Evolving Data Center Architectures: Meet the Challenge with Cisco Nexus 5000 Series Switches

What You Will Learn

Data center architectures are evolving to meet the demands and complexities imposed by increasing business requirements to stay competitive and agile. Industry trends such as data center consolidation, server virtualization, advancements in processor technologies, increasing storage demands, rise in data rates, and the desire to implement “green” initiatives is causing stress on current data center designs. However, as this document discusses, crucial innovations are emerging to address most of these concerns with an attractive return on investment (ROI). Future data centers architectures will incorporate increasing adoption of 10 Gigabit Ethernet, new technologies such as Fibre Channel over Ethernet (FCoE), and increased interaction among virtualized environments.

Industry Demands

Data centers environments are adapting to accommodate higher expectations for growth, consolidation, and security. New demands for uptime and serviceability coupled with the new technology and protocols make the design of the data center more challenging. The top trends in the data center are consolidation, growth, availability, and operational efficiency. Business needs require highly reliable applications, which in turn require more servers in the data center and secondary data centers to accommodate the need for business continuity.

New technologies such as multi-core CPU, multi-socket motherboards, inexpensive memory, and Peripheral Component Interconnect (PCI) bus technology represent a major evolution in the computing environment. These advancements provide access to greater performance and resource utilization at a time of exponential growth of digital data and globalization through the Internet. Multithreaded applications designed to use these resources are both bandwidth intensive and require higher performance and efficiency from the underlying infrastructure.

While data center performance requirements are growing, IT managers are seeking ways to limit physical expansion by increasing the utilization of current resources. Server consolidation by means of server virtualization has become an appealing option. The use of multiple virtual machines take full advantage of a physical server’s computing potential and enable a rapid response to shifting data center demands. This rapid increase in computing power coupled with the increased use of virtual machine environments is increasing the demand for higher bandwidth and at the same time creating additional challenges for the network.

Power continues to be one of the top concerns facing data center operators and designers. Data center facilities are designed with a specific power budget, in kilowatts per rack (or watts per square foot). Per-rack power consumption has steadily increased over the past several years. Growth in the number of servers and advancement in electronic components continue to consume power at an exponentially increasing rate. Per-rack power requirements constrain the number of racks a data center can support, resulting in data center that are out of capacity even though there is plenty of unused space. Approximately half of the power required by the data center is consumed by cooling.

Cabling also represents a significant portion of a typical data center budget. Cable sprawl can limit data center deployments by obstructing airflows and requiring complex cooling system solutions. IT departments around the world are looking for innovative solutions that will enable them to keep up with this rapid growth with increased efficiency and low cost.

The Cisco® Data Center 3.0 strategy, which includes the Cisco Nexus 5000 Series Switches, is designed to address these challenges and allow customers to evolve their data centers by consolidating fragmented systems to create a unified fabric. Ethernet technology and pools of disparate data center resources are combined into shared groups that are linked by an intelligent information network.

Figure 1 shows a typical current data center architecture that is subject to the data center challenges just mentioned. It has Gigabit Ethernet host connectivity to Cisco Catalyst® 6500 Series Switches in the access layer end-of-row design. Integrated services such as firewalling and load balancing are provided in service switches at the aggregation layer. Each host connects to the LAN and storage area network (SAN) over separate networks through discrete network interface cards (NICs) and host bus adapters (HBAs).

Figure 1. Current Data Center Architecture

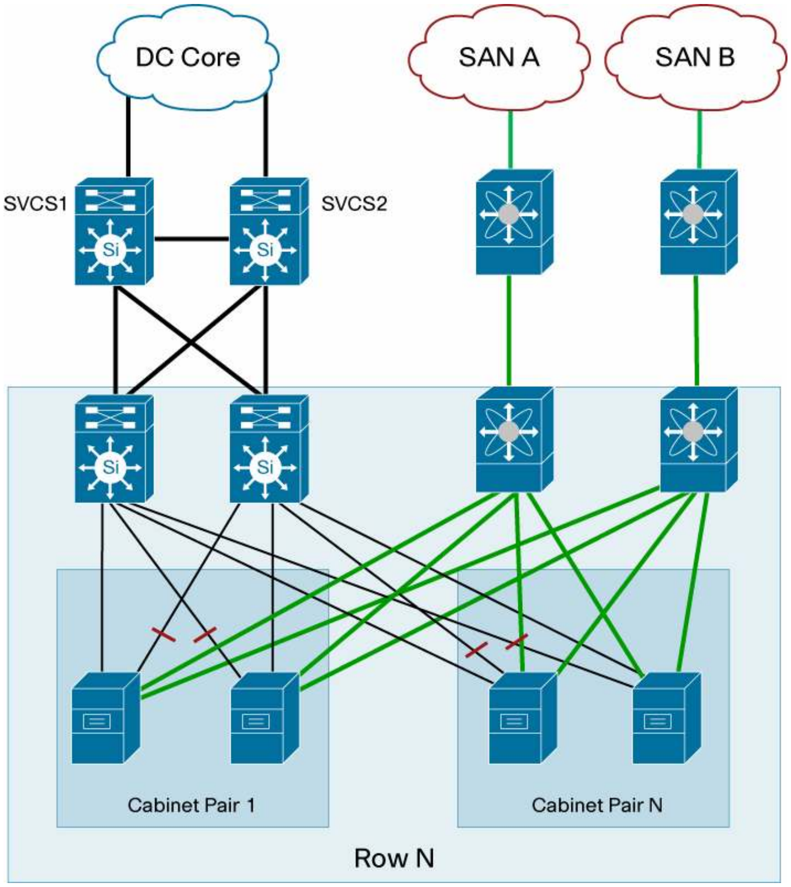
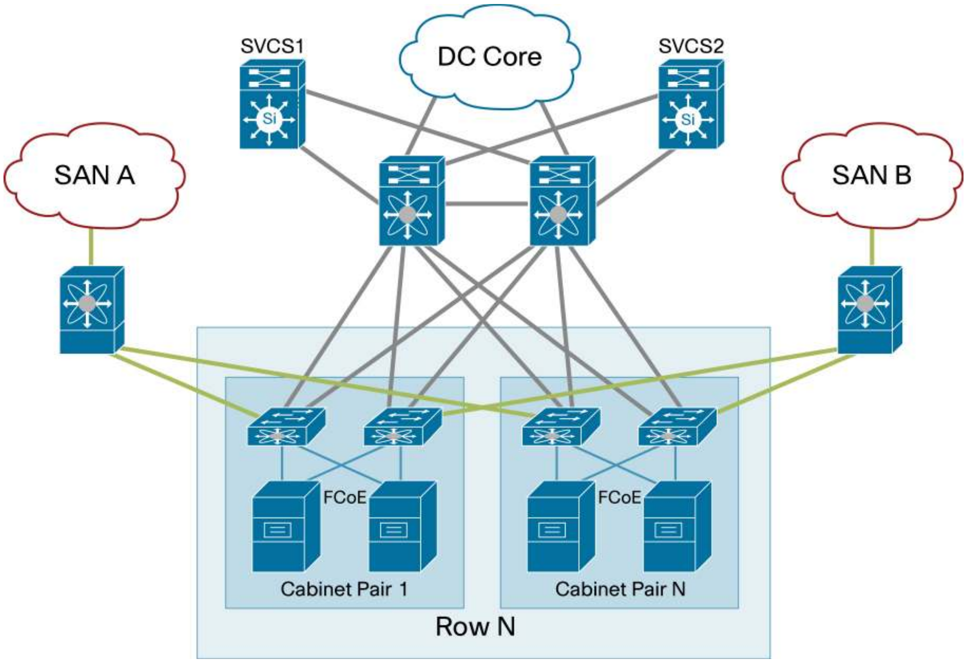


Figure 2 shows the evolution of the data center architecture, using I/O consolidation and enhanced Ethernet features. I/O consolidation is the capability to use the same physical infrastructure to carry different types of traffic, which typically have very different traffic characteristics and transmission requirements. Cisco Nexus 5000 Series Switches along with the Converged Network Adapter (CNA) in the hosts provide I/O consolidation at the access layer. The access layer switches are deployed in a top-of-rack (ToR) fashion, connecting to aggregation layer switches (Cisco Catalyst 6500 Series or Cisco Nexus 7000 Series Switches). Because of their high port density and modular design, Cisco Nexus 7000 Series Switches are targeted for the aggregation and core layers in the data center. Network services are integrated into the architecture through Cisco Catalyst 6500 Series service modules.

Figure 2. Next-Generation Data Center Architecture



Architecture Principles

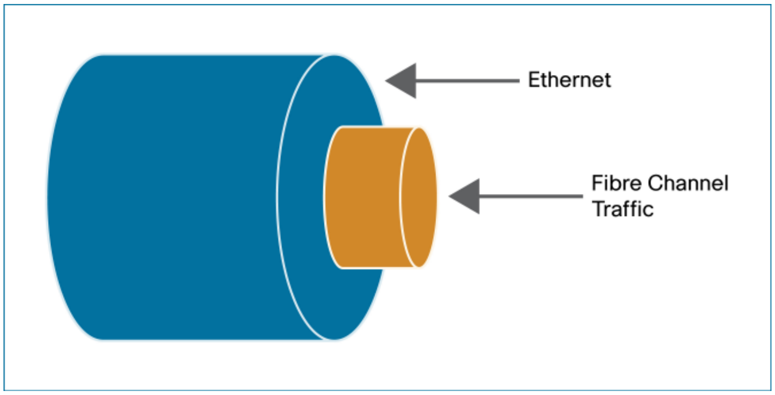
This section examines the architecture principles underlying the new data center designs.

Fibre Channel over Ethernet

FCoE is a proposed standard that allows Fibre Channel communications to run over Ethernet (Figure 3). Fibre Channel supports data communications between devices that connect servers with storage devices and between storage controllers and disk drives. Benefits of FCoE include:

- Dramatic reduction in the number of adapters, switch ports, and cables
- Low power consumption
- High-performance frame mapping instead of the use of traditional gateways
- Consolidated network infrastructure
- Effective sharing of high-bandwidth links
- Lower total cost of ownership (TCO)

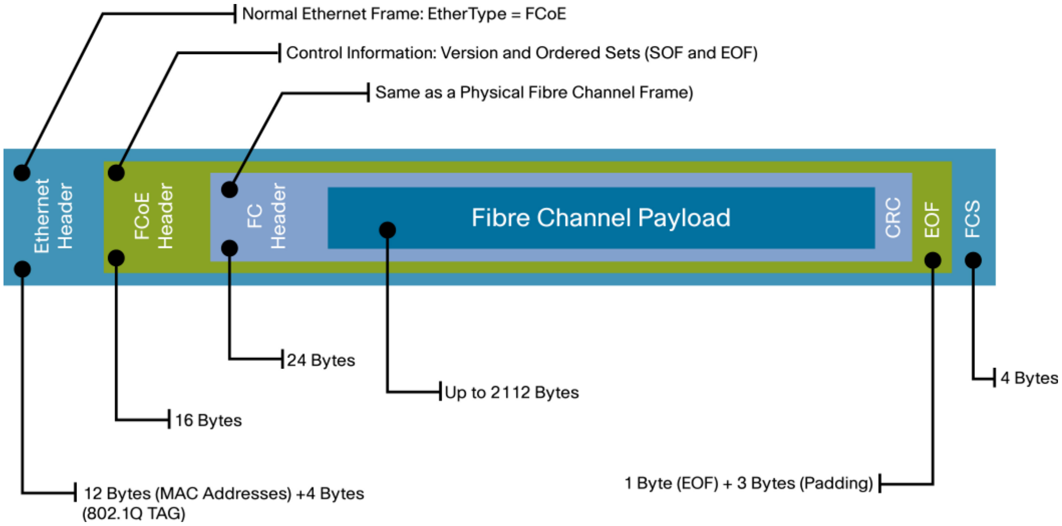
Figure 3. Fibre Channel over Ethernet Encapsulation



The first-generation implementation of FCoE at the host involves minimal change from the traditional Fibre Channel driver stack perspective. The Fibre Channel model stays consistent with FCoE to the host. FCoE termination on the Cisco Nexus 5000 Series provides transparent Fibre Channel support through the F or E port for the Fibre Channel uplinks. To support Fibre Channel over an Ethernet network, the fabric must not drop frames. The Cisco Nexus 5000 series switches are FCoE-capable switches that supports emerging IEEE Data Center Bridging standards to a deliver a lossless Ethernet service with no-drop flow control that is conceptually similar to the lossless behavior provided in Fibre Channel with buffer-to-buffer credits. FCoE standards documentation can be found at the T11 site at http://www.t11.org/fcoe.

The fabric also needs to support jumbo frames, to allow 2180-byte Ethernet frames in the fabric. A Fibre Channel frame is 2112 bytes long and can easily be transmitted without modification. Figure 4 depicts the frame structure.

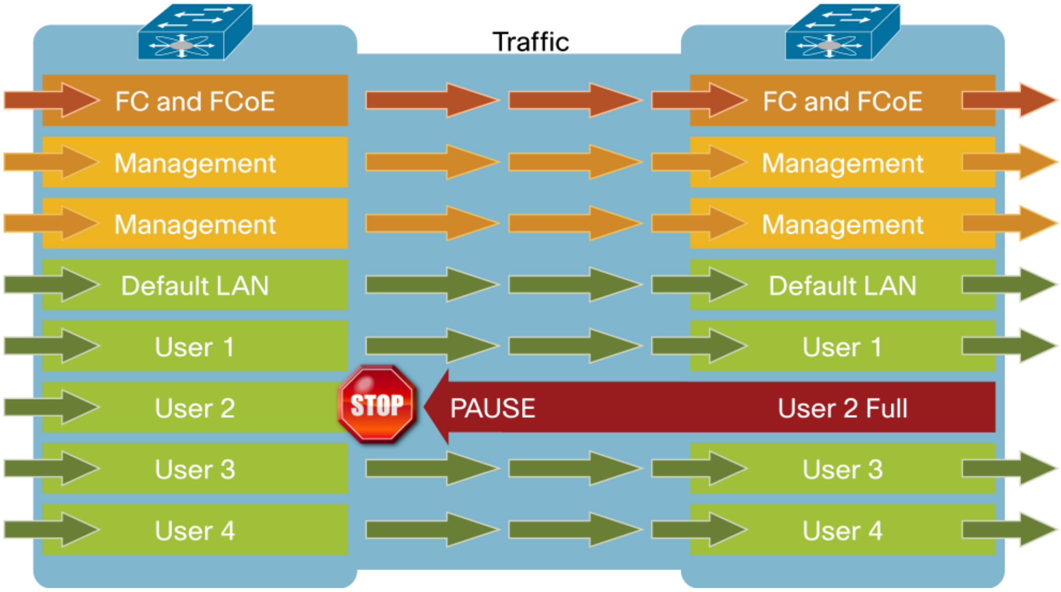
Figure 4. Fibre Channel over Ethernet Encapsulation Frame Size


Normal Ethernet Frame: EtherType = FCoE
Control Information: Version and Ordered Sets (SOF and EOF)
Same as a Physical Fibre Channel Frame)
Ethernet Header
FCoE Header
FC Header
Fibre Channel Payload
CRC
EOF
FCS
24 Bytes
Up to 2112 Bytes
4 Bytes
16 Bytes
12 Bytes (MAC Addresses) +4 Bytes (802.1Q TAG)
1 Byte (EOF) + 3 Bytes (Padding)

Priority Flow Control

To support Fibre Channel frames over Ethernet, no frames can be dropped throughout the entire transmission. IEEE 802.1Qbb Priority Flow Control (PFC) extends the granularity of IEEE 802.3x PAUSE to accommodate different priority classes. Using PFC, a link is divided into eight lanes, where PAUSE is applied on a per-lane basis such that PAUSE in one lane does not affect the other lanes. With the capability to enable PAUSE on a per-user-priority basis, a lossless lane for Fibre Channel can be created while retaining packet-drop congestion management for IP traffic. This mechanism allows storage traffic to share the same link as non-storage traffic. IEEE 802.1Qaz Enhanced Transmission Selection (ETS) is used to assign traffic to a particular virtual lane using IEEE 802.1p class of service (CoS) values to identify which virtual lane traffic belongs to. Using PFC and ETS allows administrators to allocate resources, including buffers and queues, based on user priority, which results in a higher level of service for critical traffic where congestion has the greatest effect. Figure 5 depicts the PAUSE mechanism for the User 2 traffic flow.

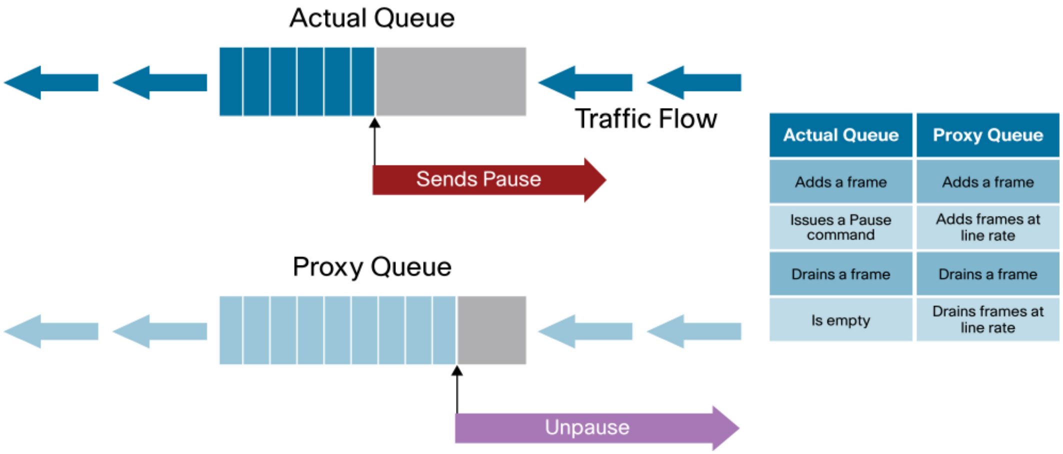
Figure 5. Priority Flow Control



Delayed Drop

Delayed drop is an Ethernet enhancement that uses the PAUSE mechanism to reduce packet drop on short-lived traffic bursts while triggering upper-layer congestion control through packet drops to handle long-term congestion. This feature allows congestion to spread only for short-term bursts, minimizing long-term congestion. Delayed drop can be enabled with per-user priority and uses a proxy queue to measure the duration of traffic bursts. During a steady-state condition, the proxy queue emulates an actual queue in which packets are added or removed. When a burst of traffic causes the actual queue to reach a particular threshold, a PAUSE frame is sent to the source. Meanwhile, the proxy queue still continues to fill. When the proxy queue is filled, the PAUSE is removed from the transmitter, which causes frame drop if the congestion persists. The condition is required to simulate the TCP flow control mechanism for long-lived streams. Delayed drop sits between the traditional Ethernet and PFC behavior. With delayed drop, a CoS can be flow controlled for a deterministic amount of time. The traditional drop behavior follows if the congestion is not resolved within that time frame. Figure 6 depicts the behavior of the queues. The Delayed Drop mechanism will be available in a future software release.

Figure 6. Delayed Drop



Actual Queue	Proxy Queue
Adds a frame	Adds a frame
Issues a Pause command	Adds frames at line rate
Drains a frame	Drains a frame
Is empty	Drains frames at line rate

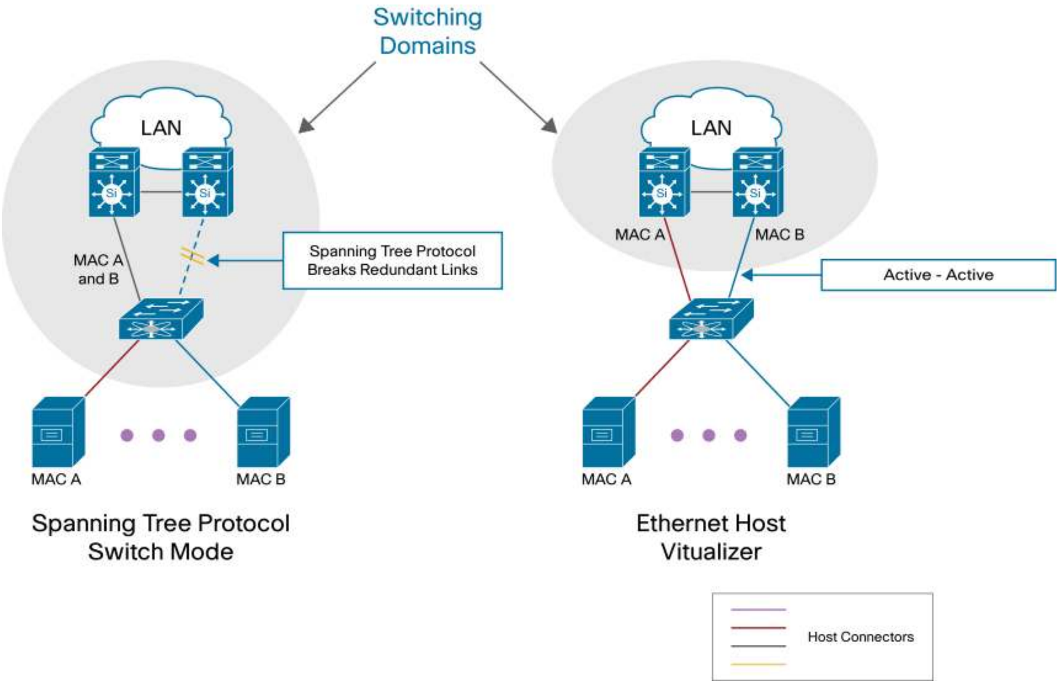
Cut-Through Switching

Switches are usually characterized by the switching method they support. Store-and-forward switching accepts the entire frame into the switch buffers and computes a cyclic redundancy check (CRC) before forwarding the frame to its determined outgoing interface. The process does add some latency at the expense of helping ensure that only valid frames that have passed CRC get forwarded to the switch fabric. Store-and-forward architecture is well suited to preventing malformed packet propagation. Cut-through switching forgoes the CRC and reads only the destination MAC address (the first 6 bytes following the preamble) to determine to which switch port to forward the traffic. This process removes some of the processing overhead and reduces transmission latency. Cut-through technology is well suited for low-latency and grid computing application environments where high-speed packet switching is required. Applications may behave differently when deployed over the different switching architectures, and appropriate testing needs to be performed to understand the behavior of the environment.

End-Host Virtualizer

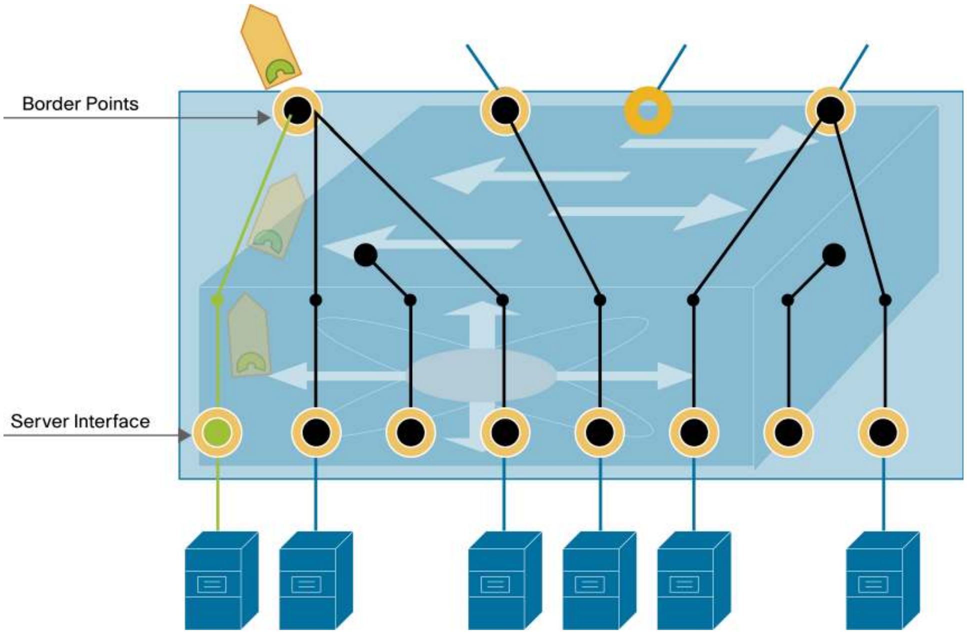
End-host virtualizer (EHV) mode allows the switch to behave like a collection of end hosts on the network, representing all the hosts (servers) connected to the switch through server-facing interfaces. Although the primary reason for implementing EHV is to offer redundancy, active-active or active-standby, EHV also allows isolation of resources from the network because the switch does not participate in the network’s control plane. Using EHV mode, the switch does not participate in Spanning Tree and avoid loops by disallowing border (network-connected) ports from forwarding network traffic to one another and by disallowing server traffic on server-facing interfaces from egressing on more than one border (network-facing) port at any given time (Figure 7).

Figure 7. Switching Methods



In EHV mode, the network-bound traffic is forwarded based on the arrival server-facing interfaces, a technique called source pinning, whereas server-bound traffic is forwarded based on the server's MAC address. All server-facing ports connecting to the Cisco Nexus 5000 Series in EHV mode are pinned to one of the border ports. Border ports provide uplink connectivity to the aggregation layer switches (Figure 8). These uplinks can also be a Cisco EtherChannel bundle if required. If one of the border ports fails, the ports will be re-pinned to an available border port as calculated by the pinning algorithm. This feature allows redundant connectivity to the network without the need for protocol to communicate the information to the network. In EHV mode, the unknown source MAC addresses from end hosts are learned and populated in the hardware forwarding tables to allow forwarding of traffic to the end hosts' MAC addresses based on the destination addresses.

Figure 8. Pinning



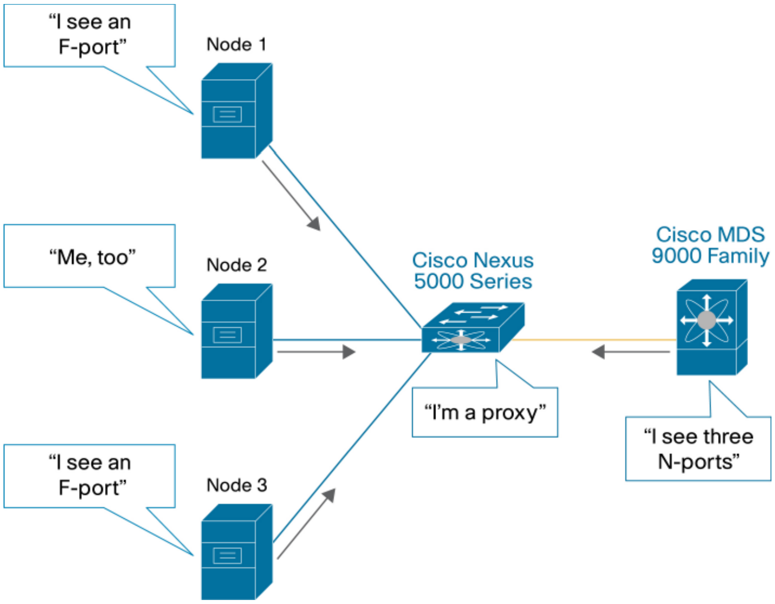
N-Port Virtualizer

From a SAN perspective, a ToR architecture in the data center introduces challenges such as a large increase in the number of domain IDs required and interoperability concerns in multivendor environments. Original Storage Manufacturers (OSMs) place an upper limit on the number of Fibre Channel domain IDs in a SAN. The number is typically less than 40, even though the theoretical limit is 239. Each ToR SAN switch uses one Fibre Channel domain ID, stretching the already low limit of available IDs in certain environments. The existing switch-to-switch interoperability between multivendor SANs based on the E-port requires configuration of special interoperability modes and requires ongoing management.

N-port Virtualizer (NPV) addresses the increase in the number of domain IDs needed by making a fabric switch appear as a host to the core Fibre Channel switch and as a Fibre Channel switch to the server edge switch. NPV aggregates multiple locally connected N-ports into one or more external N-proxy links, which share the domain ID of the upstream core switch. NPV also allows multiple devices to attach to the same port on the core switch, thereby reducing the need for more ports at the core.

- The NPV switch relays FLOGI or FDISC commands to the upstream Fibre Channel switch.
- The NPV switch has no E-port or Fibre Channel services.
- The NPV switch requires N-port identifier virtualization (NPIV) support on the E-port in the upstream Fibre Channel switch.
- The NPV switch does not provide any local switching.
- The NPV switch retries a failed login request from one uplink interface on a different interface.
- The NPV switch handles events by generating proxy LOGOs.
- The server-facing interfaces are pinned to uplink interfaces based on the lowest number of server-facing interfaces already assigned.

Figure 9. NPV Operation



NPV addresses switch-switch multivendor interoperability by appearing as a collection of host bus adapters (HBAs) rather than a Fibre Channel switch in the fabric, providing transparent interoperability. The NPV feature very effectively addresses the challenges of limited domain IDs, multivendor interoperability, and management.

N-Port Identifier Virtualization

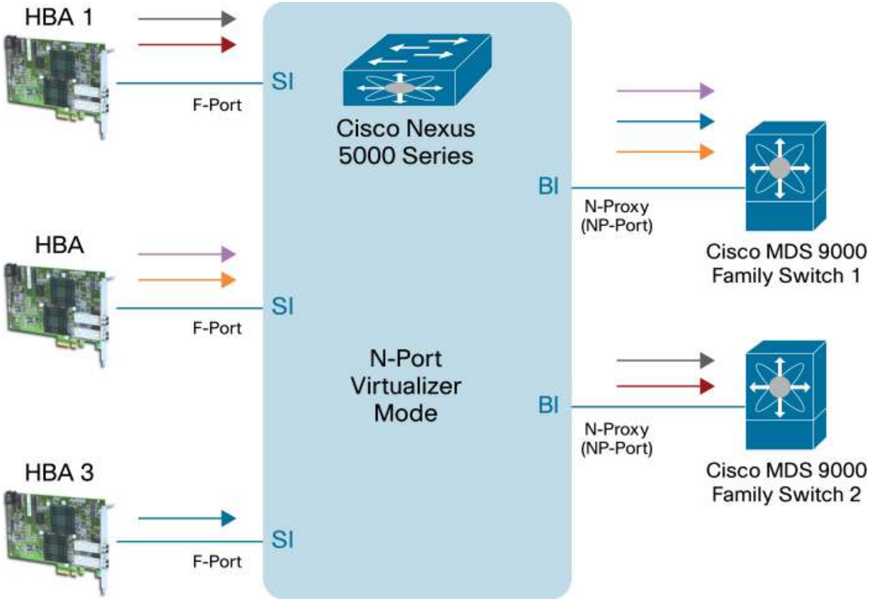
NPIV is a T11 ANSI standard that enables a fabric switch to register several worldwide port names (WWPNs) on the same physical port. The Fibre Channel standard allows only one FLOGI command per HBA; NPIV is an extension that allows multiple FLOGI commands on one F-port.

NPIV allows multiple virtual machines to share one HBA, with each virtual machine having a unique identifier on the Fibre Channel fabric.

- The first N-port uses a FLOGI command, and subsequent N-ports use the FDISC command.
- Each N-port is known by its worldwide name (WWN).
- Each N-port receives its own Fibre Channel ID (FCID).
- All N-ports share the buffer-to-buffer (BB_credits) negotiated at FLOGI.

NPIV also allows access control, zoning, and port security to be implemented at the application level. NPV uses the NPIV feature to get multiple FCIDs allocated from the upstream Fibre Channel switch on the N-proxy port (Figure 10).

Figure 10. N-Proxy

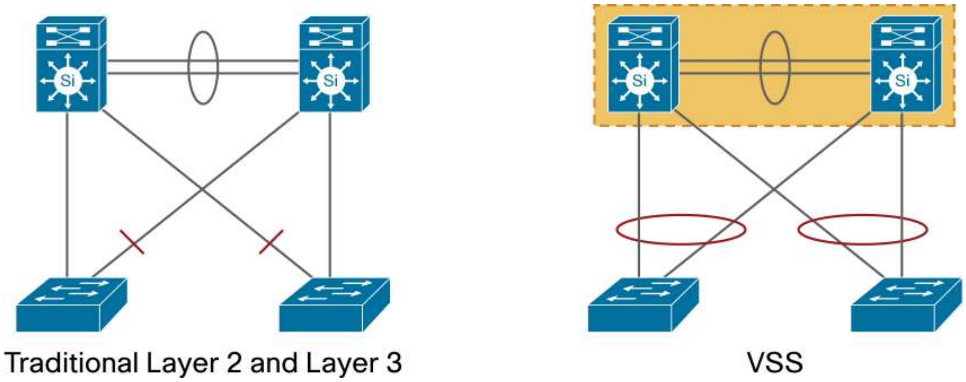


Virtual Switching System

The virtual switching system (VSS) technology allows two Cisco Catalyst 6500 Series Switches to be represented as one logical entity to the rest of the network. This design provides a single point of management, increased availability, and an efficient, scalable network design. One of the main features of VSS is multichassis Cisco EtherChannel, which allows a Cisco EtherChannel interface to span multiple physical switches.

In a traditional Layer 2 and 3 environment, one link from every access switch remains in blocking mode to prevent loops in the topology. Spanning Tree Protocol is required for this design to remain functional. In a VSS topology, all links from the access switch are in the forwarding state, allowing bisectional bandwidth and true Layer 2 multipathing capabilities. The Cisco EtherChannel links have increased resiliency since chassis-level redundancy is available in addition to the usual module- and port-level redundancy (Figure 11).

Figure 11. Traditional and VSS Topology Comparison



Host Adapters

To provide 10 Gigabit Ethernet connectivity to the host, a standards-based Ethernet adapter can be used. However, to transport LAN and SAN traffic over the same link from the host to the switch, the host requires either a CNA or a software stack with a 10 Gigabit Ethernet NIC to generate FCoE frames. Cisco has developed a broad system of partners, including QLogic and Emulex, that have developed CNAs embedding within Cisco's technology. The adapters support some of the architectural principles defined earlier, allowing applications to take advantage of enhanced Ethernet features in classic Ethernet and Unified Fabric environments. Figure 12 depicts how a typical 10GE/FCoE capable adapter provides Unified I/O to the host.

Figure 12. 10GE and FCoE Adapter

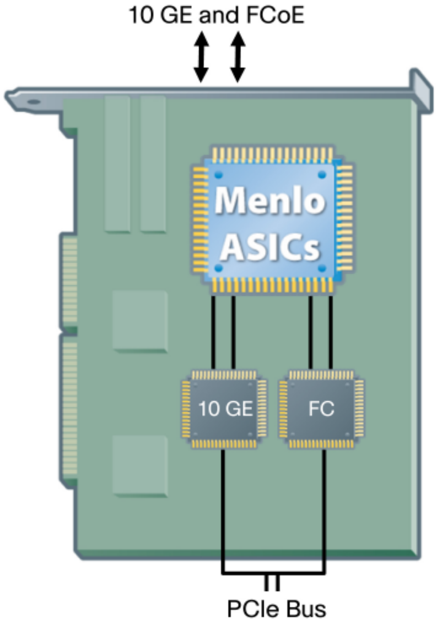
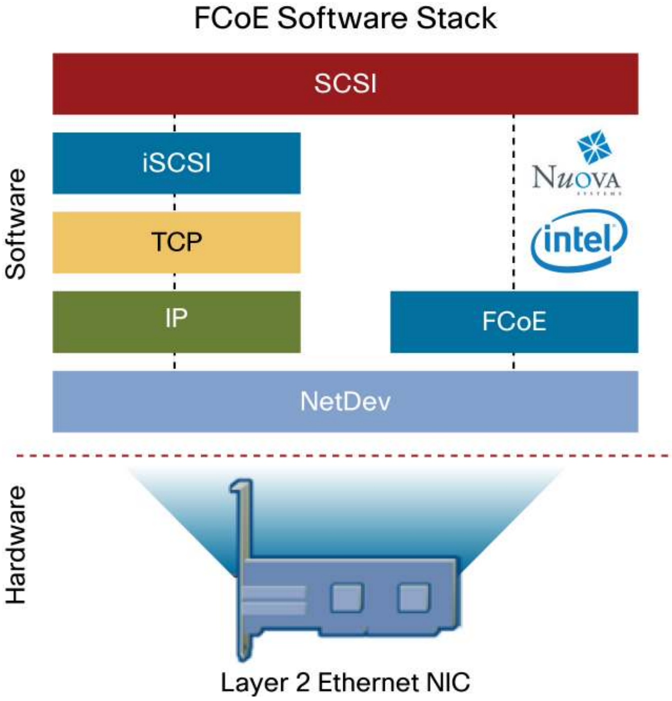


Figure 13. FCoE Software Stack



Cisco has co-developed an open source software stack with Intel that can be used in conjunction with a traditional 10 Gigabit Ethernet module to provide FCoE capability. Intel's Oplin adapter is an example of a 10 Gigabit Ethernet adapter that provides FCoE connectivity with an FCoE software stack (Figure 13).

Cabling

While the Cisco Nexus 5000 Series supports standard optical Small Form-Factor Pluggable Plus (SFP+) transceivers, it also introduces a new generation of low-latency, low-power transceivers with integrated cabling. The SFP+ direct-attached 10 Gigabit Ethernet copper solution essentially eliminates adoption barriers through its low cost. Table 1 shows the cabling and transceiver types and their power and latency budgets.

Table 1. Media Comparison Matrix

Technology	Cable	Distance	Power (Each Side)	Transceiver Latency (Link)
SFP+ Cu Copper	Twinax	10m	~0.1W	~0.1ms
SFP+ SR short reach	MM 62.5mm MM 50mm	82m 300m	1W	~0
10GBASE-T	Cat6 Cat6a/7 Cat 6a/7	55m 100m 30m	~8W ~8W ~4W	2.5us 2.5us 1.5us

Data Center Bridging Exchange Protocol

The Menlo application-specific integrated circuit (ASIC) supports the Data Center Bridging Exchange (DCBX) Protocol at the same level of type-length-values (TLVs) as the switch by using the CNAs. Intel supports a subset of TLVs for the FCoE software stack over Intel's Oplin adapter.

The DCBX Protocol is a result of a joint effort between Cisco and Intel. This ACK-based protocol is built on top of the Link Layer Discovery Protocol (LLDP). It provides:

- Discovery of peer capabilities
- Configuration of peers, with the capability to force configuration from either side of the link
- Configuration mismatch detection

The DCBX Protocol Data Unit (PDU) consists of multiple sub-TLVs called feature TLVs. Every feature TLV also defines a compatibility function. A DCBX peer can support some or all of these TLVs.

The logical link status TLV is very important in DCBX. Many features and error conditions in both Ethernet and Fibre Channel, particularly Fibre Channel, rely on bouncing the link between the initiator and the first-hop switch to cause the initiator to resend the FLOGI command. This process is a very common recovery mechanism in Fibre Channel. If the link bounces because there is a Fibre Channel error condition, this error will cause the Ethernet link to fail as well since FCoE provides both Ethernet and Fibre Channel on the same wire. Since failing both logical links comprises the high availability of the unified fabric, a protocol message is needed that the NIC and switch can use to bounce just the logical Fibre Channel link or the logical Ethernet link. Using this new message, the Fibre Channel side can recover and resend a FLOGI command, and the Ethernet side will not be affected. When the switch needs to shut down a virtual interface, a “Logical Link Status DOWN” message is sent to the peer if the peer supports DCBX. If the peer does not support DCBX, the physical link is shut down. The same operation applies if a state change occurs on the Ethernet link, in this case preserving the integrity of the Fibre Channel link.

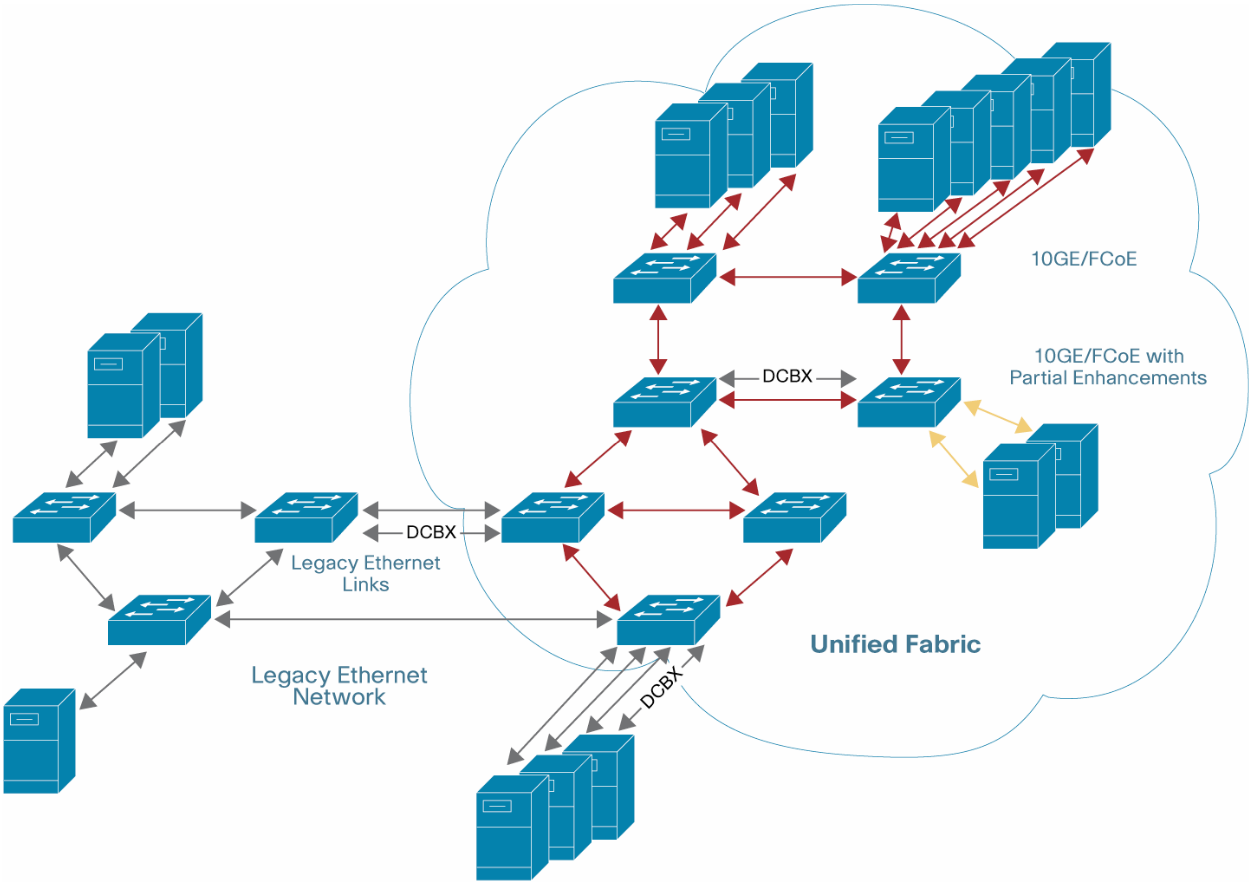
The FCoE CoS TLV indicates:

- FCoE support and FCoE CoS value being used to the peer
- If both sides are using different CoS values, the switch does not bring up Fibre Channel virtual interfaces

The two peers talking FCoE must agree on the CoS value in FCoE packets because a PAUSE frame would be created for the appropriate virtual lane for the correct CoS value to build the lossless Fibre Channel fabric within Ethernet. Some of the other TLVs that exist are the NPIV TLV, PFC TLV, and priority group TLV.

Figure 14 depicts an enhanced Ethernet cloud using DCBX between switches and switch to NIC in the host.

Figure 14. Data Center Bridging Exchange



Deployment Scenarios

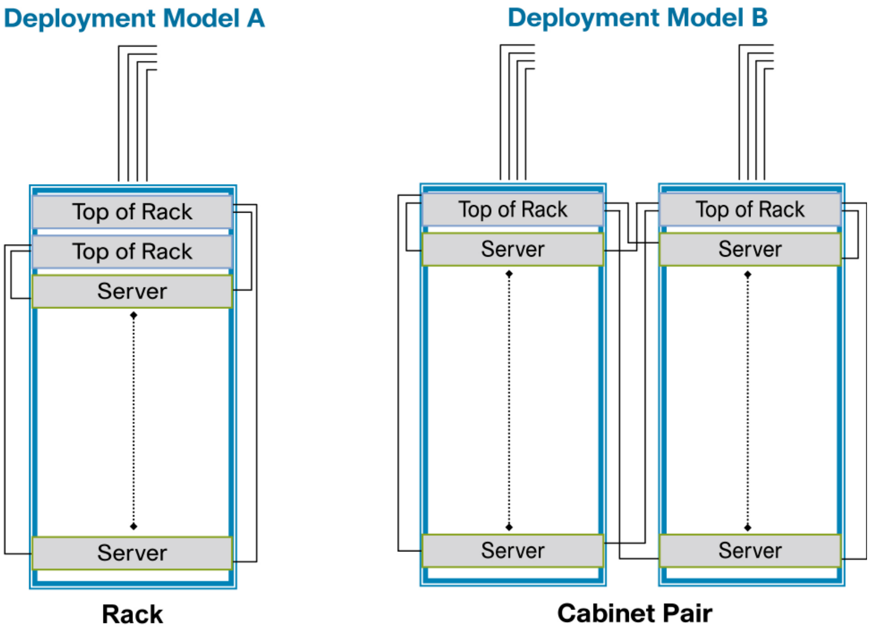
The Cisco Nexus 5000 and Nexus 7000 Series platforms are used primarily in next-generation data center architectures for high-density 10 Gigabit Ethernet and I/O consolidation designs. The Cisco Nexus 5000 Series provides a lossless Ethernet fabric that uses credit-based scheduling between egress and ingress points, port-to-port latency of 3.2 microseconds for any packet size, flexible port configuration options with Ethernet and Fibre channel expansion modules, and I/O consolidation with FCoE. In addition, several Ethernet enhancements are available in the Cisco Nexus 5000 Series that make it an optimal 10 Gigabit Ethernet server farm access layer platform. These features such as IEEE 802.1Qbb PFC, IEEE 802.1Qaz Enhanced Transmission Selection, and delayed drop make the Cisco Nexus 5000 Series the foundation for a next-generation data center. PFC coupled with a lossless switch fabric provides the lossless characteristics required to transport FCoE. In an Ethernet-only topology, the no-drop class may be configured to carry traffic other than FCoE frames, such as market data in Financial Services trading environments. Delayed drop is a crucial Ethernet enhancement that helps absorb microbursts of traffic that are a common cause of TCP retransmissions. These unique enhancements in the Cisco Nexus 5000 Series Switches make them suitable for high-throughput and low-latency environments transporting mission-critical traffic.

The Cisco Nexus 5000 Series Switches are optimized for ToR deployment scenarios in the data center, in which a computing entity can be compartmentalized to a rack. This approach offers flexibility in cabling, allowing servers within a rack to potentially be pre-cabled and then moved into the data center, where they can obtain network connectivity by simply connecting the uplinks. This deployment model enables computing resources to scale quickly and minimizes the burden on the overall cabling infrastructure across the data center.

Several deployment scenarios use the Cisco Nexus 5000 Series as a data center Access layer switch connecting to the Cisco Nexus 7000 or Catalyst 6500 Series platforms in the aggregation layer. Services such as firewalling, load balancing, and network monitoring, although not depicted in the designs here, can be deployed in the architecture through service switches at the aggregation layer. The Cisco Catalyst 6500 and Nexus 7000 Series can be used in the core layer as well, depending on the 10 Gigabit Ethernet port-density requirements. Based on the port density and model of the Cisco Nexus 5000 Series Switch, different rack deployment configurations can be used for optimal cabling requirements. Typically, intra-rack server connections to the ToR Cisco Nexus 5000 Series Switch will use a Twinax cable assembly with SFP+ connectors. Twinax cable provides some of the lowest power budget and latency characteristics for 10 Gigabit Ethernet server connectivity.

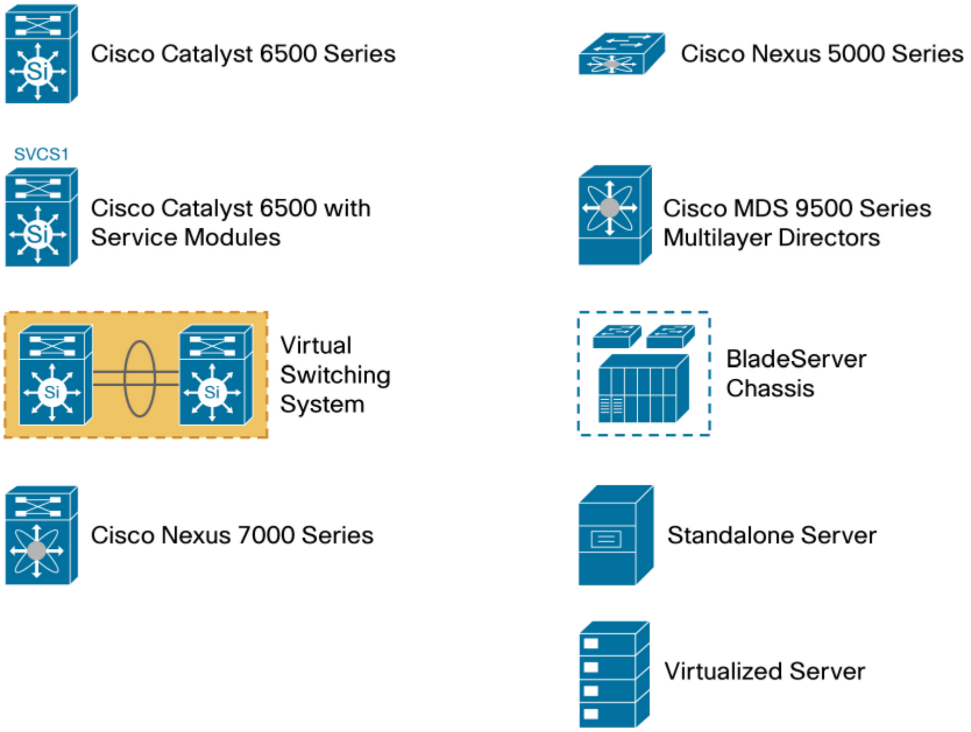
Most of the designs mentioned here are based on ToR deployment models. Deployment Model B depicted in Figure 15 is the model used for the reference designs mentioned here. Most of the designs use FCoE from the host to the Cisco Nexus 5000 Series access layer switch, providing immediate savings in power, cooling, cabling, and host adapter costs compared to traditional designs without I/O consolidation. There are multiple server rack deployment models which can be utilized to provide efficient utilization of switch ports in the Cisco Nexus 5000 Series Switch. Depending on the amount of servers installed in a rack, Figure 15 depicts a few common deployment models.

Figure 15. Rack Deployment Models



- **Rack:** A self-contained unit that encompasses standalone servers or a blade server chassis
- **Cabinet pair:** Two racks joined together as one logical entity in which each rack shares its ToR Cisco Nexus 5000 Series Switch with the adjacent rack in the pair
- **Row:** N number of cabinet pairs in a ToR design or N number of racks in an end-of-row (EoR) design
- **Pod:** A collection of all rows serviced by a pair of core switches; the typical access, aggregation, and core layer architecture is used in all the designs

Figure 16. Device Icons



Scenario 1: 10 Gigabit Ethernet Aggregation with Blade Servers

The scenario 1 design (Figure 17 and Table 2) provides a high-density 10 Gigabit Ethernet topology to support growing blade server environments. Each blade server chassis is equipped with 10 Gigabit Ethernet switches, which provide 10 Gigabit Ethernet uplinks to the EoR Cisco Nexus 5000 Series aggregation switches. The blade servers are also connected to SAN A and SAN B fabrics for Fibre Channel connectivity. Each blade server chassis contains integrated Ethernet Cisco Catalyst Blade Switch 3120 Ethernet blade switches, which are connected in multiple rings to create a virtual blade system (VBS). All eight integrated blade switches are managed as one logical entity. Four 10 Gigabit Ethernet uplinks are connected from the rack to the EoR Cisco Nexus 5000 Series Switches. SAN connectivity from the blade chassis is provided by Cisco MDS 9124e Multilayer Fabric Switch integrated Fibre Channel switches. These integrated SAN blade switches connect to an EoR pair of Cisco MDS 9500 Series switches. Every pair of Cisco Nexus 5000 Series Switches at the EoR is connected to a pair of Cisco Nexus 7000 Series core switches. This design is ideal for aggregating 10 Gigabit Ethernet blade server environments that also require SAN connectivity.

Figure 17. 10 Gigabit Ethernet Aggregation with Blade Servers (VBS)

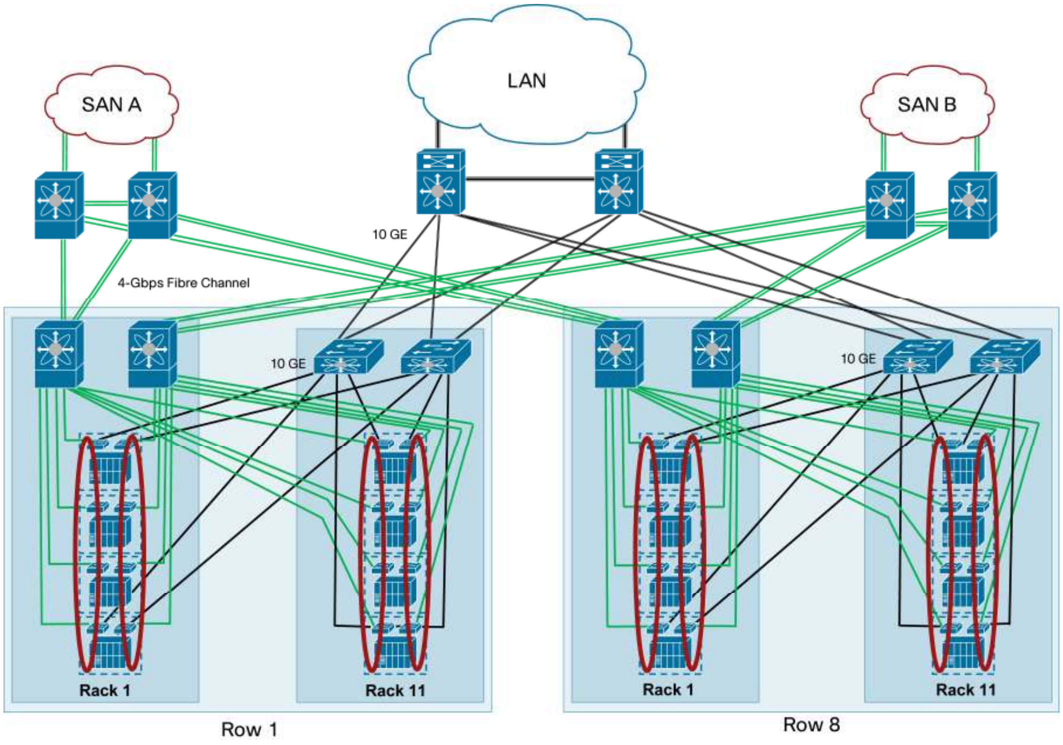


Table 2. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	HP c-Class	Cisco Nexus 5020	Cisco Nexus 7010
Modules	Cisco Catalyst Blade Switch 3120 for HP BladeSystem c-Class	• 1 40-port 10-Gigabit Ethernet/FCoE fixed • 2 6-port 10-Gigabit Ethernet/FCoE modules	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 4 per rack	Uplinks = 12 per row	• Uplinks = 4 per switch • Crosslinks = 6 per switch
Uplink Oversubscription Ratio	Server: Access = 1.6:1	Access: Aggregation = 7.67:1	Aggregation: Core = 5.4:1
Cabling	• Intra-Rack: Stackwise • Uplinks: Fiber (X2)	Fiber (SFP+)	Fiber (SFP+)
Storage			
	Edge	**Aggregation**	**Core**
Chassis Type	HP c-Class	Cisco MDS 9509	Cisco MDS 9509
Modules	Cisco MDS 9124e Fabric Switch	5 DS-X9112	5-DS-X9112
Number of Links	• 48 servers per rack • Uplinks = 8 per rack	Uplinks = 12	Crosslinks = 4

Servers per Rack	Cisco Nexus 5020 Switches per Rack	Racks per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
64	–	11	704	2

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
8	5632	16	2

Scenario 2: I/O Consolidation with Standalone Servers (Cisco Catalyst 6500 Series, Nexus 7000 Series, Nexus 5000 Series, and MDS 9000 Family)

The scenario 2 design (Figure 18 and Table 3) uses FCoE connectivity from the hosts to the Cisco Nexus 5000 Series access layer. The links from the hosts are in an active-standby configuration. The Cisco Nexus 5000 Series Switches provide connectivity to the Ethernet and SAN A and SAN B fabrics. Cisco Catalyst 6500 Series Switches are used in the aggregation layer to provide connectivity to the entire row, which is then aggregated into a pair of Cisco Nexus 7000 Series Switches. This design uses the existing Cisco Catalyst 6500 Series EoR investment and introduces new access layer switches in a gradual fashion to provide an immediate ROI, with savings in power, cabling, cooling, etc. as mentioned earlier in this document. The topology introduces FCoE and preserves existing Cisco Catalyst 6500 Series EoR environments.

Figure 18. I/O Consolidation with Standalone Servers with Cisco Catalyst 6500 Series at Aggregation Layer

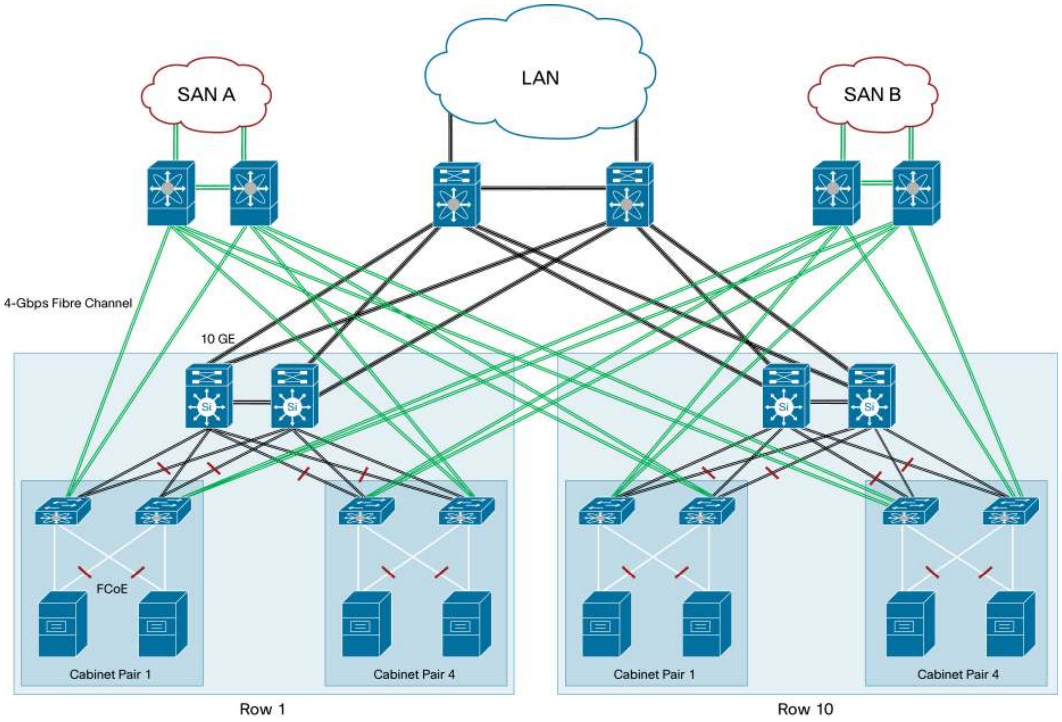


Table 3. Design Details

<table>
<tr><th colspan="4">Ethernet</th></tr>
<tr><td></td><td><b>Access Layer</b></td><td><b>Aggregation Layer</b></td><td><b>Core Layer</b></td></tr>
<tr><td><b>Chassis Type</b></td><td>Cisco Nexus 5020</td><td>Cisco Catalyst 6509</td><td>Cisco Nexus 7010</td></tr>
<tr><td><b>Modules</b></td><td>• 1 40-port 10 Gigabit Ethernet/FCoE fixed
• 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules</td><td>• 1 Cisco Catalyst 6500 Series Supervisor Engine 720 with 10 Gigabit Ethernet
• 8 WS-X6708-10G-3C modules</td><td>• 2 Supervisor 1 Engine
• 8 N7K-M132XP-12</td></tr>
<tr><td><b>Number of Links</b></td><td>Uplinks = 12 per Cabinet Pair</td><td>Uplinks = 8 per row
Crosslinks = 4 per switch</td><td>• Uplinks = 8 per switch
• Crosslinks = 8 per switch</td></tr>
<tr><td><b>Uplink Oversubscription Ratio</b></td><td>Server: Access = 1:1</td><td>Access: Aggregation = 6.67:1</td><td>Aggregation: Core = 6:1</td></tr>
<tr><td><b>Cabling</b></td><td>• Intra-Rack: Twinax (SFP+)
• Uplinks: Fiber (SFP+)</td><td>Fiber (SFP+)</td><td>Fiber (SFP+)</td></tr>
<tr><th colspan="4">Storage</th></tr>
<tr><td></td><td colspan="2"><b>Edge</b></td><td><b>Core</b></td></tr>
<tr><td><b>Chassis Type</b></td><td colspan="2">Cisco Nexus 5020</td><td>Cisco MDS 9513</td></tr>
</table>

Modules	2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	8-DS-X9112 for 32 cabinet pairs
Number of Links	● 40 servers per cabinet pair ● Uplinks = 12 per cabinet pair	Crosslinks = 4 per switch

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	4	160	8

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 6509 Switches per Pod
10	1600	80	20

Scenario 3: 10 Gigabit Ethernet Aggregation with Standalone Servers

The scenario 3 design (Figure 19 and Table 4) uses the Cisco Nexus 5000 Series as a ToR switch and the Cisco Nexus 7000 Series as aggregation and core switches. To address the growing requirements for bandwidth in high-density server farms, a full nonblocking 10 Gigabit Ethernet port is required for all servers in the rack. Two Cisco Nexus 5000 Series Switches are provided as a ToR solution, which allows all servers to access switch cabling contained within the rack. Twinax cabling provides a low-cost solution, providing 10 Gigabit Ethernet capabilities to the servers. 10 Gigabit Ethernet host connectivity also simplifies data center consolidation as more virtual machines can be installed on one physical server to increase asset utilization. Rapid Per-VLAN Spanning Tree Plus (PVST+) Spanning Tree Protocol is required to prevent loops in the topology. This design is recommended for creating a high-density 10 Gigabit Ethernet pod.

Figure 19. 10 Gigabit Ethernet Aggregation with Standalone Servers

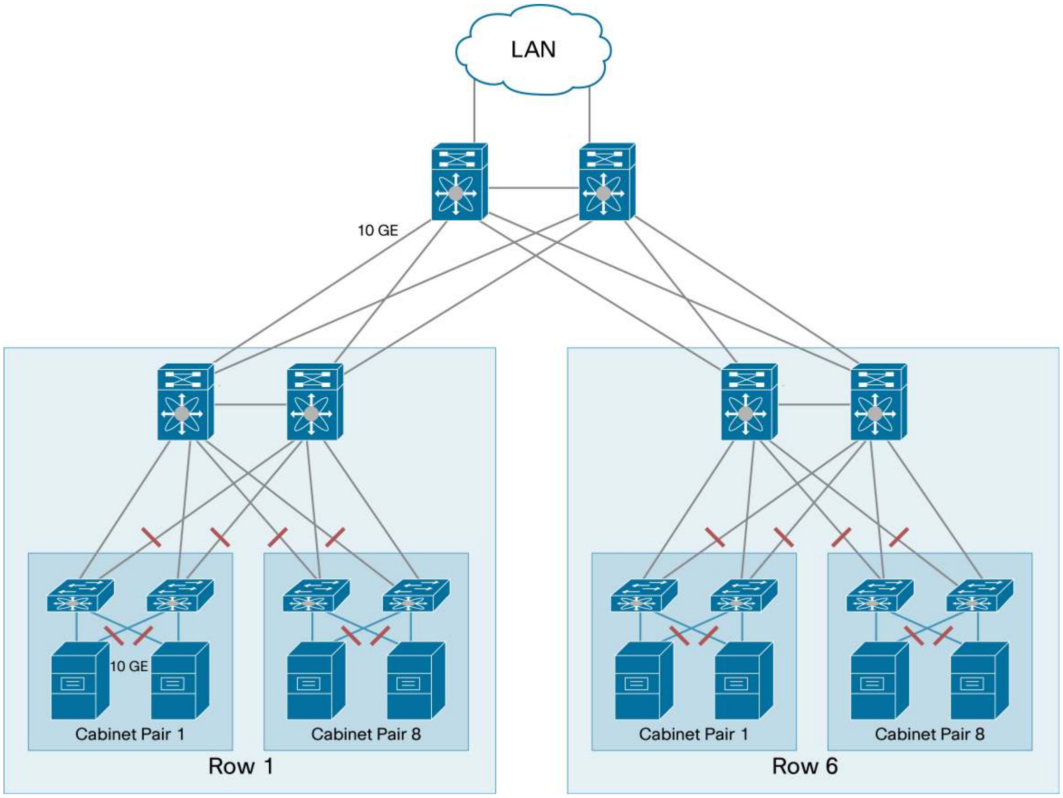


Table 4. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	Cisco Nexus 5020	Cisco Nexus 7010	Cisco Nexus 7010
Modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 6-port 10 Gigabit/FCoE modules	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 12 per cabinet pair	• Uplinks = 16 per row • Crosslinks = 8 per switch	• Uplinks = 8 per switch • Crosslinks = 8 per switch
Uplink Oversubscription Ratio	Server: Access = 1:1	Access: Aggregation = 6.67:1	Aggregation: Core = 6:1
Cabling	• Intra-Rack: Twinax (SFP+) • Uplinks: Fiber (SFP+)	Fiber (SFP+)	Fiber (SFP+)

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	8	320	16

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
6	1920	96	14

Scenario 4: I/O Consolidation with Standalone Servers

The scenario 4 design (Figure 20 and Table 5) uses Cisco Nexus 5000 Series Switches at the access and aggregation layers to construct a large Layer 2 topology. FCoE is used for connectivity to the host for I/O consolidation. Fibre Channel and Ethernet connectivity is provided by the Cisco Nexus 5000 Series at the access layer. Rapid PVST+ Spanning Tree Protocol is required in the design to prevent potential loops. The core is serviced by a pair of Cisco Nexus 7000 Series Switches servicing the entire pod. The large Layer 2 design is suitable for high-performance computing (HPC) environments where throughput and latency are the major factors in optimal performance. An end-to-end cut-through architecture with PFC provides high performance with congestion management.

Figure 20. I/O Consolidation with Standalone Servers with Cisco Nexus 5000 Series at Aggregation Layer

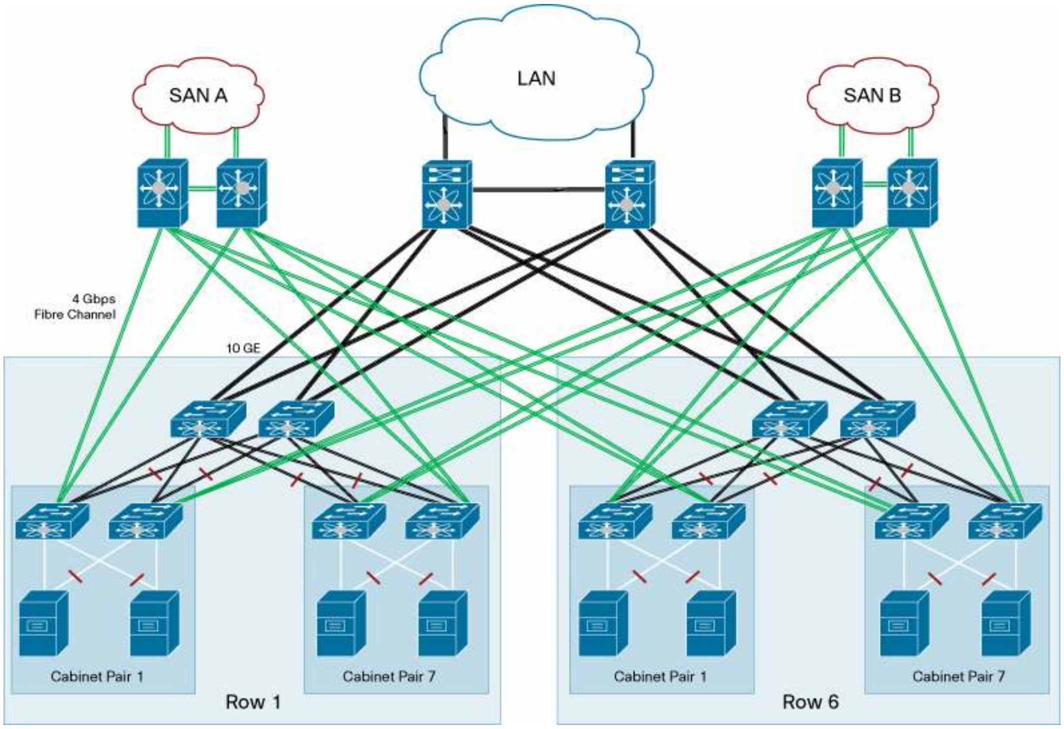


Table 5. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	Cisco Nexus 5020	Cisco Nexus 5020	Cisco Nexus 7010
Modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 6-port 10 Gigabit Ethernet/FCoE modules	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 12 per cabinet pair	Uplinks = 16 per row	• Uplinks = 8 per switch • Crosslinks = 8 per switch
Uplink Oversubscription Ratio	Server: Access = 1:1	Access: Aggregation = 6.67:1	Aggregation: Core = 5.25:1
Cabling	• Intra-Rack: Twinax (SFP+) • Uplinks: Fiber (SFP+)	Fiber (SFP+)	Fiber (SFP+)

Storage		
	Edge	**Core**
Chassis Type	Cisco Nexus 5020	Cisco MDS 9513
Modules	2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	8-DS-X9112 for 32 cabinet pairs
Number of Links	• 40 servers per cabinet pair • Uplinks = 12 per cabinet pair	Crosslinks = 4 per switch

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	7	280	14

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
6	1680	70	2

Scenario 5: I/O Consolidation with Standalone Servers

The scenario 5 design (Figure 21 and Table 6) introduces I/O consolidation at the access layer by using FCoE to carry Ethernet and Fibre Channel on a single wire. This approach eliminates the need for multiple HBAs and NICs in the hosts and separate Ethernet and Fibre Channel switches at the access layer. CNAs are used to unify LAN and SAN connectivity to Cisco Nexus 5000 Series Switches while keeping the distribution layer unchanged. The Fibre Channel uplinks from the FCoE switches connect to SAN A and SAN B, which are two separate SAN fabrics.

PFC is an important characteristic in the design, creating a lossless fabric to transport Fibre Channel frames. The delayed drop feature helps absorb microbursts within server farms. The Cisco Nexus 7000 Series Switches provide high-density 10 Gigabit Ethernet at the aggregation and core layers, and Cisco MDS 9000 family directors provide the SAN connectivity. This design is recommended for FCoE connectivity to the servers while creating a scalable 10 Gigabit Ethernet architecture to support multiple server farms that have large SAN and LAN requirements.

Figure 21. I/O Consolidation with Standalone Servers with Cisco Nexus 7000 Series at Aggregation Layer

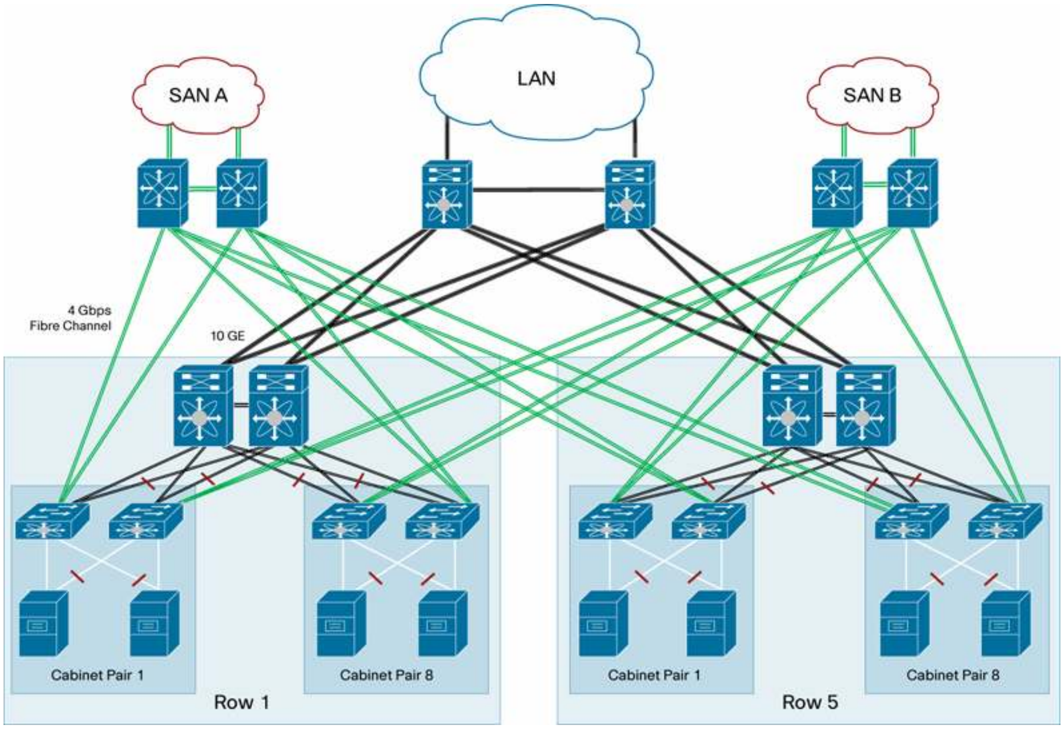


Table 6. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	Cisco Nexus 5020	Cisco Nexus 7010	Cisco Nexus 7010
Modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 12 per cabinet pair	• Uplinks = 16 per row • Crosslinks = 8 per switch	• Uplinks = 8 per switch • Crosslinks = 8 per switch
Uplink Oversubscription Ratio	Server: Access = 1:1	Access: Aggregation = 6.67:1	Aggregation: Core = 6:1
Cabling	• Intra-Rack: Twinax (SFP+) • Uplinks: Fiber (SFP+)	Fiber (SFP+)	Fiber (SFP+)
Storage			
	Edge	**Core**	

Chassis Type	Cisco Nexus 5020	Cisco MDS 9513
Modules	2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	8-DS-X9112 for 32 cabinet pairs
Number of Links	● 40 servers per cabinet pair ● Uplinks = 12 per cabinet pair	Crosslinks = 4 per switch

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	8	320	16

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
5	1600	80	12

Scenario 6: I/O Consolidation with Virtual Switch System

The scenario 6 design (Figure 22 and Table 7) provides I/O consolidation with FCoE connectivity to the hosts and uses VSS technology on the Cisco Catalyst 6500 Series platform. VSS minimizes the dependency on Spanning Tree Protocol and doubles the uplink bandwidth from each rack by using multichassis Cisco EtherChannel capability to provide dual-active links from the ToR switches to the aggregation layer. The Cisco Nexus 5000 Series Switches provide connections to the Ethernet and SAN A and SAN B fabrics. The pair of Cisco Catalyst 6500 Series aggregation switches appear as one logical node in the topology, provide a single configuration point, and appear as one routing node with a unified control plane. The entire pod is aggregated by two core Cisco Nexus 7000 Series Switches. This design is recommended for deployments where Cisco Catalyst 6500 Series Switches are used as an EoR solution and I/O consolidation at the rack level is required.

Figure 22. I/O Consolidation with Virtual Switch System

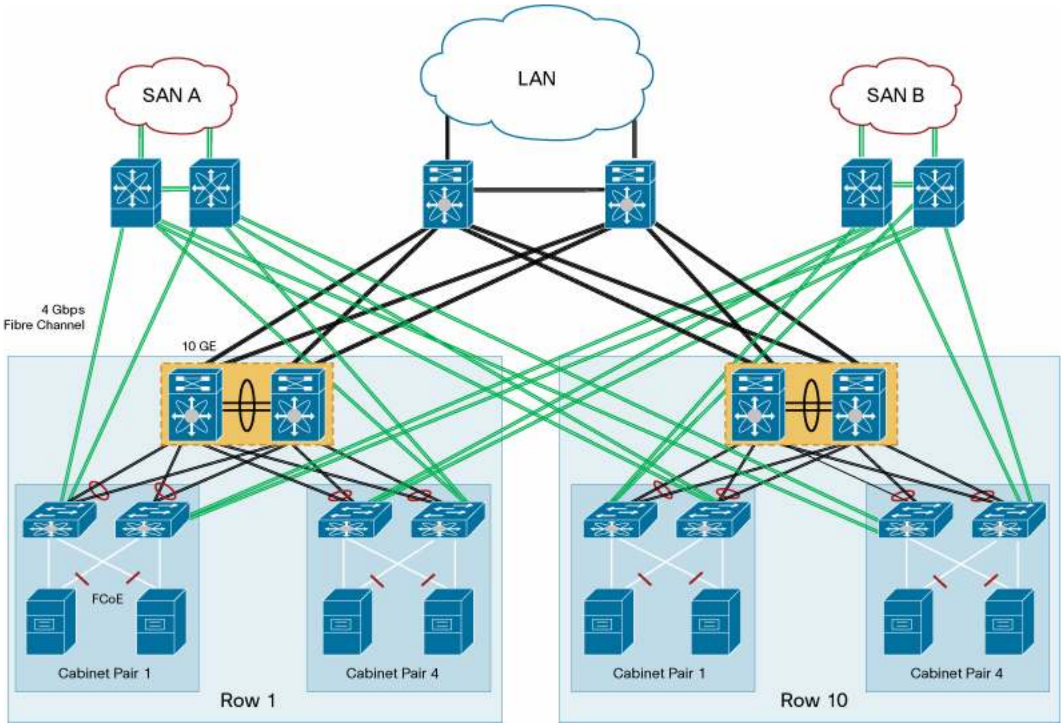


Table 7. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	Cisco Nexus 5020	Cisco Nexus 6509	Cisco Nexus 7010
Modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	• 1 Cisco Catalyst 6500 Series Supervisor Engine 720 with 10 Gigabit Ethernet • 8 WS-X6708-10G-3C modules	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 12 per cabinet pair	• Uplinks = 8 per row • Crosslinks = 4 per switch	• Uplinks = 8 per switch • Crosslinks = 8 per switch
Uplink Oversubscription Ratio	Server: Access = 1:1	Access: Aggregation = 6.67:1	Aggregation: Core = 6:1
Cabling	• Intra-Rack: Twinax (SFP+) • Uplinks: Fiber (SFP+)	Fiber (SFP+)	Fiber (SFP+)

Storage		
	Edge	**Core**
Chassis Type	Cisco Nexus 5020	Cisco MDS 9513
Modules	2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	8-DS-X9112 for 32 cabinet pairs
Number of Links	• 40 servers per cabinet pair • Uplinks = 12 per cabinet pair	Crosslinks = 4 per switch

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	4	160	8

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 6509 Switches per Pod
10	1600	80	20

Scenario 7: I/O Consolidation with End-Host Virtualizer

The scenario 7 design (Figure 23 and Table 8) highlights the EHV feature, which is used to provide an active-active topology between the access switch and the aggregation layer by reducing the dependency on Spanning Tree Protocol and minimizing the number of Spanning Tree Protocol computations. Both uplinks from the Cisco Nexus 5000 Series Switches are in the forwarding state, providing doubled bisectional bandwidth between Cisco Nexus 5000 and 7000 Series Switches. Failover of traffic to a secondary uplink is extremely fast due to minimal topology change notifications (TCNs). The Cisco Nexus 5020 Switches appear as a collection of end hosts to the aggregation switches. EHV support is required only on the Cisco Nexus 5000 Series Switches. This design is recommended for I/O consolidation of server farms where traffic forwarding on all uplinks from the Cisco Nexus 5000 Series Switches at the access layer is required.

Figure 23. I/O Consolidation with End-Host Virtualizer

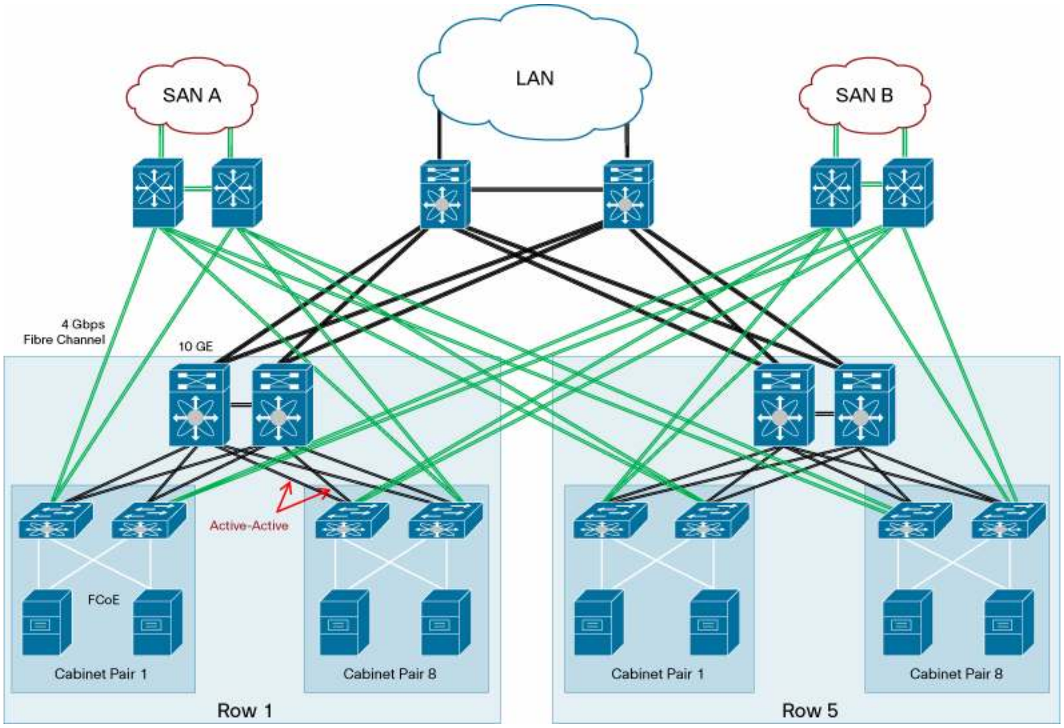


Table 8. Design Details

Ethernet			
	Access Layer	**Aggregation Layer**	**Core Layer**
Chassis Type	Cisco Nexus 5020	Cisco Nexus 7010	Cisco Nexus 7010
Modules	• 1 40-port 10 Gigabit Ethernet/FCoE fixed • 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12	• 2 Supervisor 1 Engine • 8 N7K-M132XP-12
Number of Links	Uplinks = 12 per cabinet pair	• Uplinks = 16 per row • Crosslinks = 8 per switch	• Uplinks = 8 per switch • Crosslinks = 8 per switch
Uplink Oversubscription Ratio	Server: Access = 1:1	Access: Aggregation = 6.67:1	Aggregation: Core = 6:1
Cabling	• Intra-Rack: Twinax (SFP+) • Uplinks: Fiber (SFP+)	Fiber (SFP+)	Fiber (SFP+)

Storage		
	Edge	**Core**
Chassis Type	Cisco Nexus 5020	Cisco MDS 9513
Modules	2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules	8-DS-X9112 for 32 cabinet pairs
Number of Links	• 40 servers per cabinet pair • Uplinks = 12 per cabinet pair	Crosslinks = 4 per switch

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	8	320	16

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
5	1600	80	12

Scenario 8: I/O Consolidation with NPV Mode

The scenario 8 design (Figure 24 and Table 9) uses the NPV feature of the Cisco Nexus 5000 Series Switches to address two major challenges of SAN deployment: limited domain IDs and multivendor interoperability. With NPV, the Cisco Nexus 5000 Series relays the FLOGI and FDISC commands to the upstream Fibre Channel switch. In this mode, the Cisco Nexus 5000 Series operates in N-port proxy mode. Since the Cisco Nexus 5000 Series does not provide Fibre Channel service, it does not require a Fibre Channel domain, which results in increased scalability. From the upstream Fibre Channel switch's point of view, the Cisco Nexus 5000 Series is just an N-node that sends multiple FLOGI and FDISC commands. The NPV switch requires NPIV support on the upstream F-port.

This design is especially useful when the host is virtualized and every virtual machine requires a separate FCID. NPIV support on the server-facing interfaces of the Cisco Nexus 5000 Series Switch provides a mechanism for assigning multiple FCIDs to a single N-port. This feature allows multiple applications on the N-port to use different identifiers and allows access control, zoning, and port security to be implemented at the application level. The NPIV proxy module in the Cisco Nexus 5000 Series Switch provides the proxy function of distributing FLOGI requests from the host over the available external Fibre Channel interfaces. The Fibre Channel HBAs in the servers and the upstream Fibre Channel switches act as if they are directly connected to each other using a physical cable.

Figure 24. I/O Consolidation with NPV Mode

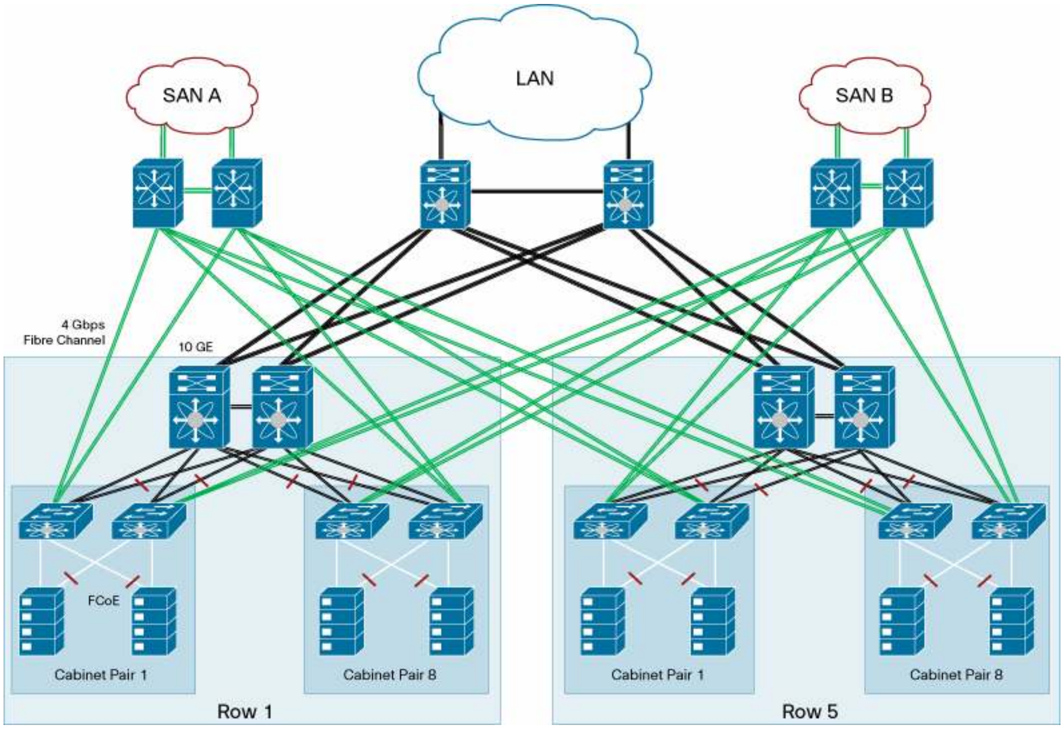


Table 9. Design Details

<table>
<tr><th colspan="5">Ethernet</th></tr>
<tr><th></th><th>Access Layer</th><th colspan="2">Aggregation Layer</th><th>Core Layer</th></tr>
<tr><td><b>Chassis Type</b></td><td>Cisco Nexus 5020</td><td colspan="2">Cisco Nexus 7010</td><td>Cisco Nexus 7010</td></tr>
<tr><td><b>Modules</b></td><td>● 1 40-port 10 Gigabit Ethernet/FCoE fixed
● 2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules</td><td colspan="2">● 2 Supervisor 1 Engine
● 8 N7K-M132XP-12</td><td>● 2 Supervisor 1 Engine
● 8 N7K-M132XP-12</td></tr>
<tr><td><b>Number of Links</b></td><td>Uplinks = 12 per cabinet pair</td><td colspan="2">● Uplinks = 16 per row
● Crosslinks = 8 per switch</td><td>● Uplinks = 8 per switch
● Crosslinks = 8 per switch</td></tr>
<tr><td><b>Uplink Oversubscription Ratio</b></td><td>Server: Access = 1:1</td><td colspan="2">Access: Aggregation = 6.67:1</td><td>Aggregation: Core = 6:1</td></tr>
<tr><td><b>Cabling</b></td><td>● Intra-Rack: Twinax (SFP+)
● Uplinks: Fiber (SFP+)</td><td colspan="2">Fiber (SFP+)</td><td>Fiber (SFP+)</td></tr>
<tr><th colspan="5">Storage</th></tr>
<tr><th></th><th colspan="2">Edge</th><th colspan="2">Core</th></tr>
<tr><td><b>Chassis Type</b></td><td colspan="2">Cisco Nexus 5020</td><td colspan="2">Cisco MDS 9513</td></tr>
<tr><td><b>Modules</b></td><td colspan="2">2 4-port 10 Gigabit Ethernet/FCoE and 4-port Fibre Channel modules</td><td colspan="2">8-DS-X9112 for 32 cabinet pairs</td></tr>
<tr><td><b>Number of Links</b></td><td colspan="2">● 40 servers per cabinet pair
● Uplinks = 12 per cabinet pair</td><td colspan="2">Crosslinks = 4 per switch</td></tr>
</table>

Servers per Cabinet Pair	Cisco Nexus 5020 Switches per Cabinet Pair	Cabinet Pairs per Row	Servers per Row	Cisco Nexus 5020 Switches per Row
40	2	8	320	16

Rows per Pod	Servers per Pod	Cisco Nexus 5020 Switches per Pod	Cisco Nexus 7010 Switches per Pod
5	1600	80	12

Conclusion

The rapidly changing data center environment is forcing architectures to evolve and address the challenges of power and cooling, server density, and virtualization. Cisco Nexus 5000 Series Switches provide support for FCoE and IEEE Data Center Bridging standards, delivering the capability to consolidate IP, storage, and interprocess communication (IPC) networks over a Ethernet based, Unified Fabric. Cisco Nexus 5000 Series Switches provide investment protection for current data center assets while maintaining the performance and operating characteristics necessary to keep pace with growing business needs for many years to come.

For More Information

- **Cisco Nexus 5000 Series:** http://www.cisco.com/en/US/products/ps9670/index.html
- **Fibre Channel over Ethernet:** http://www.fcoe.com
- **Unified Fabric:** http://www.cisco.com/en/US/netsol/ns945/index.html
- **T11:** http://www.t11.org

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-473501-01 06/09