CHAPTER 2

Data Center Multi-Tier Model Design

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

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

- Data Center Multi-Tier Design Overview
- Data Center Core Layer
- Data Center Aggregation Layer
- Data Center Access Layer
- Data Center Services Layer

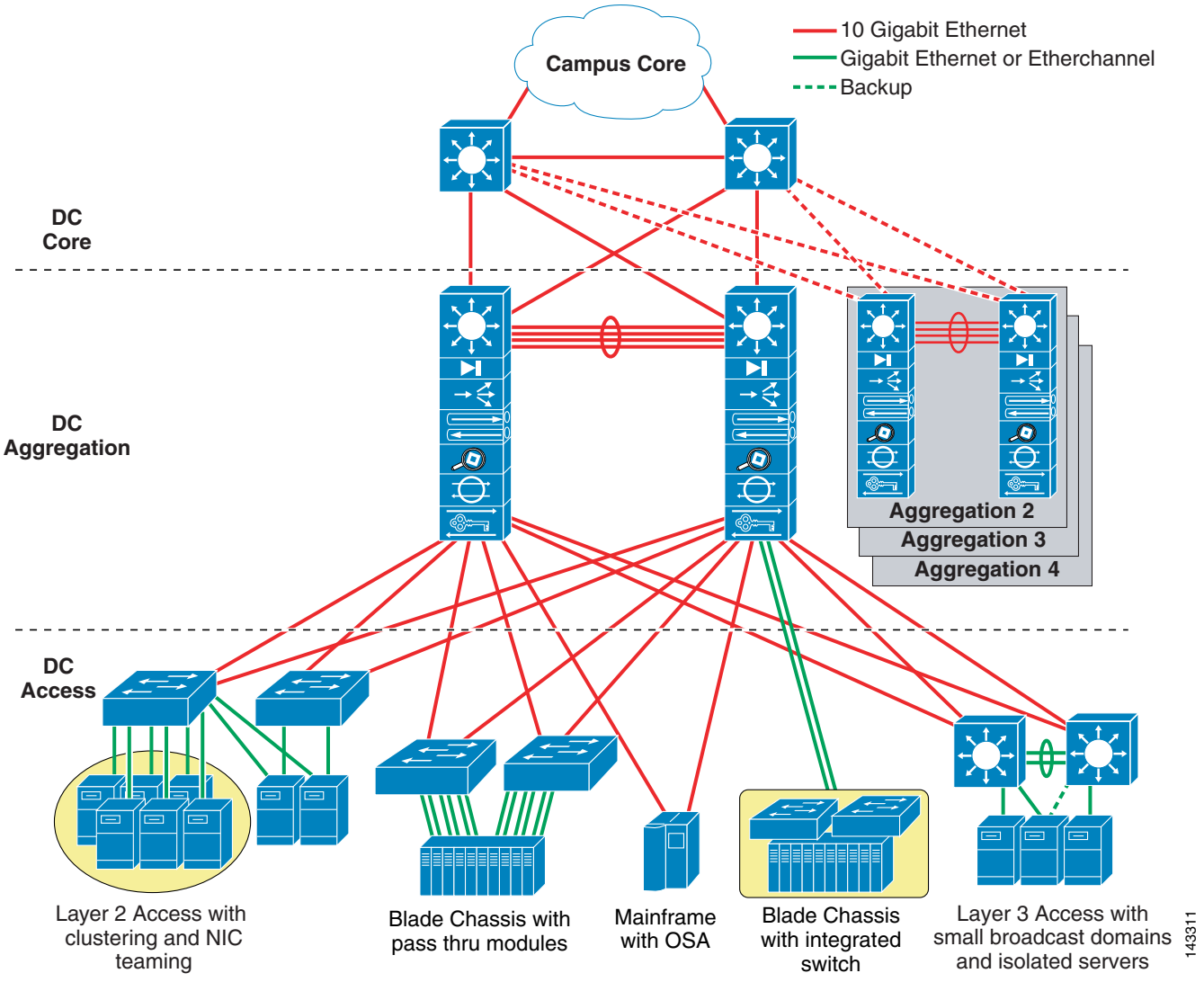
Note For a high-level overview of the multi-tier model, refer to Chapter 1, "Data Center Architecture Overview."

Data Center Multi-Tier Design Overview

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

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

Figure 2-1 Data Center Multi-Tier Model Topology



Data Center Core Layer

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

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

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

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

Recommended Platform and Modules

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

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

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

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

Distributed Forwarding

The Cisco 6700 Series line cards support an optional daughter card module called a Distributed Forwarding Card (DFC). The DFC permits local routing decisions to occur on each line card via a local Forwarding Information Base (FIB). The FIB table on the Sup720 policy feature card (PFC) maintains synchronization with each DFC FIB table on the line cards to ensure accurate routing integrity across the system. Without a DFC card, a compact header lookup must be sent to the PFC on the Sup720 to determine where on the switch fabric to forward each packet to reach its destination. This occurs for both Layer 2 and Layer 3 switched packets. When a DFC is present, the line card can switch a packet directly across the switch fabric to the destination line card without consulting the Sup720 FIB table on the PFC. The difference in performance can range from 30 Mpps system-wide to 48 Mpps *per slot* with DFCs.

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

Table 2-1 Performance Comparison with Distributed Forwarding

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

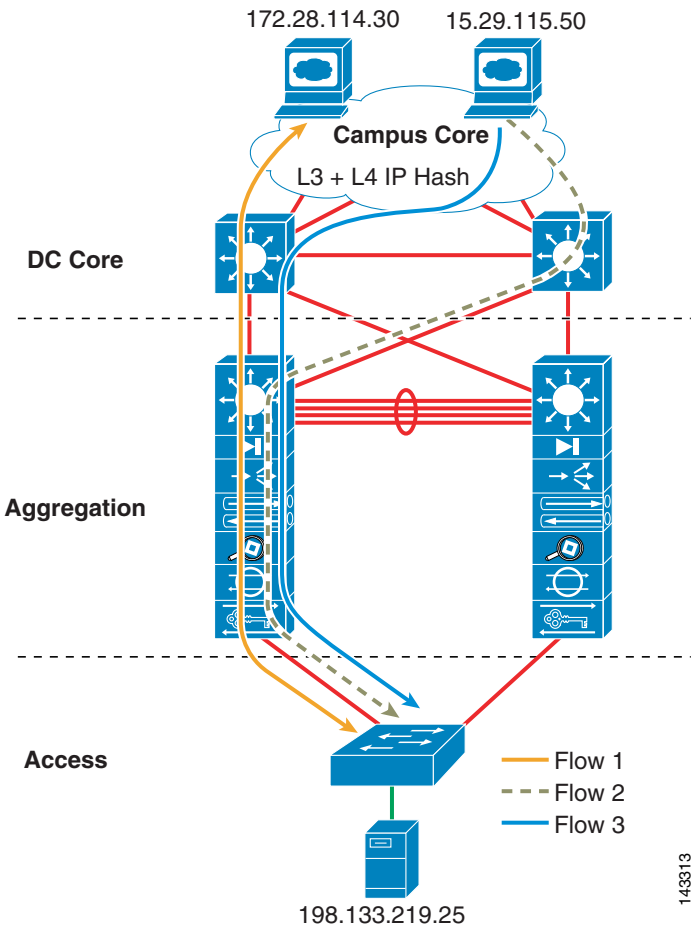
Using DFCs in the core layer of the multi-tier model is optional. An analysis of application session flows that can transit the core helps to determine the maximum bandwidth requirements and whether DFCs would be beneficial. If multiple aggregation modules are used, there is a good chance that a large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

Traffic Flow in the Data Center Core

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

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

Figure 2-2 Traffic Flow through the Core Layer



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

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

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

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

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

Data Center Aggregation Layer

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

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

Recommended Platforms and Modules

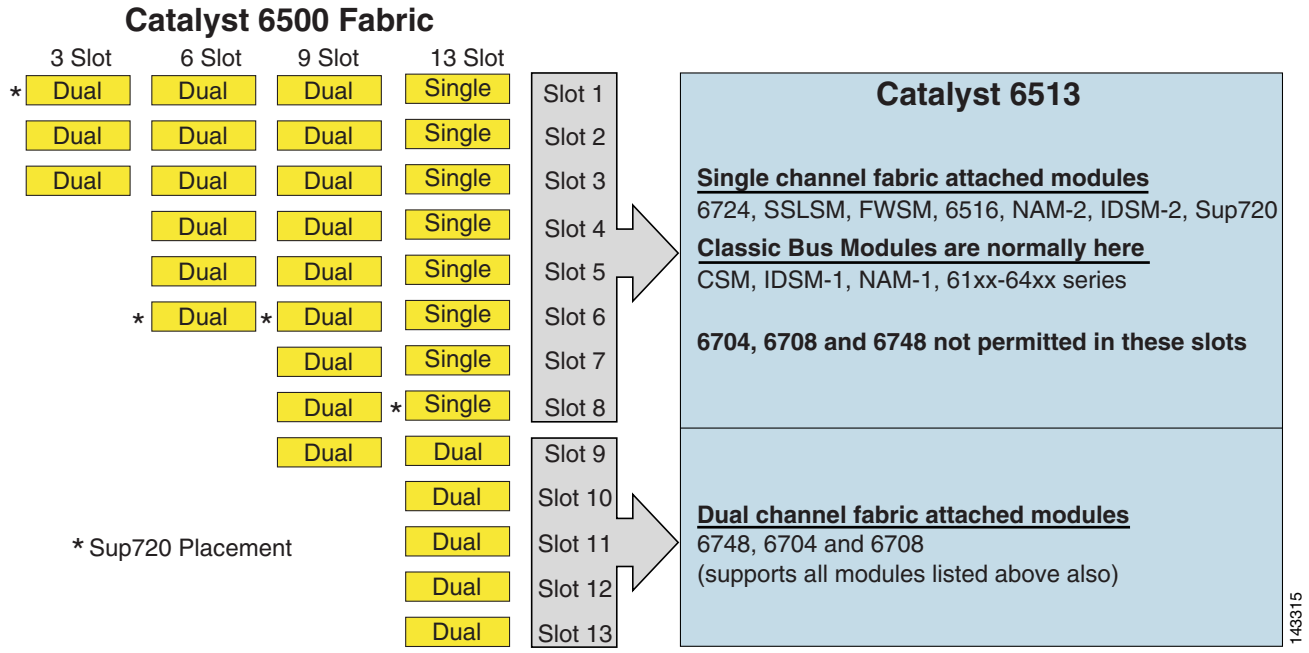
The enterprise data center contains at least one aggregation module that consists of two aggregation layer switches. The aggregation switch pairs work together to provide redundancy and to maintain session state while providing valuable services to the access layer.

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

- Cisco Firewall Services Modules (FWSM)
- Cisco Application Control Engine (ACE)
- Intrusion Detection
- Network Analysis Module (NAM)
- Distributed denial-of-service attack protection (Guard)

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

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



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

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

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

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

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

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

Distributed Forwarding

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

Note For more information on DFC operations, refer to Distributed Forwarding, page 2-4.

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

Traffic Flow in the Data Center Aggregation Layer

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

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

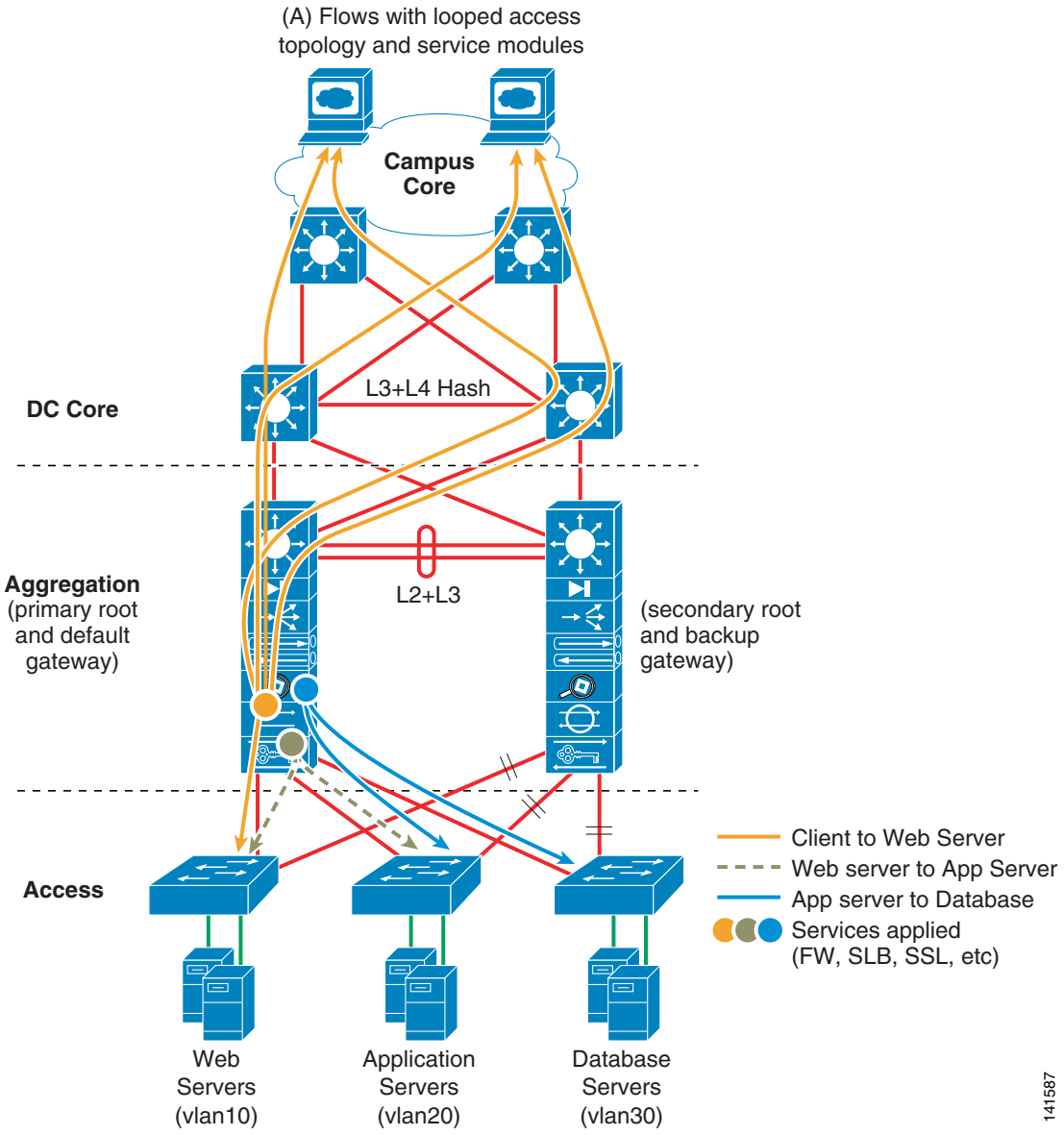
- Core layer to access layer—The core-to-access traffic flows are usually associated with client HTTP-based requests to the web server farm. At least two equal cost routes exist to the web server subnets. The CEF-based L3 plus L4 hashing algorithm determines how sessions balance across the equal cost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

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

Path Selection in the Presence of Service Modules

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

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



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

Note More detail on path preference design is provided in Chapter 7, "Increasing HA in the Data Center."

Without route tuning, the core has two equal cost routes to the server farm subnet; therefore, sessions are distributed across links to both aggregation layer switches. Because Aggregation 1 contains the active service modules, 50 percent of the sessions unnecessarily traverse the inter-switch link between Aggregation 1 and Aggregation 2. By tuning the routing configuration, the sessions can remain on symmetrical paths in a predictable manner. Route tuning also helps in certain failure scenarios that create active-active service module scenarios.

Server Farm Traffic Flow with Service Modules

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

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

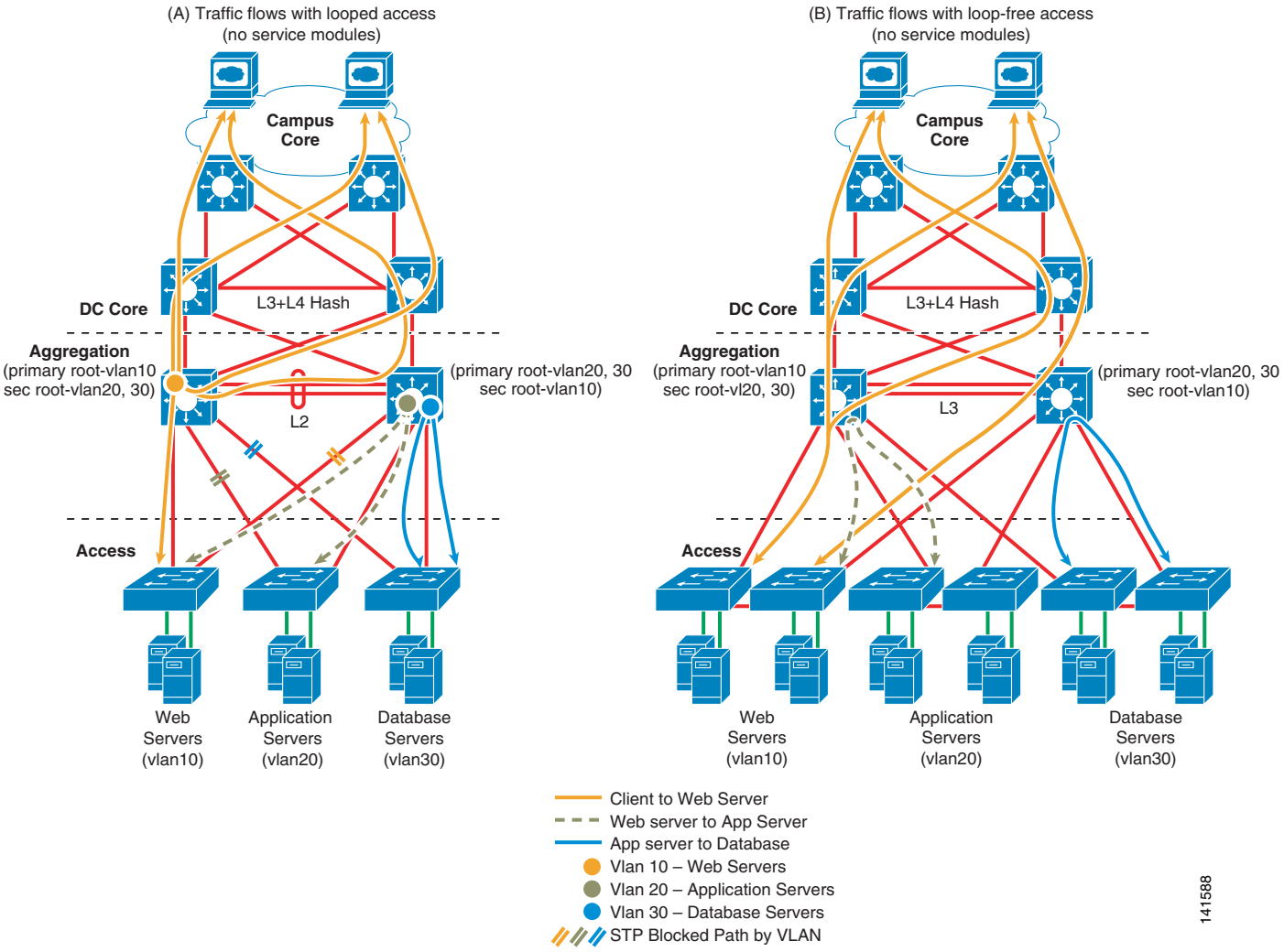
If multiple pairs of service modules are used in an aggregation switch pair, it is possible to distribute active services, which permits both access layer uplinks to be used. However, this is not usually a viable solution because of the additional service modules that are required. Future active-active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

Note The CSM and FWSM-2.x service modules currently operate in active/standby modes. These module pairs require identical configurations. The access layer design must ensure that connection paths remain symmetrical to the active service module. For more information on access layer designs, refer to Chapter 6, "Data Center Access Layer Design." The Cisco Application Control Engine (ACE) is a new module that introduces several enhancements with respect to load balancing and security services. A key difference between the CSM, FWSM Release 2.x, and Cisco ACE is the ability to support active-active contexts across the aggregation module with per context failover.

Server Farm Traffic Flow without Service Modules

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

Figure 2-5 Traffic Flow without Service Modules



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

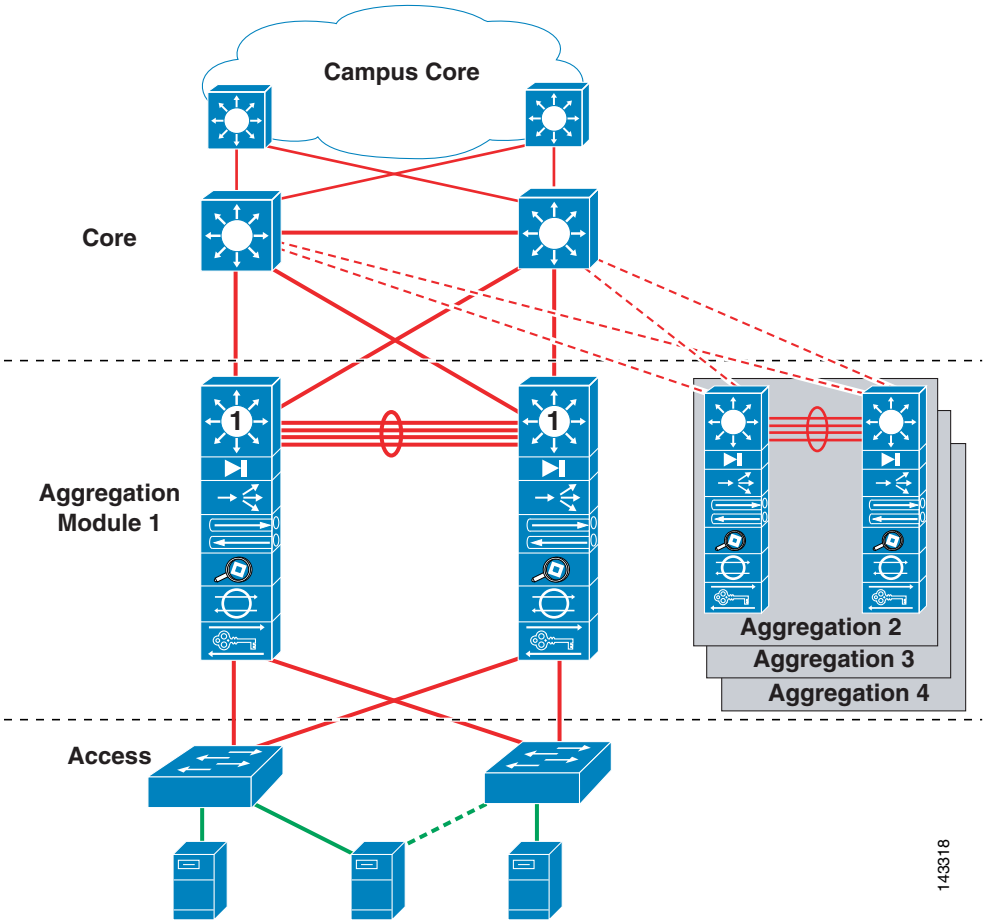
Scaling the Aggregation Layer

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

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

The aggregation layer consists of pairs of interconnected aggregation switches referred to as modules. Figure 2-6 shows a multiple aggregation module design using a common core.

Figure 2-6 Multiple Aggregation Modules



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

Layer 2 Fault Domain Size

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

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

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

Spanning Tree Scalability

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

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

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

Note More details on spanning tree scaling are provided in Chapter 5, "Spanning Tree Scalability."

10 GigE Density

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

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

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

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

Default Gateway Redundancy with HSRP

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

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

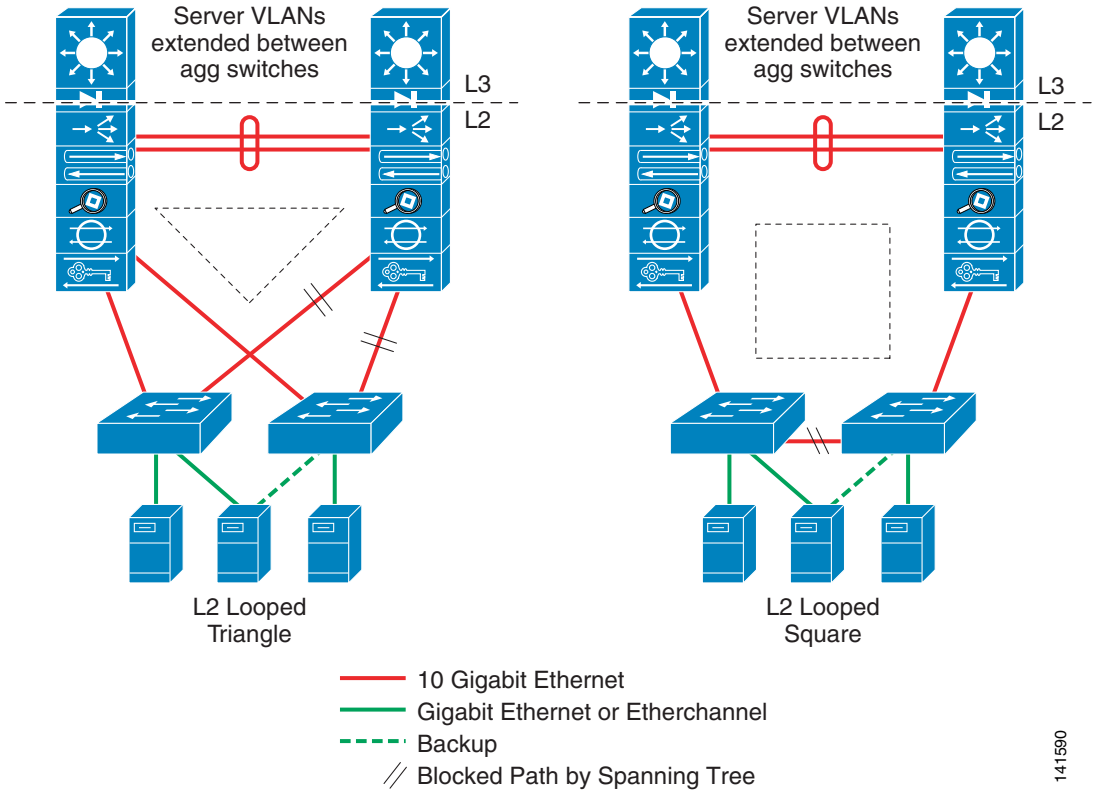
Data Center Access Layer

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

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

Figure 2-7 illustrates the access layer using the Layer 2 looped model in triangle and square loop topologies.

Figure 2-7 Access Layer Looped Topologies

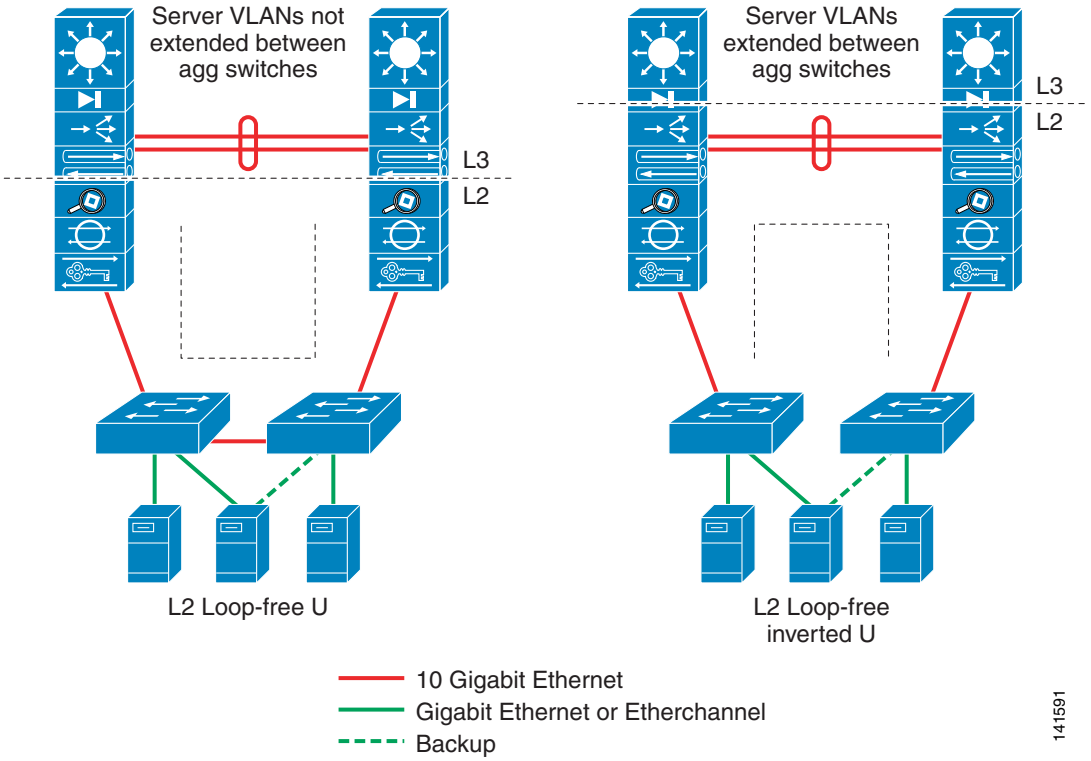


The triangle-based looped topology is the most widely used today. Looped topologies are the most desirable in the data center access layer for the following reasons:

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

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

Figure 2-8 Access Layer Loop-free Topologies



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

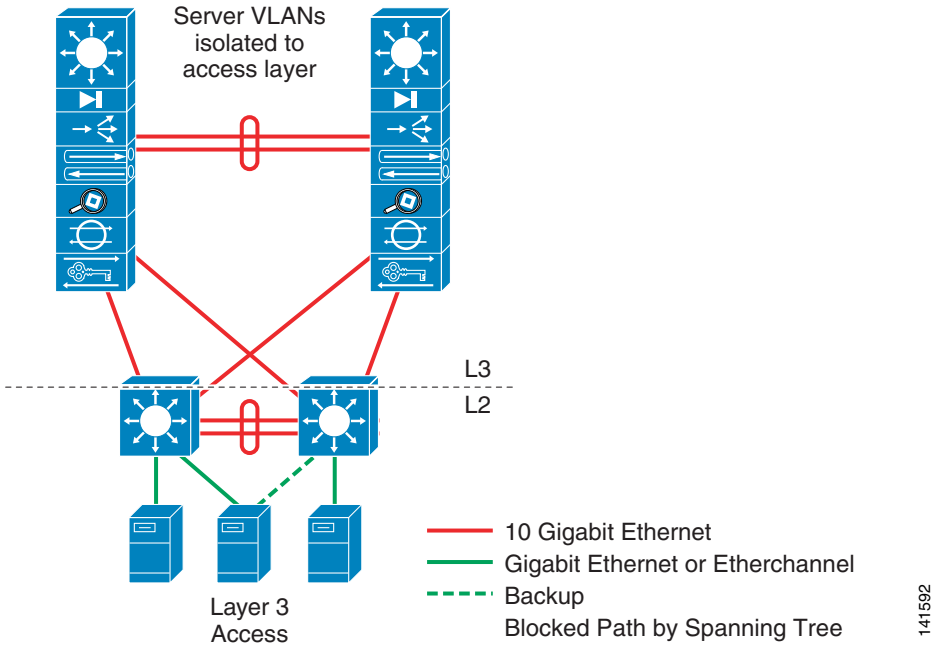
- No blocking on uplinks, all links are active in a loop-free topology
- Layer 2 adjacency for servers is limited to a single pair of access switches in a loop-free topology
- VLAN extension across the data center is not supported in a loop-free U topology but is supported in the inverted U topology.

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

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

Figure 2-9 illustrates the access layer using the Layer 3 model.

Figure 2-9 Access Layer 3 Topology



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

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

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

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

Recommended Platforms and Modules

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

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

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

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

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

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

Distributed Forwarding

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

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

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

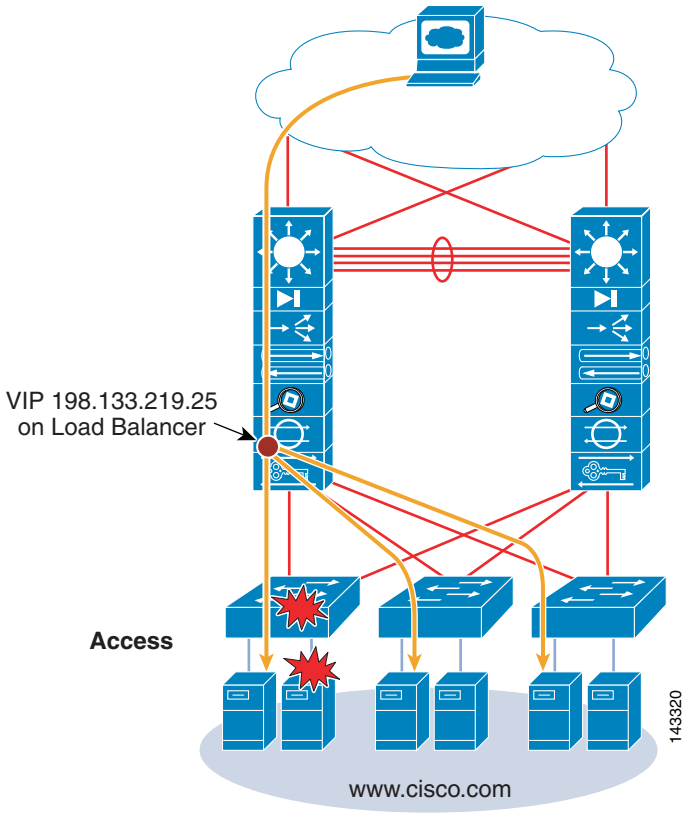
Resiliency

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

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

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

Figure 2-10 Server Redundancy with Load Balancers



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

Sharing Services at the Aggregation Layer

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

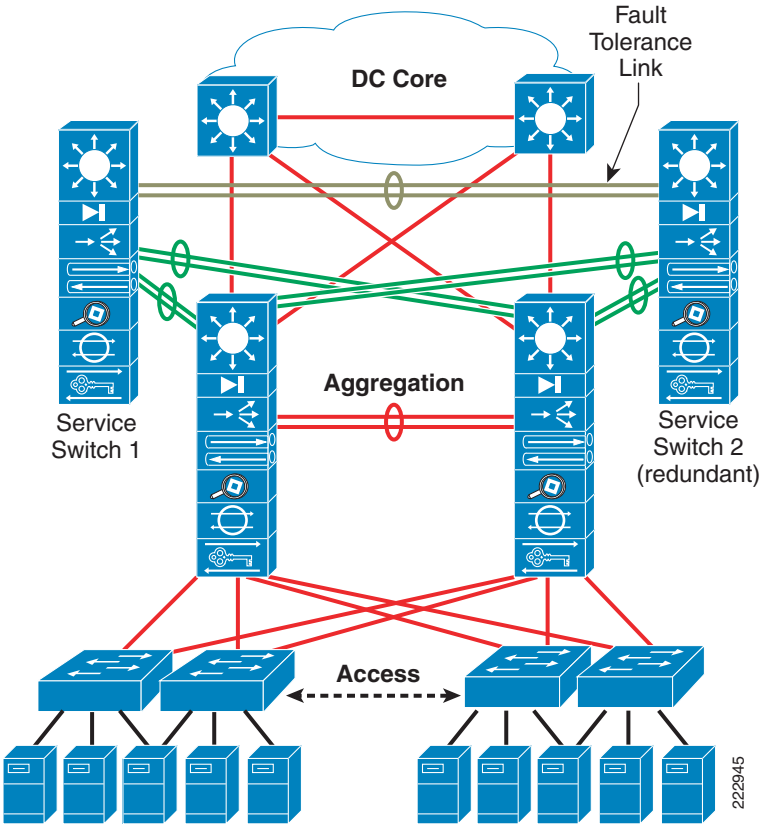
The services in the aggregation layer that can be used by the access layer servers include the following:

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

Data Center Services Layer

The service layer switch provides a method of scaling up services using service modules without using slots in the aggregation layer switch. Moving certain service modules out of the aggregation layer switch increases the number of available slots and improves aggregation layer performance. For example, this is useful when a farm of Cisco ACE and Cisco Firewall Service Modules (FWSM) are required. Figure 2-11 shows a topology with a service layer switch design.

Figure 2-11 Data Center Service Layer Switch



Recommended Platforms and Modules

Typically, theCisco ACE and Cisco FWSM service modules are located in the aggregation switch to provide services to servers across the access layer switches. By locating these service modules in a separate standalone switch connected using 802.1Q trunks, the aggregation layer can support a higher access layer uplink density. This is particularly useful when 10 GigE port density requirements increase at the aggregation layer.

Performance Implications

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

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

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

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

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

Because the CSM is classic bus-based, it must send truncated packet headers to the Sup720 PFC3 to determine packet destination on the backplane. When a single classic bus module exists in the switch, all non-DFC enabled line cards must perform truncated header lookup, which limits the overall system performance.

Table 2-2 provides an overview of switching performance by module type.

Table 2-2 Performance with Classic Bus Modules

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

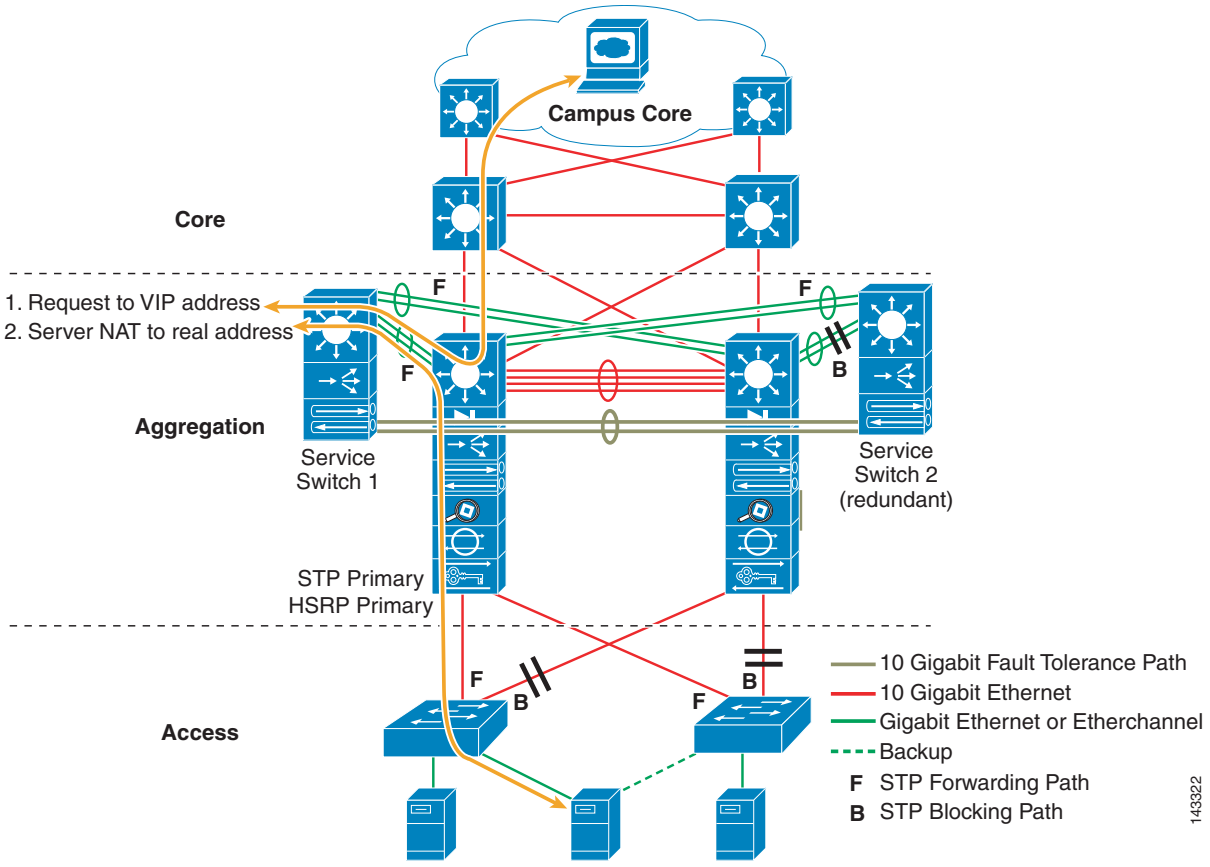
The service switch can be any of the Catalyst 6500 Series platforms that use a Sup720 CPU module. The supervisor engine choice should consider sparing requirements, future migration to next generation modules, performance requirements, and uplink requirements to the aggregation module. For example, if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

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

Traffic Flow through the Service Layer

The service switch is connected to both aggregation switches with 10 GigE links configured as 802.1Q trunks. From a logical perspective, this can be viewed as extending the service module VLANs across an 802.1Q trunk. Figure 2-12 shows the flow of a session using an ACEhg in one-arm mode on a service layer switch. The service switches themselves are interconnected with a Gigabit Ethernet or 10 GigE link to keep the fault tolerant vlans contained to the services switches only.

Figure 2-12 Service Layer Switch Traffic Flow



Note Configuration examples are provided in Chapter 8, "Configuration Reference."

The VLANs used in supporting the Cisco ACE configuration are extended across 802.1Q trunks (GEC or 10GigE) from the aggregation layer switch to each service switch. The Fault Tolerant (FT) VLANs are not extended across these trunks. Spanning tree blocks only a single trunk from the secondary service switch to the Aggregation 2 switch. This configuration provides a forwarding path to the primary service switch from both aggregation switches.

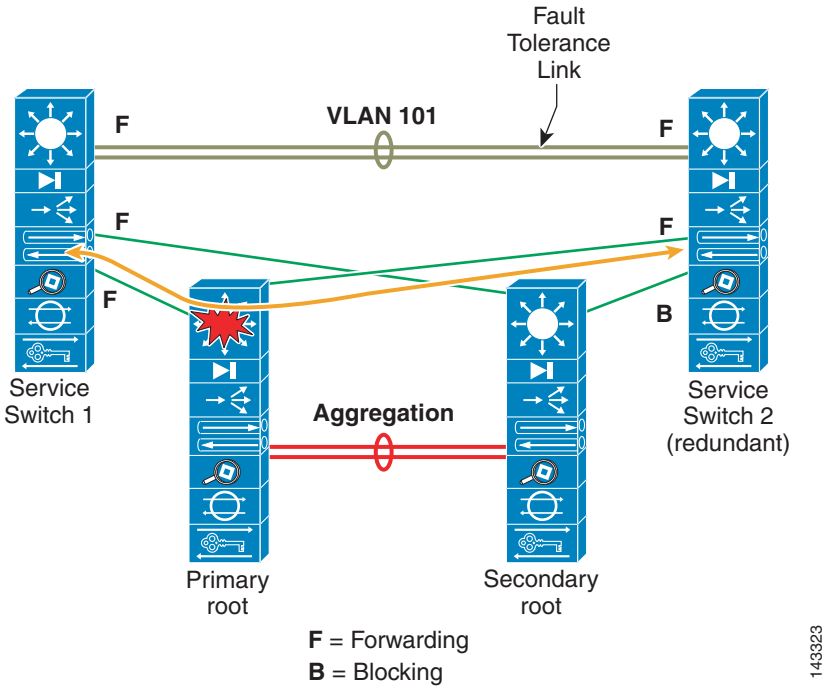
Note The bandwidth used to connect service switches should be carefully considered. The Cisco ACE is capable of supporting up to 16 Gbps of server traffic. The switch interconnect to the aggregation switches should be sized large enough to avoid congestion.

Resiliency

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

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

Figure 2-13 Redundancy with the Service Layer Switch Design



The FT VLAN is used to maintain session state between service modules. The FT VLAN is a direct 802.1Q trunk connection between the service switches. The FT VLANs are not passed across the links connected to and between the aggregation switch pairs.

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

Refer to the Service Module Design With Cisco ACE and Cisco FWSM document at the following URL for details on configuring the Cisco ACE and Cisco FWSM with virtual contexts and creating service chain environments.

http://www.cisco.com/application/pdf/en/us/guest/netsol/ns376/c649/ccmigration_09186a008078de90.pdf