CHAPTER 3

Server Cluster Designs with Ethernet

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

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

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

Technical Objectives

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

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

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

Distributed Forwarding and Latency

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

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

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

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

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

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

Catalyst 6500 System Bandwidth

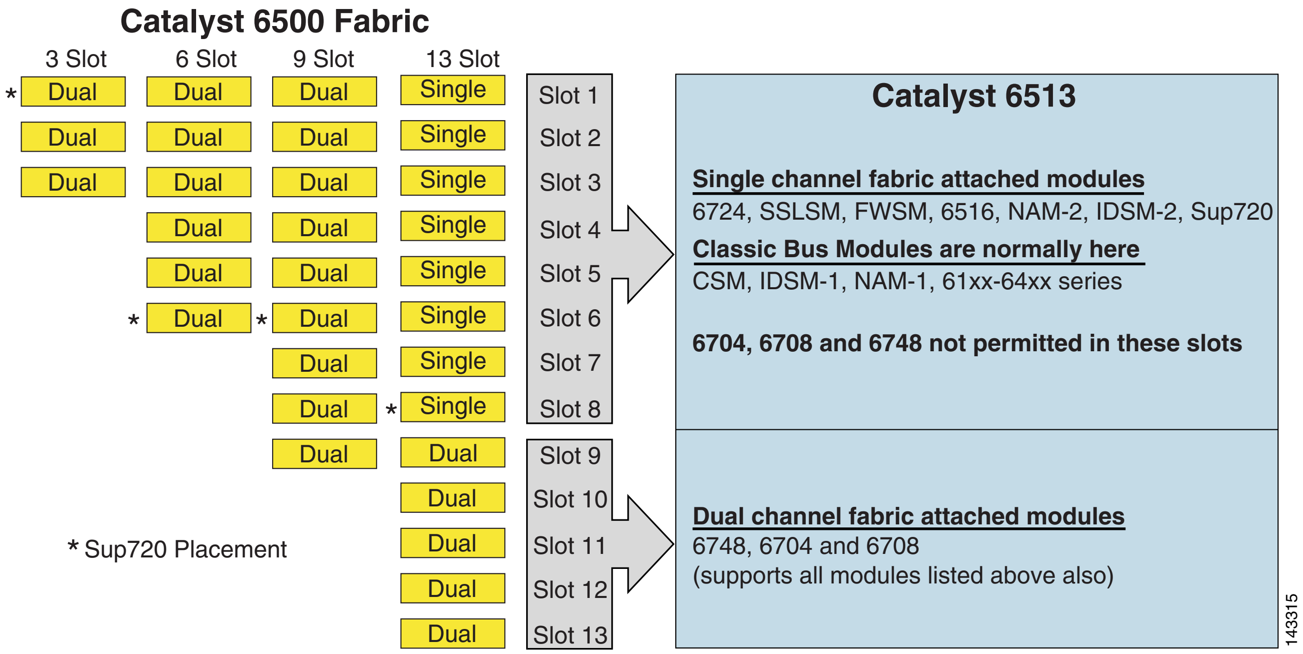
The available system bandwidth does not change when DFCs are used. The DFCs improve the packets per second (pps) processing of the overall system. Table 3-2 summarizes the throughput and bandwidth performance for modules that support DFCs, in addition to the older CEF256 and classic bus modules.

Table 3-2 Performance Comparison with Distributed Forwarding

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

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

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



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

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

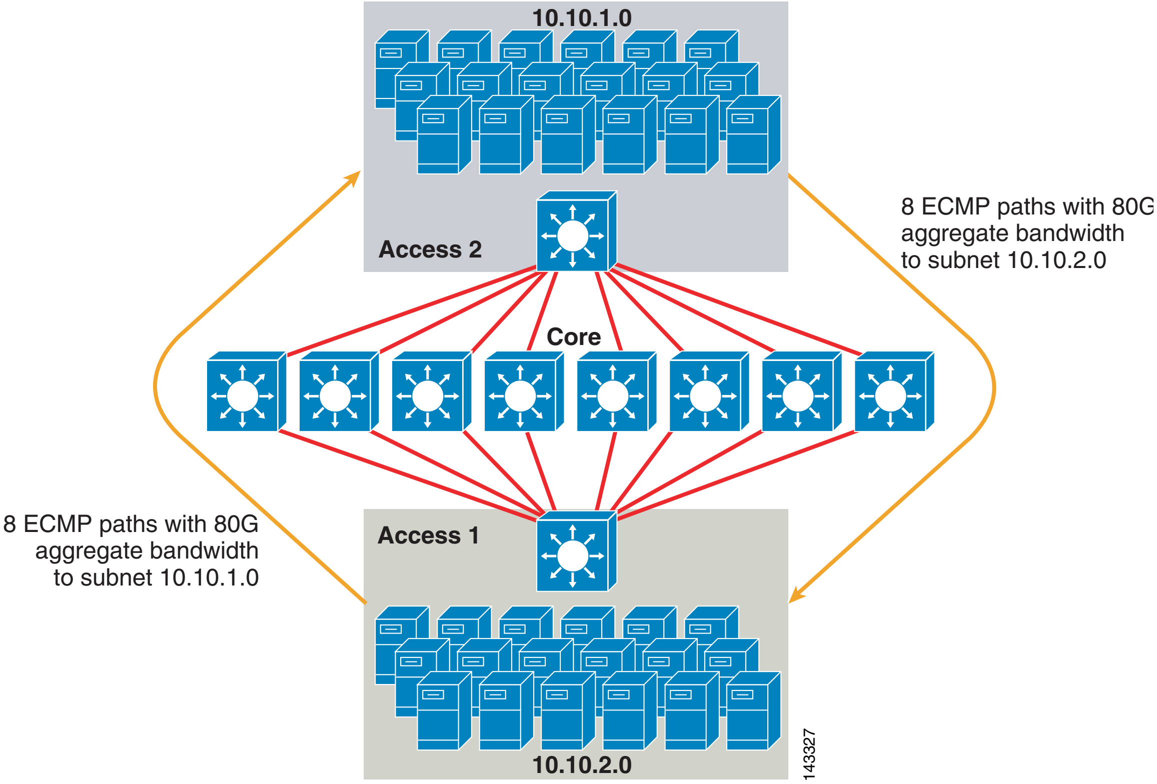
Equal Cost Multi-Path Routing

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

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

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

Figure 3-2 8-Way ECMP Server Cluster Design



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

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

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

Redundancy in the Server Cluster Design

The server cluster design is typically not implemented with redundant CPU or switch fabric processors. Resiliency is typically achieved inherently in the design and by the method the cluster functions as a whole. As described in Chapter 1, "Data Center Architecture Overview," the compute nodes in the cluster are managed by master nodes that are responsible for assigning specific jobs to each compute node and monitoring their performance. If a compute node drops out of the cluster, it reassigns to an available node and continues to operate, although with less processing power, until the node is available. Although it is important to diversify master node connections in the cluster across different access switches, it is not critical for the compute nodes.

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

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

Server Cluster Design—Two-Tier Model

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

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

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

- Maximum number of server GigE connections at scale—Cluster designers typically have an idea of the maximum scale required at initial concept. A benefit of the way ECMP designs function is that they can start with a minimum number of switches and servers that meet a particular bandwidth, latency, and oversubscription requirement, and flexibly grow in a low/non-disruptive manner to maximum scale while maintaining the same bandwidth, latency, and oversubscription values.

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

  4x10GigE Uplinks with 336 servers = 120 Mbps per server

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

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

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

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

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

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

4- and 8-Way ECMP Designs with Modular Access

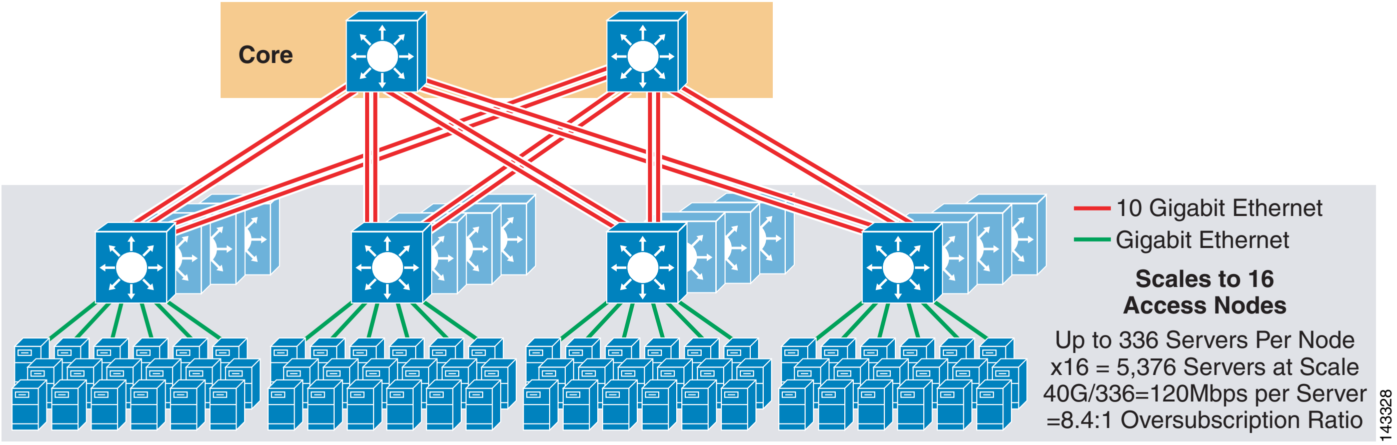
The following four design examples demonstrate various methods of building and scaling the two-tier server cluster model using 4-way and 8-way ECMP. The main issues to consider are the number of core nodes and the maximum number of uplinks, because these directly influence the maximum scale, bandwidth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

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

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

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

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


Core
10 Gigabit Ethernet
Gigabit Ethernet
Scales to 16
Access Nodes
Up to 336 Servers Per Node
x16 = 5,376 Servers at Scale
40G/336=120Mbps per Server
=8.4:1 Oversubscription Ratio
143328

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

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

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

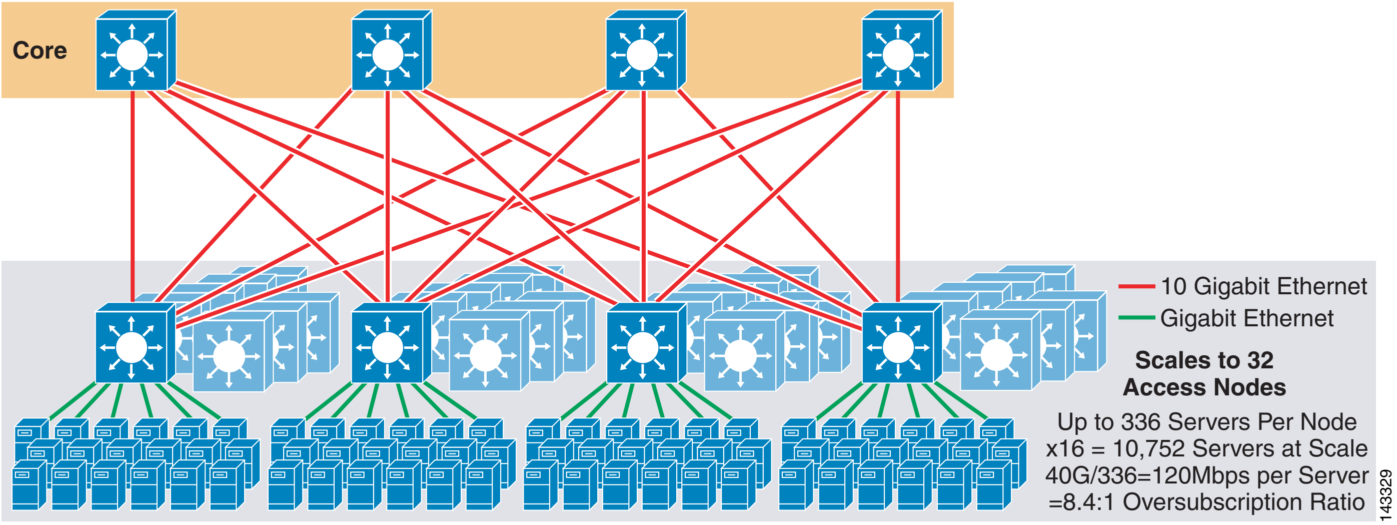
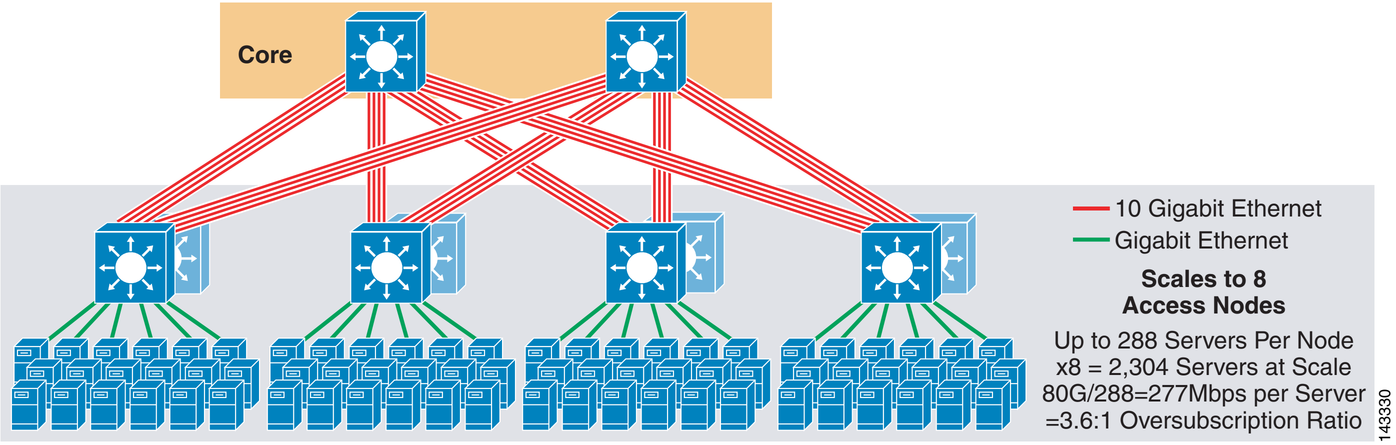

Core
10 Gigabit Ethernet
Gigabit Ethernet
Scales to 32
Access Nodes
Up to 336 Servers Per Node
x16 = 10,752 Servers at Scale
40G/336=120Mbps per Server
=8.4:1 Oversubscription Ratio
143329

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

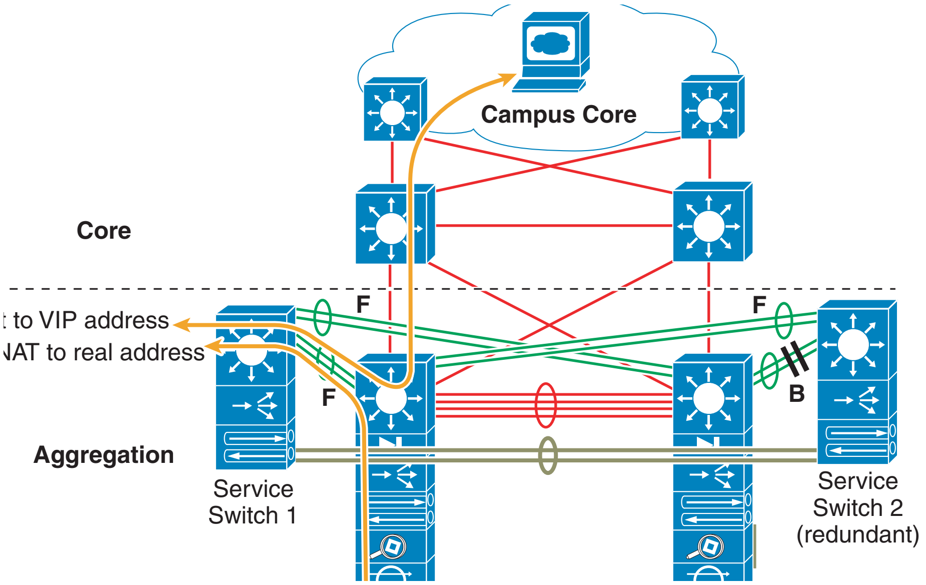
Figure 3-5 8-Way ECMP using Two Core Nodes


Core
10 Gigabit Ethernet
Gigabit Ethernet
Scales to 8
Access Nodes
Up to 288 Servers Per Node
x8 = 2,304 Servers at Scale
80G/288=277Mbps per Server
=3.6:1 Oversubscription Ratio
143330

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

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

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



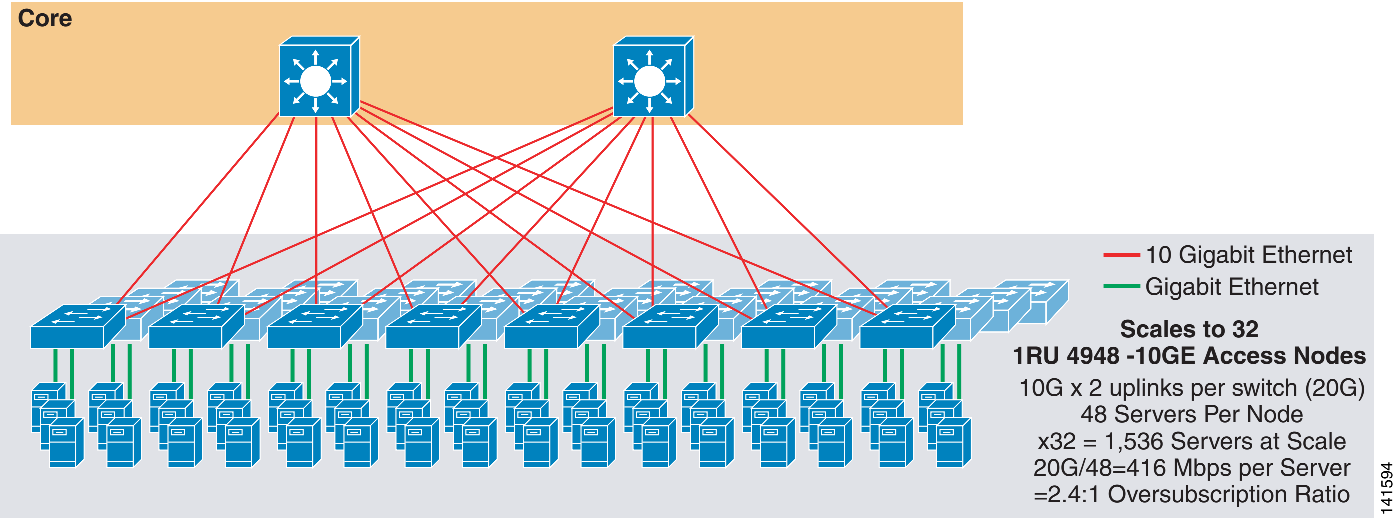
This demonstrates how adding four core nodes to the same previous design can dramatically increase the maximum scale while maintaining the same oversubscription and bandwidth per server values.

2-Way ECMP Design with 1RU Access

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

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

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


Core
🟥 10 Gigabit Ethernet
🟩 Gigabit Ethernet
Scales to 32
1RU 4948 -10GE Access Nodes
10G x 2 uplinks per switch (20G)
48 Servers Per Node
x32 = 1,536 Servers at Scale
20G/48=416 Mbps per Server
=2.4:1 Oversubscription Ratio
141594

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

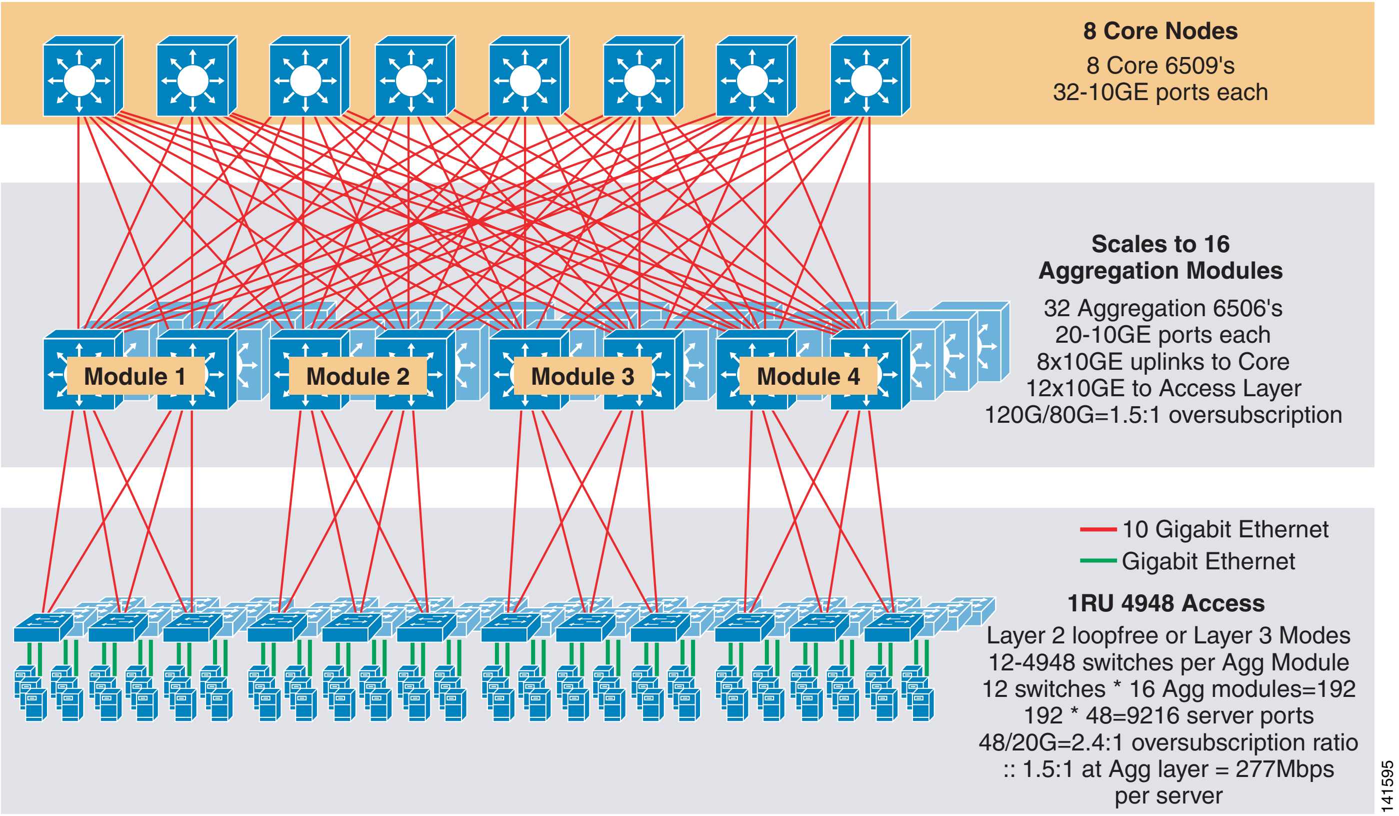
Note More information on rack-based server switching is provided in Chapter 3, “Server Cluster Designs with Ethernet.”

Server Cluster Design—Three-Tier Model

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

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

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



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

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

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

- Centralization of core and aggregation switches—With 1RU switches deployed in the racks, it is possible to centralize the larger core and aggregation modular switches. This can simplify power and cabling infrastructure and improve rack real estate usage.

- Permits Layer 2 loop-free topology—A large cluster network using Layer 3 ECMP access can use a lot of address space on the uplinks and can add complexity to the design. This is particularly important if public address space is used. The three-tier model approach lends itself well to a Layer 2 loop-free access topology that reduces the number of subnets required.

When a Layer 2 loop-free model is used, it is important to use a redundant default gateway protocol such as HSRP or GLBP to eliminate a single point of failure if an aggregation node fails. In this design, the aggregation modules are not interconnected, permitting a loop-free Layer 2 design that can leverage GLBP for automatic server default gateway load balancing. GLBP automatically distributes the servers default gateway assignment between the two nodes in the aggregation module. After a packet arrives at the aggregation layer, it is balanced across the core using the 8-way ECMP fabric. Although GLBP does not provide a Layer 3/Layer 4 load distribution hash similar to CEF, it is an alternative that can be used with a Layer 2 access topology.

Calculating Oversubscription

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

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

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

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

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

1.5*2.4 = 3.6:1

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

a:b = c:d

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

416:1 = x:1.5

x=~264 Mbps per server

Recommended Hardware and Modules

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

The following are recommended:

- Sup720—The Sup720 can consist of both PFC3A (default) or the newer PFC3B type daughter cards.
- Line cards—All line cards should be 6700 Series and should all be enabled for distributed forwarding with the DFC3A or DFC3B daughter cards.

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

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