Campus Network for High Availability Design Guide

Cisco Validated Design I

January 25, 2008

Americas 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: 408 527-0883

Customer Order Number:
Text Part Number: OL-15734-01

Cisco Validated Design

The Cisco Validated Design Program consists of systems and solutions designed, tested, and documented to facilitate faster, more reliable, and more predictable customer deployments. For more information visit www.cisco.com/go/validateddesigns.

ALL DESIGNS, SPECIFICATIONS, STATEMENTS, INFORMATION, AND RECOMMENDATIONS (COLLECTIVELY, "DESIGNS") IN THIS MANUAL ARE PRESENTED "AS IS," WITH ALL FAULTS. CISCO AND ITS SUPPLIERS DISCLAIM ALL WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE. IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THE DESIGNS, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

THE DESIGNS ARE SUBJECT TO CHANGE WITHOUT NOTICE. USERS ARE SOLELY RESPONSIBLE FOR THEIR APPLICATION OF THE DESIGNS. THE DESIGNS DO NOT CONSTITUTE THE TECHNICAL OR OTHER PROFESSIONAL ADVICE OF CISCO, ITS SUPPLIERS OR PARTNERS. USERS SHOULD CONSULT THEIR OWN TECHNICAL ADVISORS BEFORE IMPLEMENTING THE DESIGNS. RESULTS MAY VARY DEPENDING ON FACTORS NOT TESTED BY CISCO.

Campus Network for High Availability Design Guide

CONTENTS

Introduction 1-1

- Audience 1-1
- Document Objectives 1-1

Campus Network Design Overview 1-2

Design Recommendations Summary 1-2

- Tuning for Optimized Convergence 1-2
  - Access Layer Tuning 1-2
  - Distribution Layer Tuning 1-3
  - Core Layer Tuning 1-4
- Design Guidance Review 1-4
  - Layer 3 Foundations Services 1-4
  - Layer 2 Foundation Services 1-5
  - General Design Considerations 1-6

Hierarchical Network Design Model 1-7

- Hierarchical Network Design Overview 1-7
- Core Layer 1-8
- Distribution Layer 1-9
- Access Layer 1-10

Network and In-the-Box Redundancy 1-11

Foundation Services Technologies 1-14

- Layer 3 Routing Protocols 1-14
  - Using Triangle Topologies 1-14
  - Limiting L3 Peering to Transit Links 1-15
  - Ensuring Connectivity in Case of Failure 1-16
  - Tuning Load Balancing with Cisco Express Forwarding 1-19
- Layer 2 Redundancy—Spanning Tree Protocol Versions 1-22
  - Spanning Tree Protocol Versions 1-22
  - Best Practices for Optimal Convergence 1-23
- Trunking Protocols 1-25
  - Deploying Multiple VLANS on a Single Ethernet Link (Trunking) 1-26
  - Virtual Trunk Protocol 1-27
  - Dynamic Trunk Protocol 1-28
  - Preventing Double 802.1Q Encapsulated VLAN Hopping 1-29

Protecting Against One-Way Communication with UniDirectional Link Detection **1-31**

Link Aggregation—EtherChannel Protocol and 802.3ad **1-32**

Link Aggregation Protocol **1-33**

Using HSRP, VRRP, or GLBP for Default Gateway Redundancy **1-35**

Gateway Load Balancing Protocol **1-37**

Oversubscription and QoS **1-40**

Design Best Practices **1-43**

Daisy Chaining Dangers **1-43**

Asymmetric Routing and Unicast Flooding **1-45**

Designing for Redundancy **1-47**

Spanning VLANs Across Access Layers Switches **1-51**

Deploying the L2 /L3 Boundary at the Distribution Layer **1-51**

Routing in the Access Layer **1-53**

Deploying the L2/L3 Boundary at the Access Layer **1-53**

Comparing Routing Protocols **1-55**

Using EIGRP in the Access Layer **1-56**

Using OSPF in the Access Layer **1-57**

Summary **1-59**

Campus Network for High Availability Design Guide

Introduction

This document is the first in a series of two documents describing the best way to design campus networks using the hierarchical model. The second document, *High Availability Campus Recovery Analysis*, provides extensive test results showing the convergence times for the different topologies described in this document, and is available at the following website: http://www.cisco.com/application/pdf/en/us/guest/netsol/ns432/c649/cdccont_0900aecd801a89fc.pdf

This document includes the following sections:

- Campus Network Design Overview, page 2
- Design Recommendations Summary, page 2
- Hierarchical Network Design Model, page 7
- Network and In-the-Box Redundancy, page 11
- Foundation Services Technologies, page 14
- Design Best Practices, page 43
- Summary, page 59

Audience

This document is intended for customers and enterprise systems engineers who are building or intend to build an enterprise campus network and require design best practice recommendations and configuration examples.

Document Objectives

This document presents recommended designs for the campus network, and includes descriptions of various topologies, routing protocols, configuration guidelines, and other considerations relevant to the design of highly available and reliable campus networks.

Corporate Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Campus Network Design Overview

Designing a campus network may not appear as interesting or exciting as designing an IP telephony network, an IP video network, or even designing a wireless network. However, emerging applications like these are built upon the campus foundation. Much like the construction of a house, if the engineering work is skipped at the foundation level, the house will crack and eventually fail. If the foundation services and reference design in an enterprise network are not rock-solid, applications that depend on the services offered by the network like IP telephony, IP video, and wireless communications will eventually suffer performance and reliability challenges.

To continue the analogy, if a reliable foundation is engineered and built, the house will stand for years, growing with the owner through alterations and expansions to provide safe and reliable service throughout its life cycle. The same is true for an enterprise campus network. The design principles and implementation best practices described in this document are tried-and-true lessons learned over time. Your enterprise can take advantage of these lessons to implement a network that will provide the necessary flexibility as the business requirements of your network infrastructure evolve over time.

Design Recommendations Summary

This section summarizes the design recommendations presented in this document and includes the following topics:

- Tuning for Optimized Convergence, page 2
- Design Guidance Review, page 4

Tuning for Optimized Convergence

This section summarizes the recommendations for achieving optimum convergence in the access, distribution, and core layers, and includes the following topics:

- Access Layer Tuning, page 2
- Distribution Layer Tuning, page 3
- Core Layer Tuning, page 4

Access Layer Tuning

The following are the recommendations for optimal access layer convergence:

- Limit VLANs to a single closet whenever possible.

  There are many reasons why STP/RSTP convergence should be avoided for the most deterministic and highly available network topology. In general, when you avoid STP/RSTP, convergence can be predictable, bounded, and reliably tuned.

  Additionally, it should be noted that in soft failure conditions where keepalives (BPDU or routing protocol hellos) are lost, L2 environments fail open, forwarding traffic with unknown destinations on all ports and causing potential broadcast storms; while L3 environments fail closed, dropping routing neighbor relationships, breaking connectivity, and isolating the soft failed devices.

- If STP is required, use Rapid PVST+.

If you are compelled by application requirements to depend on STP to resolve convergence events, use Rapid PVST+. Rapid PVST+ is far superior to 802.1d and even PVST+ (802.1d plus Cisco enhancements) from a convergence perspective.

- Set trunks to on/on with no negotiate, prune unused VLANs, and use VTP transparent mode.

  When configuring switch-to-switch interconnections to carry multiple VLANs, set DTP to **on/on** with **no negotiate** to avoid DTP protocol negotiation. This tuning can save seconds of outage when restoring a failed link or node. Unused VLANs should be manually pruned from trunked interfaces to avoid broadcast propagation. Finally, VTP transparent mode should be used because the need for a shared common VLAN database is reduced.

- Match PAgP settings between CatOS and Cisco IOS software.

  When connecting a Cisco IOS software device to a CatOS device, make sure that PAgP settings are the same on both sides. The defaults are different. CatOS devices should have PAgP set to off when connecting to an Cisco IOS software device if EtherChannels are not configured.

- Consider EIGRP/Routing in the access layer.

  A routing protocol like EIGRP, when properly tuned, can achieve better convergence results than designs that rely on STP to resolve convergence events. A routing protocol can even achieve better convergence results than the time-tested L2/L3 boundary hierarchical design. However, some additional complexity (uplink IP addressing and subnetting) and loss of flexibility are associated with this design alternative. Additionally, this option is not as widely deployed in the field as the L2/L3 distribution layer boundary model.

Distribution Layer Tuning

The following are the recommendations for optimal distribution layer convergence:

- Use equal-cost redundant connections to the core for fastest convergence and to avoid black holes.

  While it is tempting to reduce cost by reducing links between the distribution nodes to the core in a partial mesh design, the complexity and convergence tradeoffs related to this design are ultimately far more expensive.

- Connect distribution nodes to facilitate summarization and L2 VLANs spanning multiple access layer switches where required.

  Summarization is required to facilitate optimum EIGRP or OSPF convergence. If summarization is implemented at the distribution layer, the distribution nodes must be linked or routing black holes occur.

  Additionally, in a less than optimal design where VLANs span multiple access layer switches, the distribution nodes must be linked by an L2 connection. Otherwise, multiple convergence events can occur for a single failure and undesirable traffic paths are taken after the spanning tree converges.

- Utilize GLBP/HSRP millisecond timers.

  Convergence around a link or node failure in the L2/L3 distribution boundary model depends on default gateway redundancy and failover. Millisecond timers can reliably be implemented to achieve sub-second (800 ms) convergence based on HSRP/GLBP failover.

- Tune GLBP/HSRP preempt delay to avoid black holes.

  Ensure that the distribution node has connectivity to the core before it preempts its HSRP/GLBP standby peer so that traffic is not dropped while connectivity to the core is established.

- Tune EtherChannel and CEF load balancing to ensure optimum utilization of redundant, equal-cost links.

Monitor redundant link utilization in the hierarchical model and take steps to tune both L2 (EtherChannel) and L3 (CEF) links to avoid under-utilization. Use L3 and L4 (UDP/TCP port) information as input to hashing algorithms.

When you use EtherChannel interconnections, use L3 and L4 information to achieve optimum utilization. When you use L3 routed equal-cost redundant paths, vary the input to the CEF hashing algorithm to improve load distribution. Use the default L3 information for the core nodes and use L3 with L4 information for the distribution nodes.

Core Layer Tuning

For optimum core layer convergence, build triangles, not squares, to take advantage of equal-cost redundant paths for the best deterministic convergence.

When considering core topologies, it is important to consider the benefits of topologies with point-to-point links. Link up/down topology changes can be propagated almost immediately to the underlying protocols. Topologies with redundant equal-cost load sharing links are the most deterministic and optimized for convergence measured in milliseconds.

With topologies that rely on indirect notification and timer-based detection, convergence is non-deterministic and convergence is measured in seconds.

Design Guidance Review

This section summarizes the campus network design recommendations, and includes the following topics:

- Layer 3 Foundations Services, page 4
- Layer 2 Foundation Services, page 5
- General Design Considerations, page 6

Layer 3 Foundations Services

The following are the design recommendations for Layer 3 foundation services:

- Design for deterministic convergence—triangles, not squares.

  Topologies where point-to-point physical links are deployed provide the most deterministic convergence. Physical link up/down is faster than timer-based convergence.

- Control peering across access layer links (passive interfaces).

  Unless you control L3 peering in the hierarchical campus model, the distribution nodes establish L3 peer relationships many times using the access nodes that they support, wasting memory and bandwidth.

- Summarize at the distribution.

  It is important to summarize routing information as it leaves the distribution nodes towards the core for both EIGRP and OSPF. When you force summarization at this layer of the network, bounds are implemented on EIGRP queries and OSPF LSA/SPF propagation, which optimizes both routing protocols for campus convergence.

- Optimize CEF for best utilization of redundant L3 paths.

The hierarchical campus model implements many L3 equal-cost redundant paths. Typical traffic flows in the campus cross multiple redundant paths as traffic flows from the access layer across the distribution and core and into the data center. Unless you vary the decision input for the CEF hashing algorithm at the core and distribution layers, CEF polarization can result in under-utilization of redundant paths.

Layer 2 Foundation Services

The following are the design recommendations for Layer 2 foundation services:

- Use Rapid PVST+ if you must span VLANs.

  If you are compelled by application requirements to depend on STP to resolve convergence events, use Rapid PVST+, which is far superior to 802.1d and even PVST+ (802.1d plus Cisco enhancements) from the convergence perspective.

- Use Rapid PVST+ to protect against user-side loops.

  Even though the recommended design does not depend on STP to resolve link or node failure events, STP is required to protect against user-side loops. There are many ways that a loop can be introduced on the user-facing access layer ports. Wiring mistakes, misconfigured end stations, or malicious users can create a loop. STP is required to ensure a loop-free topology and to protect the rest of the network from problems created in the access layer.

- Use the Spanning-Tree toolkit to protect against unexpected STP participation.

  Switches or workstations running a version of STP are commonly introduced into a network. This is not always a problem, such as when a switch is connected in a conference room to temporarily provide additional ports/connectivity. Sometimes this is undesirable, such as when the switch that is added has been configured to become the STP root for the VLANs to which it is attached. BDPU Guard and Root Guard are tools that can protect against these situations. BDPU Guard requires operator intervention if an unauthorized switch is connected to the network, and Root Guard protects against a switch configured in a way that would cause STP to converge when being connected to the network.

- Use UDLD to protect against one-way up/up connections.

  In fiber topologies where fiber optic interconnections are used, which is common in a campus environment, physical misconnections can occur that allow a link to appear to be up/up when there is a mismatched set of transmit/receive pairs. When such a physical misconfiguration occurs, protocols such as STP can cause network instability. UDLD detects these physical misconfigurations and disables the ports in question.

- Set trunks to **on/on** with **no negotiate**, prune unused VLANs, and use VTP transparent mode.

  When you configure switch-to-switch interconnections to carry multiple VLANs, set DTP to **on/on** with **no negotiate** to avoid DTP protocol negotiation. This tuning can save seconds of outage when restoring a failed link or node. Unused VLANs should be manually pruned from trunked interfaces to avoid broadcast propagation. Finally, VTP transparent mode should be used because the need for a shared VLAN database is lessened given current hierarchical network design.

- Match PAgP settings between CatOS and Cisco IOS software.

  When connecting a Cisco IOS software device to a CatOS device, make sure that PAgP settings are the same on both sides. The defaults are different. CatOS devices should have PAgP set to off when connecting to a Cisco IOS software device if EtherChannels are not configured.

General Design Considerations

The following are general design considerations:

- Use HSRP or GLBP for default gateway redundancy (sub-second timers).

  Default gateway redundancy is an important component in convergence in a hierarchical network design. You can reliably tune HSRP/GLBP timers to achieve 900 ms convergence for link/node failure in the L2/L3 boundary in the distribution hierarchical model.

- Deploy QoS end-to-end; protect the good and punish the bad.

  QoS is not just for voice and video anymore. Internet worms and denial of service (DoS) attacks have the ability to flood links even in a high-speed campus environment. You can use QoS policies to protect mission-critical applications while giving a lower class of service to suspect traffic.

- Avoid daisy chaining stackable switches; stacks are good, StackWise and chassis solutions are better.

  Daisy-chained fixed configuration implementations add complexity. Without careful consideration, discontinuous VLAN/subnets, routing black holes, and active/active HSRP/GLPB situations can exist. Use StackWise technology in the Cisco Catalyst 3750 family or modular chassis implementations to avoid these complications.

- Avoid asymmetric routing and unicast flooding; do not span VLANs across the access layer.

  When a less-than-optimal topology is used, a long-existing but frequently misunderstood situation can occur as a result of the difference between ARP and CAM table aging timers. If VLANs span across multiple access layer switches, return path traffic can be flooded to all access layer switches and end points. This can be easily avoided by not spanning VLANs across access layer switches. If this cannot be avoided, then tune the ARP aging timer so that it is less than the CAM aging timer.

- Keep redundancy simple.

  Protecting against double failures by using three redundant links or three redundant nodes in the hierarchical design does not increase availability. Instead, it decreases availability by reducing serviceability and determinism.

- Only span VLANs across multiple access layer switches if you must.

  Throughout this document we have discussed the challenges with an environment in which VLANs span access layer switches. This design is less than optimal from a convergence perspective. If you follow the rules, you can achieve deterministic convergence. However, there are many opportunities to increase your availability and optimize convergence with alternative designs.

- L2/L3 distribution with HSRP or GLBP is a tried-and-true design.

  A network design that follows the tried-and-true topology in which the L2/L3 boundary is in the distribution layer is the most deterministic and can deliver sub-second (900 ms) convergence. When properly configured and tuned, this design is the recommended best practice.

- L3 in the access is an emerging and intriguing option.

  Advances in routing protocols and campus hardware have made it viable to deploy a routing protocol in the access layer switches and use an L3 point-to-point routed link between the access and distribution layer switches. This design can provide improvement in several areas, most notably reliable convergence in the 60–200 ms range.

Hierarchical Network Design Model

This section includes the following topics:

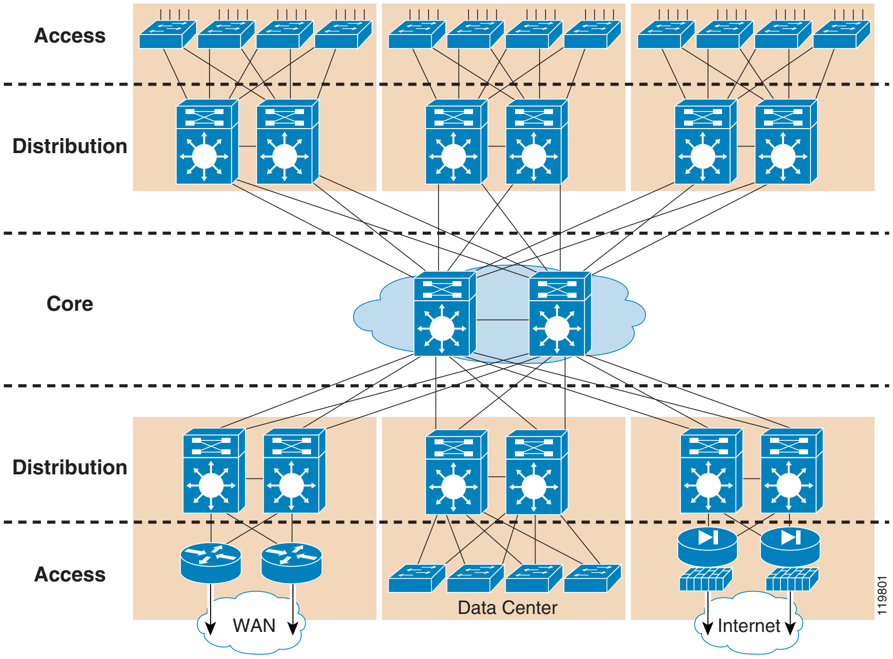
- Hierarchical Network Design Overview, page 7
- Core Layer, page 8
- Distribution Layer, page 9
- Access Layer, page 10

Hierarchical Network Design Overview

You can use the hierarchical model to design a modular topology using scalable "building blocks" that allow the network to meet evolving business needs. The modular design makes the network easy to scale, understand, and troubleshoot by promoting deterministic traffic patterns.

Cisco introduced the hierarchical design model, which uses a layered approach to network design in 1999 (see Figure 1). The building block components are the access layer, the distribution layer, and the core (backbone) layer. The principal advantages of this model are its hierarchical structure and its modularity.

Figure 1 Hierarchical Campus Network Design



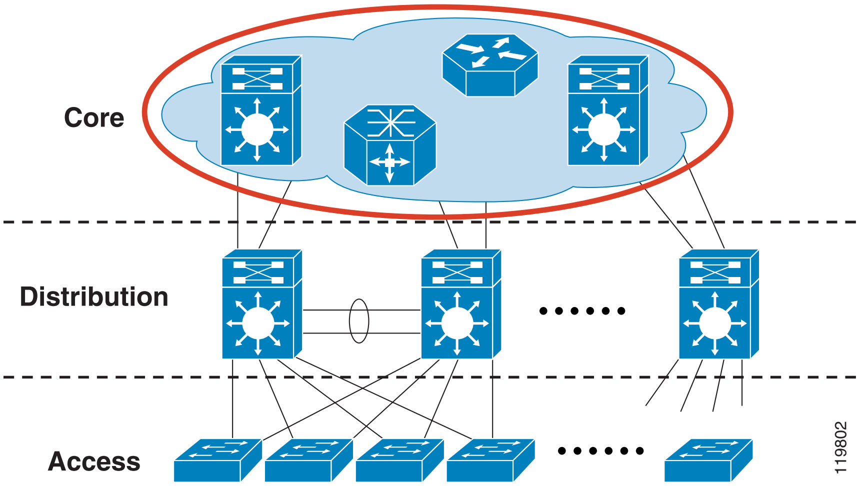
In a hierarchical design, the capacity, features, and functionality of a specific device are optimized for its position in the network and the role that it plays. This promotes scalability and stability. The number of flows and their associated bandwidth requirements increase as they traverse points of aggregation and move up the hierarchy from access to distribution to core. Functions are distributed at each layer. A hierarchical design avoids the need for a fully-meshed network in which all network nodes are interconnected.

The building blocks of modular networks are easy to replicate, redesign, and expand. There should be no need to redesign the whole network each time a module is added or removed. Distinct building blocks can be put in-service and taken out-of-service without impacting the rest of the network. This capability facilitates troubleshooting, problem isolation, and network management.

Core Layer

In a typical hierarchical model, the individual building blocks are interconnected using a core layer. The core serves as the backbone for the network, as shown in Figure 2. The core needs to be fast and extremely resilient because every building block depends on it for connectivity. Current hardware accelerated systems have the potential to deliver complex services at wire speed. However, in the core of the network a "less is more" approach should be taken. A minimal configuration in the core reduces configuration complexity limiting the possibility for operational error.

Figure 2 Core Layer



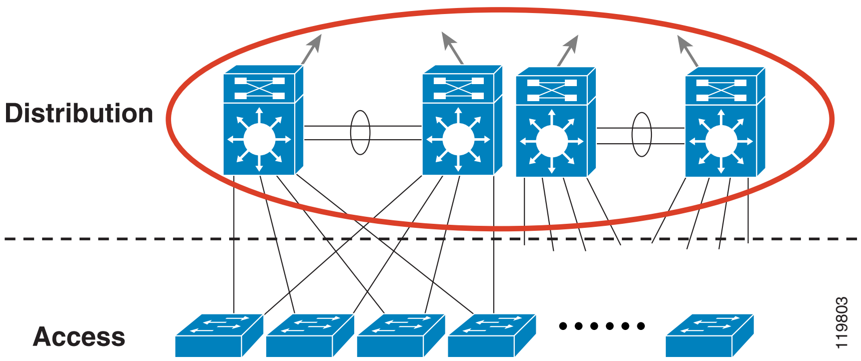
Although it is possible to achieve redundancy with a fully-meshed or highly-meshed topology, that type of design does not provide consistent convergence if a link or node fails. Also, peering and adjacency issues exist with a fully-meshed design, making routing complex to configure and difficult to scale. In addition, the high port count adds unnecessary cost and increases complexity as the network grows or changes. The following are some of the other key design issues to keep in mind:

- Design the core layer as a high-speed, Layer 3 (L3) switching environment utilizing only hardware-accelerated services. Layer 3 core designs are superior to Layer 2 and other alternatives because they provide:
  - Faster convergence around a link or node failure.
  - Increased scalability because neighbor relationships and meshing are reduced.
  - More efficient bandwidth utilization.
- Use redundant point-to-point L3 interconnections in the core (triangles, not squares) wherever possible, because this design yields the fastest and most deterministic convergence results.
- Avoid L2 loops and the complexity of L2 redundancy, such as Spanning Tree Protocol (STP) and indirect failure detection for L3 building block peers.

Distribution Layer

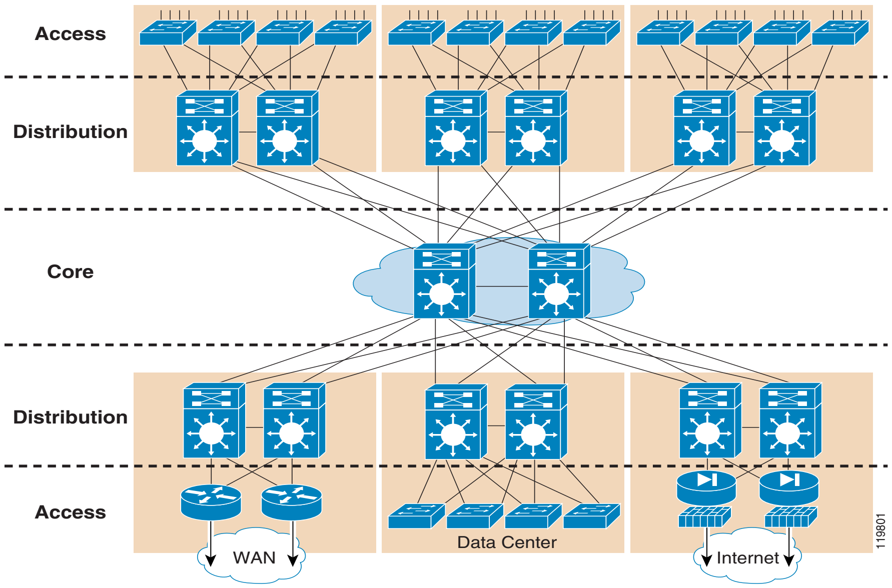
The distribution layer aggregates nodes from the access layer, protecting the core from high-density peering (see Figure 3). Additionally, the distribution layer creates a fault boundary providing a logical isolation point in the event of a failure originating in the access layer. Typically deployed as a pair of L3 switches, the distribution layer uses L3 switching for its connectivity to the core of the network and L2 services for its connectivity to the access layer. Load balancing, Quality of Service (QoS), and ease of provisioning are key considerations for the distribution layer.

Figure 3 Distribution Layer



High availability in the distribution layer is provided through dual equal-cost paths from the distribution layer to the core and from the access layer to the distribution layer (see Figure 4). This results in fast, deterministic convergence in the event of a link or node failure. When redundant paths are present, failover depends primarily on hardware link failure detection instead of timer-based software failure detection. Convergence based on these functions, which are implemented in hardware, is the most deterministic.

Figure 4 Distribution Layer—High Availability



L3 equal-cost load sharing allows both uplinks from the core to the distribution layer to be utilized. The distribution layer provides default gateway redundancy using the Gateway Load Balancing Protocol (GLBP), Hot Standby Router Protocol (HSRP), or Virtual Router Redundancy Protocol (VRRP). This allows for the failure or removal of one of the distribution nodes without affecting end point connectivity to the default gateway.

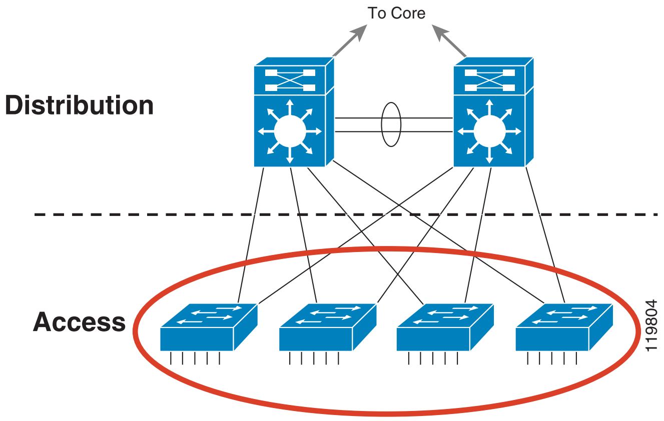
You can achieve load balancing on the uplinks from the access layer to the distribution layer in many ways, but the easiest way is to use GLBP. GLBP provides HSRP-like redundancy and failure protection. It also allows for round robin distribution of default gateways to access layer devices, so the end points can send traffic to one of the two distribution nodes.

See "Using HSRP, VRRP, or GLBP for Default Gateway Redundancy" section on page 35 for more details on default gateway redundancy.

Access Layer

The access layer is the first point of entry into the network for edge devices, end stations, and IP phones (see Figure 5). The switches in the access layer are connected to two separate distribution layer switches for redundancy. If the connection between the distribution layer switches is an L3 connection, then there are no loops and all uplinks actively forward traffic.

Figure 5 Access Layer



A robust access layer provides the following key features:

- High availability (HA) supported by many hardware and software attributes.
- Inline power (POE) for IP telephony and wireless access points, allowing customers to converge voice onto their data network and providing roaming WLAN access for users.
- Foundation services.

The hardware and software attributes of the access layer that support high availability include the following:

- System-level redundancy using redundant supervisor engines and redundant power supplies. This provides high-availability for critical user groups.
- Default gateway redundancy using dual connections to redundant systems (distribution layer switches) that use GLBP, HSRP, or VRRP. This provides fast failover from one switch to the backup switch at the distribution layer.

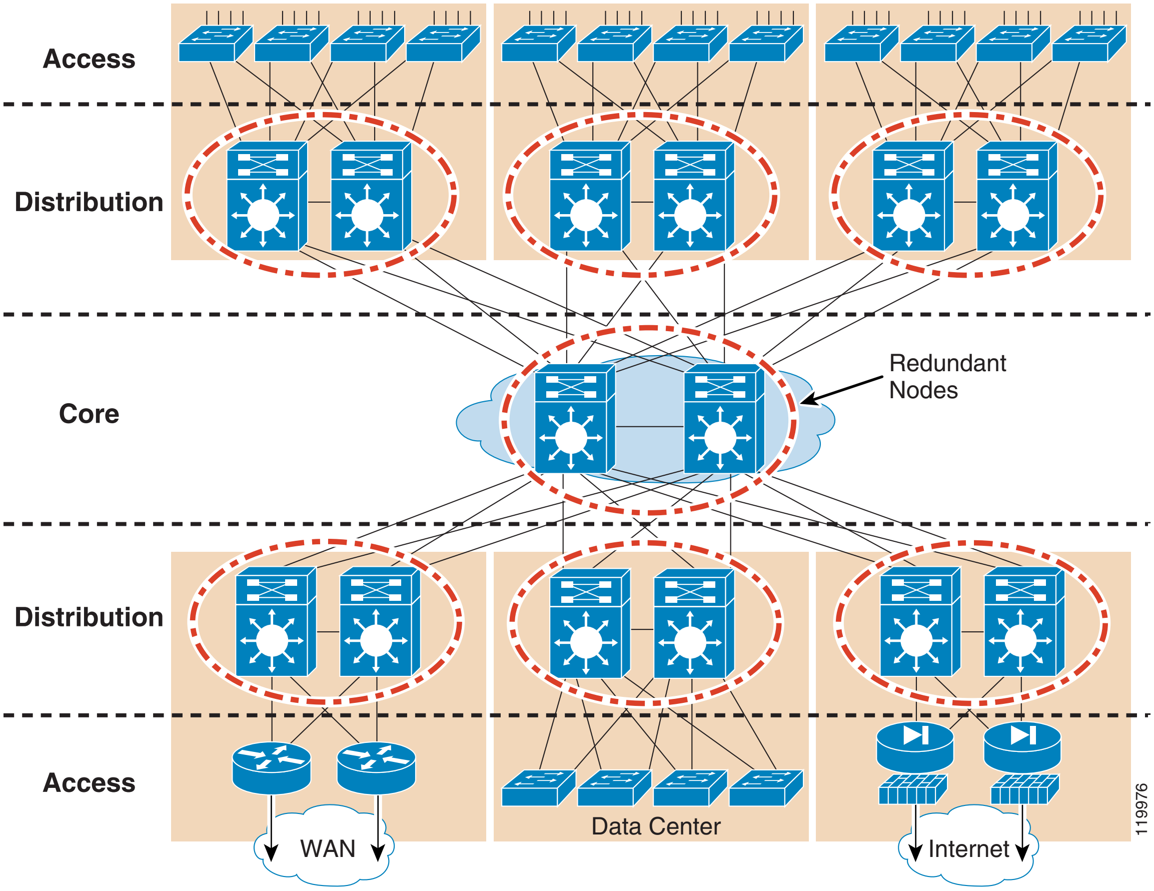
- Operating system high-availability features, such as Link Aggregation (EtherChannel or 802.3ad), which provide higher effective bandwidth while reducing complexity.
- Prioritization of mission-critical network traffic using QoS. This provides traffic classification and queuing as close to the ingress of the network as possible.
- Security services for additional security against unauthorized access to the network through the use of tools such as 802.1x, port security, DHCP snooping, Dynamic ARP Inspection, and IP Source Guard.
- Efficient network and bandwidth management using software features such as Internet Group Membership Protocol (IGMP) snooping. IGMP snooping helps control multicast packet flooding for multicast applications.

Network and In-the-Box Redundancy

When designing a campus network, the network engineer needs to plan the optimal use of the highly redundant devices. Careful consideration should be given as to when and where to make an investment in redundancy to create a resilient and highly available network.

As shown in Figure 6, the hierarchical network model consists of two actively forwarding core nodes, with sufficient bandwidth and capacity to service the entire network in the event of a failure of one of the nodes. This model also requires a redundant distribution pair supporting each distribution building block. Similarly to the core, the distribution layer is engineered with sufficient bandwidth and capacity so that the complete failure of one of the distribution nodes does not impact the performance of the network from a bandwidth or switching capacity perspective.

Figure 6 Redundant Network Nodes



Campus network devices can currently provide a high level of availability within the individual nodes. The Cisco Catalyst 6500 and 4500 switches can support redundant supervisor engines and provide L2 Stateful Switchover (SSO), which ensures that the standby supervisor engine is synchronized from an L2 perspective and can quickly assume L2 forwarding responsibilities in the event of a supervisor failure.

The Catalyst 6500 also provides L3 Non-Stop Forwarding (NSF), which allows the redundant supervisor to assume L3 forwarding responsibilities without resetting or re-establishing neighbor relationships with the surrounding L3 peers in the event of the failure of the primary supervisor.

When designing a network for optimum high availability, it is tempting to add redundant supervisors to the redundant topology in an attempt to achieve even higher availability. However, adding redundant supervisors to redundant core and distribution layers of the network can increase the convergence time in the event of a supervisor failure.

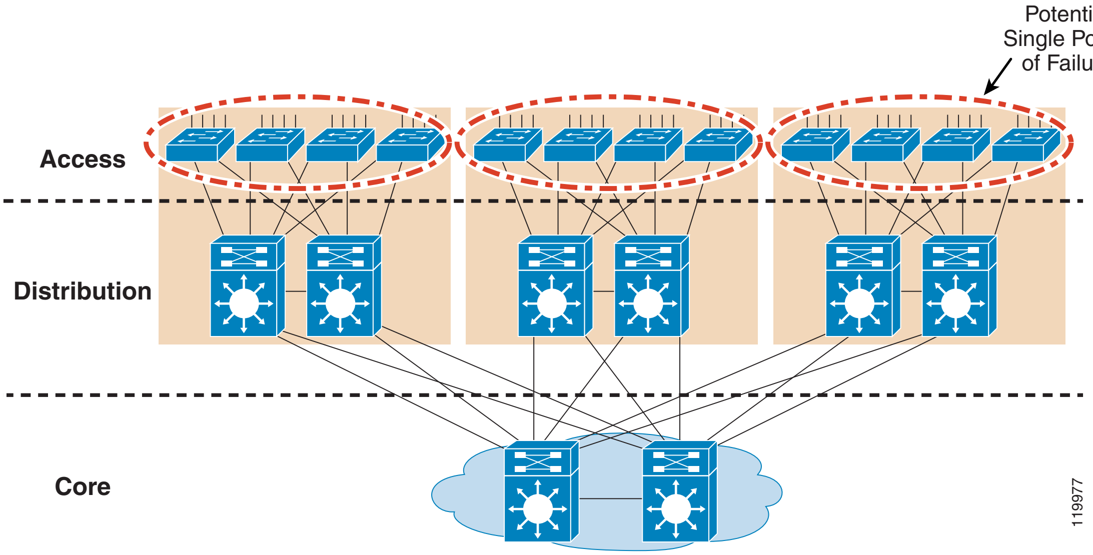
In the hierarchical model, the core and distribution nodes are connected by point-to-point L3 routed fiber optic links. This means that the primary method of convergence for core or distribution node failure is loss of link. If a supervisor fails on a non-redundant node, the links fail and the network converges around the outage through the second core or distribution node. This allows the network to converge in 60–200 milliseconds for EIGRP and OSPF.

Note For more details, refer to *High Availability Campus Recovery Analysis*.

When redundant supervisors are introduced, the links are not dropped during an SSO or NSF convergence event if a supervisor fails. Traffic is lost while SSO completes, or indirect detection of the failure occurs. SSO recovers in 1–3 seconds, depending on the physical configuration of device in question. L3 recovery using NSF happens after the SSO convergence event, minimizing L3 disruption and convergence. For the same events, where 60–200 milliseconds of packet loss occurred without redundant supervisors when dual supervisor nodes were used in the core or distribution, 1.8 seconds of loss was measured.

The access layer of the network is typically a single point of failure, as shown in Figure 7.

Figure 7 Potential Single Points of Failure



While the access nodes are dual connected to the distribution layer, it is not typical for endpoints on the network to be dual connected to redundant access layer switches (except in the data center). For this reason, SSO provides increased availability when redundant supervisors are used in the access layer and the L2/L3 boundary is in the distribution layer of the network. In this topology, SSO provides for protection against supervisor hardware or software failure with 1–3 seconds of packet loss and no network convergence. Without SSO and a single supervisor, devices serviced by this access switch would experience a total network outage until the supervisor was physically replaced or, in the case of a software failure, until the unit reloaded.

If the L2/L3 boundary is in the access layer of the network, a design in which a routing protocol is running in the access layer, then NSF with SSO provides an increased level of availability. Similarly to the L2/L3 distribution layer topology, NSF with SSO provides 1–3 seconds of packet loss without network convergence compared to total outage until a failed supervisor is physically replaced for the routed access topology.

Campus topologies with redundant network paths can converge faster than topologies that depend on redundant supervisors for convergence. NSF/SSO provide the most benefit in environments where single points of failure exist. In the campus topology, that is the access layer. If you have an L2 access layer design, redundant supervisors with SSO provide the most benefit. If you have a routed access layer design, redundant supervisors with NSF with SSO provide the most benefit.

Foundation Services Technologies

This section describes the foundation technologies used in the campus network and the recommended configurations. It includes the following topics:

- Layer 3 Routing Protocols, page 14
- Layer 2 Redundancy—Spanning Tree Protocol Versions, page 22
- Trunking Protocols, page 25
- Protecting Against One-Way Communication with UniDirectional Link Detection, page 31
- Link Aggregation—EtherChannel Protocol and 802.3ad, page 32
- Link Aggregation Protocol, page 33
- Using HSRP, VRRP, or GLBP for Default Gateway Redundancy, page 35
- Gateway Load Balancing Protocol, page 37
- Oversubscription and QoS, page 40

Layer 3 Routing Protocols

This section includes the following topics:

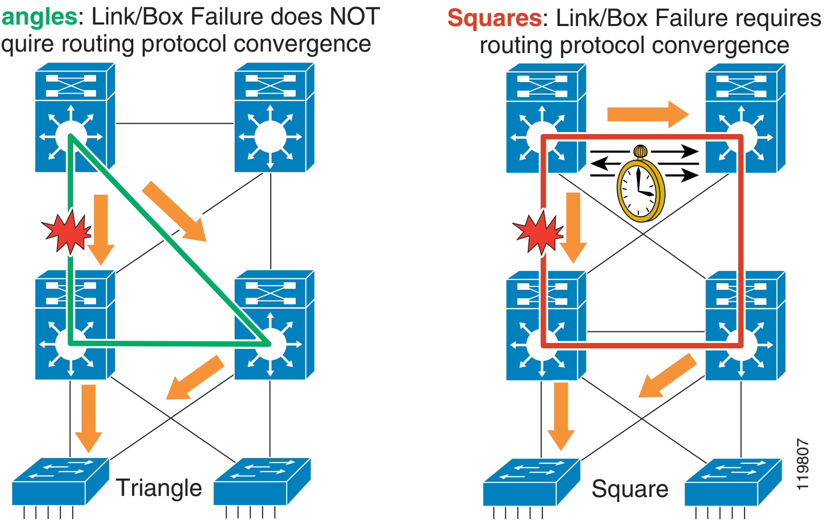
- Using Triangle Topologies, page 14
- Limiting L3 Peering to Transit Links, page 15
- Ensuring Connectivity in Case of Failure, page 16
- Tuning Load Balancing with Cisco Express Forwarding, page 19

Using Triangle Topologies

Layer 3 routing protocols are typically deployed in the core-to-core and core-to-distribution layers of the network, and can be used all the way to the access layer. However, fully-routed access layer designs are not often deployed today. See the "Routing in the Access Layer" section on page 53 for a in-depth discussion of routed access layer designs. Routing protocols are utilized in a hierarchical network design to reroute around a failed link or node.

If you build a topology using triangles, with equal-cost paths to all redundant nodes, you can avoid timer-based, non-deterministic convergence. Instead of indirect neighbor or route loss detection using hellos and dead timers, you can rely on physical link loss to mark a path as unusable and reroute all traffic to the alternate equal-cost path. Figure 8 shows both triangle and square network topologies.

Figure 8 Triangle and Square Network Topologies

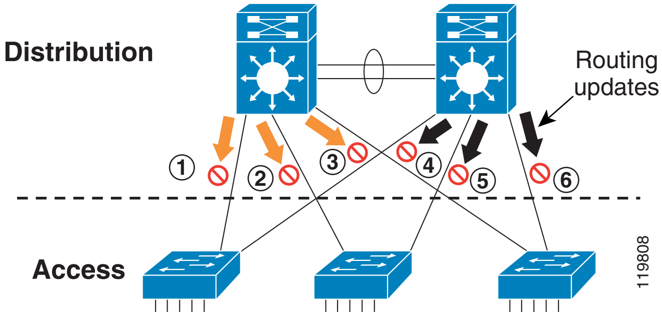

Triangles: Link/Box Failure does NOT require routing protocol convergence
Squares: Link/Box Failure requires routing protocol convergence
Triangle
Square
119807

The use of triangle rather than square topologies is only a recommendation. It is possible to build a topology that does not rely on equal-cost redundant paths to compensate for limited physical fiber connectivity or to reduce cost. However, it is not possible to achieve the same deterministic convergence in the event of a link or node failure, and for this reason the design will not be optimized for high availability.

Limiting L3 Peering to Transit Links

In the hierarchical model, the distribution routers, based on the default configuration, can establish a peer relationship through the access layer for each VLAN supported by the distribution pair (see Figure 9). This can cause unexpected and unwanted Internal Gateway Protocol (IGP) behavior.

Figure 9 Layer 3 Peer Relationships



This redundant L3 peering has no benefit from an HA perspective, and only adds load in terms of memory, routing protocol update overhead, and complexity. Additionally, in the event of a link failure, it is possible for traffic to transit through a neighboring access layer switch, which is not desirable. It is therefore recommended that only links intended for transit traffic be used to establish routing neighbor or peer relationships. To achieve this goal, you can make individual interfaces passive or make all the interfaces passive.

To make the individual interfaces passive, where a peering relationship is not desired, enter the following commands:

```
router eigrp 1
passive-interface Vlan 99
```

Alternatively, you can make all interfaces passive, and then use the **no passive** command to enable a routing neighbor relationship on the interfaces where peering is desired. This is shown in the following example:

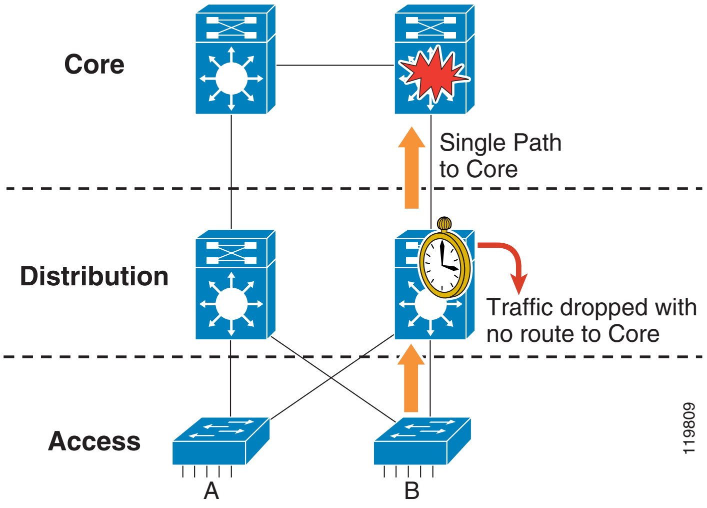
```
router eigrp 1
passive-interface default
no passive-interface Vlan 99
```

Use either technique to minimize the number of peer relationships between distribution nodes, allowing them to peer only over links intended as transit links. Use whichever technique requires the fewest lines of configuration or is the easiest for you to manage.

Ensuring Connectivity in Case of Failure

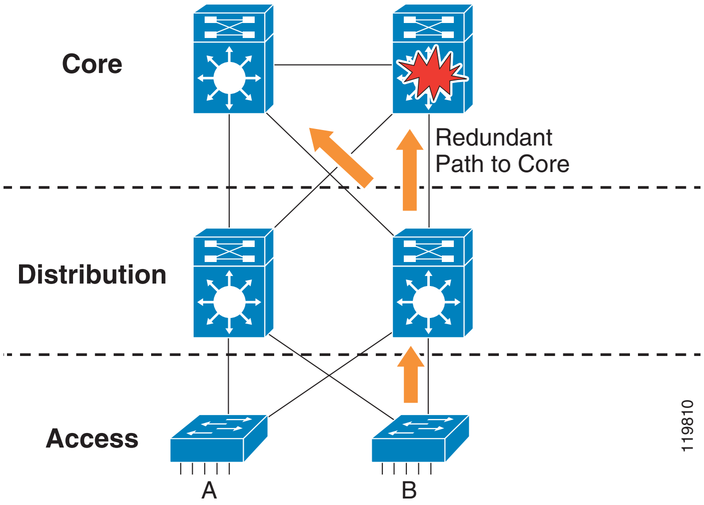
From a connectivity perspective, some network designers recommend dual distribution nodes that are individually connected to a single core node member. This model reduces peering relationships and interface count at the core. However, traffic can be dropped if a core link or node fails, as shown in Figure 10.

Figure 10 Single Path to the Core



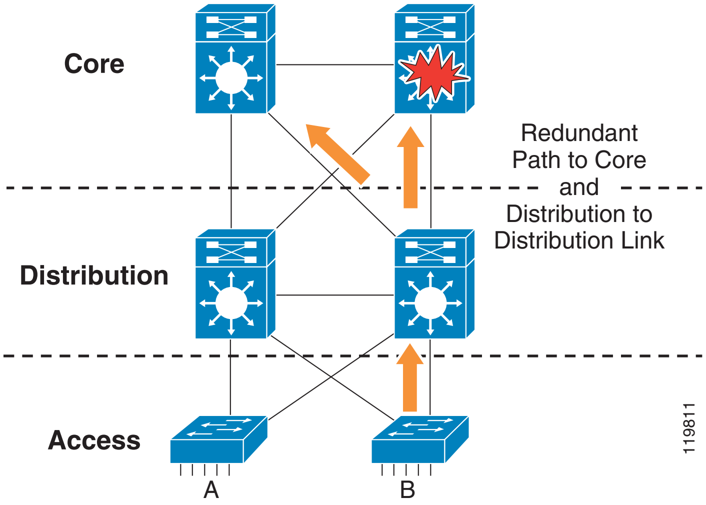
The recommended design is to provide an alternate path to the core, as shown in Figure 11.

Figure 11 Alternate Path to the Core



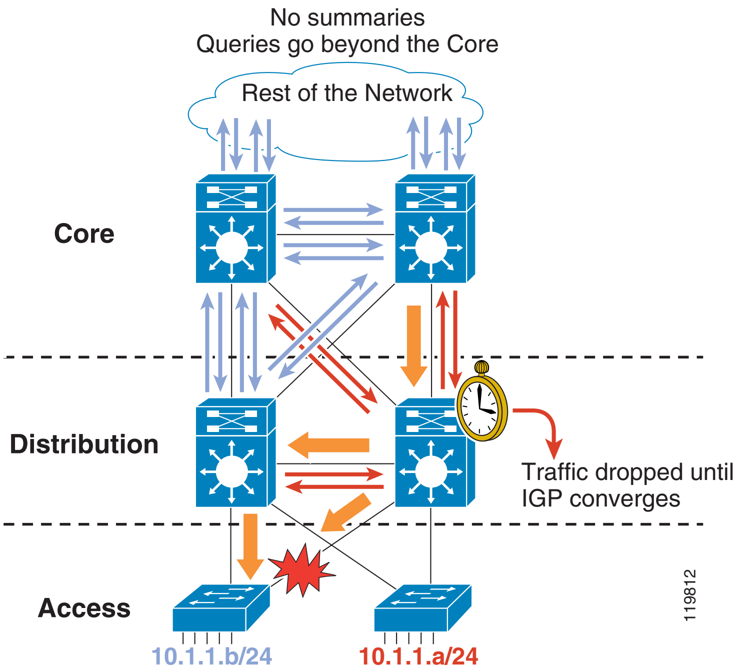
The additional link between Distribution A and Core B is not the only additional link that is required. A link between the two distribution nodes is also required. This requirement is discussed in detail in the next section. The recommended topology is shown in Figure 12.

Figure 12 Recommended Topology (Links Between Two Distribution Nodes)



The additional link between the distribution switches is required to support summarization of routing information from the distribution layer towards the core. If the routing information is not summarized towards the core, Enhanced Interior Gateway Protocol (EIGRP) and Open Shortest Path First (OSPF) require interaction with a potentially large number of peers to converge around a failed node, as shown in Figure 13.

Figure 13 Convergence Around a Failed Node



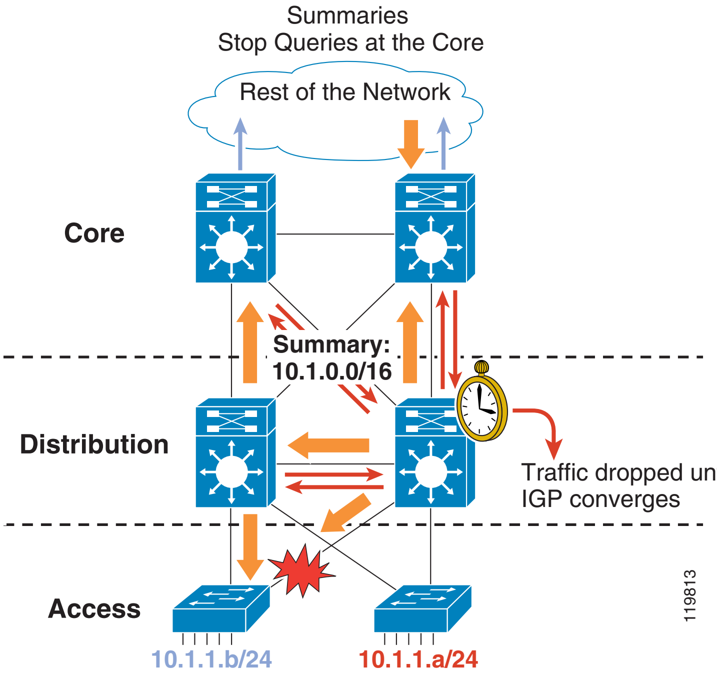
In the configuration example below, summary routes are sent towards the core:

```
interface Port-channel1
description to Core#1
ip address 10.122.0.34 255.255.255.252
ip hello-interval eigrp 100 1
ip hold-time eigrp 100 3
ip summary-address eigrp 100 10.1.0.0 255.255.0.0 5
```

When summarization is used, the distribution nodes interact with a bounded number of routing peers when converging around a link or node failure. Summarizing using EIGRP or using an area boundary for OSPF are the recommended L3 configurations for the distribution-to-core layer L3 connection.

An L3 link is required between the distribution nodes. If an L3 link between the distribution nodes is not present, return traffic (from the core to the access layer) could be dropped if an access layer link fails and the distribution nodes are not interconnected with an L3 link, as shown in Figure 14.

Figure 14 Summaries Stop Queries at the Core



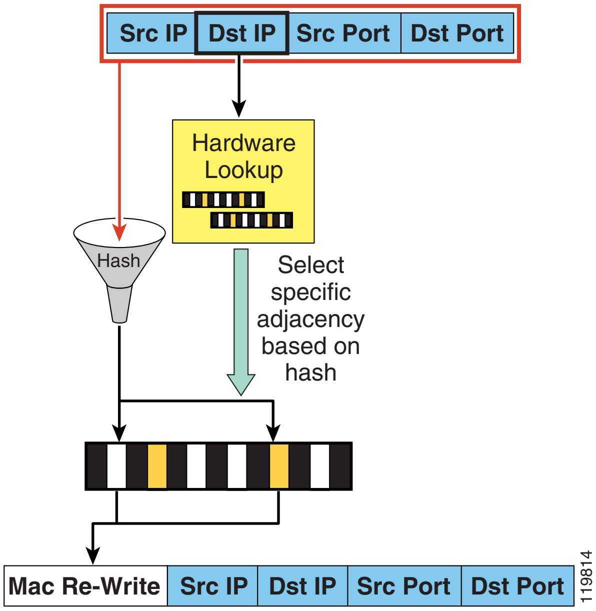
Because the distribution nodes send summarized information towards the core, an individual distribution node does not advertise loss of connectivity to a single VLAN or subnet. This means that the core does not know that it cannot send traffic to the distribution member where the link has failed. Adding an L3 link between the distribution switches allows the distribution node that loses connectivity to a given VLAN or subnet to reroute traffic across the distribution-to-distribution link. The address space selected for the distribution-to-distribution link must be within the address space being summarized to be effective.

Tuning Load Balancing with Cisco Express Forwarding

Many redundant paths are provided in the recommended network topology. From the perspective of the access layer, at least three sets of redundant links are traversed to another building block, such as the data center. Tuning of Cisco Express Forwarding (CEF) equal-cost path selection is required to prevent CEF polarization, in which redundant links may be underutilized.

CEF is a deterministic algorithm. As shown in Figure 15, when using the same information for input, the same result is always obtained.

Figure 15 CEF Load Balancing



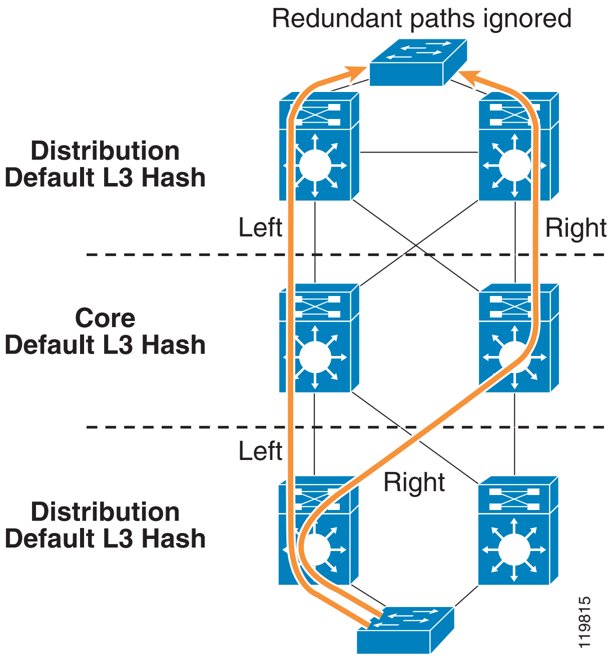
CEF uses a multistep process to make its final forwarding decision:

1. CEF determines the longest path match for the destination address using a hardware lookup.
2. Each specific index is associated with a next-hop adjacencies table.
   - By default, one of the possible adjacencies is selected by a hardware hash where the packet source and destination IP address are used.
   - As a configurable alternative, one of the possible adjacencies can also be selected by a hardware hash using L4 port information in addition to the packet source and destination IP address.
3. The new MAC address is attached and the packet is forwarded.

If you change the input to the hash, you will change the output. The default input value is L3 for source and destination. If you change this input value to L3 with L4, the output hash value also changes.

When packets traverse a network with multiple redundant paths that all use the same input value, a “go to the right” or “go to the left” decision is made for each redundant path. As a result, some redundant links are underutilized and the network is said to be experiencing CEF polarization (see Figure 16).

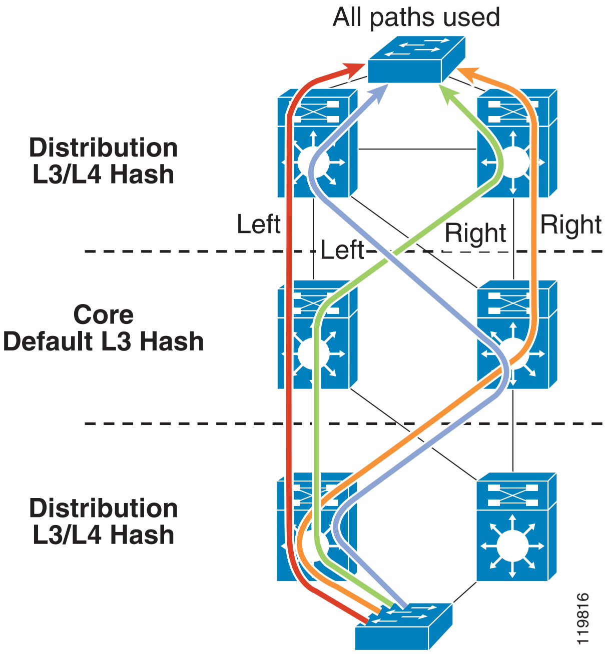
Figure 16 CEF Polarization



To avoid CEF polarization, you need to vary the input into the CEF algorithm across the layers in the network. In the distribution layer, change the default CEF load balancing behavior and use L3 and L4 information as input into the CEF hashing algorithm. To achieve this, use the **mls ip cef load-sharing full** command on the distribution nodes.

In the core layer, leave the default, which is to use only L3 information. This alternating approach eliminates the always right or always left biased decisions and helps balance the traffic over equal-cost redundant links in the network (see Figure 17).

Figure 17 Preventing CEF Polarization



Layer 2 Redundancy—Spanning Tree Protocol Versions

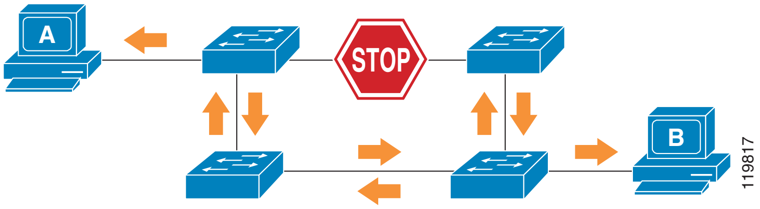
This section includes the following topics:

- Spanning Tree Protocol Versions, page 22
- Best Practices for Optimal Convergence, page 23

Spanning Tree Protocol Versions

Highly available networks require redundant paths to ensure connectivity in the event of a node or link failure. Various versions of Spanning Tree Protocol (STP) are used in environments that include redundant L2 loops. STP lets the network deterministically block interfaces and provide a loop-free topology in a network with redundant links (see Figure 18).

Figure 18 STP Operation



The following versions of STP have evolved over time:

- DEC STP pre-IEEE
- 802.1D—Classic STP
- 802.1w—Rapid STP (RSTP)
- 802.1s—Multiple STP (MST)
- 802.1t—802.1d maintenance

The following enhancements to 802.1(d,s,w) comprise the Cisco Spanning-Tree toolkit:

- PortFast—Lets the access port bypass the listening and learning phases
- UplinkFast—Provides 3-to-5 second convergence after link failure
- BackboneFast—Cuts convergence time by MaxAge for indirect failure
- Loop Guard—Prevents the alternate or root port from being elected unless Bridge Protocol Data Units (BPDUs) are present
- Root Guard—Prevents external switches from becoming the root
- BPDU Guard—Disables a PortFast-enabled port if a BPDU is received
- BPDU Filter—Prevents sending or receiving BPDUs on PortFast-enabled ports

Cisco has incorporated a number of these features into the following versions of STP:

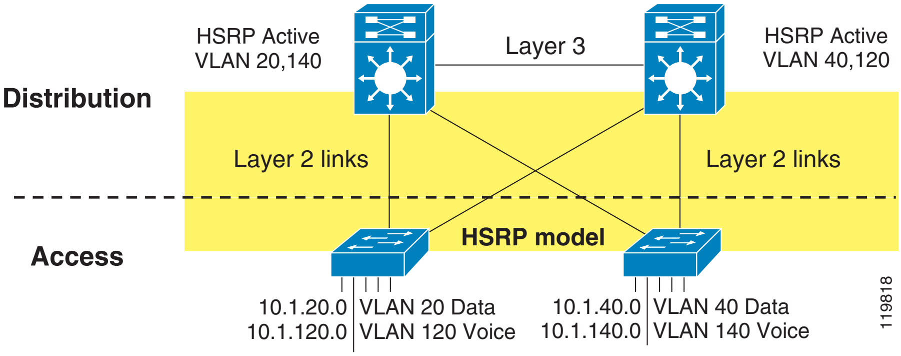
- Per-VLAN Spanning Tree Plus (PVST+)—Provides a separate 802.1D spanning tree instance for each VLAN configured in the network. This includes PortFast, UplinkFast, BackboneFast, BPDU Guard, BPDU Filter, Root Guard, and Loop Guard.
- Rapid PVST+—Provides an instance of RSTP (802.1w) per VLAN. This includes PortFast, BPDU Guard, BPDU Filter, Root Guard, and Loop Guard.
- MST—Provides up to 16 instances of RSTP (802.1w) and combines many VLANS with the same physical and logical topology into a common RSTP instance. This includes, PortFast, BPDU Guard, BPDU Filter, Root Guard, and Loop Guard.

Best Practices for Optimal Convergence

Only use L2 looped topologies if it cannot be avoided. In general practice, the most deterministic and best-performing networks in terms of convergence, reliability, and manageability are free from L2 loops and do not require STP to resolve convergence events under normal conditions. However, STP should be enabled to protect against unexpected loops on the access or user-facing interfaces.

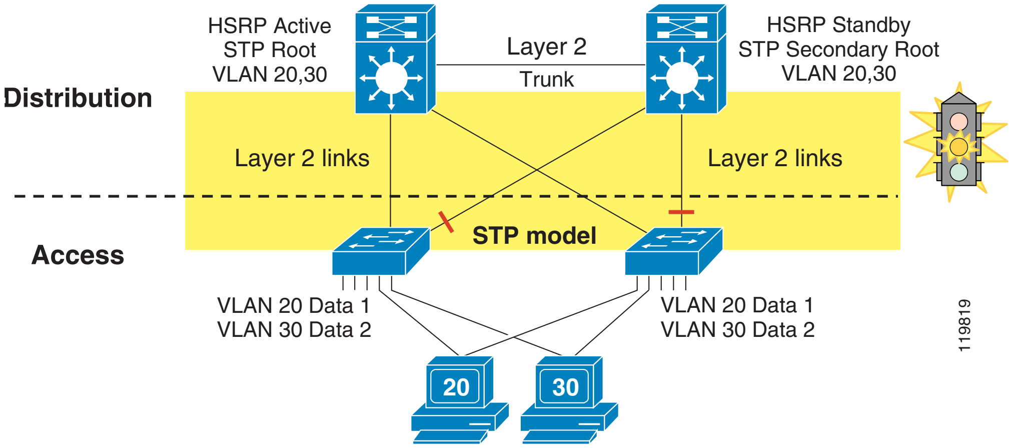
In the reference hierarchical design, L2 links are deployed between the access and distribution nodes. However, no VLAN exists across multiple access layer switches. Additionally, the distribution-to-distribution link is an L3 routed link. This results in an L2 loop-free topology in which both uplinks from the access layer are forwarding from an L2 perspective and are available for immediate use in the event of a link or node failure (see Figure 19).

Figure 19 Layer 2 Loop-Free Topology



In the data center, servers are commonly dual-attached and L2 connectivity is required, from the host perspective, to support dual attachment. In this case, L2 loops are common (see Figure 20).

Figure 20 Layer 2 Looped Topology in the Data Center



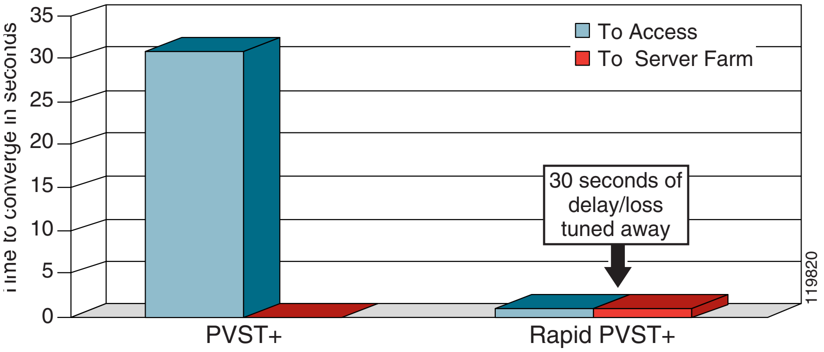
This L2 looped topology is configuration and management intensive. You must make sure that the STP root and default gateway (HSRP or VRRP) match.

Note Without additional STP configuration, GLBP load balancing behavior can cause traffic to take a two hop L2 path across the distribution-to-distribution link to its default gateway. See "Gateway Load Balancing Protocol" section on page 37 for more details on this subject.

STP/RSTP convergence is required for several convergence events. Depending on the version of STP, convergence could take as long as 90 seconds.

If you use a topology where spanning-tree convergence is required, then Rapid PVST+ is the best version. Rapid PVST+ provides the rapid convergence of 802.1w while avoiding the complexity of 802.1s. From a configuration perspective, it resembles PVST+, which Cisco customers have deployed for years. However, from a convergence perspective, it is much improved, as shown in Figure 21.

Figure 21 PVST+ and Rapid PVST+ Performance



Rapid PVST+ greatly improves the detection of indirect failures (L2 distribution-to-distribution link) or link up (uplink) restoration events.

STP is also required to protect against inadvertent loops introduced on the user side or end point-facing access layer ports. These can easily happen by accident because of misconfigured hosts. For example, by default, the Windows XP Home Networking Wizard bridges together all the interfaces on the machine.

This can result in a bridge between a wireless LAN interface and an Ethernet interface, or between two Ethernet interfaces. Loops can be introduced even if L3 is the only protocol running on uplinks in the network. For this reason you must enable STP or RSTP to ensure a loop-free topology even if it is used only as a failsafe.

You should enable the following additional STP features to protect against soft failures and rogue devices:

- Root Guard
- BPDU Guard
- Loop Guard

Enable either Root Guard or BPDU Guard on access layer ports. Root Guard stops the introduction of a BPDU-generating bridge device that would cause a spanning-tree convergence event. It prevents a port from transmitting BPDUs that would cause a change in the root port or path selection. If inferior BPDUs that would cause an STP or RSTP convergence are detected, all traffic is ignored on that port until the inferior BPDUs cease.

Use BPDU Guard to prevent the introduction of non-authorized bridging devices. When a switch or a PC running bridging software is detected, BPDU Guard error-disables the port, preventing the unauthorized device from participating in the network.

BPDU Guard requires operator intervention or the setting of error recovery mechanisms to re-enable the error-disabled port. You can use BPDU Guard to stop all bridge devices, such as switches, from being added to your network. Alternatively, you can use Root Guard to protect against an unexpected spanning-tree convergence event caused by the addition of an un-authorized bridge device. Only use BPDU Guard if you are able to intervene and re-enable error-disabled ports.

Use Loop Guard to protect the network from a soft failure where physical connectivity and packet forwarding are intact but STP (BPDU generation, forwarding, and evaluation) fails.

Trunking Protocols

This section includes the following topics:

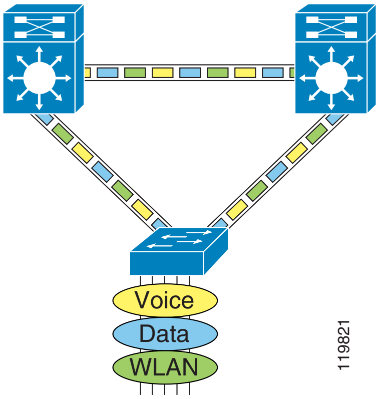
- Deploying Multiple VLANS on a Single Ethernet Link (Trunking), page 26
- Virtual Trunk Protocol, page 27
- Dynamic Trunk Protocol, page 28
- Preventing Double 802.1Q Encapsulated VLAN Hopping, page 29

Deploying Multiple VLANS on a Single Ethernet Link (Trunking)

VLANs provide the broadcast isolation, policy implementation, and fault isolation benefits that are required in highly available networks.

Trunking protocols allow network node interconnections (uplinks) to carry multiple VLANS through a single physical link, as shown in Figure 22.

Figure 22 Multiple VLANs on a Single Interconnection



Two types of trunks are currently available:

- 802.1Q
- Inter-Switch Link (ISL).

802.1Q is the Institute of Electrical and Electronics Engineers (IEEE) standard implementation. Cisco developed ISL trunking before the standard was established. In general, there is no technical reason to use one or the other. ISL does consume a small amount of additional bandwidth because of the double CRC check that it performs. The current best practice is to use 802.1Q trunks for the sake of simplicity, compatibility, and efficiency. Implement Cisco extensions to 802.1Q to avoid security concerns related to the 802.1Q non-tagged native VLAN.

The following are best practices to use when deploying multiple VLANs on a single switch-to-switch interconnection or trunk:

- Deploy VLANs on the interconnection between access and distribution layers.
- Use VLAN Trunking Protocol (VTP) in transparent mode to reduce the potential for operational error.
- Hard set the trunk mode to **on** and the encapsulation negotiate to **off** for optimal convergence.
- Assign the native VLAN to an unused ID or use the Tagged Native VLAN option to avoid VLAN hopping.
- Manually prune all VLANS except those needed.
- Disable Trunking/VLAN tagging on host ports with the following commands:

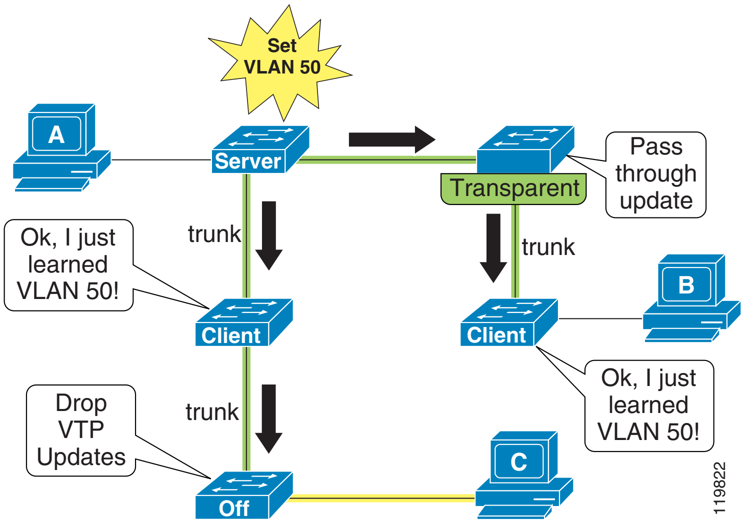
- CatOS—**set port host**
- Cisco IOS software—**switchport host**

Note The **set port host** macro disables EtherChannel, and enables STP PortFast in addition to disabling trunking.

Virtual Trunk Protocol

Virtual Trunk Protocol (VTP) is a protocol that allows network managers to centrally manage the VLAN database. VTP is an essential component of VLAN Trunking. (See Figure 23.)

Figure 23 Virtual Trunk Protocol Operation



VTP runs only on trunks and provides the following four modes:

- Server—Updates clients and servers. The VTP server switch propagates the VTP database to VTP client switches.
- Client—Receives updates but cannot make changes.
- Transparent—Lets updates pass through.
- Off—Ignores VTP updates.

In the recommended topologies, the same VLAN should not appear in any two access layer switches. Adding and removing VLANs is generally not a frequent network management practice. In most cases, VLANs are defined once during switch setup with few, if any, additional modifications to the VLAN database in an access layer switch. The benefits of dynamic propagation of VLAN information across the network are not worth the potential for unexpected behavior due to operational error. For this reason, VTP transparent mode is the recommended configuration option.

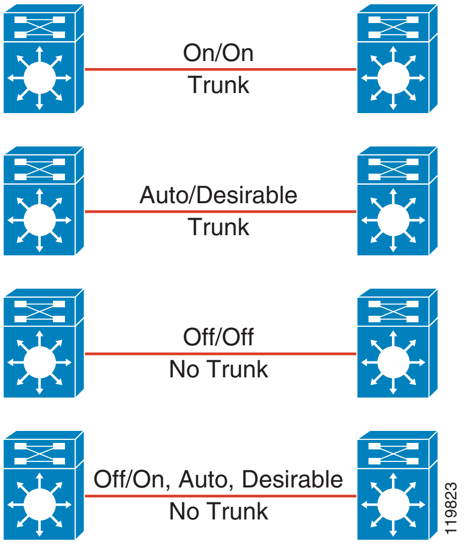
New technologies such as 802.1x and VLAN assignment and Cisco Network Admission Control with quarantined VLAN, must be used with transparent mode. These technologies require a unique VLAN database with common names in each access layer switch.

If you require a common, centrally-managed VLAN database, consider using VTP version 3. VPTv3 contains many enhancements for security and reliability.

Dynamic Trunk Protocol

Dynamic Trunk Protocol (DTP) runs over switch interconnections and allows them to form a trunking interface. (See Figure 24.)

Figure 24 DTP Settings



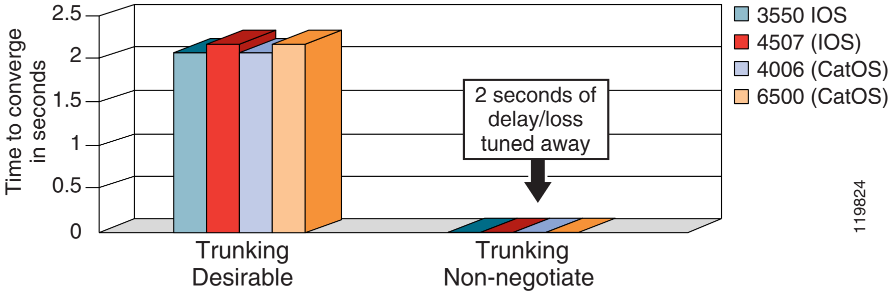
The following are the DTP settings show in Figure 24:

- Automatic formation of interconnection between trunked switch and switch:
  - On—Always form a trunk
  - Desirable—Form a trunk if the other switch will
  - Auto—Form a trunk if the other switch suggests
  - Off—Never form a trunk
- Negotiation of 802.1Q or ISL encapsulation:
  - ISL—Try to use ISL trunk encapsulation
  - 802.1Q—Try to use 802.1Q encapsulation
  - Negotiate—Negotiate ISL or 802.1Q encapsulation with peer
  - No negotiate—Always use hard-set encapsulation

A common practice is to set one side of the interconnection (typically the access) to **auto** and the other end (typically the distribution) to **desirable**. This setting allows for automatic trunk formation, with DTP running on the interconnection to protect against some rare hardware failure scenarios and software misconfigurations. Another alternative is to configure both ends of the trunk to desirable. This has the operational benefit of providing a clear indication of a functional trunking connection with show commands. However, when DTP and 802.1Q or ISL negotiation are enabled, considerable time can be spent negotiating trunk settings when a node or interface is restored. While this negotiation is happening, traffic is dropped because the link is up from an L2 perspective.

As shown in Figure 25, as much as two seconds of packet loss can be eliminated by setting the trunking interface statically to trunk mode and to never dynamically negotiate the trunk type (ISL or 802.1Q).

Figure 25 Trunking Interface Performance



Keep in mind, however, that this setting can cause loss of connectivity if the process is not performed in the correct order *and* there is no out-of-band connectivity to the farthest switch from where the in-band modifications are being made. Therefore, make sure you maintain connectivity when applying this configuration.

The following example shows how to perform this configuration:

With CatOS:

```
set trunk <port> nonegotiate dot1q <vlan>
```

With Cisco IOS software:

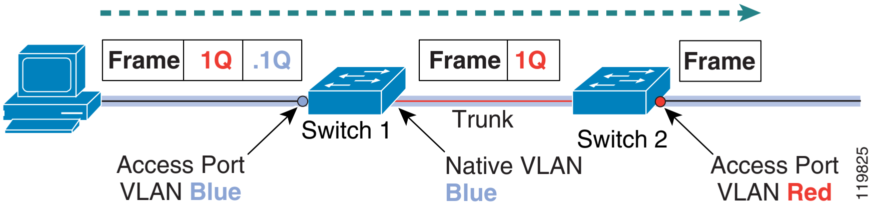
```
switchport mode trunk
switchport nonegotiate
```

This configuration optimizes convergence by setting the trunking interface to always trunk and preventing negotiation of ISL or 802.1Q trunking formats.

Preventing Double 802.1Q Encapsulated VLAN Hopping

There is a remote possibility that an attacker can create a double 802.1Q-encapsulated packet. If the attacker has specific knowledge of the 802.1Q native VLAN, a packet could be crafted that when processed, the first or outermost tag is removed when the packet is switched onto the untagged native VLAN. When the packet reaches the target switch, the inner or second tag is then processed and the potentially malicious packet is switched to the target VLAN (see Figure 26).

Figure 26 Double 802.1Q-Encapsulated Packets

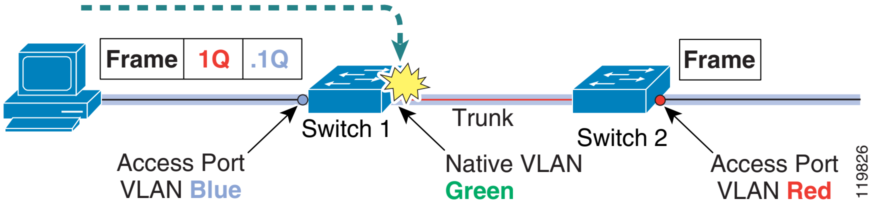


At first glance, this appears to be a serious risk. However, the traffic in this attack scenario is in a single direction and no return traffic can be switched by this mechanism. Additionally, this attack cannot work unless the attacker knows the native VLAN ID.

To reduce whatever risk this attack may pose, set the native VLAN to an obscure ID that is not used for any real traffic in the network. In addition, you should disable 802.1Q trunking on any ports that are connected to hosts (see Figure 27).

When these steps are taken, it is impossible for a double-tagged packet to enter the network, and even if one did, it is very unlikely that it would have the proper tags to be switched to the untagged native VLAN or the target VLAN.

Figure 27 Mitigating Double-Tagged Packet Attacks



The following configuration example shows how to change the 802.1Q native VLAN to something other than 1 (the default). This helps prevent the VLAN hopping attack by making it difficult to correctly tag a packet.

For CatOS:

```
set vlan 207  1/1
clear trunk 1/1  1-6,8-49,52-106,108-206,208-554,556-4094
set trunk 1/1  nonegotiate dot1q 7,50-51,107,207,555
```

For Cisco IOS software:

```
switchport trunk native vlan 207
switchport trunk allowed vlan 7,50-51,107,207,555
```

The following configuration example shows how to change the user-facing port configuration so that tagged traffic is not supported.

CatOS:

```
set port host
```

For Cisco IOS software:

```
switchport host
```

The recommended way to configure an access port is with the **host** macro. Use the CatOS **set port host** or the Cisco IOS software **switchport host** commands to disable trunking and EtherChannel, and to enable STP PortFast.

Additionally, Cisco switch operating software can now tag all native VLAN traffic. This removes any possibility that a double 802.1Q-tagged packet can hop VLANs.

The following configuration examples enforce tagging of all native VLAN traffic:

For CatOS:

```
set dot1q-all-tagged enable
```

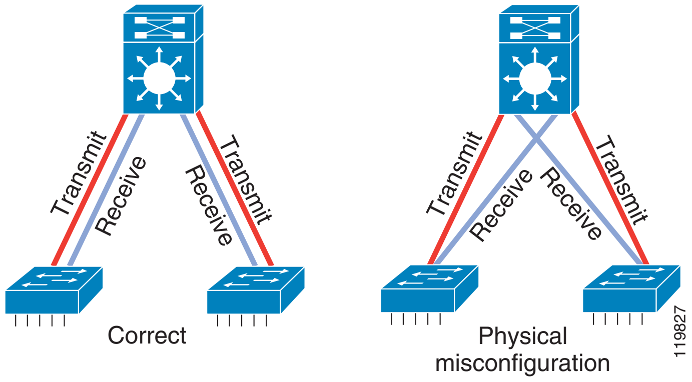
For Cisco IOS software:

```
switchport native vlan tag
```

Protecting Against One-Way Communication with UniDirectional Link Detection

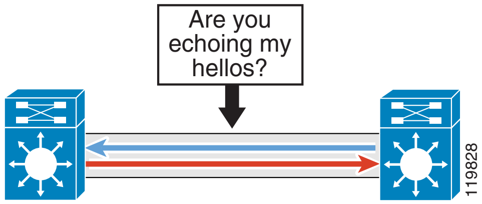
Because one-way communication is possible in fiber optic environments, mismatched transmit/receive pairs can cause a link up/up condition even though bidirectional communication has not been established. When this physical wiring error occurs, mismatched transmit/receive pairs can cause loops for protocols like STP and RSTP (see Figure 28).

Figure 28 Mismatched Transmit/Receive Pairs



UniDirectional Link Detection (UDLD) provides protection from this type of physical misconfiguration. UDLD monitors hello messages to ensure that a response is received from the destination device, as shown in Figure 29.

Figure 29 UniDirectional Link Detection



If a hello is not received that contains the port and node information of the sending machine, this indicates a misconfiguration and the port is error-disabled.

Enable UDLD aggressive mode in all environments where fiber optic interconnections are used. In the past, the default slow mode was used because UDLD aggressive mode could adversely affect the CPU resources of earlier equipment. However, this is no longer a concern in campus topologies with current hardware. You should enable UDLD in global mode so you do not have to enable it on every individual fiber optic interface.

The following configuration examples show how to enable UDLD for CatOS and Cisco IOS software.

For CatOS:

```
set udld enable
set udld aggressive-mode en <mod/port>
```

For Cisco IOS software (in global configuration mode):

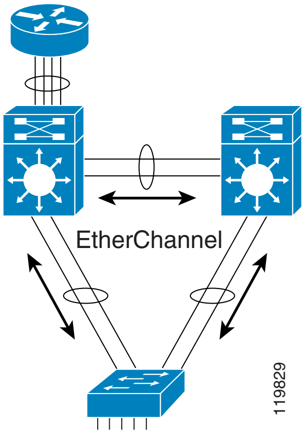
```
udld aggressive
```

Link Aggregation—EtherChannel Protocol and 802.3ad

The logical grouping of multiple redundant links into a single logical entity is called a link aggregation. There are two variants: the pre-standard Cisco EtherChannel implementation that uses Port Aggregation Protocol (PAgP) as a control mechanism, and the IEEE 802.3ad standards-based implementation that uses Link Aggregation Control Protocol (LACP) as its control mechanism. The two protocols are interoperable, with some manual configuration required. For the remainder of this document, the term EtherChannel is used to describe both variants.

An EtherChannel aggregates the bandwidth of redundant links and prevents a single point of failure. Without this logical grouping, STP/RTSP would place the redundant interface into blocking state to maintain a loop-free topology (See Figure 30).

Figure 30 EtherChannels



You can create channels containing up to eight parallel links between switches. You can also create these channels on interfaces that are on different physical line cards, which provides increased availability because the failure of a single line card does not cause a complete loss of connectivity. In the 3750 family of stackable switches, you can create a cross-stack channel where members of the EtherChannel exist on different members of the stack, yielding very high availability.

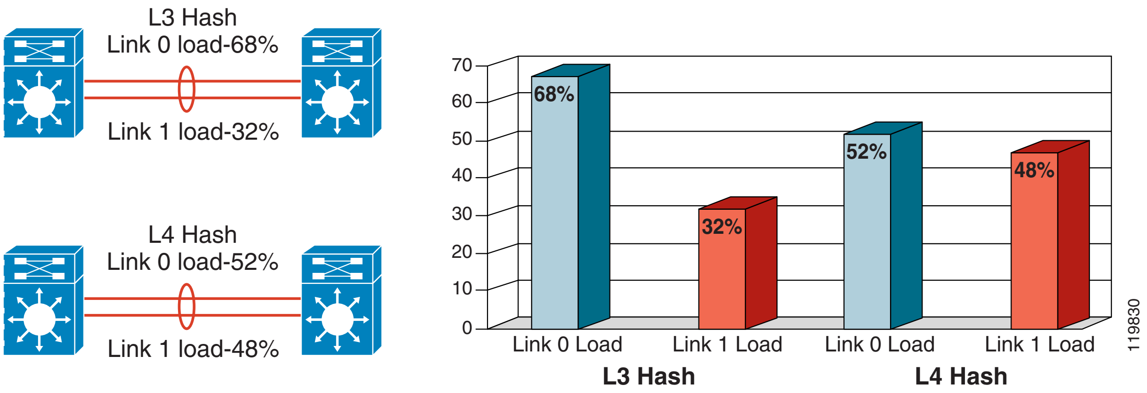
EtherChannels are typically deployed between the distribution-to-core and core-to-core interconnections where increased availability and scaled bandwidth are required. With multiple individual point-to-point L3 interfaces, the number of L3 neighbor relationships is greatly increased and this unnecessarily increases memory and configuration requirements.

Cisco switches let you tune the hashing algorithm used to select the specific EtherChannel link on which a packet is transmitted. You can use the default source/destination IP information, or you can add an additional level of load balancing to the process by adding the L4 TCP/IP port information as an input to the algorithm.

The current best practice is to use as much information as possible for input to the EtherChannel algorithm to achieve the best or most uniform utilization of EtherChannel members.

In a test environment using a typical IP addressing scheme of one subnet per VLAN and two VLANs per access switch using the RFC1918 private address space, the default L3 algorithm provided about one-third to two-thirds utilization. When the algorithm was changed to include L4 information, nearly full utilization was achieved with the same topology and traffic pattern (see Figure 31).

Figure 31 EtherChannel Testing Results



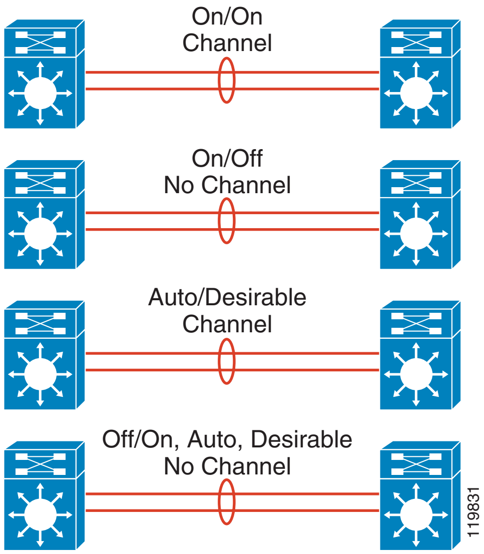
The following configuration example shows how to change the EtherChannel input algorithm on a Cisco Catalyst 6000 Series switch using CatOS.

```
port-channel load-balance src-dst-port
```

Link Aggregation Protocol

PAgP or LACP enable the automatic formation of EtherChannel tunnels between interconnected switches (see Figure 32).

Figure 32 Port Aggregation Protocol Operation



PAgP has four modes related to the automatic formation of bundled, redundant switch-to-switch interconnections:

- On—Always be an EtherChannel tunnel member
- Desirable—Request that the other side become a member
- Auto—Become a member at the request of the other side
- Off—Do not become a member

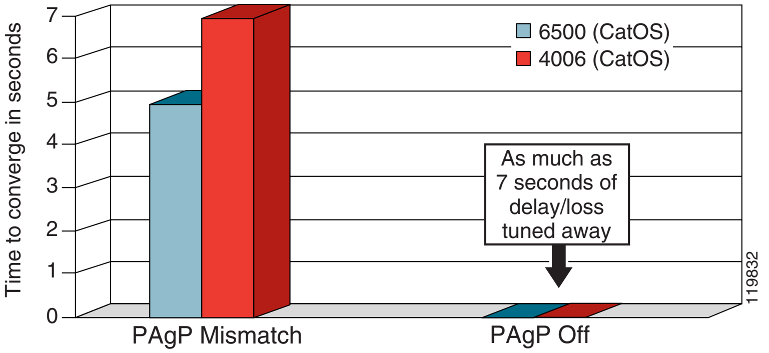
As with Trunking/DTP, the long-standing practice for EtherChannel/PAgP has been to set one side of the interconnection (typically the access switch) to **auto** and the other side (typically the distribution switch) to **desirable**. In this configuration, a trunk is established when configuration is complete, and connectivity to the remote switch is always available, even when the channel is not completely established.

Although this option allows for the *safest* deployment, there is a performance cost when a link or node is restored and channel negotiation occurs. Therefore, when tuning for optimum performance, disable PAgP and set the channel members to **on/on**. When making this optimization, there is a trade-off to be considered: when using the Auto/Desirable setting, PAgP is enabled, protecting against misconfiguration and hardware failure scenarios that can cause STP loops to form. When using the on/on setting, PAgP is not enabled on members of the bundle. Misconfiguration (mis-matched pairs) or hardware failure can result in unexpected STP behavior.

A specific situation can cause considerable periods of packet loss during channel negotiation when mixing CatOS in the access layer and Cisco IOS software in the distribution layer. The default state for PAgP in CatOS is **desirable**, meaning that a CatOS switch tries to negotiate an EtherChannel. The default state for Cisco IOS software is **off**. Unless you explicitly create a port channel interface and make the physical interface part of the EtherChannel, PAgP is not enabled and EtherChannel negotiation does not occur.

On links between a CatOS device and a Cisco IOS software device, you should disable PAgP negotiation if EtherChannel tunnels are not required. If you do not disable EtherChannel negotiation, then the mismatch between the default states of CatOS and Cisco IOS software can cause as much as seven seconds of loss during link negotiation, as shown in Figure 33.

Figure 33 PAgP Performance Results



Use the following command to disable PAgP negotiation:

For CatOS:

```
set port channel <mod/port> off
```

Additionally, port aggregation should be disabled on interfaces facing end users. This is most effectively accomplished by using the **set port host** macro which disables trunking, EtherChannel, and enables STP PortFast:

For CatOS:

```
set port host
```

For Cisco IOS software:

```
switchport host
```

The following configuration snippets demonstrate the EtherChannel configuration used to achieve optimum convergence:

For Cisco IOS software: (global configuration mode):

```
port-channel load-balance src-dst-port
```

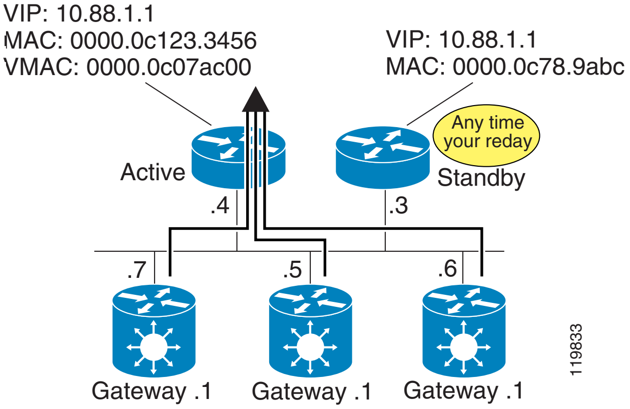
For Cisco IOS software (interface configuration mode):

```
interface GigabitEthernet2/1
description to 6k-Core-left CH#1
no ip address
mls qos trust dscp
channel-group 1 mode on
!
interface GigabitEthernet2/2
description to 6k-Core-left CH#1
no ip address
mls qos trust dscp
channel-group 1 mode on
!
interface Port-channel1
description to cr2-6500-1 CHANNEL #1
ip address 10.122.0.34 255.255.255.252
mls qos trust dscp
```

Using HSRP, VRRP, or GLBP for Default Gateway Redundancy

Default gateway redundancy (also known as first hop redundancy), allows a highly available network to recover from the failure of the device acting as the default gateway for the end stations on a physical segment (see Figure 34).

Figure 34 First Hop Default Gateway Redundancy



In the recommended hierarchical model, the distribution switches are the L2/L3 boundary and also act as the default gateway for the entire L2 domain that they support. Some form of redundancy is required because this environment can be large and a considerable outage could occur if the device acting as default gateway failed.

Cisco has developed the Hot Standby Router Protocol (HSRP) to address this need, and the IETF subsequently ratified Virtual Router Redundancy Protocol (VRRP) as the standards-based method of providing default gateway redundancy.

HSRP and VRRP with Cisco enhancements both provide a robust method of backing up the default gateway, and can provide sub-second failover to the redundant distribution switch when tuned properly. HSRP is the recommended protocol because it is a Cisco-owned standard, which allows for the rapid development of new features and functionality for HSRP before VRRP. VRRP is the logical choice when interoperability with a non-Cisco device is required. However, when interoperating with non-Cisco devices, you can use only the standard "lowest common denominator" features and you cannot take advantage of the Cisco enhancements to VRRP.

The configuration snippet below demonstrates how HSRP can be tuned in a campus environment to achieve sub-second convergence.

```
interface Vlan5
description Data VLAN for 6k-Access
ip address 10.1.5.3 255.255.255.0
ip helper-address 10.5.10.20
standby 1 ip 10.1.5.1
standby 1 timers msec 200 msec 750
standby 1 priority 150
standby 1 preempt
standby 1 preempt delay minimum 180
```

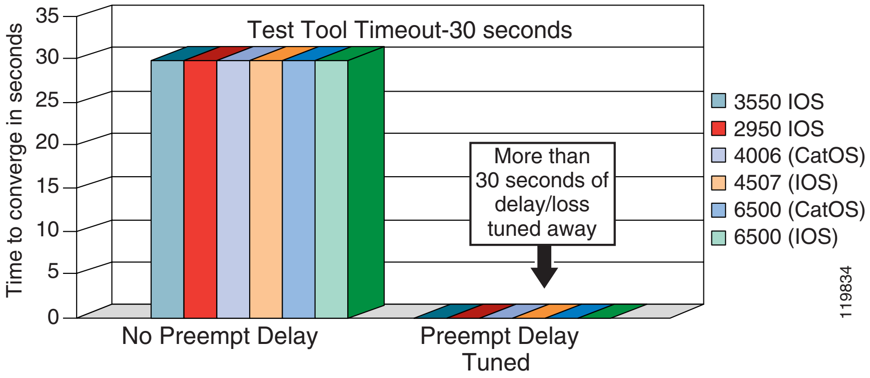
One important factor to take into account when tuning HSRP is its preemptive behavior. Preemption causes the primary HSRP peer to re-assume the primary role when it comes back online after a failure or maintenance event.

Preemption is the desired behavior because the STP/RSTP root should be the same device as the HSRP primary for a given subnet or VLAN. If HSRP and STP/RSTP are not synchronized, the interconnection between the distribution switches can become a transit link, and traffic takes a multi-hop L2 path to its default gateway.

HSRP preemption needs to be aware of switch boot time and connectivity to the rest of the network. It is possible for HSRP neighbor relationships to form and preemption to occur before the primary switch has L3 connectivity to the core. If this happens, traffic can be dropped until full connectivity is established.

The recommended best practice is to measure the system boot time, and set the HSRP preempt delay statement to 50 percent greater than this value. This ensures that the HSRP primary distribution node has established full connectivity to all parts of the network before HSRP preemption is allowed to occur (see Figure 35).

Figure 35 Preempt Delay Test Results



The best practice using Cisco IOS software is shown in the following configuration snippet:

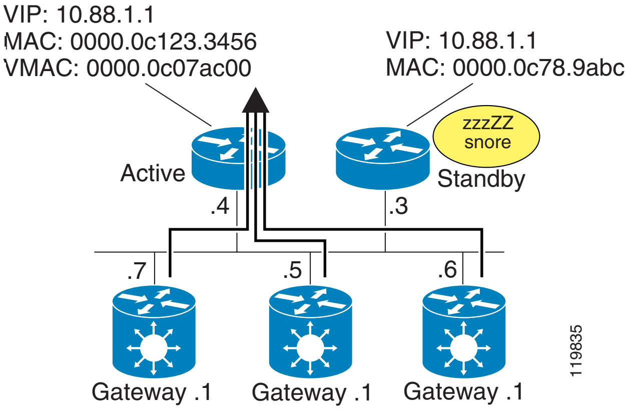
```
interface Vlan5
description Data VLAN for 6k-Access
ip address 10.1.5.3 255.255.255.0
ip helper-address 10.5.10.20
standby 1 ip 10.1.5.1
standby 1 timers msec 200 msec 750
standby 1 priority 150
standby 1 preempt
standby 1 preempt delay minimum 180
```

Gateway Load Balancing Protocol

Gateway Load Balancing Protocol (GLBP) protects data traffic from a failed router or circuit, like HSRP and VRRP, while allowing packet load sharing between a group of redundant routers.

When HSRP or VRRP are used to provide default gateway redundancy, the backup members of the peer relationship are idle, waiting for a failure event to occur for them to take over and actively forward traffic (see Figure 36).

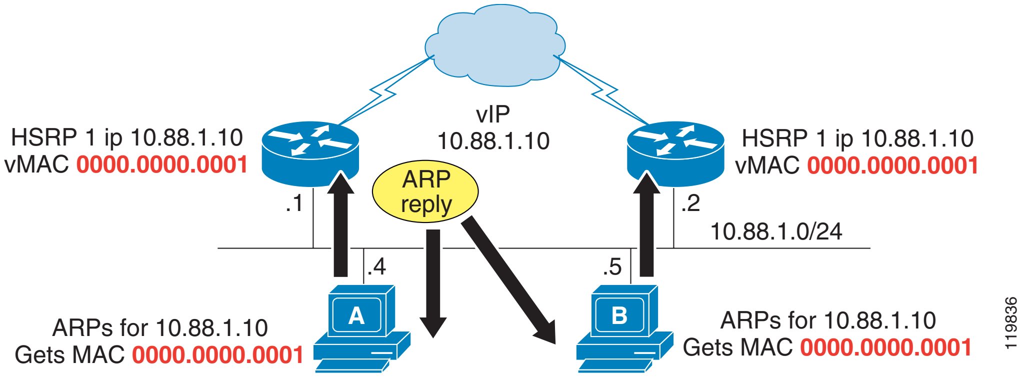
Figure 36 Idle Backup Capacity



Before the development of GLBP, methods used to utilize uplinks more efficiently were difficult to implement and manage. In one technique, the HSRP and STP/RSTP root alternated between distribution node peers, with the even VLANs homed on one peer and the odd VLANs homed on the alternate. Another technique used multiple HSRP groups on a single interface and used DHCP to alternate between the multiple default gateways. These techniques worked but were not optimal from a configuration, maintenance, or management perspective.

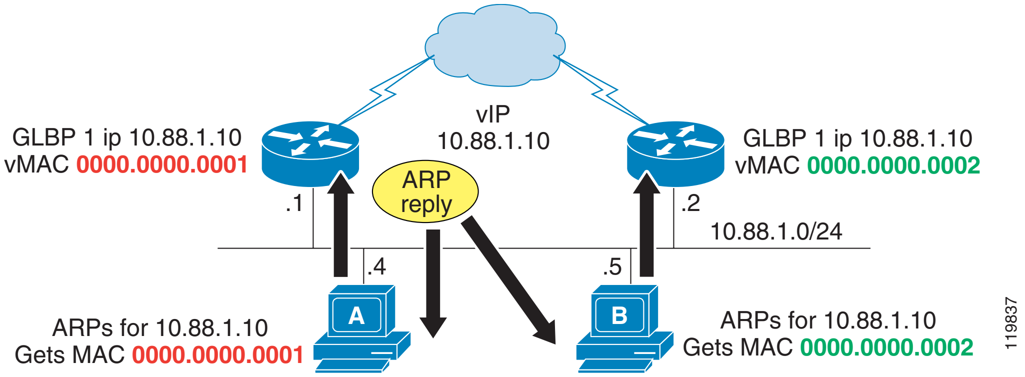
GLPB is configured and functions like HSRP. For HSRP, a single virtual MAC address is given to the end points when they use Address Resolution Protocol (ARP) to learn the physical MAC address of their default gateways (see Figure 37).

Figure 37 HSRP Operation



Two virtual MAC addresses exist with GLBP, one for each GLBP peer (see Figure 38).

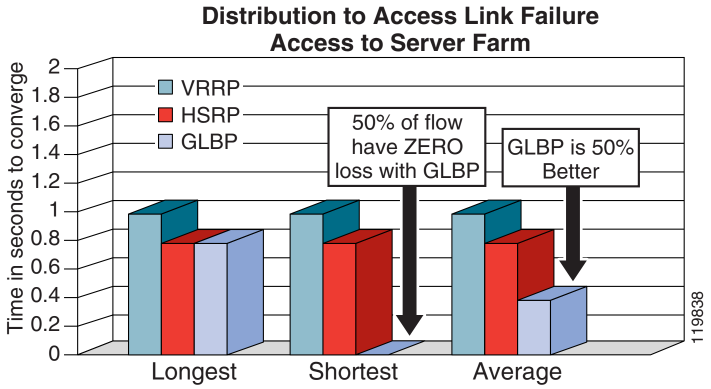
Figure 38 GLBP Operation



When an end point ARPs for its default gateway, the virtual MACs are checked out on a round-robin basis. Failover and convergence work just like HSRP. The backup peer assumes the virtual MAC of the device that has failed and begins forwarding traffic for its failed peer.

The end result is that a more equal utilization of the uplinks is achieved with minimal configuration. As a side effect, a convergence event on the uplink or on the primary distribution node affects only half as many hosts, giving a convergence event an average of 50 percent less impact (see Figure 39).

Figure 39 GLBP, HSRP, and VRRP Test Results

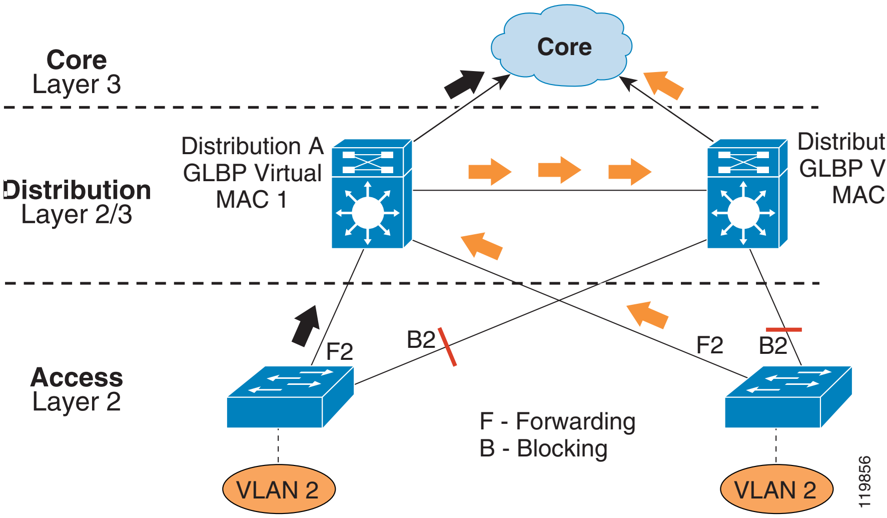


The configuration snippet below demonstrates how GLBP was configured to achieve these results.

```
interface Vlan7
description Data VLAN for 4k-Access
ip address 10.120.7.3 255.255.255.0
ip helper-address 10.121.0.5
glbp 1 ip 10.120.7.1
glbp 1 timers msec 250 msec 750
glbp 1 priority 150
glbp 1 preempt delay minimum 180
```

As shown in Figure 40, it is important to note that using GLBP in topologies where STP has blocked one of the access layer uplinks could cause a two-hop path at L2 for upstream traffic.

Figure 40 GLBP with STP Blocking Uplinks

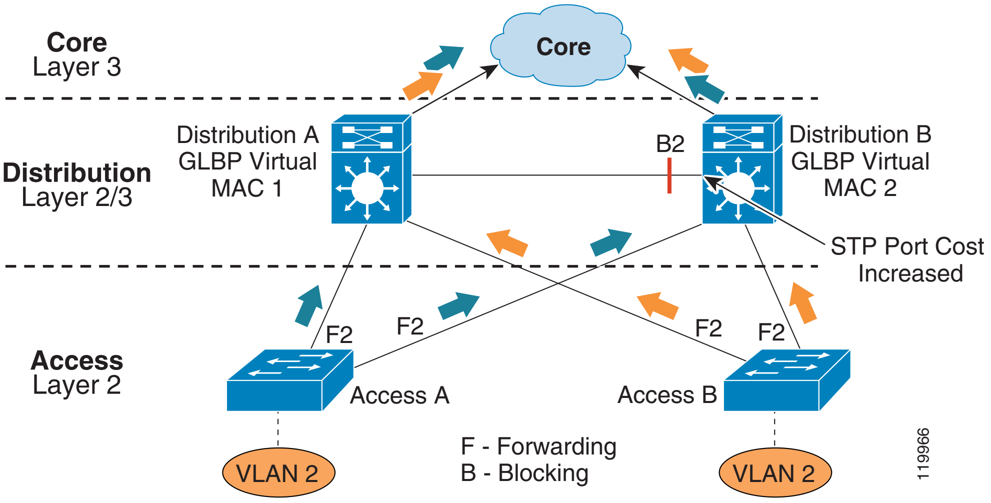


To avoid this situation the Spanning Tree environment must be tuned so that the L2 link between the distribution switches is the blocking link while the uplinks from the access layer switches are in a forwarding state. This can be most easily accomplished by changing the port cost on the interface between the distribution layer switches on the STP secondary root switch. On the interface facing the primary root switch, the following Cisco IOS command was entered in interface configuration mode to accomplish the desired effect.

```
spanning-tree cost 2000
```

Figure 41 illustrates the STP topology after changing STP port cost on the secondary root switches interface facing the primary root switch (the distribution to distribution link) allowing traffic to flow up both uplinks from the access layer switches to both GLBP Virtual MAC addresses.

Figure 41 GLBP with STP Blocking Distribution-to-Distribution Link



Oversubscription and QoS

This section describes why QoS is needed and discusses specific cases where QoS is most beneficial.

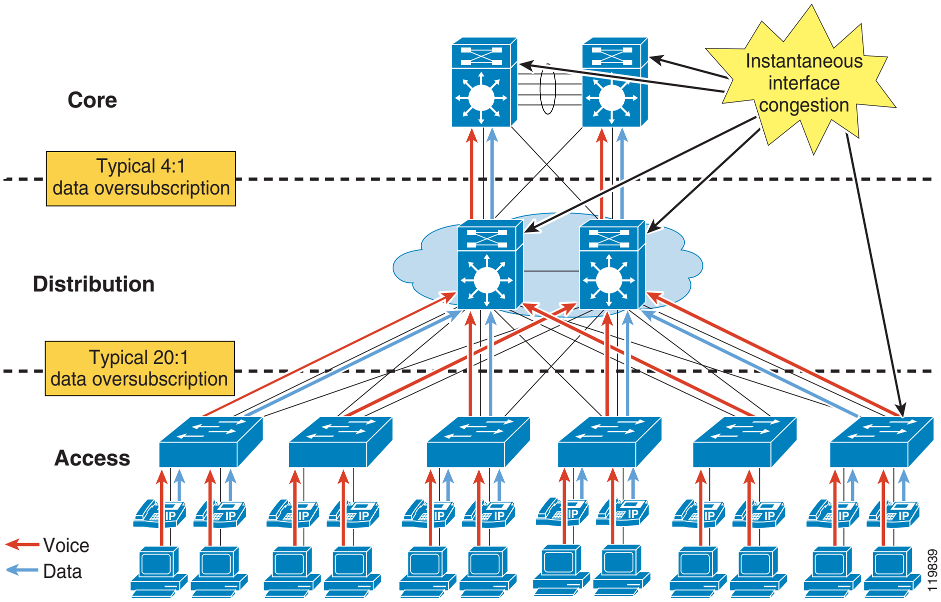
Note For more information refer to the QoS SRND (http://www.cisco.com/go/srnd), which provides configuration examples and a detailed explanation of other technical issues.

Typical campus networks are engineered with oversubscription. It is not generally practical to provide line rate for every port upstream from the access-to-distribution switch, the distribution-to-core switch, or even for core-to-core links. Even though bandwidth capacity has increased to 1 Gbps, multiples of 1 Gbps, and even 10 Gbps, it is still impractical to provide enough bandwidth to run an entire access layer switch full of ports at line rate at the same time.

The rule-of-thumb recommendation for oversubscription is 20:1 for access ports on the access-to-distribution uplink. The recommendation is 4:1 for the distribution-to-core links. In the data center, you may need a 1:1 ratio.

Using these oversubscription ratios, congestion on the uplinks occurs by design (see Figure 42).

Figure 42 Oversubscription Congestion

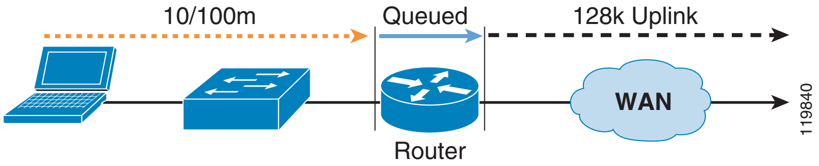


When congestion does occur, QoS is required to protect important traffic such as mission-critical data applications, voice, and video. Additionally, you can use QoS to reduce the priority of unwanted traffic. For example, an Internet worm infection, such as Slammer, can cause congestion on many links in the network, and QoS can minimize the effect of this event.

Congestion on a Cisco Catalyst switch interface is not typically caused by oversubscription or an anomaly such as an Internet worm. However, you must design for the unexpected to ensure that mission-critical applications including voice and video survive such situations.

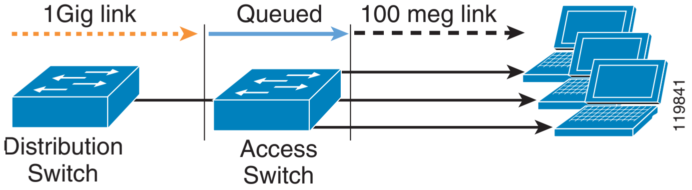
The type of congestion that is more prevalent in a campus network is called transmit queue (TX-queue) starvation. During a transition from LAN to WAN, a router has to make the rate transition from 10/100 Ethernet to WAN speeds. When this happens, the router must queue the packets and apply QoS to ensure that important traffic is transmitted first (see Figure 43).

Figure 43 WAN Rate Transitions



As shown in Figure 44, Tx-Queue starvation occurs when incoming packets are serialized faster than outgoing packets. Packets are queued as they wait to serialize out onto the slower link.

Figure 44 LAN Rate Transitions

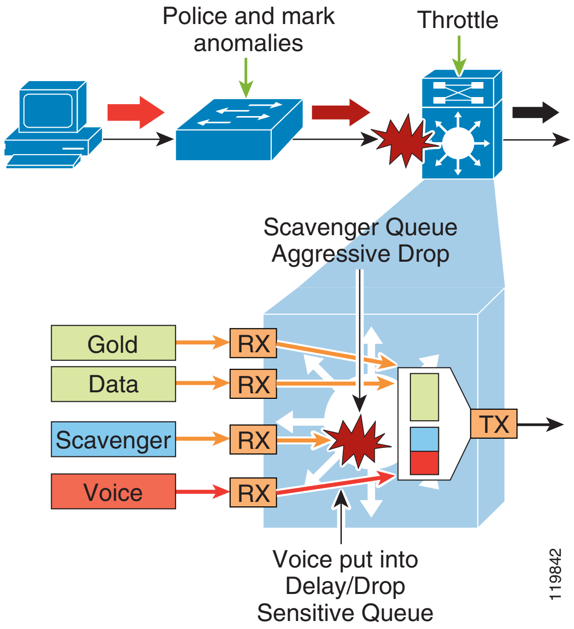


In the campus, as we transition from 10 Gbps or 1 Gbps to 10/100 Gbps to the desktop, packets must be queued as they wait to serialize out the 10 or 100 Mbps link. The difference between a WAN router and a campus switch is the number of interfaces and the amount of memory associated with each. In the campus, the amount of Tx-queue space is much smaller than the amount of memory available in a WAN router. Because of this small amount of memory, the potential for dropped traffic because of Tx-queue starvation is relatively high.

Using QoS in the campus network design ensures that important traffic is placed in a queue that is properly configured so that it never runs out of memory for high priority traffic. Under normal circumstances, the network should provide an adequate level of service for all network traffic, including lower priority best-effort traffic.

During periods of congestion, scavenger-class traffic is the first to experience Tx-queue starvation and packet loss because the bandwidth is reserved for higher priority traffic. As demand increases or capacity is reduced, best-effort traffic may also be affected. The minimum goal of high availability network design is to ensure that high-priority, mission-critical data applications and voice/video are never affected by network congestion (see Figure 45).

Figure 45 QoS Mitigation of Tx-Queue Starvation


Police and mark anomalies
Throttle
Scavenger Queue Aggressive Drop
Gold
Data
Scavenger
Voice
RX
RX
RX
RX
TX
Voice put into Delay/Drop Sensitive Queue
119842

Design Best Practices

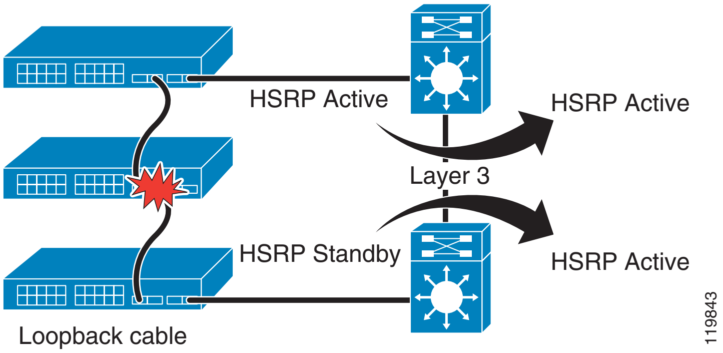
This section describes the recommended best practices for ensuring high availability in the campus network and includes the following topics:

- Daisy Chaining Dangers, page 43
- Asymmetric Routing and Unicast Flooding, page 45
- Designing for Redundancy, page 47
- Spanning VLANs Across Access Layers Switches, page 51
- Deploying the L2 /L3 Boundary at the Distribution Layer, page 51
- Routing in the Access Layer, page 53

Daisy Chaining Dangers

If multiple fixed-configuration switches are used in the access layer of the network, be careful that black holes do not occur in the event of a link or node failure. In Figure 46, an L3 connection exists between the distribution nodes. In this topology, no links are blocking from a STP/RSTP perspective, so both uplinks are available to actively forward and receive traffic.

Figure 46 L3 Connection Between Distribution Nodes

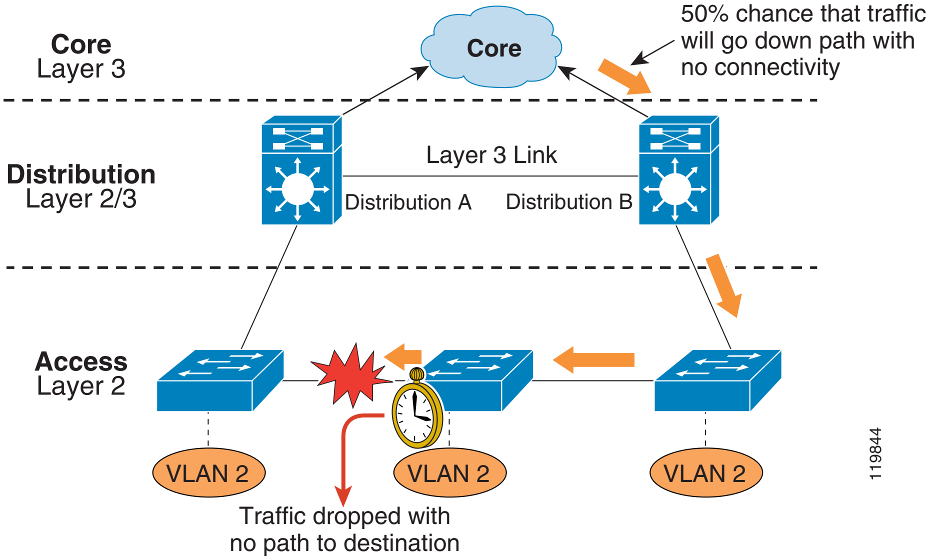


Both distribution nodes can forward return path traffic from the rest of the network towards the access layer for devices attached to all members of the stack or chain. Two things can happen if a link or node in the middle of the chain or stack fails.

In the first case, the standby HSRP peer can go active as it loses connectivity to its primary peer, forwarding traffic outbound for the devices that still have connectivity to it. The primary HSRP peer remains active and also forwards outbound traffic for its half of the stack. While this is not optimum, it is also not detrimental from the perspective of outbound traffic.

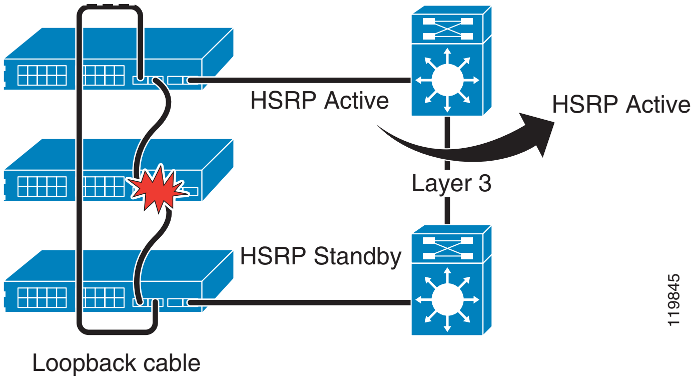
The second scenario presents a problem. Return path traffic has a 50/50 chance of arriving on a distribution switch that does not have physical connectivity to the half of the stack where the traffic is destined. Traffic is dropped when it arrives on the wrong distribution switch (see Figure 47).

Figure 47 Dropped Traffic



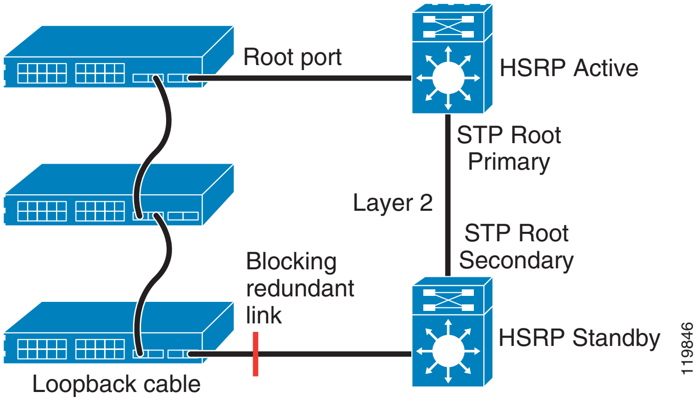
The solution to this problem is to provide alternate connectivity across the stack in the form of a loopback cable running from the top to the bottom of the stack, as shown in Figure 48.

Figure 48 Loopback Cable



If there is an L2 interconnection between the distribution switches, then you must consider the STP/RSTP root and link cost so that the correct interfaces are placed in a blocking state. A loopback cable is not required to ensure connectivity because traffic can pass over the distribution-to-distribution interconnection, as shown in Figure 49.

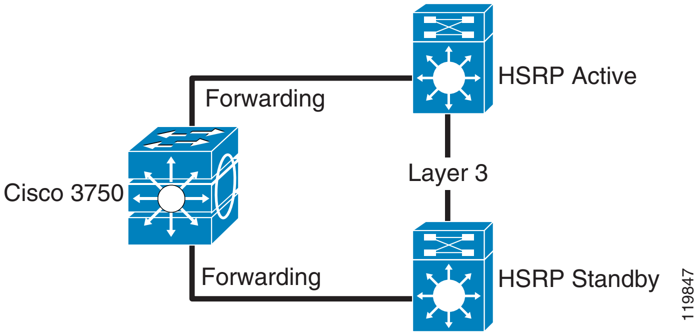
Figure 49 Layer 2 Interconnection



When stacking technology is used to interconnect the fixed configuration switches, and an L2 link is used for the interconnection between the distribution switches, it is important to use STP/RSTP enhancements such as Cross-stack UplinkFast so that the uplinks to the distribution node can rapidly transition to a forwarding state in the event of a link or node failure that would require STP/RSTP convergence.

If StackWise technology is utilized, you can follow the best practice recommendation by using an L3 connection between the distribution switches without having to use a loop-back cable or perform extra configuration. The true stack creation provided by the Cisco Catalyst 3750 family of fixed-configuration switches makes using stacks in the access layer much less complex than chains or stacks of other models (see Figure 50).

Figure 50 Cisco Catalyst 3750 Stack



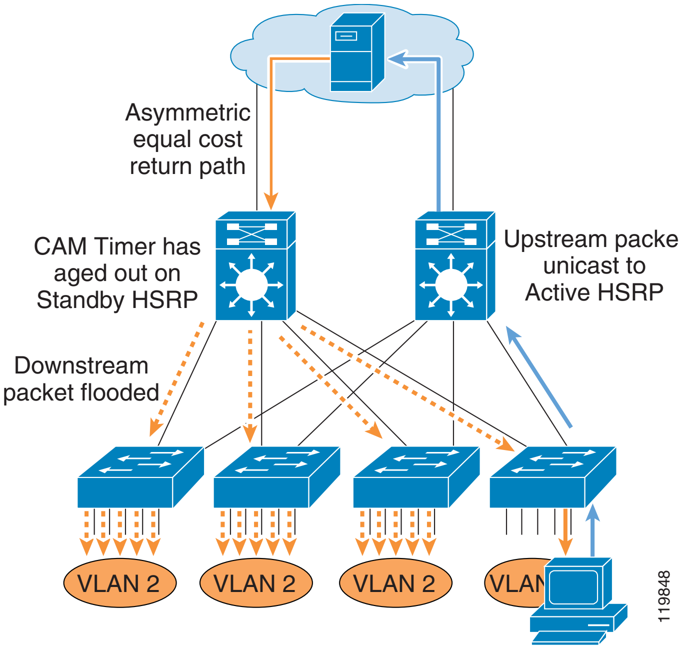
Additionally, if you use a modular chassis switch, such as the Cisco Catalyst 4500 or Catalyst 6500 family of switches, these design considerations are not required.

Asymmetric Routing and Unicast Flooding

Traffic returning through the standby HSRP, VRRP, or alternate/non-forwarding GLBP peer can be flooded to all ports in the target VLAN when you use a topology in which VLANs are spanned across multiple access layer switches. This can have significant impact on performance. Figure 51 illustrates a redundant topology where a common VLAN is shared across the access layer switches.

.

Figure 51 Asymmetric Routing with Unicast Flooding



In this topology, the CAM table entry ages out on the standby HSRP router. This occurs because the ARP and CAM aging timers are different. The CAM timer expires because no traffic is sent upstream towards the standby HSRP peer after the end point initially ARPs for its default gateway. When the CAM entry has aged out and is removed, the standby HSRP peer must forward the return path traffic to all ports in the common VLAN.

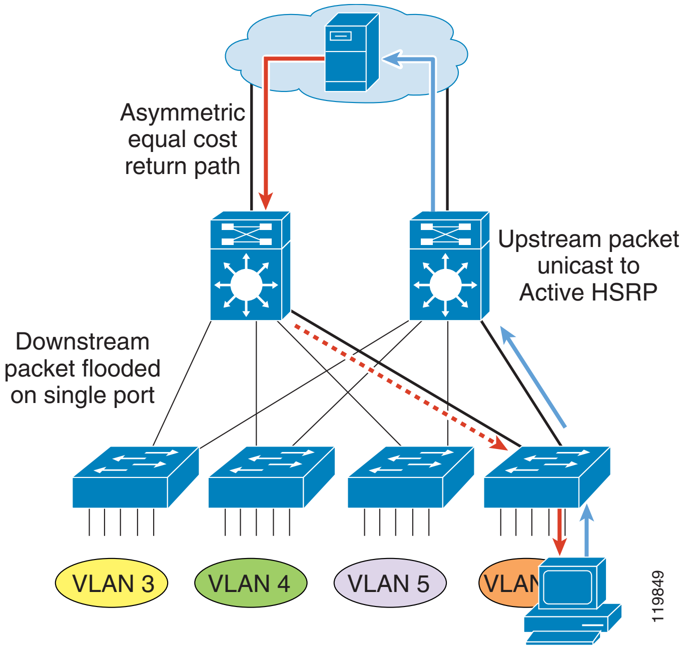
The corresponding access layer switches also do not have a CAM entry for the target MAC, and they also broadcast the return traffic on all ports in the common VLAN. This traffic flooding can have a performance impact on the connected end stations because they may receive a large amount of traffic that is not intended for them.

If you must implement a topology where VLANs span more than one access layer switch, the recommended work-around is to tune the ARP timer to be equal to or less than the CAM aging timer. A shorter ARP cache timer causes the standby HSRP peer to ARP for the target IP address before the CAM entry timer expires and the MAC entry is removed. The subsequent ARP response repopulates the CAM table before the CAM entry is aged out and removed. This removes the possibility of flooding asymmetrically-routed return path traffic to all ports.

As stated earlier, this problem only occurs in a topology where VLANs span multiple access layer switches in a large L2 domain. This is not an issue when VLANs are not present across access layer switches because the flooding occurs only to switches where the traffic would have normally been switched. Additionally, larger L2 domains have a greater potential for impact on end-station performance because the volume of potentially flooded traffic increases in larger L2 environments.

If you build a topology where VLANs are local to individual access layer switches, this type of problem is inconsequential because traffic is only flooded on one interface (the only interface in the VLAN) on the standby HSRP, VRRP, or non-forwarding GLBP peer. Traffic is flooded out the same interface that would be used normally, so the end result is the same. Additionally, the access layer switch receiving the flooded traffic has a CAM table entry for the host because it is directly attached, so traffic is switched only to the intended host. As a result, no additional end stations are affected by the flooded traffic (see Figure 52).

Figure 52 Traffic Flooding on Single Interface

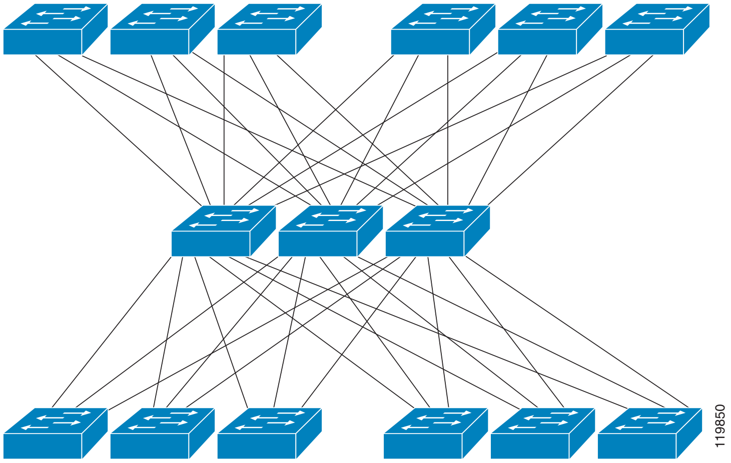

Asymmetric equal cost return path
Upstream packet unicast to Active HSRP
Downstream packet flooded on single port
VLAN 3
VLAN 4
VLAN 5
VLAN
119849

Designing for Redundancy

The hierarchical network model stresses redundancy at many levels to remove a single point of failure wherever the consequences of a failure are serious. At the very least, this model requires redundant core and distribution layer switches with redundant uplinks throughout the design. The hierarchical network model also calls for EtherChannel interconnection for key links where a single link or line card failure can be catastrophic.

When it comes to redundancy, however, you can have too much of a good thing. Take care not to over-duplicate resources. There is a point of diminishing returns when the complexity of configuration and management outweighs any benefit of the added redundancy (see Figure 53).

Figure 53 Over-Duplicated Resources



In Figure 53, the addition of a single switch to a very basic topology adds several orders of magnitude in complexity. This topology raises the following questions:

- Where should the root switch be placed?
- What links should be in a blocking state?
- What are the implications of STP/RSTP convergence?
- When something goes wrong, how do you find the source of the problem?

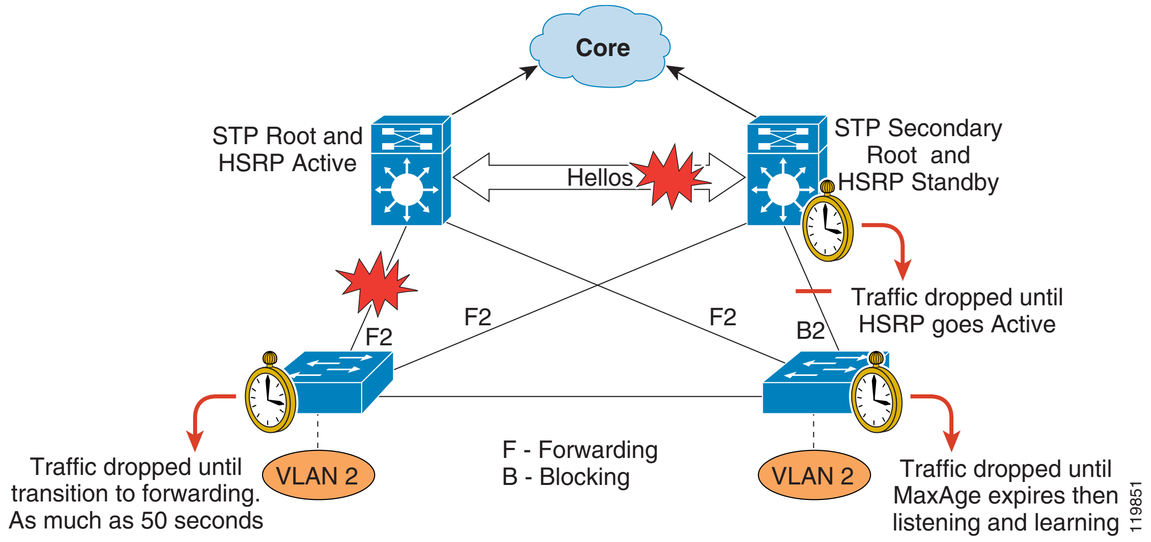
When there are only two switches in the center of this topology, the answers to those questions are straightforward and clear. In a topology with three switches, the answer depends on many factors.

However, the other extreme is also a bad thing. You might think that completely removing loops in a topology that requires the spanning of multiple VLANs across access layer switches might be a good thing. After all, this eliminates the dependence of convergence on STP/RSTP.

However, this approach can cause its own set of problems (see Figure 54), including the following:

- Traffic is dropped until HSRP becomes active.
- Traffic is dropped until the link transitions to forwarding state, taking as long as 50 seconds.
- Traffic is dropped until the MaxAge timer expires and until the listening and learning states are completed.

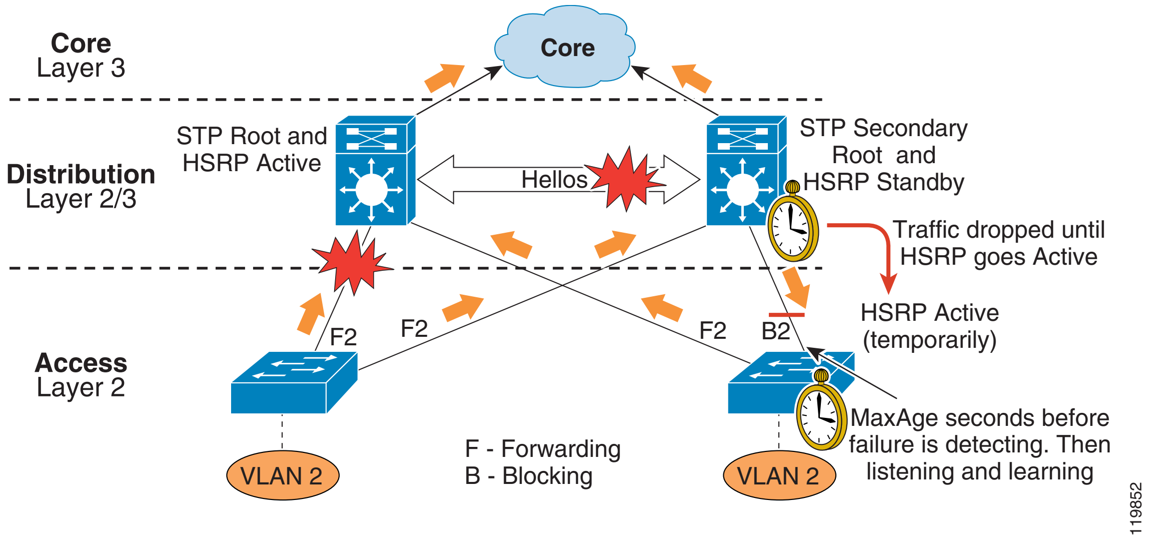
Figure 54 Removal of L2 Distribution-to-Distribution Link


Core
STP Root and HSRP Active
Hellos
STP Secondary Root and HSRP Standby
Traffic dropped until HSRP goes Active
F2
F2
F2
B2
Traffic dropped until transition to forwarding. As much as 50 seconds
VLAN 2
F - Forwarding
B - Blocking
VLAN 2
Traffic dropped until MaxAge expires then listening and learning
119851

Spanning-Tree convergence can cause considerable periods of packet loss because of the time that STP/RSTP takes to react to transition events.

Additionally, when you remove a direct path of communication for the distribution layer switches, you then become dependent on the access layer for connectivity. This can introduce unexpected behavior in the event of a failure, as demonstrated in the order of convergence events that occur when an individual uplink fails in a topology (see Figure 55).

Figure 55 Convergence Events with an Uplink Failure


Core
Layer 3
Core
Distribution
Layer 2/3
STP Root and HSRP Active
Hellos
STP Secondary Root and HSRP Standby
Traffic dropped until HSRP goes Active
HSRP Active (temporarily)
Access
Layer 2
F2
F2
F2
B2
MaxAge seconds before failure is detecting. Then listening and learning
VLAN 2
F - Forwarding
B - Blocking
VLAN 2
119852

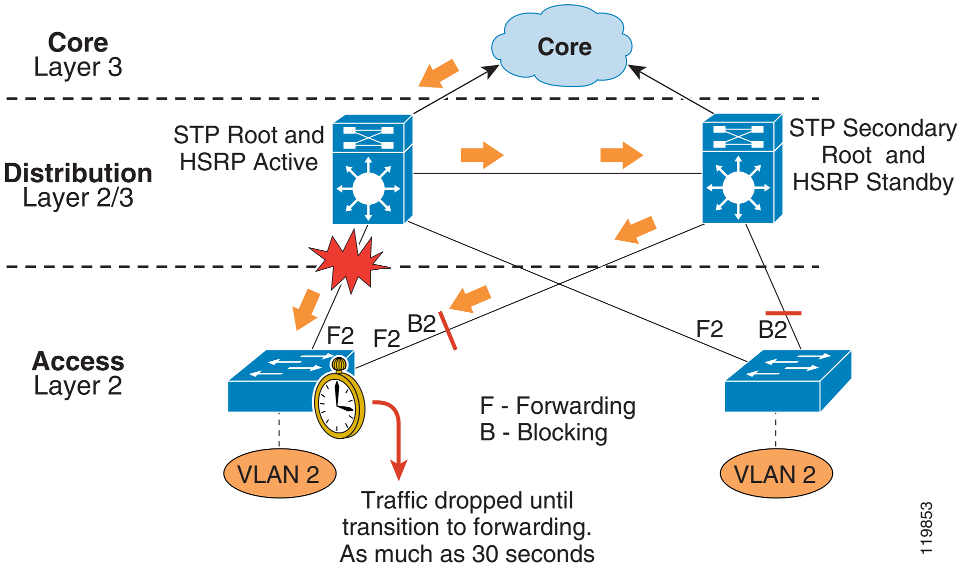
When the link from Access-a to the STP root and the HSRP primary switch fails, traffic is lost until the standby HSRP peer takes over as the default gateway. With aggressive HSRP timers (such as those previously recommended in this document), you can minimize this period of traffic loss to approximately 900 milliseconds.

Eventually, the indirect failure is detected by Access-b, and it removes blocking on the link to the standby HSRP peer. With standard STP, this can take as long as 50 seconds. If BackboneFast is enabled with PVST+, this time can be limited to 30 seconds, and Rapid PVST+ can reduce this interval to as little as one second.

When an indirect failure is detected and STP/RSTP converges, the distribution nodes reestablish their HSRP relationships and the primary HSRP peer preempts. This causes yet another convergence event when Access-a end points start forwarding traffic to the primary HSRP peer. The unexpected side effect is that Access-a traffic goes through Access-b to reach its default gateway. The Access-b uplink to the backup HSRP peer to Access-b is now a transit link for Access-a traffic, and the Access-b uplink to the primary HSRP peer must now carry traffic for both Access-b (its original intent) and for Access-a.

The behavior of the outbound traffic from the access layer to the rest of the network was described in the previous example (Figure 55). Return path traffic for the same convergence event in this topology is shown in Figure 56.

Figure 56 Convergence Events with Return Path Traffic



In the topology shown in Figure 57, the following convergence times can be observed:

- With 802.1d—Up to 50 seconds
- With PVST+ (with UplinkFast)—Up to 5 seconds
- With Rapid PVST+ (address by the protocol)—1 second

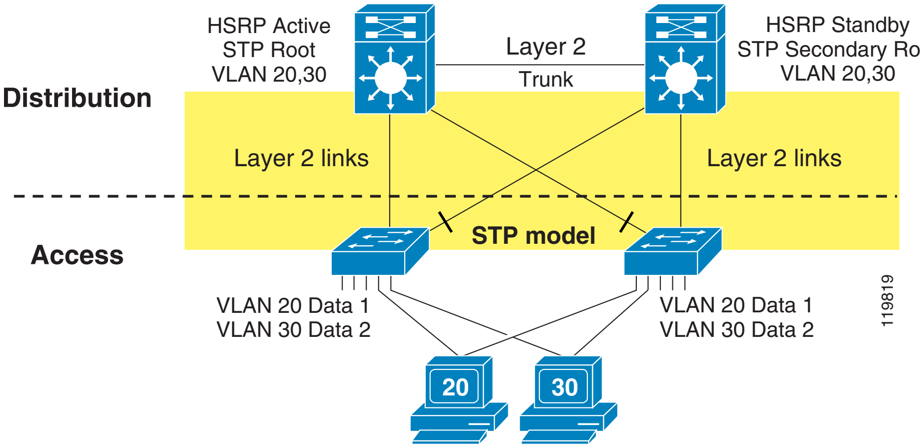
Return path traffic for hosts on Access-a arrive on Access-b and are dropped until the indirect failure is detected and the uplink to the standby HSRP peer goes active. This can take as long as 50 seconds. PVST+ with UplinkFast reduces this to 3–5 seconds, and Rapid PVST+ further reduces the outage to one second. After the STP/RSTP convergence, the Access-b uplink to the standby HSRP peer is used as a transit link for Access-a return path traffic.

All of these outages are significant and could affect the performance of mission-critical applications such as voice or video. Additionally, traffic engineering or link capacity planning for both outbound and return path traffic is difficult and complex, and you must plan to support the traffic for at least one additional access layer switch.

Spanning VLANs Across Access Layers Switches

This section describes the best way to build a topology that includes VLANs spanning access layer switches and that depend on STP/RSTP for convergence (see Figure 57).

Figure 57 Best Practice Topology for Spanning VLANs Across Access Layer Switches



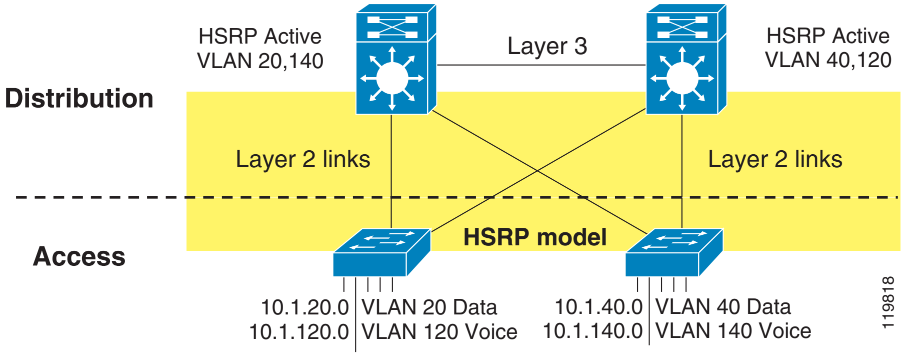
If your applications require spanning VLANs across access layer switches and using STP as an integral part of your convergence plan, take the following steps to make the best of this suboptimal situation:

- Use Rapid PVST+ as the version of STP. When spanning-tree convergence is required, Rapid PVST+ is superior to PVST+ or plain 802.1d.
- Provide an L2 link between the two distribution switches to avoid unexpected traffic paths and multiple convergence events.
- If you choose to load balance VLANs across uplinks, be sure to place the HSRP primary and the STP primary on the same distribution layer switch. The HSRP and Rapid PVST+ root should be co-located on the same distribution switches to avoid using the inter-distribution link for transit.

Deploying the L2 /L3 Boundary at the Distribution Layer

The time-proven topology that provides the highest availability does not require STP/RSTP convergence. In this topology, no VLANs span access layer switches and the distribution layer interconnection is an L3 point-to-point link. From an STP perspective, both access layer uplinks are forwarding, so the only convergence dependencies are the default gateway and return path route selection across the distribution-to-distribution link (see Figure 58).

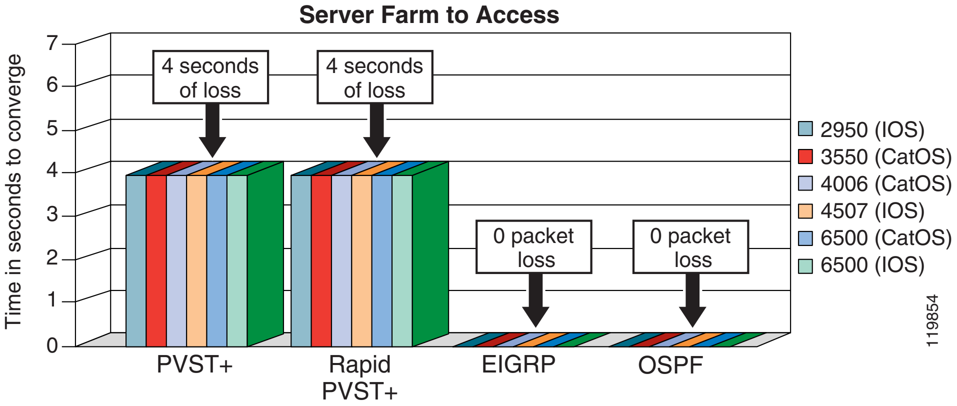
Figure 58 Best Practice Topology



You can achieve reliable default gateway failover from the HSRP primary to the HSRP standby in less than 900 ms by tuning the HSRP timers, as described in the section, "Using HSRP, VRRP, or GLBP for Default Gateway Redundancy, page 35."

EIGRP can reroute around the failure in 700-1100 ms for the return path traffic. For details, see *High Availability Campus Recovery Analysis*. This topology yields a sub-second bi-directional convergence in response to a failure event (see Figure 59).

Figure 59 HSRP Tuning Test Results



When implementing this topology, be aware that when the primary HSRP peer comes back online and establishes its L3 relationships with the core, it must ARP for all the end points in the L2 domain that it supports. This happens as equal-cost load sharing begins to occur and return path traffic starts to flow through the node, regardless of HSRP state because this is for return path traffic. ARP processing is rate limited in Cisco IOS software and in hardware to protect the CPU against DoS attacks that might overrun the CPU with an extraordinary number of ARP requests.

The end result is that for return path traffic, the distribution node that is coming back online can not resolve all the IP to MAC addresses for the L2 domain that it supports for a considerable period of time. In a 40-node access layer test, recovery times of up to four seconds were measured for all flows to be re-established during this convergence event. Results vary depending on the size of the L2 domain supported by the distribution pair.

Routing in the Access Layer

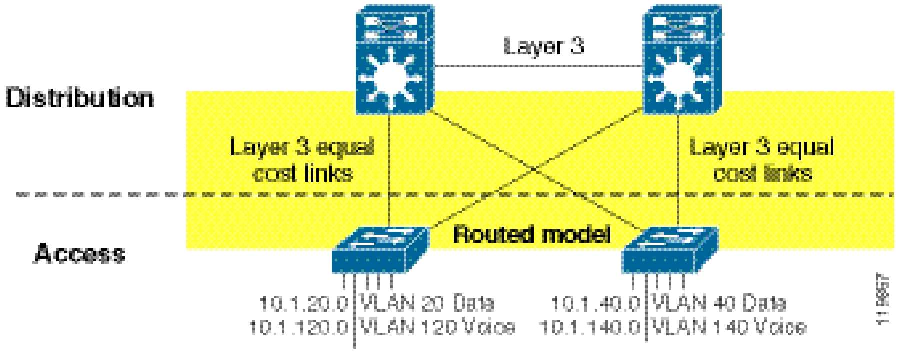
This section includes the following topics:

- Deploying the L2/L3 Boundary at the Access Layer, page 53
- Comparing Routing Protocols, page 55
- Using EIGRP in the Access Layer, page 56
- Using OSPF in the Access Layer, page 57

Deploying the L2/L3 Boundary at the Access Layer

Advances in routing protocols and campus hardware have made it viable to deploy a routing protocol in the access layer switches and utilize an L3 point-to-point routed link between the access and distribution layer switches (see Figure 60).

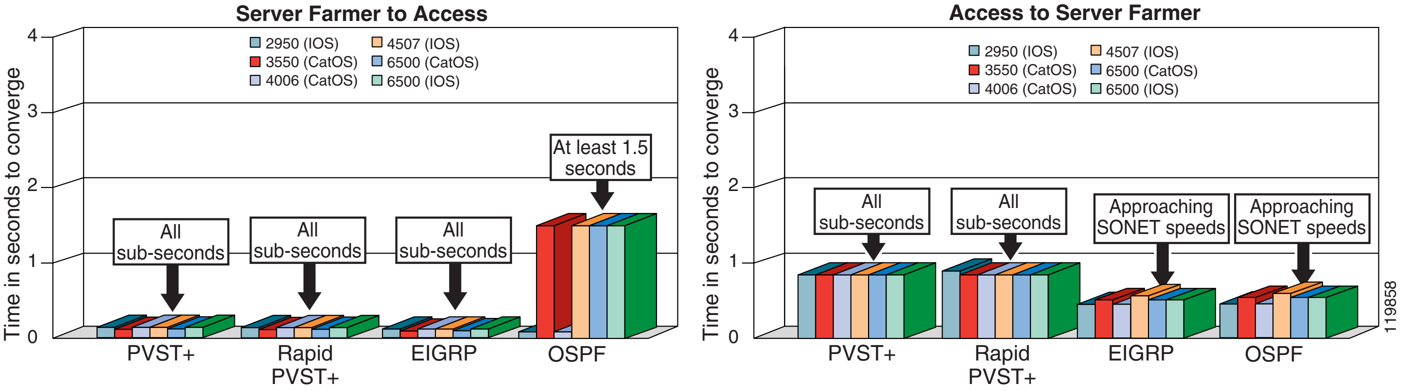
Figure 60 Fully Routed Solution with Point-to-Point L3 Links



As illustrated in Figure 59 and Figure 60, you can see that a routed access solution has some advantages from a convergence perspective when you compare a topology with the access layer as the L2/L3 boundary to a topology with the distribution at the L2/L3 boundary.

The convergence time required to reroute around a failed access-to-distribution layer uplink is reliably under 200 milliseconds as compared to 900 milliseconds for the L2/L3 boundary distribution model. Return path traffic is also in the sub-200 milliseconds of convergence time for an EIGRP re-route, again compared to 900 milliseconds for the traditional L2/L3 distribution layer model (see Figure 61).

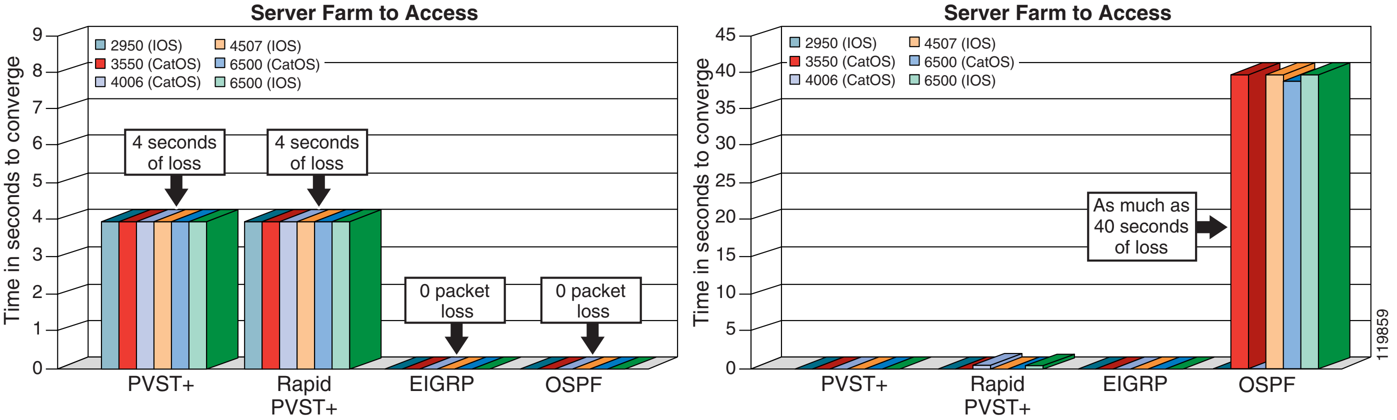
Figure 61 ***Distribution-to-Access Link Failure***



Additionally, because both EIGRP and OSPF load share over equal-cost paths, this provides a benefit similar to GLBP. Approximately 50 percent of the hosts are not affected by the convergence event because their traffic is not flowing over the link or through the failed node.

Using a routed access layer topology addresses some of the concerns discussed with the recommended topology in which the distribution switch is the L2/L3 boundary. For example, ARP processing for a large L2 domain by the distribution node is not a concern in this design, as shown in Figure 62. When a distribution is re-introduced to the environment, there is no disruption of service as compared to the four-second outage measured in the 40-node test bed for the L2/L3 distribution layer boundary topology. The previously large L2 domain and ARP processing is now distributed among the access layer switches supported by the distribution pair.

Figure 62 ***Primary Distribution Node Restoration***



However, a routed access layer topology is not a panacea. You must consider the additional IP address consumption for the point-to-point links between the access layer and distribution layer. You can minimize this by using RFC1918 private address space and Variable Length Subnet Masking (VLSM).

Additionally, this topology requires adherence to the best practice recommendation that no VLANs should span access layer switches. This is a benefit, however it makes this design less flexible than other configurations. If the design is modified to support VLANs spanning access layer switches the fast convergence benefit of the design can not be realized.

Finally, this topology has not been widely deployed and tested over time, while the design with the L2/L3 boundary at the distribution layer has.

If you want the best convergence available and you can ensure that no VLAN will need to span multiple access layer switches, then using a routed access layer topology is a viable design alternative.

Comparing Routing Protocols

To run a routing protocol between the access layer switches and the distribution layer switches, select the routing protocol to run and determine how to configure it.

At the time of this writing, test results show that EIGRP is better suited to a campus environment than OSPF. The ability of EIGRP to provide route filtering and summarization maps easily to the tiered hierarchical model, while the more rigid requirements of OSPF do not easily integrate to existing implementations and require more complex solutions.

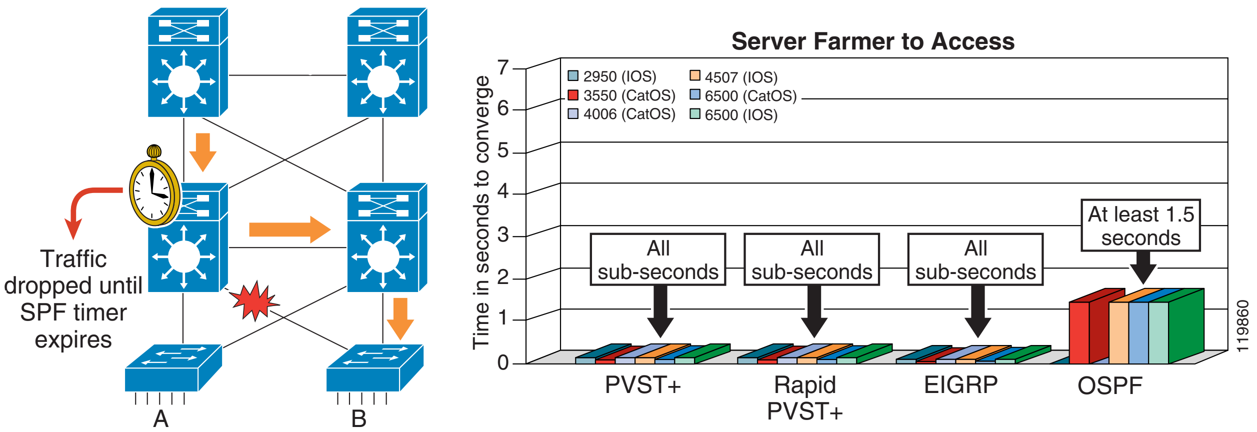
The following are additional considerations when comparing EIGRP and OSPF:

- Within the campus environment, EIGRP provides for faster convergence and greater flexibility.
- EIGRP provides for multiple levels of route summarization and route filtering that map to the multiple tiers of the campus.
- OSPF implements throttles on Link-State Advertisement (LSA) generation and Shortest Path First (SPF) calculations that limit convergence times.
- When routes are summarized and filtered, only the distribution peers in an EIGRP network need to calculate new routes in the event of link or node failure.

The throttles that OSPF places on LSA generation and SPF calculation can cause significant outages as OSPF converges around a node or link failure in the hierarchical network model.

There are two specific ways in which OSPF is limited. First, OSPF implements an SPF timer that can not currently be tuned below one second. When a link or node has failed, an OSPF peer cannot take action until this timer has expired. As a result, no better than 1.65 seconds of convergence time can be achieved in the event of an access layer to distribution layer uplink failure or primary distribution node failure (see Figure 63).

Figure 63 OSPF SPF Timer Affects Convergence Time



Return path traffic is dropped until the SPF timer has expired and normal reroute processing is completed. While PVST+, Rapid PVST+, and EIGRP all converged in less than one second (EIGRP in sub 200 ms), OSPF required at least 1.65 seconds to converge around this specific failure.

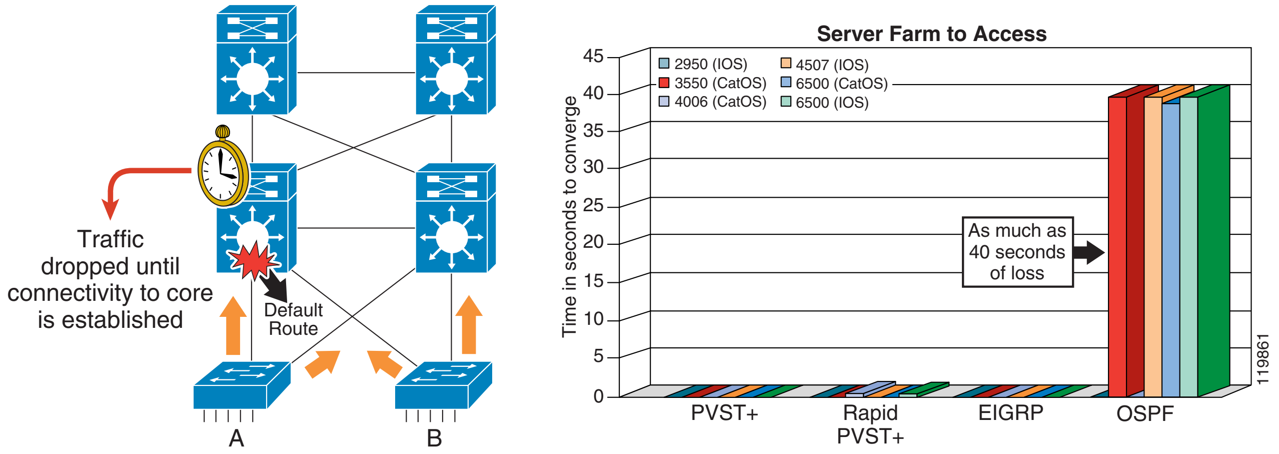
Additionally, totally stubby areas that are required to limit LSA propagation and unnecessary SPF calculation have an undesirable side effect when a distribution node is restored.

In a topology where HSRP and preemption are required for upstream traffic restoration, the HSRP process was tuned to wait until connectivity to the core had been established and the network had settled down before HSRP was allowed to take over and begin forwarding traffic upstream towards the core.

If EIGRP is utilized in the same topology, a default route is propagated from the core of the network and is therefore only distributed to the access layer switch when connectivity has been established and the network is ready to forward traffic from the access using the recovering distribution node.

With OSPF in the same topology, the default route is propagated to the totally stubby peer (the access layer switch in this case) when the neighbor relationship is established, regardless of the ability of the distribution node to forward traffic to the core. In the topology tested, the recovering distribution node had not fully established connectivity to the core, yet it was distributing a default route to the access layer switch. This behavior caused a considerable amount of traffic being dropped; more than 40 seconds in the tested topology. This occurred while the access layer switch was load sharing over the equal-cost paths on both uplinks to the distribution layer, and the recovering distribution node was unable to forward the traffic being sent its way (see Figure 64).

Figure 64 Convergence Time with OSPF Totally Stubby Areas



At the time of this writing, there is no workaround for this situation except using normal areas instead of totally stubby areas for the access layer switches. This is a less than optimal design because it lacks the protection from undesirable LSA propagation and subsequent CPU-intensive SPF calculations that totally stubby areas provide.

Using EIGRP in the Access Layer

When EIGRP is used as the routing protocol for a fully routed or routed access layer solution, take the following EIGRP tuning and best practice steps to achieve sub-200 ms convergence:

- Summarize towards the core from the distribution layer.

  As discussed earlier in this document, you should summarize at the distribution layer towards the core layer to stop EIGRP queries from propagating beyond the core of the network. When the distribution layer summarizes towards the core, queries are limited to one hop from the distribution switches, which optimizes EIGRP convergence.

- Control route propagation to access layer using distribute lists.

To conserve memory and optimize performance at the access layer, configure a distribute list outbound on the distribution switch and apply it to all interfaces facing the access layer. The distribute list allows only the default route (0.0.0.0) to be advertised to the access layer nodes.

- Configure all edge access layer switches to use EIGRP stub.

  By using the EIGRP stub option, you optimize the ability of EIGRP to converge in the access layer and also optimize its behavior from a route processing perspective. EIGRP stub nodes are not able to act as transit nodes and as such, they do not participate in EIGRP query processing. When the distribution node learns through the EIGRP hello packets that it is talking to a stub node, it does not flood queries to that node.

- Set hello and dead timers to 1 and 3, respectively.

  Tune EIGRP hello and dead timers to 1 and 3 respectively to protect against a soft failure in which the physical links remain active but hello/route processing has stopped.

The following configuration snippets demonstrate how EIGRP was configured to achieve sub-200ms convergence for link and node failure scenarios.

Access node EIGRP configuration:

```
interface GigabitEthernet1/1
ip hello-interval eigrp 100 1
ip hold-time eigrp 100 3

router eigrp 100
eigrp stub connected
```

Distribution node EIGRP configuration:

```
interface Port-channel1
description to Core Right
ip address 10.122.0.34 255.255.255.252
ip hello-interval eigrp 100 1
ip hold-time eigrp 100 3
ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
mls qos trust dscp
!

interface GigabitEthernet3/3
 description To 4500-Access (L3)
 ip address 10.120.0.198 255.255.255.252
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
 mls qos trust dscp
!
router eigrp 100
passive-interface default
no passive-interface Port-channel1
no passive-interface GigabitEthernet3/3
network 10.0.0.0
distribute-list Default out GigabitEthernet3/3
no auto-summary
!
!
ip Access-list standard Default
 permit 0.0.0.0
```

Using OSPF in the Access Layer

The following steps are recommended when using OSPF in the access layer:

- Control the number of routes and routers in each area.

- Configure each distribution block as a separate totally stubby OSPF area.
- Do not extend area 0 to the edge switch.
- Tune OSPF hello, dead-interval, and SPF timers to 1, 3, and 1, respectively.

OSPF in the access layer is similar to OSPF for WAN/Branch networks, except that you can tune for optimum convergence. With currently available hardware switching platforms, CPU resources are not as scarce in a campus environment as they might be in a WAN environment. Additionally, the media types common in the access layer are not susceptible to the same half up or rapid transitions from up to down to up (bouncing) as are those commonly found in the WAN. Because of these two differences, you can safely tune the OSPF timers (hello, dead-interval, and SPF) to their minimum allowable values of 1, 3, and 1 second, respectively.

With OSPF, you force summarization and limit the diameter of OSPF LSA propagation through the implementation of L2/L3 boundaries or Area Border Routers (ABRs). The access layer is not used as a transit area in a campus environment. As such, you can safely configure each access layer switch into its own unique totally stubby area. The distribution switches become ABRs with their core-facing interfaces in area 0 and the access layer interfaces in unique totally stubby areas for each access layer switch. In this configuration, LSAs are isolated to each access layer switch, so that a link flap for one access layer switch is not communicated beyond the distribution pairs. No additional access layer switches are involved in the convergence event.

As discussed previously, the OSPF SPF timer does not allow an OSPF environment to converge as quickly as EIGRP, PVST, or PVST+. You must consider this limitation before selecting OSPF as a routing protocol in campus environments. Additionally, you must consider the tradeoffs between totally stubby areas and regular areas for the access layer. Considerable outages can be experienced when distribution nodes are restored with totally stubby areas. However, the implications of LSA propagation and SPF calculation on the network as a whole are unknown in a campus topology where non-stubby areas are used for the access layer.

The following configuration snippets illustrate the OSPF configuration:

Access layer OSPF configuration:

```
interface GigabitEthernet1/1
ip ospf hello-interval 1
ip ospf dead-interval 3

router ospf 100
area 120 stub no-summary
timers spf 1 1
```

Distribution layer OSPF configuration:

```
mls ip cef load-sharing full
port-channel load-balance src-dst-port
!
interface GigabitEthernet2/1
description to 6k-Core-left CH#1
no ip address
mls qos trust dscp
channel-group 1 mode on
!
interface GigabitEthernet2/2
description to 6k-Core-left CH#1
no ip address
mls qos trust dscp
channel-group 1 mode on
!
```

```
interface Port-channel1
description to Channel to 6k-Core-left CH#1
ip address 10.122.0.34 255.255.255.252
ip ospf hello-interval 1
ip ospf dead-interval 3
mls qos trust dscp
!
interface GigabitEthernet3/3
 description to 4k Access
 ip address 10.120.0.198 255.255.255.252
 ip pim sparse-mode
 ip ospf hello-interval 1
 ip ospf dead-interval 3
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
!
router ospf 100
 log-adjacency-changes
 area 120 stub no-summary
 area 120 range 10.120.0.0 255.255.0.0
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
```

Summary

The design recommendations described in this design guide are best practices designed to achieve the best convergence possible. Although each recommendation should be implemented if possible, each network is unique, and issues such as cost, physical plant limitations, or application requirements may limit full implementation of these recommendations.

Following the hierarchical network model is essential for achieving high availability. In a hierarchical design, the capacity, features, and functionality of a specific device are optimized for its position in the network and the role that it plays. This promotes scalability and stability. If the foundation is not rock solid, the performance of applications that depend on network services such as IP telephony, IP video, and wireless communications will eventually suffer.

The proper configuration and tuning of foundational services is an essential component of a highly available campus network. From a design perspective, the following three alternatives exist within the hierarchical network model:

- Layer 2 Looped—Cisco does not recommend this option because of issues such as slow convergence, multiple convergence events, and the complexity and difficulty of implementation, maintenance, and operations.
- Layer 2 Loop-Free—This is the time-tested solution.
- Routed Access—This option is interesting from a convergence performance perspective, but is not yet widely deployed.

Your enterprise can take advantage of the design principles and implementation best practices described in this design guide to implement a network that will provide the optimal performance and flexibility as the business requirements of your network infrastructure evolve.