CHAPTER 4

Data Center Design Considerations

This chapter describes factors that influence the enterprise data center design. The following topics are included:

Factors that Influence Scalability

Determining scalability is never an easy task because there are always unknown factors and inter-dependencies. This section examines some of the most common scalability-related questions that arise when designing a data center network.

Why Implement a Data Center Core Layer?

Do I need a separate core layer for the data center? Can I use my existing campus core?

The campus core can be used as the data center core. The recommendation is to consider the long-term requirements so that a data center core does not have to be introduced at a later date. Advance planning helps avoid disruption to the data center environment. Consider the following items when determining the right core solution:

- 10GigE density—Will there be enough 10GigE ports on the core switch pair to support both the campus distribution as well as the data center aggregation modules?
- Administrative domains and policies—Separate cores help to isolate campus distribution layers from data center aggregation layers in terms of troubleshooting, administration, and policies (QoS, ACLs, troubleshooting, and maintenance).
- Future anticipation—The impact that can result from implementing a separate data center core layer at a later date might make it worthwhile to install it at the beginning.

Why Use the Three-Tier Data Center Design?

Why not connect servers directly to a distribution layer and avoid installing an access layer?

The three-tier approach consisting of the access, aggregation, and core layers permit flexibility in the following areas:

- Layer 2 domain sizing— When there is a requirement to extend a VLAN from one switch to another, the domain size is determined at the distribution layer. If the access layer is absent, the Layer 2 domain must be configured across the core for extension to occur. Extending Layer 2 through a core causes path blocking by spanning tree and has the risk of uncontrollable broadcast issues related to extending Layer 2 domains, and therefore should be avoided.
- Service modules—An aggregation plus access layer solution enables services to be shared across the entire access layer of switches. This lowers TCO and lowers complexity by reducing the number of components to configure and manage. Consider future service capabilities that include Application-Oriented Networking (AON), ACE, and others.
- Mix of access layer models—The three-tier approach permits a mix of both Layer 2 and Layer 3 access models with 1RU and modular platforms, permitting a more flexible solution and allowing application environments to be optimally positioned.
- NIC teaming and HA clustering support—Supporting NIC teaming with switch fault tolerance and high availability clustering requires Layer 2 adjacency between NIC cards, resulting in Layer 2 VLAN extension between switches. This would also require extending the Layer 2 domain through the core, which is not recommended.

Why Deploy Services Switch?

When would I deploy Services Switch instead of just putting Services Modules in the Aggregation Switch?

Incorporating Services Switch into the data center design is desirable for the following reasons:

- Large Aggregation Layer—If services are deployed in the aggregation layer, as this section scales it may become burdensome to continue to deploy services in every aggregation switch. The service switch allows for services to be consolidated and applied to all the aggregation layer switches without the need to physically deploy service cards across the entire aggregation layer. Another benefit is that it allows the aggregation layer to scale to much larger port densities since slots used by service modules are now able to be deployed with LAN interfaces.
- Mix of service modules and appliances—Data center operators may have numerous service modules and applicances to deploy in the data center. By using the service chassis model you can deploy all of the services in a central fashion, allowing the entire data center to use the services instead of having to deploy multiple appliances and modules across the facility.
- Operational or process simplification—Using the services switch design allows for the core, aggregation, and access layers to be more tightly controlled from a process change prespective. Security, load balancing, and other services can be configured in a central fashion and then applied across the data center without the need to provide numerous access points for the people operating those actual services.
- Support for network virtualization—As the network outside of the data center becomes more virtualized it may be advantageous to have the services chassis become the point where things like VRF-aware services are applied without impacting the overall traffic patterns in the data center.

Determining Maximum Servers

What is the maximum number of servers that should be on an access layer switch? What is the maximum number of servers to an aggregation module?

The answer is usually based on considering a combination of oversubscription, failure domain sizing, and port density. No two data centers are alike when these aspects are combined. The right answer for a particular data center design can be determined by examining the following areas:

- Oversubscription—Applications require varying oversubscription levels. For example, the web servers in a multi-tier design can be optimized at a 15:1 ratio, application servers at 6:1, and database servers at 4:1. An oversubscription ratio model helps to determine the maximum number of servers that should be placed on a particular access switch and whether the uplink should be Gigabit EtherChannel or 10GE. It is important for the customer to determine what the oversubscription ratio should be for each application environment. The following are some of the many variables that must be considered when determining oversubscription:
  - NIC—Interface speed, bus interface (PCI, PCI-X, PCI-E)
  - Server platform—Single or dual processors, offload engines

  - Application characteristics—Traffic flows, inter-process communications
  - Usage characteristics—Number of clients, transaction rate, load balancing
- Failure domain sizing—This is a business decision and should be determined regardless of the level of resiliency that is designed into the network. This value is not determined based on MTBF/MTTR values and is not meant to be a reflection of the robustness of a particular solution. No network design should be considered immune to failure because there are many uncontrollable circumstances to consider, including human error and natural events. The following areas of failure domain sizing should be considered:
  - Maximum number of servers per Layer 2 broadcast domain
  - Maximum number of servers per access switch (if single-homed)
  - Maximum number of servers per aggregation module
  - Maximum number of access switches per aggregation module
- Port density—The aggregation layer has a finite number of 10GigE ports that can be supported, which limits the quantity of access switches that can be supported. When a Catalyst 6500 modular access layer is used, thousands of servers can be supported on a single aggregation module pair. In contrast, if a 1RU Catalyst 4948 is used at the access layer, the number of servers supported is less. Cisco recommends leaving space in the aggregation layer for growth or changes in design.

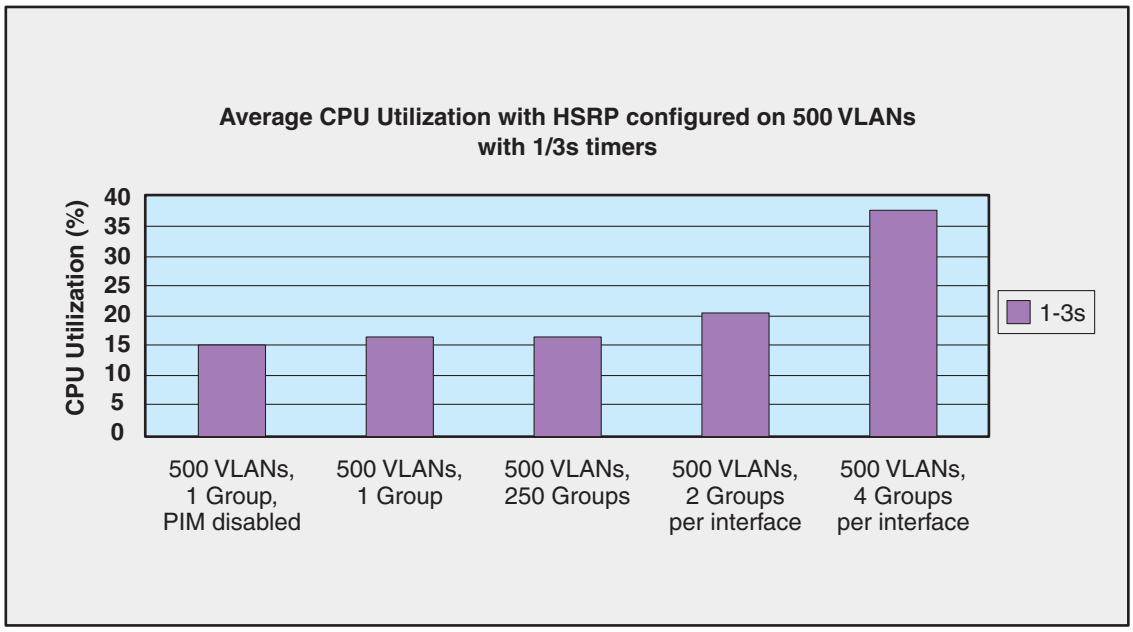
  The data center, unlike other network areas, should be designed to have flexibility in terms of emerging services such as firewalls, SSL offload, server load balancing, AON, and future possibilities. These services will most likely require slots in the aggregation layer, which would limit the amount of 10GigE port density available.

Determining Maximum Number of VLANs

What is the maximum number of VLANs that can be supported in an aggregation module?

- Spanning tree processing—When a Layer 2 looped access topology is used, which is the most common, the amount of spanning tree processing at the aggregation layer needs to be considered. There are specific watermarks related to the maximum number of system-wide active logical instances and virtual port instances per line card that, if reached, can adversely affect convergence and system stability. These values are mostly influenced by the total number of access layer uplinks and the total number of VLANs. If a data center-wide VLAN approach is used (no manual pruning on links), the watermark maximum values can be reached fairly quickly. More details and recommendations are provided in Chapter 5, “Spanning Tree Scalability.”
- Default Gateway Redundancy Protocol— The quantity of HSRP instances configured at the aggregation layer is usually equal to the number of VLANs. As Layer 2 adjacency requirements continue to gain importance in data center design, proper consideration for the maximum HSRP instances combined with other CPU-driven features (such as GRE, SNMP, and others) have to be considered. Lab testing has shown that up to 500 HSRP instances can be supported in an aggregation module, but close attention to other CPU driven features must be considered. The graph in Figure 4-1 shows test results when using 500 VLANs with one or multiple groups with the hello and holddown timer configuration at 1/3 seconds.

Figure 4-1 Graphed Average CPU Utilization with HSRP

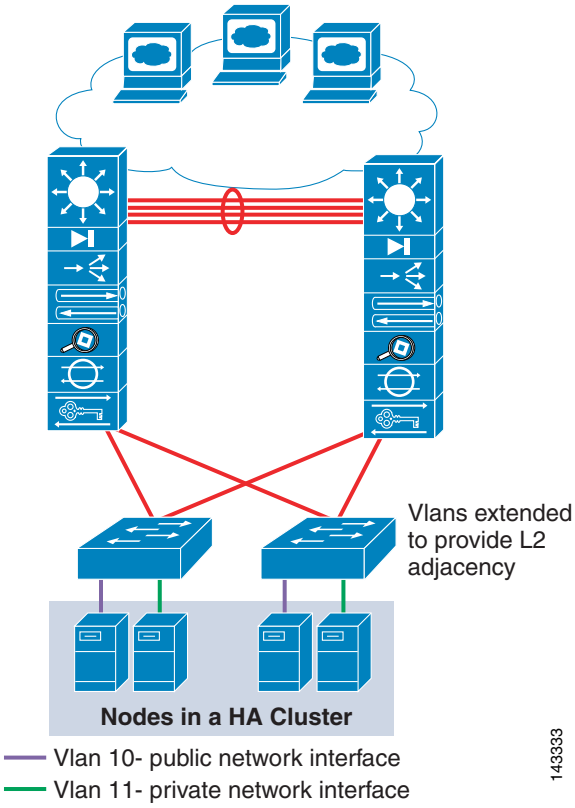


Server Clustering

The goal of server clustering is to combine multiple servers so that they appear as a single unified system through special software and network interconnects**.** "Clusters" were initially used with the Digital Equipment Corporation VAX VMS Clusters in the late 1980s. Today, "clustering" is a more general term that is used to describe a particular type of grouped server arrangement that falls into the following four main categories:

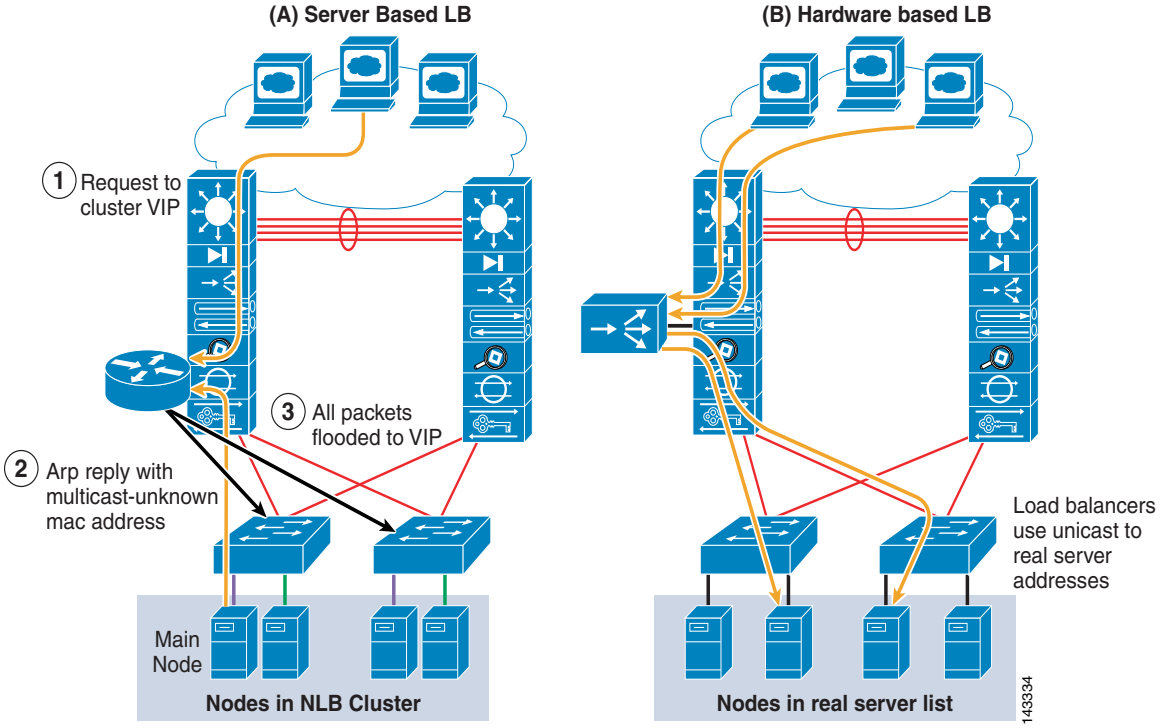
- High availability clusters—This type of cluster uses two or more servers and provides redundancy in the case of a server failure (see Figure 4-2). If one node fails, another node in the cluster takes over with little or no disruption. This type of cluster is usually up to a maximum of eight nodes and requires Layer 2 adjacency between their public and private interfaces. High availability clusters are common in the data center multi-tier model design**.**

Figure 4-2 High Availability Cluster



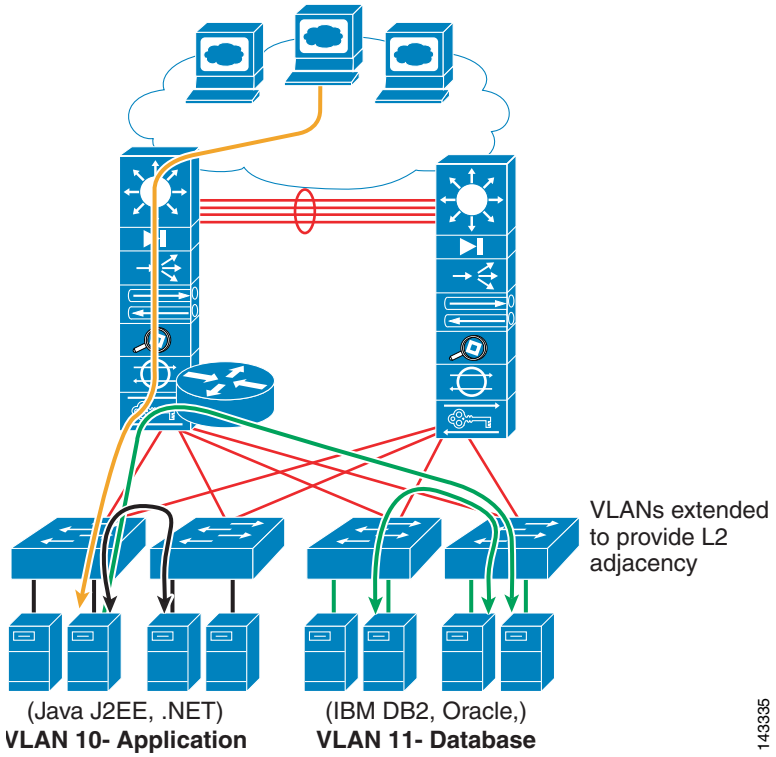
- Network load balanced clusters (NLBs)—This type of cluster typically supports up to a maximum of 32 servers that work together to load balance HTTP sessions on a website. It uses a broadcast mechanism in which the ARP reply from the main server to the gateway router is an unknown MAC address, so that all packets to the destination web site address are essentially broadcast to all servers in the VLAN. This type of implementation is usually much less robust than hardware-based load balancing solutions, and requires Layer 2 adjacency. Hardware-based server load balancers such as the Cisco CSM provide a unicast-based solution, scale beyond 32 servers, and provide many value-added features. NLB clusters are common in the data center multi-tier model design. Figure 4-3 illustrates both a server based load balancing solution and a hardware-based load balancing solution.

Figure 4-3 Network Load Balanced Clusters



- Database clusters—As databases become larger, the ability to search the database becomes more complex and time sensitive. Database clusters provide a way to enable efficient parallel scans and improve database lock times. Some examples of parallel database implementations are Oracle RAC and IBM DB2. These implementations also require Layer 2 adjacency between the servers. Database clusters are typically two to eight nodes in size and are common in the data center multi-tier model design. Figure 4-4 illustrates how the application and database layers communicate across the aggregation layer router and how the interfaces require Layer 2 adjacency within each layer, resulting in VLANs being extended across multiple access layer switches.

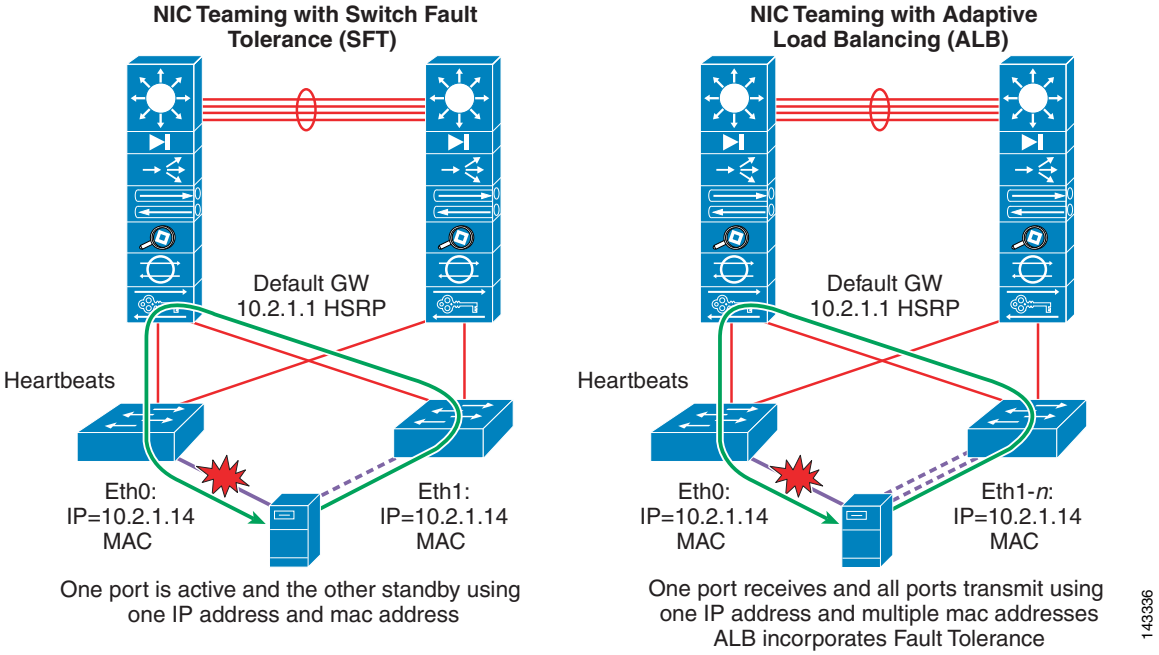
Figure 4-4 Database Clusters



NIC Teaming

Servers with a single Network Interface Card (NIC) interface can have many single points of failure. The NIC card, the cable, and the switch to which it connects are all single points of failure. NIC teaming is a solution developed by NIC card vendors to eliminate this single point of failure by providing special drivers that allow two NIC cards to be connected to two different access switches or different line cards on the same access switch. If one NIC card fails, the secondary NIC card assumes the IP address of the server and takes over operation without disruption. The various types of NIC teaming solutions include active/standby and active/active. All solutions require the NIC cards to have Layer 2 adjacency with each other. NIC teaming solutions are common in the data center multi-tier model design and are shown in Figure 4-5.

Figure 4-5 NIC Teaming Configurations



Note the following:

- Switch fault tolerance (SFT)—With SFT designs, one port is active and the other standby using one common IP address and MAC address.
- Adaptive load balancing (ALB) —With ALB designs, one port receives and all ports transmit using one IP address and multiple MAC addresses.

Pervasive 10GigE

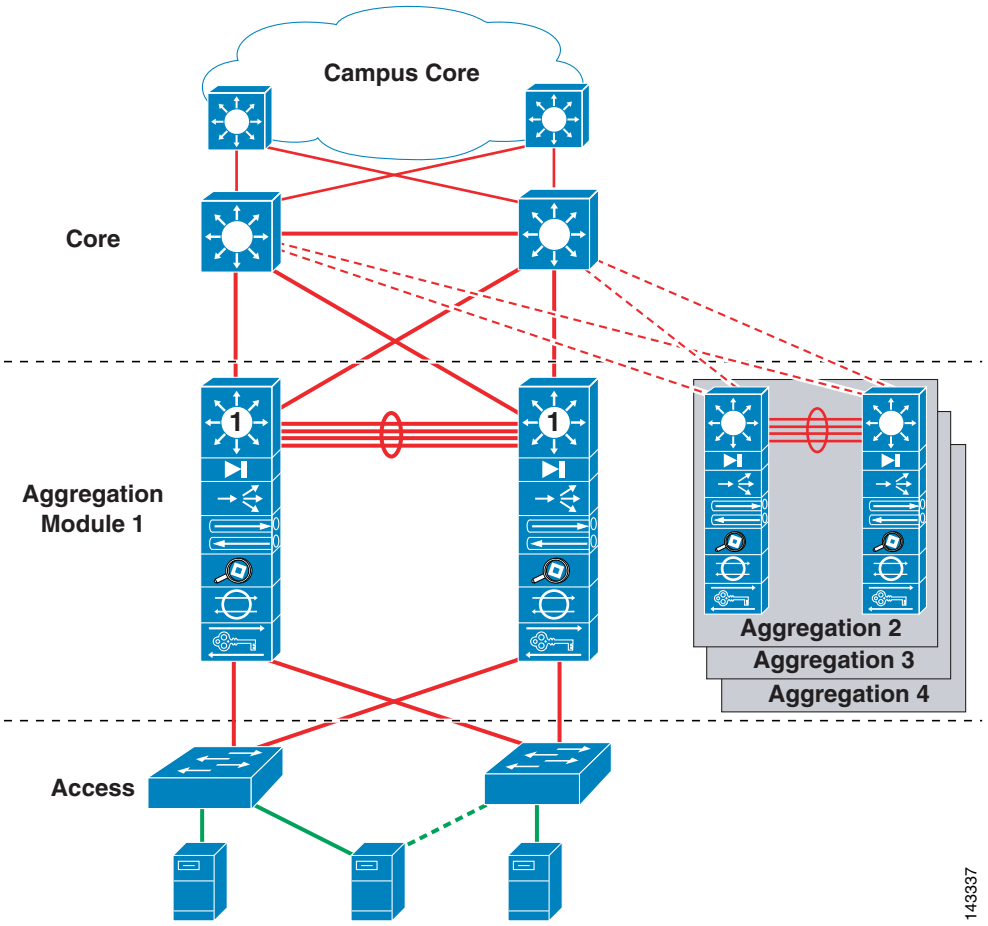
Customers are seeing the benefit of moving beyond Gigabit or Gigabit EtherChannel implementations to 10GigE, which includes benefits such as the following:

- Improving IP-based storage access (iSCSI)
- Improving network use of SMP-based servers including virtual machine implementations
- Improving access layer uplink use because of Gigabit EtherChannel hashing algorithm barriers
- Improving server backup and recovery times
- Improving NAS performance

Many customers are also moving to 10GigE access layer uplinks in anticipation of future requirements. The implications relative to this trend are usually related to density in the aggregation layer.

A proven method of increasing aggregation layer 10GigE ports is to use multiple aggregation modules, as shown in Figure 4-6.

Figure 4-6 Multiple Aggregation Modules



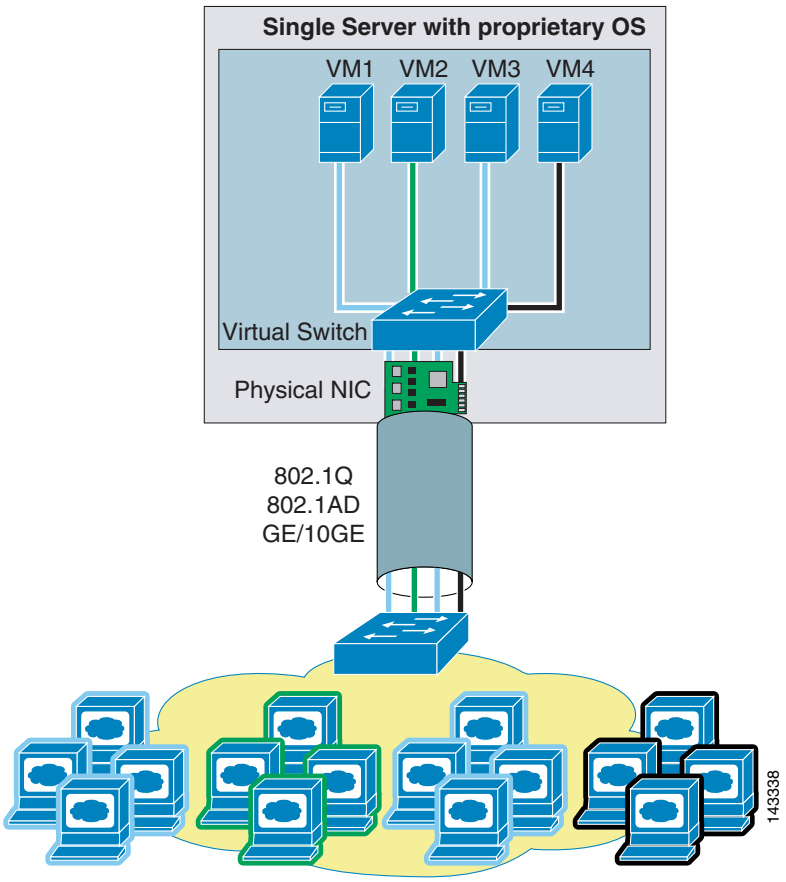
Aggregation modules provide a way to scale 10GigE port requirements while also distributing CPU processing for spanning tree and HSRP. Other methods to increase port density include using the service layer switch, which moves service modules out of the aggregation layer and into a standalone chassis, making slots available for 10GigE ports. Other methods of improving 10GE density are in the access layer design topology used such as a looped square topology. More on this subject is covered in Chapter 2, "Data Center Multi-Tier Model Design," and Chapter 6, "Data Center Access Layer Design."

Server Consolidation

The majority of servers in the data center are underutilized in terms of CPU and memory; particularly the web server tier and development environment server resources. Virtual machine solutions are being used to solve this deficiency and to improve the use of server resources.

- The virtual machine solution is a vendor software product that can install multiple server images on a single hardware server platform to make it appear the same as multiple, separate physical servers. This was initially seen in development or lab environments but is now a production solution in the enterprise data center. The virtual machine solution is supported on small single processor server platforms to large SMP platforms with greater memory support and multi-processors. This software-based solution allows over 32 virtual machines to coexist on the same physical server. Figure 4-7 shows a server with multiple virtual machine instances running on it.

Figure 4-7 Multiple Virtual Machines



Virtual machine solutions can be attached to the network with multiple GE network interfaces, one for each virtual host implementation. Other implementations include the use of 802.1Q on GE or 10GE interfaces to connect virtual hosts directly to a specific VLAN over a single interface. Although 10GE interfaces are not supported on all virtual machine solutions available today, it is expected that this will change in the near future. This requirement could have implications on access layer platform selection, NIC teaming support, and in determining uplink oversubscription values.

Top of Rack Switching

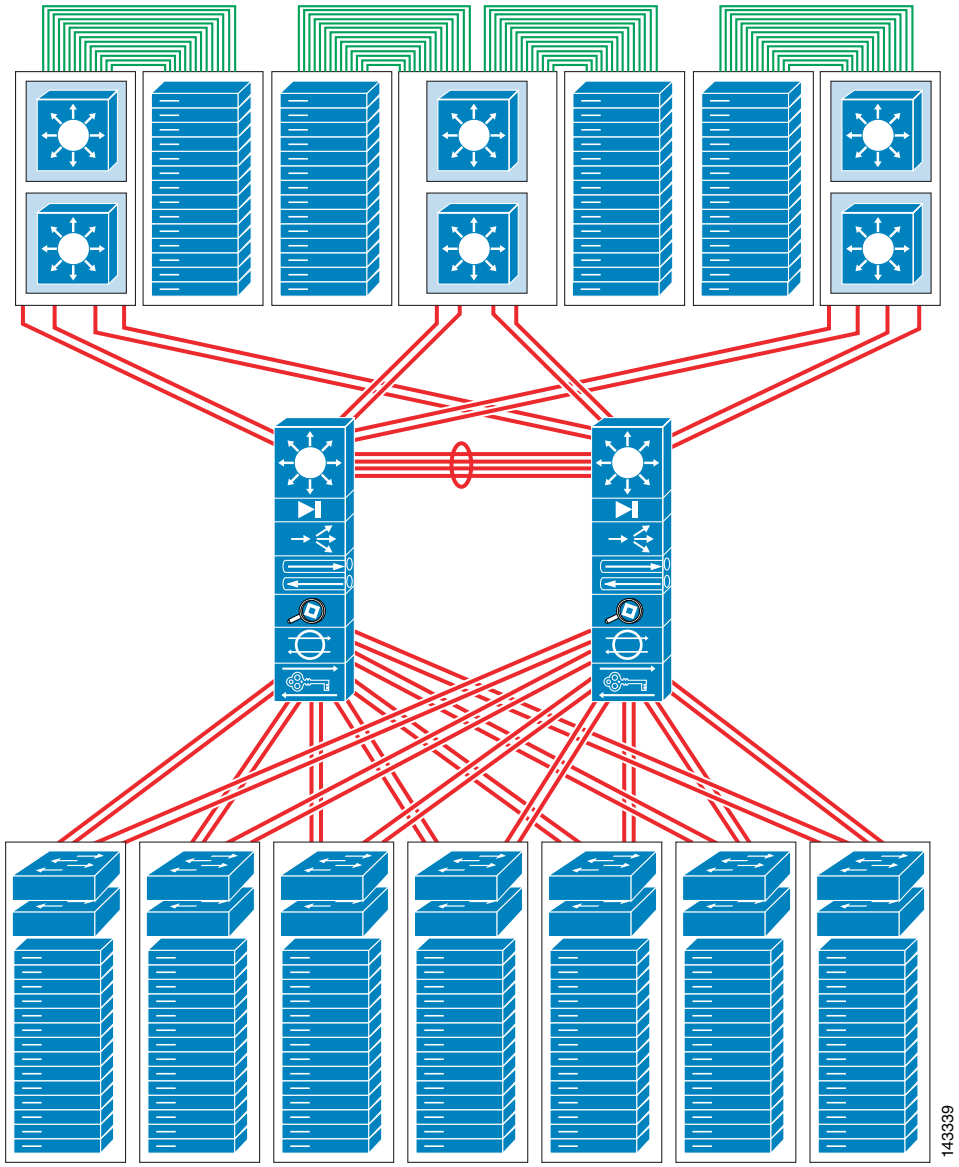
The most common access layer topology in the enterprise today is based on the modular chassis Catalyst 6500 or 4500 Series platforms. This method has proven to be a very scalable method of building out server farms that are providing high density, high speed uplinks, and redundant power and processors. Although this approach has been very successful, it has certain challenges related to the environments of data centers. The enterprise data center is experiencing a large amount of growth in the sheer number of servers while at the same time server density has been improved with 1RU and blade server solutions. Three particular challenges that result from this trend are related to the following:

- Cable bulk—There are typically three to four interfaces connected on a server. With a higher density of servers per rack, cable routing and management can become quite difficult to manage and maintain.

- Power—The increased density of components in the rack is driving a need for a larger power feed to the rack. Many data centers do not have the power capacity at the server rows to support this increase.
- Cooling—The amount of cables laying under the raised floor and the cable bulk at the cabinet base entry is blocking necessary airflow required to cool equipment in the racks. At the same time, the servers in the rack are requiring more cooling volume because of their higher density.

These challenges have forced customers to find alternative solutions by spacing out cabinets, modifying cable routes, or other means, including to not deploy high density server solutions. Another way that customers are seeking to solve some of these problems is by using a rack-based switching solution. By using 1RU top of rack switches, the server interface cables are kept in the cabinet, reducing the amount of cabling in the floor and thus reducing the cabling and cooling issues. Figure 4-8 shows both a modular (top) and rack-based (bottom) access layer approach.

Figure 4-8 Modular and 1RU Access Layers



The upper half of Figure 4-8 has the following characteristics:

- Less devices to manage
- Increased cabling and cooling challenges
- Lower spanning tree processing

The lower half of Figure 4-8 has the following characteristics:

- More devices to manage
- Less cabling and cooling challenges
- Higher spanning tree processing
- Uplink density challenge at the aggregation layer

When considering a 1RU top of rack switch implementation, the following should be considered:

- 10GigE density—The increase in uplinks requires higher 10GigE density at the aggregation layer or additional aggregation modules.
- Spanning tree virtual and logical ports—For Layer 2 looped access layer topologies, the increase in uplinks increases the STP active logical and virtual port per line card instances at the aggregation layer, creating more overhead and processing requirements.
- How many 1RU switches per rack?—The maximum number of ports that might need to be connected in a worst case scenario could create a need for three, four, or more 1RU switches in the rack. This has obvious cost issues and further impacts 10GigE density and STP overhead.
- Management—More switches mean more elements to manage, adding complexity.

Blade Servers

Blade-server chasses have become very popular in the enterprise data center, driven mostly by the IBM BladeCenter, HP BladeServer, and Dell Blade Server products. Although the blade server seeks to reduce equipment footprint, improve integration, and improve management, it has the following specific challenges related to designing and supporting the data center network:

- Administrative domains—Blade server products can support either integrated switches or pass-through modules for connecting its servers to the network. Who is responsible for configuring and managing these integral switches? Usually the system administration team is responsible for the components inside of a server product. So, who configures spanning tree? How should the trunks be configured? How are change control and troubleshooting supported? It is important for customers to address these questions before implementation.
- Interoperability—Blade servers support many different vendor-integral switches, including Cisco, Nortel, and D-Link, to name a few. Although many of the technologies in use are expected to meet interoperability standards such as spanning tree 802.1w, they must be verified and tested to ensure proper operation.
- Spanning tree scaling—The integral switch on the blade server is logically similar to the external rack-based server switching design. The same challenges apply relative to the increase in spanning tree logical/virtual ports.
- Pass-through cabling—The pass-through module option on blade servers permits customers to use their existing external access switches for connecting the servers in the blade server chassis and to avoid the integral switch option. Customers should examine the pass-through cabling system to make sure it can properly be supported in their cabinets.
- Topologies—Each vendor blade server implementation has unique internal and external switch trunk connectivity options. Careful consideration should be taken in determining the proper access layer topology that meets the requirements such as VLAN extension and NIC teaming while staying within the watermark values of spanning tree design.

Importance of Team Planning

Considering the roles of different personnel in an IT organization shows that there is a growing need for team planning with data center design efforts. The following topics demonstrate some of the challenges that the various groups in an IT organization have related to supporting a "business ready" data center environment:

- System administrators usually do not consider physical server placement or cabling to be an issue in providing application solutions. When the need arises for one server to be connected into the same VLAN as other servers, it is usually expected to simply happen without thought or concern about possible implications. The system administrators are faced with the challenge of being business-ready and must be able to deploy new or to scale existing applications in a timely fashion.
- Network administrators have traditionally complied with these requests by extending the VLAN across the Layer 2 looped topology and supporting the server deployment request. This is the flexibility of having a Layer 2 looped access layer topology, but is becoming more of a challenge now than it was in the past. The Layer 2 domain diameters are getting larger, and now the network administrator is concerned with maintaining spanning tree virtual/logical port counts, manageability, and the failure exposure that exists with a large Layer 2 broadcast domain. Network designers are faced with imposing restrictions on server geography in an effort to maintain spanning tree processing, as well as changing design methods to include consideration for Layer 2 domain sizing and maximum failure domain sizing.
- Facilities administrators are very busy trying to keep all this new dense hardware from literally burning up. They also see the additional cabling as very difficult if not impossible to install and support with current design methods. The blocked air passages from the cable bulk can create serious cooling issues, and they are trying to find ways to route cool air into hot areas. This is driving the facilities administrators to look for solutions to keep cables minimized, such as when using 1RU switches. They are also looking at ways to locate equipment so that it can be cooled properly.

These are all distinct but related issues that are growing in the enterprise data center and are creating the need for a more integrated team planning approach. If communication takes place at the start, many of the issues are addressed, expectations are set, and the requirements are understood across all groups.