Chapter 1:
Data Center Design with Cisco Nexus Switches and Virtual PortChannel: Overview
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td><strong>Current Environments</strong></td>
<td>3</td>
</tr>
<tr>
<td>Consolidation</td>
<td>3</td>
</tr>
<tr>
<td>End of the Row and Top of the Rack</td>
<td>3</td>
</tr>
<tr>
<td>Layer 2</td>
<td>4</td>
</tr>
<tr>
<td>V-Shape</td>
<td>4</td>
</tr>
<tr>
<td>Common Designs</td>
<td>4</td>
</tr>
<tr>
<td><strong>Next-Generation Data Centers</strong></td>
<td>5</td>
</tr>
<tr>
<td>All Links Forwarding</td>
<td>6</td>
</tr>
<tr>
<td>Server Connectivity at 10 Gigabit Ethernet</td>
<td>6</td>
</tr>
<tr>
<td>Fabric Extender</td>
<td>6</td>
</tr>
<tr>
<td>Unified I/O Support</td>
<td>6</td>
</tr>
<tr>
<td>Cut-through Operations and Latency</td>
<td>6</td>
</tr>
<tr>
<td>Using Cisco Nexus Switches and vPC to Design Data Centers</td>
<td>6</td>
</tr>
<tr>
<td><strong>Technology Overview</strong></td>
<td>7</td>
</tr>
<tr>
<td>Virtual PortChannel</td>
<td>7</td>
</tr>
<tr>
<td>vPC Forwarding</td>
<td>8</td>
</tr>
<tr>
<td>vPC and VSS</td>
<td>11</td>
</tr>
<tr>
<td>Spanning Tree Protocol</td>
<td>12</td>
</tr>
<tr>
<td>10 Gigabit Ethernet to the Server</td>
<td>13</td>
</tr>
<tr>
<td>Teaming on the Server</td>
<td>13</td>
</tr>
<tr>
<td>Unified I/O</td>
<td>14</td>
</tr>
<tr>
<td>FCoE</td>
<td>14</td>
</tr>
<tr>
<td>Virtual N-Ports and Virtual F-Ports</td>
<td>14</td>
</tr>
<tr>
<td>The VLAN-VSAN Relationship</td>
<td>15</td>
</tr>
<tr>
<td>FIP Protocol</td>
<td>15</td>
</tr>
<tr>
<td>IEEE DCB</td>
<td>15</td>
</tr>
<tr>
<td>DCBX</td>
<td>15</td>
</tr>
<tr>
<td>VN-Link and VN-TAG</td>
<td>16</td>
</tr>
<tr>
<td><strong>Cisco Nexus Products in This Design Guide</strong></td>
<td>17</td>
</tr>
<tr>
<td>Cisco Nexus 7000 Series Overview</td>
<td>17</td>
</tr>
<tr>
<td>Cisco Nexus 5000 Series Overview</td>
<td>18</td>
</tr>
<tr>
<td>Cisco Nexus 2000 Series Overview</td>
<td>18</td>
</tr>
<tr>
<td>Cisco Nexus 1000V Series Overview</td>
<td>19</td>
</tr>
<tr>
<td>Software Revisions</td>
<td>20</td>
</tr>
<tr>
<td><strong>Core, Aggregation, and Access Layers</strong></td>
<td>20</td>
</tr>
<tr>
<td><strong>Data Center Pods</strong></td>
<td>21</td>
</tr>
<tr>
<td>Cabling Considerations</td>
<td>21</td>
</tr>
<tr>
<td><strong>For More Information</strong></td>
<td>22</td>
</tr>
</tbody>
</table>
Introduction

This collection of white papers provides guidelines for designing and deploying the access and aggregation layers in the data center using Cisco Nexus® products.

This document covers the following topics:

- Data Center Design with Cisco Nexus Switches and Virtual PortChannel: Overview (Chapter 1, this document)
- Cisco NX-OS Software Command-Line Interface Primer (Chapter 2)
- Cisco NX-OS Software Virtual PortChannel: Fundamental Concepts (Chapter 3)
- Spanning Tree Design Guidelines for Cisco NX-OS Software and Virtual PortChannels (Chapter 4)
- Data Center Aggregation Layer Design and Configuration with Cisco Nexus Switches and Virtual PortChannel (Chapter 5)
- Data Center Access Design with Cisco Nexus 5000 Series Switches and 2000 Series Fabric Extenders and Virtual PortChannels (Chapter 6)
- 10 Gigabit Ethernet Connectivity with Microsoft Windows Servers (Chapter 7)
- Data Center Design with VMware ESX 4.0 and Cisco Nexus 5000 and 1000V Series Switches and 2000 Series Fabric Extenders (Chapter 8)

This document is a natural extension of Data Center Design—IP Network Infrastructure published at the following link: http://www.cisco.com/en/US/docs/solutions/Enterpr ise/Data_Center/DC_3_0/DC-3_0_IPInfra.html.


Current Environments

Most data centers today have been built using well-known Cisco® multilayer topology with core, aggregation, and access layers. Typical data centers use the concept of the pod, which is a building block for the access layer of the data center. A pod may be made of several racks or an entire row of machines.

Consolidation

Most data centers are the result of a consolidation process that was facilitated by the availability of bandwidth to connect remote locations to centralized data centers. During the consolidation process, enterprises and service providers simplified the network and the storage infrastructure by adopting topologies that take advantage of virtual local area networks (VLANs) and the virtual storage area network (VSAN) technology.

The same process of simplification is currently happening on the computing nodes. It involves both physical consolidation by means of blade enclosures and server virtualization at the operating system (OS) level, which makes it possible to run multiple instances of various operating systems on the same physical machine.

End of the Row and Top of the Rack

Most existing data centers are deployed according to two design philosophies:

- **End-of-the-row topologies**: This topology consists of large, director-class switching devices at the end of each row of servers. End-of-the-row topologies require significant cabling bulk to be carried from all server racks to the network rack. The main advantage of end-of-the-row topologies is the fact that fewer configuration points (switches) control a large number of server ports.
• **Top-of-the-rack topologies**: This topology consists of one-rack unit (1RU) or 2RU devices at the top or bottom of each server rack, providing server (or blade server) connectivity within each rack and aggregated by a switch at the aggregation layer. Top-of-the-rack topologies are more efficient in terms of cabling because fewer cables are required from each rack to the end-of-the-row switch. On the other hand, top-of-the-rack topologies require more switches than end-of-the-row topologies require for the same number of switch ports, which increases the management burden.

**Layer 2**
Layer 2 has been well known for its flexibility in providing VLAN access anywhere in the data center. Layer 2 is also known for the potential impact of Layer 2 loops. Overall, the benefits of Layer 2 largely outweigh the risk of loops, so most environments today consist of Layer 2 pods or even have multiple pods that are part of the same Layer 2 domain.

A Layer 3 boundary is always recommended to connect servers to clients. The placement of the Layer 3 boundary defines the scale of the Layer 2 domain, with the result that the size of Layer 2 domains varies enormously between customers.

The factors that determine the size of Layer 2 domains include:

- The percentage of high-availability, clustered servers
- The need for virtual machine mobility
- The need to quickly provision new servers in the same subnet

The availability of new application technologies will influence new data centers toward more or less use of Layer 2, but the current trend indicates a permanent need for a scalable Layer 2 infrastructure.

Cisco is delivering technology that makes Layer 2 topologies more solid and more efficient in terms of bandwidth, and that helps you extend Layer 2 where needed.

**V-Shape**
While several network designs fit the requirements of consolidation, as well as the cabling requirements of end-of-the-row or top-of-the-rack topologies, the most widely deployed network design model is referred to as “V-shape,” whereby each “access” switch is dual-homed to the aggregation layer, and Spanning Tree Protocol keeps the topology free from loops.

**Common Designs**
Figure 1 illustrates a legacy design.
As Figure 1 indicates, the legacy design is a V-shape topology with access and aggregation layers, and with well-known placement of root and secondary root switches, well-known placement of Hot Standby Router Protocol (HSRP) primary and secondary devices, forwarding and blocking links from the access layer to the aggregation layer, and various hardening features in place to guarantee deterministic spanning-tree behavior upon link failures.

**Next-Generation Data Centers**

Today’s data centers are mostly driven by the following needs:

- The need for a higher level of reliability, with minimized downtime for updates and configuration changes: Once a consolidated architecture is built, it’s critical to keep it up and running with minimum disruption.
- The need to optimize the use of the data center network infrastructure by moving towards a topology where no link is kept idle, whereas legacy topologies based on Spanning Tree Protocol are known to be inefficient because of Spanning Tree Protocol blocking links or because of active/standby network interface card (NIC) teaming. This need is addressed by Layer 2 multipathing technologies such as Virtual PortChannels (vPCs).
- The need to optimize computing resources by reducing the rate of growth of physical computing nodes. This need is addressed by server virtualization.
- The need to reduce the time that it takes to provision new servers. This need is addressed by the ability to configure server profiles, which can be easily applied to hardware.
- The need to reduce overall power consumption in the data center. This need can be addressed with various technologies, including unified fabric (which reduce the number of adapters on a given server), server virtualization, and more power-efficient hardware.
- The need to increase computing power at a lower cost: More and higher-performance computing clouds are being built to provide a competitive edge to various enterprises.

The data center design drivers just listed call for capabilities such as:

- Architectures capable of supporting a SAN and a LAN on the same network (for power use reduction and server consolidation).
• Architectures that provide an intrinsic lower latency than traditional LAN networks, so that computing cloud can be built on the same LAN infrastructure as regular transactional applications.

• Architectures that provide the ability to distribute Layer 2 traffic on all available links.

• Simplified cabling: For a more efficient airflow, lower power consumption, and lower cost of deployment of high-bandwidth networks.

• Reduction of management points: It’s important to limit the impact of the sprawl of switching points (software switches in the servers, multiple blade switches, and so on).

All Links Forwarding
The next-generation data center provides the ability to use all links in the LAN topology by taking advantage of technologies such as virtual PortChannels (vPCs). vPCs enable full, cross-sectional bandwidth utilization among LAN switches, as well as between servers and LAN switches.

Server Connectivity at 10 Gigabit Ethernet
Most rackable servers today include redundant LAN-on-motherboard (LOM) interfaces for management, an integrated-lights-out (iLO) standard-based port, and one or more Gigabit Ethernet interfaces, and redundant host bus adapters (HBA). The adoption of 10 Gigabit Ethernet on the server simplifies server configuration by reducing the number of network adapters and providing enough bandwidth for virtualized servers. The data center design can be further optimized with the use of Fibre Channel over Ethernet (FCoE) to build a unified fabric.

Cost-effective 10 Gigabit Ethernet connectivity can be achieved by using copper twinax cabling with Small Form-Factor Pluggable Plus (SFP+) connectors.

A rackable server configured for 10 Gigabit Ethernet connectivity may have an iLO port, a dual-LOM, and a dual-port 10 Gigabit Ethernet adapter (for example, a converged network adapter). This adapter would replace multiple Quad Gigabit Ethernet adapters and, in case the adapter is also a Cisco Network Adapter, it would also replace an HBA.

Fabric Extender
Fabric extender technology simplifies the management of the many LAN switches in the data center by aggregating them in groups of 10 to 12 under the same management entity. In its current implementation, Cisco Nexus 2000 Series Fabric Extenders can be used to provide connectivity across 10 to 12 racks that are all managed from a single switching configuration point, thus bringing together the benefits of top-of-the-rack and end-of-the-row topologies.

Unified I/O Support
Significant cost reduction can be achieved by replacing Quad Gigabit Ethernet cards and dual HBAs with a dual-port converged network adapter card connected to a data-center-bridging (DCB)-capable device. A device such as the Cisco Nexus 5000 Series Switch also provides Fibre Channel Forwarder (FCF) capability.

The next-generation data center provides no-drop capabilities through priority flow control (PFC) and a no-drop switch fabric for suitable traffic types.

Cut-through Operations and Latency
While not mandatory, using a cut-through-capable access layer enables low-latency communication between servers for any packet size. The Cisco Nexus 5000 Series supports deterministic 3.2- microseconds latency for any packet size with all features enabled (with access control list [ACL] filtering applied, as an example) and similarly the Cisco Nexus 4000 Series supports 1.5- microseconds latency.

Using Cisco Nexus Switches and vPC to Design Data Centers
Figure 2 illustrates how a next-generation data center would look like with Cisco Nexus Switches and vPC.
In a vPC topology, all links between the aggregation and the access layer would be forwarding and part of a virtual PortChannel.


**Figure 2.** Next-Generation Data Center Design

Connectivity from blade servers or servers connected through 10 Gigabit Ethernet can take advantage of the Cisco Nexus 5000 Series line-rate 10 Gigabit Ethernet forwarding capabilities and copper twinax support. Blade server connectivity to the Cisco Nexus 5000 Series can vary based on the type of switching device placed inside the blade server, but this architecture is flexible enough to support pass-through or other multiplexing technologies as needed.

In this topology, the Spanning Tree Protocol root and secondary root would still be placed at the aggregation layer on the Cisco Nexus 7000 Series Switches, which would also be the primary and secondary HSRP devices providing the gateway for the servers.

**Technology Overview**

**Virtual PortChannel**

Virtual PortChannel (vPC) allows links that are physically connected to two different Cisco switches to appear to a third, downstream device as coming from a single device and as part of a single port channel. The third device can be a switch, a server, or any other networking device that supports IEEE 802.3ad port channels.

The Cisco NX-OS Software Virtual PortChannel and Cisco Catalyst® 6500 Virtual Switching System (VSS) 1440 are similar technologies. With regard to Cisco EtherChannel® technology, the term "multichassis EtherChannel (MCEC)" refers to either technology interchangeably.
vPC allows the creation of Layer 2 port channels that span two switches. At the time of this writing, vPC is implemented on the Cisco Nexus 7000 and 5000 Series platforms (with or without the Nexus 2000 Series Fabric Extender).

Data centers based on Cisco Nexus technology can use vPC technology to build redundant loop-free topologies. Data centers that use Cisco Catalyst technology can use VSS technology for the same purpose.

vPC Forwarding
The main benefit of running vPC on the data center is that the traffic between client and servers or from server to server can use all the available links, as illustrated by Figures 3 and 4.

As Figure 3 shows, client-to-server traffic coming from the core would be first distributed based on Layer 3 Equal Cost Multipath (ECMP) to the two aggregation layer Cisco Nexus 7000 Series Switches, which would in turn distribute it to the access layer in the Cisco Nexus 5000 Series Switches.

Conversely, as Figure 4 shows, server-to-client traffic would take advantage of both Cisco Nexus 7000 Series servers, regardless of which one is HSRP active or HSRP standby. HSRP is configured identically to non-vPC topologies, but it has been modified to allow forwarding on both active and standby.
Adding vPC to the Cisco Nexus 5000 Series in the access layer allows further load distribution from the server to the fabric extenders to the Cisco Nexus 5000 Series.

With vPC, the proper tuning delivers the following traffic distribution:

- Client-to-server with fabric extenders in straight-through mode, and server running port channeling on two ports, resulting in four paths (Figure 5)
- Client-to-server with fabric extenders in active/active mode, resulting in four paths (Figure 6)
- Server-to-client with fabric extenders in straight-through mode and server port channeling, producing four different paths (Figure 7)
- Server-to-client with fabric extenders in active/active mode, producing four different paths (Figure 8)
- Server-to-server with fabric extenders in active/active mode, resulting in two paths per direction

Although a server connected at Gigabit Ethernet speeds with two ports will not be able to generate more than 2 gigabits worth of traffic, load distribution along a greater number of paths means a more equal spread of the data center traffic along all the available links. The result is that the data center performs more efficiently.
Figure 5. Client-to-Server Traffic Fex Straight-Through

![Diagram of Client-to-Server Traffic Fex Straight-Through]

Figure 6. Client-to-Server Traffic Flows with Fex Active/Active

![Diagram of Client-to-Server Traffic Flows with Fex Active/Active]
Mixed data centers consisting of both Cisco Catalyst and Cisco Nexus products can interoperate by using both vPC and VSS technologies. For sake of completeness, this chapter includes a comparison of the two technologies, although a full comparison of vPC and VSS is beyond the scope of this document.

Both vPC and VSS allow the creation of PortChannels that span two switches. At the time of this writing, vPC is implemented on the Cisco Nexus 7000 and 5000 Series platforms. VSS is implemented on the Cisco Catalyst 6500 Virtual Switching System 1440 platform.

The fundamental concepts underlying vPC are described at the following site:
The fundamental concepts of VSS are described at the following site:  

The main difference between VSS and vPC is that vPC doesn’t unify the control plane, while VSS does. Table 1 summarizes the differences between the two technologies.

Table 1. VSS Compared with vPC

<table>
<thead>
<tr>
<th></th>
<th>Cisco VSS</th>
<th>Cisco vPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform Support</td>
<td>Cisco Catalyst 6500 Virtual Switching System 1440</td>
<td>Cisco Nexus 7000 and Nexus 5000 Series (Cisco NX-OS Software Release 4.1(3))</td>
</tr>
<tr>
<td>Control Plane</td>
<td>One single control plane</td>
<td>Separate control planes</td>
</tr>
<tr>
<td>In-Service Software Upgrade (ISSU)</td>
<td>Across chassis</td>
<td>Within a single system on the Cisco Nexus 7000 Series, there is a mechanism to prevent unwanted vPC configurations during ISSU.</td>
</tr>
<tr>
<td>Configuration Synchronization</td>
<td>Automatic</td>
<td>Manual, but assisted by protocol verification (Cisco Fabric Services)</td>
</tr>
<tr>
<td>EtherChannel to Prefer Local Links</td>
<td>Yes (both inbound and outbound)</td>
<td>Yes (both inbound and outbound)</td>
</tr>
<tr>
<td>Dual-Active Detection</td>
<td>Yes, through Enhanced Port Aggregation Protocol (EpagP), bidirectional forwarding detection (BFD), or Dual-Fast Hello</td>
<td>Yes, through fault-tolerant link (routed)</td>
</tr>
<tr>
<td>Port Channel Protocol</td>
<td>Link Aggregation Control Protocol (LACP), PagP</td>
<td>LACP</td>
</tr>
<tr>
<td>Ports in a Single Port Channel</td>
<td>8 without LACP, 16 (8 active, 8 standby) with LACP</td>
<td>8 without LACP or 16 (8 active, 8 standby) with LACP on a single Cisco Nexus 7000 Series or 16 all active with 8 active per Cisco Nexus 7000 Series, 16 all active on Cisco Nexus 5000 Series</td>
</tr>
<tr>
<td>Number of MCEC Supported</td>
<td>Up to 128 port channels per system in standalone mode, and 512 multichassis EtherChannels in VSS mode</td>
<td>The hardware supports 768 virtual PortChannels. The control plane supports as many as indicated in the release notes. The number is increasing as new releases are introduced and higher-density configurations are tested. 480 2-port virtual PortChannels (as defined in this document) on the Cisco Nexus 5000 Series with NX-OS Software Release 4.1(3)</td>
</tr>
<tr>
<td>Spanning Tree Neighbors</td>
<td>Single switch</td>
<td>Single switch as seen on the vPC ports (but 2 switches as seen on non-vPC ports)</td>
</tr>
<tr>
<td>Cisco Discovery Protocol Neighbors</td>
<td>One single neighbor</td>
<td>Each switch appears individually</td>
</tr>
<tr>
<td>Layer 2 and Layer 3 MCEC</td>
<td>Yes</td>
<td>vPC is by default a switch port, thus Layer 2</td>
</tr>
<tr>
<td>HSRP Configuration</td>
<td>Not required</td>
<td>Standard HSRP configuration, enhanced to prefer local forwarding</td>
</tr>
<tr>
<td>Back-to-Back MCEC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dual-Layer MCEC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

For more details on the vPC technology, please see Chapter 3.

Spanning Tree Protocol

Spanning Tree Protocol doesn’t disappear in vPC-based data centers. Spanning Tree Protocol still runs, except that dual-connected switches have no port blocking.

If links that are unbundled from the vPC end up in the Individual (I) state, Spanning Tree Protocol intervenes to prevent a loop. Hence vPC-based topologies provide double protection from Layer 2 loops.

Just as with regular non-vPC designs, with Spanning Tree Protocol the choice between Multiple Spanning Tree (MST) and Rapid Per-VLAN Spanning Tree Plus (Rapid PVST+) depends on the number of VLANs requested by the design and how many of these VLANs need to be on any individual links (logical interfaces count).

Chapter 4 covers the design choice of MST versus Rapid PVST+ in a vPC-based data center.
10 Gigabit Ethernet to the Server

Servers connected at 10 Gigabit Ethernet benefit from a reduction in the number of Gigabit Ethernet adapters needed. They also increase network bandwidth and the possibility to consolidate storage traffic with LAN traffic, as described in the section “Unified I/O.”

Most servers equipped with at least four cores and with 10 Gigabit Ethernet adapters capable of receive-side scaling (RSS), large segment offload (LSO), and enough memory can generate close to 9 Gbps of transmit (Tx) traffic with memory operations. (I/O operations can, of course, reduce the performance.) Servers with a greater number of cores can more easily take advantage of the available bandwidth both in transmit and receive.

For several applications to take advantage of the available bandwidth, proper tuning is needed in order to increase the socket buffer.

OS tuning with the latest OS versions is not necessary. As an example, Microsoft Windows 2008 by default includes auto-tuning of the receive window, with various levels:

- Restricted (up 64-KB TCP Rx [receive] window)
- Highly Restricted (up to 256-KB TCP Rx window)
- Normal (up to 16-MB TCP Rx window)
- Experimental (up to 1-GB TCP Rx window)

For more information on 10 Gigabit Ethernet tuning on Microsoft Windows servers, please see Chapter 7.

Virtualized servers can more easily take advantage of the additional bandwidth. Even if applications are often throttled by the socket size, with virtualized servers, the individual virtual machine traffic gets aggregated onto the same NIC. For this reason, 10 Gigabit Ethernet proves useful even if it’s not necessary to provide enough bandwidth.

In addition to this, VMware® VMotion™ migration is a pure memory operation; hence it can take advantage of the additional bandwidth more easily, which, in turn, may enhance simultaneous VMotion migration of multiple VMs.

Teaming on the Server

Servers provide several different options for teaming with names that vary according to the vendor. The most common options include:

- Active-standby
- Active-active transmit load balancing: With this option, only one NIC can transmit and receive, and all NICs can transmit. This configuration enhances the server transmit performance, but doesn’t improve the receive bandwidth.
- Static port channeling: Equivalent to channel-group-mode on, which is PortChannels without any negotiation protocol in place.
- IEEE 802.3ad port channeling: This option enables the negotiation of the PortChannel between server and switch, thus allowing the server administrator to know if the configuration was successful. Similarly, it gives the network administrator information on whether the server administrator configured teaming properly.

With vPC support on the switch, the last two options can be deployed with network adapters split on two different access switches, thus achieving increased bandwidth and redundancy at the same time.

In addition, the teamed adapters can be virtualized with VLANs, with each VLAN showing on the server as a physically separate adapter. This allows the consolidation of multiple adapters for increased aggregated bandwidth.

The choice of the teaming option depends on the topology and the configuration of the switching infrastructure. A vPC-based data center enables both static port channeling and 802.3ad port channeling with or without 802.1q VLAN partitioning on the NIC.
In case a given switch topology has not been enabled for vPC port channeling, the other teaming options remain valid.

For additional information on network adapter teaming, please refer to Chapter 7.

Unified I/O

This design document focuses on the LAN switch design of the data center with Cisco Nexus products, and not specifically on the SAN aspects of the design. But because the Cisco Nexus product line is a key enabler for unified I/O deployments, it is important to be aware of some key Fibre Channel over Ethernet characteristics. What’s more, the designs that are presented in this document can easily be turned into a unified I/O design.

Here are some of the key technology characteristics that you need to focus on to properly design the data center:

- The role of the Fibre Channel Forwarder (FCF)
- The VLAN-VSAN relationship
- Virtual PortChannel compatibility with virtual Fibre Channel interfaces
- The support of negotiation protocols, including FCoE Initialization Protocol (FIP) and Data Center Bridging Exchange (DCBX) Protocol on network adapters

FCoE

Fibre Channel over Ethernet (FCoE) is the key technology that enables unified I/O on servers. Unified I/O is the ability to carry both storage and LAN data traffic on the same network adapter. Although FCoE is not the key topic of this document, the designs proposed for LAN switching connectivity can be extended to support unified I/O through FCoE.

FCoE (which is referenced as FC-BB_E in the FC-BB-5 specifications) achieved standard status in June 2009, and is documented in the T11 publication. You can access this publication at: http://www.t11.org/ftp/t11/pub/fc/bb-5/09-056v5.pdf.

In the simplest terms, one could say that FCoE is just an encapsulation of Fibre Channel frames into Ethernet frames.

The FCoE model consists of servers with FCoE adapters (E-nodes) normally called converged network adapters (CNAs), a lossless fabric. A simple way to think of this is as an Ethernet network with the ability to pause FCoE frames, instead of dropping them, when there is congestion. The model also contains Fibre Channel Forwarders (FCFs), devices like the Cisco Nexus 5000 Series Switches, which can forward FCoE frames.

Starting from the NX-OS Software Release 4.1(3)N1, the Cisco Nexus 5000 Series supports the FCoE standard. A “prestandard” implementation of FCoE was already available on the Cisco Nexus 5000 Series prior to this release. In order to distinguish the two, the prestandard FCoE implementation is referred to as pre-FIP, while the standard implementation is referred to as FIP-compliant (that is, FCoE Initialization Protocol compliant).

The default class of service assignment for FCoE is 3, and the default bandwidth allocation for FCoE on a CNA card is 4 Gbps.

While the current Cisco Nexus 5000 Series hardware is capable of supporting both fabric provided MAC address (FPMA) and server provided MAC address (SPMA) addressing schemes for FCoE, the software currently supports the only addressing scheme requested by the standard, which is FPMA.

Virtual N-Ports and Virtual F-Ports

When an E-node (that is, a CNA) logs into the FCF (fabric), if the login is successful it establishes VN-to-VF virtual connectivity whereby the E-node behaves like an N-port and the FCF provides the virtual F-port functionality.
The virtual F-port differs from regular Fibre Channel ports in that it's not a physical port, existing instead as a construct on the FCF logic, because the links up and down are virtual.

The VLAN-VSAN Relationship
When deploying a unified fabric design, remember that each FC fabric is associated with an FCoE fabric. As a result, each virtual SAN is uniquely associated with a virtual LAN. The E-node discovers the VLAN on which to send the FCoE traffic through the FIP protocol.

FIP Protocol
The FCoE Initialization Protocol (FIP) enables the discovery of FCFs over a lossless Ethernet cloud. The FIP protocol is used for the CNA to discover the FCoE VLANs and remote virtual FC interfaces and to log in to and log out of the FCF. FIP also supports link initialization functions (FLOGI or FDISC or ELP).

First-generation CNAs are typically not FIP-capable (hence they are called pre-FIP). Second-generation CNAs are FIP-compliant.

FIP was added to the FC-BB-5 standard at a later stage; hence in the pre-FIP implementation of FCoE, some functions that are today supported by FIP were supported through the Data Center Bridging Exchange (DCBX) Protocol. Link status is one of them. The difference between FIP keepalive and DCBX is that with FIP keepalive, the link status is extended to verify connectivity between the E-node and the FCF, as opposed to verifying the status on a specific wire as DCBX does.

A significant difference between FIP and pre-FIP implementations is that a FIP-capable switch can support nondirect connectivity of E-nodes to FCF, as in the case of a FIP snooping bridge in the path.

IEEE DCB
In order for Ethernet to be suitable to carry both LAN traffic and SAN traffic, certain enhancements are needed to help ensure a no-drop transport behavior for Fibre Channel. This set of enhancements is being standardized by IEEE as the Data Center Bridging (DCB) Task Group: http://www.ieee802.org/1/pages/dcbridges.html.

According to IEEE, the enhancements include:

- **802.1Qbb**: “Priority-based Flow Control (PFC) provides a link level flow control mechanism that can be controlled independently for each priority. The goal of this mechanism is to ensure zero loss due to congestion in DCB networks.” (from http://www.ieee802.org/1/pages/802.1bb.html)

- **802.1Qaz**: “Enhanced Transmission Selection (ETS) provides a common management framework for assignment of bandwidth to traffic classes.” (from http://www.ieee802.org/1/pages/802.1az.html)

- **Always part of 802.1Qaz is Data Center Bridging Exchange Protocol**: The DCBX Protocol is an extension of the Link Layer Discovery Protocol (LLDP, http://www.ieee802.org/1/pages/802.1ab.html). The Cisco Nexus 5000 Series uses DCBX to convey the class-of-service (CoS) value to be used for FCoE to the CNA, the 802.1p CoS values, the scheduling information, and the maximum transmission unit (MTU) size. This protocol is also used to perform a virtual link up or link down between the virtual Fibre Channel interface and the CNA.

Unified I/O deployments today make use of the three technologies just described, which are implemented on the network connecting a Converged Network Adapter to a Cisco Nexus 5000 Series Switch.

DCBX
DCBX is a protocol that serves the following purposes:

- Discovery of DCB capabilities, in particular the ability to support priority flow control (PFC_, and priority groups
- The ability to configure the peer DCBX device (for example, configure PFC and bandwidth allocation on a CNA from a DCB capable switch, such as the Cisco Nexus 5000 Series)
As previously explained, DCBX is built on top of LLDP, taking advantage of the extensibility characteristics of LLDP. DCBX conveys to the CNA information on how to partition traffic among classes on the 10 Gigabit Ethernet CNA and more specifically between FCoE and regular IP traffic. It also creates a virtual point-to-point link between the virtual Fibre Channel interfaces on the switch and the CNA.

DCBX can be categorized as:

- **Cisco, Intel, Nuova DCBX**: This is the DCBX version that the Cisco Nexus 5000 Series supported prior to the release of NX-OS 4.1(3)N1.
- **Converged Enhanced Ethernet Data Center Bridging Exchange (CEE-DCBX) Protocol**: This is the new DCBX version after Cisco NX-OS Software Release 4.1(3)N1, which reflects the fact that some of the Cisco, Intel, Nuova DCBX functions have been ported to FIP.

The Cisco, Intel, Nuova DCBX implementation took care of some functionality that is now part of FIP. When deciding the deployment mode of the Cisco Nexus 5000 Series, you need to consider whether the CNA adapters support:

- Pre-FIP and Cisco, Intel, Nuova DCBX
- FIP and CEE-DCBX
- Both of the above

Because the Cisco Nexus 5000 Series supports both, it can support Gen1 and Gen2 CNAs, and Link Layer Discovery Protocol (LLDP) can be enabled or disabled on a per-interface basis (and it is enabled by default).

**VN-Link and VN-TAG**

Bridges use the MAC address table as the basis for traffic forwarding. When a frame is received on one of the bridge's interfaces, the bridge looks up the frame's destination address in its internal table. If the table contains an association between the destination address and any of the bridge's ports aside from the one on which the frame was received, the frame is forwarded from the indicated port. If no association is found, the frame is flooded to all ports except the inbound port.

With virtualized servers, it is obvious that two virtual machines that are on the same VLAN need to be able to talk to each other. Communication between the virtual machines requires additional technology.

Cisco's solution to this problem is the introduction of VN-link technology which, at the time of this writing, is provided through the Cisco Nexus 1000V Series Switches. These switches provide a virtual distributed switch for VMware ESX servers.

As Figure 9 illustrates, in the absence of a switching component on the server, virtual-machine-to-virtual-machine communication would not be possible: a switch would not forward frames destined to MAC2 out of the same port (1/1) from which it received the frames.
Cisco Nexus Products in This Design Guide

All Cisco Nexus products run Cisco NX-OS Software, a modular operating system that provides a higher level of availability as a result of several enhancements such as process restartability and memory protection. Chapter 2 in this guide explains the characteristics NX-OS and how it compares with Cisco Catalyst IOS® Software.

Cisco Nexus 7000 Series Overview

The Cisco Nexus 7000 Series Switch is a modular switch available in a 10-slot or 18-slot configuration.

The Cisco Nexus 7010 Switch features front-to-back cooling compatible with data center hot-aisle and cold-aisle designs.

Cisco Nexus 7000 Series Switches provide high-density 10 Gigabit Ethernet port aggregation. With the 10-slot chassis, the system is capable of an aggregate density of 256 10 Gigabit Ethernet ports, including up to 64 ports of wire-rate 10 Gigabit Ethernet.

The current 32-port 10 Gigabit Ethernet modules support 80 gigabits of bandwidth per slot in the system backplane, and offer the choice to operate them in “dedicated” or “shared” mode for eight non-blocking 10 Gigabit Ethernet ports on a single I/O module. Gigabit Ethernet as well as 10 Gigabit Ethernet modules support IEEE 802.1AE MAC security with hardware-based 128-bit Advanced Encryption Standard (AES) encryption.

The Cisco Nexus 7000 Series offers control plane virtualization with virtual device contexts (VDCs).

The Cisco Nexus 7000 Series supports the ability to forward on all uplinks in typical V-shape or square topologies by using Virtual PortChannel technology.

The platform supports up to five hot-swappable, redundant switch fabric modules.

More details on the Cisco Nexus 7000 Series, VDCs, and how to design the aggregation layer with vPCs can be found in Chapter 5 of this guide.
Cisco Nexus 5000 Series Overview
The Cisco Nexus 5000 Series is a 52-port 10 Gigabit Ethernet switch capable of transporting Fibre Channel and Fibre Channel over Ethernet.

The Cisco Nexus 5000 Series features front-to-back cooling compatible with data center hot-aisle and cold-aisle designs, with all switch ports at the rear of the unit in close proximity to server ports.

The Cisco Nexus 5000 Series provides 40 fixed 10 Gigabit Ethernet ports, each capable of FCoE, of which the first eight fixed ports support both 10 Gigabit Ethernet and 1 Gigabit Ethernet. The expansion modules accept Fibre Channel ports, 10 Gigabit Ethernet ports capable of FCoE, or a hybrid module with both Fibre Channel and 10 Gigabit ports.

The Cisco Nexus 5000 Series can be used for top-of-the-rack deployments and also to aggregate the 10 Gigabit links in multiple racks. They can also be used in conjunction with the Nexus 2000 Series to provide Gigabit Ethernet top-of-the-rack connectivity across 10–12 racks.

The Cisco Nexus 5000 Series provides low-latency, cut-through forwarding independent of the packet size of traffic, and line-rate, nonoversubscribed 10 Gigabit Ethernet wired-rate forwarding for all ports. The single-stage fabric ensures no switch-induced congestion.

The Cisco Nexus 5000 Series can be used to connect servers equipped with converged network adapters and split Fibre Channel traffic from IP traffic.

The Cisco Nexus 5000 Series also supports Virtual PortChannels and can aggregate up 16 10 Gigabit Ethernet ports (all active).

The Cisco Nexus 5000 Series is also capable of switching frames based on the virtual network tag (VNTag) for further integration with I/O virtualization (IOV)-capable adapters and virtualized servers.

More details on how to design with the Cisco Nexus 5000 Series at the access layer can be found in Chapter 6.

Cisco Nexus 2000 Series Overview
The Cisco Nexus 2000 Fabric Extenders provide an extension of the Cisco Nexus 5000 Series through modules that operate as satellites of the Cisco Nexus 5000 Series Switches. At the time of this writing, the Cisco Nexus 2148T Fabric Extender provides 48 Gigabit Ethernet server ports and four times 10 Gigabit Ethernet uplink ports in a 1RU form factor. The Cisco Nexus 2148T Fabric Extender features front-to-back cooling compatible with data center hot-aisle and cold-aisle designs, with all switch ports at the rear of the unit in close proximity to server ports.

The Cisco Nexus 2148T Fabric Extender forwards all traffic to the parent Cisco Nexus 5000 Series Switches over 10 Gigabit Ethernet uplinks. When all four fabric uplinks are used, the Cisco Nexus 2148T can support an oversubscription ratio close to 1:1.

All forwarding for the fabric extender is performed on the Cisco Nexus 5000 Series Switch, which is a single point of management for all fabric extenders. The fabric extender does not run Spanning Tree Protocol even when multiple links are connected to the parent switch, providing loop-free, active-active connectivity between the two.

The Cisco Nexus 2148T Fabric Extender provides 48 Gigabit Ethernet ports supporting 1G-BASE-T through RJ-45 connectors.

It provides four 10 Gigabit Ethernet uplinks through Small Form-Factor Pluggable Plus (SFP+) ports using short-reach (SR) optics with fiber or cost-efficient, copper-based twinax (CX1 direct attach) connectivity.
Pod-based architecture can benefit from the fabric extender technology in two ways:

- Fabric extenders provide simplified cabling.
- Rack-and-roll deployment: The rack can be precabled and then attached to a Cisco Nexus 5000 Series from which the fabric extender gets the code and the configuration.

For more details on how to design with the Cisco Nexus 5000 and 2000 Series at the access layer, see Chapter 6.

Cisco Nexus 1000V Series Overview

The Cisco Nexus 1000V Series Switches are virtual distributed switches providing Cisco NX-OS Software-based switching for virtualized servers.

The Cisco Nexus 1000V Switch is modeled as a chassis switch, with a data plane component called Virtual Ethernet Modules (VEMs), which run on the individual virtualized servers and are managed by the equivalent of a supervisor, called Virtual Supervisor Module (VSM). The VSM provides a centralized point of configuration for up to 64 VEMs.

Figure 10 illustrates the abstraction model provided by the Cisco Nexus 1000V Series Switches. In this model, the hardware switching function, provided by an access switch in this example, is the functional equivalent of a fabric in a modular switch.

This list provides information on some of the advantages of using a Cisco Nexus 1000V Switch for virtual machine switching:

- The ability for the network administrator to configure port profiles (port groups) with a Cisco NX-OS Software command-line interface. These profiles appear on VMware vCenter as port groups that the server administrator can choose from.
- The ability to configure network properties through port profiles without having to configure each VMware ESX host in the same cluster independently. This means that all 64 VMware ESX hosts that are under the same Cisco Nexus 1000V Switch are automatically configured when a new port profile is defined. The hosts are therefore ready in case one virtual machine migrates from one VMware ESX host to a different one.
The ability to look at the interface counters for the virtual network adapter connecting the virtual machine to the VEM. These counters are the same ones available on any Cisco switch for physical interfaces, and they are statefully maintained in case a virtual machine moves from an ESX host to a different one.

The ability to use an Encapsulated Remote Switch Port Analyzer (ERSPAN) with a virtual Ethernet interface, even if the virtual machine migrates.

The ability to apply security policies, such as access lists, private VLANs, and port security, on virtual Ethernet interfaces.

The Cisco Nexus 1000V Switch performs all these operations and Layer 2 forwarding without introducing any Layer 2 loops. It does so without the need to run the Spanning Tree Protocol on the server.

Chapter 8 provides more details on how to design with the Cisco Nexus 1000V Series Switches on virtualized servers.

**Software Revisions**

The testing that supports this design guide was run on a topology consisting of:

- An aggregation layer of the Cisco Nexus 7000 Series Switches running Cisco NX-OS Software Release 4.1(5)
- Virtualized servers with Cisco Nexus 1000V Series Switches running Cisco NX-OS Software Release 4.0(4)SV1(1)

These code revisions do not constitute a code recommendation. New code versions add valuable enhancements and new features.

**Core, Aggregation, and Access Layers**

Data centers based on Cisco Nexus products follow the well-proven Cisco architecture of core, aggregation, and access layers. The details of the functions provided by each layer can be found in the Cisco white paper entitled Data Center Design—IP Network Infrastructure ([http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html](http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html)) In addition, design guidelines for the aggregation layer based on vPC can be found in Chapter 5, which guidelines for the access layer can be found in Chapter 6.

As a quick-summary reference, the aggregation and access layers provide the following functions:

- The aggregation layer of the data center provides a consolidation point where access layer switches are connected. This, in turn, provides connectivity between servers for multilayer applications, as well as connectivity across the core of the network to clients residing within the campus, the WAN, or the Internet. The aggregation layer typically provides the boundary between Layer 3 routed links and Layer 2 Ethernet broadcast domains in the data center.

- The access layer of the network provides connectivity for server farm end nodes residing in the data center. Traditional data center access layer designs are strongly influenced by the need to locate switches in a way that most conveniently provides cabling connectivity. The most commonly used designs for data center server farm connectivity are end-of-row and top-of-rack

With the widespread adoption of blade servers, the definition of access layer is blurred, as the blade enclosure may often include a switching component whose main role is either to provide blade-to-blade switching support or simply to funnel the traffic upstream to the “real” access layer.
Often a blade switch is preferred to a pass-through module in order to achieve more cable reduction toward the access layer. In a classic aggregation and access design, this need not be the case, as the access layer can provide for line-rate 10 Gigabit Ethernet connectivity for both LAN and SAN for all possible blade server configurations. Given this fact, it is not uncommon to have an access layer that can be used as a “Layer 2 aggregation” mechanism for additional switching that may occur within a blade server.

These design choices are often intertwined with the cabling choice and with the definition of the pod.

**Data Center Pods**

Each access layer in the data center is “divided” in pods (building blocks), where servers in each pod share similar characteristics, whether these characteristics are similar hardware, similar SLAs provided to the customer (internal or external), and so on. This is different from building separate silos for each business unit. In a pod design, multiple business units may share the same pod, and run on similar server platforms because of the similar level of service required.

A data center pod is constituted by a certain number of racks of servers that are all dependent on the same set of access switch hardware. Switch cabinets may either include top-of-rack switch with a fabric extender or a modular middle-of-the-row switch.

A pod can benefit from VLAN isolation to virtualize the LAN infrastructure, the ability to “rack-and-roll” new racks by using the fabric extender technology whereby a rack can be precabled and then attached to a Cisco Nexus 5000 Series Switch from which the fabric extender gets the code and the configuration.

With fabric extenders, it is not uncommon for customers to build 10 to 12 rack-and-roll pods.

**Cabling Considerations**

At the time of this writing, 10 Gigabit Ethernet connectivity to the server can be provided with the following options:

- **SFP+ copper over twinax CX1 cable**: This is the most cost-effective option and the most currently choice, available with the option of 1, 3, or 5 meters passive cables, although additional lengths may become available with active cables. SFP+ twinax, preterminated copper cables also provide the most power-efficient option to connect servers at 10 Gigabit Ethernet today.
- **X2 connectors with CX4 cables**: provide a form factor that is less suitable for servers (due to the space taken up by X2), and they also consume more power.
- **SFP+ short-reach (SR)** provides optical connectivity that can span longer distances (33m on OM1, 82m on OM2, and 300m on OM3 fiber), but it is less cost-effective than copper.

In this light of these issues, pods of servers connected using 10 Gigabit Ethernet tend to use twinax SFP+ cabling, shown in Figure 11.
Connectivity between Cisco Nexus 2000 Fabric Extenders and Cisco Nexus 5000 Series Switches runs on any of the SFP+ options, but to sustain longer distances between Cisco Nexus 5000 and 2000 Series, currently SFP+ SR tends to be used.

Cisco Nexus 5000 to Cisco Nexus 7000 Series connectivity can be implemented with either SFP+ SR or long reach (LR).

Cisco Nexus 5000 Series to Cisco Catalyst Switch connectivity can be implemented with SR or LR optics, to an equivalent X2 optic on the Cisco Catalyst Switch.

For More Information

For more information about cabling and transceivers, visit: