CISCO SYSTEMS

Data Center Blade Server Integration 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

CONTENTS

CHAPTER 3   **Pass-Through Technology**   3-1

CHAPTER 4   **Blade Server Integration into the Data Center with Intelligent Network Services**   4-1

Preface

Document Purpose

The data center is the repository for applications and data critical to the modern enterprise. The enterprise demands on the data center are increasing, requiring the capacity and flexibility to address a fluid business environment whilst reducing operational costs. Data center expenses such as power, cooling, and space have become more of a concern as the data center grows to address business requirements.

Blade servers are the latest server platforms that attempt to address these business drivers. Blade servers consolidate compute power and suggest that the data center bottom line will benefit from savings related to the following:

- Power
- Cooling
- Physical space
- Management
- Server provisioning
- Connectivity (server I/O)

This document explores the integration of blade servers into a Cisco data center multi-tier architecture.

Intended Audience

This guide is intended for system engineers who support enterprise customers that are responsible for designing, planning, managing, and implementing local and distributed data center IP infrastructures.

Document Organization

This guide contains the chapters in the following table.

Section	Description
Chapter 1, "Blade Servers in the Data Center—Overview."	Provides high-level overview of the use of blade servers in the data center.

Chapter 2, "Integrated Switch Technology."	Provides best design practices for deploying Cisco Intelligent Gigabit Ethernet Switch Modules (Cisco IGESM) for the IBM eServer BladeCenter (BladeCenter) within the Cisco Data Center Networking Architecture.
Chapter 3, "Pass-Through Technology."	Provides best design practices for deploying blade servers using pass-through technology within the Cisco Data Center Networking Architecture.
Chapter 4, "Blade Server Integration into the Data Center with Intelligent Network Services."	Discusses the integration of intelligent services into the Cisco Data Center Architecture that uses blade server systems.

C H A P T E R 1

Blade Servers in the Data Center—Overview

Data Center Multi-tier Model Overview

The data center multi-tier model is a common enterprise design that defines logical tiers addressing web, application, and database functionality. The multi-tier model uses network services to provide application optimization and security.

Figure 1-1 shows a generic multi-tier data center architecture.

Figure 1-1   *Data Center Multi-tier Model*
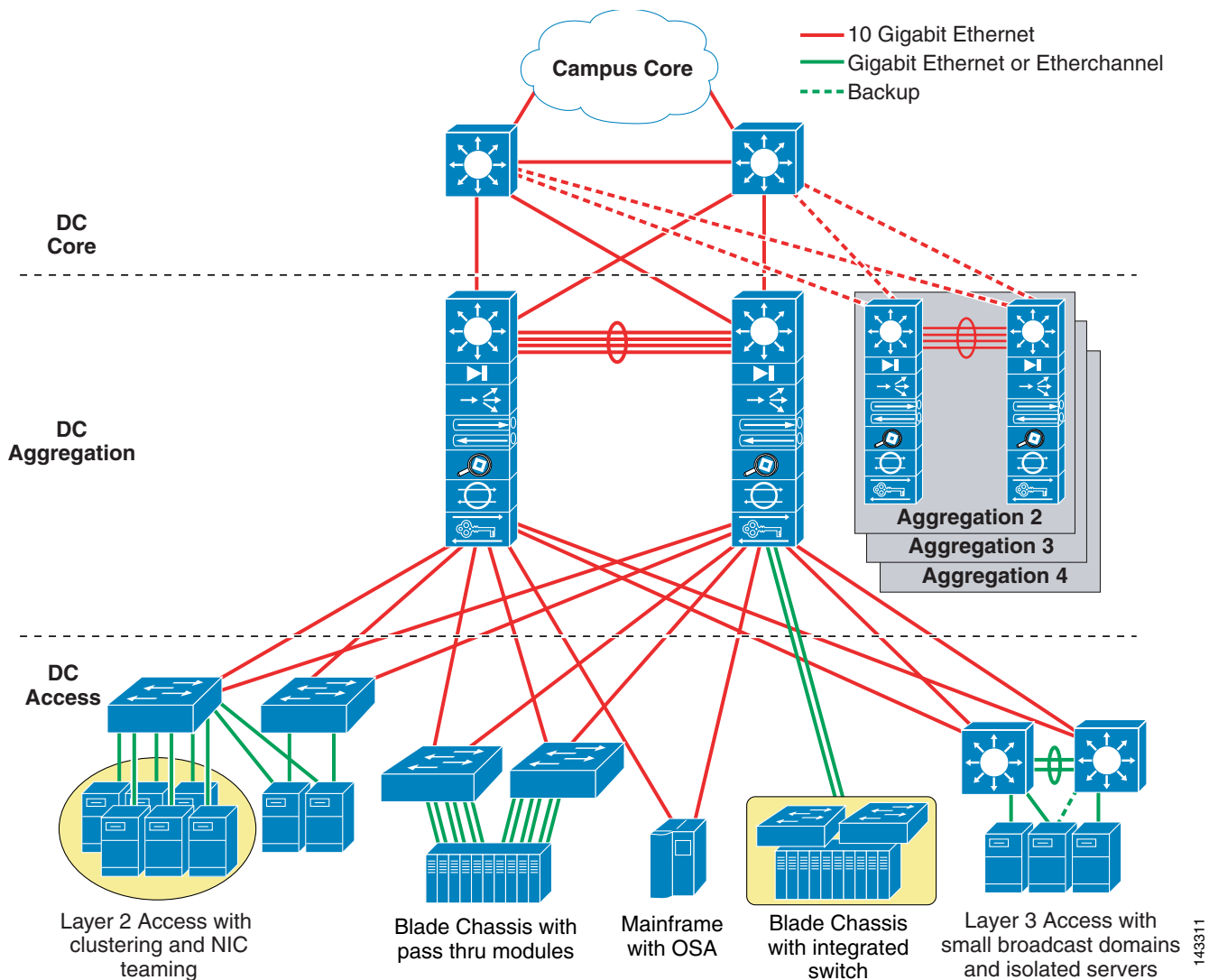
The layers of the data center design are the *core, aggregation,* and *access* layers. These layers are referred to throughout this SRND and are briefly described as follows:

- 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—Provides 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 may 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, which provide services that include content switching, firewall, SSL offload, intrusion detection, and network analysis.

- Access layer—Location 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.

The multi-tier data center is a flexible, robust environment capable of providing high availability, scalability, and critical network services to data center applications with diverse requirements and physical platforms. This document focuses on the integration of blade servers into the multi-tier data center model. For more details on the Cisco Data Center infrastructure, see the *Data Center Infrastructure SRND 2.5* at the following URL:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

Blade Server Integration Options

Blade systems are the latest server platform emerging in the data center. Enterprise data centers seek the benefits that this new platform can provide in terms of power, cooling, and server consolidation that optimize the compute power per rack unit. Consequently, successfully incorporating these devices into the data center network architecture becomes a key consideration for network administrators.

The following section is an overview of the options available to integrate blade systems into the data center. The following topics are included:

- Integrated Switches
- Pass-Through Technology

Integrated Switches

Blade systems allow built-in switches to control traffic flow between the blade servers within the chassis and the remaining enterprise network. Blade systems provide a variety of switch media types, including the following:

- Built-in Ethernet switches (such as the Cisco Ethernet Switch Modules)
- Infiniband switches (such as the Cisco Server Fabric Switch)
- Fibre Channel switches

Integrated switches are a passageway to the blade servers within the chassis and to the data center. As illustrated in Figure 1-2, each blade server connects to a backplane or a mid-plane that typically contains four dedicated signaling paths to redundant network devices housed in the chassis. This predefined physical structure reduces the number of cables required by each server and provides a level of resiliency via the physical redundancy of the network interface controllers (NICs) and I/O network devices.

> **Note**   The predefined connectivity of a blade system has NIC teaming implications. Therefore, network administrators must consider this when determining their blade server high availability strategy.

Figure 1-2   Sample Blade System Internal Connection

Example of a Blade System Backplane
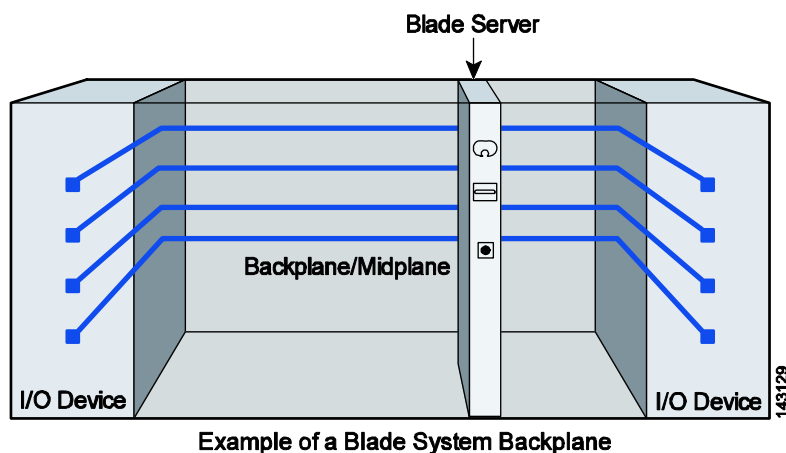
Note   The chassis illustrated in Figure 1-2 is for demonstration purposes. Chassis details differ between blade system vendors.

Introducing a blade server system that uses built-in Ethernet switches into the IP infrastructure of the data center presents many options to the network administrator, such as the following:

- Where is the most appropriate attachment point—the aggregation or access layer?
- What features are available on the switch, such as Layer 2 or trunk failover?
- What will the impact be to the Layer 2 and Layer 3 topologies?
- Will NIC teaming play a role in the high availability design?
- What will the management network look like?

These topics are addressed in Chapter 2, "Integrated Switch Technology."

Pass-Through Technology

Pass-through technology is an alternative method of network connectivity that allows individual blade servers to communicate directly with external resources. Both copper and optical pass-through modules that provide access to the blade server controllers are available.

Figure 1-3 shows two common types of pass-through I/O devices. Each of these provides connectivity to the blade servers via the backplane or mid-plane of the chassis. There is a one-to-one relationship between the number of server interfaces and the number of external ports in the access layer that are necessary to support the blade system. Using an octopus cable changes the one-to-one ratio, as shown by the lower pass-through module in Figure 1-3.

Figure 1-3   Pass-Through Module Examples
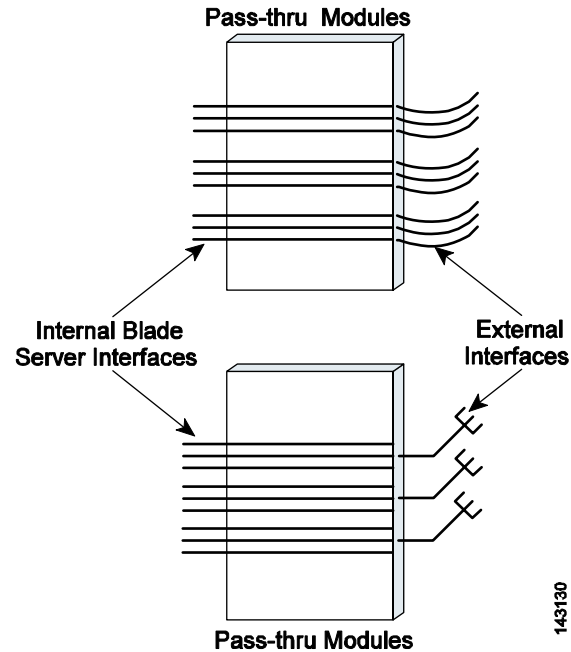
Pass-through modules are passive devices that simply expose the blade server NIC to the external network. They do not require configuration by the network administrator. These I/O devices do not require configuration and do not extend the network Layer 2 or Layer 3 topologies. In addition, the blade servers may employ any of the NIC teaming configurations supported by their drivers.

The need to reduce the amount of cabling in the data center is a major influence driving the rapid adoption of blade servers. Pass-through modules do not allow the data center to take full advantage of the cable consolidation the blade platform offers. This lack of cable reduction in the rack, row, or facility often hinders the use of a pass-through based solution in the data center.

Pass-through technology issues are addressed in Chapter 3, "Pass-Through Technology."

2

Integrated Switch Technology

This section discusses the following topics:

- Cisco Intelligent Gigabit Ethernet Switch Module for the IBM BladeCenter
- Cisco Gigabit Ethernet Switch Module for the HP BladeSystem

Cisco Intelligent Gigabit Ethernet Switch Module for the IBM BladeCenter

This section provides best design practices for deploying Cisco Intelligent Gigabit Ethernet Switch Modules (Cisco IGESMs) for the IBM eServer BladeCenter (BladeCenter) within the Cisco Data Center Networking Architecture. This section describes the internal structures of the BladeCenter and the Cisco IEGSM and explores various methods of deployment. It includes the following sections:

- Cisco Intelligent Gigabit Ethernet Switching Module
- Cisco IGESM Features
- Using the IBM BladeCenter in the Data Center Architecture
- Design and Implementation Details

Cisco Intelligent Gigabit Ethernet Switching Module

This section briefly describes the Cisco IGESM and explains how the blade servers within the BladeCenter chassis are physically connected to it.

The Cisco IGESM integrates the Cisco industry-leading Ethernet switching technology into the IBM BladeCenter. For high availability and multi-homing, each IBM BladeCenter can be configured to concurrently support two pairs of Cisco IGESMs. The Cisco IGESM provides a broad range of Layer 2 switching features, while providing a seamless interface to SNMP-based management tools, such as CiscoWorks. The following switching features supported on the Cisco IGESM help provide this seamless integration into the data center network:

- Loop protection and rapid convergence with support for Per VLAN Spanning Tree (PVST+), 802.1w, 802.1s, BDPU Guard, Loop Guard, PortFast and UniDirectional Link Detection (UDLD)
- Advanced management protocols, including Cisco Discovery Protocol, VLAN Trunking Protocol (VTP), and Dynamic Trunking Protocol (DTP)

- Port Aggregation Protocol (PAgP) and Link Aggregation Control Protocol (LACP), for link load balancing and high availability

- Support for authentication services, including RADIUS and TACACS+

- Support for protection mechanisms, such as limiting the number of MAC addresses allowed, or shutting down the port in response to security violations

Each Cisco IGESM provides Gigabit Ethernet connectivity to each of the 14 blade slots in the BladeCenter and supplies four external Gigabit Ethernet uplink interfaces. You may install from one to four Cisco IGESMs in each BladeCenter. Figure 2-1 illustrates how the BladeCenter chassis provides Ethernet connectivity.

Figure 2-1   BladeCenter Architecture for Ethernet Connectivity
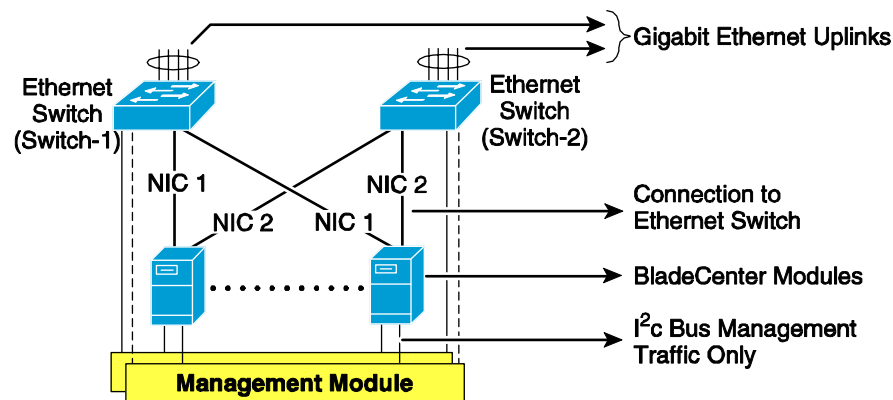
In Figure 2-1, two Ethernet switches within the BladeCenter chassis connect the blade server modules to external devices. Each Ethernet switch provides four Gigabit Ethernet links for connecting the BladeCenter to the external network. The uplink ports can be grouped to support the 802.3ad link aggregation protocol. In the illustrated example, each blade server is connected to the available Gigabit Ethernet network interface cards (NICs). NIC 1 on each blade server is connected to Cisco IGESM 1, while NIC 2 is connected to Cisco IGESM 2. The links connecting the blade server to the Cisco IGESM switches are provided by the BladeCenter chassis backplane.

Figure 2-2 provides a simplified logical view of the blade server architecture for data traffic. The dotted line between the two Cisco IGESMs shows the connectivity provided by the BladeCenter Management Module, which bridges traffic.

Figure 2-2   Logical View of BladeCenter Chassis Architecture
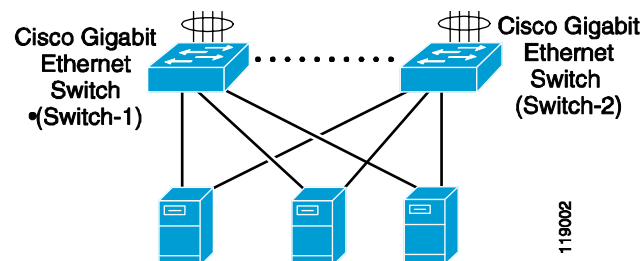
Cisco IGESM Features

This section highlights information about protocols and features provided by Cisco IGESM that help integrate the BladeCenter into the Cisco Data Center Network Architecture and the IBM On-Demand Operating environment. This section includes the following topics:

- Spanning Tree
- Traffic Monitoring
- Link Aggregation Protocols
- Layer 2 Trunk Failover

Spanning Tree

The Cisco IGESM supports various versions of the Spanning Tree Protocol (STP) and associated features, including the following:

- 802.1w
- 802.1s
- Rapid Per VLAN Spanning Tree Plus (RPVST+)
- Loop Guard
- Unidirectional Link Detection (UDLD)
- BPDU Guard

The 802.1w protocol is the standard for rapid spanning tree convergence, while 802.1s is the standard for multiple spanning tree instances. Support for these protocols is essential in a server farm environment for allowing rapid Layer 2 convergence after a failure in the primary path. The key benefits of 802.1w include the following:

- The spanning tree topology converges quickly after a switch or link failure.
- Convergence is accelerated by a handshake, known as the proposal agreement mechanism.
- There is no need to enable BackboneFast or UplinkFast.

In terms of convergence, STP algorithms based on 802.1w are much faster than traditional STP 802.1d algorithms. The proposal agreement mechanism allows the Cisco IGESM to decide new port roles by exchanging proposals with its neighbors.

With 802.1w, as with other versions of STP, bridge protocol data units (BPDUs) are still sent, by default, every 2 seconds (called the *hello time*). If three BPDUs are missed, STP recalculates the topology, which takes less than 1 second for 802.1w.

This seems to indicate that STP convergence time can be as long as 6 seconds. However, because the data center is made of point-to-point links, the only failures are physical failures of the networking devices or links. 802 1w is able to actively confirm that a port can safely transition to forwarding without relying on any timer configuration. This means that the actual convergence time is below *1 second* rather than 6 seconds.

The scenario where BPDUs are lost may be caused by unidirectional links, which can cause Layer 2 loops. To prevent this specific problem, you can use Loop Guard and UDLD. Loop Guard prevents a port from forwarding as a result of missed BPDUs, which might cause a Layer 2 loop that can bring down the network.

UDLD allows devices to monitor the physical configuration of fiber optic or copper Ethernet cables and to detect when a unidirectional link exists. When a unidirectional link is detected, UDLD shuts down the affected port and generates an alert. BPDU Guard prevents a port from being active in a spanning tree topology as a result of an attack or misconfiguration of a device connected to a switch port. The port that sees unexpected BPDUs is automatically disabled and must be manually enabled. This gives the network administrator full control over port and switch behavior.

The Cisco IGESM supports Per VLAN Spanning Tree (PVST) and a maximum of 64 spanning tree instances. RPVST+ is a combination of Cisco PVST Plus (PVST+) and Rapid Spanning Tree Protocol. Multiple Instance Spanning Tree (MST) adds Cisco enhancements to 802.1s. These protocols create a more predictable and resilient STP topology, while providing downward compatibility with simpler 802.s and 802.1w switches.

Note   By default, the 802.1w protocol is enabled when running spanning tree in RPVST+ or MST mode.

Traffic Monitoring

Cisco IGESM supports the following traffic monitoring features, which are useful for monitoring BladeCenter traffic in blade server environments:

- Switched Port Analyzer (SPAN)
- Remote SPAN (RSPAN)

SPAN mirrors traffic transmitted or received on source ports to another local switch port. This traffic can be analyzed by connecting a switch or RMON probe to the destination port of the mirrored traffic. Only traffic that enters or leaves source ports can be monitored using SPAN.

RSPAN enables remote monitoring of multiple switches across your network. The traffic for each RSPAN session is carried over a user-specified VLAN that is dedicated for that RSPAN session for all participating switches. The SPAN traffic from the source ports is copied onto the RSPAN VLAN through a reflector port. This traffic is then forwarded over trunk ports to any destination session that is monitoring the RSPAN VLAN.

Link Aggregation Protocols

The Port Aggregation Protocol (PAgP) and Link Aggregation Control Protocol (LACP) help automatically create port channels by exchanging packets between Ethernet interfaces. PAgP is a Cisco-proprietary protocol that can be run only on Cisco switches or on switches manufactured by vendors that are licensed to support PAgP. LACP is a standard protocol that allows Cisco switches to manage Ethernet channels between any switches that conform to the 802.3ad protocol. Because the Cisco IGESM supports both protocols, you can use either 802.3ad or PAgP to form port channels between Cisco switches.

When using either of these protocols, a switch learns the identity of partners capable of supporting either PAgP or LACP and identifies the capabilities of each interface. The switch dynamically groups similarly configured interfaces into a single logical link, called a channel or aggregate port. The interface grouping is based on hardware, administrative, and port parameter attributes. For example, PAgP groups interfaces with the same speed, duplex mode, native VLAN, VLAN range, trunking status, and trunking type. After grouping the links into a port channel, PAgP adds the group to the spanning tree as a single switch port.

Layer 2 Trunk Failover

Trunk failover is a high availability mechanism that allows the Cisco IGESM to track and bind the state of external interfaces with one or more internal interfaces. The four available Gigabit Ethernet uplink ports of the Cisco IGESM provide connectivity to the external network and can be characterized as "upstream" links. The trunk failover feature may track these upstream interfaces individually or as a port channel. Trunk failover logically binds upstream links together to form a link state group. The internal interfaces of the IGESM provide blade server connectivity and are referred to as "downstream" interfaces in the trunk failover configuration. This feature creates a relationship between the two interface types where the link state of the "upstream" interfaces defined in a link state group determines the link state of the associated "downstream" interfaces.

Figure 2-3 illustrates the logical view of trunk failover on the Cisco IGESM. The two external port channels of Switch-1 and Switch-2 are configured as upstream connections in a link state group local to the switch. The 14 internal blade server ports are downstream interfaces associated with each local group.

Figure 2-3   Trunk Failover Logical View
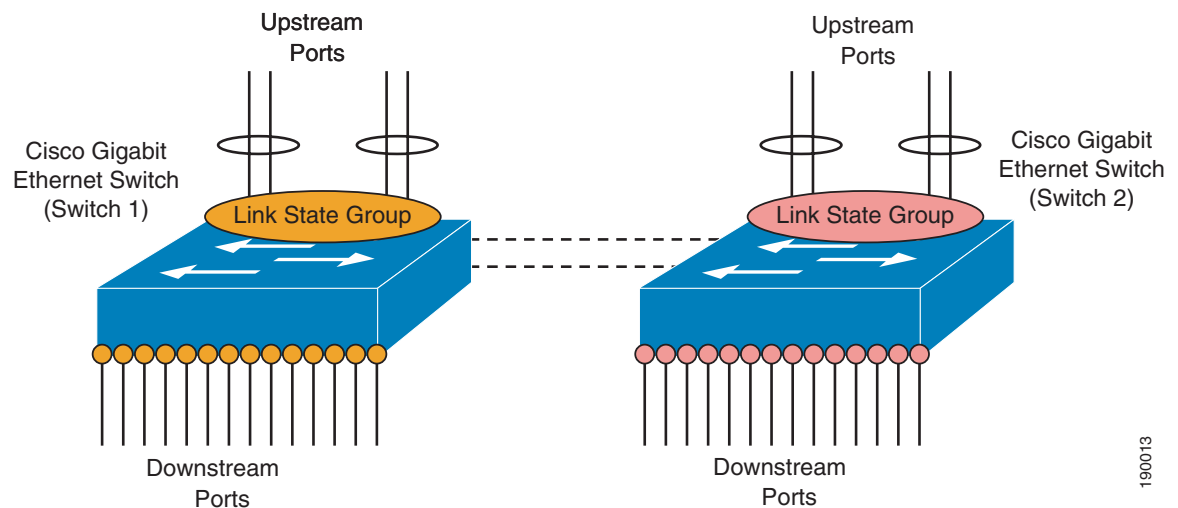
Trunk failover places downstream devices into the same link state, "up" or "down", based on the condition of the link state group. If an uplink or upstream failure occurs, the trunk failover feature places the downstream ports associated with those upstream interfaces into a link "down" or inactive state. When upstream interfaces are recovered, the related downstream devices are placed in an "up" or active state. An average failover and recovery time for network designs implementing the trunk failover feature is 3 seconds.

Consider the following when configuring the trunk failover on the Cisco IGESM:

- Internal ports (Gigabit Ethernet 0/1–14) may not be configured as "upstream" interfaces.
- External ports (Gigabit Ethernet 0/17–20) may not be configured as "downstream" interfaces.
- The internal management module ports (Gigabit Ethernet 0/15–16) may not be configured in a link state group.
- Trunk failover does not consider STP. The state of the upstream connections determines the status of the link state group not the STP state forwarding, blocking, and so on.
- Trunk failover of port channels requires that all of the individual ports of the channel fail before a trunk failover event is triggered.

- SPAN/RSPAN destination ports are automatically removed from the trunk failover link state groups.

Using the IBM BladeCenter in the Data Center Architecture

The BladeCenter chassis provides a set of internal redundant Layer 2 switches for connectivity to the blade servers. Each blade server installed in the BladeCenter can use dual NICs connected to both Layer 2 switches. The BladeCenter can be also be deployed without redundant switches or dual-homed blade servers.

Figure 2-1 illustrates the physical connectivity of the BladeCenter switches and the Blade Servers within the BladeCenter, while the logical connectivity is shown in Figure 2-2. When using the Cisco IGESM, a BladeCenter provides four physical uplinks per Cisco IGESM to connect to upstream switches. Blade servers in the BladeCenter are dual-homed to a redundant pair of Cisco IGESMs.

BladeCenters can be integrated into the data center topology in various ways. The primary design goal is a fast converging, loop-free, predictable, and deterministic design, and this requires giving due consideration to how STP algorithms help achieve these goals.

This section describes the design goals when deploying blade servers and the functionality supported by the Cisco IGESM in data centers. It includes the following topics:

- High Availability
- Scalability
- Management

High Availability

Traditionally, application availability has been the main consideration when designing a network for supporting data center server farms. Application availability is achieved through a highly available server and network infrastructure. For servers, a single point of failure is prevented through dual-homing. For the network infrastructure, this is achieved through dual access points, redundant components, and so forth.

When integrating the BladeCenter, the Cisco IGESM Layer 2 switches support unique features and functionality that help you achieve additional design considerations.

High availability, which is an integral part of data center design, requires redundant paths for the traffic to and from the server farm. In the case of a BladeCenter deployment, this means redundant blade server connectivity. The following are two areas on which to focus when designing a highly available network for integrating BladeCenters:

- High availability of the switching infrastructure provided by the Cisco IGESM
- High availability of the blade servers connected to the Cisco IGESM

High Availability for the BladeCenter Switching Infrastructure

Redundant paths are recommended when deploying BladeCenters, and you should carefully consider the various failure scenarios that might affect the traffic paths. Each of the redundant BladeCenter Layer 2 switches provides a redundant set of uplinks, and the design must ensure fast convergence of the spanning tree topology when a failure in an active spanning tree link occurs. To this end, use the simplest possible topology with redundant uplinks and STP protocols that are compatible with the BladeCenter IGESMs and the upstream switches.

To create the redundant spanning tree topology, connect each of the BladeCenter IGESMs to a set of Layer 2/3 upstream switches that support RPVST+. To establish physical connectivity between the BladeCenter IGESMs and the upstream switches, dual-home each IGESM to two different upstream Layer 3 switches. This creates a deterministic topology that takes advantage of the fast convergence capabilities of RPVST+.

To ensure that the topology behaves predictably, you should understand its behavior in both normal and failure conditions. The recommended topology is described in more detail in Design and Implementation Details, page 2-13.

Figure 2-4 illustrates a fully redundant topology, in which the integrated Cisco IGESMs are dual-homed to each of the upstream aggregation layer switches. Each Cisco IGESM has a port channel containing two Gigabit Ethernet ports connected to each aggregation switch.

Figure 2-4   ***Cisco IGESM Redundant Topology***
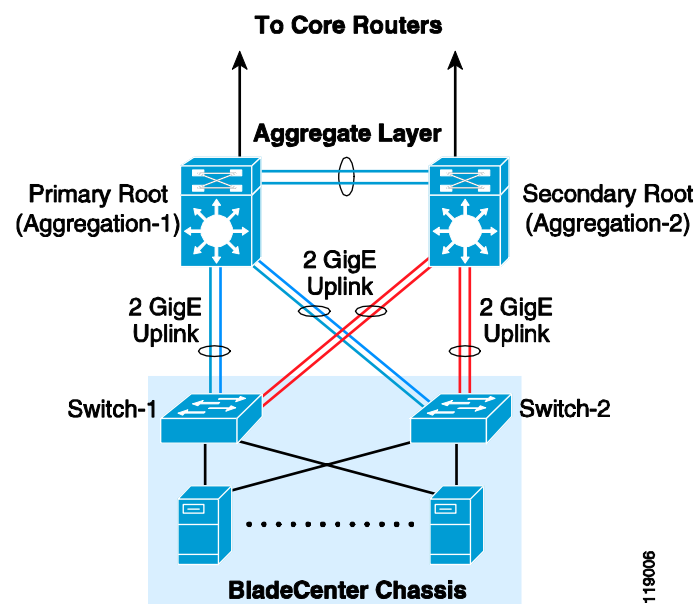
This provides a fully redundant topology, in which each BladeCenter switch has a primary and backup traffic path. Also notice that each Cisco IGESM switch has a deterministic topology in which RPVST+ provides a convergence time of less than one second after a failure. The environment is highly predictable because there is a single primary path used at all times, even when servers are dual-homed in active-standby scenarios.

Note   The aggregation switches that provide connectivity to the BladeCenter are *multilayer* switches. Cisco does not recommend connecting a BladeCenter to Layer 2-only upstream switches.

High Availability for the Blade Servers

Blade server high availability is achieved by multi-homing each blade to the integrated IGESMs employing the trunk failover feature. Multi-homing can consist of dual-homing each server to each of the Cisco IGESM switches, or using more than two interfaces per server, depending on the connectivity requirements.

Dual-homing leverages the NIC teaming features offered by the Broadcom chipset in the server NICs. These features support various teaming configurations for various operating systems. The following teaming mechanisms are supported by Broadcom:

- Smart Load Balancing
- Link Aggregation (802.3ad)
- Gigabit Cisco port channel

Smart Load Balancing is the only method of dual homing applicable to blade servers. The other two methods of teaming are not discussed in this document because they are not applicable. Although three teaming methods are supported, neither 802.3ad or Gigabit port channels can be used in the BladeCenter for high availability because the servers are connected to two different switches and the physical connectivity is dictated by the hardware architecture of the BladeCenter.

With Smart Load Balancing, both NICs use their own MAC addresses, but only the primary NIC MAC address responds to ARP requests. This implies that one NIC receives all inbound traffic. The outbound traffic is distributed across the two NICs based on source and destination IP addresses when the NICs are used in active-active mode.

The trunk failover feature available on the Cisco IGESM combined with the NIC teaming functionality of the Broadcom drivers provides additional accessibility to blade server resources. Trunk failover provides a form of "network awareness" to the NIC by binding the link state of upstream and downstream interfaces. The IGESM is capable of tracking the condition of its uplinks and placing associated "downstream" blade server ports in the same link state. If uplink failure occurs, the trunk failover feature disables the internal blade server ports, allowing a dual-homed NIC to converge using the high availability features of the NIC teaming driver. The trunk failover feature also recovers the blade server ports when uplink connectivity is re-established.

Scalability

From a design perspective, Layer 2 adjacency also allows horizontal server farm growth. You can add servers to the same IP subnet or VLAN without depending on the physical switch to which they are connected, and you can add more VLANs/IP subnets to the server farm while still sharing the services provided by the aggregation switches.

Scaling the size of BladeCenters server farms depends on the following characteristics of the network:

- Physical port count at aggregation and access layers (the access layer being the Cisco IGESMs)
- Physical slot count of the aggregation layer switches

The following sections provide some guidance for determining the number of physical ports and physical slots available.

Physical Port Count

Scalability, in terms of the number of servers, is typically determined by the number of free slots and the number of ports available per slot. With BladeCenter, this calculation changes because the blade servers are not directly connected to traditional external access layer or aggregation layer switches.

With BladeCenters, the maximum number of servers is limited by the number of BladeCenters and the number of ports in the upstream switches used to connect to the BladeCenters.

In the topology illustrated in Figure 2-1, for every 14 servers per BladeCenter, each aggregation switch needs to provide four Gigabit Ethernet ports (two to each Cisco IGESM).

The port count at the aggregation layer is determined by the number of slots multiplied by the number of ports on the line cards. The total number of slots available is reduced by each service module and supervisor installed.

Table 2-1 summarizes the total number of blade servers that can be supported for various line cards on a Cisco Catalyst 6500 switch on a per-line card basis. Keep in mind that the uplinks are staggered between two distinct aggregation switches, as shown in Figure 2-4.

Table 2-1   BladeCenters Supported Based on Physical Port Count

Type of Line Card	Cisco IGESM per BladeCenter	Uplinks per Cisco IGESM	Total Uplinks	BladeCenters per Line Card
8-port Gigabit Ethernet	2	2	4	4
		4	8	2
	4	2	8	2
		4	16	1
16-port Gigabit Ethernet	2	2	4	8
		4	8	4
	4	2	8	4
		4	16	2
48-port Gigabit Ethernet	2	2	4	24
		4	8	12
	4	2	8	12
		4	16	6

Slot Count

Your design should be flexible enough to quickly accommodate new service modules or BladeCenters without disruption to the existing operating environment. The slot count is an important factor in planning for this goal because the ratio of servers to uplinks dramatically changes as the number of BladeCenters increases.

This scaling factor is dramatically different than those found in traditional server farms where the servers are directly connected to access switches and provide very high server density per uplink. In a BladeCenter environment, a maximum of 14 servers is supported over as many as eight uplinks per BladeCenter. This creates the need for higher flexibility in slot/port density at the aggregation layer.

A flexible design must be able to accommodate growth in server farm services along with support for higher server density, whether traditional or blade servers. In the case of service modules and blade server scalability, a flexible design comes from being able to increase slot count rapidly without changes to the existing architecture. For instance, if firewall and content switching modules are required, the slot count on each aggregation layer switch is reduced by two.

Cisco recommends that you start with a high-density slot aggregation layer and then consider the following two options to scale server farms:

- Use a pair of service switches at the aggregation layer.
- Use data center core layer switches to provide a scaling point for multiple aggregation layer switches.

Using service switches for housing service modules maintains the Layer 2 adjacency and allows the aggregation layer switches to be dedicated to provide server connectivity. This uses all available slots for line cards that link to access switches, whether these are external switches or integrated IGESMs. This type of deployment is illustrated in Figure 2-4.

Figure 2-5 illustrates traditional servers connected to access switches, which are in turn connected to the aggregation layer.

Figure 2-5 **Scaling With Service Switches**
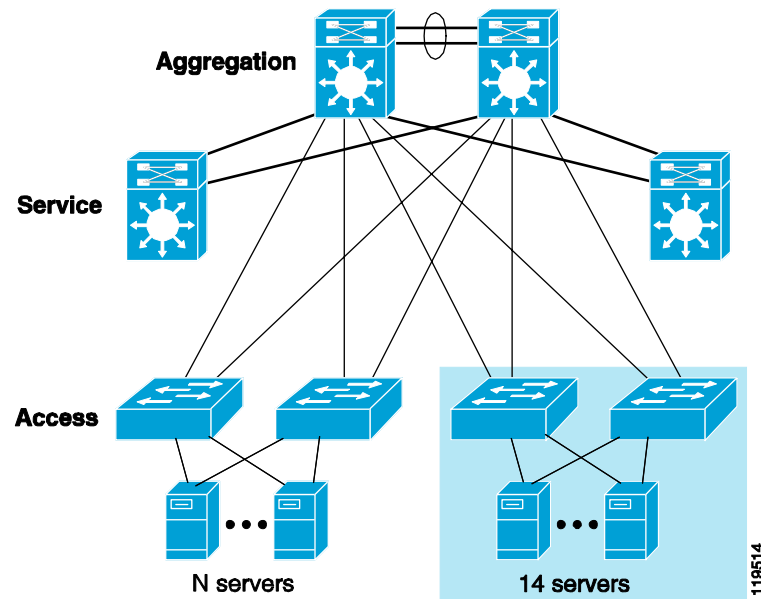
Blade servers, on the other hand, are connected to the integrated IGESMs, which are also connected to the aggregation switches. The slot gained by moving service modules to the service layer switches lets you increase the density of ports used for uplink connectivity.

Using data center core layer switches allows scaling the server farm environment by sizing what can be considered a single module and replicating it as required, thereby connecting all the scalable modules to the data center core layer. Figure 2-6 illustrates this type of deployment.

Figure 2-6   Scaling With Data Center Core Switches
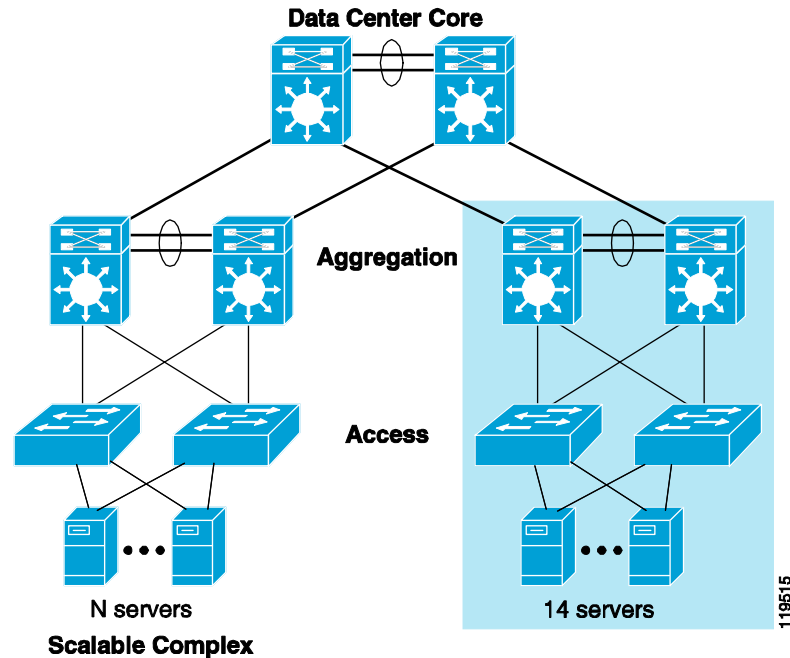
In the topology displayed in Figure 2-6, all service modules are housed in the aggregation layer switches. These service modules support the server farms that share the common aggregation switching, which makes the topology simple to implement and maintain. After you determine the scalability of a single complex, you can determine the number of complexes supported by considering the port and slot capacity of the data center core switches. Note that the core switches in this topology are Layer 3 switches.

Management

You can use the BladeCenter Management Module to configure and manage the blade servers as well as the Cisco IGESMs within the BladeCenter without interfering with data traffic. To perform configuration tasks, you can use a browser and log into the management module.

Within the BladeCenter, the server management traffic (typically server console access) flows through a different bus, called the I^2C bus. The I^2C bus and the data traffic bus within the BladeCenter are kept separate.

The BladeCenter supports redundant management modules. When using redundant management modules, the backup module automatically inherits the configuration of the primary module. The backup management module operates in standby mode.

You can access the management module for configuring and managing the Cisco IGESMs using the following three methods, which are described in the following sections:

- Out-of-Band Management
- In-Band Management
- Serial/Console Port

Out-of-Band Management

Out-of-band management refers to the common practice of dedicating a separate interface on each manageable device for management traffic.

Each management module in the BladeCenter supports an Ethernet interface, which is used to manage the blade servers and the Ethernet switches as well as the management module itself.

By default, the Ethernet switch management ports are placed in VLAN 1. It is very important that you put the management interface in a VLAN that is not shared by any of the blade server interfaces. The BladeCenter switch does not let you change the native VLAN on the interface connected to the management module.

Figure 2-7 presents a logical representation of an out-of-band management environment.

Figure 2-7   Out-Of-Band Management
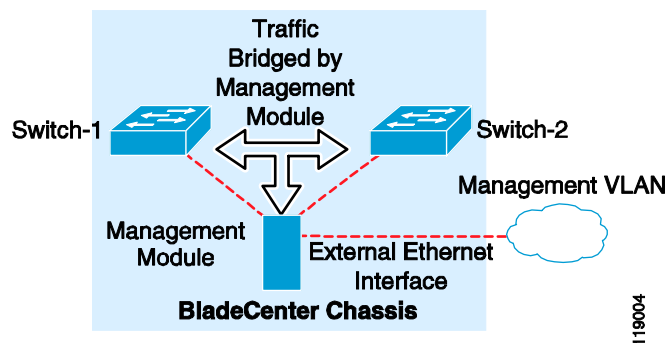
Note that the dotted line shows control traffic between the external Ethernet interface and the Cisco IGESMs.

The management module Ethernet interface connects to the out-of-band management network. The management traffic destined for the Cisco IGESM comes through the management module Ethernet interface.

Note Do *not* configure VLAN 1 on any Cisco IGESM interface that is connected to a blade server. The Cisco IGESM management interface is in VLAN 1, by default, and all traffic from the blade server would be broadcast to VLAN 1.

In-Band Management

With in-band management, Telnet or SNMP traffic uses the same path that is used by data, which limits the bandwidth available over uplinks. However, in-band management is common, especially where application management and network management traffic is separated in different VLANs or subnets.

Therefore, in addition to the VLAN 1 consideration, it is important to keep all management traffic on a different VLAN than non-control traffic.

Serial/Console Port

Like other Cisco products, the Cisco IGESM has a serial port that can be used for management purposes. The serial port is typically connected to a terminal server through which the IGESMs can be managed remotely. You can establish a management session over a Telnet connection to the terminal server IP address and to the port to which the IGESM is connected. See the following URL for details about how

to set up console access to the switch through the serial/console port:
http://www.cisco.com/en/US/docs/switches/lan/catalyst2950/software/release/12.1_9_ea1/configuratio
n/guide/swadmin.html.

Design and Implementation Details

This section provides design and implementation details for a server farm design that integrates BladeCenters. The following topics are included:

- Network Management Recommendations
- Layer 2 Looped Access Layer Design—Classic "V"
- Layer 2 Loop-Free Access Layer Design—Inverted "U"
- Configuration Details

Network Management Recommendations

As described in the previous sections, there are various ways to manage the Cisco IGESM switches. Out-of-band management is the recommended option (see Figure 2-7) because it is simpler and management traffic can be kept on a separate VLAN. Additionally, setting the IP addresses on the IGESMs is a straightforward process using the graphic user interface (GUI) provided by the management module.

By default, the Cisco IGESM management ports are placed in VLAN 1. As discussed previously, it is very important that you put the management interface in a VLAN that is not shared by any of the blade server interfaces. The Cisco IGESM does not let you change the native VLAN on the management module interface. By using the default VLAN for the management module interfaces on each Cisco IGESM, management traffic and data traffic are kept in separate VLANs.

In-band management is not recommended for the following reasons:

- Management traffic must share the limited bandwidth between the aggregate switch and the Cisco IGESM switches.
- A loop is introduced by the management module.
- Broadcast traffic can conceivably overload the CPU.

When you configure the BladeCenter switches for in-band management, the management interface is automatically assigned to the management VLAN (VLAN 1), which creates a loop. The management module has three interfaces and it bridges traffic received on these three interfaces. A loop is introduced by the management module because two of the interfaces are connected to the redundant Cisco IGESM switches. Although the Cisco IGESM has hardware filters to discard all the traffic between switches except spanning tree BPDUs and Cisco Discovery Protocol (CDP) packets, this mode is still not recommended. Although STP blocks potential loops, broadcast traffic can conceivably overload the CPU, which can lead to serious problems.

When you assign an IP address to the management interface of the Cisco IGESM, which for management purposes is a VLAN, the management module interface VLAN is automatically changed to the VLAN used along with IP address configuration. This helps ensure connectivity to the switches through the management module.

Layer 2 Looped Access Layer Design—Classic "V"

Figure 2-8 illustrates the topology where the uplinks from the BladeCenter switches are distributed between two aggregate switches. This topology supports high availability by providing redundant Layer 2 links, so there is no single point of failure. The switches can be configured to provide a deterministic traffic path. In this topology, the BladeCenter switches must run STP to block Layer 2 loops.

Figure 2-8   Uplinks Distributed Between Two Aggregate Switches
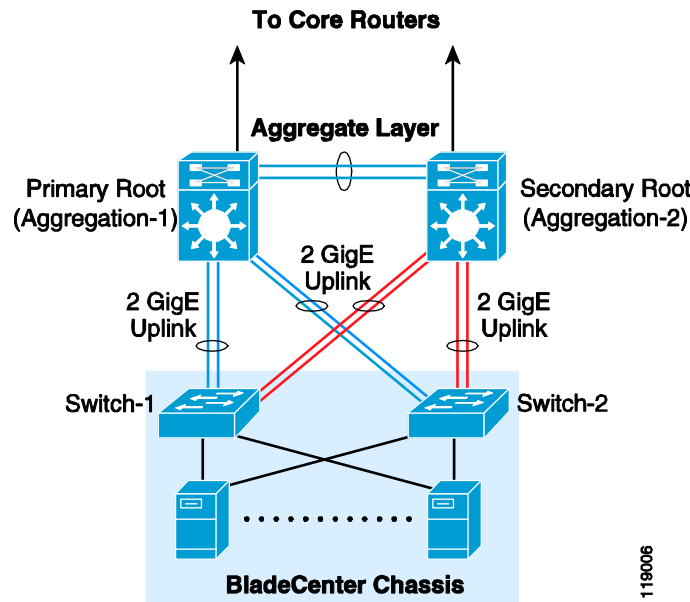
The recommendation is to use RPVST+ because it provides convergence of less than one second in case of device or link failures. In addition to rapid convergence, RPVST+ incorporates many enhanced Cisco Layer 2 features, including BackboneFast and UplinkFast, which are used by default when you enable RPVST+.

To validate this topology, failure times were tested by sending traffic from the blade servers to a Layer 3 device on the aggregate switch at increasingly smaller intervals, and then measuring the number of packets lost. The following failure and recovery scenarios were tested:

- Uplink failure and recovery between Switch-1 and the primary root
- Uplink failure and recovery between Switch-2 and the primary root
- Switch-1 failure and recovery
- Switch-2 failure and recovery
- Primary root switch failure and recovery
- Secondary root switch failure and recovery

In most test cases, the failover and recovery times were a few hundred milliseconds. To allow a margin of error, the failover times can safely be rounded to one second. When the test case involves the failure of a switch that is the active HSRP device, the failover time is dependent on the HSRP failover time. Although HSRP can be configured to converge in sub-second times, a conservative estimate for recovery time when multiple components are involved is approximately five to six seconds.

These failover times are for Layer 2 and Layer 3 failures with HSRP at Layer 3. If the default gateway is on a different device, such as a firewall, the failover time for aggregate switch failure may change.

If a link fails within a port channel with two Ethernet links, the spanning tree topology does not change. The port channel simply stays up with a single link. This helps ensure that the BladeCenter traffic flow is not affected by the link failure.

The recommended topology provides redundant paths to BladeCenter traffic under all failure scenarios except for one case. This particular case is when all the links fail between a BladeCenter switch and the aggregation switches, and the NIC on the blade server is unaware of the uplink failure. The NIC teaming drivers cannot detect this condition and the servers are isolated until the links are restored. The trunk failover feature, available on the Cisco IGESM, addresses this situation. Trunk failover places blade server switch ports in a "down" state when their associated upstream uplinks fail. By doing so, the dual-homed server relies on the high availability features of the NIC teaming software to bypass the network failure and re-establish network connectivity in three seconds.

Cisco recommends alternating the active blade server interfaces on different Cisco IGESMs. This configuration helps prevent server isolation in the absence of trunk failover, and overall provides for better bandwidth utilization with this design. In addition, it places content switches in front of the server farm. The content switch can detect the failure or isolation of servers and can reroute requests to available resources.

It is also possible to monitor traffic on the Cisco IGESM switches in the topology shown in Figure 2-8. Under normal conditions, the backup links that are blocking can carry RSPAN traffic to the aggregate switches on a VLAN specifically used for mirrored traffic. This VLAN is configured only on Aggregation-2, on the Cisco IGESMs, and the backup link connecting these switches. This means that the topology is loop-free, and all ports are forwarding for this VLAN only. A network analysis probe should be attached to Aggregation-2. In this design, under failure conditions, the mirrored traffic shares the data traffic path.

> **Note** The RSPAN VLAN is used only when it is required. For traffic mirroring, the Cisco IGESM switches require that one of the internal ports be configured as a reflector port, which means that one of the internal server ports is used for RSPAN. This is the recommended topology when a dedicated port for SPAN is not required.

To monitor traffic in this topology, perform the following steps:

Step 1 Configure a VLAN for RSPAN and allow that VLAN on the backup port channel (blocking spanning tree link) on both the aggregate switch and the Cisco IGESM switch.

Step 2 Configure traffic monitoring using the VLAN created in Step 1.

Given the current example, to configure traffic monitoring for the server connected to Gi0/5, enter the following commands:

```
monitor session 1 source interface Gi0/5
monitor session 1 destination remote vlan 300 reflector-port int Gi0/14
```

Configuring the Aggregate Switches

Complete the following sequence of tasks on the aggregate switches:

1. VLAN configuration
2. RPVST+ configuration

3. Primary and secondary root configuration

4. Configuration of port channels between aggregate switches

5. Configuration of port channels between aggregate and Cisco IGESM switches

6. Trunking the port channels between aggregate switches

7. Configuration of default gateway for each VLAN

These tasks might be performed on a different device or a service module instead of the MSFC on the aggregate switch, depending on the architecture.

Configuring the Cisco IGESM Switches

Complete the following sequence of tasks on the Cisco IGESM switches:

1. VLAN configuration

2. RPVST+ configuration

3. Configuration of port channels between the Cisco IGESM and aggregate switches

4. Trunking port channels between the Cisco IGESM and aggregate switches

5. Configuration of server ports on the Cisco IGESM

6. Configure trunk failover on the Cisco IGESM

Configuration Details, page 2-21 provides details for each of these steps. Most of these steps are required for either the primary recommended topology or the alternate topologies. In addition to these configuration steps, the following recommendations should be followed to ensure the successful integration of Cisco IGESM switches.

Additional Aggregation Switch Configuration

Step 1 Enable RootGuard on both the aggregate switch links that are connected to the Cisco IGESMs in the BladeCenter.

This prevents the Cisco IGESMs from assuming the STP root role by shutting down the interfaces in the aggregation switch that are connected to the Cisco IGESM. This safeguards against network meltdown in case of Cisco IGESM misconfiguration or misbehavior.

RootGuard can be enabled on both the aggregate switches using the **spanning-tree guard root** command in interface configuration mode. Enter this command on all the port channel interfaces between the aggregate switch and the Cisco IGESM switches.

Step 2 Limit the VLANs on the port channel between the aggregate and Cisco IGESMs to those that are required.

Use the **switchport trunk allowed vlan** *<vlandID>* command on both the Cisco IGESM and aggregation switches in interface configuration mode. Enter this command on the Gigabit Ethernet interfaces and port channels.

Additional Cisco IGESM Configuration

Note If BPDU Guard is enabled and a bridge protocol data unit (BPDU) is received on a port, the port is shut down. For this reason, do *not* enable BPDU Guard or BPDU filtering on the internal port connected to the management module because connectivity to the management module would be lost.

If connectivity to the management module is lost because BPDU Guard is enabled, you must reload the switch to recover from this condition. If the faulty configuration was saved, you must remove the connection from the management module external interface before reloading the switch.

Step 1 Remove BPDU filtering.

By default, BPDU filter is enabled on all the internal ports of Cisco IGESM. To disable this, use the **spanning-tree bpdufilter disable** command in interface configuration mode.

Step 2 Remove BPDU Guard.

To remove BPDU Guard, use the **spanning-tree bpduguard disable** command. However, BPDU guard can be enabled on the internal access ports connected to the blade servers.

Step 3 Restrict VLANs on the port channel between the aggregate and Cisco IGESMs to those required.

Use the **switchport trunk allowed vlan** *<vlanID>* command on both the Cisco IGESM and aggregation switches in interface configuration mode. This should be applied on the Gigabit Ethernet interfaces and port channels.

Note Do not use VLAN 1 on any of the interfaces on the Cisco IGESM connected to the blade servers. By default, the interface connected to the management module is in VLAN 1 and all the management traffic flows to the CPU of the Cisco IGESM.

Layer 2 Loop-Free Access Layer Design—Inverted "U"

The topology in Figure 2-9 has the four individual uplinks from the BladeCenter switches connected to a single aggregate aggregation switch. This design is considered a loop-free topology. Switch-1 and Switch-2 have no interconnect; therefore, no loop exists. A redundant network path to the blade server NICs connected to each BladeCenter switch does not exist. Despite the absence of designed loops, Cisco recommends enabling spanning tree, preferably RPVST+, to manage potential loops created by misbehaving devices or human error.

Figure 2-9   IGESM Uplinks Connected To One Aggregate Switch
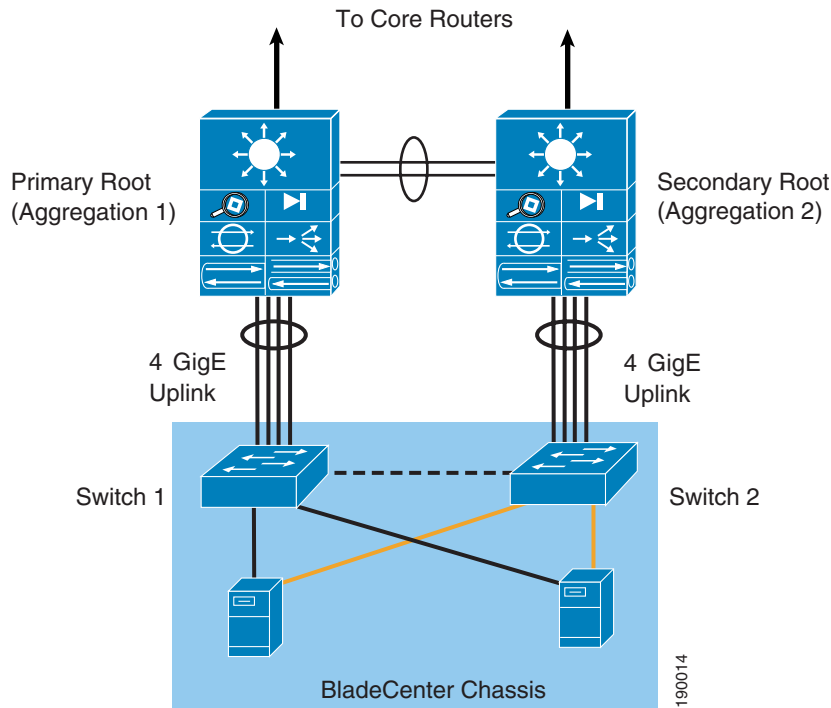
As noted with the previous design, a link failure between the BladeCenter switch and its associated aggregate switch isolates the blade server. The blade server is not aware of the network connectivity issues occurring beyond its own NIC. The trunk failover feature allows the Cisco IGESM and the blade server NIC teaming drivers to account for this scenario. The tests performed with this design achieved an average convergence of network traffic in three seconds.

The topology shown in Figure 2-9 has the advantage of higher bandwidth utilization when compared to other designs. Each Cisco IGESM uses the four Gigabit uplinks available. This design implies that blade server traffic that may pass through IGESM Switch-2 uses the aggregate inter-switch links to reach the primary root and the network services it may host. The links between the two aggregate switches should provide sufficient bandwidth to support all of the BladeCenters present in the data center. As a result, the configuration of the dual-homed blade server NICs becomes an important consideration.

The primary interface assignment of a NIC team influences the traffic patterns within the data center. To avoid oversubscription of links between the aggregation switches and to simplify server deployments, assign the primary NIC on the IGESM homed to the primary root switch (Switch-1 in Figure 2-9). The oversubscription ratio of a fully-populated IBM BladeCenter with this NIC configuration is 3.5:1. The secondary NIC, assigned to the second IGESM (Switch-2 in Figure 2-10) is inactive unless activated by a failover. If oversubscribing the inter-switch links is not a concern, the blade server primary NIC interfaces may be evenly distributed between the two IGESMs, creating an oversubscription ratio on each switch of 1.75:1 that permits a higher bandwidth solution.

The configuration of this design requires the same steps as the recommended topology with the exception of redundant links to the aggregation layer from the IGESM.

Figure 2-10 and Figure 2-11 show two different topologies that provide a dedicated port for traffic monitoring. As mentioned in Network Management Recommendations, page 2-13, RSPAN requires a Cisco IGESM switch port to be configured as a reflector port. This port is then used for traffic monitoring instead of being available for a blade server. You can avoid using a valuable Cisco IGESM port by dedicating an external port to SPAN traffic.

Note This topology in which the link to the secondary root from Switch-2 is non-blocking is useful when performing any kind of maintenance function to the primary root that can affect the BladeCenter connectivity, thus providing the active server NICs on Switch-2 a primary forwarding path. This presumes that servers are both dual-homed and that half the servers are active on Switch-1 and the other half are active on Switch-2.

Figure 2-10   First Topology with SPAN Ports
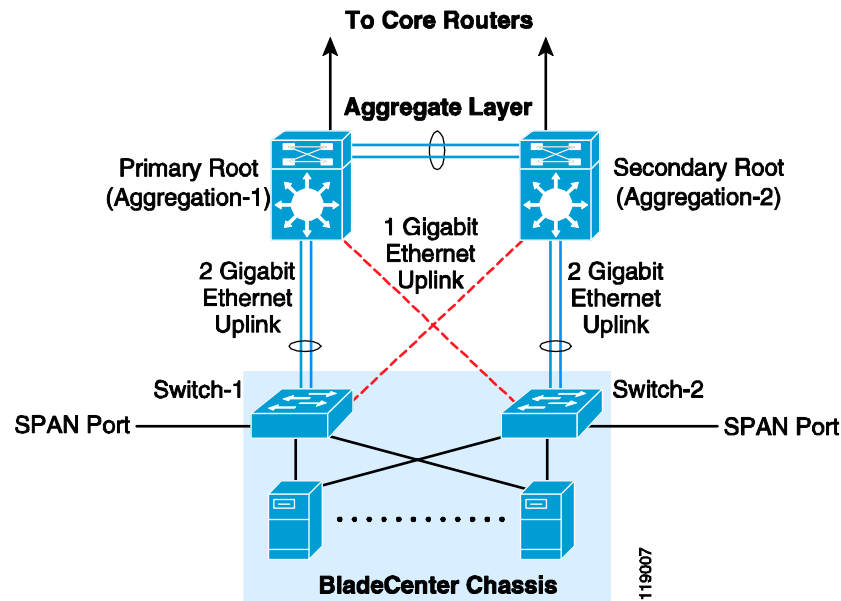
In this example, to monitor Blade Server 2 traffic, enter the following commands:

```
monitor session 1 source interface Gi0/2
monitor session 1 destination int Giga0/20
```

The topology presented in Figure 2-11 can be considered a variation of a topology in which the port channel is always non-blocking to the primary root, and the single link from either Cisco IGESM is to the secondary root, thus having a direct primary path from the primary root to each Cisco IGESM.

These topologies are recommended only if dedicated ports are required for monitoring traffic to and from the servers connected to Cisco IGESMs. In both topologies shown in Figure 2-10 and Figure 2-11, a dedicated port is used to monitor traffic on both the Cisco IGESM switches.

Figure 2-11   Second Topology with SPAN Ports

BladeCenter Chassis
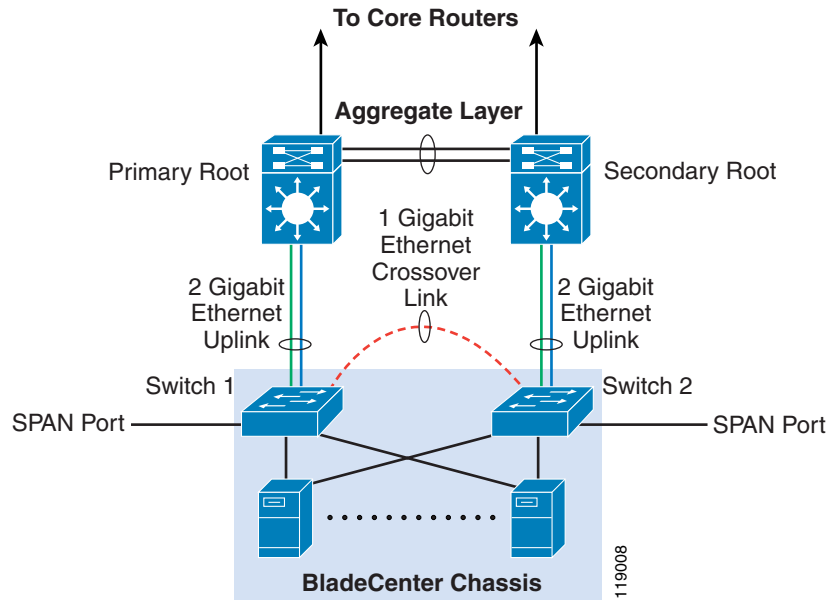
The disadvantage in these topologies is that in the event of device or link failures external to the BladeCenter, the redundant traffic path uses lower bandwidth links. You must monitor the link and device failures so you can quickly restore the higher bandwidth traffic path.

The difference between these two topologies is the traffic path during link or device failures. The advantage of the first topology (Figure 2-10) is that when the primary root switch fails, the traffic switches over to the secondary root directly (one hop versus two hops to get to the secondary root).

Note   For the topology in Figure 2-10, you must change the path cost so that under normal circumstances, the traffic from Switch-2 in the BladeCenter always takes the port channel (2-port Gigabit Ethernet uplink) to the secondary root. See Configuration Details, page 2-21 for details about the commands required.

The second topology (Figure 2-11) saves some cabling effort and requires no modification to the spanning tree path cost. However, from the standpoint of providing an optimal traffic path, the first topology is recommended when dedicated ports are required for traffic monitoring.

For testing the failover times, the blade servers were dual-homed to both the switches and the RedHat Linux operating system was used. To test failover times for different links and devices, the following failure and recovery scenarios were tested:

- Uplink failure and recovery from Switch-1 to primary root
- Uplink failure and recovery from Switch-2 to secondary root
- Failure and recovery of Switch-1 and Switch-2 of BladeCenter
- Failure and recovery of the aggregation switches

In most cases, the failover time can still be rounded up to one second. However, as before, an active HSRP switch failure increases the failover time.

The best practices for this topology are the same as those described previously in Network Management Recommendations, page 2-13. For implementing this topology, the physical topologies may change and corresponding configuration changes will be required. However, the majority of the implementation details are similar to the recommended topology described previously.

The implementation steps for the recommended topology apply here as well. The differences in implementation are as follows:

- Inter-switch links can have single links, so port channel configuration is not required for every inter-switch link.

- The spanning tree path cost might have to be changed to ensure that the traffic follows the high bandwidth links under normal circumstances.

The details of the trunk configuration can be found in Configuration Details, page 2-21. For trunks between switches, ensure that only the required VLANs are allowed and that all traffic is tagged.

In the topology shown in Figure 2-10, you need to change the spanning tree path cost. Change the path cost of the trunk between Switch-2 and the primary root to a higher value (such as 8) on the Switch-2 interface. By changing this path cost, the traffic is forced to take the high bandwidth path if the traffic originates at Switch-2. Enter the following commands in global configuration mode to change the path cost.

```
interface interface-id
spanning-tree cost cost
spanning-tree vlan vlan-id cost cost
```

To verify the path cost, enter the following commands:

```
Show spanning-tree interface interface-id
Show spanning-tree vlan vlan-id
```

For the topology shown in Figure 2-11, you do not need to change the spanning tree path cost. In this topology, the only difference between that described in Figure 2-11 and in Network Management Recommendations, page 2-13, is that you must configure a trunk on the Gigabit Ethernet link between the Cisco IGESM switches and the aggregation switches.

Configuration Details

This section describes the configuration steps required for implementing the topologies discussed in this design guide. It includes the following topics:

- VLAN
- RPVST+
- Inter-Switch Link
- Port Channel
- Trunking
- Server Port
- Verifying Connectivity Between Cisco IGESMs
- Server Default Gateway
- Changing Spanning Tree Path Cost
- Layer 2 Trunk Failover

Figure 2-12 illustrates the topology to which this configuration applies and shows the primary root, secondary root, and where the port channel and link aggregation configurations are used.

Figure 2-12   Example Topology
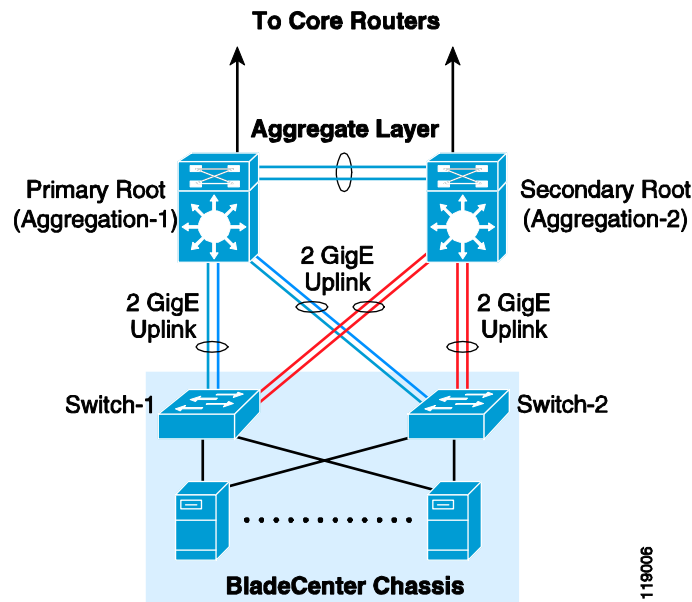
VLAN

Before configuring VLANs, you need to define the VTP mode for the switch. Enter the following commands to configure VTP:

```
vtp domain domain name
vtp mode transparent
```

Use the same VTP domain name everywhere in the data center. The configuration in this section covers only the server VLANs.

Other VLANs become involved when more services are added on top of the server VLANs. For more information about the configuration required for these additional services, see the *Data Center Infrastructure SRND* at the following URL:
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

To create the VLANs, enter the following commands:

```
(config)# VLAN 10
(config-vlan)# name Blade1Vlan
```

VLANs 1–1002 are normal VLANs, while the VLANs numbered 1006 and higher are extended VLANs. After the VLANs are configured, name them appropriately and use the following command to place each VLAN in active state.

```
(config-vlan)# state active
```

RPVST+

The following configuration details are for the aggregation switches. One aggregation switch is the primary root (by convention, the switch on the left) and the other aggregation switch is configured as the secondary root.

The examples in this design document show one aggregation switch (Aggregation-1) as the primary root and the other aggregation switch (Aggregation-2) as the secondary root for all the instances. This is because the design best practice is to not distribute traffic on uplinks to maintain a deterministic topology.

To configure RPVST+ in Cisco IOS Software Native Mode, enter the following command:

```
spanning tree mode rapid-pvst
```

A key recommendation is to have a single spanning tree topology for the VLANs in a set of redundant access and aggregation switches in the data center. If it is necessary to distribute the load through uplinks between access and aggregation switches, and to assign different priorities for the even and odd VLANs.

The next configuration step is to assign the root and the secondary root switches.

The configuration on Aggregation-1 for Cisco IOS Software Native Mode is as follows:

```
spanning-tree vlan 10,20,30 root primary
```

The configuration on Aggregation-2 for Cisco IOS Software Native Mode is as follows:

```
spanning-tree vlan 10,20,30 root secondary
```

With the **mac address reduction** option enabled, these commands assign the following election priorities:

- Root bridge priority—24576 (instead of 8192 without **mac address reduction** enabled)
- Secondary root bridge priority—28672 (instead of 16384 without **mac address reduction** enabled)
- Regular bridge priority—32768

Note   With RPVST+, there is no need for UplinkFast and BackboneFast. Just configure RPVST+ in all the devices that belong to the same VTP domain.

Inter-Switch Link

An inter-switch link here refers to the link between the switches and is not the same as ISL. There are two components in an inter-switch link:

- Port channel configuration
- Trunking configuration

The following inter-switch links are required for the topologies in this design guide:

- Inter-switch link between the aggregate switches (port-channel + trunk)
- Inter-switch link from both BladeCenter Cisco IGESMs to both aggregate switches (port-channel + trunk)
- In some topologies, you need to create a trunk link between the two Cisco IGESMs in the BladeCenter.

Port Channel

To configure a port channel between the two aggregate switches, follow these guidelines:

- Use multiple ports from different line cards to minimize the risk of losing connectivity between the aggregation switches.
- Configure LACP active on Aggregation-1.

- Configure LACP passive on Aggregation-2.

The following example shows the channel configuration between the ports Giga1/1, Giga1/2, Giga6/1, and Giga6/2 for Aggregation-1.

```
interface GigabitEthernet1/1
description to_Aggregation-2
channel-group 2 mode active
channel-protocol lacp
interface GigabitEthernet1/2
description to_Aggregation-2
channel-group 2 mode active
channel-protocol lacp
interface GigabitEthernet6/1
description to_Aggregation-2
channel-group 2 mode active
channel-protocol lacp
interface GigabitEthernet6/2
description to_Aggregation-2
channel-group 2 mode active
channel-protocol lacp
```

The configuration for Aggregation-2 is the same with the exception that the channel mode is passive, configured with the following command*:*

```
channel-group 2 mode passive
```

For more information about configuring port channels, see the following URL:
http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.1E/native/configuration/guide/channel.html.

Trunking

To configure trunking between the two switches, follow these guidelines:

- Do not configure the trunks to allow all VLANs. Allow only those that are used.
- Use 802.1q trunking.
- Tag all VLANs over a trunk from the aggregation switches.

To define the VLANs allowed on a trunk, enter the following command:

```
switchport trunk allowed vlan 10,20
```

To modify the list of the VLANs allowed on a trunk, enter the following commands in Cisco IOS Software Native Mode:

```
switchport trunk allowed vlan add vlan number
switchport trunk allowed vlan remove vlan number
```

The recommended trunk encapsulation is 802.1q because it is the standard. The configuration in Catalyst 6500 IOS is as follows:

```
switchport trunk encapsulation dot1q
```

With some software versions, 802.1q is the default encapsulation and the **dot1q** option is not required.

You can force a port to be a trunk by entering the following command:

```
switchport mode trunk
```

This mode puts the port into permanent trunk mode and sends DTP frames to turn the neighboring port into a trunk as well. If the trunk does not form, verify the VTP domain configuration. VTP domain names must match between the neighboring switches.

The port channels that connect the aggregation switch to the BladeCenter switches are also trunked and should have Root Guard configured as shown in the following configuration. This configuration also shows a port channel trunking from Aggregation-1 to Switch-1

```
interface GigabitEthernet6/5
 description to-BladeCenterSW1/1
 no ip address
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 2
 switchport trunk allowed vlan 10,20
 switchport mode trunk
 spanning-tree guard root
 channel-group 5 mode active
 channel-protocol lacp

interface GigabitEthernet6/6
 description to-BladeCenterSW1/1
 no ip address
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk native vlan 2
 switchport trunk allowed vlan 10,20
 switchport mode trunk
 spanning-tree guard root
 channel-group 5 mode active
 channel-protocol lacp
```

Note Configuring Root Guard on the port channel interface between the two Catalyst 6500 switches ensures that the primary root (Aggregation-1) is always the root for all the VLANs as long as the switch is up and running.

Server Port

When a blade server is inserted into the BladeCenter, the blade server NIC attaches to specific Gigabit Ethernet interfaces on both BladeCenter switches based on the slot into which the blade is inserted. For example, when a blade server is inserted into slot 8, it is connected to Gigabit Ethernet interface 0/8 on both switches.

Trunking and PortFast are enabled by default on the access ports. It is also useful to trunk the server ports if trunking is configured on the blade server NICs. If trunking is not enabled on the server NICs (whether teamed or not), you can change the configuration to non-trunking and configure the port in access mode.

In addition, BPDU filtering is enabled by default. If the server drivers and operating systems do not bridge BPDUs and do not have to see BPDUs, there is no need to change this default configuration. However, if you enable BPDU Guard and disable BPDU filtering, the BPDUs are allowed to pass through from the Cisco IGESM interface to the blade server modules, and if the blade server modules bridge the BPDUs, BPDU Guard shuts down the port.

Note Do not enable BPDU Guard or BPDU filtering on the internal port connected to the management module.

If BPDU Guard is enabled and a BPDU is received on this interface, the port shuts down and connectivity to the management module is lost. To recover from this condition, you have to reload the switch. If the faulty configuration was saved, you must remove the connection from the management module external interface before reloading the switch.

Enable trunk failover on the "downstream" internal blade server ports and assign the port to a specific trunk failover group. Port security can also be enabled on the access ports. The new non-trunking (access port) configuration is as follows:

```
interface GigabitEthernet0/8
 description blade8
 switchport access vlan 20
 switchport mode access
 switchport port-security maximum 3
 switchport port-security aging time 20
 link state group 1 downstream
 spanning-tree portfast
 spanning-tree bpduguard enable
 no cdp enable
end
```

Note No explicit port speed settings or duplex settings are shown here because auto-negotiation is reliable between the blade servers and the BladeCenter Cisco IGESM. However, if the environment requires ports to come up faster, configure explicit port speed and duplex settings.

For more information about PortFast, BPDU Guard, and TrunkFast, see the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.1E/native/configuration/guide/stp_enha.html.

By default, the Cisco IGESM port connected to the management module is in VLAN 1. Putting any of the blade servers in this VLAN causes broadcast traffic to show up in the Cisco IGESM CPU queue. This might overload the CPU and cause other undesired results. Also, as mentioned previously, the management module bridges traffic coming from its external port to both the Cisco IGESM switches.

Verifying Connectivity Between Cisco IGESMs

This section explains how to verify the connectivity between two Cisco IGESM switches in the same chassis.

To confirm the connectivity between the two switches, enter the **show cdp neighbor** command on either of the switches. The following shows sample output from this command:

```
Switch-1# show cdp neighbors Gigabit Ethernet 0/15 detail
-------------------------
Device ID: Switch-1.example.com
Entry address(es):
  IP address: 192.26.208.14
Platform: cisco OS-Cisco IGESM-18,  Capabilities: Switch IGMP
Interface: GigabitEthernet0/15,  Port ID (outgoing port): GigabitEthernet0/15
Holdtime : 134 sec

Version :
Cisco Internetwork Operating System Software
IOS (tm) Cisco IGESM Software (Cisco IGESM-I6Q4L2-M), Version 12.1(0.0.42)AY, CISCO DEVELO
PMENT TEST VERSION
Copyright (c) 1986-2004 by Cisco Systems, Inc.
Compiled Mon 08-Mar-04 10:20 by antonino
advertisement version: 2
Protocol Hello:  OUI=0x00000C, Protocol ID=0x0112; payload len=27, value=0000000
0FFFFFFFF010221FF000000000000000ED7EE0900FF0000
VTP Management Domain: 'blade-centers'
Native VLAN: 1
Duplex: full
```

Note Do not use the management VLAN for any blade server module. Always keep the management VLAN and the blade server VLANs separate.

Server Default Gateway

The default gateway for the blade servers is typically configured on a Layer 3 device, which can be a Firewall Service Module (FWSM) or the Multilayer Switch Feature Card (MSFC).

Assuming that the default gateway is on the MSFC, you can use HSRP to configure a highly available Layer 3 interface. The Layer 3 interface is on the aggregation switch. A Layer 3 interface exists on the aggregation switch for each server VLAN on the BladeCenter Cisco IGESM.

You also configure the server VLAN on the aggregate switch, and the trunks between the two aggregate switches carry this server VLAN. The VLANs on these aggregate switch trunks carry HSRP heartbeats between the active and standby HSRP interfaces. The HSRP configuration on aggregate1 (HSRP Active) is shown below.

```
interface Vlan10
 description BladeServerFarm1
 ip address 10.10.10.2 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.10.1
 standby 1 timers 1 3
 standby 1 priority 110
 standby 1 preempt delay minimum 60
 standby 1 authentication cisco
end

interface Vlan20
 description BladeServerFarm2
 ip address 10.10.20.2 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.20.1
 standby 1 timers 1 3
 standby 1 priority 110
 standby 1 preempt delay minimum 60
 standby 1 authentication cisco
end

The configuration on Aggregation-2 (HSRP Standby) is as shown below.
interface Vlan10
 description BladeServerFarm1
 ip address 10.10.10.3 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.10.1
 standby 1 timers 1 3
 standby 1 priority 100
 standby 1 preempt delay minimum 60
 standby 1 authentication cisco
end

interface Vlan20
 description BladeServerFarm2
 ip address 10.10.20.3 255.255.255.0
```

```
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.20.1
 standby 1 timers 1 3
 standby 1 priority 100
 standby 1 preempt delay minimum 60
 standby 1 authentication cisco
end
```

Changing Spanning Tree Path Cost

This section is applicable to the topology introduced in Figure 2-10 only.

Changing spanning tree parameters is not required in most cases. However, when dedicated ports are used for traffic monitoring, change the path cost to force the traffic through higher bandwidth links.

For example, to implement the topology shown in Figure 2-10, on the Switch-2 interface connected to the primary root, change the path cost to 8. Use the following commands in global configuration mode to change the path cost.

```
interface interface-id
spanning-tree cost cost
spanning-tree vlan vlan-id cost cost
```

To verify the path cost, enter the following commands:

```
Show spanning-tree interface interface-id
Show spanning-tree vlan vlan-id
```

Layer 2 Trunk Failover

The trunk failover feature may track an upstream port or a channel. To assign an interface to a specific link state group, use the following command in the interface configuration sub mode:

```
link state group <1-2> upstream
```

Note   Gigabit Ethernet interfaces 0/17–20 may be configured only as "upstream" devices.

Enable the Trunk Failover feature for the downstream ports of the internal blade server interfaces for a specific link state group.

```
interface GigabitEthernet0/1
 description blade1
 link state group <1-2> downstream

interface GigabitEthernet0/2
 description blade2
 link state group <1-2> downstream
```

Note   Gigabit Ethernet interfaces 0/1–14 may be configured only as "downstream" devices.

Globally enable the trunk failover feature for a specific link state group:

```
link state track <1-2>
```

To validate the trunk failover configuration, use the following command:

```
show link state group detai
```

Cisco Gigabit Ethernet Switch Module for the HP BladeSystem

This section provides best design practices for deploying the Cisco Gigabit Ethernet Switch Modules (CGESM) for the HP BladeSystem p-Class enclosures within the Cisco Data Center Networking Architecture. This document describes the internals of the blade enclosure and CGESM and explores various methods of deployment. It includes the following sections:

- Cisco Gigabit Ethernet Switching Module
- CGESM Features
- Using the HP BladeSystem p-Class Enclosure in the Data Center Architecture
- Design and Implementation Details

Cisco Gigabit Ethernet Switching Module

This section briefly describes the Cisco Gigabit Ethernet Switching Module (CGESM) and explains how the blade servers within the HP BladeSystem are physically connected to the switching module.

The CGESM provides enhanced Layer 2 services (known as L2+ or Intelligent Ethernet) to the HP BladeSystem p-Class. The CGESM extends the capabilities of a Layer 2 Ethernet switch to include Cisco proprietary protocols, ACLs, and QoS based on Layer 3 information. With SNMP, CLI, or HTTP management options available and a robust set of IOS switching features, the CGESM naturally integrates into the data center environment. The following features highlight this capacity:

- Loop protection and rapid convergence with support for Trunk Failover, Per VLAN Spanning Tree (PVST+), 802.1w, 802.1s, BDPU Guard, Loop Guard, PortFast, UplinkFast, and UniDirectional Link Detection (UDLD)
- Advanced management protocols, including Cisco Discovery Protocol (CDP), VLAN Trunking Protocol (VTP), and Dynamic Trunking Protocol (DTP)
- Port Aggregation Protocol (PAgP) and Link Aggregation Control Protocol (LACP) for link load balancing and high availability
- Support for authentication services, including RADIUS and TACACS+ client support
- Support for protection mechanisms, such as limiting the number of MAC addresses allowed, or shutting down the port in response to security violations

The HP BladeSystem p-Class enclosure consists of eight server bays and two network-interconnect bays. Figure 2-13 shows the BladeSystem p-Class architecture using two CGESMs housed in the network interconnect bays and eight BL20P G3 servers.

Figure 2-13   BladeSystem p-Class Switch Architecture
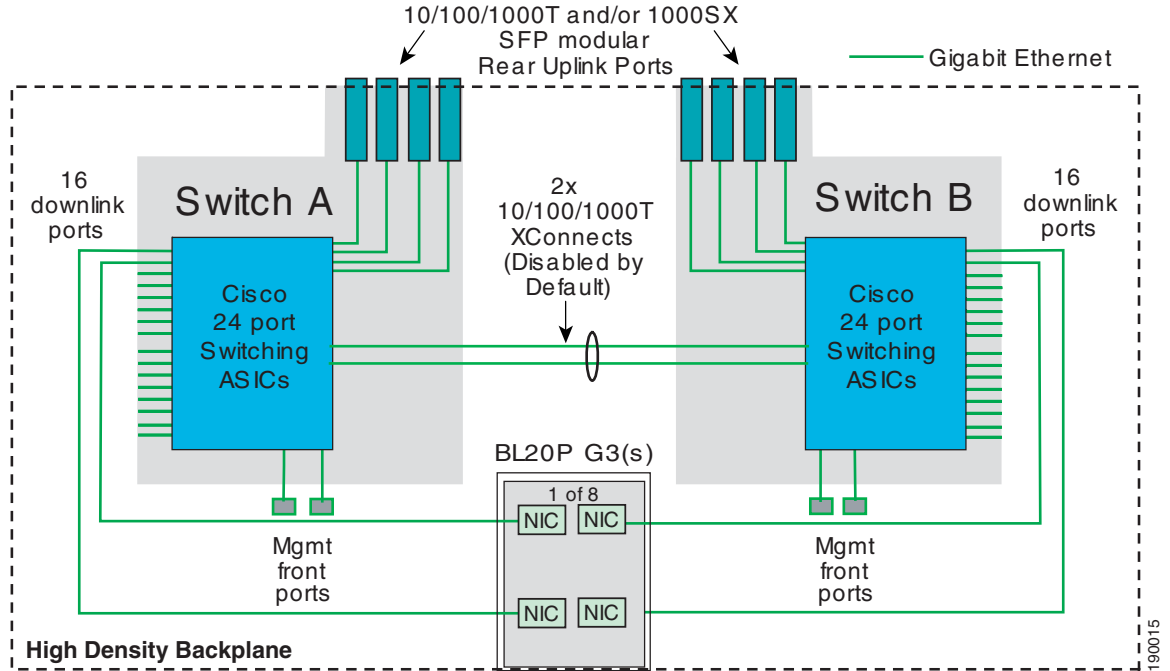
The HP BladeSystem p-Class backplane provides power and network connectivity to the blades. The interconnect bays house a pair of CGESMs, which provide a highly available and multi-homed environment where each server blade is Gigabit attached to each CGESM. Figure 2-14 illustrates how the HP BladeSystem p-Class logically provides Ethernet connectivity.

Figure 2-14   Blade Enclosure Ethernet Connectivity
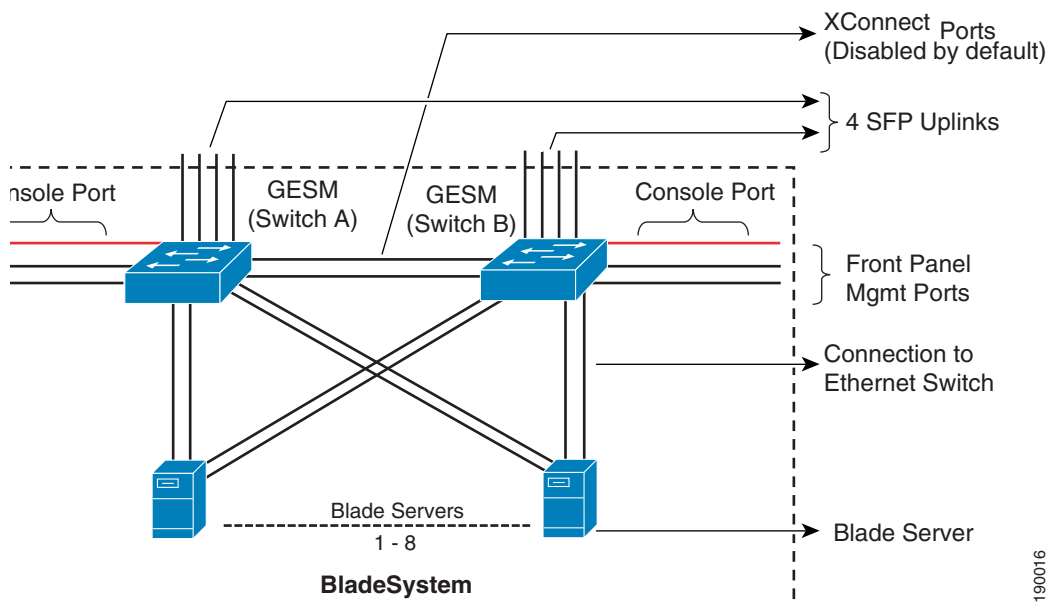
Note   Figure 2-14 is based on the use of the HP BladeSystem p-Class using BL20P G3 servers. The remainder of this document uses the BL20P G3 server for all figures.

In Figure 2-14, two CGESMs within the blade enclosure connect the blade server modules to external network devices such as aggregation layer switches. Each Ethernet switch provides six external Ethernet ports for connecting the blade enclosure to the external network. Four SFP ports provide 1000 Base-SX and 10/100/1000 Base-T links on the rear of the enclosure and two 10/100/1000 Base-T ports provide connectivity on the front panel. All six of these ports can be grouped to support the 802.3ad link aggregation protocol. In Figure 2-14 above, each blade server is connected to the backplane via the available Gigabit Ethernet network interface cards (NICs). The number of NICs on each blade server varies. Table 2-2 provides more detail on the connectivity options available with each HP blade server and the maximum number of blade servers a single enclosure can support.

Note   This list is not comprehensive; more detail on HP blade server connectivity can be found at the following URL: http://www.hp.com.

Table 2-2   Blade Server Options

Blade Server	Maximum Number of Server Blades per Enclosure	NICs Available
BL20P G2	8	3 10/100/1000T NICs 1 dedicated iLO interface
BL20P G3	8	4 10/100/1000T NICs 1 dedicated iLO interface
BL30P	16	2 10/100/1000T NICs 1 dedicated iLO interface
BL40P	2	5 x 10/100/1000T NICs 2 Slots for SAN connectivity 1 dedicated iLO interface

Note   In Table 2-2, "iLO" refers to the Integrated Lights-Out interface. It supports the iLO management subsystem that resides on each server blade. For more information on the iLO system, see Management, page 2-43.

In Figure 2-13 and Figure 2-14, two NICs on each blade server connect to CGESM A and CGESM B. The blade servers connect to the CGESM switches over the HP BladeSystem p-Class backplane. There are sixteen 10/100/1000 internal ports on each CGESM dedicated to the blade server modules.

Figure 2-13 and Figure 2-14 also show two internal 10/100/1000 ports interconnecting the two blade enclosure switches over the backplane. These ports are disabled by default, but are configurable to carry all traffic types between the two blade enclosure switches. These ports support trunking and can be configured as channels.

Note   In Figure 2-13, if the BL20p G1 and G2 servers are used, each server dedicates one NIC connected to the CGESM B side to the iLO port. This NIC is capable only of 100 MB. The enclosure with an enhanced backplane provides blade server connectivity to an embedded iLO module on the backplane. See Management, page 2-43 for more details.

The HP BladeSystem p-Class enclosure with enhanced backplane consists of eight server bays and two network-interconnect bays. The HP BladeSystem p-Class sleeve option allows the enclosure to support 16 Proliant BL30p servers or two BL30p servers per bay. Figure 2-15 illustrates the Gigabit Ethernet connectivity of an enhanced backplane enclosure and 16 BL30p servers.

Figure 2-15   HP BladeSystem p-Class with 16 Servers
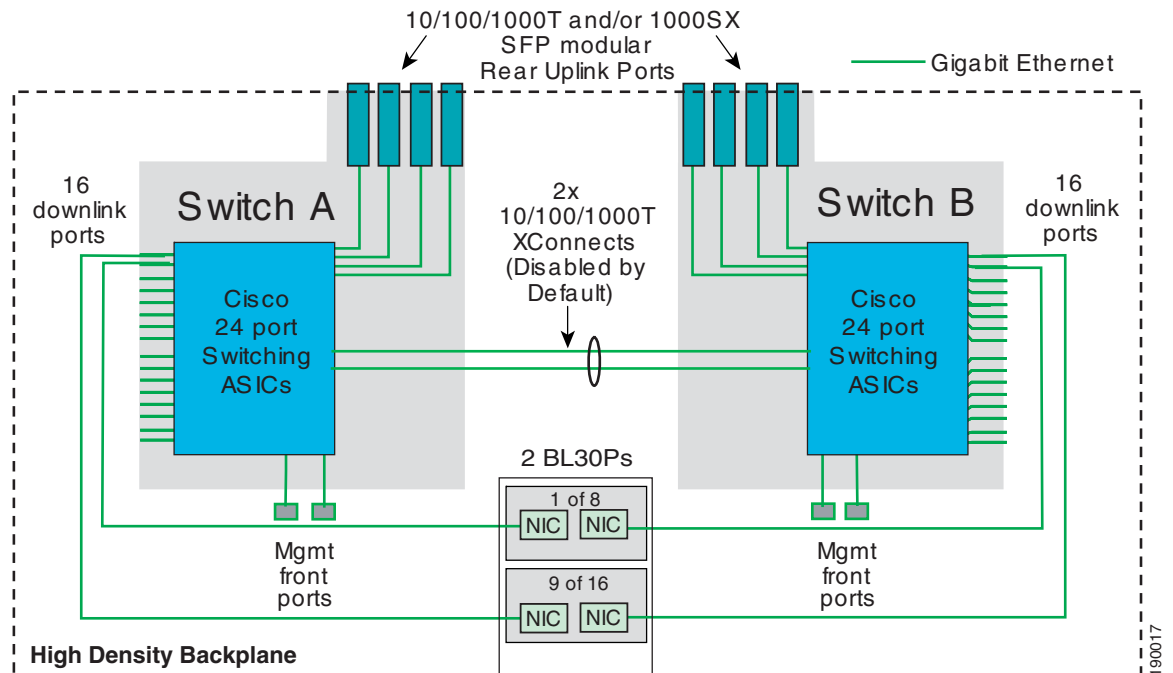
For more information about the HP BladeSystem p-Class, see the following URL:
http://h18004.www1.hp.com/products/servers/proliant-bl/p-class/documentation.html.

CGESM Features

This section provides information about the protocols and features provided by the CGESM that help to integrate the HP BladeSystem p-Class enclosure into the Cisco Data Center Network Architecture. This section includes the following topics:

- Spanning Tree
- Traffic Monitoring
- Link Aggregation Protocols
- Layer 2 Trunk Failover

Spanning Tree

The CGESM supports various versions of the Spanning Tree Protocol (STP) and associated features, including the following:

- Rapid Spanning Tree (RSTP) based on 802.1w
- Multiple Spanning Tree (MST) based on 802.1s with 802.1w
- Per VLAN Spanning Tree Plus (PVST+)
- Rapid Per VLAN Spanning Tree Plus (RPVST+)
- Loop Guard
- Unidirectional Link Detection (UDLD)
- BPDU Guard
- PortFast
- UplinkFast (Cisco proprietary enhancement for 802.1d deployments)
- BackboneFast (Cisco proprietary enhancement for 802.1d deployments)

The 802.1w protocol is the standard for rapid spanning tree convergence, while 802.1s is the standard for multiple spanning tree instances. Support for these protocols is essential in a server farm environment for allowing rapid Layer 2 convergence after a failure occurs in the primary path. The key benefits of 802.1w include the following:

- The spanning tree topology converges quickly after a switch or link failure.
- Convergence is accelerated by a handshake, known as the proposal agreement mechanism.

Note   There is no need to enable BackboneFast or UplinkFast

In terms of convergence, STP algorithms based on 802.1w are much faster than the traditional STP 802.1d algorithms. The proposal agreement mechanism allows the CGESM to decide new port roles by exchanging proposals with its neighbors.

With 802.1w, as with other versions of the STP, BPDUs are by default sent every two seconds (called the *hello time*). If three BPDUs are missed, STP recalculates the topology, which takes less than one second for 802.1w.

This seems to indicate that STP convergence time can be as long as six seconds. However, because the data center is made of point-to-point links, the only failures are physical failures of the networking devices or links. 802.1w is able to actively confirm that a port can safely transition to forwarding without relying on any timer configuration. This means that the actual convergence time is below *one second* rather than six seconds.

A scenario where BPDUs are lost may be caused by unidirectional links, which can cause Layer 2 loops. To prevent this problem, you can use Loop Guard and UDLD. Loop Guard prevents a port from forwarding as a result of missed BPDUs, which might cause a Layer 2 loop that can bring down the network.

UDLD allows devices to monitor the physical configuration of fiber optic or copper Ethernet cables and to detect when a unidirectional link exists. When a unidirectional link is detected, UDLD shuts down the affected port and generates an alert. BPDU Guard prevents a port from being active in a spanning tree topology as a result of an attack or misconfiguration of a device connected to a switch port. The port that sees unexpected BPDUs is automatically disabled and must then be manually enabled. This gives the network administrator full control over port and switch behavior.

The CGESM supports PVST and a maximum of 128 spanning tree instances. RPVST+ is a combination of Cisco PVST Plus (PVST+) and Rapid Spanning Tree Protocol. RPVST+ provides the flexibility of one spanning tree instance per VLAN and fast convergence benefits of 802.1w. Multiple Spanning Tree (MST) allows the switch to map several VLANs to one spanning tree instance, reducing the total number of spanning tree topologies the switch processor must manage. A maximum of 16 MST instances are supported. In addition, MST uses 802.1w for rapid convergence. MST and RPVST+ create a more predictable and resilient spanning tree topology, while providing downward compatibility for integration with devices that use 802.1d and PVST+ protocols.

Note   The 802.1w protocol is enabled by default when running spanning tree in RPVST+ or MST mode on the CGESM. CGESM enables PVST+ for VLAN 1 by default.

Spanning tree uses the path cost value to determine the shortest distance to the root bridge. The port path cost value represents the media speed of the link and is configurable on a per interface basis, including EtherChannels. To allow for more granular STP calculations, enable the use of a 32-bit value instead of the default 16-bit value. The *longer* path cost better reflects changes in the speed of channels and allows STP to optimize the network in the presence of loops.

Note   The CGESM supports IEEE 802.1t, which allows for spanning tree calculations based on a 32-bit path cost value instead of the default 16 bits. For more information about the standards supported by the CGESM, see *Cisco Gigabit Ethernet Switch Module (CGESM) Overview.*

Traffic Monitoring

The CGESM supports the following traffic monitoring features, which are useful for monitoring blade enclosure traffic in data center environments:

- Switched Port Analyzer (SPAN)
- Remote SPAN (RSPAN)

SPAN mirrors traffic transmitted or received on source ports or source VLANs to another local switch port. This traffic can be analyzed by connecting a switch or RMON probe to the destination port of the mirrored traffic. Only traffic that enters or leaves source ports or source VLANs can be monitored using SPAN.

RSPAN enables remote monitoring of multiple switches across your network. The traffic for each RSPAN session is carried over a user-specified VLAN that is dedicated for that RSPAN session for all participating switches. The SPAN traffic from the source ports or source VLANs is copied to the RSPAN VLAN. This mirrored traffic is then forwarded over trunk ports to any destination session that is monitoring the RSPAN VLAN.

Note   RSPAN does not require a dedicated reflector port to mirror traffic from either a source port or source VLAN.

Link Aggregation Protocols

Fast EtherChannel (FEC) and Gigabit EtherChannel (GEC) are logically bundled physical interfaces that provide link redundancy and scalable bandwidth between network devices. Port Aggregation Protocol (PAgP) and Link Aggregation Control Protocol (LACP) help automatically create these channels by exchanging packets between Ethernet interfaces and negotiating a logical connection. PAgP is a

Cisco-proprietary protocol that can be run only on Cisco switches or on switches manufactured by vendors that are licensed to support PAgP. LACP is a standard protocol that allows Cisco switches to manage Ethernet channels between any switches that conform to the 802.3ad protocol. Because the CGESM supports both protocols, you can use either 802.3ad or PAgP to form port channels between Cisco switches.

When using either of these protocols, a switch learns the identity of partners capable of supporting either PAgP or LACP and identifies the capabilities of each interface. The switch dynamically groups similarly configured interfaces into a single logical link, called a channel or aggregate port. The interface grouping is based on hardware, administrative, and port parameter attributes. For example, PAgP groups interfaces with the same speed, duplex mode, native VLAN, VLAN range, trunking status, and trunking type. After grouping the links into a port channel, PAgP adds the group to the spanning tree as a single switch port.

Layer 2 Trunk Failover

Layer 2 Trunk failover is a high availability mechanism that allows the CGESM to track and bind the state of external interfaces with one or more internal interfaces. The four available Gigabit Ethernet uplink ports of the CGESM provide connectivity to the external network and are characterized as "upstream" links. The trunk failover feature may track these upstream interfaces individually or as a port channel. Trunk failover logically binds upstream links together to form a link state group. The internal interfaces of the CGESM provide blade server connectivity and are referred to as "downstream" interfaces in the trunk failover configuration. This feature creates a relationship between the two interface types where the link state of the "upstream" interfaces defined in a link state group determines the link state of the associated "downstream" interfaces.

Figure 2-16 illustrates the logical view of trunk failover on the CGESM. The two external port channels of Switch-1 and Switch-2 are configured as upstream connections in a link state group local to the switch. The 16 internal blade server ports are downstream interfaces associated with each local group.

Figure 2-16   Trunk Failover Logical View
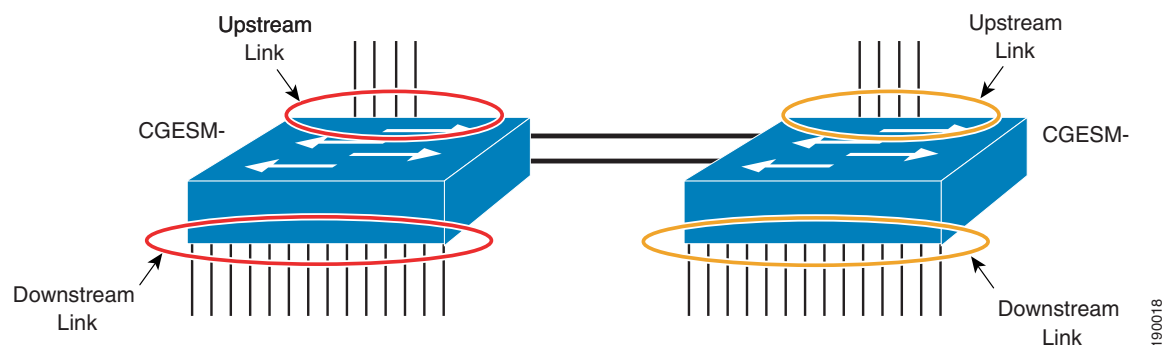
Trunk failover places downstream devices into the same link state, "up" or "down", based on the condition of the link state group. If an uplink or upstream failure occurs, the trunk failover feature places the downstream ports associated with those upstream interfaces into a link "down" or inactive state. When upstream interfaces are recovered, the related downstream devices are placed in an "up" or active state. An average failover and recovery time for network designs implementing the trunk failover feature is three seconds.

Consider the following when configuring the trunk failover on the CGESM:

 • Internal ports (Gigabit Ethernet 0/1–16) may not be configured as "upstream" interfaces.

- External ports (Gigabit Ethernet 0/19–24) may not be configured as "downstream" interfaces.

- Trunk failover does not consider STP. The state of the upstream connections determines the status of the link state group, not the STP state forwarding, blocking, and so on.

- Trunk failover of port channels requires that all of the individual ports of the channel fail before a trunk failover event is triggered.

- Interfaces cannot be members of more than one link state group.

- Do not configure an Etherchannel as a downstream interface.

- SPAN/RSPAN destination ports are automatically removed from the trunk failover link state groups.

- The CGESM is capable of defining two link state groups. This flexibility allows the administrator to define two different failover conditions for the downstream blade servers. This may be necessary when a server is not using NIC teaming, or two different uplink paths are defined on the switch.

Using the HP BladeSystem p-Class Enclosure in the Data Center Architecture

The HP BladeSystem p-Class enclosure supports a maximum of two internal CGESM enhanced Layer 2 switches. Each switch provides 16 internal Gigabit Ethernet ports to support the blade servers within the enclosure. The HP BladeSystem p-Class supports up to eight blade servers (see Figure 2-13), each having multiple ProLiant NC series NICs (see Table 2-2). Note that the enclosure with enhanced backplane supports up to 16 blade servers (see Figure 2-15). Figure 2-17 and Figure 2-18 illustrate the physical layout of the chassis. The two interconnect bays house the CGESM switches that are connected via two 10/100/1000 cross connects on the backplane. Each switch also has separate dual-backplane connections to the individual server blade bays. This indicates that each server blade is dual-homed to the two internal switches.

Figure 2-17   HP BladeSystem p-Class Enclosure with Proliant BL20p G3 Servers
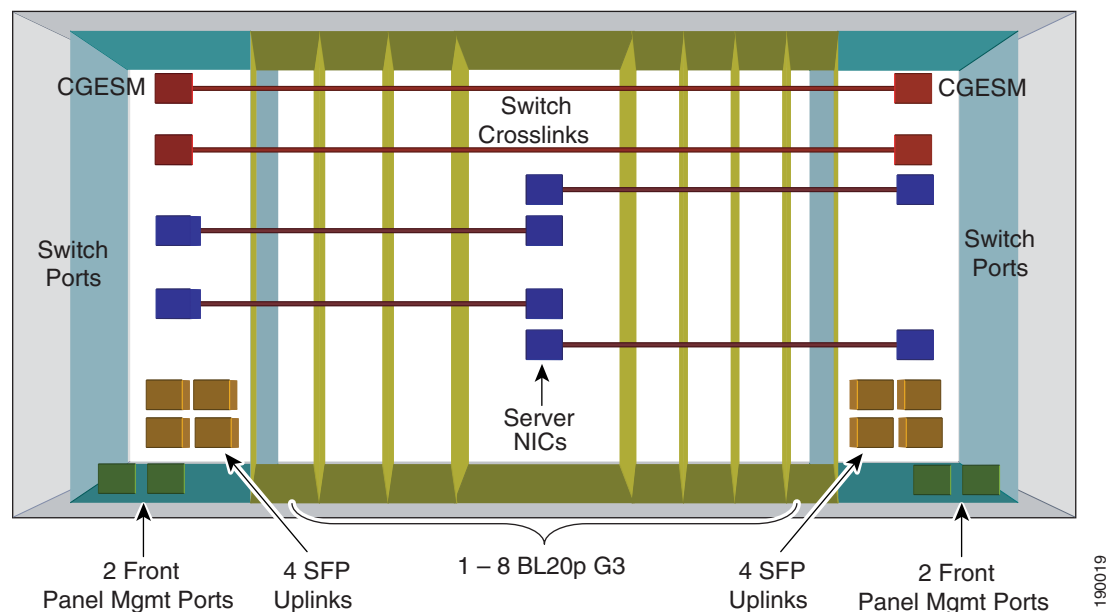
Figure 2-18   *HP BladeSystem p-Class Enclosure with Proliant BL30p Servers*
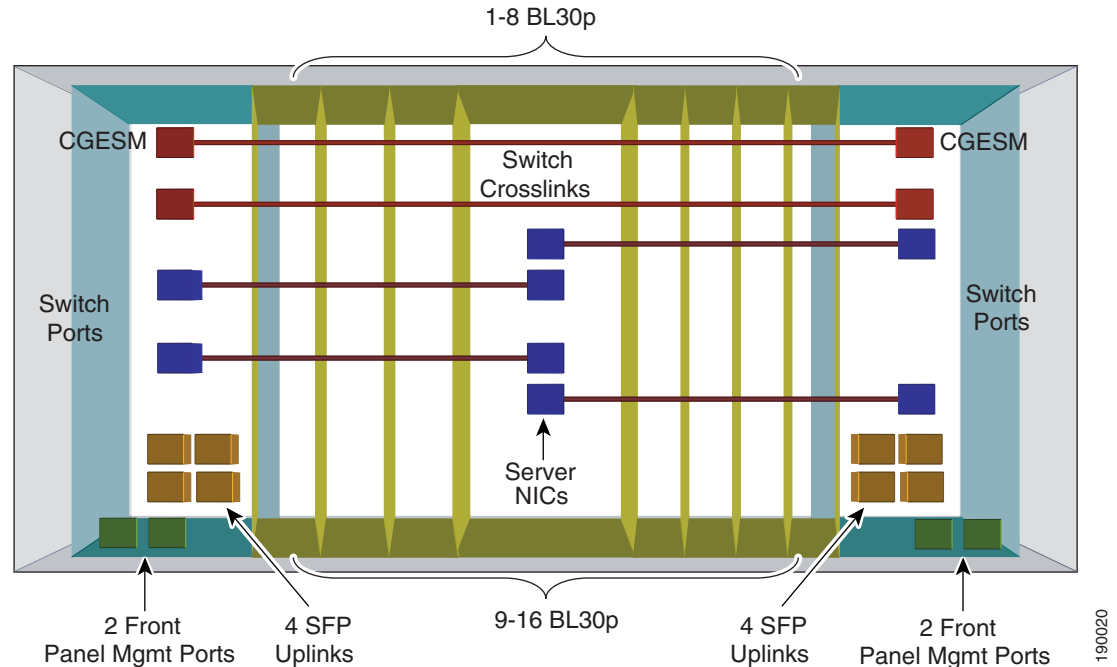
Each CGESM has four external SFP ports supporting 1000Base-SX and 10/100/1000Base-T on the rear of the enclosure and two external 10/100/1000BaseT ports on the front panel. These six ports provide connectivity to the data center or other external network. For more information, see the HP Blade Systems at the following URL:

http://h71028.www7.hp.com/enterprise/cache/80632-0-0-0-121.aspx#Servers

This section describes the design goals when deploying blade servers and the functionality supported by the CGESM in data centers. It includes the following topics:

- High Availability

- Scalability

- Management

High Availability

Data centers are the repository of critical business applications that support the continual operation of an enterprise. Some of these applications must be accessible throughout the working day during peak times, and others at all times. The infrastructure of the data center, network devices, and servers must address these diverse requirements. The network infrastructure provides device and link redundancy combined with a deterministic topology design to achieve application availability requirements. Servers are typically configured with multiple NIC cards and dual-homed to the access layer switches to provide backup connectivity to the business application.

High availability is an important design consideration in the data center. An HP BladeSystem p-Class, using the CGESM, has a number of features and characteristics that contribute to a reliable, highly available network.

High Availability for the Blade Enclosure Switching Infrastructure

High availability between the HP BladeSystem p-Class CGESMs and the aggregation layer switches requires link redundancy. Each CGESM in the HP BladeSystem p-Class uses four SFP uplinks for connectivity to the external network, which allows for redundant paths using two links each for more redundancy. Redundant paths implemented between the HP BladeSystem p-Class and each aggregation layer switch when each path uses two links provide a highly resilient design. However, this introduces the possibility of Layer 2 loops; therefore, a mechanism is required to manage the physical topology. The implementation of Rapid Spanning Tree Protocol (RSTP) ensures a fast converging, predictable Layer 2 domain between the aggregation layer and access switches (the CGESMs) when redundant paths are present.

The recommended design is a triangle topology, which delivers a highly available environment through redundant links and a spanning tree. It allows for multiple switch or link failures without compromising the availability of the data center applications.

As shown in Figure 2-19, each CGESM switch has two direct port channel connections to the Layer 2/Layer 3 aggregation layer where the primary STP root switch resides.

Figure 2-19   *Blade Enclosure Redundant Topology*
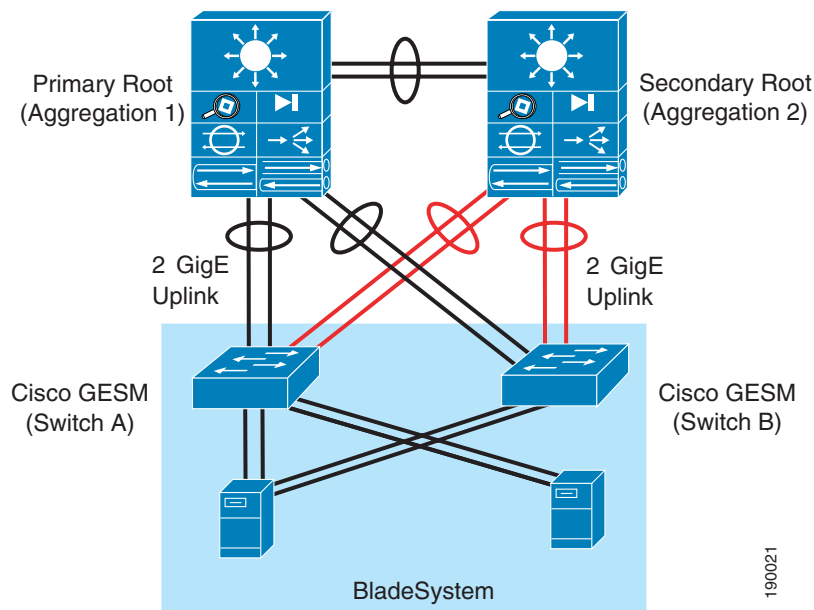
These channels support the publicly available subnets in the data center and traffic between servers. The server-to-server traffic that uses these uplinks is logically segmented through VLANs and may take advantage of network services available in the aggregation layer. There is also a port channel defined between the two blade enclosure switches. This path provides intra-chassis connectivity between the servers for VLANs defined locally on the blade enclosure switches. Clustering applications that require Layer 2 communication may take advantage of this traffic path, as well as mirrored traffic. Each of these port channels are composed of two-Gigabit Ethernet ports.

Cisco recommends using RPVST+ as the method for controlling the Layer 2 domain because of its predictable behavior and fast convergence. A meshed topology combined with RPVST+ allows only one active link from each blade enclosure switch to the root of the spanning tree domain. This design creates a highly available server farm through controlled traffic paths and the rapid convergence of the spanning tree.

The details of the recommended design are discussed in a later section.

High Availability for the Blade Servers

The HP BladeSystem p-Class enclosure provides high availability to blade servers by multi-homing each server to the CGESMs. The two CGESMs housed in the interconnect bays are connected to the blade server over the backplane. Four backplane Gigabit Ethernet connections are available to every blade-server slot.

Multi-homing the server blades allows the use of network adapter (NIC) teaming driver, which provides another high availability mechanism to failover and load balance at the server level. The ProLiant NC series NICs support three modes of teaming:

- Network Fault Tolerance (NFT)—Creates a virtual interface by grouping the blade server network adapters into a team. One adapter is the primary active interface and all other adapters are in a standby state. The virtual adapter uses a single MAC address and a single Layer 3 address. NFT provides adapter fault tolerance by monitoring the state of each team member network connection. The standby NICs become active only if the primary NIC loses connectivity to the network.

- Transmit Load Balancing (TLB)—Supports adapter fault tolerance (NFT) and adds more functionality in the server for load balancing egress (transmit) traffic across the team. Note that a TLB team uses only one NIC to receive traffic. The load balancing algorithm is based on either the destination MAC or IP address. This teaming method provides better use of the bandwidth available for egress traffic in the network than NFT.

- Switch Assisted Load Balancing (SLB)—Extends the functionality of TLB by allowing the team to receive load balanced traffic from the network. This requires that the switch can load balance the traffic across the ports connected to the server NIC team. The CGESM supports the IEEE 802.3ad standard and Gigabit port channels.

For more information about the ProLiant NIC teaming features, see the following URL:
http://h18000.www1.hp.com/products/servers/networking/whitepapers.html

Layer 2 Trunk Failover combined with NIC teaming provides a complete high availability solution in a blade server environment. Trunk Failover allows teamed NICs to converge by disabling downstream server ports when upstream network failures are detected. This systematic approach makes the dual-homed architecture of the HP BladeSystem even more robust.

Scalability

The ability of the data center to adapt to increased demands without compromising its availability is a key design consideration. The aggregation layer infrastructure and the services it provides must accommodate future growth in the number of servers or subnets it supports.

When deploying blade servers in the data center, there are the following two primary factors to consider:

- Number of physical ports in the aggregation and access layers
- Number of slots in the aggregation layer switches

Physical Port Count

The introduction of blade systems into the data center requires greater port density at the aggregation layer. Blade systems, deployed with internal switches, provide their own access layer. The cabling and maximum number of servers per enclosure is predetermined. Scaling the aggregation layer ports to accommodate the blade system uplinks is an area that requires attention.

As shown in Figure 2-19, each CGESM requires four Gigabit Ethernet ports from the aggregation layer switches. The number of physical ports that an aggregation-layer switch can support equals the number of ports per slot times the number of available slots.

It is important to remember that aggregation switches provide data center services such as load balancing, security, and network analysis that may require dedicated ports for appliances or slots for integrated services. This directly affects the number of ports available for access layer connectivity.

Table 2-3 lists the number of blade systems supported by a single line card with varying port counts. This table is based on the recommended topology shown in Figure 2-4, where each blade system is dual-homed to the aggregation layer over two Gigabit Ethernet port channels.

Table 2-3   Blade System Support per Aggregate Switch Line Card

Type of Line Card	Uplinks per CGESM	Total Uplinks /Blade System Enclosure (Two CGESM/Enclosure)	Blade Systems per Line Card
8-port Gigabit Ethernet	2	4	2
16-port Gigabit Ethernet	2	4	4
48-port Gigabit Ethernet	2	4	12

Table 2-3 highlights the finite number of BladeSystems supported by a single aggregate switch line card. This table implies that the aggregate layer must provide line card density for a scalable BladeSystem environment.

Slot Count

The data center infrastructure must be flexible enough to allow growth both in server capacity and service performance. Connecting a blade system directly into the aggregation layer places more significance on the number of slots available to accommodate blade system uplinks and integrated services.

Traditionally, the access layer provides the port density necessary to allow the physical growth of server farms. Modular access layer switches offer connectivity to densely packed server farms over a few uplinks. The aggregation layer switches support a limited number of uplinks from the access layer. With this model, the number of servers supported per uplink is high.

Blade systems use more aggregation layer resources per server than this traditional deployment model. Each uplink from a blade enclosure provides connectivity to a maximum of 16 servers. The aggregation layer must be flexible enough to manage the increased demand for ports and slots in this blade server system environment.

To scale the server farm, use an aggregation layer switch that provides an ample number of slots for line cards and/or service module expansion.

In addition, consider using the following two options (which are not mutually exclusive):

- Deploying service switches in the aggregation layer (as shown in Figure 2-20)
- Using a data center core to accommodate multiple aggregation layer modules

Service switches are deployed in the aggregation layer to host integrated data center services such as load balancing, intrusion detection, and network analysis. Relocating these services to a separate switch frees ports and slots in the aggregation layer switches. This design allows the aggregation switches to commit more slots and ultimately, more ports to the Layer 2 connectivity of the server farms. Figure 2-20 shows a service switch deployment.

Figure 2-20   Data Center Scaling with Service Switches
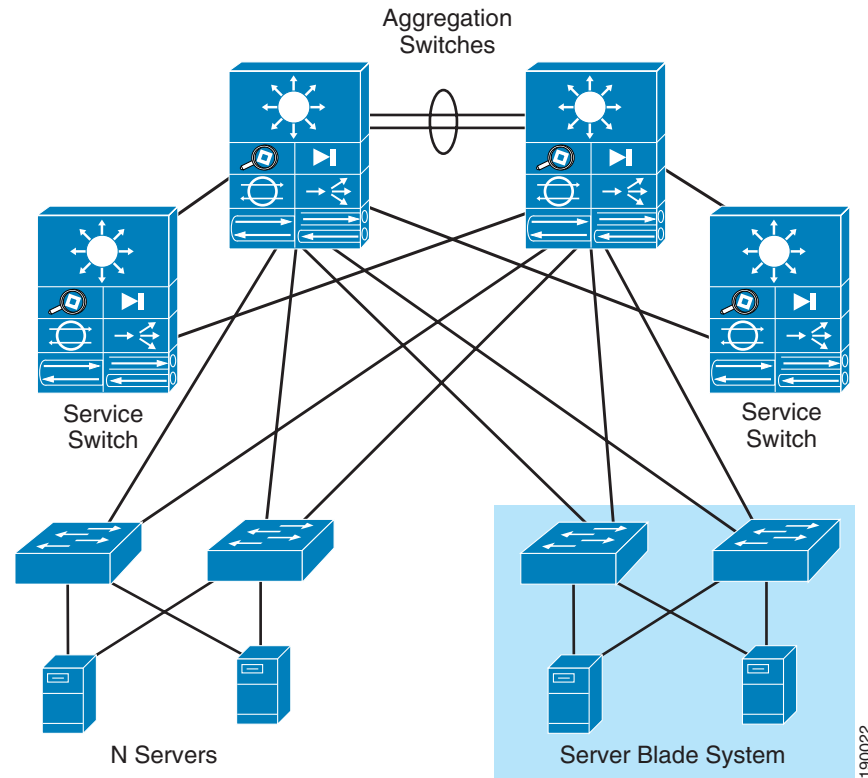
The data center core is a mechanism to replicate and horizontally scale the data center environment. In the recommended design, the aggregation and access layer is regarded as a module that can be duplicated to extend the enterprise. Each data center module provides its own network services locally in the aggregation switches. This approach allows the network administrator to determine the limits of each data center module and to replicate as necessary.

Figure 2-21 shows the data center core design. The aggregation switches for each data center module are Layer 3-attached to the core. In addition, the aggregation switches house the service modules required to support the server farms.

Figure 2-21   Data Center Core Design
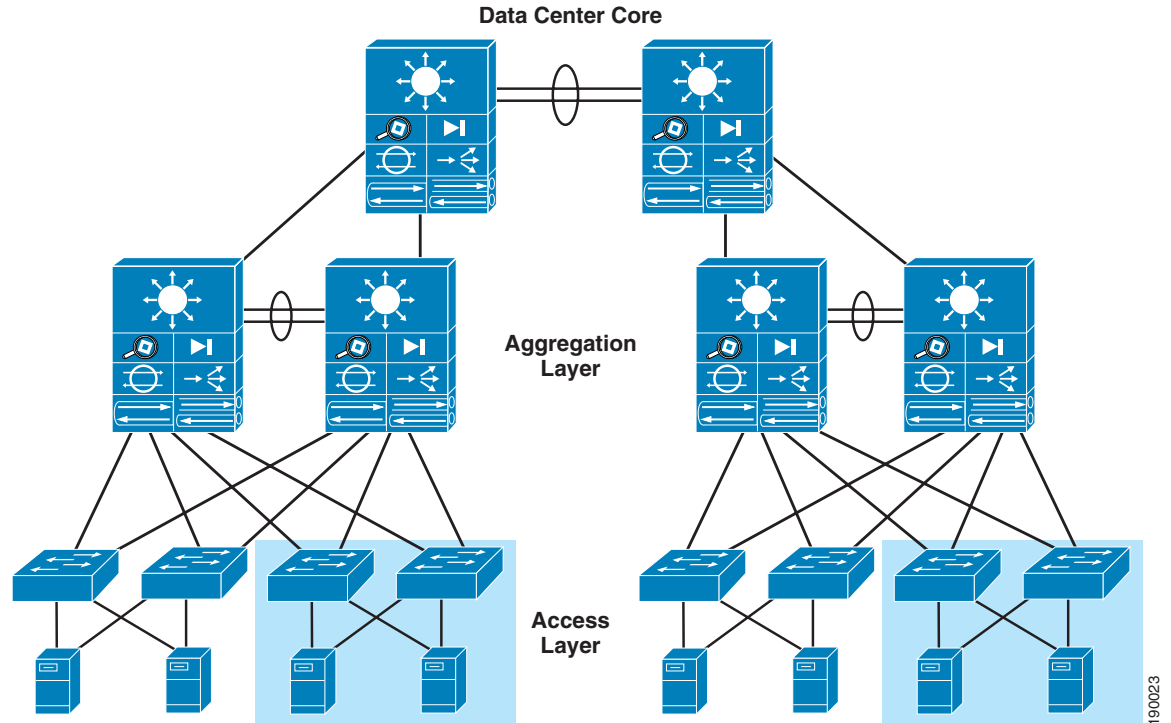
Management

The CGESM is accessible for management and configuration by any of the following traffic paths:

- Out-of-band management
- In-band management
- Serial/console port

These traffic paths provide three management options for network administration and support various user and application interfaces to the CGESM. The remote management of the blade servers within the HP BladeSystem p-Class enclosure is critical to an efficient and scalable data center. The iLO connectivity options provided via the enclosure to the blade servers are also discussed.

Out-of-Band Management

Out-of-band management is the practice of dedicating an interface on the managed device for carrying management traffic. It is also the recommended management method for HP BladeSystems. Out-of-band management isolates the management and data traffic and provides a more secure environment.

Figure 2-22 illustrates the interfaces available for connectivity. Cisco recommends using the front panel ports for connectivity to the management domain.

Figure 2-22   Blade Enclosure
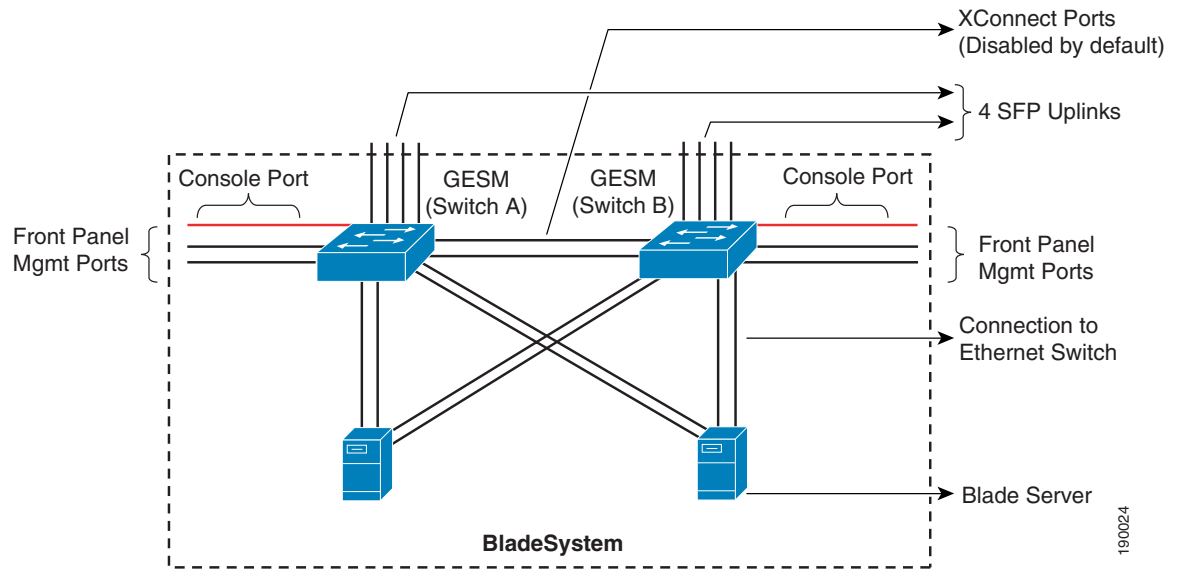
The CGESM allows only one switched virtual interface (SVI) to be active. By default, the SVI is created as VLAN 1 and it is disabled in an administratively down state. Cisco recommends that a VLAN other than VLAN 1 be used as the management VLAN. Therefore, it is important to create an SVI with another VLAN and to allow this VLAN on the external front panel ports.

For best practices in selecting the management VLAN, see the following URL:
http://www.cisco.com/en/US/products/hw/switches/ps700/products_white_paper09186a00801b49a4.shtml

In-Band Management

In-band management uses logical isolation to separate management traffic from data traffic. VLANs segregate the two traffic types that are sharing the bandwidth of the uplink ports. This practice is common where applications running on the servers must be managed along with the network infrastructure devices.

In-band management traffic uses the uplink trunk ports located on the rear of the CGESMs for management. Cisco recommends using a VLAN other than VLAN 1 for management.

Serial/Console Port

The front panel of the CGESM has a single RJ-45 serial port that can be used to manage the switch through the command-line interface (CLI). The CLI can be accessed by connecting directly to the console port with the serial port of a workstation or remotely by using terminal servers and IP connectivity protocols such as Telnet.

Management Options

The CGESM switch is manageable through the following methods:

- HTTP-based device manager GUI
- SNMP-based management applications
- Cisco IOS software CLI

The embedded device manager on the CGESM provides a GUI interface to configure and monitor the switch through a web browser. This requires using either in-band or out-of-band management and enabling the HTTP/HTTPS server on the switch. The HTTP server and SSL are enabled by default.

SNMP-compatible management utilities are supported through a comprehensive set of MIB extensions and through four remote monitoring (RMON) groups. CiscoWorks2000 and HP OpenView are two such management applications. SNMP versions 1, 2, and 3 are available on the switch (Cisco IOS software crypto image).

The CLI delivers the standard Cisco IOS software interface over Telnet or the console port. The use of SSH for CLI access is recommended.

Note   For more information about the embedded device manger, see the online help on the switch CLI.

For more information about Cisco MIBs, see the following URL:
http://www.cisco.com/public/sw-center/netmgmt/cmtk/mibs.shtml

For more information about the management options for the HP BladeSystem, see the following URL:
http://h18004.www1.hp.com/products/blades/components/management.html

HP BladeSystem p-Class iLO Connectivity

The iLO provides remote management capabilities and is standard with BL p-Class server blades. Remote power, console, and diagnostics are just a few of the advanced functions iLO provides. Table 2-2 shows that each of the blade servers supports a dedicated iLO NIC. The HP BladeSystem p-Class enclosure provides two methods to access this management interface:

- Through the CGESM located in interconnect bay B
- Through an enhanced BladeSystem enclosure

An HP BladeSystem p-Class enclosure without the enhanced backplane features provides connectivity to the dedicated iLO NIC through the CGESM located in interconnect bay B. The iLO NIC auto-negotiates to 100 Mbps and uses one of the CGESM ports assigned to that server bay. Figure 2-23 illustrates the iLO interface connection in the absence of an enhanced backplane.

Figure 2-23   *HP BladeSystem p-Class iLO Connectivity*
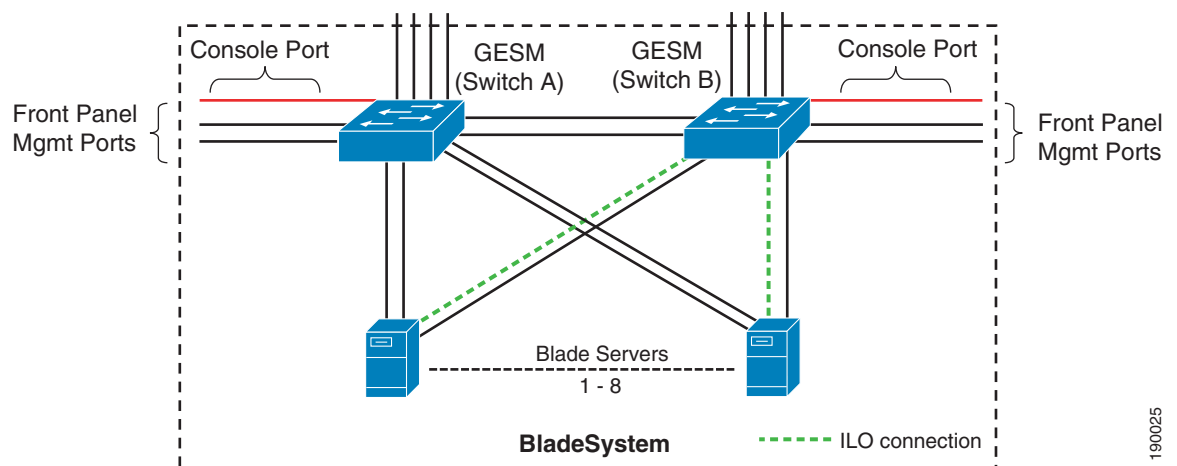
The enhanced backplane of the HP BladeSystem p-Class enclosure allows each server to use a dedicated Ethernet port for iLO connectivity. As shown in Figure 2-24, the iLO connection is independent of the CGESM. The blade server management module located on the rear of the chassis provides access to each of the iLO interfaces through a single RJ-45 cable.

Figure 2-24   HP BladeSystem pClass with Enhanced Backplane iLO Connectivity
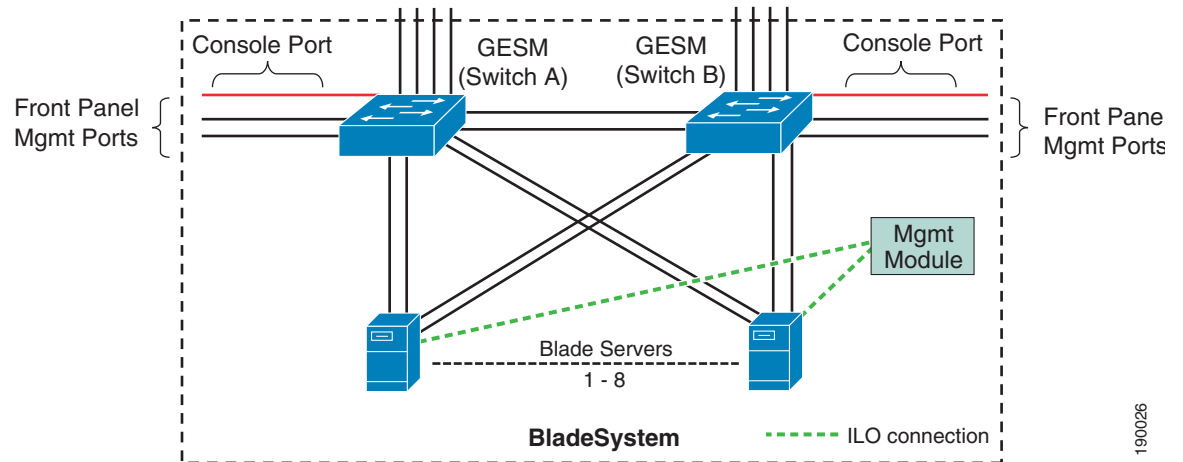
Note   The Proliant BL30p server blade requires the use of an enhanced backplane.

Design and Implementation Details

Network Management Recommendations

An out-of-band (OOB) network is recommended for managing the CGESM switch. OOB management provides an isolated environment for monitoring and configuring the switch. Isolation is achieved by deploying a physically separate management network or by logically separating the traffic with management VLANs.

The CGESM switch has two external Gigabit Ethernet ports located at the front of the chassis that may be used to support network monitoring devices and network management traffic. The use of secure protocols, such as SSH or HTTPS, maintains the integrity of communications between the switch and the management station. The console port positioned at the front of the CGESM is another option for connectivity to the OOB network.

Network Topologies using the CGESM

The following physical topologies are discussed in this section:

- Layer 2 Looped Access Layer Design—Classic "V"
- Layer 2 Looped Access Layer Design—"Square"
- Layer 2 Loop-Free Access Layer Design—Inverted "U"

These network designs emphasizes high availability in the data center by eliminating any single point of failure, and by providing deterministic traffic patterns, and predictable behavior during times of network convergence. The configuration example included uses a pair of Cisco Catalyst 6513 as the aggregation layer platform. This Layer 2/Layer 3 switching platform supports slot density and integrated network services required by data centers deploying blade systems. An HP BladeSystem p-Class server blade enclosure with two CGESMs composes the Layer 2 access layer.

Layer 2 Looped Access Layer Design—Classic "V"

Figure 2-25 shows a blade system deployment in the data center using the classic triangle topology. There is no single point of failure in this deployment model. The CGESMs are dual-homed to the aggregation layer, providing link redundancy. STP manages the physical loops created by the uplinks between the aggregation and access switches, assuring a predictable and fast converging topology. In Figure 2-25, the black links are in spanning tree forwarding state and the red links are in blocking state. The green links represent the internal cross connects that are disabled by default.

Figure 2-25   Recommended Topology HP BladeSystem p-Class with CGESMs
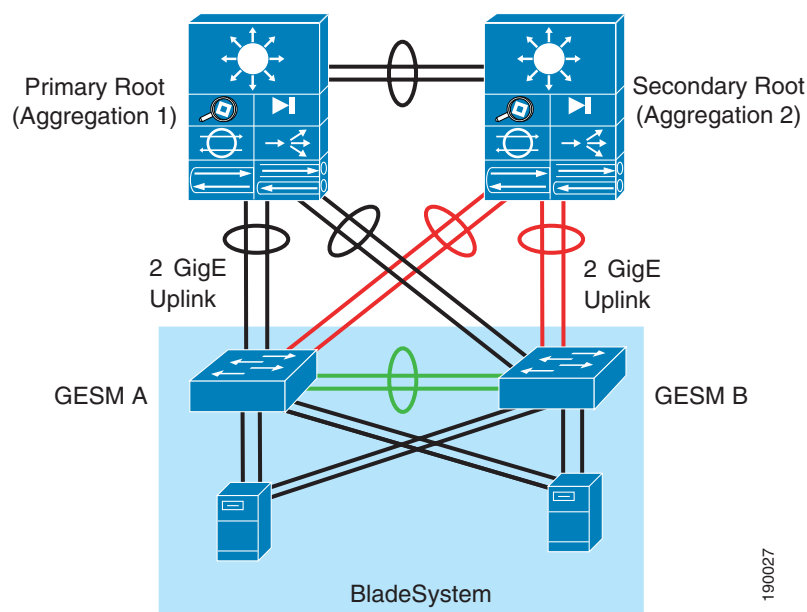
RPVST+ fulfills the high availability requirements of this design and is the recommended mode of spanning tree operation. RPVST+ provides fast convergence (less than one second) in device or uplink failure scenarios. In addition, RPVST+ offers enhanced Layer 2 features for the access layer with integrated capabilities equivalent to UplinkFast and BackboneFast.

The connection between the two internal blade switches in Figure 2-25 supports local traffic limited to the HP BladeSystem; for example, clustering applications, or management traffic such as remotely mirrored (RSPAN) traffic. This connection does not carry a publicly accessible subnet (for example, a VLAN that exists on the uplinks to the aggregation switches). If this were the case, another set of interfaces would have to be accounted for in the STP calculations. Therefore, to create a less complex STP domain, these cross-connect interfaces are removed from the equation by clearing the public VLANs from the links.

The HP BladeSystem p-Class server blade NICs support the logical separation of VLANs via trunking. This allows each NIC to accommodate the public and the private VLANs on the CGESMs. In addition, servers such as the BL20P G3 series are dual-homed to each of the two CGESMs in the HP BladeSystem enclosure (see Figure 2-19). This structural design allows for the physical separation of public and private VLANs between two NICs homed to the same CGESM.

A series of network convergence tests were performed to verify and characterize the high availability features of the recommended design. These tests consisted of passing traffic between an external client device and the blade servers while monitoring packet loss. The following test cases were used:

- Uplink failure and recovery between Switch-A and the primary root
- Uplink failure and recovery between Switch-B and the primary root
- Switch-A failure and recovery
- Switch-B failure and recovery
- Primary root switch failure and recovery
- Secondary root switch failure and recovery

These tests revealed the intricacies of fast convergence in the data center and the necessity for a holistic approach to high availability.

Test cases that did not involve the failure of the active HSRP aggregation switch resulted in an average failover time of about one second. Failing the active HSRP device requires convergence at Layer 3 and resulted in a recovery time that reflected the settings of the HSRP timers.

It is possible to tune the HSRP timers for sub-second convergence. However, when multiple HSRP devices are involved, the recovery time is typically in the five-second range.

In this topology, two Gigabit Ethernet links comprise the port channel uplinks between the access and aggregation layers. This configuration allows a single link to fail without triggering STP convergence.

Note The default gateway for the servers is the HSRP address of the Layer 3 aggregation switches. Failover times may be affected if the default gateway of the server is located on another device, such as a load balancer or firewall.

The recommended topology provides a high level of availability to the blade servers except in one failure scenario. If both the uplinks to each of the aggregation switches from a single CGESM are unavailable, the server NICs homed to that CGESM are not notified. The blade servers are unaware of the disconnection between the access layer switches (CGESMs) and the aggregation layer switches and continue to forward traffic. To address this breakdown in network connectivity, use one of the following methods:

- Use the NIC teaming features of the ProLiant blade servers with Layer 2 Trunk Failover.
- Deploy load balancing in front of the blade server farm.

In addition, the NIC teaming features of the ProLiant blade servers provide redundancy at the network adapter level. Stagger the preferred primary NICs between the two Cisco switches in the enclosure to increase server availability. Assigning the primary NIC is a straightforward process. The NIC teaming

software provides a GUI interface or a small configuration file, depending on the operating system, to construct the team. HP also offers network-aware teaming software to verify and detect network routes. For more information about these features, see the ProLiant Essential Intelligent Network Pack at the following URL: http://h18004.www1.hp.com/products/servers/proliantessentials/inp/index.html

The ProLiant blade server NIC teaming functionality combined with the Cisco Layer 2 Trunk Failover feature provides a comprehensive high availability solution to the blade servers. On detection of an upstream link failure from the CGESM, the associated downstream server ports are disabled by the CGESM, which allows the NIC team to redirect traffic over the remaining active NICs in the team homed to another CGESM. Multiple link state groups may be defined on the CGESM to allow for redundant uplink paths.

By monitoring the health of a server farm, a load balancer can bypass the network failure by redirecting traffic to available servers. This helps ensure fulfillment of end user requests despite the network failure.

The recommended network topology illustrated in Figure 2-25 allows for traffic monitoring either locally or remotely using Switched Port Analyzer (SPAN). Local SPAN supports monitoring of network traffic within one switch, while remote SPAN (RSPAN) allows the destination of mirrored traffic to be another switch within the data center. The source of mirrored traffic for a SPAN or RSPAN session may be one or more ports or VLANs.

Local SPAN is readily supported by the CGESM over one of the two external Gigabit Ethernet ports located on the front panel of the switch. This RJ-45 connection is an ideal location to attach intrusion detection or other network analysis device.

RSPAN requires a VLAN to carry the mirrored traffic to the remote destination switch. In the recommended topology, the secondary aggregation switch is the RSPAN destination, where an analysis device, such as the integrated Network Analysis Module (NAM), resides.

Figure 2-26 illustrates the traffic path of the RSPAN VLAN. The RSPAN VLAN uses the uplink between the blade switch and the secondary aggregation switch. This uplink is blocking under normal conditions for regular VLANs. As a result, bandwidth utilization is only a concern when the uplink is forwarding and sharing the path with production traffic.

Figure 2-26   RSPAN Traffic Path

Configuring the Aggregate Switches

Complete the following sequence of steps on the aggregate switches:

1. VLAN configuration
2. RPVST+ configuration
3. Primary and secondary root configuration
4. Configuration of port channels between aggregate switches
5. Configuration of port channels between aggregate and CGESM switches
6. Trunking the port channels between aggregate switches
7. Configuration of default gateway for each VLAN

Note   Configuration Details, page 2-53 describes each of these steps.

Configuring the CGESM Switches

Complete the following sequence of steps on the CGESM switches:

1. VLAN configuration
2. RPVST+ configuration
3. Configuration of port channels between the CGESM and aggregate switches
4. Trunking port channels between the CGESM and aggregate switches
5. Configuration of server ports on the CGESM
6. Configure Layer 2 Trunk Failover

Note   Configuration Details, page 2-53 describes each of these steps.

Additional Aggregation Switch Configuration

The following recommendations help integrate the CGESM switches into the data center:

1. Enable Root Guard on the aggregate switches links connected to the switches in the blade enclosure.

   The spanning tree topology is calculated, and one of the primary parameters involved in this equation is the location of the root switch. Determining the position of the root switch in the network allows the network administrator to create an optimized forwarding path for traffic. Root Guard is a feature designed to control the location of the root switch.

   The aggregation switches should employ the **spanning-tree guard root** command on the port channel interfaces connected to the blade switches.

2. Allow only those VLANs that are necessary on the port channel between the aggregate switch and the blade switches.

   Use the **switchport trunk allowed vlan** *vlanID* command to configure the port channel interfaces of the aggregate switch to allow only those VLANs indicated with the *vlanID* option.

Additional CGESM Configuration

1. Enable BPDU Guard on the internal server ports of the switch

Use the **spanning-tree bpduguard enable** command to shut down a port that receives a BPDU when it should not be participating in the spanning tree.

2. Allow only those VLANs that are necessary on the port channels between the aggregate switches and the blade switches.

Use the **switchport trunk allowed vlan** *vlanID* command to configure the port channel interfaces of the switch to allow only those VLANs indicated with the *vlanID* option.

Layer 2 Looped Access Layer Design—"Square"

Figure 2-27 illustrates an alternative topology that relies on RPVST+ to account for redundant paths in the network. The two aggregate switches connect to each other via a port channel supporting the server farm VLANs. The four external uplinks of each CGESM are channeled and connected to one of the two aggregate switches. The internal connections between the two CGESMs complete the loop.

Figure 2-27   Alternate Topology HP BladeSystem p-Class with CGESMs
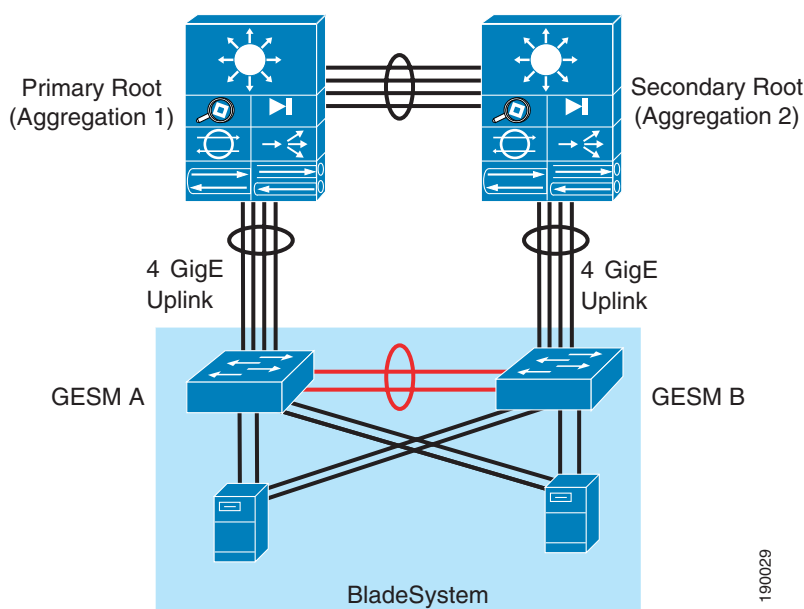
This design uses the links between the two CGESMs as a redundant path for blade server traffic. In Figure 2-27, the black links are in spanning tree forwarding state and the red links are in blocking state. These links are in blocking state by default. The use of a longer path cost value provides for a more granular calculation of the topology based on the available link bandwidth (see Cisco IGESM Features, page 2-3). This feature is enabled with the **spanning-tree pathcost method long** command. RPVST+ should be used in this network design for its fast convergence and predictable behavior.

Note This design uses a lower bandwidth path when an uplink failure occurs on either CGESM A or CGESM B. To increase the bandwidth of the redundant path between the CGESMs, consider using the external ports of CGESM A and B in the EtherChannel.

The following convergence tests were conducted with this alternate topology:

- Uplink failure and recovery between Switch-A and the primary root

- Uplink failure and recovery between Switch-B and the secondary root

- Failure and recovery of Switch-A and Switch-B

- Failure and recovery of the primary and secondary root switches

These tests yielded results similar to the recommended topology. Layer 2 convergence occurs in approximately one second. As stated previously, recovery at Layer 3 is dependent on the HSRP settings of the aggregate switches (see Network Management Recommendations, page 2-46). In the Cisco test bed, the failure of the active HSRP device typically increased the convergence time to five seconds.

The design in Figure 2-27 supports traffic monitoring via SPAN and/or RSPAN. For example, a network analysis device connected to the external ports on the front of the CGESM may capture locally mirrored traffic. Alternatively, RSPAN traffic may be carried on the CGESM uplinks if bandwidth utilization is not a concern. For the steps to configure traffic monitoring, see Configuration Details, page 2-53.

Layer 2 Loop-Free Access Layer Design—Inverted "U"

Figure 2-28 shows a Layer 2 loop-free access layer design commonly referred to as an inverted "U". The two aggregate switches connect to each other via a port channel supporting the server farm VLANs. The four external uplinks of each CGESM are channeled and connected to one of the two aggregate switches. The internal connections between the two CGESMs remain disabled (the default state of these links) to disrupt the loop. This design requires that the connections between the two CGESMs remain disabled to prevent a loop condition. Cisco recommends that RPVST+ be enabled to account for inadvertent loops created through human errors during configuration or uplink wiring.

Figure 2-28   Layer 2 Loop-Free Access Layer Design—Inverted "U"
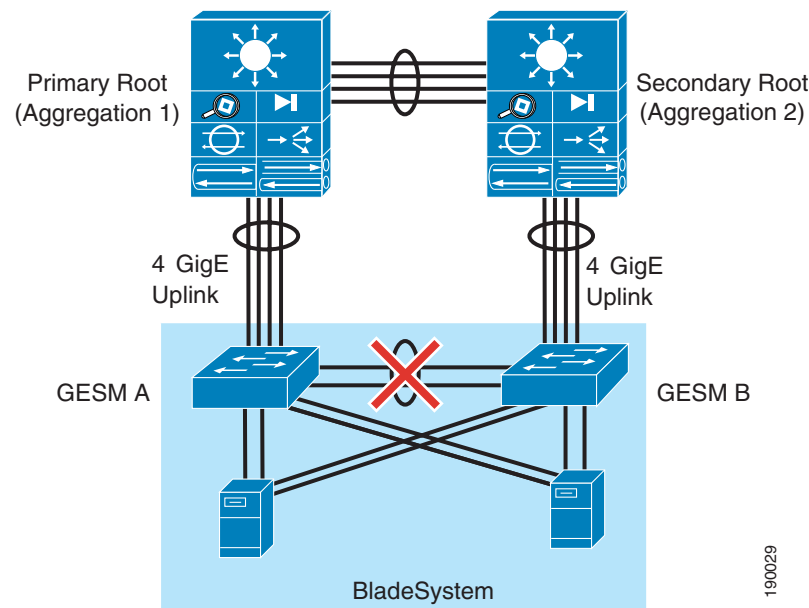
The inverted "U" design provides a highly available blade server environment using NIC teaming and Layer 2 trunk failover. The CGESM Trunk Failover feature disables downstream server ports when the uplink port channel fails. Disabling the downstream server ports of the CGESM allows a properly configured ProLiant blade server NIC team to converge traffic to the remaining active NICs homed to the other CGESM in the BladeSystem.

Convergence tests with this topology revealed that approximately three seconds of downtime is experienced when uplink failures occur on the CGESM with the active server NICs. The CGESM, employing Layer 2 Trunk Failover, properly disabled downstream server ports when the uplink failure condition occurred.

Configuring the Aggregate Switches

Complete the following sequence of steps on the aggregate switches:

1. VLAN configuration
2. RPVST+ configuration
3. Primary and secondary root configuration
4. Configuration of port channels between aggregate switches
5. Configuration of port channels between aggregate and CGESM switches
6. Trunking the port channels between aggregate switches
7. Configuration of default gateway for each VLAN

> **Note**   Configuration Details, page 2-53 describes each of these steps.

Configuring the CGESM Switches

Complete the following sequence of steps on the CGESM switches:

1. VLAN configuration
2. RPVST+ configuration
3. Configuration of port channels between the CGESM and aggregate switches
4. Trunking port channels between the CGESM and aggregate switches
5. Configuration of server ports on the CGESM
6. Configure Layer 2 Trunk Failover

> **Note**   Configuration Details, page 2-53 describes each of these steps.

Configuration Details

This section describes the configuration steps required for implementing the topologies discussed in this guide. The configuration for the following are discussed:

- VLAN
- RPVST+
- Inter-Switch Link
- Port Channel
- Trunking
- Server Port
- Server Default Gateway

- RSPAN
- Layer 2 Trunk Failover

VLAN

To configure the VLANs on the switches, complete the following tasks:

Step 1   Set the VLAN trunking protocol administrative domain name and mode as follows:

```
(config)# vtp domain <domain name>
(config)# vtp mode transparent
```

Step 2   Configure the server farm VLANs as follows:

```
(config)# vlan 60
(config-vlan)# name bladeservers
(config-vlan)# state active
```

RPVST+

Step 1   Configure STP to manage the physical loops in the topology. Cisco recommends using RPVST+ for its fast convergence characteristics. Set the STP mode an each aggregation switch as follows:

```
(config)# spanning-tree mode rapid-pvst
```

Step 2   Configure the path cost to use 32 bits in the STP calculations:

```
(config)# spanning-tree pathcost method long
```

Step 3   Configure the root switch as follows:

```
(config)# spanning-tree vlan <vlan range> root primary
```

Step 4   Configure the secondary root switch as follows:

```
(config)# spanning-tree vlan <vlan range> root secondary
```

Inter-Switch Link

The topologies discussed in this guide require connectivity between the switches. The following three types of inter-switch connections exist:

- Aggregate-1 to Aggregate-2
- Aggregate-1 or Aggregate-2 to Blade Enclosure Switch-A or Switch-B
- HP BladeSystem Switch-A to Switch-B

Each of these connections are Layer 2 EtherChannels consisting of multiple physical interfaces bound together as a channel group or port channel. These point-to-point links between the switches should carry more than one VLAN; therefore, each is a trunk.

Port Channel

Link Aggregate Control Protocol (LACP) is the IEEE standard for creating and managing EtherChannels between switches. Each aggregate switch uses this feature to create a port channel across the line cards. The use of multiple line cards within a single switch reduces the possibility of the point-to-point port channel becoming a single point of failure in the network.

Step 1   Configure the active LACP members on Aggregate-1 to CGESM Switch-A as follows:

```
(config)# interface GigabitEthernet12/1
(config-if)# description <<** Connected to Switch-A **>>
(config-if)# channel-protocol lacp
(config-if)# channel-group 1 mode active
(config)# interface GigabitEthernet11/1
(config-if)# description <<** Connected to Switch-A **>>
(config-if)# channel-protocol lacp
(config-if)# channel-group 1 mode active
```

Step 2   Configure the passive LACP members on CGESM Switch-A as follows:

```
(config) # interface GigabitEthernet0/19
(config-if)# description <<** Connected to Aggregation-1 **>>
(config-if)# channel-group 1 mode on
(config) # interface GigabitEthernet0/20
(config-if)# description <<** Connected to Aggregation-1 **>>
(config-if)# channel-group 1 mode on
```

Trunking

Use the following guidelines when configuring trunks:

- Allow only those that are necessary on the trunk
- Use 802.1q trunking
- Tag all VLANs over a trunk from the aggregation switches

Step 1   Configure trunks using the standard encapsulation method 802.1q as follows:

```
(config-if)# switchport trunk encapsulation dot1q
```

Step 2   Define the VLANs permitted on a trunk as follows:

```
(config-if)# switchport trunk allowed vlan <VLAN IDs>
```

Step 3   Modify the VLANs allowed on a trunk using one of the following commands:

```
(config-if)# switchport trunk allowed vlan add <VLAN IDs>
(config-if)# switchport trunk allowed vlan remove <VLAN IDs>
```

Step 4   Define a port as a trunk port as follows:

```
(config-if)# switchport mode trunk
```

Note   The auto-negotiation of a trunk requires that the ports be in the same VTP domain and be able to pass DTP frames.

Step 5   To secure and enforce a spanning tree topology, configure the Root Guard feature on the aggregate switch interfaces that connect to the blade switches.

The following is an example of the interface configuration between the aggregate and blade switch with root guard enabled:

```
(config)# interface GigabitEthernet12/13
config-if)# description <text>
config-if)# no ip address
config-if)# switchport
config-if)# switchport trunk encapsulation dot1q
config-if)# switchport trunk native vlan <vlan id>
config-if)# switchport trunk allowed vlan <vlan id>
config-if)# switchport mode trunk
config-if)# spanning-tree guard root
config-if)# channel-protocol lacp
config-if)# channel-group <group id> mode active
```

Server Port

A blade server is assigned a specific port on the blade switch. This is pre-determined by the physical slot the blade server occupies in the chassis. Table 2-3 correlates the server and switch ports.

Table 2-4   Correlation of Server and Switch Ports

IOS CLI Identifier	Actual Port Location in 8-Slot Server Chassis	Actual Port Location in 16-Slot Server Chassis
GigabitEthernet 0/1	Server Slot 1	Server Slot 1
GigabitEthernet 0/2	Server Slot 1	Server Slot 2
GigabitEthernet 0/3	Server Slot 2	Server Slot 3
GigabitEthernet 0/4	Server Slot 2	Server Slot 4
GigabitEthernet 0/5	Server Slot 3	Server Slot 5
GigabitEthernet 0/6	Server Slot 3	Server Slot 6
GigabitEthernet 0/7	Server Slot 4	Server Slot 7
GigabitEthernet 0/8	Server Slot 4	Server Slot 8
GigabitEthernet 0/9	Server Slot 5	Server Slot 9
GigabitEthernet 0/10	Server Slot 5	Server Slot 10
GigabitEthernet 0/11	Server Slot 6	Server Slot 11
GigabitEthernet 0/12	Server Slot 6	Server Slot 12
GigabitEthernet 0/13	Server Slot 7	Server Slot 13
GigabitEthernet 0/14	Server Slot 7	Server Slot 14
GigabitEthernet 0/15	Server Slot 8	Server Slot 15
GigabitEthernet 0/16	Server Slot 8	Server Slot 16
GigabitEthernet 0/17	Cross Connect Port 1	Cross Connect Port 1
GigabitEthernet 0/18	Cross Connect Port 2	Cross Connect Port 2
GigabitEthernet 0/19	SFP/Uplink Port 1	SFP/Uplink Port 1
GigabitEthernet 0/20	SFP/Uplink Port 2	SFP/Uplink Port 2
GigabitEthernet 0/21	SFP/Uplink Port 3	SFP/Uplink Port 3
GigabitEthernet 0/22	SFP/Uplink Port 4	SFP/Uplink Port 4

Table 2-4 *Correlation of Server and Switch Ports*

GigabitEthernet 0/23	RJ45/Front Panel Port 1	RJ45/Front Panel Port 1
GigabitEthernet 0/24	RJ45/Front Panel Port 2	RJ45/Front Panel Port 2

The server ports on the blade switch support a single VLAN access and trunk configuration modes. The operational mode chosen should support the server NIC configuration (that is, a trunking NIC is attached to a trunking switch port). Enable PortFast for the edge devices.

The BPDU Guard feature disables a port that receives a BPDU. This feature protects the STP topology by preventing the blade server from receiving BPDUs. A port disabled via the BPDU Guard feature must be recovered by an administrator manually. Enable the BPDU Guard feature on all server ports that should not be receiving BPDUs.

Port Security limits the number of MAC addresses permitted to access the blade switch port. Configure the maximum number of MAC addresses expected on the port.

Note The NIC teaming driver configuration (that is, the use of a virtual MAC address) must be considered when configuring Port Security.

```
interface GigabitEthernet0/1
 description <<** BladeServer-1 **>>
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 10,60
 switchport mode trunk
 switchport port-security aging time 20
 switchport port-security maximum 1 vlan 10,60
 no cdp enable
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
end
```

Server Default Gateway

The default gateway for a server is a Layer 3 device located in the aggregation layer of the data center. This device may be a firewall, a load balancer, or a router. Using protocols such as HSRP protect the gateway from being a single point of failure and create a highly available date center network. HSRP allows the two aggregate switches to act as a single virtual router by sharing a common MAC and IP address between them. Define a switched VLAN interface on each aggregate switch and use the HSRP address as the default gateway of the server farm.

Step 1 Configure Aggregation-1 as the active HSRP router. The **priority** command helps to select this router as the active router because it has a greater value.

```
interface Vlan10
 description <<**  BladeServerFarm – Active  **>>
 ip address 10.10.10.2 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.10.1
 standby 1 timers 1 3
 standby 1 priority 51
 standby 1 preempt delay minimum 60
 standby 1 authentication <password>
end
```

Step 2   Configure Aggregation-2 as the standby HSRP router as follows:

```
interface Vlan10
 description <<**  BladeServerFarm – Standby  **>>
 ip address 10.10.10.3 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.10.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 preempt delay minimum 60
 standby 1 authentication <password>
end
```

RSPAN

RSPAN allows for remote traffic monitoring in the data center. Define source and destination sessions to mirror interesting traffic to a remote VLAN captured by network analysis tools.

Step 1   Configure a VLAN for RSPAN on the CGESM and the aggregate switch as follows:

```
(config)# vlan <vlanID>
(config-vlan)# name <vlan name>
(config-vlan)# remote-span
```

Step 2   Create a source session as follows. This is the interface or VLAN that contains interesting traffic.

```
(config) # monitor session <session id> source vlan <VLAN IDs>
```

Step 3   Configure the RSPAN VLAN as the target for the mirrored traffic as follows:

```
(config) # monitor session <session ID> destination remote vlan <remote vlan ID>
```

Layer 2 Trunk Failover

The trunk failover feature may track an upstream port or a channel.

Step 1   To assign an interface to a specific link state group, use the following command in the interface configuration sub mode:

```
(config-if)#link state group <1-2> upstream
```

> **Note**   Gigabit Ethernet interfaces 0/19–24 may only be configured as "upstream" devices.

Step 2   Enable the Trunk Failover feature for the internal blade server interfaces, downstream ports, for a specific link state group.

```
interface GigabitEthernet0/1
 description blade1
 link state group <1-2> downstream

interface GigabitEthernet0/2
 description blade2
 link state group <1-2> downstream
```

> ✎
>
> **Note**   Gigabit Ethernet interfaces 0/1–16 may be configured only as "downstream" devices.

Step 3   Globally enable the trunk failover feature for a specific link state group:

```
(config)#link state track <1-2>
```

Step 4   To validate the trunk failover configuration, use the following command:

```
show link state group detail
```

Pass-Through Technology

This chapter provides best design practices for deploying blade servers using pass-through technology within the Cisco Data Center Networking Architecture, describes blade server architecture, and explores various methods of deployment. It includes the following sections:

- Blade Servers and Pass-Through Technology
- Design Goals
- Design and Implementation Details
- Configuration Details

Blade Servers and Pass-Through Technology

A blade server is an independent server that includes an operating system, memory, one or more processors, network controllers, and optional local storage. Blades servers are designed to reduce the space, power, and cooling requirements within the data center by providing these services within a single chassis. Blade server systems are a key component of data center consolidation that help reduce costs and provide a platform for improving virtualization, automation, and provisioning capabilities.

A primary consideration in any blade server deployment is how the blade server system is connected to the data center network. There are several I/O options for blade server systems, including the following:

- Built-in Ethernet switches (such as the Cisco Ethernet Switch Modules)
- Infiniband switches (such as the Cisco Server Fabric Switch)
- Fibre Channel switches
- Blade Server Chassis Pass-through Modules

Each of these I/O technologies provides a means of network connectivity and consolidation.

This chapter focuses on integrating blade server systems within the Cisco Data Center Architecture using pass-through technology, which allows individual blade servers to communicate directly with resources external to the blade chassis. Both copper and optical pass-through modules are available that provide access to the blade server controllers. It is therefore important to understand the internal connectivity provided by the blade server chassis before discussing the external ramifications of pass-through deployments.

Currently, there is no industry-standard design for blade servers or blade server enclosures. Various blade system architectures are available from various vendors. The following section describes two generic blade server systems, which illustrate many of the design features found in these various architectures:

- System A—Uses octopus cabling to interconnect the blade servers with the data center architecture.
- System B—Passes the blade server signaling to the external network port-to-port.

System A in Figure 3-1 illustrates the front and rear view of a typical blade server chassis.

Figure 3-1   Example Blade Server Architecture—System A
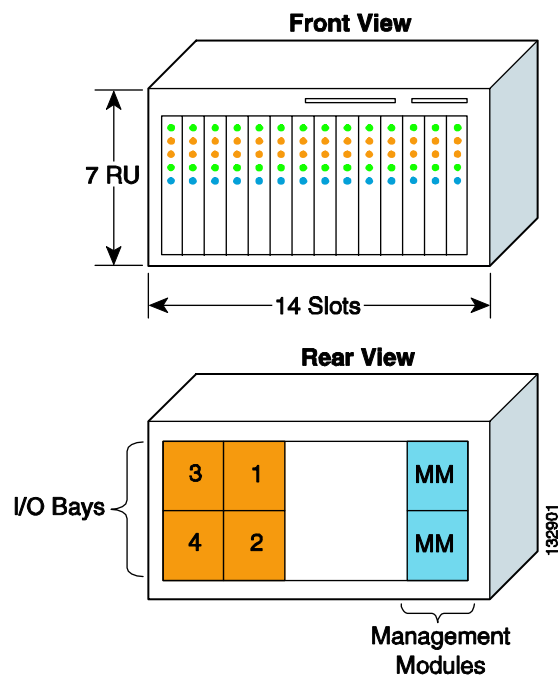
System A is seven rack units (RUs) in height and provides 14 slots to house individual blade servers. The rear of the chassis allows for four individual I/O modules for network connectivity and two management modules to administer the blade system. The blade servers and I/O modules communicate over a midplane, as shown in Figure 3-2.

Figure 3-2   System A Internal Blade Server Connectivity
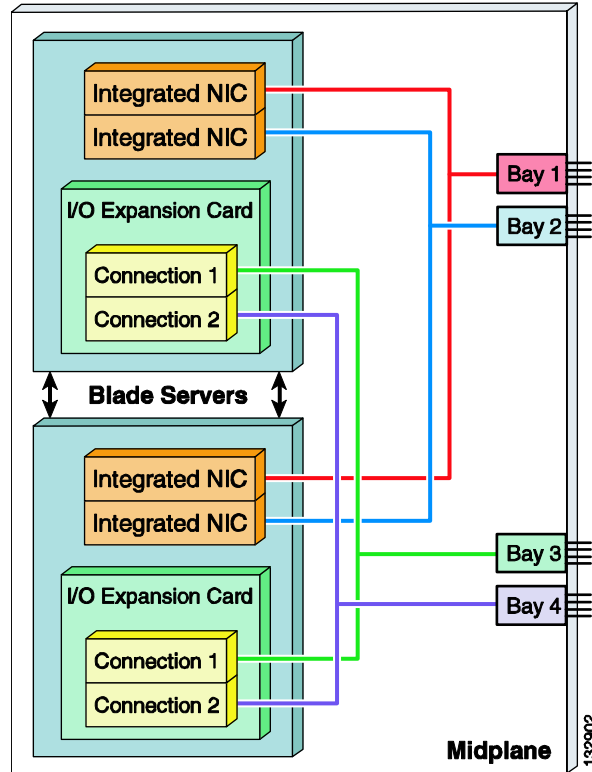
Each network interface controller (NIC) has a dedicated channel on the midplane connecting it to a specific I/O module bay. Typically, the integrated NICs are Gigabit Ethernet by default, while the I/O expansion card supports host bus adapters (HBA), host channel adapters (HCA), or Gigabit Ethernet NICs.

Note   Note that with this architecture, the I/O expansion card on each blade server must be compatible with the I/O modules installed in Bay 3 and Bay 4.

System B (see Figure 3-3) illustrates another common blade server architecture.

Figure 3-3   Example Blade Server Architecture—System B
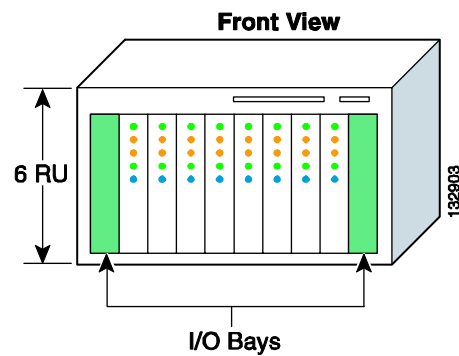
This six-RU blade server chassis has ten slots; two for I/O modules, and eight dedicated for server use. The blade server chassis provides four dedicated backplane channels to each server slot. Figure 3-4 shows this design, which supports a total of 32 independent channels for the eight blade server slots.

Figure 3-4   System B Blade Server Connectivity

Backplane Connectors
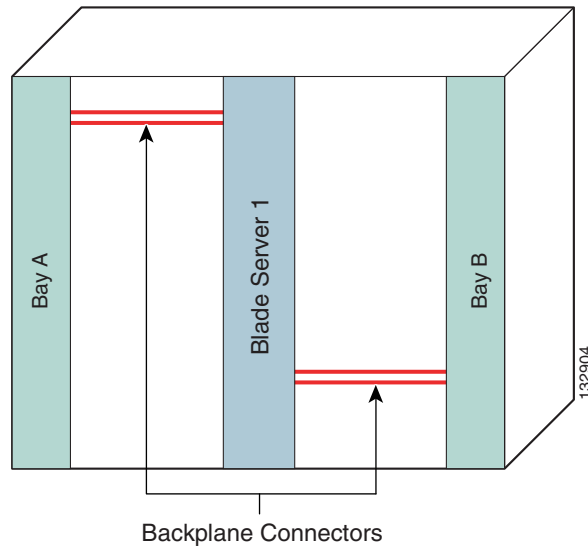
The architectures represented by System A and System B use different types of pass-through module technology, as shown in Figure 3-5.

Figure 3-5   Pass-Through Module Examples
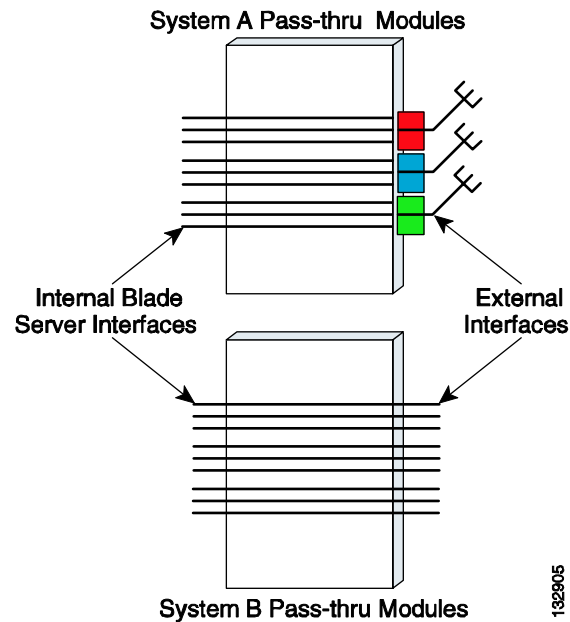
The pass-through module technology used in System A depends on octopus cables to connect to the external network. This octopus cable allows multiple servers to be supported by a single output cable that connects to the external network with transmit and receive pairs dedicated to each blade server controller.

The architecture represented by System B does not use any physical cabling consolidation. Instead, it simply passes the blade server signaling to the external network port-to-port.

Both systems provide redundant dedicated connections to the I/O modules over the midplane or backplane. By default, each blade server is dual- or multi-homed to the I/O modules deployed on the blade system. This physical design provides an increased level of availability for the services deployed on the blade servers. In addition, the redundant I/O modules can be used for establishing redundant connections to the external network, which provides an even higher level of availability.

Design Goals

This section describes the key design goals when deploying blade servers with pass-through technology in data centers. It includes the following topics:

- High Availability
- Pass-through technology is a flexible solution that provides blade server high availability by supporting all three NIC teaming modes of operation.
- Manageability

High Availability

This section describes key issues to consider when making design choices for the overall data center architecture as well as for the blade servers.

Achieving Data Center High Availability

Data centers house the critical business applications of the enterprise, which must be accessible for use either at specific times or continuously, and without interruption. The network infrastructure provides the level of availability required by these applications through device and link redundancy and a deterministic topology. Servers are typically configured with multiple NIC cards and dual-homed to the access layer switches to provide backup connectivity to the business applications.

The implementation of blade servers does not change the high availability requirements of the data center. Implementing blade servers with pass-through technology allows non-disruptive deployment. Pass-through deployments do not alter the fast convergence and deterministic traffic patterns provided by Layer 2 and Layer 3 technologies. The connection established between the external network device and the blade server by the pass-through module is neither switched nor blocked. The modules simply expose the blade server NICs to the network without affecting the Layer 2 or Layer 3 network topologies created through spanning tree or routing protocols. When using pass-through modules, the blade servers function as servers that happen to be located inside a blade server chassis.

Achieving Blade Server High Availability

Blade server enclosures provide high availability to local blade servers through a multi-homed architecture. Each blade server achieves an increased level of accessibility by using NIC teaming software. NIC teaming lets you create a virtual adapter consisting of one to eight physical NIC interfaces, which can typically support up to 64 VLANs. NIC teaming is a high availability mechanism that can provide both failover and local load balancing services. There are the following three primary modes of NIC teaming:

- Fault tolerant mode

- Load balancing mode

- Link aggregation mode

Fault tolerant mode, also known as *active/standby*, creates a virtual NIC by grouping one or more network controllers in a team. One adapter is the primary or active NIC, leaving all other network controllers in the team as secondary or standby interfaces. The standby interface becomes active if the primary NIC fails because of probe or physical link problems. Fault tolerance can be achieved in the following two ways:

- Adapter fault tolerance (AFT)—The NIC team is homed to a single switch.

- Network or switch fault tolerance (NFT/SFT)—The NIC team is dual-homed to two different switches.

Pass-through I/O modules support both configurations. Figure 3-6 illustrates AFT and NFT configurations with blade servers using pass-through.

Figure 3-6   Pass-Through Module with Fault Tolerant NIC Teaming

Blade Server with Pass-thru I/O Modules
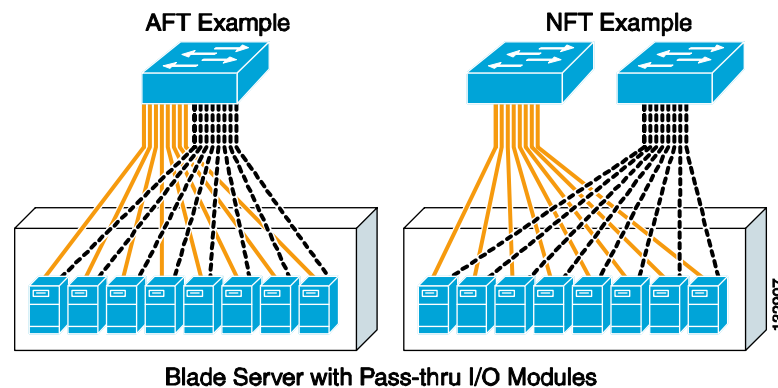
The solid yellow links represent the active links, and the dotted grey links represent the standby links that are not being used by the server. NFT provides a greater level of network availability. However, neither configuration optimizes the bandwidth available to the blade server.

The second method of NIC teaming (load balancing mode) builds upon the high availability features of NFT by configuring all the NICs in a team to participate in the transmission of data. This feature lets the server utilize more of the available bandwidth.

In load balancing mode, a single primary NIC receives all incoming traffic while egress traffic is load balanced across the team by a Layer 2- or Layer 3-based algorithm. Pass-through technology permits this configuration, as shown in Figure 3-7.

Figure 3-7   Pass-Through Module with Load Balance NIC Teaming

Load Balancing Mode

Blade Server with Pass-thru I/O Modules
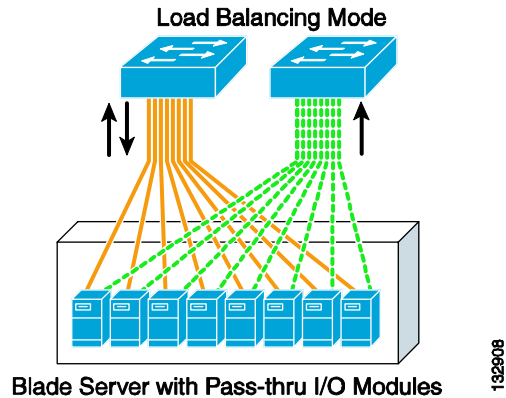
The solid yellow lines indicate the primary interfaces of the blade server that both receive and transmit traffic. The dotted green lines are standby interfaces that only transmit traffic. A hashing algorithm, usually based on the source and destination IP addresses, determines which NIC is responsible for transmitting the traffic for any given transaction. The standby controller becomes responsible for both ingress and egress traffic only if the primary NIC fails.

The third method of NIC teaming, link aggregation mode (channeling), extends the load balancing functionality of NIC teaming by allowing all interfaces to receive and transmit traffic. This requires the switch to load balance traffic across the ports connected to the server NIC team. This mode of NIC teaming provides link redundancy and the greatest bandwidth utilization. However, the access switch in this design represents a single point of failure. Figure 3-8 shows a channel established between the blade servers and the access switch.

Figure 3-8   Pass-Through Module with Link Aggregation NIC Teaming

Link Aggregation

Blade Server with Pass-thru I/O Modules
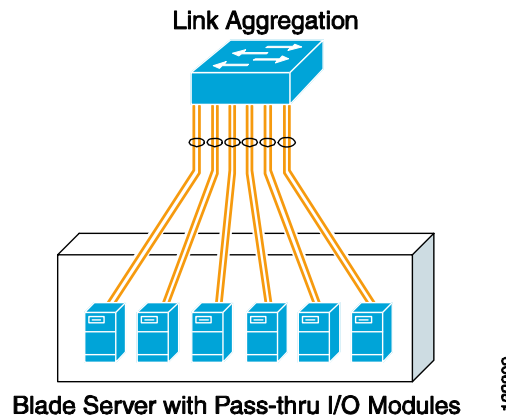
NIC teaming software typically supports the configuration of Gigabit EtherChannel (GEC) or IEEE 802.3ad (LACP) channels.

Pass-through technology is a flexible solution that provides blade server high availability by supporting all three NIC teaming modes of operation.

Scalability

Network designs incorporating blade server devices must ensure network scalability to allow for increases in server density. Scalability allows increases in services or servers without requiring fundamental modifications to the data center infrastructure. The choice of I/O module used by the blade servers (integrated switches or pass-through modules) is a critical deployment factor that influences the overall data center design.

Note Pass-through modules allow blade servers to connect directly to a traditional access layer. The access layer should provide the port density to support the data center servers and the flexibility to adapt to increased demands for bandwidth or server capacity. For more information on data center scalability, see Design Goals, page 3-5, or the *Cisco Data Center Infrastructure SRND* at the following URL: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

Manageability

Management of the data center, including administration of software versions and hardware configurations, is also a key design consideration. From this point of view, blade server pass-through modules reduce the administrative complexity of the data center. Pass-through modules do not require configuration, which eliminates configuration errors on the devices and reduces the need for configuration backups. The blade server chassis may also provide limited diagnostic information and the ability to enable or disable external ports on the module. The availability of these features and the level of diagnostic information depends on the manufacturer of the blade system.

Note Serial Over LAN (SOL) is a blade server management feature available on the IBM BladeCenter chassis. However, SOL requires the use of an integrated switch and is not currently available with the IBM pass-through modules. SOL leverages the trunk that exists between the management module and the Ethernet blade switch to allow console access to the individual blade servers.

Design and Implementation Details

The following section discusses the following two access switch attachment options available when using blade server pass-through I/O modules:

- Modular Access Switches
- One Rack Unit Access Switches

These network designs emphasize high availability in the data center by eliminating any single point of failure, by providing deterministic traffic patterns, and through predictable network convergence behavior. The configuration examples use Cisco Catalyst 6500s as the aggregation layer platform. This Layer 2/Layer 3 switching platform supports high slot density and integrated network services, which are important features for data centers deploying blade systems.

The introduction of pass-through modules into an existing Layer 2/Layer 3 design does not require much modification to the existing data center architecture, which allows blade server systems to be easily inserted into the network. However, the disadvantage is that cable consolidation and use of shared interconnects, which are important benefits that can be provided by blade systems, are not fully realized.
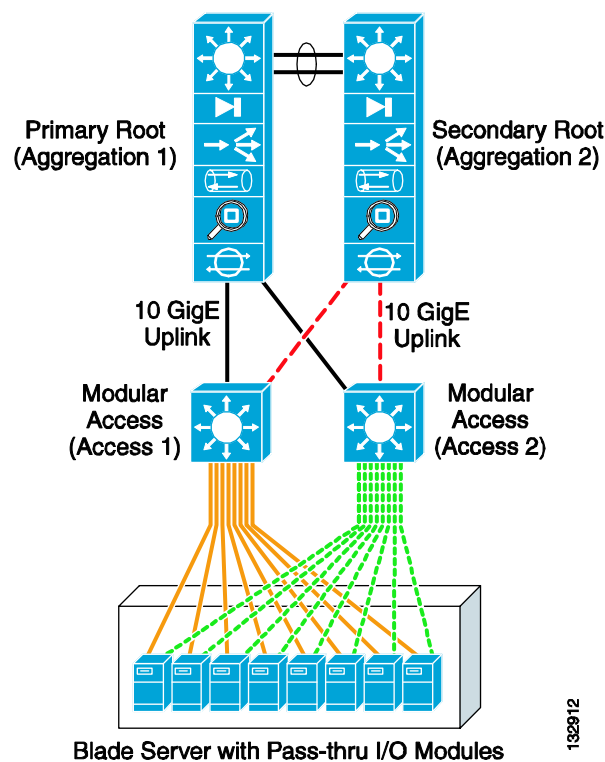
Modular Access Switches

Figure 3-9 illustrates the use of a modular access switch with pass-through I/O modules in a blade server system. The blade servers are dual-homed to the access layer switches. The modular access provides port density and 10 Gigabit Ethernet uplinks to the aggregation layer where intelligent services such as security, load balancing, and network analysis reside.

The advantages of this design include the following:

- Proven high availability design
- Scalable approach to blade server deployments
- Improved operational support

Figure 3-9   Modular Access Design with Pass-Through I/O Modules

Blade Server with Pass-thru I/O Modules

The design in Figure 3-9 uses a classic triangular topology with the modular access layer switches. There is no single point of failure in this network topology. The access layer switches are dual-homed to the aggregation layer switches, which provide redundant network paths. Spanning tree manages the physical loops created by the uplinks between the aggregation and access switches, assuring a predictable and fast converging topology.

In Figure 3-9, the solid black lines represent uplinks in a spanning tree forwarding state, while the dotted red lines represent uplinks in blocking state. Rapid Per VLAN Spanning Tree Plus (RPVST+) is recommended for this design because of its high availability features. RPVST+ provides fast convergence (less than 1 second) in device or uplink failure scenarios. In addition, RPVST+ offers enhanced Layer 2 features for the access layer with integrated capabilities equivalent to the UplinkFast and BackboneFast features in the previous version of Spanning Tree Protocol (STP).

Note   For more information on Layer 2 design and RPVST+, see the *Data Center Infrastructure SRND* at the following URL:
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

In addition to the high availability attributes of this proven access layer design, the modular access switch provides a very high level of port density. This allows the server farm to scale as it addresses future data center needs without an exponential increase in administrative overhead (see Figure 3-10).

Figure 3-10   *Scalability with a Modular Access Switch Design*
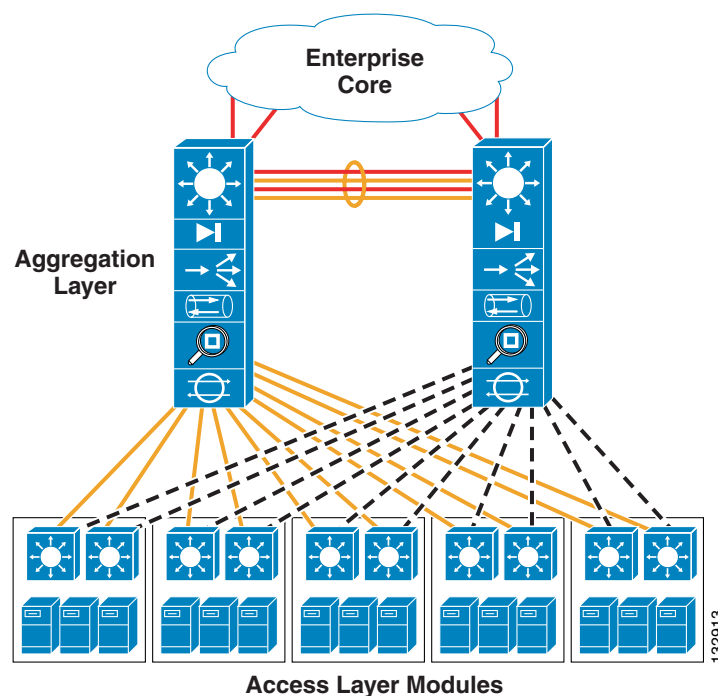
In Figure 3-10, ten modular access switches are connected to a pair of aggregation layer switches. Each pair of access switches supports 12 fully populated blade server systems housed in a set of three racks. In this example, each blade system requires 32 ports on the access switch for the pass-through links from the blade servers.

With four blade systems per rack, 128 access layer ports are required to support a single rack. Dual-homing the blade servers to each modular access switch means that each access layer switch must provide 64 ports per rack or 192 ports for three racks. Four 48-port line-cards (192/48 = 4) are required on each modular access switch to support this configuration.

Note   This example does not consider the scalability of the spanning tree, which depends on the number of active logical interfaces and virtual ports supported per line card. In addition, the acceptable oversubscription ratio for the applications must be taken into account. For more information on scaling the Layer 2 and Layer 3 topologies in the data center, see the *Data Center Infrastructure SRND* at the following URL:
http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

A modular access switch design reduces the total number of switches in the network. In the previous example, 120 integrated blade switches would be required, assuming two blade switches per chassis, to support an equivalent number of blade servers. The ramifications of introducing this number of devices into the data center network are obvious. Specifically, the Layer 2 and Layer 3 topologies expand and become more difficult to manage. In addition, there is an increase in the network administration required to support the integrated switch. Using a modular access layer switch reduces these operational and logical complexities.

However, pass-through technology does not provide the benefits of cable consolidation and the use of shared interconnects provided by integrated blade switches. Pass-through modules do not reduce the I/O cabling volume within the rack or lessen the cable bulk from the rack to the modular access switches. Cabling represents an obstruction that restricts the airflow within the data center and may adversely affect the temperature of the room. When using pass-through technology and blade servers, the design and use of an effective cable management system within the facility is necessary to mitigate these issues.

One Rack Unit Access Switches

Figure 3-11 illustrates the use of a 1-RU access switch with pass-through I/O modules in a blade server system. The blade servers are dual-homed to the 1-RU access layer switches. The 1-RU access layer switches provide the port density and uplink connectivity to the aggregation layer required by the blade servers.

Figure 3-11   1-RU Access Layer Switch with Pass-Through Technology
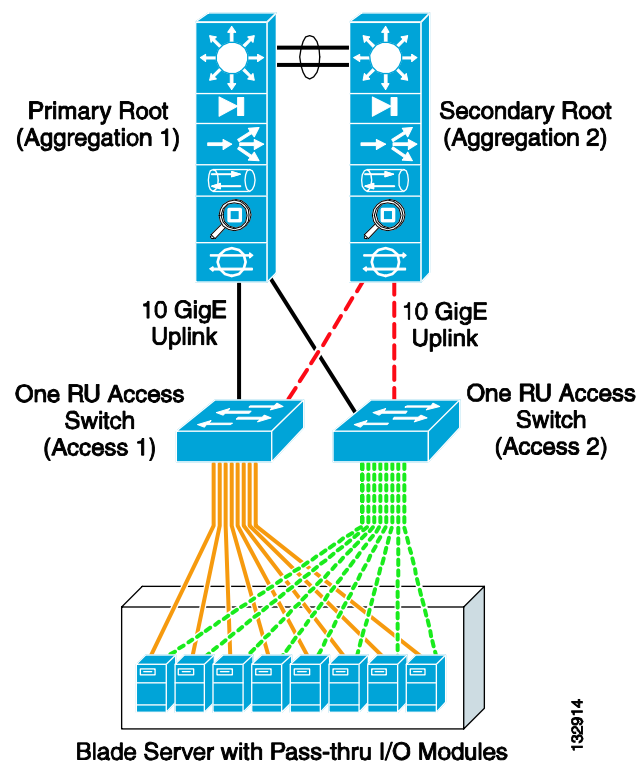
Typically, 1-RU access switches are located at the top of the racks housing the blade server units. This design allows the cable density created with the pass-through modules to remain within the rack, which helps contain the potential problems. This is a distinct advantage compared to the modular access layer

model discussed previously. The uplinks from the 1-RU access layer switches provide a common, consolidated connection to the aggregation layer and its services. This essentially reduces the number of cable runs required between the rack and the aggregation layer switches.

This design also provides a highly available and predictable server farm environment. In Figure 3-11, redundant paths for network traffic are managed with a Layer 2 protocol such as RPVST+ that provides sub-second convergence. In Figure 3-11, Aggregation-1 is the primary root switch with all links forwarding. Aggregation-2 is the secondary root and provides an alternative traffic path for application traffic in the event of a failure. The solid black lines represent links that are forwarding and the dotted red lines represent links in a spanning tree blocking state.

The blade servers provide another level of protection by using the high availability features of NIC teaming, in addition to the high availability functions of the network. The sold yellow lines represent each link to the active NIC while the dotted green lines show the links to each interface in a standby state. In this configuration, the NIC teaming software offers sub-second convergence at the server.

> **Note** It is also important to consider the high availability features available on the 1-RU switch platform when this is deployed in the data center. Redundant power and hot-swappable fans are recommended to improve Layer 1 availability.

The scalability of 1-RU switches is limited compared to the use of modular switches. A 1-RU switch cannot provide the same level of port density for server connectivity as a modular switch. As a result, the number of switching devices in the data center is increased, compared to the solution using modular switches. This in turn increases the spanning tree domain as well as the administrative overhead required to implement and maintain the solution.

For example, Figure 3-12 demonstrates the potential scale of a 1-RU access switch deployment.

As stated previously, the access switches must provide the local port density required by the blade servers. In this instance, each data center rack houses three blade systems providing 32 individual pass-through connections to the internal blade servers. The blade servers are dual-homed over these pass-through connections to a pair of 1-RU access switches located in the rack. Three blade systems with 16 dual-homed servers per chassis require 96 ports. To provide for network fault tolerance, each 1-RU access layer rack switch should supply 48 ports for server connectivity.

In addition, the access layer switches provide connectivity to the aggregation layer services. The modular aggregation layer switch must furnish the uplink port density for the 1-RU access layer. A Catalyst 6509 would suffice in this scenario. A pair of aggregation layer switches can support 12 1-RU access switches that are dual-homed over ten Gigabit Ethernet uplinks. In this example, 288 servers are supported in the 1-RU access switch model with modular aggregation layer support.

> **Note** This example does not consider the scalability of spanning tree, which depends on the number of active logical interfaces and virtual ports supported per line card. In addition, the acceptable oversubscription ratio for the applications must be taken into account. For more information on scaling the Layer 2 and Layer 3 topologies in the data center, see the *Data Center Infrastructure SRND 2.0* at the following URL:
> http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

Figure 3-12 *Scalability with a 1-RU Access Switch Design*
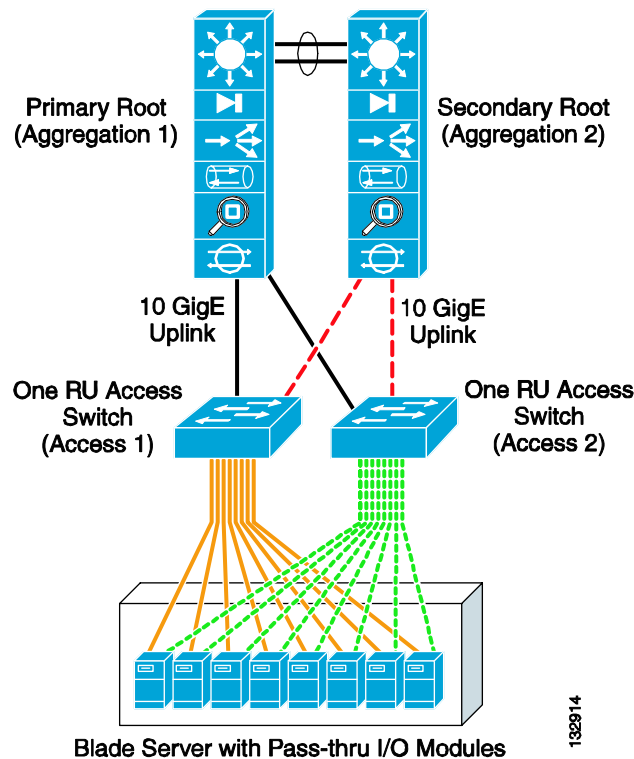
Note The use of three or four 1-RU switches per rack may be necessary depending on the number of ports available on the switching platform and the number of blade server interfaces that must be supported. This affects the scalability of this design by requiring greater uplink port density to connect to the aggregation layer switches.

The 1-RU access switch design reduces the total number of switches in the network. The example shown in Figure 3-12 would require 36 integrated blade switches, assuming two blade switches per chassis, to support an equivalent number of blade servers. The 1-RU access layer switch also reduces the operational and logical complexities (L2/L3 topologies) of the data center when compared to the integrated switch solution. In addition, the design reduces the number of cables required to provide external connectivity to the rack. As previously discussed, the modular access layer switch design requires fewer network devices and topology changes but uses more cabling.

Configuration Details

This section describes the switch configurations necessary to integrate pass-through technology into the Cisco Data Center Architecture. The following configurations are described:

- VLAN Configuration
- RPVST+ Configuration
- Inter-Switch Link Configuration

- Port Channel Configuration
- Trunking Configuration
- Server Port Configuration
- Server Default Gateway Configuration

VLAN Configuration

To configure the VLANs on the switches, complete the following steps:

Step 1   Set the VLAN trunking-protocol administrative domain name and mode as follows:

```
(config)# vtp domain domain name
(config)# vtp mode transparent
```

Step 2   Configure the server farm VLANs as follows:

```
(config)# vlan VLAN ID
(config-vlan)# name VLAN name
(config-vlan)# state active
```

RPVST+ Configuration

Configure STP to manage the physical loops in the topology. RPVST+ is recommended for its fast convergence characteristics.

Step 1   To set the STP mode an each aggregation switch, enter the following command:

```
(config)# spanning-tree mode rapid-pvst
```

The port path cost value represents the media speed of the link and is configurable on a per interface basis, including EtherChannels. To allow for more granular STP calculations, enable the use of a 32-bit value instead of the default 16-bit value. The longer path cost better reflects changes in the speed of channels and allows STP to optimize the network in the presence of loops.

Step 2   To configure STP to use 32 bits in port part cost calculations, enter the following command:

```
(config)# spanning-tree pathcost method long
```

Step 3   To configure the root switch, enter the following command:

```
(config)# spanning-tree vlan vlan range root primary
```

Step 4   To configure the secondary root switch, enter the following command:

```
(config)# spanning-tree vlan vlan range root secondary
```

Inter-Switch Link Configuration

The topologies discussed in this guide require connectivity between the switches. The following two types of inter-switch connections can be used to provide this connectivity:

- Aggregation Switch to Aggregation Switch
- Aggregation Switch to Access Layer Switch

Each of these connections are Layer 2 EtherChannels consisting of multiple physical interfaces bound together as a channel group or port channel. Each of these point-to-point links between the switches is a trunk because they typically carry more than one VLAN.

Port Channel Configuration

Link Aggregate Control Protocol (LACP) is the IEEE standard for creating and managing port channels between switches. Each aggregate switch uses this feature to create a port channel across the line cards. The use of multiple line cards within a single switch reduces the possibility of the point-to-point port channel becoming a single point of failure in the network.

Step 1   Configure the active LACP members on the aggregation switches that connect to each access layer switch as follows:

```
(config)# interface GigabitEthernet12/1
(config-if)# description Connection to Access Layer Switch
(config-if)# channel-protocol lacp
(config-if)# channel-group 1 mode active
(config)# interface GigabitEthernet11/1
(config-if)# description Connected to Access Layer Switch
(config-if)# channel-protocol lacp
(config-if)# channel-group 1 mode active
```

Step 2   Configure the passive LACP members on the access layer switch as follows:

```
(config) # interface GigabitEthernet0/19
(config-if)# description Connected to Aggregation Layer Switch
(config-if)# channel-group 1 mode on
(config) # interface GigabitEthernet0/20
(config-if)# description Connected to Aggregation Layer Switch
(config-if)# channel-group 1 mode on
```

Trunking Configuration

Use the following guidelines when configuring trunks:

- Allow only those VLANs that are necessary on the trunk
- Use 802.1q trunking
- Tag all VLANs over a trunk from the aggregation switches

Step 1   Configure trunks using the standard encapsulation method 802.1q by entering the following command:

```
(config-if)# switchport trunk encapsulation dot1q
```

Step 2   Define the VLANs permitted on a trunk by entering the following command:

```
(config-if)# switchport trunk allowed vlan VLAN IDs
```

Step 3   Modify the VLANs allowed on a trunk by using the following commands:

```
(config-if)# switchport trunk allowed vlan add VLAN IDs
(config-if)# switchport trunk allowed vlan remove VLAN IDs
```

Step 4   Define a port as a trunk port by entering the following command:

```
(config-if)# switchport mode trunk
```

✎
Note   The auto-negotiation of a trunk requires that the ports be in the same VTP domain and be able to exchange DTP frames.

Step 5   To secure and enforce a spanning tree topology, configure the Root Guard feature on the aggregate switch interfaces that connect to the access layer switches. The following is an example of the interface configuration between the aggregate and access layer switch with Root Guard enabled:

```
(config)# interface GigabitEthernet12/13
config-if)# description text
config-if)# no ip address
config-if)# switchport
config-if)# switchport trunk encapsulation dot1q
config-if)# switchport trunk native vlan <vlan id>
config-if)# switchport trunk allowed vlan <vlan id>
config-if)# switchport mode trunk
config-if)# spanning-tree guard root
config-if)# channel-protocol lacp
config-if)# channel-group group id mode active
```

Server Port Configuration

The server ports on the access layer switch may support single VLAN access and/or trunk configuration modes. The operational mode chosen should support the server NIC configuration. In other words, a trunking NIC should be attached to a trunking switch port. Enable PortFast for the edge devices in the spanning tree domain to expedite convergence times.

BPDU Guard disables a port that receives a BPDU. This feature protects the STP topology by preventing the blade server from receiving BPDUs. A port disabled using BPDU Guard must be recovered by an administrator manually. Enable BPDU Guard on all server ports that should not receive BPDUs. The commands required to enable this feature are as follows:

```
interface GigabitEthernet6/1
 description  blade server port
 speed 1000
 duplex full
 switchport
 switchport access vlan VLAN ID
 switchport mode access
 spanning-tree portfast
end
```

Port Security limits the number of MAC addresses permitted to access the blade switch port. To configure Port Security, configure the maximum number of MAC addresses expected on the port. The NIC teaming driver configuration (the use of a virtual MAC address) must be considered when configuring Port Security.

To enable Port Security, enter the following command:

```
(config)# switchport port-security maximum maximum addresses
```

Server Default Gateway Configuration

The default gateway for a server is a Layer 3 device located in the data center aggregation layer. This device may be a firewall, a load balancer, or a router. Using a redundancy protocol, such as HSRP, protects the gateway from becoming a single point of failure and improves data center network availability. HSRP allows the two aggregate switches to act as a single virtual router by sharing a common MAC and IP address between them. To enable HSRP, define a Switched VLAN Interfaces (SVI) on each aggregate switch and use the HSRP address as the default gateway of the server farm.

Step 1   Configure one aggregation switch as the active HSRP router. The **priority** command helps to select this router as the active router by assigning it a higher value.

```
interface Vlan10
 description    Primary Default Gateway
 ip address IP address subnet mask
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip IP address
 standby 1 timers 1 3
 standby 1 priority 51
 standby 1 preempt delay minimum 60
 standby 1 authentication password
end
```

Step 2   Configure the second aggregation switch as the standby HSRP router as follows:

```
interface Vlan10
 description Standby Default Gateway
 ip address 10.10.10.3 255.255.255.0
 no ip redirects
 no ip proxy-arp
 arp timeout 200
 standby 1 ip 10.10.10.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 preempt delay minimum 60
 standby 1 authentication password
end
```

4

Blade Server Integration into the Data Center with Intelligent Network Services

This chapter discusses the integration of intelligent services into the Cisco Data Center Architecture that uses blade server systems. It includes the following topics:

- Blade Server Systems and Intelligent Services
- Data Center Design Overview
- Design and Implementation Details

Blade Server Systems and Intelligent Services

A blade server is an independent server that includes an operating system, memory, one or more processors, network controllers, and optional local storage. Blade servers reduce data center space, power, and cooling requirements by providing these services within a single chassis. Blade server systems are a key component of data center consolidation that help reduce costs and improve virtualization, automation, and provisioning capabilities.

Blade system vendors offer high availability, scalability, and security services such as load balancing, firewalls, and intrusion detection to enterprise applications. Depending on the vendor, these services are made available through the blade system integrated network devices or individual blade servers using service specific software. Figure 4-1 shows a service- enabled blade system. In this example, two software-based firewalls are located on two of the eight-blade servers in the chassis. This example also shows two integrated network switches, one on each side of the device, connecting to the outside network as well as to Layer 4 through Layer 7 services.

Figure 4-1   *Blade Server System Service Example*
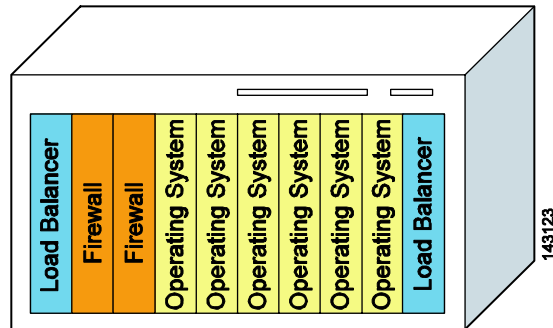
Example Blade System

■ L4 – L7 Load Balancer via Integrated Switch
■ Firewall located on Blade Server
■ Operating System on Blade Server

Traditionally, these services have been provided by appliances or integrated service modules located in the data center network. Modern blade systems are challenging this approach.

Data Center Design Overview

A data center design must consider application architecture and network services. This section reviews these needs and includes the following topics:

- Application Architectures
- Network Services in the Data Center

Application Architectures

The data center is a repository for enterprise software applications. These applications are continuously changing to meet business requirements and to accommodate the latest technological advances and methods. Consequently, the logical and physical structure of the data center server farm and of the network infrastructure hosting these software applications is changing.

The server farm has evolved from the classic client/server model to an N-tier approach. The N-tier model logically or physically separates the enterprise application by creating functional areas. These areas are generally defined as the web front end, the application business logic, and the database tiers. Figure 4-2 illustrates the progression of the enterprise application from the client/server to the N-tier paradigm.

Figure 4-2   Client/Server and N-Tier Model
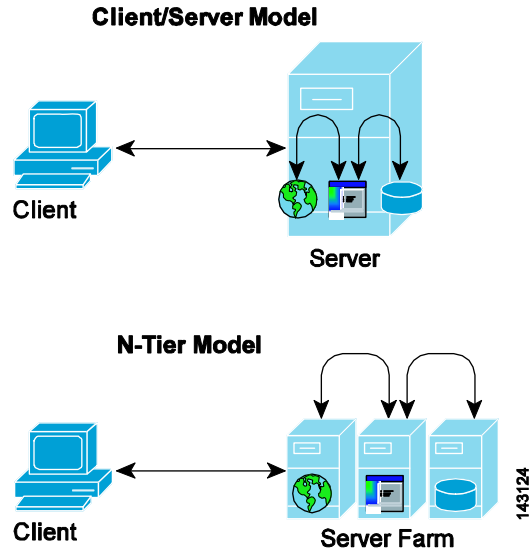
The N-tier model provides a more scalable and manageable enterprise application environment because it creates distinct serviceable areas in the software application. The application is distributed and becomes more resilient because single points of failure are removed from the design. In addition, the N-tier application architecture impacts the design of the data center network and the services it renders. The data center network must be able to adapt and meet the requirements of a multi-tiered server farm. It must also provide a reliable means for distributed servers to communicate and to benefit from other data center services such as the following:

- Infrastructure services (VLANs, RPVST+, UDLD, HSRP, VRRP, OSPF, QoS)
- Application optimization (load balancing, caching, SSL offloading)
- Security (ACLs, firewalls, intrusion detection/prevention)

The N-tier application architecture is the current standard for enterprise data center application deployments.

> **Note**   For more information on a flexible data center design for the N-tier application architecture, see the *Data Center Infrastructure SRND v 2.0* at the following URL:
> http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

There is a movement toward creating service-oriented architectures (SOAs) in the enterprise application space. SOAs advocate the ability to connect and integrate multi-tiered enterprise applications based on standards. Standard transport, standard messaging and standard data formats are used throughout the enterprise and between partners. The objective of a SOA architecture is to readily incorporate new functionality and provide flexible solutions to business requirements; ultimately creating a competitive advantage for the enterprise.

The data center network transports the messaging created by these distinct application services. In a SOA environment, reliable messaging is critical to the success of the enterprise. The network provides this end-to-end functionality, managing the messaging between services. If standard forms of communication are used by a service-oriented architecture, the network simply needs the ability to interpret those messages to make informed decisions and provide increasing value to the business.

> **Note**   One common type of SOA is using web services. A web service can be defined as "any piece of software that makes itself available over the Internet and uses a standardized XML messaging system" –Ethan Cerami, O'Reilly Press. The IP network supports the transport protocols typically used by web services such as HTTP, HTTPS, or SMTP. Messaging between web services is typically performed with a non-proprietary messaging protocol such as Simple Object Access Protocol (SOAP) allowing data exchange and remote process communication between dissimilar applications. The Extensible Markup Language (XML) provides a structured data format for these disparate services, allowing independent operating systems and programming languages to communicate and provide services to one another.

Network Services in the Data Center

The data center network is the end-to-end transport for multi-tiered applications and service-based architecture. The network is in an ideal position to deploy intelligent services providing resiliency, security, and application optimization because it already manages communication between enterprise applications. Thus, the following two fundamental questions must be answered:

- How should the network services be deployed, such as in a consolidated or dispersed design?
- What hardware platforms should be considered or used for those network services?

Figure 4-3 illustrates a collapsed multi-tier design. This is the recommended design for data center networks. In Figure 4-2, the access layer switches provide port connectivity for the server farms and uplink connectivity to an aggregation layer. The aggregation layer provides a communication path between servers and a link to the remaining enterprise via the core.

Figure 4-3   Collapsed Multi-tier Design
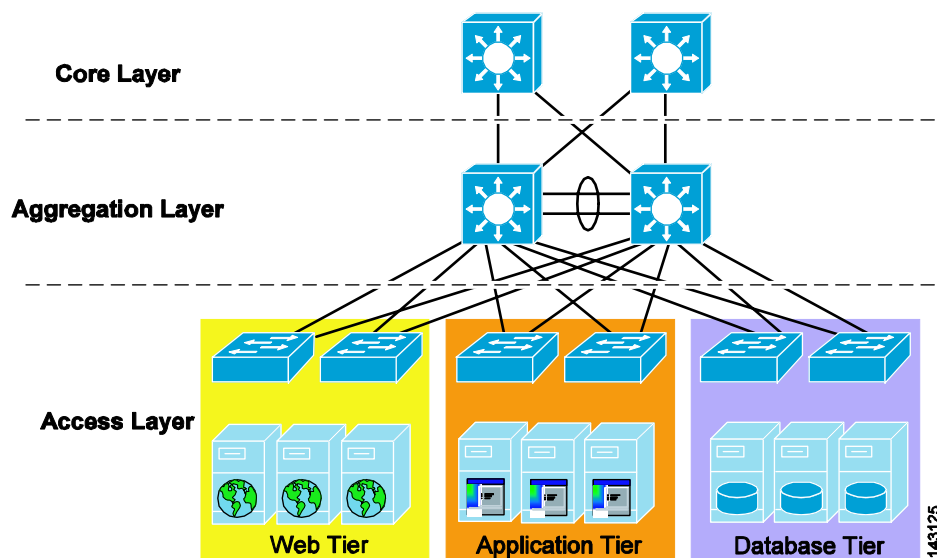
This design is flexible and scalable, allowing many different operating systems and applications to share the same network infrastructure. Segmentation of the web, application and database tiers is achievable using VLANs; providing a logical boundary between the respective enterprise applications.

> **Note**   For more information about the Cisco Data Center Infrastructure, see the *Cisco Data Center Infrastructure SRND v. 2.0* at the following URL: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCI_SRND_2_5_book.html.

Centralized or Distributed Services

The aggregation layer of the data center is an ideal location to provide centralized network services because it transports client-to-server and server-to-server communication within the data center. The aggregation layer is a control point where network services can be deployed and shared across server farms and their software applications. The maturing set of standards defining enterprise service-oriented architectures improve the effectiveness of centralized network-based services. These standards improve effectiveness by providing well-understood protocols and message formats supporting application communications.

Figure 4-4 shows the collapsed multi-tier architecture with network services available at the aggregation layer. These services include the following:

- Server load balancing
- Firewalls
- Network analysis
- SSL offloading
- Intrusion detection
- Content caching
- VPN termination

Figure 4-4   Aggregation Layer Services
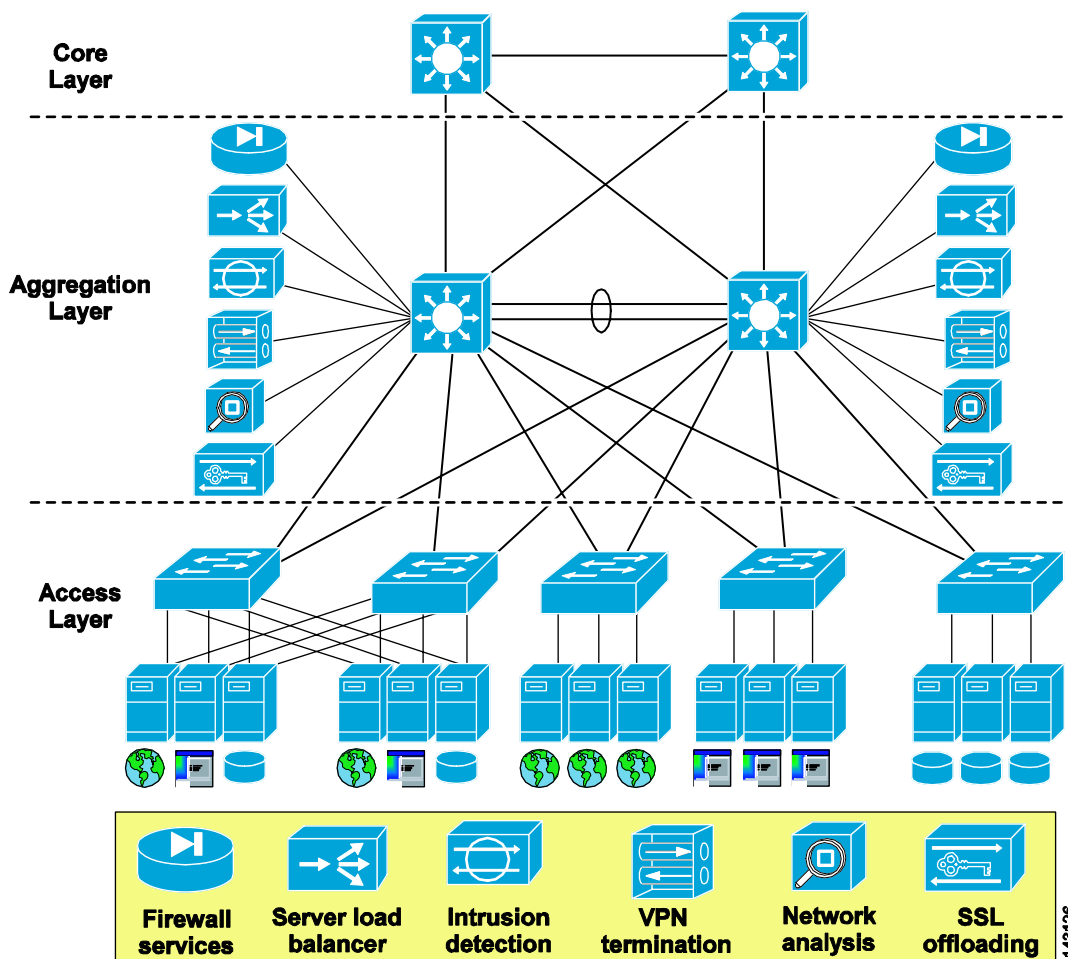
The shared data center network infrastructure provides segmentation and services where they are needed in a manageable and scalable fashion. These services may be provided by appliances or integrated service modules. If services are incorporated into the switching platform, as with the Catalyst 6500 service modules, the following significant advantages are gained by the enterprise:

- Services modules provide a single service configuration point
- Service modules are aware of the logical network topology and ease the consolidation and virtualization of the data center
- Service modules provide enterprise class performance and hardware specifically designed for the service function
- Service modules optimize rack space and require no additional cabling, power, or cooling

Distributed services are another approach to optimizing application performance in the data center. Figure 4-5 illustrates the distributed approach concept. In this design, network services are located at the edge of the enterprise data center adjacent to the server farms. As shown in Figure 4-5, server load balancing and firewall services are made available using network devices and dedicated software platforms.

Figure 4-5   Distributed Service Design
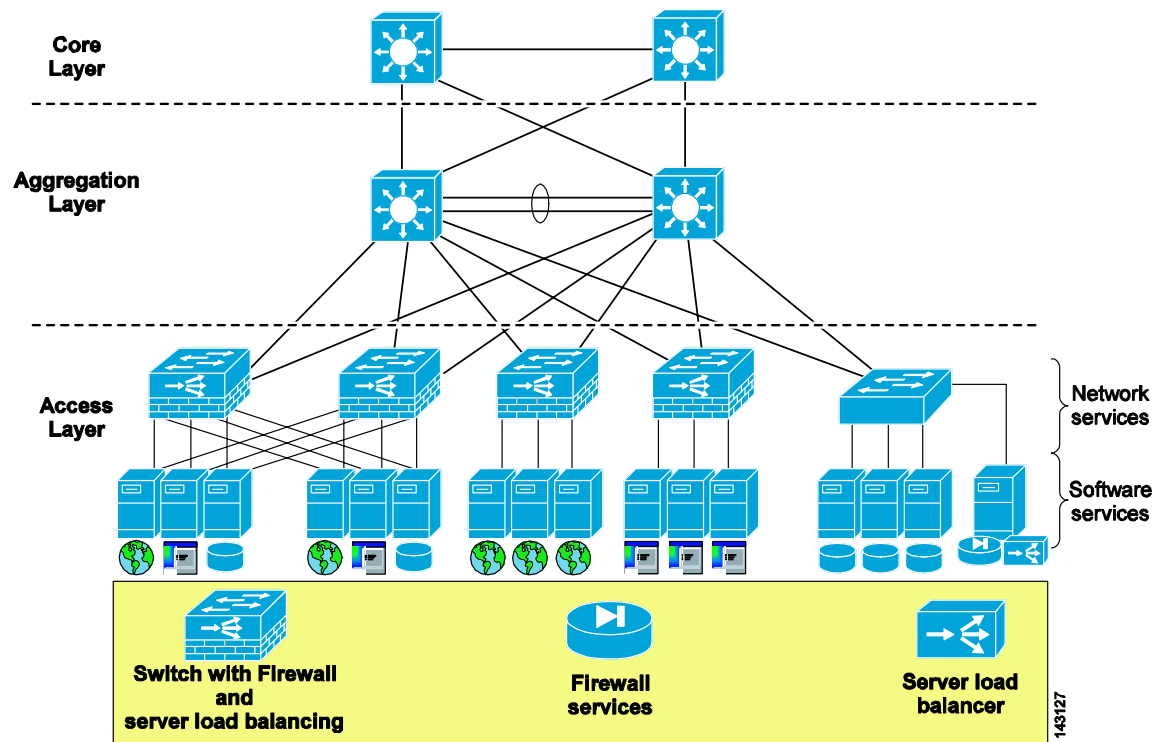
When compared to a centralized model, the distributed service design requires more resources and devices to provide equivalent service coverage to the server farm. This device proliferation has an impact on the manageability and flexibility of the data center. For example, instead of using a single server load balancer in a centralized design, the enterprise may need to use separate load balancers that manage traffic for servers in their immediate locale. The distributed service design does not take advantage of the traffic aggregation occurring at the aggregation layer. As more applications, servers, and services are required the enterprise is compelled to add more devices in the access layer to meet the demand.

Typically, software-based services such as server load balancing and firewalls do not provide the same level of performance as a dedicated hardware-based solution. As service-oriented architectures continue to grow, the increase in inter-server messaging requires greater performance and intelligence from the service providing devices in the data center.

Design and Implementation Details

This section details the integration of the Cisco Content Switching Module (CSM) and the Cisco Firewall Services Module providing centralized services in the enterprise data center. It includes the following topics:

- CSM One-Arm Design in the Data Center
- Architecture Details
- Configuration Details
- Configuration Listings

The design specifically addresses a multi-tier deployment of an IBM WebSphere application on a blade system platform. The CSM one-arm design provides centralized server load balancing and firewall services for the application.

CSM One-Arm Design in the Data Center

The availability, scalability, and security of enterprise applications are typically dependent on the services provided by devices such as load balancers and firewalls. The CSM one-arm design promotes these goals and efficiently integrates these services within an N-tier application architecture. The design permits server load balancing and firewall functionality to be selectively applied at each tier of the architecture via the CSM and FWSM. This efficiency improves the overall performance of the data center by providing application services only where those services are required.

Figure 4-6 shows a logical view of the CSM one-arm design. The CSM and the FWSM physically reside within the same Catalyst 6500 chassis and are logically positioned to optimize the performance and services of the data center. The CSM is removed from any direct traffic flows, and relies on the advanced routing capabilities of the MSFC to make its services available. Conversely, situating the FWSM in front of the server farm and between the application and database tiers allows the use of security services where they are most often required.

Figure 4-6   CSM One-Arm Design Logical View
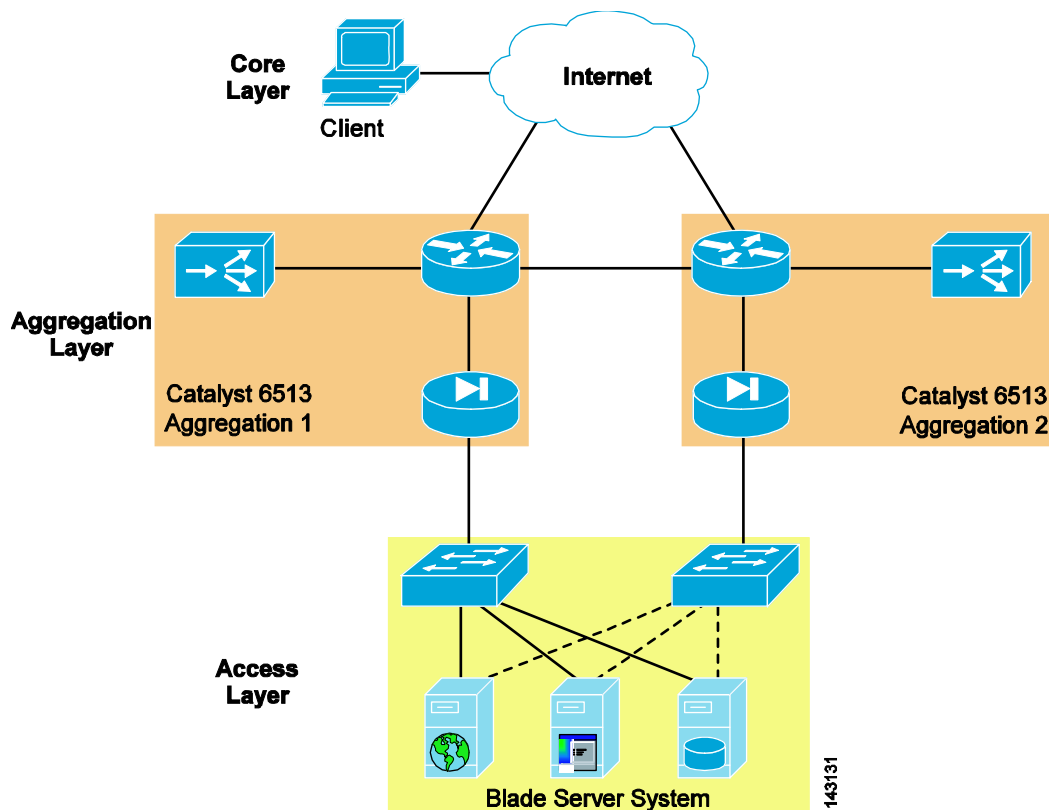
The CSM one-arm design follows the principal of centralized data center services, allowing web, application, and database servers to use the same service device located at the aggregation layer of the data center.

Traffic Pattern Overview

This section describes the traffic pattern in the data center for the following flows:

- Client-to-server
- Server-to-server

Client-to-Server

Figure 4-7 shows the client-to-server traffic flow through the data center when using the CSM in one-arm mode. The client is attempting to reach a web page located on a blade server in the enterprise data center.

Figure 4-7   Client-to-Server Traffic Flow
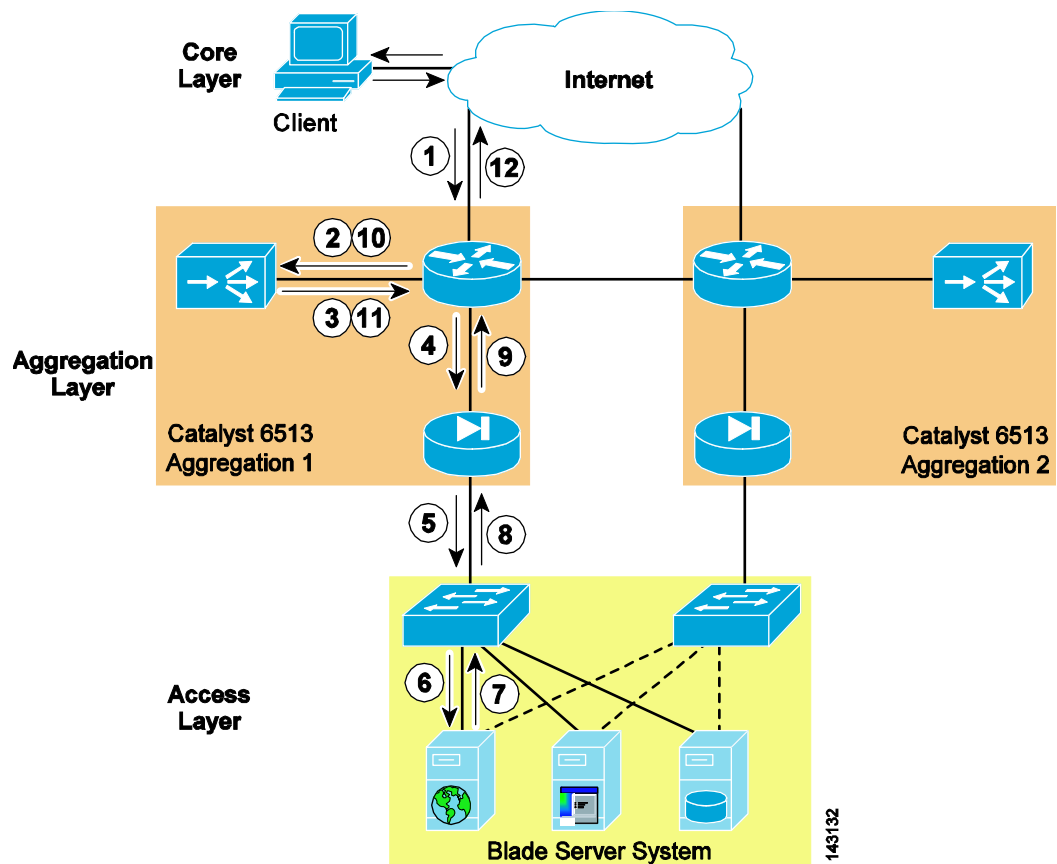
A successful transaction with the one-arm data center design includes the following sequence of events:

1. Client requests a URL associated with the CSM virtual IP (VIP) address.

2. The MSFC routes the request to the CSM VIP.

3. The CSM makes a load balancing decision and selects a real server. At this point, the CSM replaces the VIP address with the IP address of the real server if the **nat server** command is present in the virtual server configuration. The CSM forwards the request to its default gateway on the MSFC using the destination IP address of the real server and the source address of the client.

4. The MSFC routes the request to the real server protected by the FWSM.

5. The FWSM is bridging traffic between the "inside" and "outside" networks, applying the appropriate security policies to the network segment.

6. The switch forwards the traffic to the real server.

7. The real server forwards a reply to its default gateway the MSFC.

8. The FWSM receives the traffic from the access switch.

9. The FWSM forwards the traffic to the MSFC.

10. The MSFC uses policy-based routing (PBR) on the interface to forward the return traffic to the CSM.

11. The CSM rewrites the source IP address of the return traffic from the real server IP address to the VIP of the CSM. The rebuilt packet is sent to the default gateway of the CSM, the MSFC.

12. The MSFC forwards the reply to the client.

Server-to-Server

Figure 4-8 illustrates the traffic flow through the data center when using the CSM server load balancing with server-to-server traffic. In this scenario, a blade server hosting a web application is connecting through the load balancer in one-arm mode to another blade server hosting a middle-tier application.

Figure 4-8   Traffic Flow with Server-to-Server Load Balancing via the CSM
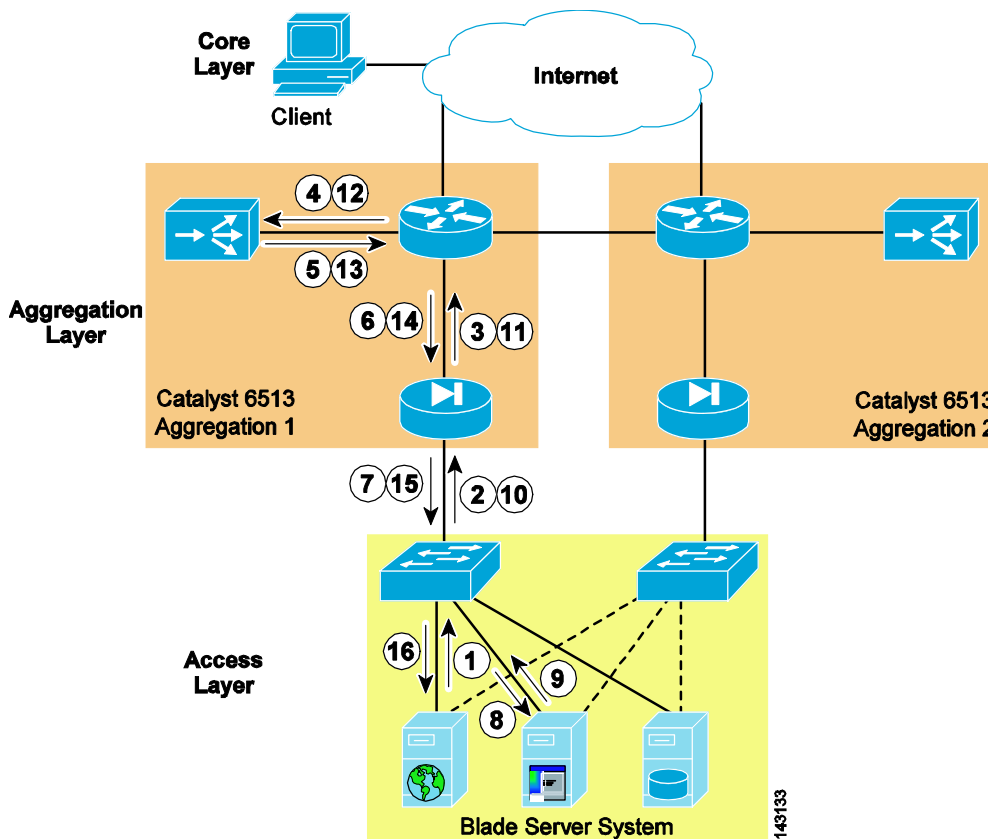
The following sequence of events result in a successful connection for the scenario shown in Figure 4-8:

1. The web server initiates a connection to the CSM VIP.

2. The firewall receives the traffic on its inside interface.

3. The firewall bridges the traffic to the default gateway of the web server, the MSFC.

4. The MSFC routes the traffic to the CSM alias address that is the static route for the CSM VIP.

5. The CSM selects a real application server based on a load balancing algorithm. It then performs server NAT and forwards the traffic to its default gateway located on the MSFC, using the destination IP address of the real application server and the source address of the web server.

6. The MSFC routes the traffic to the application tier through the FWSM.

7. The FWSM receives the packet and applies the appropriate security policies for that application tier network segment.

8. The switch forwards the traffic to the real application server.

9. The application server sends its response to the connection request to its default gateway located on the MSFC.

10. The FWSM receives the response on its inside interface of the application tier network segment.

11. The FWSM forwards the response to the MSFC located on the outside of the application network segment.

12. The MSFC applies PBR based on the Layer 4 port information and routes the traffic to the MSFC.

13. The CSM rewrites the source IP address of the return traffic from the real server IP address to the VIP of the CSM. The rebuilt packet is sent to the default gateway of the CSM, the MSFC.

14. The MSFC routes the traffic to the web tier through the outside interface of the FWSM.

15. The FWSM performs its packet filtering functions and bridges the traffic to its inside interface on the web tier network segment.

16. The packet is sent to the web server that initiated the transaction to the application tier.

Architecture Details

This section documents the application and network topology of the test bed and includes the following topics:

- WebSphere Solution Topology
- WebSphere Solution Topology with Integrated Network Services
- Additional Service Integration Options

WebSphere Solution Topology

This section is an overview of a test application topology. It identifies the hardware, software, and applications used during testing.

Software

Red Hat AS 3.0 is the operating system on each of the blade servers in the test environment. The WebSphere implementation used the following IBM software:

- IBM HTTP Server version 2.0.47
- IBM WebSphere Application Server (WAS) 5.1.0.3
- IBM DB2 UDB v8.1

The applications used to verify the functionality of the integrated network services design and a multi-tier blade server deployment were the following:

- Trade3
- Petstore

IBM provides each of these sample applications with their WebSphere installations for benchmarking performance and the functionality of a WebSphere-based solution.

Hardware

The following server platforms are used to host the WebSphere environment:

- A single IBM BladeCenter
- Seven HS20 blade servers with two Gigabit Ethernet adapters
  - Blades 1–3 host the IBM HTTP servers
  - Blades 4–6 host the WAS servers
  - Blade 7 hosts the DB2 database

Topology

Figure 4-9 illustrates the WebSphere application topology. A user request made via the web browser reaches the HTTP server. The HTTP server uses an XML file referred to as a plug-in file. This plug-in is part of the HTTP server process. The plug-in decides which traffic the HTTP server should handle locally and which traffic to direct to the application servers.

Figure 4-9   *WebSphere Solution Topology*
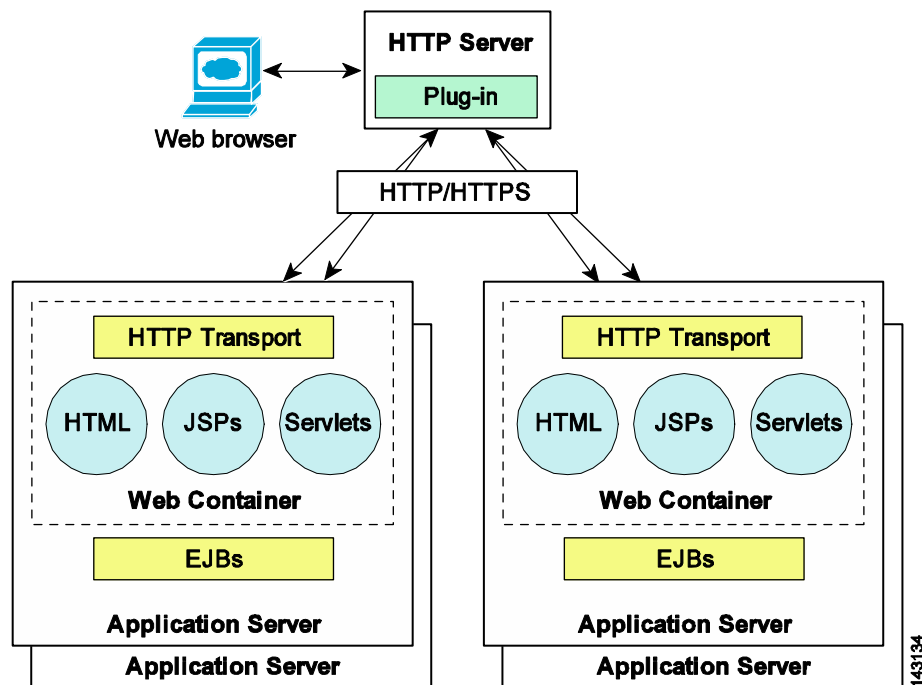
The plug-in file routes request to the application servers. In addition, the plug-in file can load balance the traffic to application servers using a round robin, weighted round robin, or random algorithm. The transport method between the HTTP server and the web container defined on the application server is HTTP or HTTPS. The plug-in creates persistent connections between itself and the web container for service requests.

Note The plug-in does not actively monitor the health of the application servers. The plug-in-cfg.xml file is generated on one WebSphere application server and imported into the HTTP server. This assumes that all of the application servers have identical configurations.

WebSphere Solution Topology with Integrated Network Services

This section discusses the introduction of network services into the WebSphere solution topology. The main topics include the following:

- Hardware
- Software
- Topology
- Test topology

Hardware

The network equipment used for this testing includes the following:

- Cisco Catalyst 6500 with Supervisor 720 for wire speed PBR
- Cisco Intelligent Gigabit Ethernet Switch (CIGESM) for blade server access layer connectivity
- Cisco CSM for load balancing and health checking functionalities
- Cisco FWSM for security between the application tiers

Software

The images used for this testing includes the following:

- Cisco Native Internetwork Operating System (IOS) software version 12.2(18)SXD5 for the Catalyst 6500 Supervisor 720s
- Cisco IOS software version 12.1(22)AY1 for the CIGESMs
- Cisco CSM software version 4.2(2)
- Cisco FWSM software version 2.3(2) allowing the creation of virtual firewall instances

Note The WebSphere application environment remains unchanged; see WebSphere Solution Topology, page 4-11 for more details on the applications used.

Topology

The integration of advanced load balancing and security functionality into the WebSphere application environment is achievable with the Cisco data center architecture. Figure 4-10 illustrates the logical position of these services within a WebSphere solution. In Figure 4-10, positioning a load balancer and a firewall in front of the HTTP server farm and the WAS servers provides an increased level of security, scalability, and high availability to the WebSphere solution.

Figure 4-10   WebSphere Solution Topology with Integrated Services
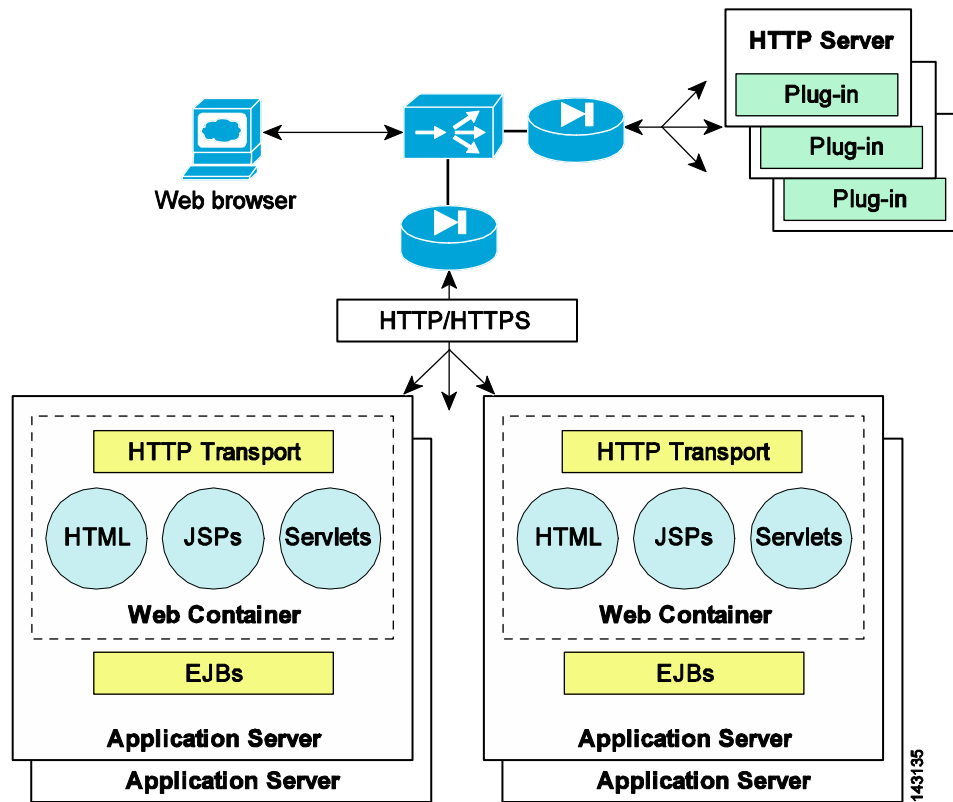
The load balancer, such as a CSM, provides advanced health monitoring techniques to track the state of both the HTTP and WAS servers. The CSM supports the following methods to verify server availability:

- Probes (ICMP, SMTP, DNS, HTTP, FTP)
- Return code checks (monitors the success of TCP transactions between clients and servers)
- In-band monitoring (examines HTTP return codes from the servers)

The load balancing algorithms supported by the CSM include the following:

- Source/destination-based hashes
- URL hashing
- Least number of connections
- Round robin
- Weighted round robin

These algorithms allow the server farm to share the traffic load efficiently and scale effectively.

In addition, the CSM is able to provide session affinity because it understands how to read the cookie or URL data created by the WebSphere applications. With this piece of information, the CSM effectively binds a user of an application to a single server for the duration of their session. Session affinity to a single device allows the application to use the local cache of the server to retrieve session state information and not the resources of a shared session state database or another instance of the application server.

Typically, the HTTP server provides session persistence in a WebSphere solution, using processor resources to ensure affinity between clients and applications servers. By positioning the CSM logically in front of the web and application tiers, the network is able to provide session affinity. In addition to health monitoring and server load balancing, the CSM relieves the web server of its session affinity and load balancing responsibilities. This CSM functionality delivers an increase in the overall performance of the application and a decrease in its complexity.

Note   The cookie or URL associated with J2EE applications is also known as a *Clone ID*. In either case, the name of the cookie or URL is "jsession" or "JSESSION." The clone ID/jsession definitions can be found in the plug-in-cfg.xml file generated by the WebSphere Application Server and used by the HTTP server.

Figure 4-10 indicates firewall services between the web and application tiers. The logical segmentation of the network via VLANs provides this firewall. Administrators can apply granular security policies and traffic filtering rules to each network segment: web, application, and database.

The load balancer provides an increased level of availability by removing unresponsive servers from its load balancing calculations, and by making server farm transactions more efficient. Combined with the security services of a firewall, WebSphere application traffic can be secure and optimized.

Test Topology

Figure 4-11 illustrates the logical topology of the test WebSphere application solution. A single blade system houses the web, application, and database tiers of the server farm. Each blade server is dual-homed to an integrated Ethernet switch. The network interface controllers (NICs) form a team in an active/standby configuration.

The integrated blade switches connect to the aggregation layer Catalyst 6500s using LACP (802.3ad) to bundle four Gigabit Ethernet ports into a single logical channel.

The channel uses the trunk failover feature of the integrated blade switch. Trunk failover is a high availability mechanism that binds the link state of an external uplink with the internal server ports on the blade switch.

A pair of Catalyst 6500 switches comprises the aggregation layer in the test network. The MSFC of each switch is the default gateway for the web, application, and database servers. In addition, the switches host the CSM and FWSM to provide integrated network services to the server farms.

Figure 4-11   Test Bed Topology
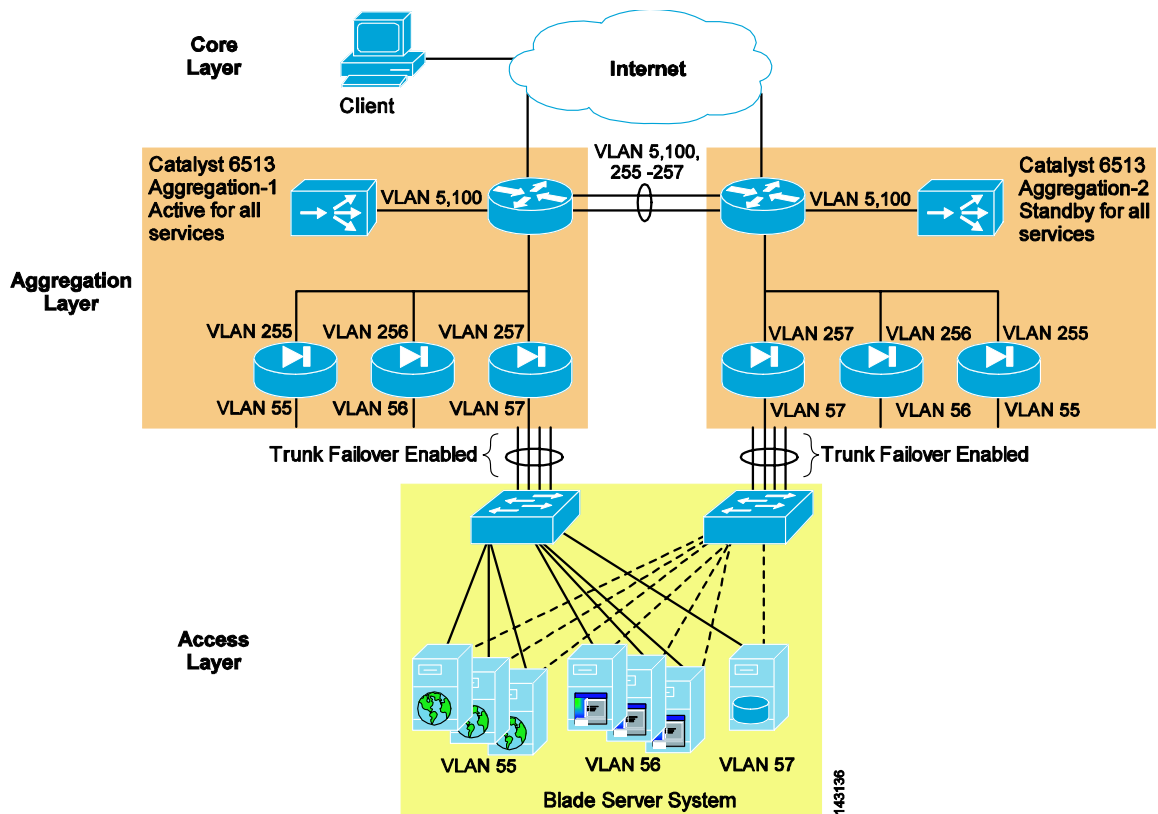
The CSM is in one-arm mode. The primary benefit of deploying the CSM in this fashion is the ability to select via routing which traffic uses the load balancer.

The FWSM has virtual capabilities, meaning there is a single physical device logically divided into three individual firewalls, called *contexts*. Each virtual firewall context bridges traffic between the inside and outside interfaces of this N-tier environment. In this manner, the FWSM provides granular traffic filtering that is independently applied by each virtual firewall context.

For more information, see the following topics:

- For policy-based routing, see *Cisco IOS Quality of Service Solutions Configuration Guide, Release 12.2* at the following URL:
  http://www.cisco.com/en/US/docs/ios/12_2/qos/configuration/guide/qcfclass.html.

Additional Service Integration Options

This document addresses the integration of load balancing and security services in an N-tier environment, specifically a WebSphere application solution. Server load balancing and security are fundamental services used by data center applications. However, these are not the only integrated network services available for the enterprise. The following network services are also available as service modules and/or appliances:

- SSL offloading
- Intrusion prevention systems
- Intrusion detection systems

- Network analysis devices
- Caching devices

Configuration Details

This section details the modifications necessary in the web and application servers to use the load balancing services of the CSM. The firewall services are transparently provided via the FWSM and do not require any modifications at the server level.

IBM HTTP Server

The IBM HTTP Server (IHS) can route requests to WebSphere application servers located on remote machines. The plug-in file defines the application servers and available services. The plug-in file provides the following advantages:

- XML-based configuration file
- Standard protocol (HTTP) recognized by firewalls
- Secure transport using HTTPS

The WebSphere application server creates the plug-in file that must be manually installed into the IHS server.

The example below is a sample of a plug-in-cfg.xml file used by the IHS server. The virtual host group identifies requests that should be handled by the WebSphere application servers. In this example, all requests received on port 80 and 9080 are routed to an application server and are not serviced locally by the web server. A host name, IP address, or wildcard character may be used to define the HTTP header hosts.

```
<VirtualHostGroup Name="default_host">
     <VirtualHost Name="*:9080"/>
     <VirtualHost Name="*:80"/>
     <VirtualHost Name="www.example.com:9443"/>
</VirtualHostGroup>
```

The ServerCluster element identifies WebSphere application servers that are configured to service the same type of requests. A cluster may contain one server or multiple servers and use either a round robin or random load balancing algorithm to direct requests to WebSphere application servers.

In the following sample of the plug-in file, the server cluster contains a single server, the CSM. The host names "LoadBalancer-WAS" and "LoadBalancer-WASSSL" refer to the DNS names given to the virtual IP addresses of the CSM listening on ports 9080 and 9443. This configuration allows the CSM to load balance traffic between the IHS web server and the WebSphere application servers.

```
<ServerCluster LoadBalance="Round Robin" Name="CSMVIPS">
 <Server ConnectTimeout="0" MaxConnections="-1" Name="CSM">
  <Transport Hostname="LoadBalancer-WAS" Port="9080" Protocol="http"/>
  <Transport Hostname="LoadBalancer-WASSSL" Port="9443" Protocol="https">
    <Property Name="keyring" Value="/opt/WebSphere/AppServer/etc/plugin-key.kdb"/>
    <Property Name="stashfile" Value="/opt/WebSphere/AppServer/etc/plugin-key.sth"/>
        </Transport>
    </Server>
</ServerCluster>
```

Uniform Resource Identifiers (URI) are strings that identify application servers. The HTTP request includes the URI as a cookie or as part of the URL. This piece of information provides session persistence and affinity when used by the HTTP server or the CSM. In the case of WebSphere applications, the cookie and the URL are named "JSESSIONID" or "jsessionid" as shown in this snippet of the plug-in-cfg.xml.

```
<UriGroup Name="default_CSMVIPS_URIs">
  <Uri AffinityCookie="JSESSIONID" AffinityURLIdentifier="jsessionid" Name="/petstore/*"/>
</UriGroup>
```

When present, the jsessionid allows the IHS server to send inbound requests to the originating WebSphere application server that is maintaining session affinity between the client and the application server.

Note The CSM may use the URI (jsessionid) of the application servers to provide session persistence, simplifying the configuration of the IHS server.

Note For more information on the plug-in-cfg.xml file, see *WebSphere Redbook Domain* at the following URL: http://www.redbooks.ibm.com/Redbooks.nsf/portals/WebSphere

IBM WebSphere Application Server

The IBM WebSphere server requires no changes to interoperate with the CSM or FWSM. Policy-based routing on the MSFC manages the direction of traffic between the web and application tiers of the example topology. The application server is unaware of these actions.

Note PBR on the MSFC is performed in hardware.

Configuration Listings

The configuration listings here detail only the portions relevant to the topology described in this document (shown in Figure 4-11.)

Aggregation1 (Primary Root and HSRP Active)

```
firewall multiple-vlan-interfaces
firewall module 2 vlan-group 123
firewall vlan-group 123  55-58,77,255-258,260
vtp domain lab
vtp mode transparent
udld enable

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
spanning-tree vlan 5,13,55-57,77,255-257 priority 24576
port-channel load-balance src-dst-port
!
```

```
vlan internal allocation policy ascending
vlan dot1q tag native
!
vlan 2
!
vlan 5
 name csm
!
vlan 13
 name CSMft
!
vlan 55
 name ihs_Blades
!
vlan 56
 name was_Blades
!
vlan 57
 name db2_blades
!
vlan 77
 name FWSMft
!
vlan 255
 name IHS_Blades
!
vlan 256
 name WAS_Blades
!
vlan 257
 name DB2_Blades
!
interface Port-channel4
 description <<** Channel between two aggregation switches **>>
 no ip address
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 5,13,55-57,77,255-257
 switchport mode trunk
!
!
interface GigabitEthernet9/16
 description <<**  Port-channel 4 **>>
 no ip address
 logging event link-status
 load-interval 30
 speed 1000
 switchport
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 5,13,55-57,77,255-257
 switchport mode trunk
 channel-protocol lacp
 channel-group 4 mode active
!
!
interface Vlan5
 description <<**  CSM VLAN **>>
 ip address 10.5.1.2 255.255.0.0
 no ip redirects
 no ip proxy-arp
 logging event link-status
 load-interval 30
 standby 1 ip 10.5.1.1
 standby 1 timers 1 3
```

```
 standby 1 priority 51
 standby 1 preempt delay minimum 120
 standby 1 name csm
!
interface Vlan255
 description <** IHS Web Server Default Gateway **>>
 ip address 10.55.1.2 255.255.0.0
 no ip redirects
 no ip proxy-arp
 ip policy route-map server-client-traffic
 logging event link-status
 load-interval 30
 standby 1 ip 10.55.1.1
 standby 1 timers 1 3
 standby 1 priority 51
 standby 1 preempt delay minimum 120
 standby 1 name ibmwebservers
!
interface Vlan256
 description <<** IBM WAS Default Gateway  **>>
 ip address 10.56.1.2 255.255.0.0
 no ip redirects
 no ip proxy-arp
 ip policy route-map was-csm-traffic
 logging event link-status
 load-interval 30
 standby 1 ip 10.56.1.1
 standby 1 timers 1 3
 standby 1 priority 51
 standby 1 preempt delay minimum 120
 standby 1 name ibmWAS
!
interface Vlan257
 description <<** IBM DB2 Server Default Gateway **>>
 ip address 10.57.1.2 255.255.0.0
 no ip redirects
 no ip proxy-arp
 logging event link-status
 load-interval 30
 standby 1 ip 10.57.1.1
 standby 1 timers 1 3
 standby 1 priority 51
 standby 1 preempt delay minimum 120
 standby 1 name ibmdb2
!
access-list 156 permit tcp 10.56.1.0 0.0.0.255 eq 9080 10.55.1.0 0.0.0.255
access-list 156 deny   ip any any
!
route-map server-client-traffic permit 10
 set ip default next-hop 10.5.1.6
!
route-map was-csm-traffic permit 10
 match ip address 156
 set ip next-hop 10.5.1.6
!
end
```

Aggregation2 (Secondary Root and HSRP Standby)

The configuration of the second aggregation switch is similar to the primary aggregation switch except for the following:

```
!
spanning-tree vlan 5,13,55-57,77,255-257 priority 28672
!
interface Vlan5
 description <<**  CSM VLAN **>>
 ip address 10.5.1.3 255.255.0.0
 no ip redirects
 no ip proxy-arp
 logging event link-status
 load-interval 30
 standby 1 ip 10.5.1.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 name csm
!
interface Vlan255
 description <** IHS Web Server Default Gateway **>>
 ip address 10.55.1.3 255.255.0.0
 no ip redirects
 no ip proxy-arp
 ip policy route-map server-client-traffic
 logging event link-status
 load-interval 30
 standby preempt delay minimum 120 sync 120
 standby 1 ip 10.55.1.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 name ibmwebservers
!
interface Vlan256
 description <<** IBM WAS Default Gateway  **>>
 ip address 10.56.1.3 255.255.0.0
 no ip redirects
 no ip proxy-arp
 ip policy route-map was-csm-traffic
 logging event link-status
 load-interval 30
 standby preempt delay minimum 120 sync 120
 standby 1 ip 10.56.1.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 name ibmWAS
!
interface Vlan257
 description <<** IBM DB2 Server Default Gateway **>>
 ip address 10.57.1.3 255.255.0.0
 no ip redirects
 no ip proxy-arp
 logging event link-status
 load-interval 30
 standby preempt delay minimum 120 sync 120
 standby 1 ip 10.57.1.1
 standby 1 timers 1 3
 standby 1 priority 50
 standby 1 name ibmdb2
!
```

CSM (Active)

```
module ContentSwitchingModule 3
 variable NO_TIMEOUT_IP_STICKY_ENTRIES 1
```

```
!
 ft group 13 vlan 13
  priority 11
  heartbeat-time 1
  failover 3
  preempt
!
 vlan 5 server
  ip address 10.5.1.4 255.255.0.0
  gateway 10.5.1.1
  alias 10.5.1.6 255.255.0.0
!
probe IHS_BLADES http
  request method get url /petstore/
  expect status 200  205
  interval 5
  failed 3
!
serverfarm IHS_BLADES
  nat server
  no nat client
  real 10.55.1.101
   inservice
  real 10.55.1.102
   inservice
  real 10.55.1.103
   inservice
  probe IHS_BLADES
!
 serverfarm WAS_BLADES
  nat server
  no nat client
  real 10.56.1.102
   inservice
  real 10.56.1.103
   inservice
  real 10.56.1.101
   inservice
!
 sticky 1 cookie JSESSIONID timeout 10
  cookie secondary jsessionid
!
policy APP_TIER
  sticky-group 1
  serverfarm WAS_BLADES
!
 policy FRONT_END
  serverfarm IHS_BLADES
!
vserver VBLADES
  virtual 10.10.10.55 tcp www
  replicate csrp sticky
  replicate csrp connection
  persistent rebalance
  slb-policy FRONT_END
  inservice
!
 vserver VWAS
  virtual 10.10.10.56 any
  replicate csrp sticky
  replicate csrp connection
  persistent rebalance
  slb-policy APP_TIER
  inservice
```

```
 !
```

CSM (Standby)

The same configuration exists on the standby CSM located on the second aggregation switch except for
the following:

```
 ft group 13 vlan 13
  priority 9
  heartbeat-time 1
  failover 3
  preempt
!
 vlan 5 server
  ip address 10.5.1.5 255.255.0.0
  gateway 10.5.1.1
  alias 10.5.1.6 255.255.0.0
!
```

FWSM (Active)

```
firewall transparent
!
failover
failover lan unit primary
failover lan interface fover vlan 77
failover polltime unit msec 500 holdtime 3
failover polltime interface 3
failover interface-policy 1%
failover replication http
failover link fover vlan 77
failover interface ip fover 10.77.1.1 255.255.0.0 standby 10.77.1.2



admin-context admin
context admin
  allocate-interface vlan146-vlan147 int1-int2
  config-url disk:admin.cfg
!
context tp255-55
  description <** Bridges vlan 255 - 55 (IHS servers) **>>
  allocate-interface vlan55
  allocate-interface vlan255
  config-url disk:/tp255-55.cfg
!

context tp256-56
  description <** Bridges vlan 256 - 56 (WAS servers) **>>
  allocate-interface vlan56
  allocate-interface vlan256
  config-url disk:/tp256-56.cfg
!

context tp257-57
  description <** Bridges vlan 257 - 57 (DB2 servers) **>>
  allocate-interface vlan57
  allocate-interface vlan257
  config-url disk:/tp257-57.cfg
```

```
!
: end
Sample of one of the virtual firewall contexts:
FWSM/tp255-55# sh run
: Saved
:
FWSM Version 2.3(2) <context>
firewall transparent
nameif vlan255 outside security0
nameif vlan55 inside security100
enable password 8Ry2YjIyt7RRXU24 encrypted
passwd 2KFQnbNIdI.2KYOU encrypted
hostname tp255-55
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 101 extended permit ip any any
access-list 102 ethertype permit bpdu
no pager
logging on
logging buffer-size 4096
mtu outside 1500
mtu inside 1500
ip address  10.55.1.4 255.255.0.0
icmp permit any outside
icmp permit any inside
pdm location 10.55.0.0 255.255.0.0 inside
pdm logging notifications 100
pdm history enable
arp timeout 14400
static (inside,outside) 10.55.0.0 10.55.0.0 netmask 255.255.0.0
access-group 102 in interface outside
access-group 101 in interface outside
access-group 102 in interface inside
access-group 101 in interface inside
!
interface outside
!
interface inside
!
route outside 0.0.0.0 0.0.0.0 10.55.1.1 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 outside
fragment chain 24 outside
fragment size 200 inside
fragment chain 24 inside
telnet timeout 5
ssh timeout 5
terminal width 511
: end
```

> **Note** The security context listed is an example configuration and does not adhere to security best practices.

FWSM (Standby)

The configuration on the standby FWSM located on the second aggregation switch is the same except for the following:

```
failover
failover lan unit secondary
failover lan interface fover vlan 77
failover polltime unit msec 500 holdtime 3
failover polltime interface 3
failover interface-policy 1%
failover replication http
failover link fover vlan 77
failover interface ip fover 10.77.1.1 255.255.0.0 standby 10.77.1.2
```

Access Layer (Integrated Switch)

The integrated blade server switches use the same configuration, providing a 4 GigE uplink using the trunk failover feature to the aggregation layer.

```
vtp mode transparent
link state track 1
!
port-channel load-balance src-ip
!
spanning-tree mode rapid-pvst
no spanning-tree optimize bpdu transmission
spanning-tree extend system-id
spanning-tree pathcost method long
!
vlan 2
 name operational
!
vlan 55
 name IHS_Blades
!
vlan 56
 name WAS_Blades
!
vlan 57
 name DB2_Blades
!
interface Port-channel1
 description <<** Channel to Aggregation Layer **>>
```

```
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55-57
 switchport mode trunk
 load-interval 30
 link state group 1 upstream
!
interface GigabitEthernet0/1
 description blade1
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/2
 description blade2
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/3
 description blade3
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/4
 description blade4
 switchport trunk native vlan 2
 switchport trunk allowed vlan 56
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/5
 description blade5
 switchport trunk native vlan 2
 switchport trunk allowed vlan 56
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/6
 description blade6
 switchport trunk native vlan 2
 switchport trunk allowed vlan 56
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
```

```
 spanning-tree bpduguard enable
!
interface GigabitEthernet0/7
 description blade7
 switchport trunk native vlan 2
 switchport trunk allowed vlan 57
 switchport mode trunk
 load-interval 30
 link state group 1 downstream
 spanning-tree portfast trunk
 spanning-tree bpduguard enable
!
!
interface GigabitEthernet0/17
 description <<** Uplink Channel **>>
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55-57
 switchport mode trunk
 load-interval 30
 channel-group 1 mode passive
!
interface GigabitEthernet0/18
 description <<** Uplink Channel **>>
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55-57
 switchport mode trunk
 load-interval 30
 channel-group 1 mode passive
!
interface GigabitEthernet0/19
 description <<** Uplink Channel **>>
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55-57
 switchport mode trunk
 load-interval 30
 channel-group 1 mode passive
 channel-protocol lacp
!
interface GigabitEthernet0/20
 description <<** Uplink Channel **>>
 switchport trunk native vlan 2
 switchport trunk allowed vlan 55-57
 switchport mode trunk
 load-interval 30
 channel-group 1 mode passive
 channel-protocol lacp
!
end
```