on the inter-switch trunk. At first, this might not appear to be a real problem because the core is going to use the preferred paths directly to agg1 anyway. However, certain link failures can create a scenario where sessions come through the agg2 switch, which would then need to be routed to agg1, so it makes sense to keep the optimal path via the inter-switch link rather than hopping around unnecessarily back to the core. The following configuration reflects the bandwidth changes to show this:

```
interface Vlan3
 description AGG1_to_AGG2_L3-RP
 bandwidth 10000000
 ip address 10.10.110.1 255.255.255.0
 no ip redirects
 no ip proxy-arp
 ip pim sparse-dense-mode
 ip ospf authentication message-digest
 ip ospf message-digest-key 1 md5 C1sC0!
 ip ospf network point-to-point
 ip ospf hello-interval 1
 ip ospf dead-interval 3
 logging event link-status
```

Aggregation 2 Route Map Configuration

The aggregation 2 switch requires the same configurations as those outlined for aggregation 1. The only difference is with the route map configuration for RHI. Because you want path preference to be toward the active VIP in the Aggregation 1 switch, you need to adjust the metric for the advertised RHI route to be less favorable in agg2. The reason this is necessary is to make sure that in an active-active service module scenario, there is symmetry in the connections to prevent asymmetrical flows that would break through the CSM and Cisco Firewall Service Module (FWSM). The following route map adds cost to the OSPF route advertised by using the **set metric +** command. Note that the type-1 is also set for the same reasons mentioned previously.

```
route-map rhi permit 10
  match ip address 44
  set metric +30
  set metric-type type-1
```

The following command can be used to view the status of an RHI probe configuration:

```
Aggregation-1#sh module contentswitchingModule 3 probe detail
```

probe	type	port	interval	retries	failed	open	receive
RHI	icmp	3	3	10			
real		vserver		serverfarm		policy	status
10.20.6.56:0		SERVER200		SERVER200		(default)	OPERABLE
10.20.6.25:0		SERVER201		SERVER201		(default)	OPERABLE

Aggregation-1#

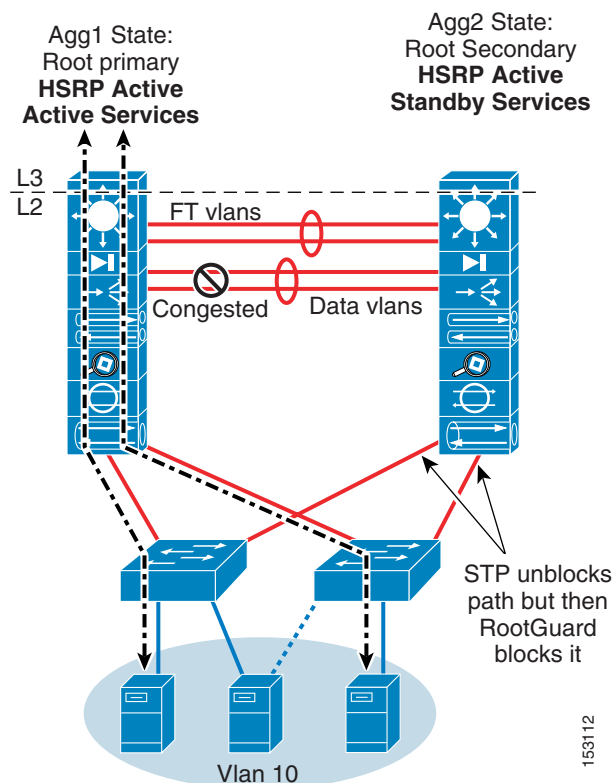
Service Module FT Paths

Service module redundant pairs monitor each other to ensure availability as well as to maintain state for all sessions that are currently active. The availability is provided by a hello protocol that is exchanged between the modules across a VLAN. Session state is provided by the active service module replicating the packet headers to the standby across the same VLAN as the hellos, as in the case of the CSM, or on a separate VLAN as in the case of the FWSM.

If the availability path between the redundant pairs becomes heavily congested or misconfigured, it is likely that the service modules believe the other has failed. This can create a split-brain scenario where both service modules move into an active state, which creates undesirable conditions including asymmetric connection attempts.

The CSM exchanges hello and session state on a common VLAN. The FWSM uses separate VLANs for both hello exchange and state replication. The standby FWSM assumes the active role when it no longer sees its redundant peer on at least two VLAN interfaces. This VLAN could be a combination of the context, failover, or state VLANs.

If a second inter-switch link were added with only service module FT vlans provisioned across it as shown in [Figure 7-2](#), the chance of a split-brain scenario because of these conditions is reduced.

Figure 7-2 Congestion on FT Path

The bandwidth required for the FT link must be considered. The maximum possible required can be equal to the CSM bus interface (4G) or the FWSM bus interface (6G) as a worst case. The required bandwidth is based on the amount and type of sessions that are being replicated.

**Note**

More detail on access layer design is covered in [Chapter 6, “Data Center Access Layer Design.”](#)

NSF-SSO in the Data Center

**Note**

The testing performed in support of this section included the use of CSM one-arm mode and FWSM transparent mode design, which influences the behavior of NSF/SSOs failover characteristics.

The data center solutions that are covered in this and other data center guides are designed for a high level of resiliency. For example, the core and aggregation layer switches are always in groups of two, and are interconnected such that no individual module or full system failure can bring the network down. The service modules and other software and hardware components are configured in redundant pairs that are located in each of the aggregation switches to further remove any single point of failure. The access layer also has many options that permit dual homing to the aggregation layer and leverage spanning tree, FlexLinks, and NIC teaming to achieve high availability.

The main objective in building a highly available data center network design is to avoid TCP session breakage while providing convergence that is unnoticeable, or as fast as possible. Each of the TCP/IP stacks that are built into the various operating systems have a different level of tolerance for determining

when TCP will break a session. The least tolerant are the Windows Server and Windows XP client stacks, which have been determined to have a ~9 second tolerance. Other TCP/IP stacks such as those found in Linux, HP, and IBM are more tolerant and have a longer window before tearing down a TCP session.

This does not necessarily mean that the data center network should be designed to converge in less than 9 seconds, but it could serve as a guideline to a worst case. The most optimal acceptable failure convergence time is zero, of course. Although each network has its own particular convergence time requirements, many network designers usually break acceptable convergence times into the following categories:

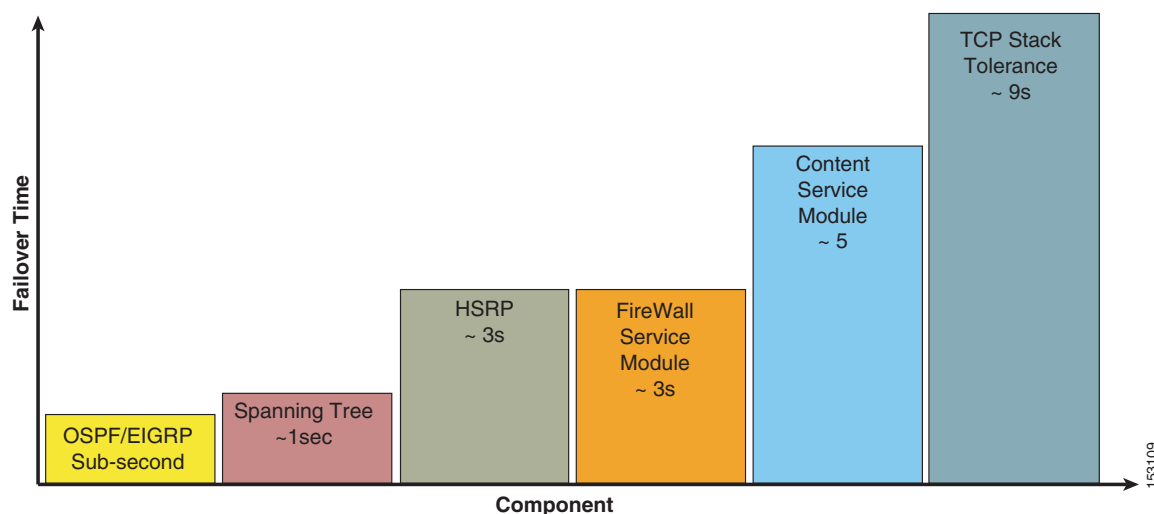
- **Minor failures**—Failures that would be expected to happen because of more common events, such as configuration errors or link outages because of cable pulls or GBIC/Xenpak failure, might be considered to be a minor failure. A minor failure convergence time is usually expected to be in the sub-second to 1 second range.
- **Major failures**—Any failure that can affect a large number of users or applications is considered a major failure. This could be because of a power loss, supervisor, or module failure. This type of failure usually has a longer convergence time and is usually expected to be under 3–5 seconds.

The following describes the various recovery times of the components in the Cisco data center design:

- **HSRP**—With recommended timers of Hello=1/Holddown=3, convergence occurs in under 3 seconds. This can be adjusted down to sub-second values, but CPU load must be considered.
- **Routing protocols**—OSPF and EIGRP can both achieve sub-second convergence time with recommended timer configurations.
- **Rapid PVST+ Spanning Tree**—802.1W permits sub-second convergence time for minor failures when logical ports are under watermarks, and usually 1–2 seconds for major failure conditions.
- **CSM**—Convergence time is ~5 seconds with recommended timers.
- **FWSM**—Convergence time is ~3 seconds with recommended timers.

Figure 7-3 shows the various failover times of the components of the Cisco data center design.

Figure 7-3 Failover Times



The worst case convergence time for an individual component failure condition is the CSM at ~5 seconds. In the event of a Supervisor720 failure on the Aggregation 1 switch, all of these components must converge to the Aggregation 2 switch, resulting in a minimum of ~5 second convergence time. This convergence time will most likely be more because of the tables that have to be rebuilt (ARP, CAM tables, and so on), so maximum convergence time can approach the 9 second limit of the Windows TCP/IP stack. This convergence time and possible lost sessions can be avoided by using dual Sup720s with NSF-SSO on the primary aggregation switch of the data center.

Supervisor 720 supports a feature called Non-Stop Forwarding with Stateful Switch-Over (NSF-SSO) that can dramatically improve the convergence time in a Sup720 failure condition. NSF with SSO is a supervisor redundancy mechanism on the Supervisor Engine 720 in Cisco IOS Release 12.2(18)SXD that provides intra-chassis stateful switchover. This technology demonstrates extremely fast supervisor switchover with lab tests resulting in approximately 1.6–2 seconds of packet loss.

The recommended data center design that uses service modules has a minimum convergence time of ~6–7 seconds primarily because of service modules. With NSF/SSO, the service modules do not converge. This alone represents a large reduction in convergence time, making dual supervisors with NSF-SSO a tool for achieving increased high availability in the data center network.

Possible Implications

HSRP

With the current 12.2.(18) release train, HSRP state is not maintained by NSF-SSO. The HSRP static MAC address is statefully maintained between supervisors but the state of HSRP between aggregation nodes is not. This means that if the primary Sup720 fails, the SSO-enabled secondary Sup720 takes over and continues to forward traffic that is sent to the HSRP MAC address, but the HSRP hellos that are normally communicated to the standby HSRP instance on agg2 are not communicated. This means that during a switchover on aggregation 1, the aggregation 2 switch HSRP instances take over as primary during the SSO control plane recovery. Because the HSRP MAC address was statefully maintained on the agg1 standby Sup720 module, the sessions continue to flow through agg1, regardless of the active state that appears on agg2. After the control plane comes up on the agg1 switch, the HSRP hello messages begin to flow, and preemptively move the active state back to the agg1 switch. The control plane recovery time is ~2 minutes.

With looped access layer topologies (triangle and square) that align HSRP, STP primary root, and active service modules on the agg1 switch, this does not create an issue because the access layer to aggregation layer traffic flow continues to be directed to the agg1 switch. If a loop-free access layer design is used, the active HSRP default gateway instance on agg2 responds to ARP requests.



Note

A square looped access also has active-active uplinks, but the FWSM transparent mode active context on agg1 prevents packets from reaching the active HSRP default gateway on agg2. The VLAN on the south side of the context follows the spanning tree path from agg2, across the inter-switch link to the active FWSM on agg1, then to the HSRP default gateway instance on agg1.

IGP Timers

It is possible that IGP timers can be tuned low enough such that NSF/SSO is defeated because the failure is detected by adjacent nodes before it is determined to be an SSO stateful switchover.

Slot Usage versus Improved HA

Placing two Sup720s in a data center aggregation node also has its drawbacks in terms of available slot density. Particularly with 10GE port density challenges, using an available slot that could be available for other purposes might not be a necessary trade-off. Consideration of using dual Sup720s should be based on actual customer requirements. The recommendation is to use the dual supervisor NSF-SSO solution in the primary aggregation node with service modules when slot density is not an issue or when HA is critical at this level. If a Network Analysis Module is used, it can be placed in the agg2 switch while the dual Sup720s are in the agg1 switch, balancing the slot usage across the two aggregation nodes.

Recommendations

NSF-SSO can provide a very robust solution to data centers that require a very high level of resiliency and are willing to use available slots to achieve it. The processes and functions that are involved for NSF/SSO to work correctly are very complex and have many inter-related dependencies. These dependencies involve the access layer design and service module modes used, for example. This guide provides a solution that permits NSF/SSO to provide a lower convergence time than the base recommendation design can provide based on service module failover times. Cisco recommends testing NSF/SSO to ensure that it works as expected in a specific customer design.



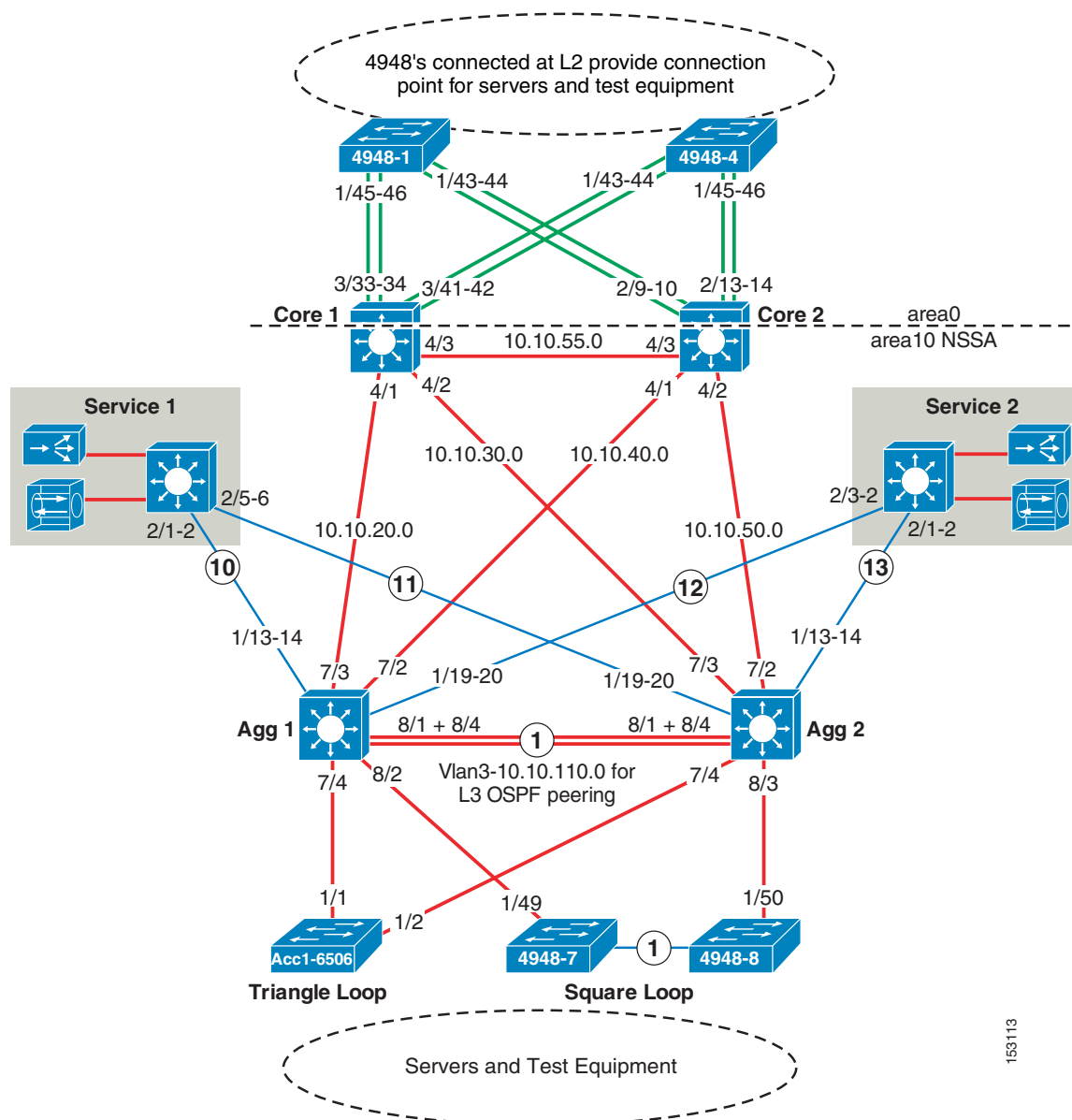
Configuration Reference

This chapter provides the test bed diagram and configurations used in tests to support this guide. It includes the following configuration listings:

- [Core Switch 1](#)
- [Aggregation Switch 1](#)
- [Core Switch 2](#)
- [Aggregation Switch 2](#)
- [Access Switch 4948-7](#)
- [Access Switch 4948-8](#)
- [Access Switch 6500-1](#)
- [FWSM 1-Aggregation Switch 1 and 2](#)

[Figure 8-1](#) shows the test bed used.

Figure 8-1 Test Bed



153113

Core Switch 1

```

version 12.2
no service pad
service timestamps debug datetime msec localtime
service timestamps log datetime msec localtime
no service password-encryption
service counters max age 10
!
hostname CORE1
!
boot system sup-bootflash:s720_18SXD3.bin

```

```
logging snmp-authfail
enable secret 5 $1$30jN$1/80W4JIQJf7l7fRlS7A2.
!
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
vtp domain datacenter
vtp mode transparent
udld enable
ip subnet-zero
no ip source-route
!
!
no ip ftp passive
no ip domain-lookup
ip domain-name cisco.com
!
no ip bootp server
ip multicast-routing
mls ip cef load-sharing full
!
spanning-tree mode rapid-pvst
spanning-tree loopguard default
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
!
vlan internal allocation policy descending
vlan dot1q tag native
vlan access-log ratelimit 2000
!
vlan 2
!
vlan 15
    name testgear
!
vlan 16
    name testgear2
!
vlan 20
    name DNS-CA
!
vlan 802
    name mgmt_vlan
!
!
interface Loopback0
    ip address 10.10.3.3 255.255.255.0
!
interface Port-channel1
    description to 4948-1 testgear
    no ip address
    logging event link-status
    switchport
    switchport trunk encapsulation dot1q
    switchport trunk native vlan 2
    switchport mode trunk
!
interface Port-channel2
    description to 4948-4 testgear
    no ip address
    logging event link-status
    switchport
    switchport trunk encapsulation dot1q
```

```
switchport trunk native vlan 2
switchport mode trunk
!
interface GigabitEthernet3/33
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode active
!
interface GigabitEthernet3/34
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode active
!
interface GigabitEthernet3/41
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 2 mode active
!
interface GigabitEthernet3/42
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 2 mode active
!
interface TenGigabitEthernet4/1
description to Agg1
ip address 10.10.20.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet4/2
description to Agg2
ip address 10.10.30.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
```



```
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet4/3
description to core2
ip address 10.10.55.1 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface GigabitEthernet6/1
no ip address
shutdown
!
interface GigabitEthernet6/2
*****
!
interface Vlan1
no ip address
shutdown
!
interface Vlan15
description test_client_subnet
ip address 10.20.15.1 255.255.255.0
no ip redirects
no ip proxy-arp
!
interface Vlan16
description test_client_subnet2
ip address 10.20.16.2 255.255.255.0
no ip redirects
no ip proxy-arp
!
router ospf 10
log-adjacency-changes
auto-cost reference-bandwidth 1000000
nsf
area 10 authentication message-digest
area 10 nssa default-information-originate
timers throttle spf 1000 1000 1000
passive-interface default
no passive-interface TenGigabitEthernet4/1
no passive-interface TenGigabitEthernet4/2
no passive-interface TenGigabitEthernet4/3
network 10.10.3.0 0.0.0.255 area 10
network 10.10.20.0 0.0.0.255 area 10
network 10.10.30.0 0.0.0.255 area 10
network 10.10.55.0 0.0.0.255 area 10
network 10.20.15.0 0.0.0.255 area 0
network 10.20.16.0 0.0.0.255 area 0
!
ip classless
no ip http server
ip pim send-rp-discovery scope 2
!
!
```

```
control-plane
!
!
line con 0
  exec-timeout 0 0
line vty 0 4
  exec-timeout 60 0
  password 7 05080F1C2243
  login local
  transport input telnet ssh
!
ntp authentication-key 1 md5 02050D480809 7
ntp trusted-key 1
ntp clock-period 17180053
ntp master 1
ntp update-calendar
end
```

Aggregation Switch 1

```
Current configuration : 22460 bytes
!
! No configuration change since last restart
!
upgrade fpd auto
version 12.2
service timestamps debug datetime msec localtime
service timestamps log datetime msec localtime
no service password-encryption
service counters max age 10
!
hostname Aggregation-1
!
boot system disk0:s720_18SXD3.bin
logging snmp-authfail
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
clock calendar-valid
firewall multiple-vlan-interfaces
firewall module 4 vlan-group 1
firewall vlan-group 1 5-6,20,100,101,105-106
analysis module 9 management-port access-vlan 20
analysis module 9 data-port 1 capture allowed-vlan 5,6,105,106
analysis module 9 data-port 2 capture allowed-vlan 106
ip subnet-zero
no ip source-route
ip icmp rate-limit unreachable 2000
!
!
!
ip multicast-routing
udld enable
udld message time 7

vtp domain datacenter
vtp mode transparent
mls ip cef load-sharing full
mls ip multicast flow-stat-timer 9
no mls flow ip
```

```
no mls flow ipv6
mls acl tcam default-result permit
no mls acl tcam share-global
mls cef error action freeze
!
redundancy
mode sso
main-cpu
  auto-sync running-config
  auto-sync standard
!
spanning-tree mode rapid-pvst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
spanning-tree vlan 1-4094 priority 24576
module ContentSwitchingModule 3
  ft group 1 vlan 102
  priority 20
  heartbeat-time 1
  failover 3
  preempt
!
vlan 44 server
  ip address 10.20.44.42 255.255.255.0
  gateway 10.20.44.1
  alias 10.20.44.44 255.255.255.0
!
probe RHI icmp
  interval 3
  failed 10
!
serverfarm SERVER200
  nat server
  no nat client
  real 10.20.6.56
  inservice
  probe RHI
!
serverfarm SERVER201
  nat server
  no nat client
  real 10.20.6.25
  inservice
  probe RHI
!
vserver SERVER200
  virtual 10.20.6.200 any
  vlan 44
  serverfarm SERVER200
  advertise active
  sticky 10
  replicate csrp sticky
  replicate csrp connection
  persistent rebalance
  inservice
!
vserver SERVER201
  virtual 10.20.6.201 any
  vlan 44
  serverfarm SERVER201
  advertise active
  sticky 10
  replicate csrp sticky
```

```

        replicate csrp connection
        persistent rebalance
        inservice
    !
    port-channel load-balance src-dst-port
    !
    vlan internal allocation policy descending
    vlan dot1q tag native
    vlan access-log ratelimit 2000
    !
    vlan 3
        name AGG1_to_AGG2_L3-OSPF
    !
    vlan 5
    !
    vlan 6
        Webapp Inside
    !
    vlan 7
    !
    vlan 10
        name Database Inside
    !
    vlan 20
    !
    vlan 44
        name CSM_Onearm_Server_VLAN
    !
    vlan 45
        name Service_switch_CSM_Onearm
    !
    vlan 46
        name SERV-CSM2-onearm
    !
    vlan 100
        name AGG_FWSM_failover_interface
    !
    vlan 101
        name AGG_FWSM_failover_state
    !
    vlan 102
        name AGG_CSM_FT_Vlan
    !
    vlan 106
        name WebappOutside
    !
    vlan 110
        name DatabaseOutside
    !
    interface Loopback0
        ip address 10.10.1.1 255.255.255.0
    !
    interface Null0
        no ip unreachable
    !
    interface Port-channel1
        description ETHERCHANNEL_TO_AGG2
        switchport
        switchport trunk encapsulation dot1q
        switchport trunk native vlan 2
        switchport trunk allowed vlan 1-19,21-4094
        switchport mode trunk
        no ip address
        logging event link-status

```

```
    arp timeout 200
    spanning-tree guard loop
!
interface Port-channel10
  description to SERVICE_SWITCH1
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport mode trunk
  no ip address
  logging event link-status
  spanning-tree guard loop

!
interface Port-channel12
  description to SERVICE_SWITCH2
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport mode trunk
  no ip address
  logging event link-status
  spanning-tree guard loop

!
!
interface GigabitEthernet1/13
  description to Service_1
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport mode trunk
  no ip address
  channel-protocol lacp
  channel-group 10 mode active
!
interface GigabitEthernet1/14
  description to Service_1
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport mode trunk
  no ip address
  channel-protocol lacp
  channel-group 10 mode active
!
interface GigabitEthernet1/19
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport trunk allowed vlan 1-5,7-105,107-300,1010-1110
  switchport mode trunk
  no ip address
  channel-protocol lacp
  channel-group 12 mode active
!
!
interface GigabitEthernet5/1
  *****
!
interface GigabitEthernet5/2
  *****
!
interface GigabitEthernet6/1
```

```
no ip address
shutdown
!
interface GigabitEthernet6/2
no ip address
shutdown
media-type rj45
!
interface TenGigabitEthernet7/2
description to Core2
ip address 10.10.40.1 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 7 112A481634424A
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet7/3
description to Core1
ip address 10.10.20.1 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 7 15315A1F277A6A
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet7/4
description TO_ACCESS1
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 105
switchport mode trunk
no ip address
logging event link-status
!
interface TenGigabitEthernet8/1
description TO_AGG2
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 1-19,21-4094
switchport mode trunk
no ip address
logging event link-status
channel-protocol lacp
channel-group 1 mode active
!
interface TenGigabitEthernet8/2
description TO_4948-7
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 106
switchport mode trunk
no ip address
```

```
logging event link-status
spanning-tree guard root
!
interface TenGigabitEthernet8/3
description TO_4948-8
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 106
switchport mode trunk
no ip address
logging event link-status
spanning-tree guard root
!
interface TenGigabitEthernet8/4
description TO_AGG2
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 1-19,21-4094
switchport mode trunk
no ip address
logging event link-status
channel-protocol lacp
channel-group 1 mode active
!
interface Vlan1
no ip address
shutdown
!
interface Vlan3
description AGG1_to_AGG2_L3-RP
bandwidth 10000000
ip address 10.10.110.1 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface Vlan6
description Outside_Webapp_Tier
ip address 10.20.6.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip policy route-map csmpr
ntp disable
standby 1 ip 10.20.6.1
standby 1 timers 1 3
standby 1 priority 120
standby 1 preempt delay minimum 60
!
!
interface Vlan44
description AGG_CSM_Onearm
ip address 10.20.44.2 255.255.255.0
no ip redirects
no ip proxy-arp
standby 1 ip 10.20.44.1
standby 1 timers 1 3
```

```
standby 1 priority 120
standby 1 preempt delay minimum 60
!
router ospf 10
log-adjacency-changes
auto-cost reference-bandwidth 1000000
nsf
area 10 authentication message-digest
area 10 nssa
timers throttle spf 1000 1000 1000
redistribute static subnets route-map rhi
passive-interface default
no passive-interface Vlan3
no passive-interface TenGigabitEthernet7/2
no passive-interface TenGigabitEthernet7/3
network 10.10.1.0 0.0.0.255 area 10
network 10.10.20.0 0.0.0.255 area 10
network 10.10.40.0 0.0.0.255 area 10
network 10.10.110.0 0.0.0.255 area 10
distribute-list 1 in TenGigabitEthernet7/2 (for PBR testing purposes)
distribute-list 1 in TenGigabitEthernet7/3 (for PBR testing purposes)
!
ip classless
ip pim accept-rp auto-rp
!
access-list 1 deny 10.20.16.0
access-list 1 deny 10.20.15.0
access-list 1 permit any
access-list 44 permit 10.20.6.200 log
access-list 44 permit 10.20.6.201 log
!
route-map csmnbr permit 10
set ip default next-hop 10.20.44.44
!
route-map rhi permit 10
match ip address 44
set metric-type type-1
!
privilege exec level 1 show
!
line con 0
exec-timeout 0 0
password 7 110D1A16021F060510
login local
line vty 0 4
no motd-banner
exec-timeout 0 0
password 7 110D1A16021F060510
login local
transport input telnet ssh
!
!
no monitor session servicemodule
ntp authentication-key 1 md5 104D000A0618 7
ntp authenticate
ntp trusted-key 1
ntp clock-period 17179928
ntp update-calendar
ntp server *****.42 key 1
end
```


Core Switch 2

```
Current configuration : 10867 bytes
!
version 12.2
no service pad
service timestamps debug datetime msec localtime
service timestamps log datetime msec localtime
no service password-encryption
service counters max age 10
!
hostname CORE2
!
boot system sup-bootflash:s720_18SXD3.bin
enable secret 5 $1$k2Df$vfht/CMz0IqFqluRCENw//
!
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
vtp domain datacenter
vtp mode transparent
udld enable
!
ip subnet-zero
no ip source-route
!
!
no ip domain-lookup
ip domain-name cisco.com
!
no ip bootp server
ip multicast-routing
mls ip multicast flow-stat-timer 9
no mls flow ip
no mls flow ipv6
mls cef error action freeze
!
power redundancy-mode combined
!
spanning-tree mode rapid-pvst
spanning-tree loopguard default
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
!
vlan internal allocation policy descending
vlan dot1q tag native
vlan access-log ratelimit 2000
!
vlan 2,15-16
!
!
interface Loopback0
 ip address 10.10.4.4 255.255.255.0
!
interface Port-channel1
 description to 4948-1
 no ip address
 logging event link-status
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 2
 switchport mode trunk
```

```
!  
interface Port-channel2  
  description to 4948-4  
  no ip address  
  logging event link-status  
  switchport  
  switchport trunk encapsulation dot1q  
  switchport trunk native vlan 2  
  switchport mode trunk  
!  
interface GigabitEthernet2/9  
  no ip address  
  logging event link-status  
  switchport  
  switchport trunk encapsulation dot1q  
  switchport trunk native vlan 2  
  switchport mode trunk  
  channel-protocol lacp  
  channel-group 1 mode active  
!  
interface GigabitEthernet2/10  
  no ip address  
  logging event link-status  
  switchport  
  switchport trunk encapsulation dot1q  
  switchport trunk native vlan 2  
  switchport mode trunk  
  channel-protocol lacp  
  channel-group 1 mode active  
!  
interface GigabitEthernet2/13  
  no ip address  
  logging event link-status  
  switchport  
  switchport trunk encapsulation dot1q  
  switchport trunk native vlan 2  
  switchport mode trunk  
  channel-protocol lacp  
  channel-group 2 mode active  
!  
interface GigabitEthernet2/14  
  no ip address  
  logging event link-status  
  switchport  
  switchport trunk encapsulation dot1q  
  switchport trunk native vlan 2  
  switchport mode trunk  
  channel-protocol lacp  
  channel-group 2 mode active  
!  
interface TenGigabitEthernet4/1  
  description to Agg1  
  ip address 10.10.40.2 255.255.255.0  
  no ip redirects  
  no ip proxy-arp  
  ip pim sparse-dense-mode  
  ip ospf authentication message-digest  
  ip ospf message-digest-key 1 md5 C1sC0!  
  ip ospf network point-to-point  
  ip ospf hello-interval 2  
  ip ospf dead-interval 6  
  logging event link-status  
!  
interface TenGigabitEthernet4/2
```

```
description to Agg2
ip address 10.10.50.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet4/3
description to core1
ip address 10.10.55.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface GigabitEthernet6/1
no ip address
shutdown
!
interface GigabitEthernet6/2
*****
!
interface Vlan1
no ip address
shutdown
!
interface Vlan15
ip address 10.20.15.2 255.255.255.0
!
interface Vlan16
description test_client_subnet
ip address 10.20.16.1 255.255.255.0
no ip redirects
no ip proxy-arp
!
router ospf 10
log-adjacency-changes
auto-cost reference-bandwidth 1000000
nsf
area 10 authentication message-digest
area 10 nssa default-information-originate
timers throttle spf 1000 1000 1000
passive-interface default
no passive-interface TenGigabitEthernet4/1
no passive-interface TenGigabitEthernet4/2
no passive-interface TenGigabitEthernet4/3
no passive-interface TenGigabitEthernet4/4
network 10.10.4.0 0.0.0.255 area 10
network 10.10.40.0 0.0.0.255 area 10
network 10.10.50.0 0.0.0.255 area 10
network 10.10.55.0 0.0.0.255 area 10
network 10.20.15.0 0.0.0.255 area 0
network 10.20.16.0 0.0.0.255 area 0
!
```

```
ip classless
no ip http server
ip pim send-rp-discovery scope 2
!
!
line con 0
  exec-timeout 0 0
line vty 0 4
  exec-timeout 60 0
  password cisco
  login local
  transport input telnet ssh
!
ntp authentication-key 1 md5 104D000A0618 7
ntp authenticate
ntp trusted-key 1
ntp clock-period 17179940
ntp update-calendar
ntp server ***** key 1
end
```

Aggregation Switch 2

```
Current configuration : 18200 bytes
version 12.2
service timestamps debug datetime msec localtime
service timestamps log datetime msec
no service password-encryption
service counters max age 10
!
hostname Aggregation-2
!
boot system disk0:s720_18SXD3.bin
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
clock calendar-valid
firewall multiple-vlan-interfaces
firewall module 4 vlan-group 1
firewall vlan-group 1 5,6,20,100,101,105,106
vtp domain datacenter
vtp mode transparent
udld enable
!
udld message time 7
!
ip subnet-zero
no ip source-route
ip icmp rate-limit unreachable 2000
!
!
ip multicast-routing
no ip igmp snooping
mls ip cef load-sharing full
mls ip multicast flow-stat-timer 9
no mls flow ip
no mls flow ipv6
mls acl tcam default-result permit
mls cef error action freeze
!
```

```
!
spanning-tree mode rapid-pvst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
spanning-tree vlan 1-4094 priority 28672
port-channel load-balance src-dst-port
module ContentSwitchingModule 3
  ft group 1 vlan 102
  priority 10
  heartbeat-time 1
  failover 3
  preempt
!
vlan 44 server
  ip address 10.20.44.43 255.255.255.0
  gateway 10.20.44.1
  alias 10.20.44.44 255.255.255.0
!
probe RHI icmp
  interval 3
  failed 10
!
serverfarm SERVER200
  nat server
  no nat client
  real 10.20.6.56
  inservice
  probe RHI
!
serverfarm SERVER201
  nat server
  no nat client
  real 10.20.6.25
  inservice
  probe RHI
!
vserver SERVER200
  virtual 10.20.6.200 any
  vlan 44
  serverfarm SERVER200
  advertise active
  sticky 10
  replicate csrp sticky
  replicate csrp connection
  persistent rebalance
  inservice
!
vserver SERVER201
  virtual 10.20.6.201 any
  vlan 44
  serverfarm SERVER201
  advertise active
  sticky 10
  replicate csrp sticky
  replicate csrp connection
  persistent rebalance
  inservice
!
!
vlan internal allocation policy descending
vlan dot1q tag native
vlan access-log ratelimit 2000
!
```

```

vlan 3
  name AGG1_to_AGG2_L3-RP
  !
vlan 5
  name Outside_Webapp
  !
vlan 6
  name Outside_Webapp
  !
  !
vlan 10
  name Outside_Database_Tier
  !
vlan 20
  !
vlan 44
  name AGG_CSM_Onearm
  !
vlan 45
  name Service_switch_CSM_Onearm
  !
vlan 46
  name SERV-CSM2-onearm
  !
vlan 100
  name AGG_FWSM_failover_interface
  !
vlan 101
  name AGG_FWSM_failover_state
  !
vlan 102
  name AGG_CSM_FT_Vlan
  !
vlan 105
  name Inside_Webapp_Tier
  !
vlan 106
  name Inside_Webapp
  !
vlan 110
  name Inside_Database_Tier
  !
  !
interface Loopback0
  ip address 10.10.2.2 255.255.255.0
  !
interface Null0
  no ip unreachable
  !
interface Port-channel1
  description ETHERCHANNEL_TO_AGG1
  no ip address
  logging event link-status
  switchport
  switchport trunk encapsulation dot1q
  switchport trunk native vlan 2
  switchport trunk allowed vlan 1-19,21-299,301-4094
  switchport mode trunk
  arp timeout 200
  spanning-tree guard loop
  !
interface Port-channel11
  description to SERVICE_SWITCH1
  no ip address

```

```
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
!
interface Port-channel13
description to SERVICE_SWITCH2
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
!
interface GigabitEthernet1/13
description to Service_2
no ip address
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 13 mode active
!
interface GigabitEthernet1/14
description to Service_2
no ip address
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 13 mode active
!
interface GigabitEthernet1/19
description to Service_1
no ip address
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 11 mode active
!
interface GigabitEthernet1/20
description to Service_1
no ip address
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 11 mode active
!
interface GigabitEthernet5/1
!
interface GigabitEthernet5/2
*****
!
interface TenGigabitEthernet7/2
description to Core2
ip address 10.10.50.1 255.255.255.0
no ip redirects
```

```
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet7/3
description to Core1
ip address 10.10.30.1 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface TenGigabitEthernet7/4
description TO_ACCESS1
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 5,6
switchport mode trunk
channel-protocol lacp
!
interface TenGigabitEthernet8/1
description TO_AGG1
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 1-19,21-299,301-4094
switchport mode trunk
channel-protocol lacp
channel-group 1 mode passive
!
!
interface TenGigabitEthernet8/3
description TO_4948-8
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport trunk allowed vlan 106
switchport mode trunk
spanning-tree guard root
!
interface TenGigabitEthernet8/4
description TO_AGG1
no ip address
logging event link-status
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
```



```
switchport trunk allowed vlan 1-19,21-299,301-4094
switchport mode trunk
channel-protocol lacp
channel-group 1 mode passive
!
interface Vlan1
no ip address
shutdown
!
interface Vlan3
description AGG1_to_AGG2_L3-RP
bandwidth 10000000
ip address 10.10.110.2 255.255.255.0
no ip redirects
no ip proxy-arp
ip pim sparse-dense-mode
ip ospf authentication message-digest
ip ospf message-digest-key 1 md5 C1sC0!
ip ospf network point-to-point
ip ospf hello-interval 2
ip ospf dead-interval 6
logging event link-status
!
interface Vlan5
description Outside_Webapp_Tier
no ip address
no ip redirects
ntp disable
standby 1 ip 10.20.5.1
standby 1 timers 1 3
standby 1 priority 115
standby 1 preempt delay minimum 60
!
interface Vlan6
ip address 10.20.6.3 255.255.255.0
no ip redirects
no ip proxy-arp
ip policy route-map csmpr
ntp disable
standby 1 ip 10.20.6.1
standby 1 timers 1 3
standby 1 priority 115
standby 1 preempt delay minimum 60
!
interface Vlan44
description AGG_CSM_Onearm
ip address 10.20.44.3 255.255.255.0
no ip redirects
no ip proxy-arp
standby 1 ip 10.20.44.1
standby 1 timers 1 3
standby 1 priority 115
standby 1 preempt delay minimum 60
!
!
router ospf 10
log-adjacency-changes
auto-cost reference-bandwidth 1000000
nsf
area 10 authentication message-digest
area 10 nssa
timers throttle spf 1000 1000 1000
redistribute static subnets route-map rhi
passive-interface default
```

```

no passive-interface Vlan3
no passive-interface TenGigabitEthernet7/2
no passive-interface TenGigabitEthernet7/3
network 10.10.2.0 0.0.0.255 area 10
network 10.10.30.0 0.0.0.255 area 10
network 10.10.50.0 0.0.0.255 area 10
network 10.10.110.0 0.0.0.255 area 10
distribute-list 1 in TenGigabitEthernet7/2
distribute-list 1 in TenGigabitEthernet7/3
!
ip classless
ip pim accept-rp auto-rp
!
access-list 1 deny 10.20.16.0
access-list 1 deny 10.20.15.0
access-list 1 permit any
access-list 44 permit 10.20.6.200 log
access-list 44 permit 10.20.6.201 log
!
route-map csmpr permit 10
 set ip default next-hop 10.20.44.44
!
route-map rhi permit 10
 match ip address 44
 set metric +40
 set metric-type type-1
!
line con 0
 exec-timeout 0 0
 password dcsummit
 login local
line vty 0 4
 exec-timeout 0 0
 password dcsummit
 login local
 transport input telnet ssh
 transport output pad telnet ssh acercon
!
no monitor session servicemodule
ntp authentication-key 1 md5 08701C1A2D495547335B5A5572 7
ntp authenticate
ntp clock-period 17179998
ntp update-calendar
ntp server *****key 1
end

```

Access Switch 4948-7

```

Current configuration : 4612 bytes
version 12.2
no service pad
service timestamps debug datetime localtime
service timestamps log datetime localtime
no service password-encryption
service compress-config
!
hostname 4948-7
!
boot-start-marker
boot system bootflash:cat4000-i5k91s-mz.122-25.EWA2.bin
boot-end-marker

```

```
!
logging snmp-authfail
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
clock calendar-valid
vtp domain datacenter
vtp mode transparent
udld enable

ip subnet-zero
no ip source-route
no ip domain-lookup
ip domain-name cisco.com
!
!
spanning-tree mode rapid-pvst
spanning-tree loopguard default
spanning-tree portfast bpduguard default
spanning-tree extend system-id
spanning-tree pathcost method long
port-channel load-balance src-dst-port
power redundancy-mode redundant
!
!
!
vlan internal allocation policy descending
vlan dot1q tag native
!
vlan 5-6
!
vlan 105
    name Outside_Webapp
!
vlan 106
    name Outside_Webapp
!
vlan 110
    name Outside_Database_Tier
!
interface Port-channel1
    description inter_4948
    switchport
    switchport trunk encapsulation dot1q
    switchport trunk native vlan 2
    switchport mode trunk
    logging event link-status
!
interface GigabitEthernet1/1    (all ports)
    switchport access vlan 106
    switchport mode access
    no cdp enable
    spanning-tree portfast
!
interface GigabitEthernet1/45
    description to 4948-8
    switchport trunk encapsulation dot1q
    switchport trunk native vlan 2
    switchport mode trunk
    channel-protocol lacp
    channel-group 1 mode active
!
interface GigabitEthernet1/46
    switchport trunk encapsulation dot1q
```

```
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode active
!
interface GigabitEthernet1/47
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode active
!
interface GigabitEthernet1/48
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode active
!
interface TenGigabitEthernet1/49
description to_AGG1
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
!
interface TenGigabitEthernet1/50
shutdown
!
interface Vlan1
no ip address
shutdown
!
!
line con 0
exec-timeout 0 0
stopbits 1
line vty 0 4
exec-timeout 0 0
password dcsummit
login local
!
ntp authenticate
ntp trusted-key 1
ntp update-calendar
ntp server ***** key 1
!
end
```

Access Switch 4948-8

```
Current configuration : 4646 bytes
!
version 12.2
no service pad
service timestamps debug datetime localtime
service timestamps log datetime localtime
no service password-encryption
service compress-config
!
hostname 4948-8
!
```

```
boot-start-marker
boot system bootflash:cat4000-i5k91s-mz.122-25.EWA2.bin
boot-end-marker
!
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
clock calendar-valid
vtp domain datacenter
vtp mode transparent
udld enable
!
ip subnet-zero
no ip source-route
no ip domain-lookup
ip domain-name cisco.com
!
no ip bootp server
!
no file verify auto
!
spanning-tree mode rapid-pvst
spanning-tree loopguard default
spanning-tree portfast bpduguard default
spanning-tree extend system-id
spanning-tree pathcost method long
port-channel load-balance src-dst-port
power redundancy-mode redundant
!
!
!
vlan internal allocation policy descending
vlan dot1q tag native
!
vlan 2,5-6
!
vlan 105
    name Outside_Webapp_Tier
!
vlan 106
    name Outside_Webapp_Tier
!
vlan 110
    name Outside_Database_Tier
!
interface Port-channel1
    description inter_4948
    switchport
    switchport trunk encapsulation dot1q
    switchport trunk native vlan 2
    switchport mode trunk
    logging event link-status
!
interface GigabitEthernet1/1 (all ports)
    switchport access vlan 106
    switchport trunk encapsulation dot1q
    switchport mode access
    no cdp enable
    spanning-tree portfast
!
interface GigabitEthernet1/45
    switchport trunk encapsulation dot1q
    switchport trunk native vlan 2
    switchport mode trunk
```

```

channel-protocol lacp
channel-group 1 mode passive
!
interface GigabitEthernet1/46
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode passive
!
interface GigabitEthernet1/47
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode passive
!
interface GigabitEthernet1/48
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
channel-protocol lacp
channel-group 1 mode passive
!
interface TenGigabitEthernet1/49
shutdown
!
interface TenGigabitEthernet1/50
description to_AGG2
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
!
interface Vlan1
no ip address
shutdown
!
!
line con 0
exec-timeout 0 0
stopbits 1
line vty 0 4
exec-timeout 0 0
password dcsummit
login local
!
ntp authenticate
ntp trusted-key 1
ntp update-calendar
ntp server ***** key 1
!
end

```

Access Switch 6500-1

```

ACCESS1-6500#
Building configuration...

Current configuration : 11074 bytes
!
! Last configuration change at 13:33:08 PST Thu Feb 9 2006

```

```
! NVRAM config last updated at 16:58:39 PST Thu Nov 17 2005
!
upgrade fpd auto
version 12.2
no service pad
service timestamps debug datetime localtime
service timestamps log datetime localtime
service password-encryption
service counters max age 10
!
hostname ACCESS1-6500
!
boot system sup-bootflash:s720_18SXD3.bin
no aaa new-model
clock timezone PST -8
clock summer-time PDT recurring
clock calendar-valid
ip subnet-zero
no ip source-route
!
!
!
no ip bootp server
ip domain-list cisco.com
no ip domain-lookup
ip domain-name cisco.com
udld enable
!
udld message time 7
!
vtp domain datacenter
vtp mode transparent
no mls acl tcam share-global
mls cef error action freeze
!
spanning-tree mode rapid-pvst
spanning-tree loopguard default
spanning-tree portfast bpduguard default
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
!
power redundancy-mode combined
no diagnostic cns publish
no diagnostic cns subscribe
fabric buffer-reserve queue
port-channel load-balance src-dst-port
!
vlan internal allocation policy descending
vlan dot1q tag native
vlan access-log ratelimit 2000
!
vlan 5
    name Outside_Webapp_Tier
!
vlan 105
    name Outside_Webapp_Tier
!
vlan 110
    name Outside_Database_Tier
!
interface TenGigabitEthernet1/1
    description to_AGG1
    switchport
```

```

switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
no ip address
logging event link-status
!
interface TenGigabitEthernet1/2
description to_AGG2
switchport
switchport trunk encapsulation dot1q
switchport trunk native vlan 2
switchport mode trunk
no ip address
logging event link-status
logging event spanning-tree status
!!
interface GigabitEthernet2/1 (all test ports)
description webapp_penguin_kvm5
switchport
switchport access vlan 5
switchport mode access
no ip address
no cdp enable
spanning-tree portfast
!
!
interface Vlan1
no ip address
shutdown
!
no ip http server
!
line con 0
exec-timeout 0 0
line vty 0 4
exec-timeout 0 0
password 7 05080F1C2243
login local
transport input telnet ssh
!
no monitor event-trace timestamps
ntp authentication-key 1 md5 110A1016141D 7
ntp authenticate
ntp trusted-key 1
ntp clock-period 17179938
ntp update-calendar
ntp server *****key 1
no cns aaa enable
end

```

FWSM 1-Aggregation Switch 1 and 2

```

FWSM Version 2.3(2) <system>
firewall transparent
resource acl-partition 12
enable password 2KFQnbNIdI.2KYOU encrypted
passwd 2KFQnbNIdI.2KYOU encrypted
hostname FWSM1-AGG1and2
ftp mode passive
pager lines 24
logging buffer-size 4096

```



```

logging console debugging
class default
  limit-resource PDM 5
  limit-resource All 0
  limit-resource IPSec 5
  limit-resource Mac-addresses 65535
  limit-resource SSH 5
  limit-resource Telnet 5
!

failover
failover lan unit primary
failover lan interface failover vlan 100
failover polltime unit msec 500 holdtime 3
failover polltime interface 3
failover interface-policy 100%
failover replication http
failover link state vlan 101
failover interface ip failover 10.20.100.1 255.255.255.0 standby 10.20.100.2
failover interface ip state 10.20.101.1 255.255.255.0 standby 10.20.101.2
arp timeout 14400

!

timeout xlate 3:00:00
timeout conn 1:00:00 half-closed 0:10:00 udp 0:02:00 icmp 0:00:02 rpc 0:10:00 h323 0:05:00
  h225 1:00:00 mgcp 0:05:00 sip 0:30:00 sip_media 0:02:00
timeout uauth 0:05:00 absolute
sysopt nodnsalias inbound
sysopt nodnsalias outbound
terminal width 511

admin-context admin
context admin
  allocate-interface vlan20 outside
  config-url disk:/admin.cfg
!

context vlan6-106
  description vlan6-106 context
  allocate-interface vlan6 outside
  allocate-interface vlan106 inside
  config-url disk:/vlan6-106.cfg
!

Cryptochecksum:a73fe039e4dbeb45a9c6730bc2a55201
: end
[OK]

FWSM1-AGG1and2# ch co vlan6-106
FWSM1-AGG1and2/vlan6-106# wr t
Building configuration...
: Saved
:
FWSM Version 2.3(2) <context>
firewall transparent
nameif outside vlan6 security0
nameif inside vlan106 security100
enable password 8Ry2YjIyt7RRXU24 encrypted
passwd 2KFQnbNIdI.2KYOU encrypted
hostname vlan6-106
fixup protocol dns maximum-length 512

```

```

fixup protocol ftp 21
fixup protocol h323 H225 1720
fixup protocol h323 ras 1718-1719
fixup protocol rsh 514
fixup protocol sip 5060
no fixup protocol sip udp 5060
fixup protocol skinny 2000
fixup protocol smtp 25
fixup protocol sqlnet 1521
names
access-list deny-flow-max 4096
access-list alert-interval 300
access-list IP extended permit ip any any
access-list IP extended permit icmp any any
access-list BPDU ethertype permit bpdv
pager lines 24
logging on
logging timestamp
logging buffer-size 4096
logging trap informational
logging device-id hostname
mtu vlan6 1500
mtu vlan106 1500
ip address 10.20.6.104 255.255.255.0 standby 10.20.6.105
icmp permit any vlan6
icmp permit any vlan106
no pdm history enable
arp timeout 14400
access-group BPDV in interface vlan6
access-group IP in interface vlan6
access-group BPDV in interface vlan106
access-group IP in interface vlan106
!
interface vlan6
!
!
interface vlan106
!

!

route vlan6 0.0.0.0 0.0.0.0 10.20.6.1 1
timeout xlate 3:00:00
timeout conn 1:00:00 half-closed 0:10:00 udp 0:02:00 icmp 0:00:02 rpc 0:10:00 h323 0:05:00
h225 1:00:00 mgcp 0:05:00 sip 0:30:00 sip_media 0:02:00
timeout uauth 0:05:00 absolute
aaa-server TACACS+ protocol tacacs+
aaa-server TACACS+ max-failed-attempts 3
aaa-server TACACS+ deadtime 10
aaa-server RADIUS protocol radius
aaa-server RADIUS max-failed-attempts 3
aaa-server RADIUS deadtime 10
aaa-server LOCAL protocol local
no snmp-server location
no snmp-server contact
snmp-server community public
snmp-server enable traps snmp
floodguard enable
fragment size 200 vlan6
fragment chain 24 vlan6
fragment size 200 vlan106
fragment chain 24 vlan106

```

```

telnet timeout 5
ssh 0.0.0.0 0.0.0.0 vlan6
ssh timeout 60
terminal width 511
Cryptochecksum:00000000000000000000000000000000
: end
[OK]

FWSM1-AGG1and2/vlan6-106# ch co admin
FWSM1-AGG1and2/admin# wr t
Building configuration...
: Saved
:
FWSM Version 2.3(2) <context>
firewall transparent
nameif outside vlan20 security0
enable password 8Ry2YjIyt7RRXU24 encrypted
passwd 2KFQnbNIdI.2KYOU encrypted
hostname admin
domain-name example.com
fixup protocol dns maximum-length 512
fixup protocol ftp 21
fixup protocol h323 H225 1720
fixup protocol h323 ras 1718-1719
fixup protocol rsh 514
fixup protocol sip 5060
fixup protocol sip udp 5060
fixup protocol skinny 2000
fixup protocol smtp 25
fixup protocol sqlnet 1521
names
access-list deny-flow-max 4096
access-list alert-interval 300
access-list IP extended permit ip any any
access-list IP extended permit icmp any any
access-list IP extended permit udp any any
access-list BPDU ethertype permit bpdu
pager lines 24
logging on
logging timestamp
logging buffer-size 4096
logging trap informational
logging device-id hostname
mtu vlan20 1500
ip address *****.34 255.255.255.0 standby *****.35
icmp permit any vlan20
no pdm history enable
arp timeout 14400
access-group IP in interface vlan20
!
interface vlan20
!

!

route vlan20 0.0.0.0 0.0.0.0 *****.1 1
timeout xlate 3:00:00
timeout conn 1:00:00 half-closed 0:10:00 udp 0:02:00 icmp 0:00:02 rpc 0:10:00 h323 0:05:00
h225 1:00:00 mgcp 0:05:00 sip 0:30:00 sip_media 0:02:00
timeout uauth 0:05:00 absolute
username mshinn password fgXai3fBCmTT1r2e encrypted privilege 15
aaa-server TACACS+ protocol tacacs+

```

```
aaa-server TACACS+ max-failed-attempts 3
aaa-server TACACS+ deadtime 10
aaa-server RADIUS protocol radius
aaa-server RADIUS max-failed-attempts 3
aaa-server RADIUS deadtime 10
aaa-server LOCAL protocol local
http server enable
http 0.0.0.0 0.0.0.0 vlan20
no snmp-server location
no snmp-server contact
snmp-server community public
snmp-server enable traps snmp
floodguard enable
fragment size 200 vlan20
fragment chain 24 vlan20
sysopt nodnsalias inbound
sysopt nodnsalias outbound
telnet timeout 5
ssh 0.0.0.
```

Additional References

See the following URL for more information:

- Cisco Catalyst 6500—<http://www.cisco.com/en/US/products/hw/switches/ps708/index.html>