Enterprise Data Center Design Considerations

This section describes general Enterprise Data Center design considerations and provides an overview of the general technologies and models used in an Enterprise Data Center.

The Enterprise Data Center

This section describes technology and trends influencing the Enterprise Data Center. For large enterprises with a significant number of servers, a dedicated Enterprise Data Center provides employees, partners, and customers with access to data and resources to effectively work, collaborate, and interact. Historically, most Enterprise Data Centers grew rapidly as organizational requirements expanded. Applications were implemented as needed, often resulting in underutilized, isolated infrastructure silos. Each silo was designed based on the specific application being deployed, so a typical data center supported a broad assortment of operating systems, computing platforms, and storage systems, resulting in various application "islands" that were difficult to change or expand and expensive to manage, integrate, secure, and back up.

This *server*-centric data center model is evolving to a *service*-centric model, as illustrated in Figure 4-19. This evolution includes the following:

- The deployment of virtual machine software, such as VMware and Xen, which breaks the one-to-one relationship between applications and the server hardware and operating system on which they run. Virtual machine software allows multiple applications to run on a single server, independent of each other and of the underlying operating system.

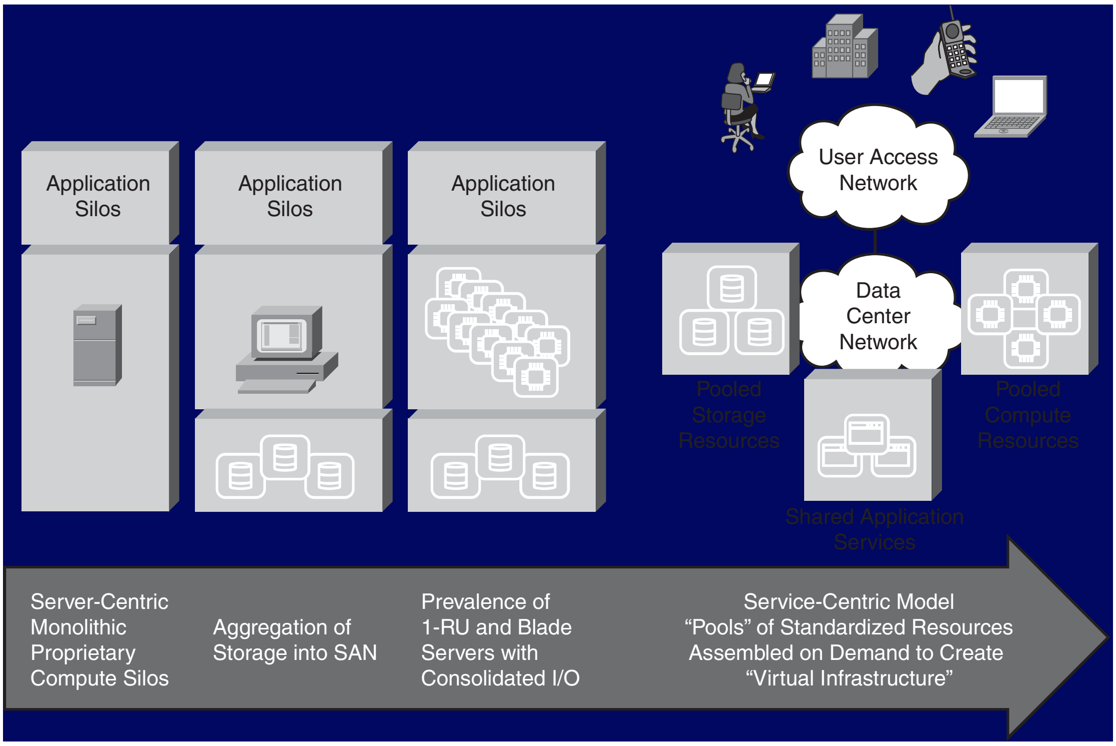
NOTE VMware information is available at http://www.vmware.com/. Xen information is available at http://www.xensource.com/.

- The removal of storage from the server, consolidating it in storage pools. Networked storage (such as storage area networks [SAN]) allows easier management, provisioning, improved utilization, and consistent recovery practices.
- The creation of pools of one-way, two-way, or four-way servers that can be pooled and provisioned, on demand.

NOTE One-way servers have a single processor, two-way servers have two processors, and four-way servers have four processors.

- The consolidation of I/O resources so that the I/O can be pooled and provisioned on demand for connectivity to other servers, storage, and LAN pools.

Figure 4-19 *Evolution from Server-Centric to Service-Centric Data Center*



The resulting service-centric data center has pooled compute, storage, and I/O resources that are provisioned to support applications over the data center network. Because the network touches and can control all the components, the network can be used to integrate all the applications and services; network technology actively participates in the delivery of applications to end users.

The Cisco Enterprise Data Center Architecture Framework

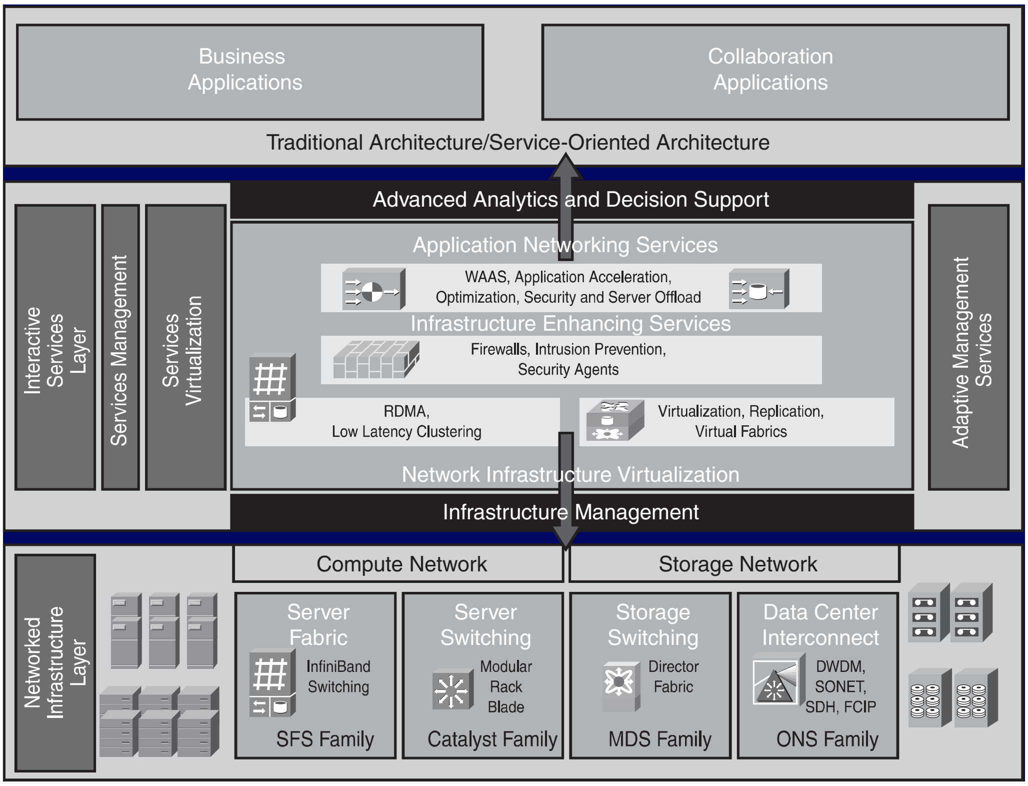
The consolidation and virtualization of data center resources requires a highly scalable, resilient, secure data center network foundation.

As described in Chapter 2, “Applying a Methodology to Network Design,” the Cisco Service-Oriented Network Architecture (SONA) framework defines how enterprises can evolve toward intelligence in the network that optimizes applications, business processes, and resources. The Cisco Enterprise Data Center Architecture, based on SONA, provides organizations with a framework to address immediate data center demands for consolidation and business continuance while enabling emerging service-oriented architectures (SOA), virtualization, and on-demand computing technologies in the data center.

The Cisco Enterprise Data Center Architecture, as illustrated in Figure 4-20, aligns data center resources with business applications and provides multiple resources to end users in an enterprise. The Cisco Enterprise Data Center Architecture has the following layers:

- **Networked Infrastructure layer**: Meets all the bandwidth, latency, and protocol requirements for user-to-server, server-to-server, and server-to-storage connectivity and communications in a modular, hierarchical infrastructure.
- **Interactive Services layer**: Provides the infrastructure services that ensure the fast and secure alignment of resources with application requirements and Cisco Application Networking Services that optimize application integration and the delivery of applications to end users.

Figure 4-20 *Cisco Enterprise Data Center Network Architecture Framework*



WAAS = Wide-Area Application Services; RDMA = Remote Data Memory Access; SFS = Server Fabric Switching; MDS = Multilayer Directors and Fabric Switches; ONS = Optical Networking Solutions; DWDM = Dense Wave Division Multiplexing; SONET = Synchronous Optical Network; SDH = Synchronous Digital Hierarchy; FCIP = Fiber Channel over IP

The Cisco Enterprise Data Center Architecture provides a scalable foundation that allows data centers to host a variety of legacy and emerging systems and technologies, including the following:

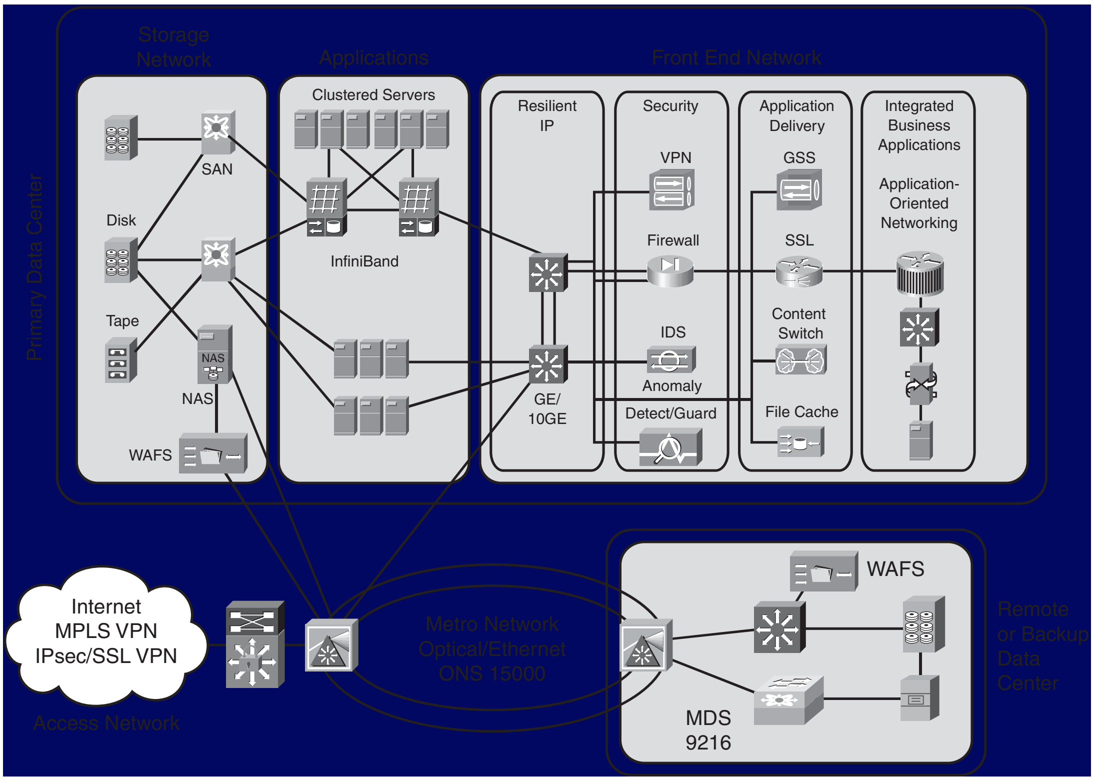
- ***N*-tier applications**: Secure network zones support two, three, or *n*-tier application environments with techniques that optimize application availability and server and storage utilization.
- **Web applications**: Application acceleration and server optimization technologies provide improved scalability and delivery of web applications to end users, wherever they are.
- **Blade servers**: As self-contained servers, blade servers, housed in a blade enclosure, have all the functional components required to be considered computers but have reduced physical components, so they require less space, power, and so forth. The Cisco Enterprise Data Center Architecture provides an intelligent network foundation using integrated Ethernet and InfiniBand switching technology that helps optimize blade server availability, security, and performance.
- **Clustering, high-performance computing and grid**: The Cisco high-performance data, server, and storage switching solutions, whether based on Ethernet, InfiniBand, or Fiber Channel, enable the deployment of data- and I/O-intensive applications that make use of these distributed compute and storage architectures.
- **SOA and web service**: The Cisco Enterprise Data Center Architecture facilitates the reliable, secure, and rapid deployment of an SOA by enabling dynamic deployment and scaling of secure infrastructures and by enhancing application integration with message-based services.
- **Mainframe computing**: Cisco offers a comprehensive set of technologies supporting Systems Network Architecture (SNA), SNA-to-IP migration, fiber connection, and native IP mainframe services.

The Cisco Enterprise Data Center Architecture is supported by networking technologies and solutions that allow organizations to evolve their data center infrastructures through the following phases:

- **Consolidation**: Integration of network, server, application, and storage services into a shared infrastructure enhances scalability and manageability while reducing cost and complexity.
- **Virtualization**: Network-enabled virtualization of computing and storage resources and virtual network services increase utilization and adaptability while reducing overall costs.
- **Automation**: Dynamic monitoring, provisioning, and orchestration of data center infrastructure resources resulting from changing loads, disruptions, or attacks increases overall IT agility while minimizing operational requirements.

Figure 4-21 illustrates a sample high-performance data center network topology that requires many technologies and connectivity options among applications and data centers. This network topology provides connectivity services for networked elements within the data center, such as servers and storage, as well as to external users or other data centers.

Figure 4-21 *Sample Data Center Network Topology*



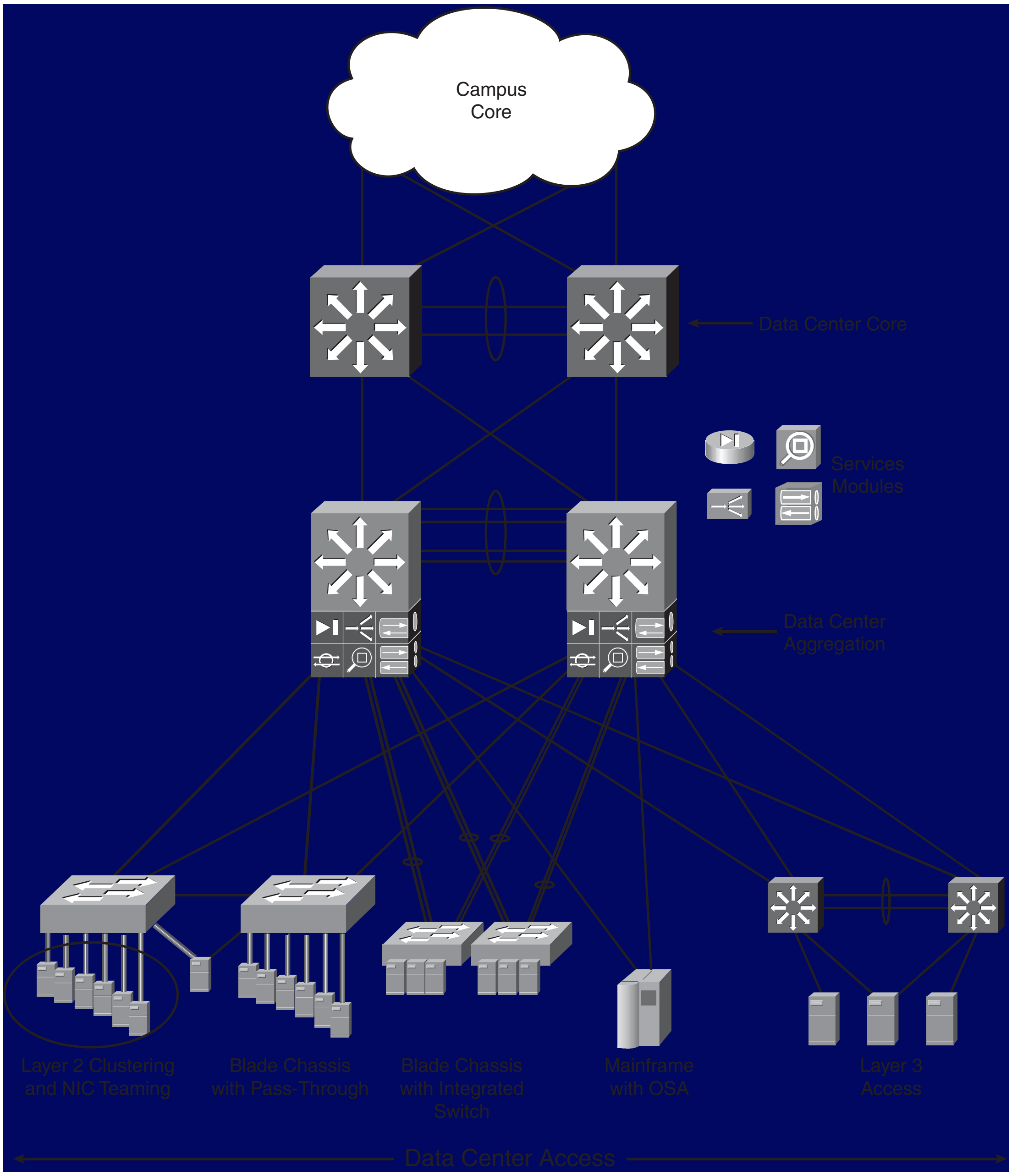
NAS = Network Attached Storage; WAFS = Wide-Area File Services; GE = Gigabit Ethernet; VPN = Virtual Private Network; IDS = intrusion detection system; GSS = Global Site Selector; SSL = Secure Sockets Layer

Enterprise Data Center Infrastructure

Figure 4-22 shows a typical large Enterprise Data Center infrastructure design. The design follows the Cisco multilayer infrastructure architecture, including core, aggregation, and access layers.

NOTE In the Enterprise Data Center, the distribution layer is known as the *aggregation* layer.

Figure 4-22 *Sample Data Center Infrastructure*



OSA = Open Systems Adapter

The data center infrastructure must provide port density and Layer 2 and Layer 3 connectivity for servers at the access layer, while supporting security services provided by ACLs, firewalls, and intrusion detection systems (IDS) at the data center aggregation layer. It must support Server Farm services, such as content switching, caching, and Secure Sockets Layer (SSL) offloading while integrating with multitier Server Farms, mainframes, and mainframe services (such as TN3270, load balancing, and SSL offloading). Network devices are often deployed in redundant pairs to avoid a single point of failure.

The following sections describe the three layers of the Enterprise Data Center infrastructure.

Data Center Access Layer

The Data Center Access layer provides Layer 2, Layer 3, and mainframe connectivity. The design of the Data Center Access layer varies depending on whether Layer 2 or Layer 3 access switches are used; it is typically built with high-performance, low-latency Layer 2 switches, allowing better sharing of service devices across multiple servers and allowing the use of Layer 2 clustering, which requires the servers to be Layer 2–adjacent. With Layer 2 access switches, the default gateway for the servers can be configured at the access or aggregation layer.

Servers can be single- or dual-attached; with dual-attached NICs in the servers, a VLAN or trunk is required between the two redundant access layer switches to support having a single IP address on the two server links to two separate switches. The default gateway is implemented at the access layer.

A mix of both Layer 2 and Layer 3 access switches using one rack unit (1RU) and modular platforms results in a flexible solution and allows application environments to be optimally positioned.

Data Center Aggregation Layer

The Data Center Aggregation (distribution) layer aggregates the uplinks from the access layer to the Data Center Core layer and is the critical point for control and application services.

Security and application service devices (such as load-balancing devices, SSL offloading devices, firewalls, and IDS devices) provide Layer 4 through Layer 7 services and are often deployed as a module in the aggregation layer. This highly flexible design takes advantage of economies of scale by lowering the total cost of ownership (TCO) and reducing complexity by reducing the number of components to configure and manage. Service devices deployed at the aggregation layer are shared among all the servers, whereas service devices deployed at the access layer benefit only the servers that are directly attached to the specific access switch.

Although Layer 2 at the aggregation (distribution) layer is tolerated for legacy designs, new designs should have Layer 2 only at the Data Center Access layer. With Layer 2 at the Data Center Aggregation layer, physical loops in the topology would have to be managed by STP; in this case, as for other designs, RPVST+ is a recommended best practice to ensure a logically loop-free topology over the physical topology.

The Data Center Aggregation layer typically provides Layer 3 connectivity from the data center to the core and maintains the connection and session state for redundancy. Depending on the requirements and the design, the boundary between Layer 2 and Layer 3 at the Data Center Aggregation layer can be in the multilayer switches, the firewalls, or the content-switching devices in the aggregation layer. Depending on the data center applications, the aggregation layer might also need to support a large STP processing load.

Data Center Core Layer

Implementing a Data Center Core layer is a best practice for large data centers. The following should be taken into consideration when determining whether a core is appropriate:

- **10-Gigabit Ethernet density**: Without a Data Center Core, will there be enough 10-Gigabit Ethernet ports on the Campus Core switch pair to support both the campus Building Distribution layer and the Data Center Aggregation layer?
- **Administrative domains and policies**: Separate campus and data center cores help isolate the campus Building Distribution layers from Data Center Aggregation layers for troubleshooting, maintenance, administration, and implementation of policies (using QoS and ACLs).
- **Anticipation of future development**: The impact that could result from implementing a separate Data Center Core layer at a later date might make it worthwhile to install it at the beginning.

The data center typically connects to the Campus Core using Layer 3 links. The data center network addresses are summarized into the Campus Core, and the Campus Core injects a default route into the data center network. Key Data Center Core layer characteristics include the following:

- A distributed forwarding architecture
- Low-latency switching
- 10-Gigabit Ethernet scalability
- Scalable IP multicast support

Density and Scalability of Servers

Some scaling issues in the data center relate to the physical environment.

The most common access layer in enterprises today is based on the modular chassis Cisco Catalyst 6500 or 4500 Series switches. This topology has also proven to be a very scalable method of building Server Farms that provide high-density, high-speed uplinks and redundant power and processors. Although this approach has been very successful, it results in challenges when used in Enterprise Data Center environments. The typical Enterprise Data Center experiences high growth in the sheer number of servers; at the same time, server density has been improved with 1RU and blade server solutions. Three particular challenges result from this trend:

- **Cable bulk**: Typically, three to four interfaces are connected on a server. With a higher density of servers per rack, cable routing and management can become quite difficult.
- **Power**: The increased density of components in a 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 number of cables lying under the raised floor and the cable bulk at the cabinet base entry is blocking the airflow required to cool equipment in the racks. At the same time, the servers in the rack require more cooling volume because of their higher density.

These challenges have forced customers to find alternative solutions by spacing cabinets, modifying cable routes, or other means, including not deploying high-density server solutions. Another way that customers seek to solve some of these problems is by using a rack-based switching solution. Using 1RU top-of-rack switches keeps the server interface cables in the cabinet, reducing the amount of cabling in the floor and thus reducing the cabling and cooling issues. Another option is to place Cisco Catalyst 6500 Series switches like bookends near the ends of the row of racks so that there are fewer switches to manage.

Summary

In this chapter you learned about campus and data center network design, with a focus on the following topics:

- The effects of the characteristics of the following on the campus network design:
  - — Application: Including peer-peer, client–local server, client–Server Farm, and client–Enterprise Edge server
  - — Environment: Including the location of the network nodes, the distance between the nodes, and the transmission media used
  - — Infrastructure devices: Including Layer 2 or multilayer switching, convergence time, type of multilayer switching, IP multicast, QoS, and load sharing