High Availability Campus Network Design—

Routed Access Layer using EIGRP or OSPF

1 Introduction

This document provides design guidance for implementing a routed (Layer 3 switched) access layer using EIGRP or OSPF as the campus routing protocol. It is an accompaniment to the hierarchical campus design guides, *Designing a Campus Network for High Availability* and *High Availability Campus Recovery Analysis*, and includes the following sections:

Note For design guides and more information on high availability campus design, see the following URL: http://www.cisco.com/en/US/netsol/ns815/networking_solutions_program_home.html.

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 related to implementing EIGRP or OSPF as a routing protocol in the access layer of the campus network.

Document Objectives

This document presents designs guidance and configuration examples for the campus network when it is desirable to implement a routed access layer using EIGRP or OSPF as the Internal Gateway Protocol (IGP).

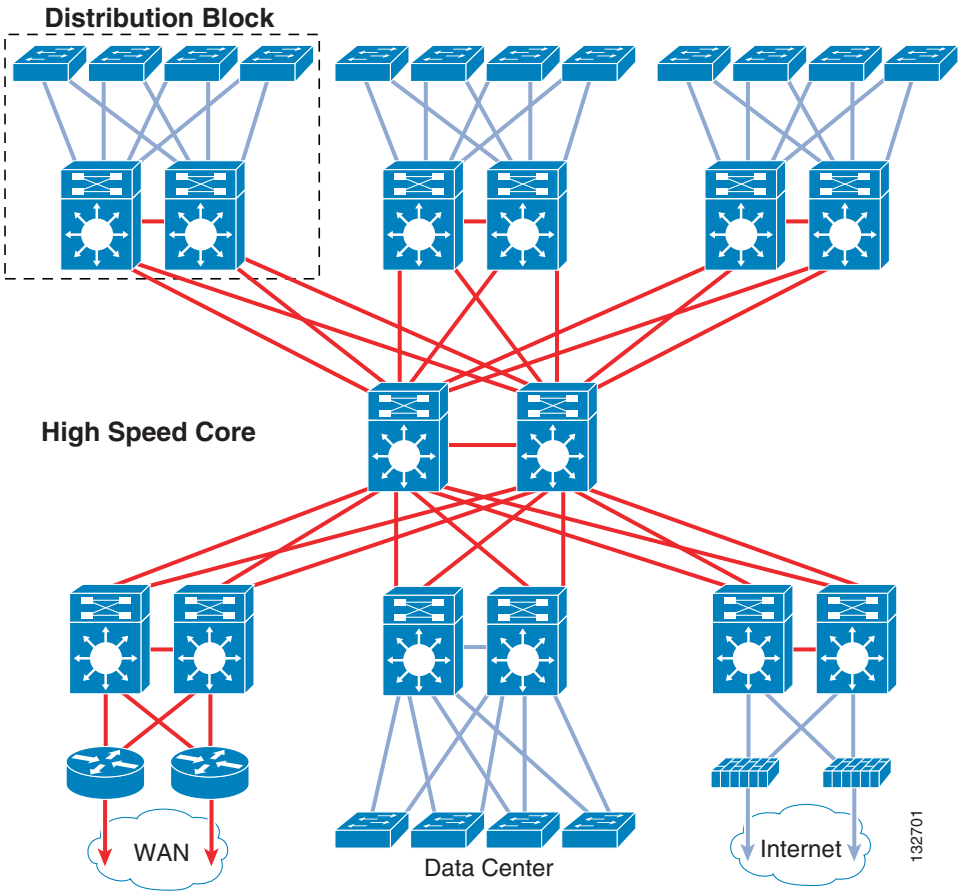
Overview

Both small and large enterprise campuses require a highly available and secure, intelligent network infrastructure to support business solutions such as voice, video, wireless, and mission-critical data applications. The use of hierarchical design principles provides the foundation for implementing campus networks that meet these requirements. The hierarchical design uses a building block approach leveraging a high-speed routed core network layer to which are attached multiple independent distribution blocks. The distribution blocks comprise two layers of switches: the actual distribution nodes that act as aggregators, and the wiring closet access switches.

The hierarchical design segregates the functions of the network into these separate building blocks to provide for availability, flexibility, scalability, and fault isolation. The distribution block provides for policy enforcement and access control, route aggregation, and the demarcation between the Layer 2 subnet (VLAN) and the rest of the Layer 3 routed network. The core layers of the network provide for high capacity transport between the attached distribution building blocks.

Figure 1 shows an example of a hierarchical campus network design using building blocks.

Figure 1 Hierarchical Campus Network Design using Building Blocks



Each building block within the network leverages appropriate switching technologies to best meet the architecture of the element. The core layer of the network uses Layer 3 switching (routing) to provide the necessary scalability, load sharing, fast convergence, and high speed capacity. Each distribution block uses a combination of Layer 2 and Layer 3 switching to provide for the appropriate balance of policy and access controls, availability, and flexibility in subnet allocation and VLAN usage.

For those campus designs requiring greater flexibility in subnet usage (for instance, situations in which VLANs must span multiple wiring closets), distribution block designs using Layer 2 switching in the access layer and Layer 3 switching at the distribution layer provides the best balance for the distribution block design.

For campus designs requiring simplified configuration, common end-to-end troubleshooting tools and the fastest convergence, a distribution block design using Layer 3 switching in the access layer (routed access) in combination with Layer 3 switching at the distribution layer provides the fastest restoration of voice and data traffic flows.

For those networks using a routed access (Layer 3 access switching) within their distribution blocks, Cisco recommends that a full-featured routing protocol such as EIGRP or OSPF be implemented as the campus Interior Gateway Protocol (IGP). Using EIGRP or OSPF end-to-end within the campus provides faster convergence, better fault tolerance, improved manageability, and better scalability than a design using static routing or RIP, or a design that leverages a combination of routing protocols (for example, RIP redistributed into OSPF).

2 Routing in the Access

This section includes the following topics:

Routing in the Campus

The hierarchical campus design has used a full mesh equal-cost path routing design leveraging Layer 3 switching in the core and between distribution layers of the network for many years. The current generation of Cisco switches can “route” or switch voice and data packets using Layer 3 and Layer 4 information with neither an increase in latency nor loss of capacity in comparison with a pure Layer 2 switch. Because in current hardware, Layer 2 switching and Layer 3 routing perform with equal speed, Cisco recommends a routed network core in all cases. Routed cores have numerous advantages, including the following:

- High availability
  - Deterministic convergence times for any link or node failure in an equal-cost path Layer 3 design of less than 200 msec
  - No potential for Layer 2 Spanning Tree loops
- Scalability and flexibility
  - Dynamic traffic load balancing with optimal path selection
  - Structured routing permits for use of modular design and ease of growth
- Simplified management and troubleshooting
  - Simplified routing design eases operational support

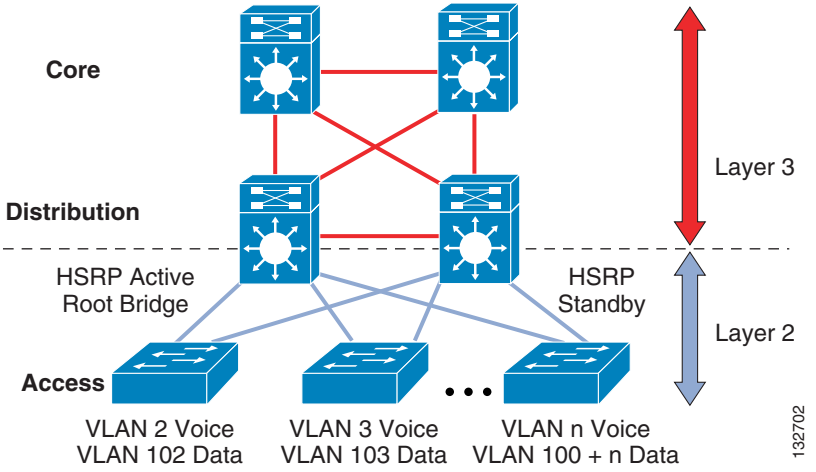
- Removal of the need to troubleshoot L2/L3 interactions in the core

The many advantages of Layer 3 routing in the campus derive from the inherent behavior of the routing protocols combined with the flexibility and performance of Layer 3 hardware switching. The increased scalability and resilience of the Layer 3 distribution/core design has proven itself in many customer networks over the years and continues to be the best practice recommendation for campus design.

Migrating the L2/L3 Boundary to the Access Layer

In the typical hierarchical campus design, distribution blocks use a combination of Layer 2, Layer 3, and Layer 4 protocols and services to provide for optimal convergence, scalability, security, and manageability. In the most common distribution block configurations, the access switch is configured as a Layer 2 switch that forwards traffic on high speed trunk ports to the distribution switches. The distribution switches are configured to support both Layer 2 switching on their downstream access switch trunks and Layer 3 switching on their upstream ports towards the core of the network, as shown in Figure 2.

Figure 2 Traditional Campus Design Layer 2 Access with Layer 3 Distribution

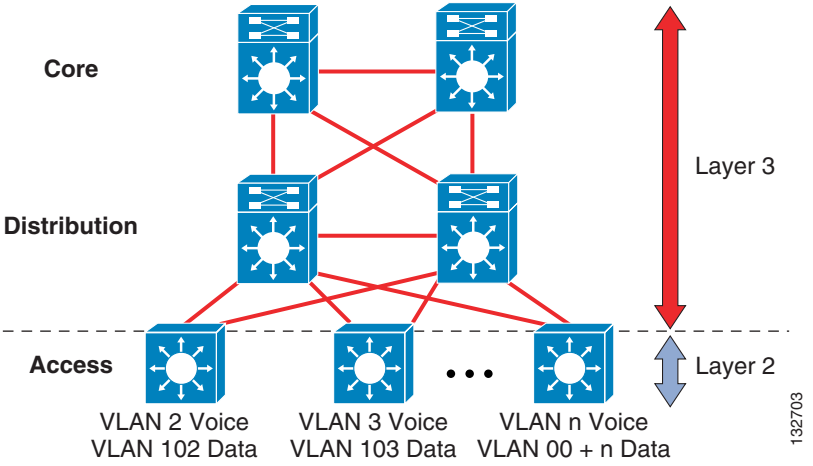


The function of the distribution switch in this design is to provide boundary functions between the bridged Layer 2 portion of the campus and the routed Layer 3 portion, including support for the default gateway, Layer 3 policy control, and all the multicast services required.

Note Although access switches forward data and voice packets as Layer 2 switches, in the Cisco campus design they leverage advanced Layer 3 and 4 features supporting enhanced QoS and edge security services.

An alternative configuration to the traditional distribution block model illustrated above is one in which the access switch acts as a full Layer 3 routing node (providing both Layer 2 and Layer 3 switching), and the access-to-distribution Layer 2 uplink trunks are replaced with Layer 3 point-to-point routed links. This alternative configuration, in which the Layer 2/3 demarcation is moved from the distribution switch to the access switch (as shown in Figure 3) appears to be a major change to the design, but is actually simply an extension of the current best practice design.

Figure 3 Routed Access Campus Design—Layer 3 Access with Layer 3 Distribution



In both the traditional Layer 2 and the Layer 3 routed access design, each access switch is configured with unique voice and data VLANs. In the Layer 3 design, the default gateway and root bridge for these VLANs is simply moved from the distribution switch to the access switch. Addressing for all end stations and for the default gateway remain the same. VLAN and specific port configuration remains unchanged on the access switch. Router interface configuration, access lists, “ip helper”, and any other configuration for each VLAN remain identical, but are now configured on the VLAN Switched Virtual Interface (SVI) defined on the access switch, instead of on the distribution switches. There are several notable configuration changes associated with the move of the Layer 3 interface down to the access switch. It is no longer necessary to configure an HSRP or GLBP virtual gateway address as the “router” interfaces for all the VLANs are now local. Similarly with a single multicast router, for each VLAN it is not necessary to perform any of the traditional multicast tuning such as tuning PIM query intervals or to ensure that the designated router is synchronized with the active HSRP gateway.

Note For details on the configuration of the Layer 3 access, see Campus Routing Design, page 9, Implementing Layer 3 Access using EIGRP, page 19, and Implementing Layer 3 Access using OSPF, page 27.

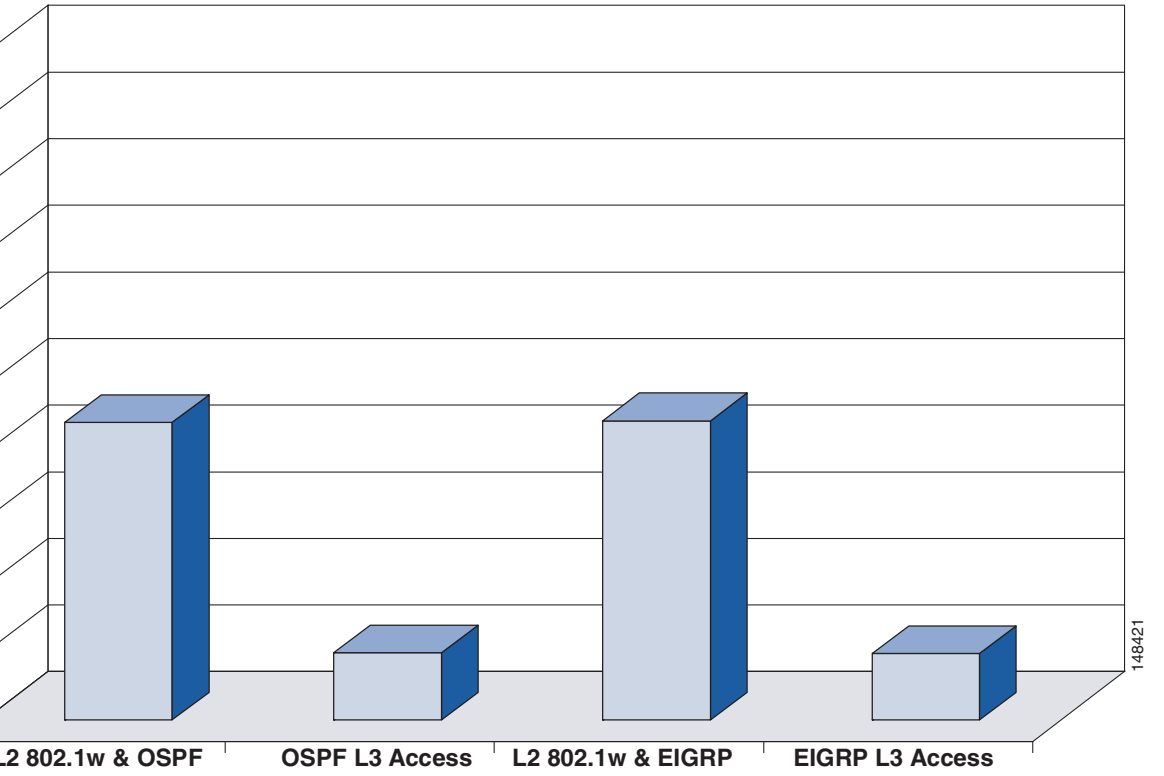
Routed Access Convergence

The many potential advantages of using a Layer 3 access design include the following:

- Improved convergence
- Simplified multicast configuration
- Dynamic traffic load balancing
- Single control plane
- Single set of troubleshooting tools (for example, ping and traceroute)

Of these, perhaps the most significant is the improvement in network convergence times possible when using a routed access design configured with EIGRP or OSPF as the routing protocol. Comparing the convergence times for an optimal Layer 2 access design (either with a spanning tree loop or without a loop) against that of the Layer 3 access design, you can obtain a four-fold improvement in convergence times, from 800–900msec for the Layer 2 design to less than 200 msec for the Layer 3 access. (See Figure 4.)

Figure 4 Comparison of Layer 2 and Layer 3 Convergence



Although the sub-second recovery times for the Layer 2 access designs are well within the bounds of tolerance for most enterprise networks, the ability to reduce convergence times to a sub-200 msec range is a significant advantage of the Layer 3 routed access design. To achieve the convergence times in the Layer 2 designs shown above, you must use the correct hierarchical design and tune HSRP/GLBP timers in combination with an optimal L2 spanning tree design. This differs from the Layer 3 campus, where it is necessary to use only the correct hierarchical routing design to achieve sub-200 msec convergence. The routed access design provides for a simplified high availability configuration. The following section discusses the specific implementation required to meet these convergence times for the EIGRP and OSPF routed access design.

Note For additional information on the convergence times shown in Figure 4, see the *High Availability Campus Recovery Analysis* design guide, located at the following URL: http://www.cisco.com/en/US/netsol/ns815/networking_solutions_program_home.html.

3 Campus Routing Design

This section includes the following topics:

Hierarchical Design

When implementing a routed access campus, it is important to understand both how the campus routing design fits into the overall network routing hierarchy, and how to best configure the campus switches to achieve the following:

- Rapid convergence because of link and/or switch failures
- Deterministic traffic recovery
- Scalable and manageable routing hierarchy

Adding an additional tier of routers into the hierarchical design does not change any of the fundamental rules of routing design. The IP addressing allocation should map onto a tiered route summarization scheme. The summarization scheme should map onto the logical building blocks of the network and provide isolation for local route convergence events (link and/or node failures within a building block should not result in routing updates being propagated to other portions of the network).

The traditional hierarchical campus design using Layer 2 access switching follows all of these rules. The distribution building block provides route summarization and fault isolation for access node and link failures and provides a summarization point for access routes up into the core of the network. Extending Layer 3 switching to the access does not require any change in this basic routing design. The distribution switches still provide a summarization point and still provide the fault domain boundary for local failure events.

Extending routing to the access layer requires only the logical structure of the distribution block itself be modified, and to do this you can use proven design principles established in the EIGRP or OSPF branch WAN environment. The routing architecture of the branch WAN has the same topology as the distribution block: redundant aggregation routers attached to edge access routers via point-to-point Layer 3 links. In both cases, the edge router provides access to and from the locally-connected subnets, but is never intended to act as a transit path for any other network traffic. The branch WAN uses a combination of stub routing, route filtering, and aggregation route summarization to meet the design requirements. The same basic configuration is used to optimize the campus distribution block.

The basic topology of the routed campus is similar to but not exactly the same as the WAN environment. Keep in mind the following differences between the two environments when optimizing the campus routing design:

- Fewer bandwidth limitations in the campus allow for more aggressive tuning of control plane traffic (for example, hello packet intervals)
- The campus typically has lower neighbor counts than in the WAN and thus has a reduced control plane load
- Direct fiber interconnects simplify neighbor failure detection
- Lower cost redundancy in the campus allow for use of the optimal redundant design
- Hardware L3 switching ensures dedicated CPU resources for control plane processing

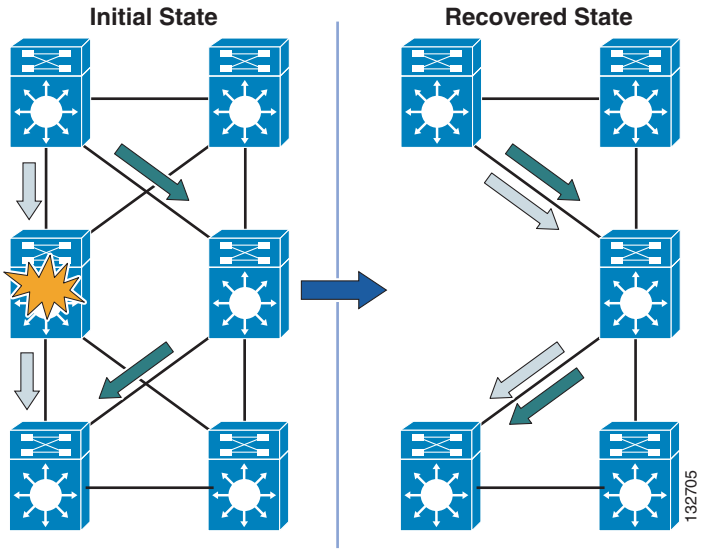
Within the routed access campus distribution block, the best properties of a redundant physical design are leveraged in combination with a hierarchical routing design using stub routing, route filtering, and route summarization to ensure consistent routing protocol convergence behavior. Each of these design requirements is discussed in more detail below.

Redundant Links

The most reliable and fastest converging campus design uses a tiered design of redundant switches with redundant equal-cost links. A hierarchical campus using redundant links and equal-cost path routing provides for restoration of all voice and data traffic flows in less than 200 msec in the event of either a link or node failure without having to wait for a routing protocol convergence to occur for all failure conditions except one (see Route Convergence, page 12 for an explanation of this particular case).

Figure 5 shows an example of equal-cost path traffic recovery.

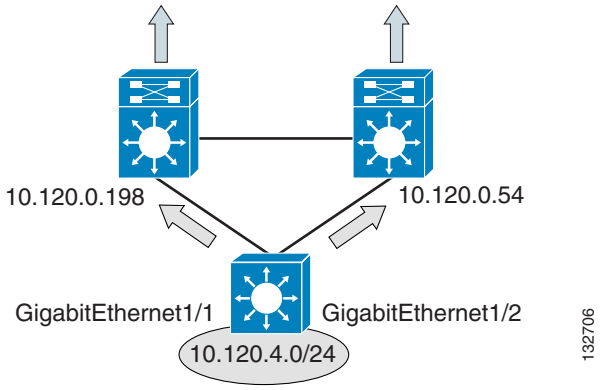
Figure 5 Equal-Cost Path Traffic Recovery



In the equal-cost path configuration, each switch has two routes and two associated hardware Cisco Express Forwarding (CEF) forwarding adjacency entries. Before a failure, traffic is being forwarded using both of these forwarding entries. On failure of an adjacent link or neighbor, the switch hardware and software immediately remove the forwarding entry associated with the lost neighbor. After the removal of the route and forwarding entries associated with the lost path, the switch still has a remaining valid route and associated CEF forwarding entry. Because the switch still has an active and valid route, it does not need to trigger or wait for a routing protocol convergence, and is immediately able to continue forwarding all traffic using the remaining CEF entry. The time taken to reroute all traffic flows in the network depends only on the time taken to detect the physical link failure and to then update the software and associated hardware forwarding entries.

Cisco recommends that Layer 3 routed campus designs use the equal-cost path design principle for the recovery of upstream traffic flows from the access layer. Each access switch needs to be configured with two equal-cost uplinks, as shown in Figure 6. This configuration both load shares all traffic between the two uplinks as well as provides for optimal convergence in the event of an uplink or distribution node failure.

Figure 6 Equal-Cost Uplinks from Layer 3 Access to Distribution



In the following example, the Layer 3 access switch has two equal-cost paths to the default route 0.0.0.0.

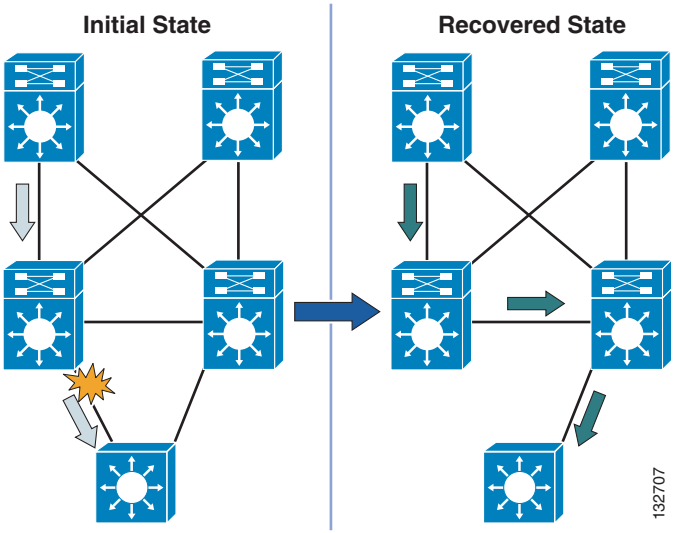
```
Layer3-Access#sh ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - OSPF, EX - OSPF external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is 10.120.0.198 to network 0.0.0.0

     10.0.0.0/8 is variably subnetted, 5 subnets, 3 masks
C       10.120.104.0/24 is directly connected, Vlan104
C       10.120.0.52/30 is directly connected, GigabitEthernet1/2
C       10.120.4.0/24 is directly connected, Vlan4
C       10.120.0.196/30 is directly connected, GigabitEthernet1/1
D*EX 0.0.0.0/0 [170/5888] via 10.120.0.198, 00:46:00, GigabitEthernet1/1
               [170/5888] via 10.120.0.54, 00:46:00, GigabitEthernet1/2
```

Route Convergence

The use of equal-cost path links within the core of the network and from the access switch to the distribution switch allows the network to recover from any single component failure without a routing convergence, except one. As in the case with the Layer 2 design, every switch in the network has redundant paths upstream and downstream except each individual distribution switch, which has a single downstream link to the access switch. In the event of the loss of the fiber connection between a

distribution switch and the access switch, the network must depend on the control plane protocol to restore traffic flows. In the case of the Layer 2 access, this is either a routing protocol convergence or a spanning tree convergence. In the case of the Layer 3 access design, this is a routing protocol convergence.

Figure 7 Traffic Convergence because of Distribution-to-Access Link Failure



To ensure the optimal recovery time for voice and data traffic flows in the campus, it is necessary to optimize the routing design to ensure a minimal and deterministic convergence time for this failure case.

The length of time it takes for EIGRP, OSPF, or any routing protocol to restore traffic flows within the campus is bounded by the following three main factors:

- The time required to detect the loss of a valid forwarding path
- The time required to determine a new best path (which is partially determined by the number of routers involved in determining the new path, or the number of routers that must be informed of the new path before the network can be considered converged)
- The time required to update software and associated CEF hardware forwarding tables with the new routing information

In the cases where the switch has redundant equal-cost paths, all three of these events are performed locally within the switch and controlled by the internal interaction of software and hardware. In the case where there is no second equal-cost path, EIGRP or OSPF must determine a new route, and this process plays a large role in network convergence times.

In the case of EIGRP, the time is variable and primarily dependent on how many EIGRP queries the switch needs to generate and how long it takes for the response to each of those queries to return to calculate a feasible successor (path). The time required for each of these queries to be completed depends on how far they have to propagate in the network before a definite response can be returned. To minimize the time required to restore traffic flows, in the case where a full EIGRP routing convergence is required, it is necessary for the design to provide strict bounds on the number and range of the queries generated.

In the case of OSPF, the time required to flood and receive Link-State Advertisements (LSAs) in combination with the time to run the Djikstra Shortest Path First (SPF) computation to determine the Shortest Path Tree (SPT) provides a bound on the time required to restore traffic flows. Optimizing the network recovery involves tuning the design of the network to minimize the time and resources required to complete these two events.

Link Failure Detection Tuning

The recommended best practice for campus design uses point-to-point fiber connections for all links between switches. In addition to providing better electromagnetic and error protection, fewer distance limitations and higher capacity fiber links between switches provide for improved fault detection. In a point-to-point fiber campus design using GigE and 10GigE fiber, remote node and link loss detection is normally accomplished using the remote fault detection mechanism implemented as a part of the 802.3z and 802.3ae link negotiation protocols. In the event of physical link failure, local or remote transceiver failure, or remote node failure, the remote fault detection mechanism triggers a link down condition that then triggers software and hardware routing and forwarding table recovery. The rapid convergence in the Layer 3 campus design is largely because of the efficiency and speed of this fault detection mechanism.

Note See IEEE standards 802.3ae and 802.3z for details on the remote fault operation for 10GigE and GigE respectively.

Link Debounce and Carrier-Delay

When tuning the campus for optimal convergence, it is important to review the status of the link debounce and carrier delay configuration. By default, GigE and 10GigE interfaces operate with a 10 msec debounce timer which provides for optimal link failure detection. The default debounce timer for 10/100 fiber and all copper link media is longer than that for GigE fiber, and is one reason for the recommendation of a high speed fiber deployment for switch-to-switch links in a routed campus design. It is good practice to review the status of this configuration on all switch-to-switch links to ensure the desired operation.

```
DistributionSwitch1#show interfaces tenGigabitEthernet 4/2 debounce
```

```
Port    Debounce time    Value(ms)
Te4/2    disable
```

The default and recommended configuration for debounce timer is “disabled”, which results in the minimum time between link failure and notification of the upper layer protocols.

Note For more information on the configuration and timer settings of the link debounce timer, see the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SXF/native/configuration/guide/intrface.html.

Similarly, it is advisable to ensure that the carrier-delay behavior is configured to a value of zero (0) to ensure no additional delay in the notification of link down. In the current Cisco IOS levels, the default behavior for Catalyst switches is to use a default value of 0 msec on all Ethernet interfaces for the carrier-delay time to ensure fast link detection. It is still recommended as best practice to hard code the carrier-delay value on critical interfaces with a value of 0 msec to ensure the desired behavior.

```
interface GigabitEthernet1/1
 description Uplink to Distribution 1
 dampening
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 0
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
```

Confirmation of the status of carrier-delay can be seen by looking at the status of the interface.

```
GigabitEthernet1/1 is up, line protocol is up (connected)
  . . .

  Encapsulation ARPA, loopback not set
  Keepalive set (10 sec)
  Carrier delay is 0 msec
  Full-duplex, 1000Mb/s, media type is SX
input flow-control is off, output flow-control is off

. . .
```

Hello/Hold and Dead Timer Tuning

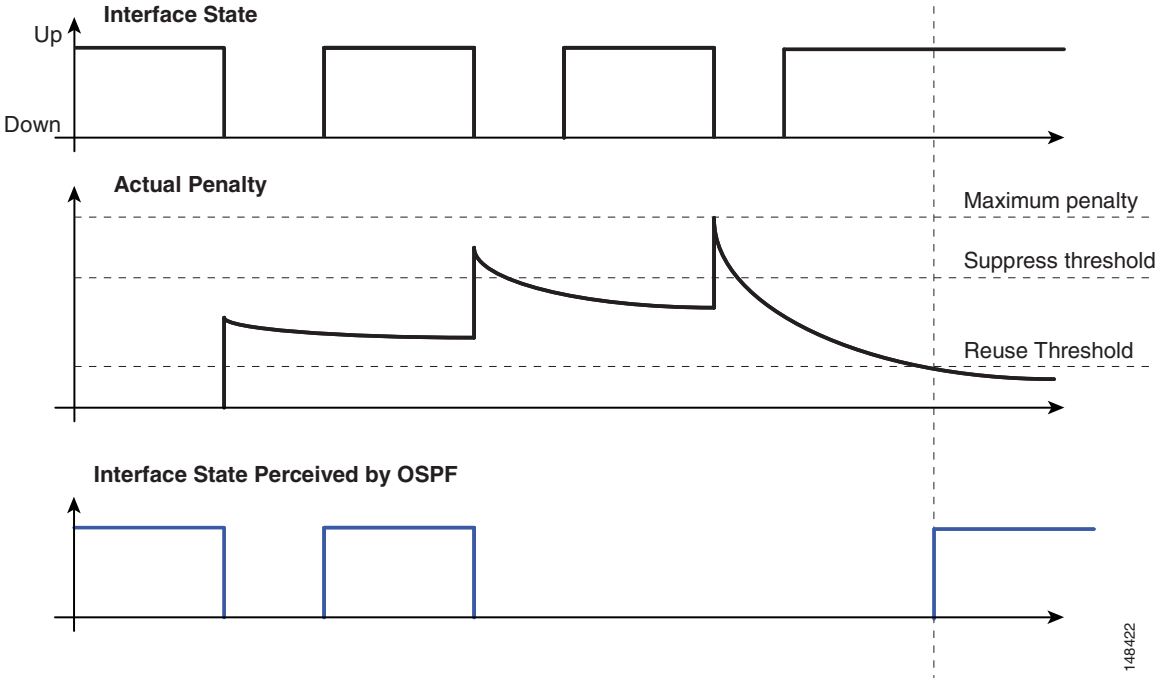
Although recovery from link failures in the campus depends primarily on 802.3z and 802.3ae remote fault detection, Cisco still recommends that the EIGRP hold and dead or OSPF hello and dead timers be reduced in the campus. The loss of hellos and the expiration of the dead timer provide a back-up to the L1/2 remote fault detection mechanisms. Tuning the EIGRP hold and dead or the OSPF hello and hold timers provides for a faster routing convergence in the rare event that L1/2 remote fault detection fails to operate.

Note See the EIGRP and OSPF design sections below for detailed guidance on timer tuning.

IP Event Dampening

When tightly tuning the interface failure detection mechanisms, it is considered a best practice to configure IP event dampening on any routed interfaces. IP event dampening provides a mechanism to control the rate at which interface state changes are propagated to the routing protocols in the event of a flapping link condition. It operates in a similar fashion to other dampening mechanisms, providing a penalty and penalty decay mechanism on link state transitions. In the event of a rapid series of link status changes, the penalty value for an interface increases until it exceeds a threshold, at which time no additional interface state changes are propagated to the routing protocols until the penalty value associated with the interface is below the reuse threshold. (See Figure 8.)

Figure 8 IP Event Dampening



IP event dampening operates with default values for the suppress, reuse, and maximum penalty values. It should be configured on every routed interface on all campus switches.

```
interface GigabitEthernet1/1
 description Uplink to Distribution 1
 dampening
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 0
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
```

Confirmation of the status of event dampening can be seen by looking at the status of the interface.

```
GigabitEthernet1/1 Uplink to Distribution 1
  Flaps Penalty    Supp ReuseTm   HalfL  ReuseV   SuppV  MaxSTm    MaxP Restart
      0       0   FALSE       0       5    1000    2000      20   16000       0
```

Note For more information on IP event dampening, see the following URL: http://www.cisco.com/en/US/docs/ios/12_2s/feature/guide/fsipevdp.html.

4 Implementing Layer 3 Access using EIGRP

This section includes the following topics:

- EIGRP Stub
- Distribution Summarization
- Route Filters
- Hello and Hold Timer Tuning

As discussed above, the length of time it takes for EIGRP or any routing protocol to restore traffic flows within the campus is bounded by the following three main factors:

- The time required to detect the loss of a valid forwarding path
- The time required to determine a new best path
- The time required to update software and associated hardware forwarding tables

In the cases where the switch has redundant equal-cost paths, all three of these events are performed locally within the switch and controlled by the internal interaction of software and hardware. In the case where there is no second equal-cost path nor a feasible successor for EIGRP to use, the time required to determine the new best path is variable and primarily dependent on EIGRP query and reply propagation across the network. To minimize the time required to restore traffic in the case where a full EIGRP routing convergence is required, it is necessary to provide strict bounds on the number and range of the queries generated.

Note For more details on the EIGRP feasible successor and the query process, see the following URL: http://www.cisco.com/en/US/tech/tk365/technologies_white_paper09186a0080094cb7.shtml .

Although EIGRP provides a number of ways to control query propagation, the two main methods are route summarization and the EIGRP stub feature. In the routed access hierarchical campus design, it is necessary to use both of these mechanisms.

EIGRP Stub

As noted previously, the design of the Layer 3 access campus is very similar to a branch WAN. The access switch provides the same routing functionality as the branch router, and the distribution switch provides the same routing functions as the WAN aggregation router. In the branch WAN, the EIGRP stub feature is configured on all of the branch routers to prevent the aggregation router from sending queries to the edge access routers. In the campus, configuring EIGRP stub on the Layer 3 access switches also prevents the distribution switch from generating downstream queries.

Access Switch EIGRP Routing Process Stub Configuration

```
router eigrp 100
 passive-interface default
 no passive-interface GigabitEthernet1/1
 no passive-interface GigabitEthernet1/2
 network 10.0.0.0
 no auto-summary
 eigrp router-id 10.120.4.1
 eigrp stub connected
```

By configuring the EIGRP process to run in "stub connected" state, the access switch advertises all connected subnets matching the network 10.0.0.0 0.255.255.255 range. It also advertises to its neighbor routers that it is a stub or non-transit router, and thus should never be sent queries to learn of a path to any subnet other than the advertised connected routes. With the design in Figure 9, the impact on the distribution switch is to limit the number of queries generated to "3" or less for any link failure.

Figure 9 EIGRP Stub Limits the Number of Queries Generated to "3"

To confirm that the distribution switch is not sending queries to the access switches, examine the EIGRP neighbor information for each access switch and look for the flag indicating queries being suppressed.

```
Distribution#sh ip eigrp neighbors detail gig 3/3
IP-EIGRP neighbors for process 100
H   Address                 Interface       Hold Uptime   SRTT   RTO  Q  Seq Type
                                           (sec)         (ms)       Cnt Num
10  10.120.0.53             Gi3/3              2 06:08:23    1   200  0  12
```

```
Version 12.2/1.2, Retrans: 1, Retries: 0
  Stub Peer Advertising ( CONNECTED REDISTRIBUTED ) Routes
Suppressing queries
```

Configuring the access switch as a "stub" router enforces hierarchical traffic patterns in the network. In the campus design, the access switch is intended to forward traffic only to and from the locally connected subnets. The size of the switch and the capacity of its uplinks are specified to meet the needs of the locally-connected devices. The access switch is never intended to be a transit or intermediary device for any data flows that are not to or from locally-connected devices. The hierarchical campus is designed to aggregate the lower speed access ports into higher speed distribution uplinks, and then to aggregate that traffic up into high speed core links. The network is designed to support redundant capacity within each of these aggregation layers of the network, but not to support the re-route of traffic through an access layer. Configuring each of the access switches as EIGRP stub routers ensures that the large aggregated volumes of traffic within the core are never forwarded through the lower bandwidth links in the access layer, and also ensures that no traffic is ever mistakenly routed through the access layer, bypassing any distribution layer policy or security controls.

Each access switch in the routed access design should be configured with the EIGRP stub feature to aid in ensuring consistent convergence of the campus by limiting the number of EIGRP queries required in the event of a failure, and to enforce engineered traffic flows to prevent the network from mistakenly forwarding transit traffic through the access layer.

Note For more information on the EIGRP stub feature, see the following URL: http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/eigrpstb.html.

Distribution Summarization

Configuring EIGRP stub on all of the access switches reduces the number of queries generated by a distribution switch in the event of a downlink failure, but it does not guarantee that the remaining queries are responded to quickly. In the event of a downlink failure, the distribution switch generates three queries; one sent to each of the core switches, and one sent to the peer distribution switch. The queries generated ask for information about the specific subnets lost when the access switch link failed. The peer distribution switch has a successor (valid route) to the subnets in question via its downlink to the access switch, and is able to return a response with the cost of reaching the destination via this path. The time to complete this event depends on the CPU load of the two distribution switches and the time required to transmit the query and the response over the connecting link. In the campus environment, the use of hardware-based CEF switching and GigE or greater links enables this query and response to be completed in less than a 100 msec.

This fast response from the peer distribution switch does not ensure a fast convergence time, however. EIGRP recovery is bounded by the longest query response time. The EIGRP process has to wait for replies from all queries to ensure that it calculates the optimal loop free path. Responses to the two

queries sent towards the core need to be received before EIGRP can complete the route recalculation. To ensure that the core switches generate an immediate response to the query, it is necessary to summarize the block of distribution routes into a single summary route advertised towards the core.

```
interface TenGigabitEthernet4/1
 description Distribution 10 GigE uplink to Core 1
 ip address 10.122.0.26 255.255.255.254
 ip pim sparse-mode
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 mls qos trust dscp
```

The summary-address statement is configured on the uplinks from each distribution switch to both core nodes. In the presence of any more specific component of the 10.120.0.0/16 address space, it causes EIGRP to generate a summarized route for the 10.120.0.0/16 network, and to advertise only that route upstream to the core switches.

```
Core-Switch-1#sh ip route 10.120.4.0
Routing entry for      10.120.0.0/16
  Known via "eigrp 100", distance 90, metric 768, type internal
  Redistributing via eigrp 100
  Last update from 10.122.0.34 on TenGigabitEthernet3/2, 09:53:57 ago
  Routing Descriptor Blocks:
  * 10.122.0.26, from 10.122.0.26, 09:53:57 ago, via TenGigabitEthernet3/1
      Route metric is 768, traffic share count is 1
      Total delay is 20 microseconds, minimum bandwidth is 10000000 Kbit
      Reliability 255/255, minimum MTU 1500 bytes
      Loading 1/255, Hops 1
    10.122.0.34, from 10.122.0.34, 09:53:57 ago, via TenGigabitEthernet3/2
      Route metric is 768, traffic share count is 1
      Total delay is 20 microseconds, minimum bandwidth is 10000000 Kbit
      Reliability 255/255, minimum MTU 1500 bytes
      Loading 1/255, Hops 1
```

With the upstream route summarization in place, whenever the distribution switch generates a query for a component subnet of the summarized route, the core switches reply that they do not have a valid path (cost = infinity) to the subnet query. The core switches are able to respond within less than 100 msec if they do not have to query other routers before replying back to the subnet in question.

Figure 10 shows an example of summarization toward the core.

Figure 10 Summarization toward the Core Bounds EIGRP Queries for Distribution Block Routes

Using a combination of stub routing and summarizing the distribution block routes upstream to the core both limits the number of queries generated and bounds those that are generated to a single hop in all directions. Keeping the query period bounded to less than 100 msec keeps the network convergence similarly bounded under 200 msec for access uplink failures. Access downlink failures are the worst case scenario because there are equal-cost paths for other distribution or core failures that provide immediate convergence.

Note To ensure a predictable EIGRP convergence time, you also need to protect the network against anomalous events such as worms, distributed denial-of-service (DDoS) attacks, and Spanning Tree loops that may cause high CPU on the switches. The use of Cisco Catalyst security features such as hardware rate limiters, QoS, CEF, and CISFs in conjunction with network security best practices as described in the SAFE design guides is a necessary component in a high availability campus design. For more information on SAFE, see the following URL: http://www.cisco.com/en/US/netsol/ns744/networking_solutions_program_home.html.

Route Filters

The discussion on EIGRP stub above noted that in the structured campus model, the flow of traffic follows the hierarchical design. Traffic flows pass from access through the distribution to the core and should never pass through the access layer unless they are destined to a locally attached device. Configuring EIGRP stub on all the access switches aids in enforcing this desired traffic pattern by preventing the access switch from advertising transit routes. As a complement to the use of EIGRP stub, Cisco recommends applying a distribute-list to all the distribution downlinks to filter the routes received by the access switches. The combination of "stub routing" and route filtering ensures that the routing protocol behavior and routing table contents of the access switches are consistent with their role, which is to forward traffic to and from the locally connected subnets only.

Cisco recommends that a default or "quad zero" route (0.0.0.0 mask 0.0.0.0) be the only route advertised to the access switches.

```
router eigrp 100
 network 10.120.0.0 0.0.255.255
 network 10.122.0.0 0.0.0.255
 . . .
 distribute-list Default out GigabitEthernet3/3
 . . .
 eigrp router-id 10.120.200.1

!
ip Access-list standard Default
 permit 0.0.0.0
```

Note No mask is required in the configuration of this access list because the assumed mask, 0.0.0.0, permits only the default route in the routing updates. It is also possible to use a prefix list to filter out all the routes other than the default route in place of an access list.

In addition to enforcing consistency with the desire for hierarchical traffic flows, the use of route filters also provides for easier operational management. With the route filters in place, the routing table for the access switch contains only the essential forwarding information. Reviewing the status and/or troubleshooting the campus network is much simpler when the routing tables contain only essential information.

```
Layer3-Access#sh ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route
```

```
Gateway of last resort is 10.120.0.198 to network 0.0.0.0

     10.0.0.0/8 is variably subnetted, 5 subnets, 3 masks
C       10.120.104.0/24 is directly connected, Vlan104
C       10.120.0.52/30 is directly connected, GigabitEthernet1/2
C       10.120.4.0/24 is directly connected, Vlan4
C       10.120.0.196/30 is directly connected, GigabitEthernet1/1
D*EX 0.0.0.0/0 [170/5888] via 10.120.0.198, 00:46:00, GigabitEthernet1/1
               [170/5888] via 10.120.0.54, 00:46:00, GigabitEthernet1/2
```

If the network does not contain a default route, it may be acceptable to use an appropriate full network summary route in its place; that is, 10.0.0.0/8, or a small subset of summary routes that summarize all possible destination addresses within the network.

Note As a design tip, unless the overall network design dictates otherwise, it is highly recommended that the network be configured with a default route (0.0.0.0) that is sourced into the core of the network, either by a group of highly available sink holes routers or by Internet DMZ routers.

The sink-hole router design is most often used by networks that implement an Internet proxy architecture that requires all traffic outbound and inbound to Internet sites be forwarded via an approved proxy. When using a proxy-based network design, Cisco recommends that the sink-hole routers also be configured to use Netflow, access lists, and/or "ip accounting" to track packets routed to the sink hole. The sink-hole routers should also be monitored by the network operations team looking for unusually high volumes of packets being forwarded to the sink hole. In normal day-to-day operations, few devices should ever generate a packet without a valid and routable destination address. End stations generating a high volume of packets to a range of un-allocated addresses are a typical symptom of a network worm-scanning behavior. By monitoring any increase in scanned random addresses in the sink-hole routers, it is possible to quickly track and identify infected end systems and take action to protect the remainder of the network.

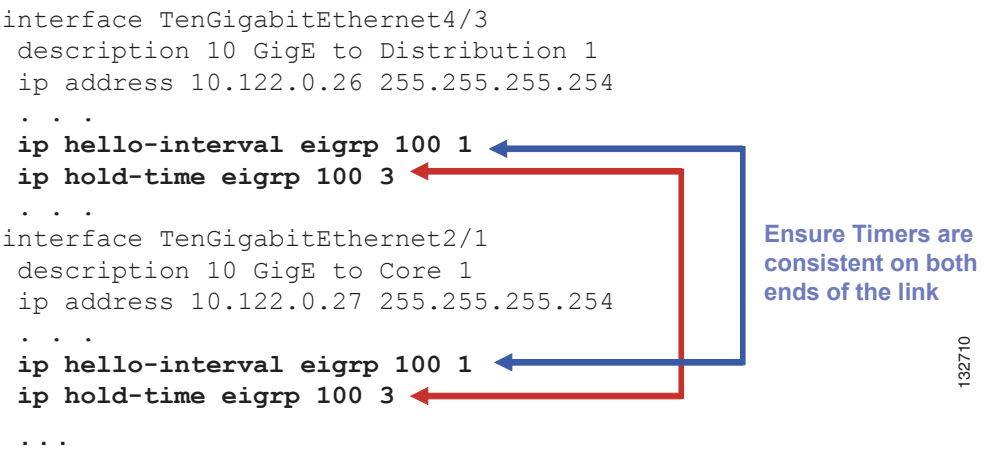
In the cases where the network uses a DMZ sourced default route to directly forward traffic to the Internet, Cisco recommends that an alternative approach be used to monitor for the presence of scanning traffic. This can be accomplished via Netflow tools such as Arbor Networks Peakflow, monitoring of connection rate on the Internet Firewall, or IPS systems.

Hello and Hold Timer Tuning

As discussed above, the recommended best practice for campus design uses point-to-point fiber connections for all links between switches. Link failure detection via 802.3z and 802.3ae remote fault detection mechanism provide for recovery from most campus switch component failures.

Cisco still recommends in the Layer 3 campus design that the EIGRP hello and dead timers be reduced to 1 and 3 seconds, respectively (see Figure 11). The loss of hellos and the expiration of the dead timer does provide a backup to the L1/2 remote fault detection mechanisms. Reducing the EIGRP hello and hold timers from defaults of 5 and 15 seconds provides for a faster routing convergence in the rare event that L1/2 remote fault detection fails to operate, and hold timer expiration is required to trigger a network convergence because of a neighbor failure.

Figure 11 Reducing EIGRP Hello and Dead Timers

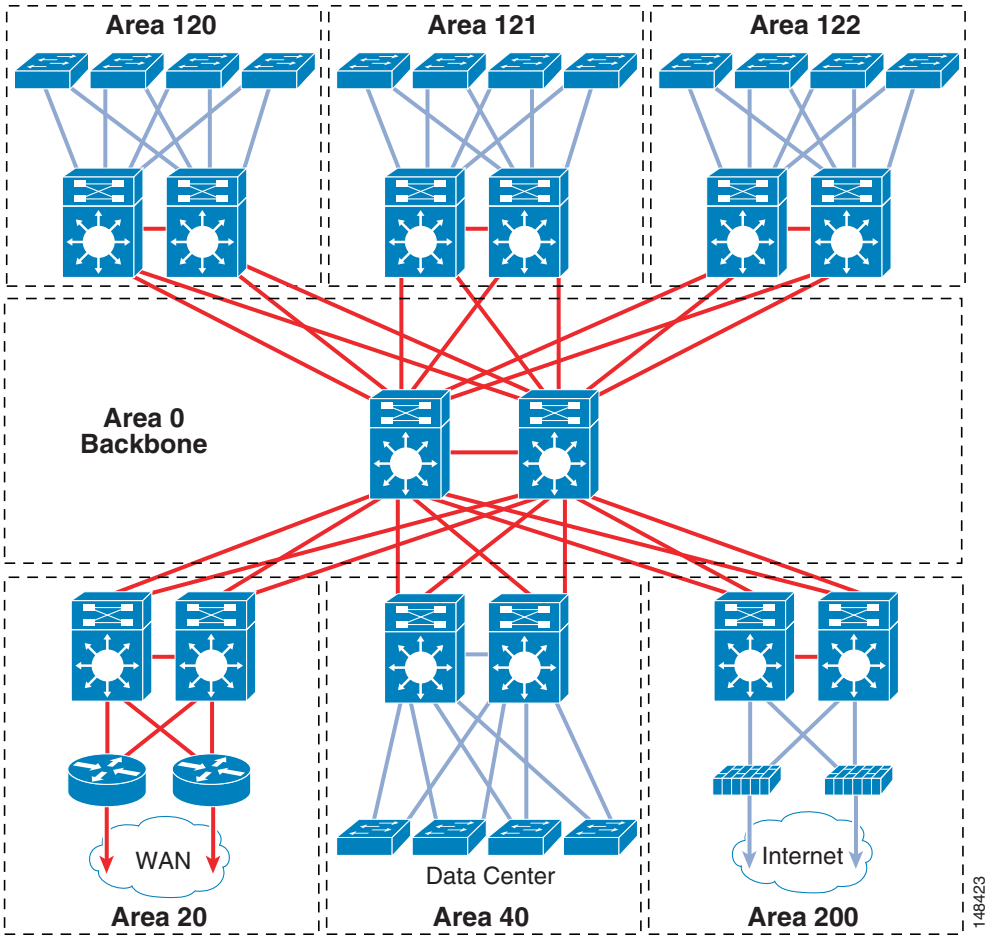


5 Implementing Layer 3 Access using OSPF

OSPF Area Design

Although ensuring the maximum availability for a routed OSPF campus design requires the consideration of many factors, the primary factor is how to implement a scalable area design. The convergence, stability, and manageability of a routed campus and the network as a whole depends on a solid routing design. OSPF implements a two-tier hierarchical routing model that uses a core or backbone tier known as area zero (0). Attached to that backbone via area border routers (ABRs) are a number of secondary tier areas. The hierarchical design of OSPF areas is well-suited to the hierarchical campus design. The campus core provides the backbone function supported by OSPF area 0, and the distribution building blocks with redundant distribution switches can be configured to be independent areas with the distribution switches acting as the ABRs, as shown in Figure 12.

Figure 12 Campus OSPF Area Design



In many OSPF area designs, the question of the optimal size of the area (number of nodes and links) is often a primary consideration in specifying the OSPF area boundaries, but does not play as key a role in the campus as it can in a general design. The desire to map the area boundaries to the hierarchical physical design, enforce hierarchical traffic patterns, minimize the convergence times, and maximize the stability of the network are more significant factors in designing the OSPF campus than is optimizing the number of nodes in the area or the number of areas in the network.

Mapping a unique OSPF area to each distribution block directly maps the basic building block of the OSPF routing design (the area) onto the basic building block of the campus network (the distribution block). The function of the distribution switch as a point of control for traffic to and from all access

segments is directly supported by the functions of the ABR to control routing information into and out of the area. The boundary for route convergence events provided by the ABR supports the desire to have the distribution block provide for fault containment, and also serves to aid in controlling the time required for routing convergence by restricting the scope of that routing convergence. Additionally, leveraging the properties of an OSPF stub area makes it relatively simple to enforce the rule that traffic not destined to an address within the distribution block is never forwarded into or through the area. As mentioned above in Implementing Layer 3 Access using EIGRP, page 19, the capacity of the access switches and their uplinks are specified to meet the needs of the locally-connected devices only. Configuring each distribution block as a unique area ensures that the large aggregated volumes of traffic within the core are never forwarded through the lower bandwidth links in the access layer, and also ensures that no traffic is ever mistakenly routed through the access layer, bypassing any distribution layer policy or security controls.

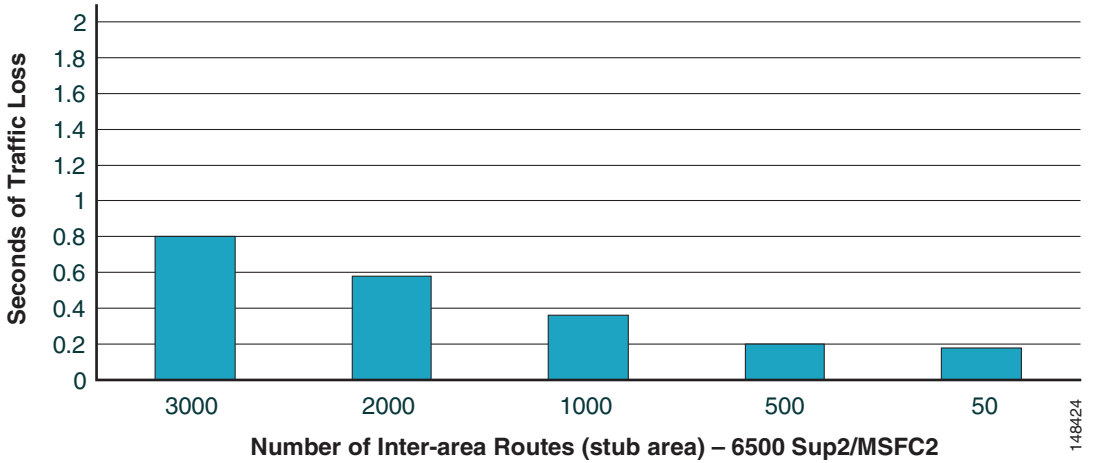
OSPF Stubby and Totally Stubby Distribution Areas

Within each distribution stub area, convergence of traffic is best optimized through a combination of equal-cost path design and the use of stub OSPF areas. Although there are many types of stub areas, Cisco recommends that "totally stubby" area configurations be used for the campus distribution blocks to minimize the size of the routing and forwarding tables in the access switches.

As discussed above, the convergence for traffic flows upstream from the access switches depends on equal-cost path recovery. As discussed above in the section on hierarchical design, one of the key mechanisms for ensuring very fast network recovery is to leverage the fast convergence behavior of equal cost path routing. When equal cost paths exist, the failure of one path requires only local hardware and software routing and forwarding updates to restore all traffic flows. However, the time to complete these routing and forwarding table updates is not constant, but varies depending on the specific hardware platform and more importantly on the number of routes or forwarding entries in the system.

The update of the CEF hardware FIB and adjacency entries is performed by the system software engine, and the entries in the tables are processed in a linear fashion. The greater the number of entries that need to be modified, the longer it takes for all entries to be modified. It is necessary to assume that the last entry updated impacts some traffic flows, and convergence time is calculated based on the time taken for the last entry to be updated. As shown in Figure 13, it can be seen that as the number of routes in the access switches is increased, the time taken to ensure all traffic flows have been restored increases.

Figure 13 Convergence Results



The time taken for convergence with 3000 inter-area routes in addition to the intra-area routes is still sub-second; however, to meet the design goals of sub-200 msec recovery, it is necessary to reduce the number of total routes in the distribution block access switches. Controlling the summarization of routes in the network as a whole aids in the reduction of the number of inter-area and external routes. The use of stub area configuration for the distribution block area prevents the propagation of external routes into the distribution block. However, Cisco recommends configuring the distribution block areas as totally stub areas, which also stops the propagation of all inter-area routes into the access switches. In this configuration, each of the distribution switch ABRs creates a default route that provides the forwarding path for the majority of traffic in the distribution block. As shown in Figure 14, the use of the **no-summary** command creates a totally stub area that contains only a single default inter-area route, and reduces the total number of routes in the hardware forwarding tables significantly.

Note The stub parameter in the area configuration command blocks "external" LSAs from entering the area through the ABR. The no-summary with the stub parameter blocks inter-area "summary" LSAs from entering the area. The ABRs also inject a default route (0.0.0.0) into the stub area to provide access to the routes not propagated into the area.

Figure 14 shows the default OSPF area configuration and associated route table impact.

Figure 14 Default OSPF Area Configuration and Associated Route Table Impact

```
router ospf 100
 router-id 10.120.250.6
 ispf
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
 network 10.120.0.0 0.0.255.255 area 120
```

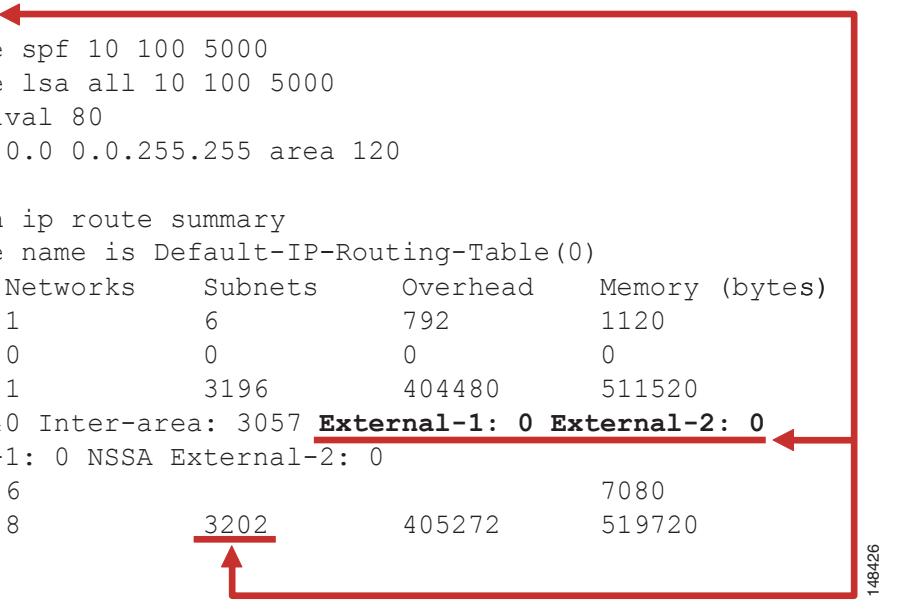
```
Access-Switch#sh ip route summary
IP routing table name is Default-IP-Routing-Table(0)
Route Source    Networks    Subnets     Overhead    Memory (bytes)
connected       1           6           776         1120
static          0           0           0           0
ospf 100        2           3626        459648      580480
  Intra-area: 70 Inter-area: 3055 External-1: 1 External-2: 502
  NSSA External-1: 0 NSSA External-2: 0
internal        7                                   8260
Total           10          3632        460424      589860
```

148425

Figure 15 shows the OSPF stub area configuration and associated route table impact.

Figure 15 OSPF Stub Area Configuration and Associated Route Table Impact

```
router ospf 100
 router-id 10.120.250.6
 ispf
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
 network 10.120.0.0 0.0.255.255 area 120

Access-Switch#sh ip route summary
IP routing table name is Default-IP-Routing-Table(0)
Route Source    Networks    Subnets     Overhead    Memory (bytes)
connected       1           6           792         1120
static          0           0           0           0
ospf 100        1           3196        404480      511520
  Intra-area: 140 Inter-area: 3057 External-1: 0 External-2: 0
  NSSA External-1: 0 NSSA External-2: 0
internal        6                                   7080
Total           8           3202        405272      519720
```

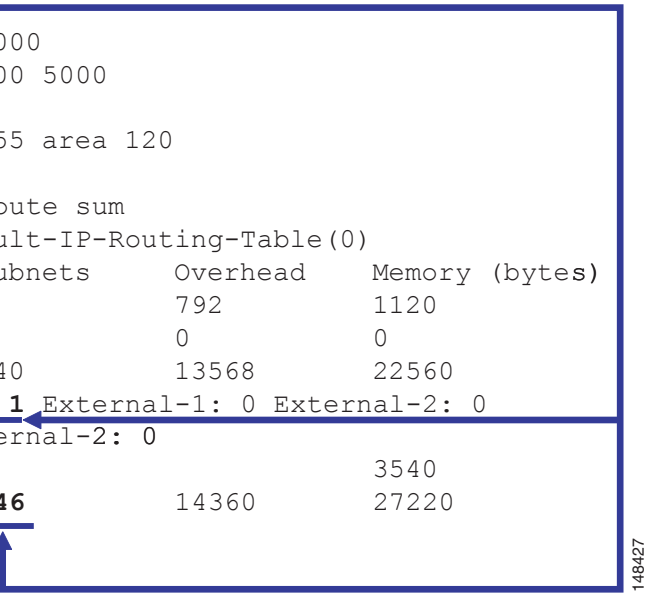


148426

Figure 16 shows the OSPF “totally stub” area configuration and associated route table impact.

Figure 16 OSPF "Totally Stub" Area Configuration and Associated Route Table Impact

```
router ospf 100
 router-id 10.120.250.6
 ispf
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
 network 10.120.0.0 0.0.255.255 area 120

Access-Totally-Stubby#sh ip route sum
IP routing table name is Default-IP-Routing-Table(0)
Route Source    Networks    Subnets     Overhead    Memory (bytes)
connected       1           6           792         1120
static          0           0           0           0
ospf 100        1           140         13568       22560
  Intra-area: 140 Inter-area: 1 External-1: 0 External-2: 0
  NSSA External-1: 0 NSSA External-2: 0
internal        3                                   3540
Total           5           146         14360       27220
```

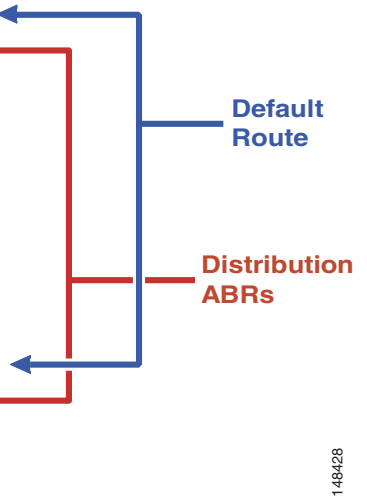


148427

Although the use of a stub, or better yet a totally stubby area, can have a positive impact on convergence times by reducing route table size, configuration of stub and in particular a totally stubby areas requires some attention. The stub area concept operates by creating an artificial default route sourced from each of the distribution ABRs, which is propagated as a type-3 network summary route into the stub area. This can be seen in Figure 16 as the single inter-area route in the "show ip route sum" output. This default route is created to represent the "rest" of the network to the stub area routers. It is used to build a forwarding path back to the distribution ABRs for all traffic external to the OSPF domain in the case of a stub configuration, and both the external and inter-area routes in the case of a totally stub area. (See Figure 17.)

Figure 17 Default Route and Distribution ABRs

```
Access-Switch#sh ip ospf data summary

OSPF Router with ID (10.120.250.6) (Process ID 100)

Summary Net Link States (Area 120)

Routing Bit Set on this LSA
LS age: 1122
Options: (No TOS-capability, DC, Upward)
LS Type: Summary Links(Network)
Link State ID: 0.0.0.0 (summary Network Number)
Advertising Router: 10.122.102.1
LS Seq Number: 8000051A
Checksum: 0x6FC4
Length: 28
Network Mask: /0
      TOS: 0  Metric: 1

Routing Bit Set on this LSA
LS age: 1120
Options: (No TOS-capability, DC, Upward)
LS Type: Summary Links(Network)
Link State ID: 0.0.0.0 (summary Network Number)
Advertising Router: 10.122.102.2
LS Seq Number: 80000001
Checksum: 0xAAA6
Length: 28
Network Mask: /0
      TOS: 0  Metric: 1
```




```
Access-Switch#sh ip route 0.0.0.0
Routing entry for 0.0.0.0/0, supernet
  Known via "ospf 100", distance 110, metric 11, candidate default path, type inter area
  Last update from 10.120.0.206 on GigabitEthernet2/1, 00:10:41 ago
  Routing Descriptor Blocks:
  * 10.120.0.204, from 10.122.102.1, 00:10:41 ago, via GigabitEthernet1/1
      Route metric is 11, traffic share count is 1
    10.120.0.206, from 10.122.102.2, 00:10:41 ago, via GigabitEthernet2/1
      Route metric is 11, traffic share count is 1
```

It is important to note that this default route is created on activation of any area "0" interface on the distribution switch, not because of the presence of any valid routing information from the rest of the network. This route is created with the assumption that the distribution ABR has connectivity for any valid destination in the network. In the normal case where all distribution switch interfaces configured

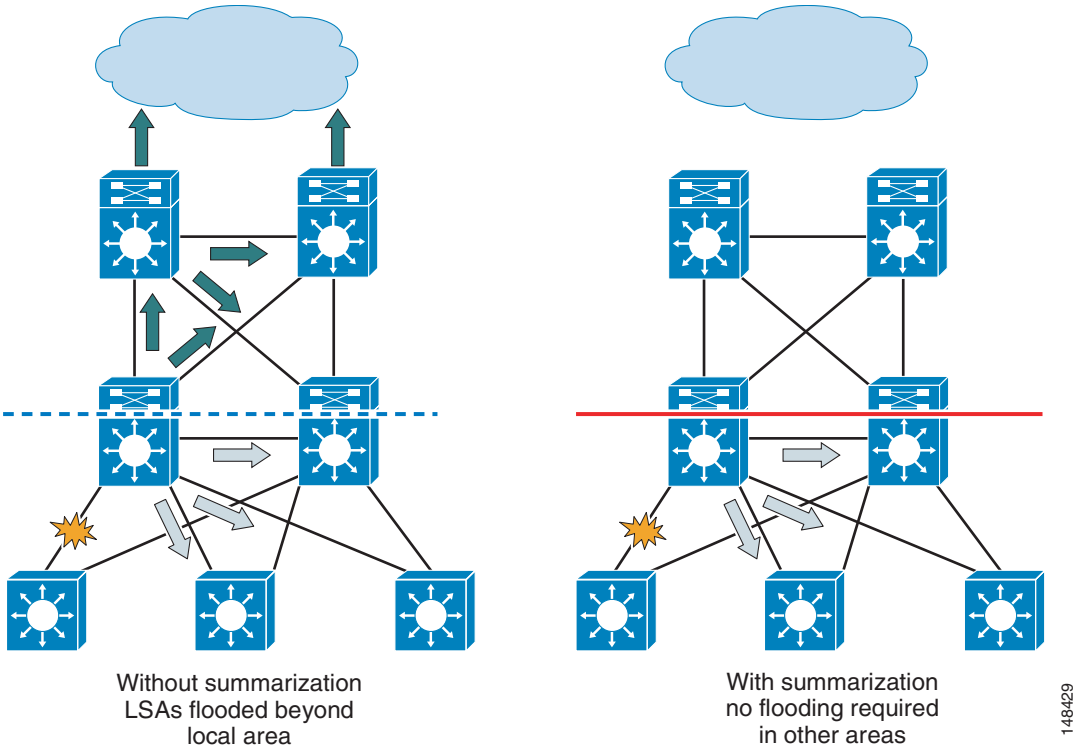
in area 0 are directly connected to the core switches, this design works very effectively. Traffic from all the access switches follows the default route and is forwarded to the distribution switches, which either forward back to the correct access switch or forward to the core switches. However, in the case where a loopback or any other interface not providing connectivity to the backbone (area 0) is incorrectly configured to reside in area "0", the distribution switch may incorrectly advertise a default route to the access switches, potentially creating a routing black hole.

Distribution ABR Route Summarization

Controlling the extent of topology changes and the number of routes advertised throughout the network are two key criteria of good routing design. As discussed above, reducing the number of active routes that need to be maintained in the forwarding tables can also help fine tune the time taken for the network to converge. Additionally, implementing a well-structured summarization scheme reduces the scope of network topology updates in the event of link or node failures. In a well-designed network, the scope of propagation of any routing update (addition or removal) is well-defined. The hierarchical campus design provides an excellent opportunity to minimize the scope of topology changes while simultaneously reducing the route count in the network. As discussed above, the hierarchical design of the campus maps physical design to logical routing design and to traffic flow. The network is designed for traffic to flow within each layer, for local traffic to remain within each distribution area, and for all other traffic to flow to the core and then to the correct destination area. This design philosophy provides an ideal environment in which to implement route controls such as stub area design as discussed above, and to complement it with route summarization.

Route summarization aids in route control by reducing the number of routes and associated topology table information (LSAs) in each network node. Without route summarization in place at the distribution ABR, each intra-area network prefix LSA in the local area is converted to a matching type-3 summary network LSA in the backbone area. A change in the cost or the deletion of a link in the local area needs to be propagated throughout the network as a summary route LSA update. With the appropriate route summarization in place, a single summary LSA is advertised into the backbone from the local area. Changes to specific intra-area LSAs within a summary range are not propagated to the backbone and the rest of the network unless a catastrophic change occurs in which all networks contained within the summary disappear, in which case a single LSA is flooded, minimizing the impact of the rest of the network. As shown in Figure 18, without summarization configured at the distribution ABRs, any link change within the local area is propagated to the rest of the network via an LSA flood.

Figure 18 OSPF Route Summarization Limits Propagation of LSAs



Summarization of distribution area routes is accomplished through the use of the **area range** command. The **range** command defines which subnets within the specified distribution area are summarized into a single outbound summary network advertisement. The cost of this route is calculated in one of the following ways,

- Based on the minimum metric/cost of any of the component routes being summarized. This is the behavior as defined in the earlier OSPF version 1 (RFC 1583). This is no longer the default behavior for a Cisco router.
- Based on the maximum metric/cost of any of the summarized components. This is the behavior as defined in the version 2 OSPF specification (RFC 2328). This is the default behavior for all Cisco IOS versions supported in the current generation Catalyst switches being used in campus designs.
- As an explicitly defined static cost. This cost is defined as a parameter in the **area range** command.

Cisco recommends that the cost of the advertised summary network route be specified with a static or hard coded cost, as shown in the following configuration example.

```
router ospf 100
```

```
router-id 10.122.102.2
ispf
log-adjacency-changes
auto-cost reference-bandwidth 10000
area 120 stub no-summary
area 120 range 10.120.0.0 255.255.0.0 cost 10
timers throttle spf 10 100 5000
timers throttle lsa all 10 100 5000
timers lsa arrival 80
network 10.120.0.0 0.0.255.255 area 120
network 10.122.0.0 0.0.255.255 area 0
```

The use of a static cost is recommended to provide a guaranteed equal-cost path to the distribution block via both of the distribution ABRs. In normal operations, the use of the default behavior (network summary cost being set equivalent to the maximum component cost) is sufficient to meet this need, but the use of a static cost provides for a guaranteed behavior, over-riding any operational or network errors. The area range statement is configured under the appropriate OSPF process as shown above. The presence of any more specific components of the 10.120.0.0/16 address space causes OSPF to generate a summary network LSA for the 10.120.0.0/16 network, and advertises only that distribution area LSA upstream to the core switches, as follows.

```
Core-Switch-1#sh ip route 10.120.0.0
Routing entry for 10.120.0.0/16
  Known via "ospf 100", distance 110, metric 11, type inter area
  Last update from 10.122.0.26 on TenGigabitEthernet3/1, 02:54:37 ago
  Routing Descriptor Blocks:
  * 10.122.0.32, from 10.122.102.2, 02:54:37 ago, via TenGigabitEthernet3/2
      Route metric is 11, traffic share count is 1
    10.122.0.26, from 10.122.102.1, 02:54:37 ago, via TenGigabitEthernet3/1
      Route metric is 11, traffic share count is 1
```

The use of summarized routes is consistent with the design principles of the campus network with its tiered traffic flows. In the structured hierarchical campus design, there is no need to propagate any specific routing information from a distribution area into the rest of the network. All traffic within the distribution area is routed via the distribution switches, and all traffic to and from the access subnets is also routed through the distribution nodes. The need to advertise specific subnet routes from an OSPF area into the backbone of the network is usually a result of the need to engineer traffic flows to ensure an optimal path and/or manage traffic volumes. In the case of the hierarchical campus, highly granular traffic load balancing is achieved via hardware CEF forwarding on the core and distribution switches, which allows flows to be load balanced on a per session basis. Additionally, the recommended structured design provides direct connectivity between both ABRs and all destination subnets, ensuring that no sub-optimal traffic paths are used. The design also uses redundant high capacity GigE and 10 GigE links to ensure sufficient capacity on all traffic paths. The need to engineer traffic flows is significantly reduced in the hierarchical campus design, and the advantages in terms of route reduction and controls on LSA flooding outweigh most any advantages gained through a more granular routing design.

The use of stub areas and route summarization reduces the number of routes and LSA entries in the topology database, which in turn also reduces the amount of memory and CPU requirements on the campus switches. Although memory and CPU utilization are not as large a concern as they have been in the past (given the use of current generation of hardware platforms), the reduction in total number of routes may still need to be managed. Hardware ASIC-based forwarding uses TCAM-based technology to manage hardware forwarding, ACL, and QoS policy information. In a very large and complex network, it may be possible to over-allocate the TCAM resources in the non-chassis-based switches (chassis-based Cisco 4500s and 6500s have larger TCAMs to handle more complex and larger network designs). In the event that it is not possible to tune the route table size, it may be necessary to tune the 3xx0 series switches TCAM allocation to meet specific network needs. The reallocation of TCAM resources via pre-defined templates can be found at the following URL: http://www.cisco.com/en/US/docs/switches/lan/catalyst3750/software/release/12.2_44_se/configuration/guide/swsdm.html.

SPF and LSA Throttle Tuning

The optimal OSPF design for routed access must include improving the convergence time for the OSPF routing protocol itself. As discussed above, the time required to restore traffic flows depends on the following three factors:

- Time to detect the failure
- Time to determine the new optimal path
- Time to update the software and hardware forwarding tables

Of these three times, the first and third can be controlled through a combination of physical design and routing design. Although the second factor is also partially controlled through the use of good area and routing design, optimal design by itself is not always sufficient to meet convergence requirements. In these cases, Cisco recommends tuning of the OSPF timers and process itself. OSPF operates using a link-state routing algorithm that uses LSAs to propagate information about the links and nodes in the network, and the Djiskstra SPF algorithm to calculate the network topology. Updates to the switches routing table involve a process of flooding and receiving updated LSAs, followed by an update of the network topology by the SPF algorithm. The time taken to complete this process depends on multiple factors, including the following:

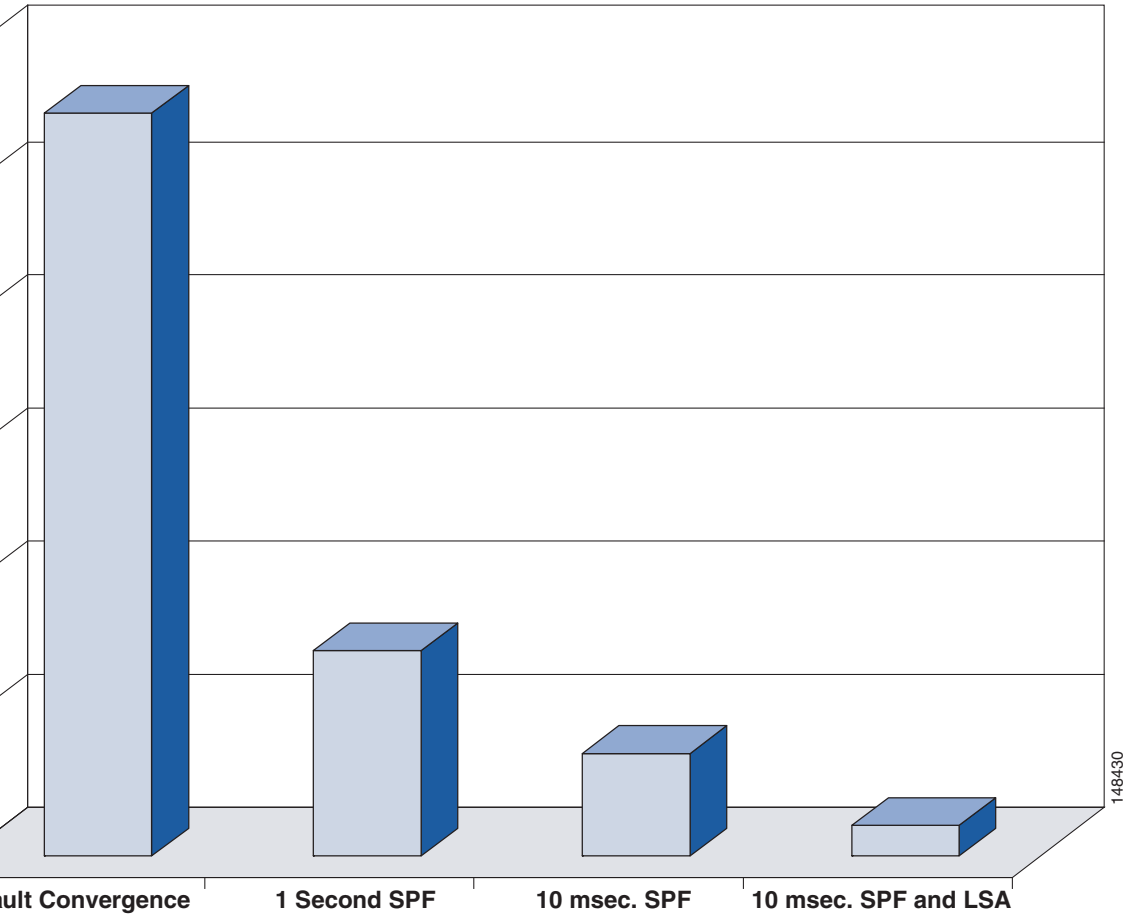
- Number of LSAs
- Number of nodes that need to receive the LSAs
- Time required to transmit the LSAs
- Time required run the SPF calculation

Note The removal of a single route from a group of equal-cost path routes because of the loss of a direct attached link does not require an SPF recalculation to be completed before the routing and forwarding tables are updated. Although LSAs are flooded and SPF is run in this type of event, the network is able to recover from this type of failure without needing to wait for OSPF. In this case, the time to complete the second step in the network recovery process (determining new forwarding paths) has no impact on network recovery.

The asynchronous nature of LSA generation and SPF calculation has historically required the application of some bounds on how often LSAs are generated, and how often the SPF algorithm has run. If allowed to run in an unconstrained fashion, a flapping link or any other network anomaly can have a significant impact on the stability of the OSPF routing environment. For this reason, the OSPF routing process has been implemented with built-in throttling mechanisms to provide a tunable control on the impact of anomaly conditions on the stability of the network. The consequence of having these throttle mechanisms in place is that the convergence times for OSPF without tuning can be longer than that of other routing protocols.

Within the hierarchical campus environment, a number of inherent advantages permit tuning of the SPF and LSA throttling parameters to provide improved convergence without impacting overall network stability. The hierarchical campus design provides well-defined boundaries on neighbor counts, scope of LSA propagation, and LSA database and route table size. Additionally, the campus network uses hardware-based forwarding engines and is primarily constructed with highly available point-to-point fiber connections. All of these factors serve to provide tighter bounds on the potential maximum load experienced by the OSPF routing processes running in the campus network. Given these advantages, it is possible to make use of the enhanced SPF and LSA throttle tuning capabilities in the current generation of Cisco IOS running in Catalyst switches. As shown in Figure 19, via a series of tuning changes, campus network convergence was reduced from 5.68 seconds for a default configuration to 200 msec for a well-tuned configuration.

Figure 19 Impact of Tuning LSA and SPF Throttle on Recovery of Voice Flows



Note All test results used in this document are based on testing in the Cisco campus test bed. For more information on the details of the test bed, traffic types and load, along with detailed results analysis, see the *High Availability Campus Recovery Analysis* at the following URL: http://www.cisco.com/en/US/netsol/ns815/networking_solutions_program_home.html. These values are representative of what can be achieved in a production network following the design recommendations and using similar hardware and software as that described in the analysis document.

The differences in convergence times shown in the four test cases are a result of the impact of the varying degrees of throttle timer on both LSA and SPF processing. In the initial case, using the default initial SPF throttle value of 5 seconds and the LSA throttle value of 500 msec, the overall convergence was 5.68 seconds.

```
router ospf 100
 router-id 10.120.250.101
 log-adjacency-changes
 area 120 stub no-summary
 <timers spf 5 10>          ! Default Values
 network 10.120.0.0 0.0.255.255 area 120
```

In the second case with an identical network scenario, by reducing the SPF timer to an initial SPF throttle of 1 second, the convergence time drops to 1.7 seconds.

```
router ospf 100
 router-id 10.120.250.101
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers spf 1 5
 network 10.120.0.0 0.0.255.255 area 120
```

SPF Throttle Tuning

Before the release of the OSPF throttling enhancements in Cisco IOS 12.2S, it was difficult to improve the convergence of an OSPF campus network much beyond this. Although it was possible to set the initial SPF timer to 0, thereby allowing a sub-second convergence, this was not considered to be best practice. By setting the initial timer value to 0, there was always the probability that a subsequent LSA would arrive at the same switch, requiring a second calculation of the SPF algorithm. This second SPF had to wait for the expiration of a secondary timer (the SPF hold timer). Although it was possible to also reduce the hold timer to 0, this would effectively remove any protection mechanism against the switch thrashing in the event of some network anomaly. The SPF process would be allowed to run continuously over and over again and the network might never stabilize. The throttle timers existed to protect the network and the practice of disabling them was never encouraged. As a consequence of this requirement, it was best practice to tune the initial timer to 1 second to allow for all LSAs flooded as a result of a network failure to be received by the switch before running the initial SPF calculation. This allowed the hold timer to be set at a large enough value, usually 5 seconds, to prevent any catastrophic system overload. To improve on OSPF convergence and provide stable sub-second convergence, it is necessary to use the newer SPF throttle capabilities.

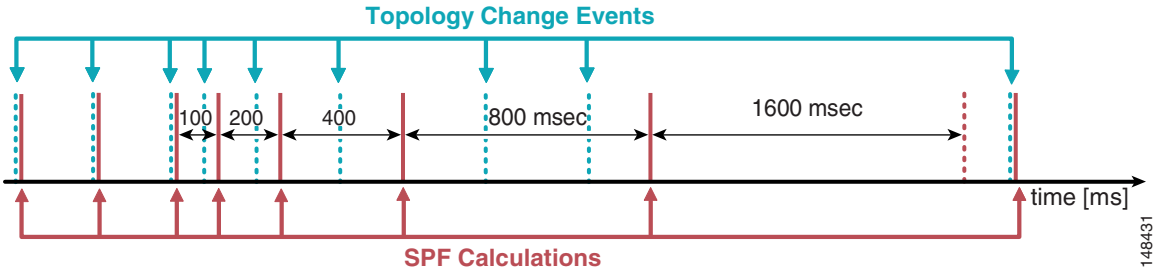
With the introduction of the new SPF throttle mechanism, the interaction of the throttle timers has been improved to implement an exponential back-off mechanism in the event of multiple triggers for sequential SPF runs. The throttle timer is now configured with three values: *spf-start, spf-hold, spf-max-wait*.

```
router ospf 100
```

```
 router-id 10.120.250.101
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers throttle spf <spf-start> <spf-hold> <spf-max-wait>
 network 10.120.0.0 0.0.255.255 area 120
```

The three parameters inter-operate to determine how long it takes for an SPF calculation to be run after notification of a topology change event (arrival of an LSA). On the arrival of the first topology notification, the spf-start or initial hold timer controls how long to wait before starting the SPF calculation. If no subsequent topology change notification arrives (new LSA) during the hold interval, the SPF is free to run again as soon as the next topology change event is received. However, if a second topology change event is received during the hold interval, the SPF calculation is delayed until the hold interval expires. Additionally, in this second case, the hold interval is temporarily doubled. In the event of more topology changes occurring during this new hold interval, the hold interval continues to grow until the maximum period configured is reached. After the expiration of any hold interval, the timer is reset and any future topology changes trigger an SPF again based on the initial timer. This sequence of events is shown in Figure 20.

Figure 20 Interaction of SPF Throttle Timers



With the introduction of the new SPF throttle timers, it is now possible to safely reduce the initial SPF timer to a sub-second value and improve the convergence time of the network further, as shown in the third column of Figure 20. The presence of the exponential back-off for the hold timers also allows for the safe reduction of the hold timer.

The recommended values are as follows:

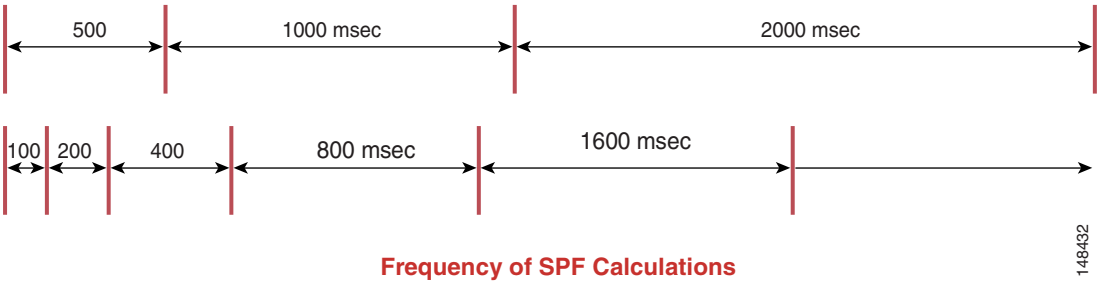
```
spf-start:     10 msec
spf-hold:      100 to 500 msec
spf-max-wait:   5 seconds

router ospf 100
 router-id 10.120.250.101
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
```

```
area 120 stub no-summary
timers throttle spf 10 100 5000
network 10.120.0.0 0.0.255.255 area 120
```

When designing a routed access campus to achieve sub-second convergence, the most important of the three timers to modify is the spf-start, or initial wait timer. By reducing this timer from the previously recommended 1 second value to the new throttle feature value of 10 msec, the convergence time can be reduced from 1.7 seconds to 0.72 seconds. It is still not recommended to set this timer to 0. Providing a slight wait interval provides a window in which multiple LSA events caused by multiple interface changes can be processed together. Cisco recommends that the wait interval be at least 10 msec (equivalent to the interface debounce timer on GigE and 10GigE fiber links). Configuring a short but non-zero initial wait timer should allow most local interface changes occurring simultaneously because of a major fiber cut or line card failure to be processed concurrently.

When considering tuning the hold timer, it is advisable to consider the stability of the campus infrastructure. Increasing the value of the hold timer reduces the number of iterations of the SPF algorithm in the event of a flapping link. Figure 21 shows how increasing the hold timer from 100 to 500 msec reduces the number of SPF calculations occurring over a fixed period of time (assuming that new LSAs are continuing to arrive).

Figure 21 Impact of Decreasing the SPF Hold Timer



In a stable campus environment with a well-summarized area design, there are mitigating factors on the probability of a flapping condition such that lowering the hold timer should not normally present a problem. However, this should be balanced against the understanding that in the hierarchical campus design, reducing the hold timer normally has little impact on the ability of the network to restore traffic flows. As described above in Hierarchical Design, page 9, the only time the campus network relies on OSPF processing to install a new route and restore traffic is in the event of a failure of an access-to-distribution uplink. This is the one case in a fully redundant hierarchical campus design where it is not possible to failover to a second equal-cost path route, and it is only in this specific failure case that OSPF must determine and install a new route.

In this specific failure case, link failure detection and LSA flooding is required; however, the distribution switch with the failed link is a primary source of that new LSA. On failure of the locally-connected access-distribution link, the distribution switch removes the forwarding entries associated with the lost link, and also generates an LSA update indicating the topology change and flooding it throughout the area. This same LSA is then used to trigger a local SPF calculation that installs a new route for the affected access switch subnets via the peer distribution node, correctly restoring all traffic flows. In this single failure scenario, no other topology information is required to complete the restoration of traffic flows, and the distribution switch effectively completes a "self-contained" recovery. Any subsequently received LSAs generated by the access switch at the other end of the failed link are received and processed but do not contain new topology information, and do not impact the already converged network. The impact of this behavior is that if the second SPF is delayed because of using a longer hold time, the network convergence is not impacted.

Testing in the Cisco ESE campus test bed has shown that in the recommended hierarchical design with OSPF, the network remains stable and meets expected convergence criteria using a hold timer value of 100 msec even under stress test cases including persistent route flaps, demonstrating that this value is a suitable setting for networks concerned about dual failure scenarios. On the other hand, increasing the hold timer to 500 msec does not impact the convergence times because of single component failures, and may be useful for networks that may experience any regular link loss and wish to take a more conservative configuration approach.

The criteria used in determining the configuration of the spf-max-wait timer is again dependent on the design of the network, the stability of the fiber plant, the number of LSA/routes, and the hardware platforms in use. Testing in the Cisco ESE lab demonstrated that a maximum hold time of 5 seconds is sufficient to ensure that a 6500 Sup2 or Sup720 running with 3400 routes continues to successfully pass traffic and prevent the CPU from exceeding an average of 40 percent, with 3000 of those routes continually flapping. (The traffic flows being measured were passing between source and destination addresses not being impacted by the flapping routes.) In a network with larger routing tables or smaller hardware platforms, it may be necessary to increase the maximum wait interval.

For more information on the configuration of SPF throttle timers, see the following URL: http://www.cisco.com/en/US/docs/ios/12_2s/feature/guide/fs_spftrl.html.

Caution ⚠ The use of sub-second SPF throttle timers is recommended only in an environment using a well-structured area, route summarization design, and link flap damping features. Without the appropriate mechanisms to control the rate and scope of any network convergence event, it is not advisable to tightly tune any timer-based routing control mechanism.

LSA Throttle Tuning

The use of SPF throttle timer tuning can aid in improving the convergence of the campus network to within the sub-second threshold, but is not sufficient to ensure optimal convergence times. Two factors impact the ability of OSPF to converge: the time waiting for an SPF calculation, and the time waiting for an LSA to be received indicating a network topology change. Before the 12.2S release, Cisco IOS implemented two internal timers affecting the generation of LSAs. The first was an internal delay timer that throttled the generation of router (type-1) and network (type-2) LSAs for 500 msec after a network interface change. A second timer throttled the generation of any specific updated LSA for at least five seconds after having sent the same LSA. These two timers could impact the speed at which the network was able to converge. On the detection of any interface change, OSPF would not generate an LSA indicating the link status change for 500 msec, thus preventing the SPF process from responding to the link failure for at least 500 additional msec. After this occurred, any additional change such as link restoration was throttled for a further five seconds, also potentially impacting recovery. The presence of these delay timers, like the SPF timers, was based on a need to ensure the stability of the network and mitigate against OSPF thrashing in the event of a flapping link or other network problem.

The same design and physical factors that allow for SPF tuning in the campus environment also make it amenable to tuning of the LSA timers. The use of routed point-to-point interfaces in the campus removes the need to consider the loss of multiple logical links in the event of a single interface failure (as is the case in a multi-point WAN environment). The use of direct fiber connections between devices also reduces the probability for link loss and ensures a higher degree of accurate link status detection (no LMI or other soft WAN-like failures need to be considered). Interface-specific features such as debounce timers and IP event dampening also lessen the probability of false or flapping interface conditions. The combination of these factors serves to mitigate the factors with which the LSA timers were initially designed to address.

Tuning LSA throttle timers uses an approach similar to that described above for SPF. Three configuration values are used: an initial delay timer, a hold timer, and a maximum hold timer. Using a similar approach to that discussed above results in the use of the same timer values for the LSA configuration as for the SPF configuration.

The recommended values are as follows:

```
lsa-start:     10 msec
lsa-hold:      100 to 500 msec
lsa-max-wait: 5 seconds

router ospf 100
 router-id 10.120.250.101
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 network 10.120.0.0 0.0.255.255 area 120
```

Using this configuration for both LSA and SPF throttle timers produces a further reduction in network convergence from 0.72 to 0.24 seconds. The combination of tuning LSA and SPF timers from their defaults values down to the values shown above provides an overall improvement in convergence time from 5.68 seconds to 0.24 seconds.

In tuning the throttle timer controlling the generation of LSAs, it is necessary to make a similar configuration to the throttle timer controlling the receipt of LSAs. The "lsa arrival" timer controls the rate at which a switch accepts a second LSA with the same LSA ID. If distribution switch A is configured to generate LSAs with a hold time of 100 msec, it is necessary for the adjacent switches, such as distribution switch B for example, to be configured to accept LSAs at a rate at least equal to that with which they are generated. It is considered best practice to tune the arrival rate at some value less than the generated rate to accommodate for any buffering or internal process timer scheduling delays. Using a hold time of 100 msec, an LSA arrival value of 80 msec is considered sufficient.

```
router ospf 100
 router-id 10.120.250.1
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
 network 10.120.0.0 0.0.255.255 area 120
```

Note For more information on configuration of LSA throttle timers, see the following URL: http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/fsolsath.html.

The configuration of the OSPF LSA and SPF configuration for switches can be verified by reviewing the output of the **show ip ospf** command. On changing the configuration of an active OSPF process, if the configuration does not appear to be active, it may be necessary to either restart the OSPF process by issuing a **clear ip ospf process** command or reloading the router.

```
Distribution-Switch#sh ip ospf
 Routing Process "ospf 100" with ID 10.120.250.1
 Supports only single TOS(TOS0) routes
 Supports opaque LSA
 Supports Link-local Signaling (LLS)
 Supports area transit capability
 Initial SPF schedule delay 10 msecs
 Minimum hold time between two consecutive SPFs 100 msecs
 Maximum wait time between two consecutive SPFs 5000 msecs
 Incremental-SPF disabled
 Initial LSA throttle delay 10 msecs
 Minimum hold time for LSA throttle 100 msecs
 Maximum wait time for LSA throttle 5000 msecs
```

```
Minimum LSA arrival 80 msecs
LSA group pacing timer 240 secs
Interface flood pacing timer 33 msecs
Retransmission pacing timer 66 msecs
. . . .
```

Note As was mentioned in regards to EIGRP above, to ensure a predictable OSPF convergence time, you also need to protect the network against anomalous events such as worms, DDoS attacks, and Spanning Tree loops that may cause high CPU on the switches. The use of Cisco Catalyst security features such as hardware rate limiters, QoS, CEF, and CISFs in conjunction with network security best practices as described in the SAFE design guides is a necessary component in a high availability campus design. For more information on SAFE, see the following URLhttp://www.cisco.com/en/US/netsol/ns744/networking_solutions_program_home.html.

Interface Timer Tuning

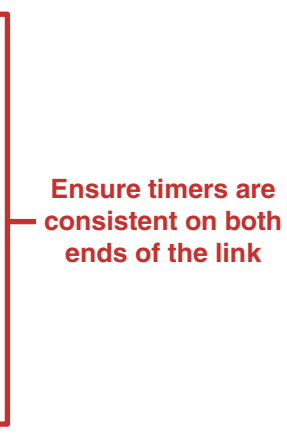
Hello and Dead Timer

As described above in Link Failure Detection Tuning, page 14, the recommended best practice for campus design uses point-to-point GigE and 10GigE fiber connections for all links between switches. In this environment, remote node and link loss detection is normally accomplished using the remote fault detection mechanism implemented as a part of the 802.3z and 802.3ae link protocols. It is still recommended in the Layer 3 campus design that the OSPF hello and dead timers be reduced to 250 msec and 1 second, respectively. The loss of hellos and the expiration of the dead timer is not the primary fault detection mechanism in the campus, but does provide a backup to the L1/2 remote fault detection mechanisms. In the rare case where a routed interface remains up after link loss, OSPF hello and dead timers are needed to detect neighbor loss to initiate convergence around failed link or neighbor.

In the configuration example shown in Figure 22, the hello interval is not explicitly configured but is calculated by taking the minimal dead interval, 1 second, as specified by the **minimal** keyword, and then dividing by the hello-multiplier value configured.

Figure 22 Hello Interval Configuration

```
interface GigabitEthernet3/11
 description Link to Access Switch
 dampening
 ip address 10.120.0.204 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 255
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp


interface GigabitEthernet1/1
 description Uplink to Distribution 1
 dampening
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 0
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
```



148433

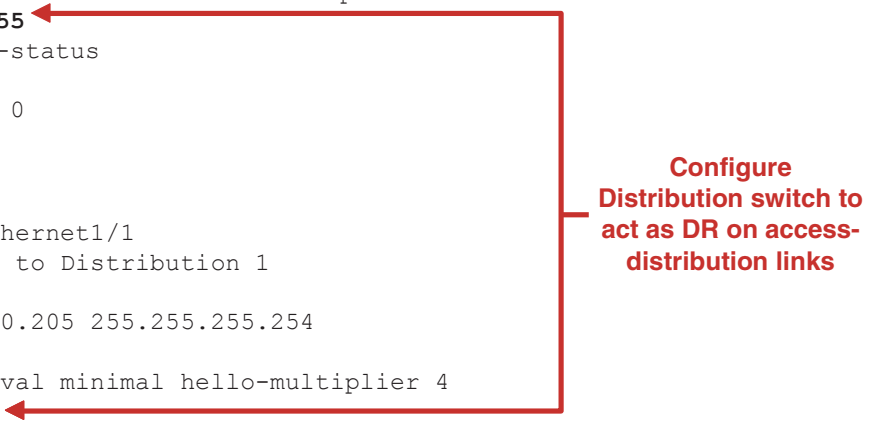
In the above example, 1 second divided by 4 intervals gives an interval value of 250 msec. The selection for the number of hellos sent in the 1 second interval should be at least 3 to allow for enough resiliency in the case of potential packet loss (a very low probability in a fiber-based campus network but still possible).

Designated Router

Although the campus is configured with point-to-point GigE links, OSPF still negotiates the designated and backup designated router on each of the switch-to-switch links. In the campus environment, the selection of which switch is selected as DR has no impact on the stability or speed of convergence of the network. However, it is still recommended that the distribution switch be configured to act as DR on each of the access-distribution uplinks. In the event of an access-distribution uplink fiber failure, the distribution switch acting as the DR can directly propagate updated network LSAs to all connected access and distribution switch peers in the area.

Figure 23 Designated Router Configuration

```
interface GigabitEthernet3/11
 description Link to Access Switch
 dampening
 ip address 10.120.0.204 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 255
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp


interface GigabitEthernet1/1
 description Uplink to Distribution 1
 dampening
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 0
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
```



148434

6 Routed Access Design Considerations

IP Addressing

The implementation of a routed access design requires the allocation of additional IP addresses to the point–to-point subnets between the distribution and the access switches. For each access switch, this entails the allocation of two new subnets, one for each uplink. These subnets should be contained within the summarized address block advertised upstream to the core of the network, and do not increase the number of routes contained within the core of the network.

Addressing Option 1—VLSM Addressing using /30 Subnets

Using 30-bit subnet masking (255.255.255.252) provides an efficient use of VLSM address space. A single class C address block chosen out of the summarized address range for the distribution block addresses links to 32 access switches, which is sufficient for all but the largest distribution block.

```
interface GigabitEthernet3/3
    description Distribution Downlink
    ip address 10.120.0.198 255.255.255.252

interface GigabitEthernet1/1
    description Access Uplink
    ip address 10.120.0.197 255.255.255.252
```

Addressing Option 2—VLSM Addressing using /31 Subnets

It may be desirable to use 31-bit masking (255.255.255.254) on the distribution-to-access point to provide an even more efficient usage of address space. 31-bit prefixes as defined in RFC 3021 provide for twice as many subnets to be created out of the same block of addresses as would be available using /30 addressing.

```
interface GigabitEthernet3/3
    description Distribution Downlink
    ip address 10.120.0.196 255.255.255.254

interface GigabitEthernet1/1
    description Access Uplink
    ip address 10.120.0.197 255.255.255.254
```

Note For more information on /31 addressing, see RFC 3021 at the following URL: http://www.faqs.org/rfcs/rfc3021.html

VLAN Usage

In the traditional Layer 2 access design, it is the best practice recommendation to not span any VLANs between access switches. Each access switch is configured with a unique data, voice, and native trunk VLAN. The routed access design uses the same VLAN assignment policy. Each access switch is configured with a unique data and voice VLAN. The replacement of the uplink trunks with point-to-point routed links removes the need for a dedicated trunk VLAN. If there is a business requirement to span a VLAN between access switches, it is not possible to use a Layer 3 routed access configuration.

Switch Management VLAN

In the Layer 2 access design, it was traditionally considered best practice to define a unique VLAN for network management. This VLAN was often spanned between multiple access switches, and the switch management or "sc0" interface was assigned to this VLAN. The use of a distinct VLAN for switch management was originally intended to provide a distinct Layer 3 interface that could be configured to control access to the switch management interface as well as to control the amount of end user broadcast traffic the switch CPU was required to process. More current generations of switch hardware and software can provide this same access control and CPU protection without the need to define a unique switch management VLAN. However, a unique VLAN is often still used in many customer networks. In the routed access design, it is no longer desirable to create a separate switch management VLAN, but rather to configure a dedicated loopback interface with a /32 network:

```
interface Loopback0
 description Dedicated Switch Management
 ip address 10.120.254.1 255.255.255.255
```

The /32 network defined for the loopback interface should be a specific network included in the summarized distribution block route advertised to the network core. It should be configured to be a passive interface in the EIGRP or OSPF router configuration, and access control lists applied to meet specific network security requirements.

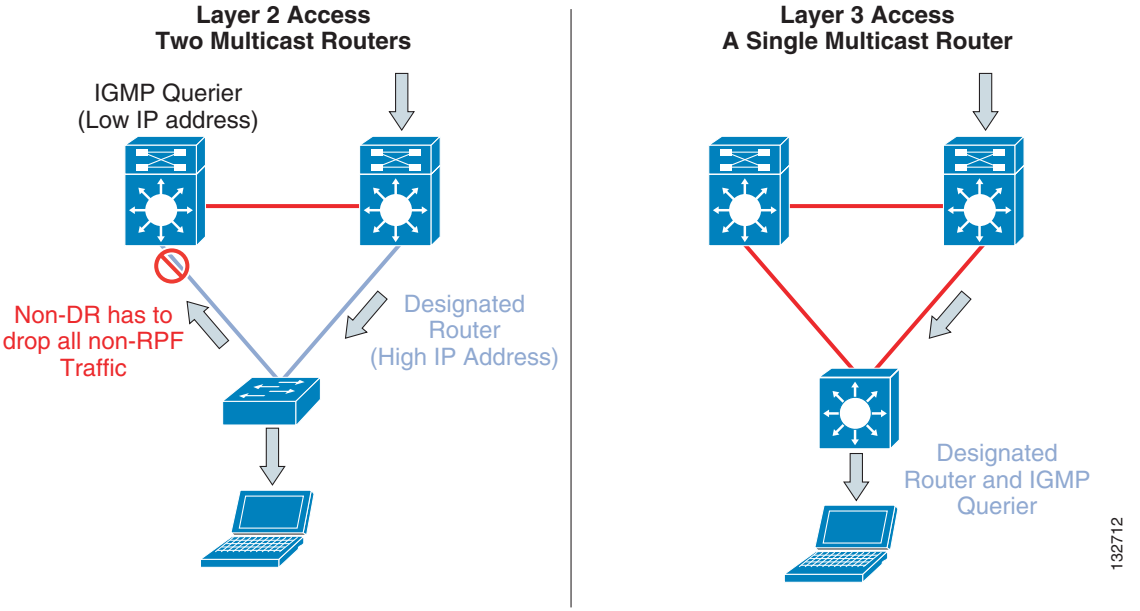
Multicast Considerations

The routed access design simplifies portions of the multicast configuration because it reduces the number of routers connected to the access VLANs. In the Layer 2 design, both the distribution routers participate as multicast routers and share the edge multicast Layer 3 functions. The switch with the lower IP address assumes the role of the IGMP querier and the switch with the higher address is elected as the PIM DR. It is the best practice recommendation that the PIM DR also be the active HSRP peer for the access subnet, which requires that three configuration parameters be synchronized in the Layer 2 access designs. The root bridge, the HSRP active node, and the PIM DR should all be on the same distribution switch for each specific VLAN. In the routed access design, this need for synchronized configuration is lessened because there is only one router on the local segment, which by default results in synchronization of the unicast and multicast traffic flows. Additionally, with the migration of the multicast router from the distribution to the access, there is no longer a need to tune the PIM hello timers to ensure rapid convergence between the distribution nodes in the case of a failure. The same remote fault indicator mechanisms that trigger rapid unicast convergence drive the multicast software and hardware recovery processes, and there is no need for Layer 3 detection of path or neighbor failure across the Layer 2 access switch.

The presence of a single router for each access VLAN also removes the need to consider non-reverse path forwarding (non-RPF) traffic received on the access side of the distribution switches. A multicast router drops any multicast traffic received on a non-RPF interface. If there are two routers for a

subnet, the DR forwards the traffic to the subnet, and the non-DR receives that traffic on its own VLAN interface. This is not its shortest path back to the source and so the traffic fails the RPF check (see Figure 24).

Figure 24 Multicast Traffic Flows and Router Functions



In the Layer 3 access design, there is a single router on the access subnet and no non-RPF traffic flows. Although the current generation of Cisco Catalyst switches can process and discard all non-RPF traffic in hardware with no performance impact or access list configuration required, the absence of non-RPF traffic simplifies operation and management.

The following summarizes the campus multicast configuration recommendation:

- The access layer switches have IGMP snooping enabled.
- The RPs are located on the two core layer switches.
- PIM-SM is configured on all access layer, distribution layer, and core-layer switches.
- Anycast RP is configured for fast recovery of IP multicast traffic.
- PIM-SM and MSDP are enabled on all core layer switches.
- Each access layer switch points to the anycast RP address as its RP.
- MSDP is used to synchronize source active (SA) state between the core switches.

Note For complete details on the recommended campus multicast design, see the Cisco *AVVID Network Infrastructure IP Multicast Design SRND* at the following URL: http://www.cisco.com/application/pdf/en/us/guest/tech/tk363/c1501/ccmigration_09186a008015e7cc.pdf.

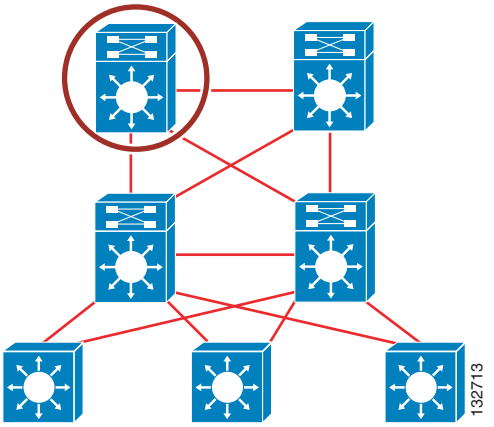
7 Summary

For those enterprise networks that are seeking to reduce dependence on spanning tree and a common control plane, are familiar with standard IP troubleshooting tools and techniques, and desire optimal convergence, a routed access design (Layer 3 switching in the access) using EIGRP or OSPF as the campus routing protocol is a viable option. To achieve the optimal convergence for the routed access design, it is necessary to follow basic hierarchical design best practices and to use advanced EIGRP and OSPF functionality, including Sstub routing, route summarization, and route filtering for EIGRP, and LSA and SPF throttle tuning, totally stubby areas, and route summarization for OSPF as defined in this document.

8 Appendix A—Sample EIGRP Configurations for Layer 3 Access Design

Core Switch Configuration (EIGRP)

Figure 25 Core Switch




```
key chain eigrp
 key 100
  key-string 7 01161501
!
! Enabled spanning tree as a fail-safe practice
spanning-tree mode rapid-pvst
!
redundancy
 mode sso
 main-cpu
  auto-sync running-config
  auto-sync standard
!
! Configure necessary loopback interfaces to support Multicast MSDP and Anycast for
! RP redundancy
interface Loopback0
 description MSDP PEER INT
 ip address 10.122.10.2 255.255.255.255
!
interface Loopback1
 description ANYCAST RP ADDRESS
```

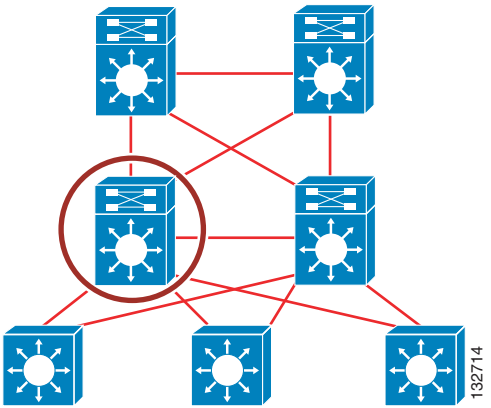
```
 ip address 10.122.100.1 255.255.255.255
!
interface Loopback2
 description Garbage-CAN RP
 ip address 2.2.2.2 255.255.255.255
!
! Configure point to point links to Distribution switches
interface TenGigabitEthernet3/1
 description 10GigE to Distribution 1
! Use of /31 addressing on point to point links optimizes use of IP address space in
! the campus
 ip address 10.122.0.27 255.255.255.254
 ip pim sparse-mode
! Reduce EIGRP hello and hold timers to 1 and 3 seconds. In a point-point L3 campus
! design the EIGRP timers are not the primary mechanism used for link and node
! failure detection. They are intended to provide a fail-safe mechanism only.
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 load-interval 30
! Reduce carrier delay to 0. Tuning carrier delay no longer has an impact on GigE and
! 10GigE interfaces but is recommended to be configured as a best practice for network
! operational consistency
 carrier-delay msec 0
! Configure trust DSCP to provide for maximum granularity of internal QoS queuing
 mls qos trust dscp
!
router eigrp 100
! Passive all interfaces not intended to form EIGRP neighbors
 passive-interface Loopback0
 passive-interface Loopback1
 passive-interface Loopback2
 network 10.0.0.0
 no auto-summary
! Explicitly configure the EIGRP router id as a best practice when using Anycast and/or
! any identical loopback address on multiple routers.
 eigrp router-id 10.122.0.1
!
! Multicast route point and MSDP configuration.
! For a detailed explanation on the specifics of the configuration below please see
! the campus chapter of the multicast design guides.
ip pim rp-address 2.2.2.2
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim accept-register list PERMIT-SOURCES
ip msdp peer 10.122.10.1 connect-source Loopback0
ip msdp description 10.122.10.1 ANYCAST-PEER-6k-core-left
ip msdp cache-sa-state
ip msdp originator-id Loopback0
```

```
!
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
!
ip access-list extended PERMIT-SOURCES
 permit ip 10.121.0.0 0.0.255.255 239.192.240.0 0.0.3.255
 permit ip 10.121.0.0 0.0.255.255 239.192.248.0 0.0.3.255
```

Distribution Node EIGRP

Note Symmetrical configuration on both distribution switches.

Figure 26 Distribution Node



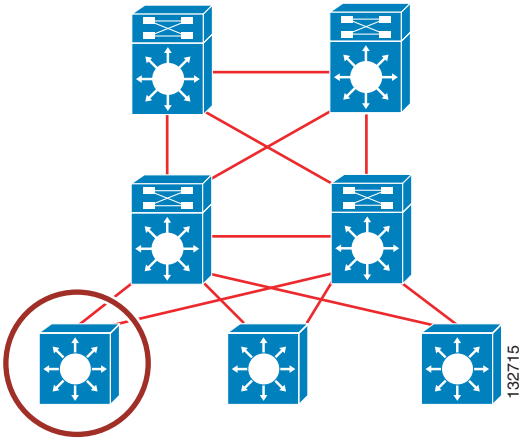

```
key chain eigrp


key-string 7 01161501
!
<Configure spanning tree as a redundant protective mechanism>
spanning-tree mode rapid-pvst
spanning-tree loopguard default
!
<Configure point to point Layer 3 links to each of the access switches>
interface GigabitEthernet3/1
 description Link to Access Switch
```

```
<configure the switch to switch link using a /30 or /31 subnet>
 ip address 10.120.0.204 255.255.255.254
 ip pim sparse-mode
<specify the use of 1 second hello and 3 second dead timers for EIGRP>
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
<enable eigrp MD5 authentication>
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 logging event link-status
 load-interval 30
<Set carrier delay to 0. On Catalyst 6500 this will have no effect on GigE ports however
it is necessary on 3x50 series switches and should be consistently configured for best
practices>
 carrier-delay msec 0
<Trust the dscp settings in all packets sourced from the access. We are extending the
trust boundary to the access switch>
 mls qos trust dscp
!
!
<Configure point to point L3 links to each of the core switches>
interface TenGigabitEthernet4/1
 description 10 GigE to Core 1
<configure the switch to switch link using a /30 or /31 subnet>
 ip address 10.122.0.26 255.255.255.254
 ip pim sparse-mode
<specify the use of 1 second hello and 3 second dead timers for EIGRP>
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
<Configure EIGRP authentication on all links>
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
<Advertise a summary address for the entire distribution block upstream to the core>
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
 logging event link-status
 load-interval 30
 carrier-delay msec 0
<Trust all DSCP markings from the core of the network>
 mls qos trust dscp
!
!
<Configure a point to point Layer 3 link between distribution switches>
interface TenGigabitEthernet4/3
 description 10 GigE to Distribution 2
<configure the switch to switch link using a /30 or /31 subnet>
 ip address 10.122.0.21 255.255.255.254
 ip pim sparse-mode
<specify the use of 1 second hello and 3 second dead timers for EIGRP>
 ip hello-interval eigrp 100 1
```

```
 ip hold-time eigrp 100 3
<Configure EIGRP authentication on all links>
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 logging event link-status
 load-interval 30
 mls qos trust dscp
!
!
router eigrp 100
<Passive all interfaces not connected to another Layer 3 switch>
 passive-interface GigabitEthernet2/1
<Specify which networks should be routed by EIGRP. Include the distribution block and the
core links>
 network 10.120.0.0 0.0.255.255
 network 10.122.0.0 0.0.0.255
<Apply a distribute list filtering all routes other than select default(s) to the access
switches>
 distribute-list Default out GigabitEthernet3/1
 distribute-list Default out GigabitEthernet3/2
 . . .
 distribute-list Default out GigabitEthernet9/14
 distribute-list Default out GigabitEthernet9/15
 no auto-summary
! Explicitly configure the EIGRP router id as a best practice when using Anycast and/or
! any identical loopback address on multiple routers.
 eigrp router-id 10.122.0.3

!
ip classless
no ip http server
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim rp-address 2.2.2.2
ip pim spt-threshold infinity
!
!
ip access-list standard Default
 permit 0.0.0.0
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```

Access Node EIGRP

Figure 27 Access Node




```
key chain eigrp
 key 100

!
<Configure spanning tree as a redundant protective mechanism>
spanning-tree mode rapid-pvst
spanning-tree loopguard default
!
redundancy
 mode sso
 main-cpu
  auto-sync running-config
  auto-sync standard
!
<Create a local Data and Voice VLAN>
vlan 6
 name Access-Data-VLAN
!
vlan 106
 name Access-Voice-VLAN
!
<Configure an RP sink hole for non-authorized Multicast groups>
interface Loopback22
 ip address 2.2.2.2 255.255.255.255
!
<Define the uplink to the Distribution switches as a point to point Layer 3 link>
```

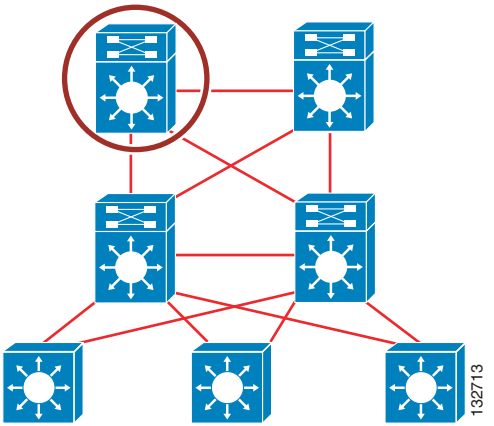
```
interface GigabitEthernet1/1
 description Uplink to Distribution 1
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
<Reduce EIGRP hello and dead timers to 1 and 3 seconds>
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
<Enable EIGRP MD5 authentication>
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
!
interface GigabitEthernet2/1
 description Uplink to Distribution 2
 ip address 10.120.0.61 255.255.255.252
 ip pim sparse-mode
 ip hello-interval eigrp 100 1
 ip hold-time eigrp 100 3
 ip authentication mode eigrp 100 md5
 ip authentication key-chain eigrp 100 eigrp
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
!
<Define Switched Virtual Interfaces's for both access Data and Voice VLANs>
interface Vlan6
 ip address 10.120.6.1 255.255.255.0
 ip helper-address 10.121.0.5
 no ip redirects
 ip pim query-interval 250 msec
 ip pim sparse-mode
 load-interval 30
!
interface Vlan106
 ip address 10.120.106.1 255.255.255.0
 ip helper-address 10.121.0.5
 no ip redirects
 ip pim query-interval 250 msec
 ip pim sparse-mode
 load-interval 30
!
<Configure EIGRP as an EIGRP stub router, advertising connected routes upstream to the
distribution>
router eigrp 100
 network 10.120.0.0 0.0.255.255
```

```
 no auto-summary
 eigrp stub connected
 eigrp router-id 10.122.0.22

!
ip classless
no ip http server
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim rp-address 2.2.2.2
ip pim spt-threshold infinity
!
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```

9 Appendix B—Sample OSPF Configurations for Layer 3 Access Design

Core Switch Configuration (OSPF)

Figure 28 Core Switch




```
! Enabled spanning tree as a fail-safe practice
spanning-tree mode rapid-pvst
spanning-tree loopguard default
```

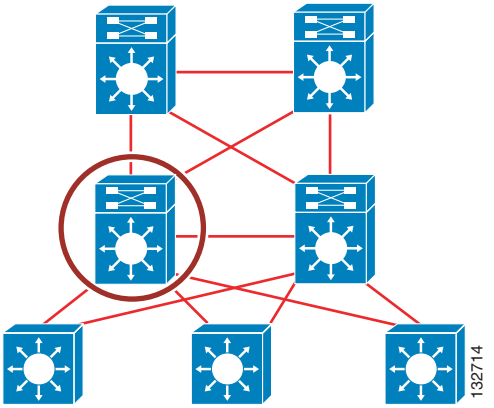
```
!
!
redundancy
 mode sso
 main-cpu
  auto-sync running-config
  auto-sync standard
!
! Configure necessary loopback interfaces to support Multicast MSDP and Anycast for
! RP redundancy
interface Loopback0
 description MSDP PEER INT
 ip address 10.122.10.2 255.255.255.255
!
interface Loopback1
 description ANYCAST RP ADDRESS
 ip address 10.122.100.1 255.255.255.255
!
interface Loopback2
 description Garbage-CAN RP
 ip address 2.2.2.2 255.255.255.255
!
! Configure point to point links to Distribution switches
interface TenGigabitEthernet3/1
 description 10GigE link to HA Distribution 1
! Configure IP Event Dampening on all links using sub-second timers and/or switches
configured with sub-second  ! LSA or SPF throttle timers
 Dampening
! Use of /31 addressing on point to point links optimizes use of IP address space in
! the campus
 ip address 10.122.0.27 255.255.255.254
 ip pim sparse-mode
! Reduce OSPF hello and dead timers to 250 msec and 1 second. In a point-point L3 campus
! design the OSPF timers are not the primary mechanism used for link and node
! failure detection. They are intended to provide a fail-safe mechanism only.
 ip ospf dead-interval minimal hello-multiplier 4
 logging event link-status
 logging event spanning-tree status
 logging event bundle-status
 load-interval 30
! Reduce carrier delay to 0. Tuning carrier delay no longer has an impact on GigE and
! 10GigE interfaces but is recommended to be configured as a best practice for network
! operational consistency
 carrier-delay msec 0
! Configure trust DSCP to provide for maximum granularity of internal QoS queuing
 mls qos trust dscp
!
!
router ospf 100
```

```
! Explicitly configure the OSPF router id as a best practice when using Anycast and/or
! any identical loopback address on multiple routers.
 router-id 10.122.10.1
 log-adjacency-changes
! Modify the reference BW to support 10GigE links
 auto-cost reference-bandwidth 10000
! Reduce the SPF and LSA Throttle timers (see explanation in design guide above for
details)
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
! Passive all interfaces not intended to form OSPF neighbors
 passive-interface Loopback0
 passive-interface Loopback1
 passive-interface Loopback2
 network 10.122.0.0 0.0.255.255 area 0.0.0.0
!
!
! Multicast route point and MSDP configuration.
! For a detailed explanation on the specifics of the configuration below please see
! the campus chapter of the multicast design guides.
ip pim rp-address 2.2.2.2
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim accept-register list PERMIT-SOURCES
ip msdp peer 10.122.10.1 connect-source Loopback0
ip msdp description 10.122.10.1 ANYCAST-PEER-6k-core-left
ip msdp cache-sa-state
ip msdp originator-id Loopback0
!
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
!
ip access-list extended PERMIT-SOURCES
 permit ip 10.121.0.0 0.0.255.255 239.192.240.0 0.0.3.255
 permit ip 10.121.0.0 0.0.255.255 239.192.248.0 0.0.3.255
```

Distribution Node OSPF

Note Symmetrical configuration on both distribution switches.

Figure 29 Distribution Node



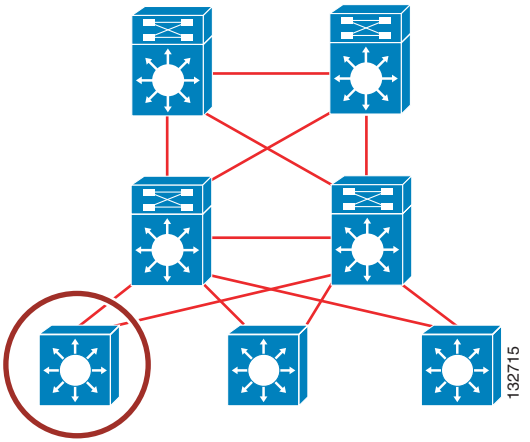

```
!
interface GigabitEthernet3/1
 description Access Switch
! Configure IP Event Dampening on all links using sub-second timers and/or switches
configured with sub-second  ! LSA or SPF throttle timers
 Dampening
! Use of /31 addressing on point to point links optimizes use of IP address space in
! the campus
 ip address 10.120.0.204 255.255.255.254
 ip pim sparse-mode
! Reduce OSPF hello and dead timers to 250 msec and 1 second. In a point-point L3 campus
! design the OSPF timers are not the primary mechanism used for link and node
! failure detection. They are intended to provide a fail-safe mechanism only.
 ip ospf dead-interval minimal hello-multiplier 4
 ip ospf priority 255
 logging event link-status
 logging event spanning-tree status
 logging event bundle-status
 logging event trunk-status
 load-interval 30
! Reduce carrier delay to 0. Tuning carrier delay no longer has an impact on GigE and
! 10GigE interfaces but is recommended to be configured as a best practice for network
! operational consistency
 carrier-delay msec 0
! Configure trust DSCP to provide for maximum granularity of internal QoS queuing
 mls qos trust dscp
!
!
! Configure point to point L3 links to each of the core switches. Follow same interface
configuration as
```

```
! specified on links to access switches
interface TenGigabitEthernet4/1
 description 10 GigE to Core 1
 dampening
 ip address 10.122.0.26 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 logging event link-status
 logging event spanning-tree status
 logging event bundle-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
!
!
! Configure point to point L3 links to the peer distribution switch. Follow same interface
configuration as
! specified on links to access switches
interface TenGigabitEthernet4/3
 description L3 link to peer distribution
 dampening
 ip address 10.120.0.23 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 logging event link-status
 logging event spanning-tree status
 logging event bundle-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp
!
!
router ospf 100
! Explicitly configure the OSPF router id as a best practice when using Anycast and/or
! any identical loopback address on multiple routers.
 router-id 10.122.102.1
 ispf
 log-adjacency-changes
! Modify the reference BW to support 10GigE links
 auto-cost reference-bandwidth 10000
! Configure distribution block area as a totally stubby area to reduce the number of LSA
and routes in the
! access switches
 area 120 stub no-summary
! Summarize the distribution block subnets into a single route advertized into area 0 core
 area 120 range 10.120.0.0 255.255.0.0 cost 10
! Reduce the SPF and LSA Throttle timers (see explanation in design guide above for
details)
 timers throttle spf 10 100 5000
```

```
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
! Define distribution block area and core area
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
!
!
ip classless
no ip http server
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim rp-address 2.2.2.2
ip pim spt-threshold infinity
!
!
ip access-list standard Default
 permit 0.0.0.0
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```

Access Node OSPF

Figure 30 Access Node




```
!
<Configure spanning tree as a redundant protective mechanism>
spanning-tree mode rapid-pvst
```

```
spanning-tree loopguard default
!
redundancy
 mode sso
 main-cpu
  auto-sync running-config
  auto-sync standard
!
! Create a local Data and Voice VLAN
vlan 6
 name Access-Data-VLAN
!
vlan 106
 name Access-Voice-VLAN
!
! Configure an RP sink hole for non-authorized Multicast groups
interface Loopback22
 ip address 2.2.2.2 255.255.255.255
!
! Define the uplink to the Distribution switches as a point to point Layer 3 link

interface GigabitEthernet1/1
 description Uplink to Distribution 1
! Configure IP Event Dampening on all links using sub-second timers and/or switches
configured with sub-second  ! LSA or SPF throttle timers
 Dampening
! Use of /31 addressing on point to point links optimizes use of IP address space in
! the campus
 ip address 10.120.0.205 255.255.255.254
 ip pim sparse-mode
! Reduce OSPF hello and dead timers to 250 msec and 1 second. In a point-point L3 campus
! design the OSPF timers are not the primary mechanism used for link and node
! failure detection. They are intended to provide a fail-safe mechanism only.
 ip ospf dead-interval minimal hello-multiplier 4
 logging event link-status
 load-interval 30
 carrier-delay msec 0
 mls qos trust dscp


interface GigabitEthernet2/1
 description Uplink to Distribution 2
 dampening
 ip address 10.120.0.207 255.255.255.254
 ip pim sparse-mode
 ip ospf dead-interval minimal hello-multiplier 4
 logging event link-status
 load-interval 30
 carrier-delay msec 0
```

```
 mls qos trust dscp
!
!
! Define Switched Virtual Interfaces's for both access Data and Voice VLANs
interface Vlan6
 ip address 10.120.6.1 255.255.255.0
 ip helper-address 10.121.0.5
 no ip redirects
 ip pim query-interval 250 msec
 ip pim sparse-mode
 load-interval 30
!
interface Vlan106
 ip address 10.120.106.1 255.255.255.0
 ip helper-address 10.121.0.5
 no ip redirects
 ip pim query-interval 250 msec
 ip pim sparse-mode
 load-interval 30
!
! Configure the access switch as a member of the totally stubby area
router ospf 100
 router-id 10.120.250.6
 ispf
 log-adjacency-changes
 auto-cost reference-bandwidth 10000
 area 120 stub no-summary
 timers throttle spf 10 100 5000
 timers throttle lsa all 10 100 5000
 timers lsa arrival 80
 network 10.120.0.0 0.0.255.255 area 120
!
ip classless
no ip http server
ip pim rp-address 10.122.100.1 GOOD-IPMC override
ip pim rp-address 2.2.2.2
ip pim spt-threshold infinity
!
ip access-list standard GOOD-IPMC
 permit 224.0.1.39
 permit 224.0.1.40
 permit 239.192.240.0 0.0.3.255
 permit 239.192.248.0 0.0.3.255
```