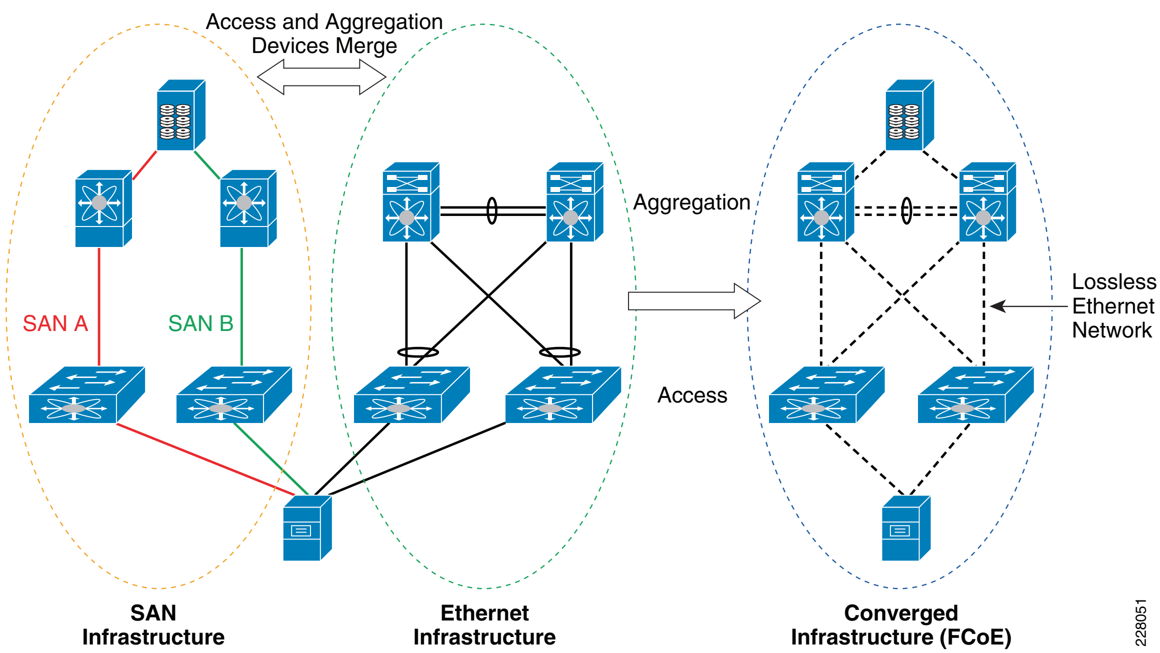
Unified Fabric White Paper—Fibre Channel over Ethernet (FCoE)

Contents

Unified fabric is the convergence of all the various data center I/O technologies over Ethernet (see Figure 1). One of the most disruptive technologies introduced in the Cisco Nexus family of switches in this area is the ability to transport SAN traffic with Fibre Channel over Ethernet (FCoE). The benefits of FCoE are easy to understand. In a typical data center, the backend servers are connected to a separate network infrastructure dedicated to storage traffic. Merging the devices that handle Fibre Channel (FC) and Ethernet traffic provides savings in term of equipment purchase as well as space, power, and cooling costs.

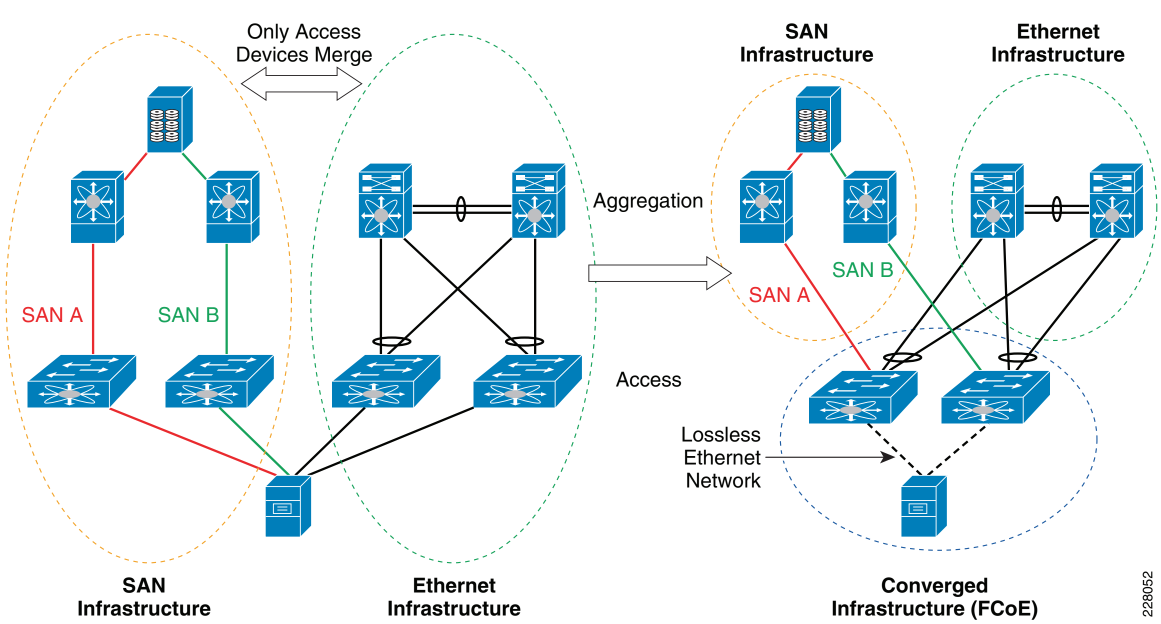
Americas Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Figure 1 Aggregation and Access FC Switches Merged with Ethernet Counterparts



An FC network is lossless, meaning that the protocol has a built-in mechanism that prevents frame drops caused by congestion. FC can be transported only over a lossless Ethernet network. The technology required to implement such a multi-hop lossless Ethernet network is not trivial or finalized. However, conservative SAN administrators should not disregard FCoE because of this apparent bleeding-edge complexity. You can enjoy the benefits of FCoE with simple and proven topologies where the span of lossless Ethernet network is reduced to a single point-to-point link. Figure 2 shows an example of FCoE pushed to the edge of the network.

Figure 2 FCoE Benefits Provided when Deployed at the Network Edge



In this scenario, FCoE is used only between the hosts and the access switches. Most of the savings in operating expenses (OPEX) and capital expenses (CAPEX) promised by FCoE are already available because the edge of the network features the highest density of FC switches, host bus adapters (HBAs), and dedicated cabling. Focusing uniquely on this very simple and efficient design, this document does the following:

- Provides an overview of the standards developed for FCoE
- Proposes a validated design, comparing the configuration and maintenance of the FCoE solution and the traditional separate FC fabric

Solution Components

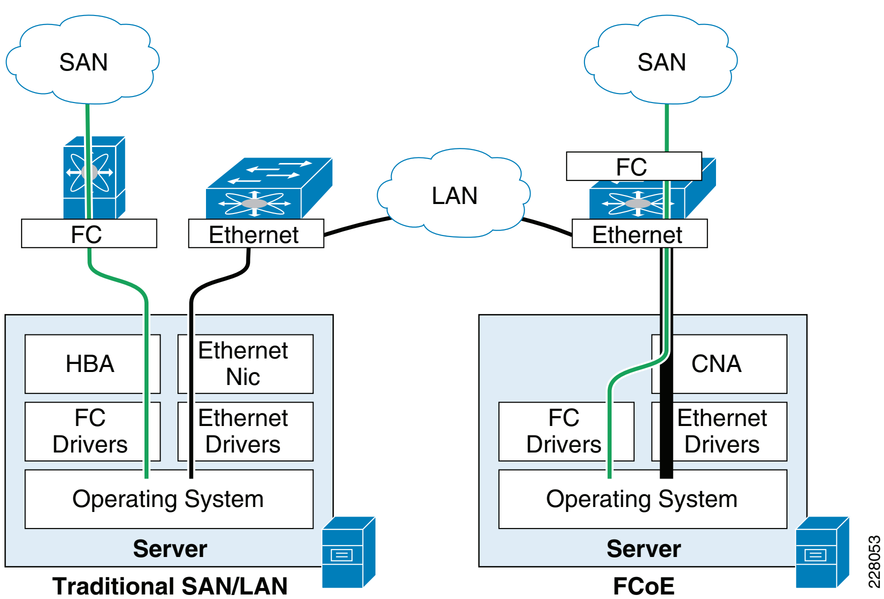
This section introduces the various elements necessary for implementing FCoE at the edge of the network.

Note This section is not intended to be a reference on FCoE. For additional details, see *I/O Consolidation in the Data Center*, Silvano Gai and Claudio De Santi, Cisco Press, ISBN: 158705888X.

Connection Host/Switch

A server connects to an FC fabric via an HBA. These servers also feature another set of network adapters to connect to an Ethernet network. For each server, separate cards with separate cables connect to separate switches, as shown on the left side of Figure 3.

Figure 3 Host-Switch Connection



The solution suggested in this guide (shown on the right side of Figure 3) consists in replacing the HBA and the Ethernet card with a converged network adapter (CNA). This unique CNA connects via a single 10 Gigabit link to an Ethernet switch that is also an FC forwarder (FCF); that is, an Ethernet switch with FCoE capabilities.

For each server, at least two HBAs, two cables (typically with Small Form-Factor Pluggable [SFP] and a transceiver), and two FC ports on an access FC switch are saved with the converged solution. This immediate CAPEX reduction is further complemented by OPEX savings because the model requires fewer devices and thus less power and cooling.

Figure 3 also shows that the CNA is controlled by two different drivers for FC and Ethernet. This means that the virtualization of FC traffic over Ethernet is entirely transparent to the operating system of the server because the CNA appears as two distinct cards. The FC traffic, represented by the green line, is transported over Ethernet to a virtual FC interface on the access switch. From there, it is forwarded using native FC interfaces in the particular FCoE at the edge scenario described in this document.

The following section introduces the various standard components involved:

- The FCoE standard specifies the FCF and defines how the traffic is encapsulated in Ethernet frames.
- The FCoE standard also introduces the FCoE Initiation Protocol (FIP), which is run between the CNAs to discover the FCF automatically.

To support FCoE, the underlying Ethernet must be lossless. The FCoE specification does not address how this must be achieved, so this design relies on the following IEEE standards:

- On the point-to-point connection between the CNA and the Ethernet switch, priority-based flow control (PFC), as defined in 802.1Qbb, provides a mechanism similar to the buffer-to-buffer credit traditionally used by FC.
- The implementation of the above PFC requires coordination between the switch and the CNA. IEEE 802.1Qaz, a standard based on Link Layer Discovery Protocol (LLDP), is used to discover and set up the various Ethernet capabilities required.

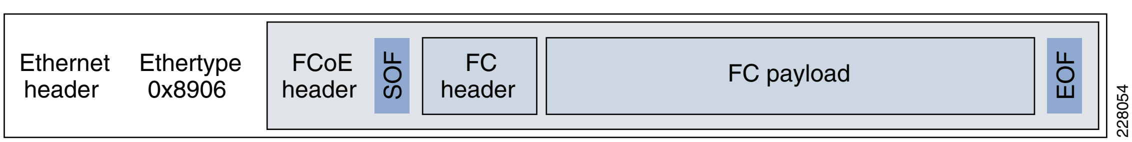
T11 Standard

This section discusses FCoE encapsulation as defined by the T11 standard, as well as FC layers and the FCoE Initialization Protocol (FIP).

FCoE Encapsulation

FCoE is defined in the same T11 FC-BB-5 that also defines Fibre Channel over IP (FCIP), another standard that is now well accepted in the SAN community. Figure 4 shows the format of an FCoE frame.

Figure 4 FCoE Encapsulation



An entire unmodified FC frame is included in a single Ethernet frame; there is no fragmentation. As a result, jumbo frames are required on links supporting FCoE because the FC payload can go up to 2180 bytes. Fields such as the start-of-frame (SOF) and end-of-frame (EOF) use symbols specific to FC lower layers. They are re-encoded following the existing model used by FCIP and carried in the FCoE header enclosing the FC frame.

From the perspective of Ethernet, FCoE is just another upper layer protocol, such as IP, identified by a specific EtherType (0x8906). From the perspective of FC, FCoE is just a different way of implementing the lower layers of the FC stack.

Fibre Channel Layers

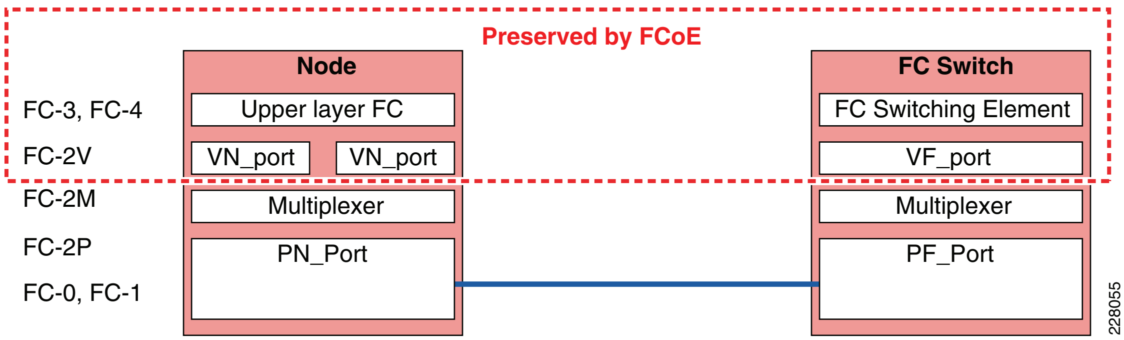
To highlight the changes introduced by FCoE, it is necessary to provide a high-level overview of the various layers in the FC protocol stack.

The bottom layers (FC-0 and FC-1) are responsible for the encoding and the error control on a particular medium.

FC-2P is responsible for the frame format and the control functions necessary for information transfer. In particular, this layer implements the buffer-to-buffer credit mechanism that makes FC lossless. Up to this level, there is only point-to-point communication between a single physical N (PN_Port) and a single physical F port (PF_Port).

Figure 5 shows the FC stack at a high level on a node (a host with an HBA) with an N port on the left connecting to the F port of a FC switch on the right.

Figure 5 FC Stack

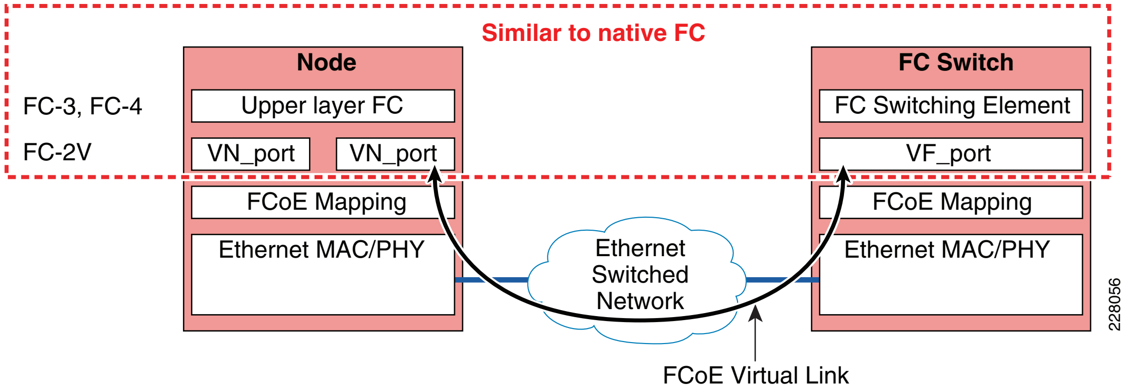


When a node connects to the fabric, it performs a fabric login during which it is assigned a 3-byte FC address, or FCID. This FCID is used for routing frames across the fabric. The entity that received the address is a virtual N port (VN_Port), and the address is provided by the virtual F port (VF_Port) to which it connects. Because it is useful to allow a node to initiate several connections and receive several addresses, there might be more than one VN_Port. The multiplexing of the traffic to and from those different VN_Ports is handled by the FC-2M layer.

FCoE preserves the FC layers FC-2V and above. Because the lower layers that were replaced performed link only specific operations, there is no significant difference between FC and FCoE from the perspective of configuration and maintenance.

Figure 6 shows the FCoE stack in the case of the connection of an E-Node (an FC node using FCoE) to an FCF; that is, an FCoE switch.

Figure 6 FCoE Stack



The VN_Port logs into the fabric and is provided an FCID by the VF_Port, just like in the native FC case. However, establishing the connection and ensuring the multiplexing are achieved quite differently. The physical point-to-point connection between a PN_Port and a PF_Port has been replaced by a multi-access Ethernet switched network. The VF_Port is assigned a unique MAC address, allowing the VN_Port to contact it to log into the fabric. During the login process, the VN_Port not only receives an FCID from the VF_Port, but also a specific MAC address derived from this FCID. The triplet consisting of MAC address of the VN_Port and VF_Port as well as the VLAN on which the exchange is taking place defines an FCoE virtual link over the Ethernet network.

FCoE Initialization Protocol

The data plane of FCoE described in the above section is relatively straightforward. However, the setup of the FCoE virtual links between the VN_Ports and VF_Ports would require significant configuration if performed manually by the administrator. To maintain the plug-and-play aspect of FC, the FCoE standard also includes an FCoE Initialization Protocol (FIP). FIP is an Ethernet protocol, with a dedicated EtherType different from FCoE, which performs the following duties:

- Identifying the FCoE VLAN
- Discovering the FCFs to which the E-Node can connect
- Establishing the virtual links between VN_Ports and VF_Ports
- Monitoring these virtual links

Identifying the FCoE VLAN

Cisco introduced the concept of a virtual storage area network (VSAN), which is to FC what a VLAN is to Ethernet. VSANs allow the creation of independent SANs from the same physical SAN infrastructure. Cisco recommends dedicating a different VLAN for the FCoE traffic of each VSAN. This way, FCoE traffic for the VSAN can be easily identified, and the administrator can control precisely the span of this traffic over the Ethernet network. The FCoE VLAN should furthermore be dedicated to FCoE traffic; for example, it should not carry IP traffic.

FIP performs FCoE VLAN discovery by sending untagged frames to its neighbor. Note that the neighbor does not have to implement a full FCoE stack; FIP is a control protocol that can be run by an Ethernet-only switch. However, in the simple case of FCoE at the access covered in this document, the E-Node directly reaches the FCF for the FCoE VLAN information. Without FIP negotiation, VLAN 1002 is considered the default for FCoE traffic.

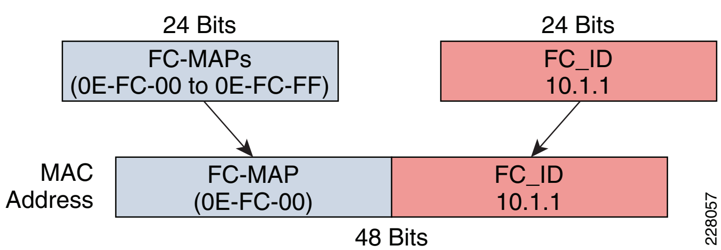
Discovering the FCFs to which the E-Node Can Connect

After FIP has discovered an FCoE VLAN, it helps to identify an available FCF. In the most generic case, this is achieved by sending multicast FIP messages to which FCFs respond with unicast. In the specific scenario described in this document, the E-Node directly reaches the unique FCF to which it is attached.

Establishing the Virtual Link between VN_Ports and VF_Ports

In addition to performing the VLAN and FCF discoveries (operations that are specific to FCoE), FIP is also responsible for performing the fabric login of the VN_Port. Although this can be performed by FC transparently over FCoE, it is done by FIP because at this stage, the virtual link has not been established yet. As previously mentioned, the VN_Port MAC address (required to set up the virtual link) is built from the FCID, and the FCID is available only after login. FIP is thus solving this chicken-and-egg problem by handling the whole fabric login process. The VN_Port is assigned a fabric-provided Mac address (FPMA) that is built by concatenating a 24-bit FCoE MAC address prefix (FC-MAP), ranging from 0x0E-FC-00 to 0x0E-FC-FF, to the 24-bit FCID, as shown in Figure 7. Being able to build a unique MAC address for the VN_Port directly from its FCID saves the switch from having to maintain a table that associates FCID and MAC addresses.

Figure 7 Fabric-Provided Mac Address



The FC-MAP range was introduced so that different values can be assigned to different SANs. For example, SAN A would be associated to 0x0EFC00 and SAN B to 0x0EFC01. This additional configuration ensures the uniqueness of the produced FPMA in the whole network. FC-MAPs are different for different SANs, FCIDs are uniquely assigned within a SAN, and the resulting FC-MAP and FCID are unique across the different SANs in the entire network.

To minimize the risk of crosstalk between SANs, Cisco recommends that each SAN uses a different VLAN for its FCoE traffic. This makes the configuration of the FC-MAP less critical, because the FPMAs have to be unique only on a per-VLAN basis.

Monitoring the Virtual Link

In native FC, both ends of a physical links are implementing an FC stack. If the link fails, an indication is provided to FC. When virtual links are created over a multi-hop Ethernet network, this function is gone, and it is now possible that the connectivity between a VN_Port and a VF_Port gets lost while the direct-attached physical link remains up. FIP allows for detecting these kinds of failures by providing a basic keepalive mechanism. However, in the case of FCoE to the access, where FCoE is only running over a single point-to-point link, this capability has little use.

FCoE Standard Summary

FCoE is a finalized standard, adopted by all the players in the SAN market. It is a solution with no gateway: just like FCIP, it encapsulates FC traffic over a different medium while maintaining all the characteristics of native FC. The additional setup complexity that results from using a potentially multi-hop Ethernet network to create virtual links between VN_Ports and VF_Ports is handled by FIP, which is a new standard control protocol. However, the FCF is an access switch, connecting directly to E-Node, so most of duties of FIP are straightforward and the FCoE operation is extremely similar to plain FC.

FCoE assumes a lossless Ethernet network. This fundamental function is provided with the help of the IEEE standards introduced in the next section.

Relevant IEEE Standards

This section describes the IEEE standards relevant to providing a lossless Ethernet network.

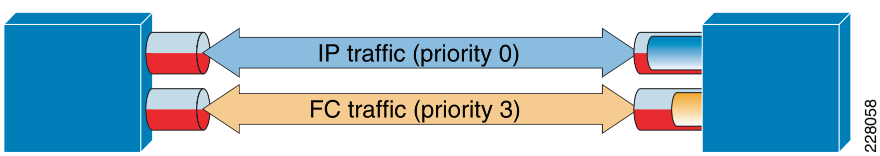
Enhanced Transmission Selection (IEEE 802.1Qaz)

FCoE is not about replacing native FC with Ethernet, but rather about transmitting both FC and other traffic such as IP over the single Ethernet medium. This means that the converged network must be able to simulate independent links for different classes of traffic. For example, you might want to reserve a certain bandwidth for FC traffic so that a burst of IP packets cannot cause a loss of FC frames.

Enhanced Transmission Selection (ETS, IEEE 802.1Qaz) formalizes basic functionality that most switching vendors have been providing for years. ETS defines priority groups to which various frames can be assigned based on their priority and how to allocate bandwidth to those priority groups as a percentage of the maximum bandwidth available on the link. Practically, this means that the port hardware must be able to support several queues for several class of traffic, identified by 802.1p priorities.

Typically, the FC and regular IP traffic is assigned to different priority groups, as shown in Figure 8. A bridge supporting ETS must support between two and eight groups, with at least one of them having no bandwidth limit.

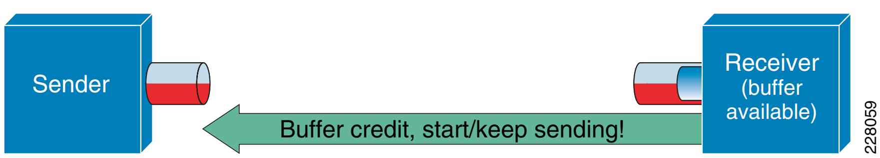
Figure 8 Different Priority Groups



Per-priority Flow Control (802.1Qbb)

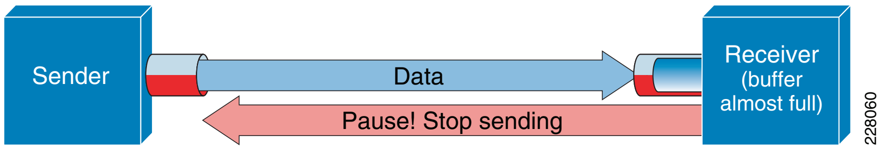
FC creates a lossless fabric using buffer-to-buffer credit (see Figure 9). On a point-to-point FC link, a port can transmit data only if the remote peer has explicitly acknowledged with a buffer credit that it has the buffer capacity to receive this data.

Figure 9 Buffer Credit Mechanism



Ethernet uses the opposite logic. The receiving port can issue a pause frame to stop the transmission from the remote peer when its buffer is about to be exhausted (see Figure 10). Note that the pause must be generated soon enough so that the buffer is not exhausted while the feedback mechanism is shutting down the remote port transmission.

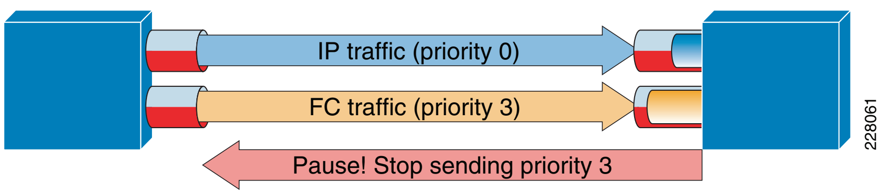
Figure 10 Pause Mechanism



The pause mechanism, as defined more than ten years ago in 802.3x, provides no granularity; when a pause frame is received, the sender stops all transmission on the port. This is not adapted to the goals of unified fabric, as mentioned in the previous section. For example, if FC is exhausting its bandwidth cap and needs to be paused, the IP traffic has no reason to be impacted. The pause mechanism was thus enhanced in a very simple way to support per-priority flow control (PFC, IEEE 802.1Qbb). An additional field within the pause frame specifies which priorities (as defined with 802.1p) are to be paused.

This mechanism allows creating both drop and no drop classes of traffic on the link. For example, because FC requires a lossless Ethernet, it is mapped to a priority group making use of the pause mechanism to stop data transmission before a drop can occur. Figure 11 shows a pause frame that applies only to the priority group to which FC is mapped.

Figure 11 Per-Priority Flow Control



IP traffic belonging to a different priority group is not affected. Note that IP does not require a lossless Ethernet; in case of congestion, the overflowing frames are simply dropped without generating a pause.

Note 802.1Qbb has not been ratified yet. However, there is already a consensus on the format of the pause format. No significant change is expected at that stage.

DCBX (IEEE802.1Qaz)

To minimize the configuration overhead, IEEE 802.1Qaz also defines the Data Center Bridging Configuration Exchange (DCBX) protocol, which can do the following:

- Discover the capability of a peer and detect peer misconfiguration. For example, devices may support a different number of priority groups, so discrepancies need to be detected.
- Configure a peer; a parameter can be set locally so that it is overridden by the peer if necessary. A server can be configured this way by the network when connected.

DCBX is simply implemented on top of the Link Layer Discovery Protocol (IEEE 802.1AB) by using some additional type-length-values (TLVs). DCBX started as a proprietary protocol known as Cisco-Intel-Nuova DCBX. It has currently evolved and is more widely adopted under the name Converged Enhanced Ethernet DCBX (CEE-DCBX). Among other things, CEE-DCBX handles the coordination of the following features:

- Priority flow control (PFC)
- Bandwidth management (ETS)
- Logical link down, a way of bringing down the FCoE interface in a logical way without disrupting other Ethernet traffic

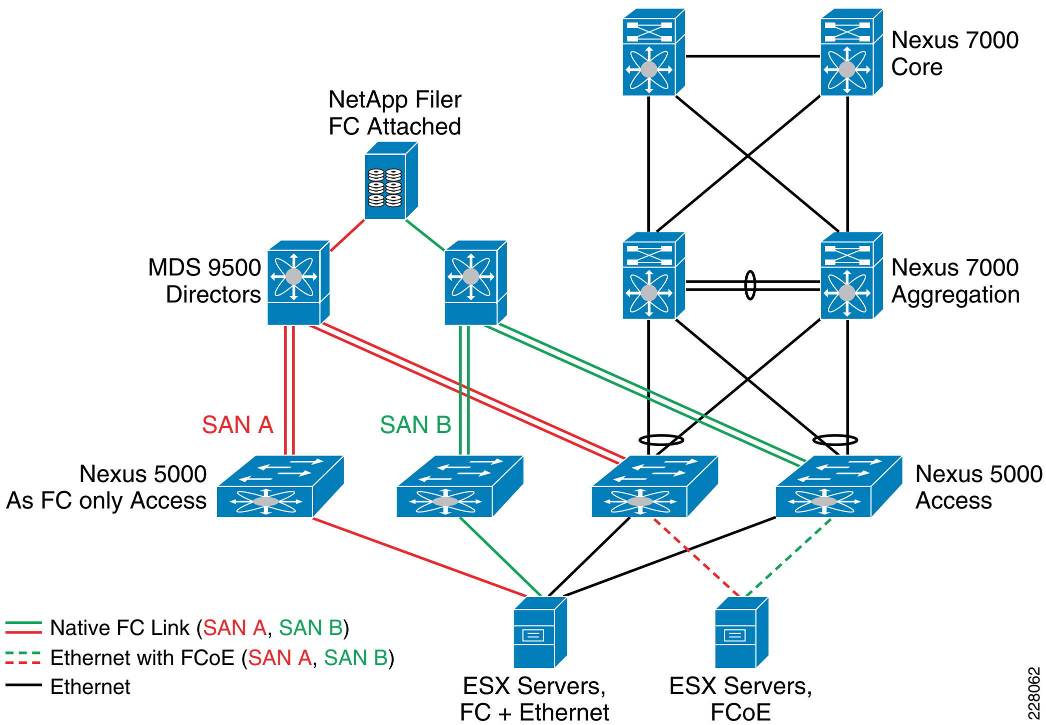
These functionalities are expected to be adopted in the IEEE standard. Because DCBX is a control protocol, any significant deviation from CEE-DCBX will be resolved by a software update.

Data Center Design with FCoE at the Access Layer

Introducing FCoE at the edge of the network does not require any change in the configuration of the FC director or the aggregation layer of the data center, so this document does not repeat the design recommendations above the access layer.

Current SAN design is based on the use of two independent fabrics (SAN A/SAN B) to which initiators and targets are dual-attached. This form of redundancy is possible only because the FC nodes have a lot of intelligence built-in to their adapter and are tightly interacting with the services provided by the fabric. The fabric informs a particular node of the reachability of others. This allows the node to balance loads across fabrics or failover from one fabric to the other. Meeting the requirement for two strictly independent fabrics is not straightforward in an end-to-end converged approach, as shown in the right side of Figure 1. However, when keeping FCoE at the edge, this function is still preserved. Figure 12 shows a logical view of the test network used for this document. The Cisco Nexus 5000s used at the access layer are either part of SAN A (in red), or SAN B (in green).

Figure 12 Test Lab—Logical View



The Cisco Nexus 5000s in Figure 12 are also used in the following ways:

- As a regular FC access switch. In this case, the hosts are equipped with an HBA and connected to the SAN via a 4 Gbps FC link, and to the Ethernet side of the network via Gigabit Ethernet.
- As an FCF: the hosts are then equipped with a CNA, providing two 10 Gbps Ethernet links on which both LAN and SAN traffic are forwarded.

This hybrid setup was not chosen as a design recommendation, but rather as a way of showing the following:

- The operation and configuration of FCoE versus the traditional model does not introduce any significant change.
- The two solutions interoperate seamlessly.

The test network furthermore included a mix of Nexus 5000s used in NPV mode (N-Port Virtualizer, introduced below) or switch mode, as well as the host connecting via first generation CNAs and second generation CNAs.

FC Connection Access/Core

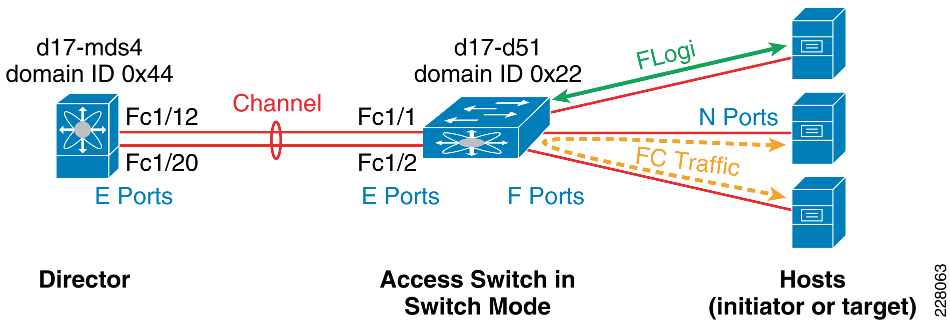
This section details the configuration steps used to connect the access FC switch to the core directors. Whether the access switch is a Cisco Nexus 5000 or a conventional FC switch does not matter here. Cisco NX-OS Software provides exactly the same interface for configuration and maintenance in both cases, because native FC uplinks are used in the scenario described in this document.

The access FC switch can be used in either switch or NPV mode. These two options imply different configuration on the director and access switch, as detailed in the following sections.

FC Access Switch in Switch Mode

In switch mode, as shown in Figure 13, the ports between the director and access switches are expansion ports (E ports).

Figure 13 Access Switch in Switch Mode



The attached FC hosts log onto the fabric by contacting the F port on the access switch and get an FCID directly from this F port.

Traffic between N ports is directly switched by the access FC switch. Because of this function, this mode is especially recommended when there is a mix of initiators and targets at the access.

It is furthermore best practice to configure the redundant connections between the access and the director as a channel. Channeling provides granular load balancing between the uplinks as well as minimal disruption to the fabric if one link fails.

The main drawback of the switch mode is that the access switch is assigned a domain ID. There is a maximum of 239 domain IDs in a fabric. Practically, supported implementations rarely go over 50 domain IDs, which makes the use of switch mode difficult at the access for large deployments.

The configuration for the access switch is as follows. It is similar to the port configuration on the director side.

```
vsan database
  vsan 10 interface san-port-channel 1
interface san-port-channel 1
  switchport mode E
  switchport trunk off

interface fc2/1
  switchport mode E
  channel-group 1
  no shutdown
interface fc2/2
  switchport mode E
  channel-group 1
  no shutdown
```

SAN switches use the keyword **port-channel**. The Nexus 5000 has a special **san-port-channel** keyword to differentiate with Ethernet port channels.

Even if the switch can negotiate the port mode, it is best to hardcode it so that improper cabling will most likely result in a link staying down.

Trunking is currently possible only for E ports. In the above example, trunking is not necessary but is possible if several VSANs are required at the access. NPV mode cannot currently achieve this.

The following console output shows the content of the fabric login (FLOGI) database on switch *d17-n51*. This switch is connected to the director *d17-mds4*, as shown in Figure 13. The FLOGI database shows the list of FC nodes that have performed a fabric login to the switch. Those nodes get an FCID assigned by d17-n51, and as a result, the first byte of the FCID is the domain ID of d17-n51, which is 0x22. Notice also that hosts connected via FCoE are treated exactly the same way as those directly attached with native FC. The only apparent difference is the type of interface. FCoE-attached nodes are connected to a virtual Fibre Channel (VFC) interface (detailed in FCoE Connection Access/Host, page 15), while the native nodes are connected with an FC interface.

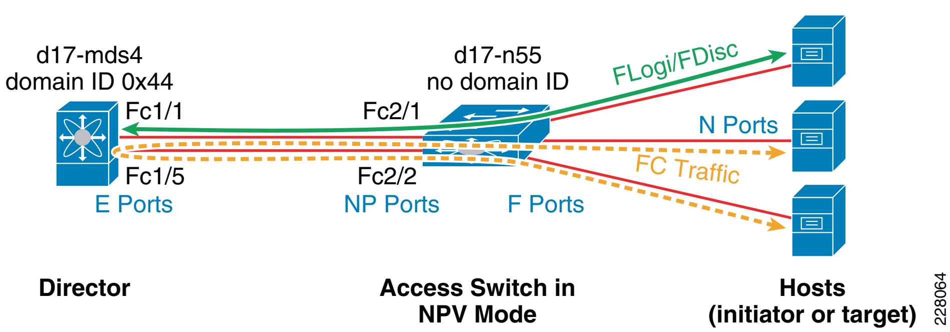
```
d17-n51# sh flogi database
--------------------------------------------------------------------------------
INTERFACE        VSAN    FCID           PORT NAME               NODE NAME
--------------------------------------------------------------------------------
fc2/3            10     0x220000  10:00:00:00:c9:80:05:03 20:00:00:00:c9:80:05:03
                            [mc5-hba-a]
vfc5             10     0x220002  10:00:00:00:c9:76:ec:44 20:00:00:00:c9:76:ec:44
                            [mc5-cna-a]
vfc6             10     0x220001  10:00:00:00:c9:76:ec:45 20:00:00:00:c9:76:ec:45
                            [mc6-cna-a]
vfc7             10     0x220003  21:00:00:c0:dd:11:08:61 20:00:00:c0:dd:11:08:61
                            [mc7-cna-a]
Total number of flogi = 4
```

FC Access Switch in NPV Mode

N-Port Virtualizer (NPV) mode is a Cisco feature that takes advantage of N-Port ID virtualization (NPIV) defined in the FC standard. As previously discussed, the FC stack provides a way of multiplexing the connection of several VN_Ports to an VF_Port over a single FC link. This is a capability provided by NPIV. The initial goal was to allow independent processes on a server to get their own VN_Port; each VN_Port gets an independent FCID and can then be serviced differently by the fabric. In NPV mode, a Cisco device connects to the fabric as if it was a host with several such VN_Ports.

Figure 14 shows the access switch in NPV mode.

Figure 14 Access Switch in NPV Mode



When the first host performs a fabric login to the F port of the access switch, this FLOGI is converted to a FLOGI to the F port of the director. The director assigns an FCID that is then returned by the access switch to the host. Further host FLOGIs are converted into fabric discoveries (FDISCs, defined by NPIV as a way of logging into additional VN_Ports) and sent to the F port of the director.

In NPV mode, the access switch is connected as a host to the F port of the director. As a result, it does not require a domain ID. This function allows for better scalability, because the available domain IDs are limited in a fabric, as mentioned previously. The NPV mode is thus preferred in large deployments. Interaction between switches from different vendors is also often problematic. Because NPIV is an FC standard designed to interact with hosts, it is also provides better compatibility in the case of a mixed-vendor SAN.

The link between the director and the access switch currently does not support channeling or trunking. If those features are critical, switch mode is required.

To perform load balancing, the access switch in NPV mode is distributing the host connections to the director among the available uplinks. If an uplink fails, all the hosts that were associated with it are disconnected and need to login again. Another drawback of NPV mode is that there is no local switching on the access switch. If two hosts need to connect to each other, they do so through the director.

The configuration of the director and the access switch is now asymmetric in the case of NPV. The ports on the NPIV-capable director are F ports, while the access switch uplinks are NP ports.

The following is the configuration on the director side:

```
feature npiv
vsan database
  vsan 10 interface fc1/1
  vsan 10 interface fc1/5
interface fc1/1
  switchport mode F
  no shutdown
interface fc1/5
  switchport mode F
  no shutdown
```

The director must support the NPIV feature. In the configuration example above, the two F ports of the director are simply assigned to the VSAN 10, representing SAN A in the test network. The configuration on the side of the access switch in NPV mode is as follows. Uplinks are configured as NP ports and also simply assigned to VSAN 10.

The following is the configuration on the access switch in NPV mode:

```
feature npv
npv enable      <----  The configuration of NPV mode is disruptive;
                       the switch is rebooted and its configuration
vsan database
  vsan 10 interface fc2/1
  vsan 10 interface fc2/2   <----  The uplinks appear as individual ports: no
                                   channel possible.
interface fc2/1
  switchport mode NP
  no shutdown
interface fc2/2
  switchport mode NP
  no shutdown
```

When the access switch is running in NPV mode, it is not an FC switch. As a result, there is no local FLOGI database. In Figure 14, the switch *d17-n55* is connected in NPV mode to the same director *d17-mds4* that was represented in Figure 13. The nodes physically connected to d17-n55 logically appear as directly attached to d17-mds4. The following output from d17-mds4 shows that end nodes have been distributed across the links between d17-mds4 and d17-n55 (*fc1/1* and *fc1/5*) by the NPV feature.

```
d17-mds4# sh flogi database vsan 10
--------------------------------------------------------------------------------
INTERFACE        VSAN    FCID           PORT NAME               NODE NAME
--------------------------------------------------------------------------------
fc1/1            10     0x440001  20:41:00:0d:ec:b2:c4:80 20:0a:00:0d:ec:b2:c4:81
fc1/1            10     0x440002  50:0a:09:87:87:49:34:c6 50:0a:09:80:87:49:34:c6
                            [NetApp-cna-a]
fc1/1            10     0x440004  10:00:00:00:c9:76:ec:30 20:00:00:00:c9:76:ec:30
                            [mc12-cna-a]
fc1/1            10     0x440006  21:00:00:c0:dd:10:e4:59 20:00:00:c0:dd:10:e4:59
                            [mc10-cna-a]
fc1/1            10     0x440100  50:06:0b:00:00:66:0a:8e 50:06:0b:00:00:66:0a:8f
                            [mc12-hba-a]
fc1/5            10     0x440003  10:00:00:00:c9:76:ed:2d 20:00:00:00:c9:76:ed:2d
                        [mc11-cna-a]
fc1/5            10     0x440005  20:42:00:0d:ec:b2:c4:80 20:0a:00:0d:ec:b2:c4:81
fc1/5            10     0x440200  50:06:0b:00:00:65:71:3e 50:06:0b:00:00:65:71:3f
```

```
                           [mc11-hba-a]
fc1/13           10    0x440000  50:0a:09:83:87:49:34:c6 50:0a:09:80:87:49:34:c6
                           [NetApp-hba-a]
Total number of flogi = 9
```

The FC fabric includes a name server to which all the nodes in the network are registered. The following output from d17-mds4 shows all the known devices in VSAN 10. The FC nodes have an FCID that is derived from the domain ID of the switch to which they performed their FLOGI. As a result, the nodes attached to d17-n51, operating in switch mode, got an FCID starting with 0x22, the domain ID of d17-n51. On the other hand, the devices connected to d17-n55 were assigned an FCID starting with 0x44, derived from the domain of d17-mds4.

```
d17-mds4# show fcns database
VSAN 10:
--------------------------------------------------------------------------
FCID        TYPE  PWWN                    (VENDOR)        FC4-TYPE:FEATURE
--------------------------------------------------------------------------
0x220000    N     10:00:00:00:c9:80:05:03 (Emulex)        scsi-fcp:init
                  [mc5-hba-a]
0x220001    N     10:00:00:00:c9:76:ec:45 (Emulex)
                  [mc6-cna-a]
0x220002    N     10:00:00:00:c9:76:ec:44 (Emulex)
                  [mc5-cna-a]
0x220003    N     21:00:00:c0:dd:11:08:61 (Qlogic)        scsi-fcp:init
                  [mc7-cna-a]
0x440000    N     50:0a:09:83:87:49:34:c6 (NetApp)        scsi-fcp:target
                  [NetApp-hba-a]
0x440001    N     20:41:00:0d:ec:b2:c4:80 (Cisco)         npv
0x440002    N     50:0a:09:87:87:49:34:c6 (NetApp)
                  [NetApp-cna-a]
0x440003    N     10:00:00:00:c9:76:ed:2d (Emulex)
                  [mc11-cna-a]
0x440004    N     10:00:00:00:c9:76:ec:30 (Emulex)
                  [mc12-cna-a]
0x440005    N     20:42:00:0d:ec:b2:c4:80 (Cisco)         npv
0x440006    N     21:00:00:c0:dd:10:e4:59 (Qlogic)        scsi-fcp:init
                  [mc10-cna-a]
0x440100    N     50:06:0b:00:00:66:0a:8e (HP)            scsi-fcp:init
                  [mc12-hba-a]
0x440200    N     50:06:0b:00:00:65:71:3e (HP)            scsi-fcp:init
                  [mc11-hba-a]
Total number of entries = 13
```

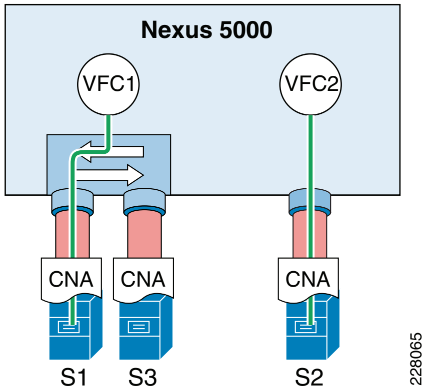
The following summarizes the configuration recommendations for the access/core connection:

- Hardcode the FC port mode in the configuration instead of relying on negotiation.
- Use channels to connect the access switch in switch mode and the director.
- Use different VSAN IDs for the different fabrics.

FCoE Connection Access/Host

In the scenario described in this document, the connection between the access switch and the director is achieved over native FC. Because the Cisco Nexus 5000 provides an interface configuration consistent with the Cisco MDS series of FC switches, there was nothing specific to FCoE in the previous section. This section, however, is dedicated to the setup of an FCoE connection between a host equipped with a CNA and a Nexus 5000. Figure 15 is another version of the right side of Figure 3.

Figure 15 Virtual Fibre Channel Interfaces



Three servers are connected via 10 Gigabit Ethernet to a Nexus 5000. A VLAN dedicated to FCoE is defined on the Nexus 5000 and then automatically advertised by FIP to the E-Nodes. The VN_Ports on the E-Nodes must then connect to the VF_Port of an FCF on this VLAN. The Nexus 5000 allows the creation of VFCs for this purpose. A VFC is to FC what a VLAN interface is to IP: a logical interface for an upper layer protocol. However, as opposed to IP, FC expects its interfaces to be connected point-to-point to a single FC neighbor. This property is not respected because the underlying Ethernet network is multi-access. In Figure 15, VFC1 can receive FC traffic from S1 or S3, because both E-Nodes have access to the FCoE VLAN. To associate a VFC to a particular device, a VFC can be defined in the following two ways:

- A VFC can be bound to the MAC address of a particular E-Node. In that case, only traffic transmitted on the FCoE VLAN and originated from the CNA is forwarded to the VFC. This method is configuration-intensive (because a MAC address must be entered on the Nexus 5000) and could in theory introduce some security concerns. Suppose that in Figure 15, S1 is associated to VFC1 by its MAC address. S3 could impersonate S1 by using the MAC address of S1 and compromise the security of the FC transactions from/to S1. To mitigate this problem, the VLANs reserved for FCoE traffic do not perform MAC address learning or unknown unicast flooding. By interacting with FIP during the fabric login of the FCoE E-Nodes, the Ethernet switches can thus enforce some sort of access lists that prevent MAC address spoofing.
- A VFC can be bound to a physical Ethernet interface. This simple solution provides the same level of security as a native FC interface: any FCoE traffic received on a particular physical Ethernet port (traffic identified by VLAN and EtherType) is forwarded to the VFC. It does not require specifying a MAC address on the switch, making the solution more flexible (the host, or the CNA in the host, can be swapped with no additional configuration).

Even when mapping a VFC to a physical port, it is best practice to minimize the span of the VLAN associated to FCoE traffic; remove it from any port that is not directly attached to a host using FCoE. As previously mentioned, although mechanisms implemented at the FIP level make sure that an E-Node connects only to the desired FCF, limiting the extension of the FCoE VLAN enforces additional security in the data plane. Because of the particular properties of the FCoE VLAN, it should only carry FCoE traffic.

The following configuration snippet shows the definition of an FCoE VLAN and a VFC.

```
feature fcoe
vlan 1010                  <----  VLAN 1010 will carry traffic for VSAN 10 over
  fcoe vsan 10                    Ethernet
vsan database
  vsan 10 name "SAN A (0xA)"
interface vfc5                    VFC5 is associated to physical interface
  bind interface Ethernet1/5 <--  Ethernet1/5. The interface is an F port by
  no shutdown                     default
```

```
vsan database
  vsan 10 interface vfc5     <-----------   VFC5 is put into VSAN 10, like a regular native
                                            FC interface

fcoe fcmap 0xefc0a     <----------------    PC-MAP specific to fabric A
```

The recommended configuration for the physical interface is as follows:

```
interface Ethernet1/5
  switchport mode trunk
  spanning-tree port type edge trunk
```

FCoE traffic must be 802.1Q tagged. This is essential to create a lossless Ethernet by the means of ETS and PFC. The Ethernet interface is thus configured as a trunk port. The native VLAN must not be the FCoE VLAN to make sure that FCoE traffic is tagged.

The port is connected to a host and must be identified as an edge port to the Spanning Tree Protocol (STP). This configuration is mandatory because Rapid per VLAN Spanning Tree (Rapid-PVST) or Multiple Spanning Tree (MST) could attempt to sync (that is, put temporarily in a discarding state) the port during their convergence, which would have catastrophic consequence on the FCoE traffic.

The following summarizes the configuration recommendations for the FCoE configuration section:

- Bind VFCs to physical interfaces (in this scenario), not to MAC addresses.
- Limit the span of the FCoE VLAN to where E-Nodes are directly connected.
- Use different FCoE VLANs for the different fabrics.
- Use only the FCoE VLANs for FCoE traffic.
- Configure an FC-MAP on the switch depending on the fabric to which it is connected.
- Configure Ethernet ports as trunks.
- Make sure that the FCoE VLAN is not the native VLAN on the Ethernet interfaces.
- Configure the Ethernet ports connecting to the E-Nodes as the edge for STP.

Comparing FCoE and Native FC from an Operational Point of View

From the perspective of the network administration, the Cisco Nexus 5000 behaves just like a regular FC switch in terms of connecting it to the core of the network. This similarity is even more obvious when using the Fabric and Device Managers. Figure 16 shows the Device Manager view for the director *d17-mds4* (Cisco MDS 9500), and Figure 17 shows the access switch *d17-n55* in NPV mode (Nexus 5000).

Figure 16 Device Manager View—Cisco MDS 9500

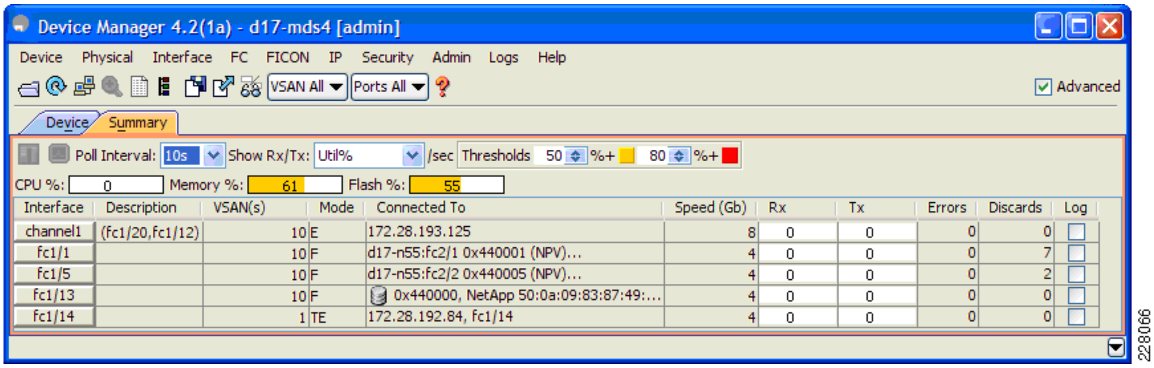
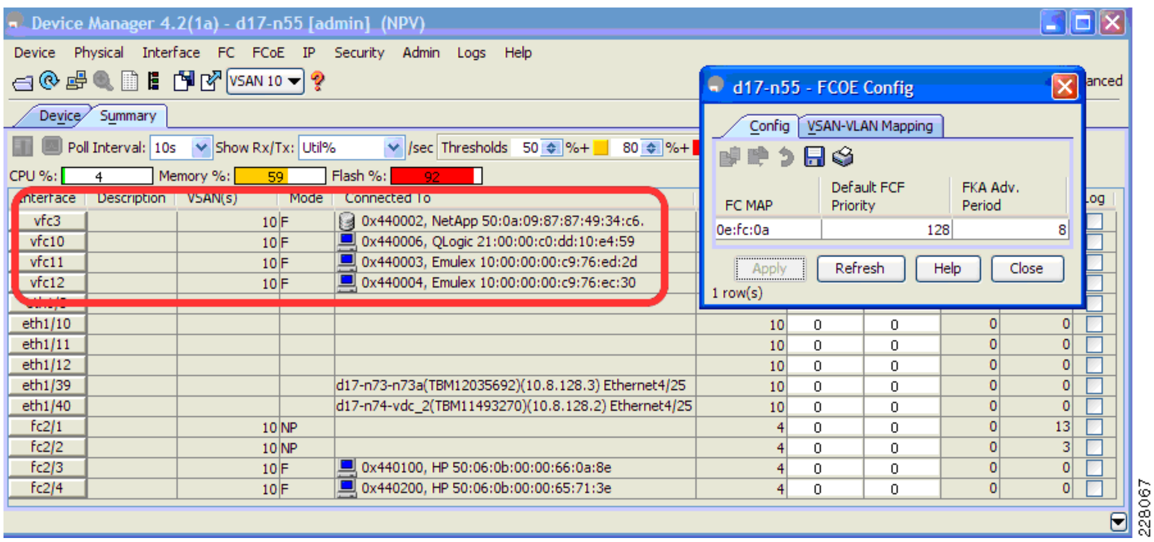


Figure 17 Device Manager View—Cisco Nexus 5000



One of the fundamental differences between the two devices is that the Nexus 5000 includes some VFC interfaces, shown highlighted in a red box. Also, the menu bar provides access to some additional configuration windows that allow the setting of the FCoE-specific parameter. That aside, the Nexus 5000 is just a FC switch.

Similarly, at the level of the hosts, there is very little from an operational point of view between a device using a traditional HBA to connect to the FC fabric and a host relying on a CNA to perform the same operation. Figure 18 shows the storage adapter view from vSphere 4.0 on a server connected via an HBA.

Figure 18 Storage Adapter View—Server Connected via an HBA

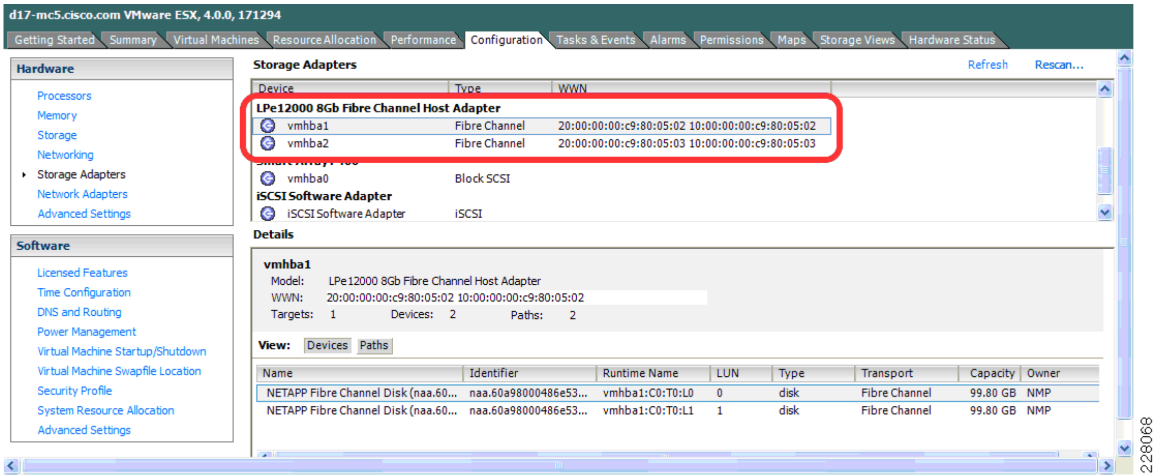
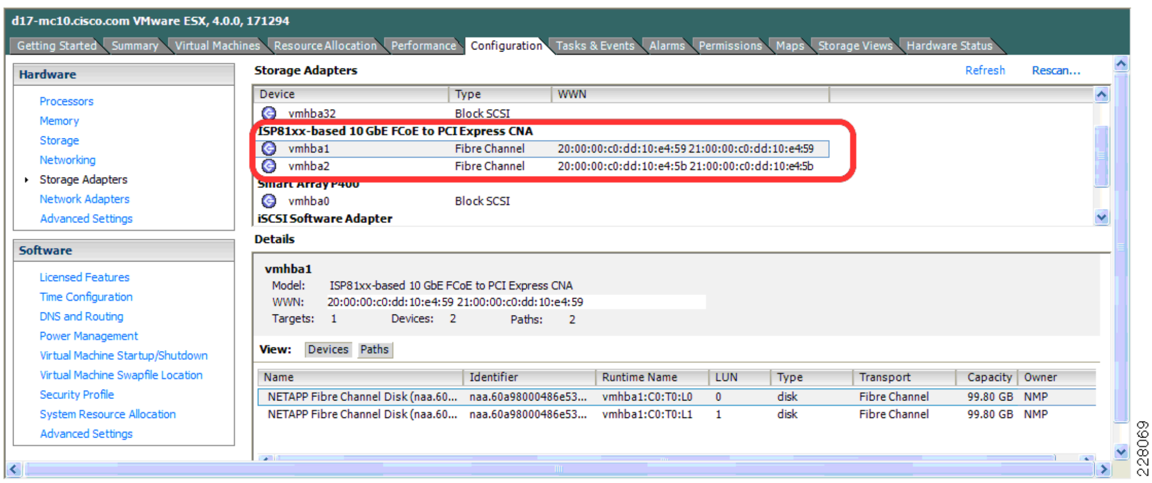


Figure 19 shows the same view for a host attached with a CNA.

Figure 19 Storage Adapter View—Server Connected via a CNA

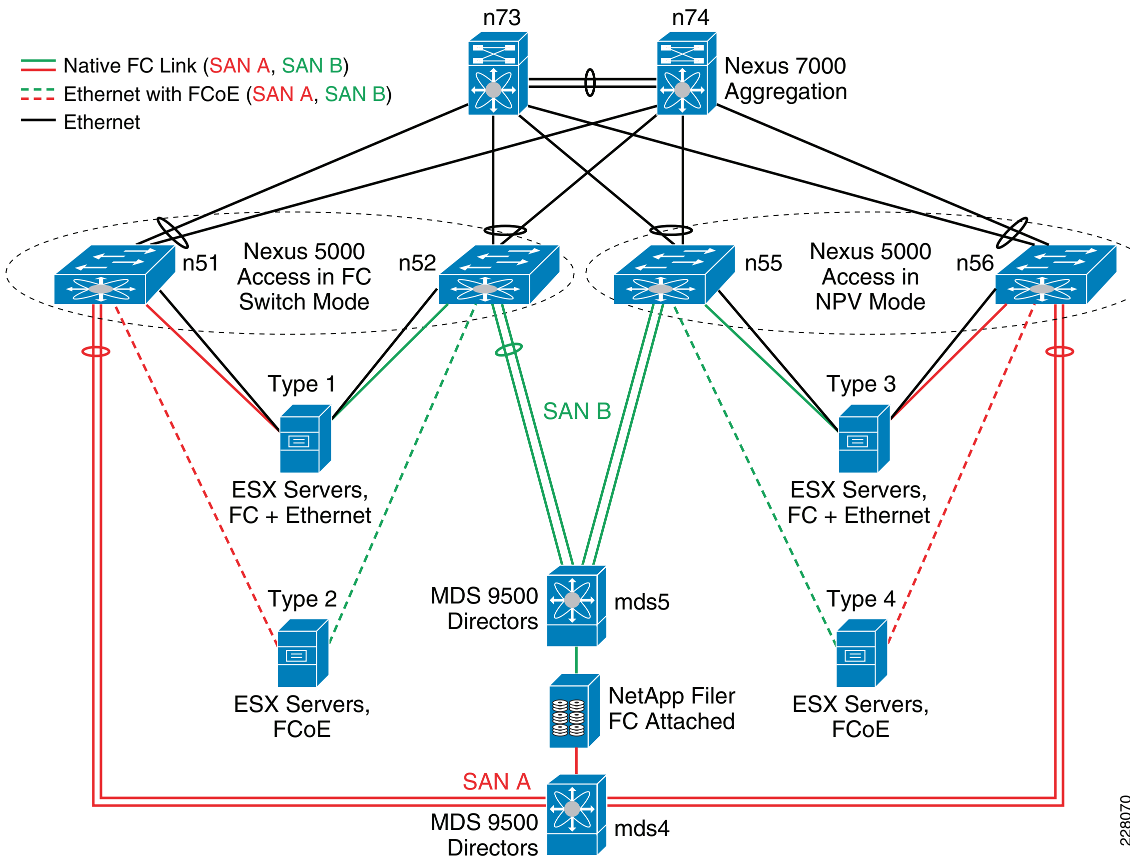


The main difference between the two models is the name of the adapters. From an operational perspective, both servers are configured the same way and see the same target on the NetApp filer available in the network.

Flexibility of the FCoE Model

The topology of the network used for the test is shown in Figure 20. It is very similar to Figure 12, except that hosts connected using the "traditional" method (HBA and Gigabit Ethernet) and the hosts using FCoE are in fact connected to the same pair of Nexus 5000s. A pair of Nexus 5000s (n51 and n52) is operating in switch mode, while the others (n55, n56) are in NPV mode. This combination of server connectivity and access switch mode defines four types of hosts (label type1 to type4 in Figure 20).

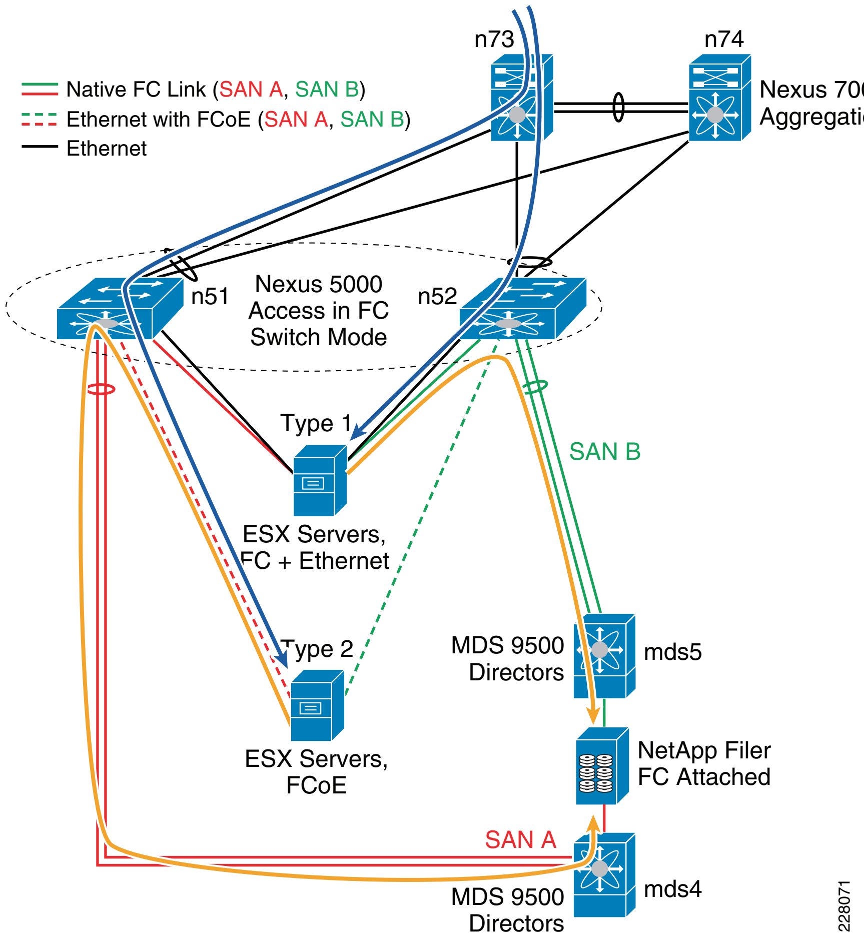
Figure 20 *Physical Lab Cabling*



The hosts are running VMware ESX 4.0 Enterprise Plus. Several virtual machines based on Windows 2003 Server are furthermore running on these hosts. The storage for the virtual machines is a datastore on the Netapp filer connected to the Cisco MDS 9500 (mds4 and mds5 in Figure 20).

To test both the IP and SAN infrastructure, some external servers off the core of the data center (not represented on the diagram) were configured to send traffic to some virtual machines through the aggregation and access switches. This traffic causes the virtual machines to store some data on their datastore, via the SAN. On virtual machines located on a host running FCoE, both IP and SAN traffic is carried over a common 10 Gigabit link, while it is split between a Gigabit Ethernet link and a 4-Gigabit FC link on hosts with "traditional" connectivity, as shown in Figure 21.

Figure 21 External IP Traffic Eventually Generating SAN Traffic



ESX provides a feature, called VMotion, which allows the dynamic migration of a virtual machine from one host to another. The storage and memory associated to a virtual machine is synchronized between the source and destination host; the virtual machine is then frozen at the initial location while it is started at the target host. Because the virtual machines used in this example are using SAN-attached storage, only the memory associated to a virtual machine needs to be synchronized when it moves from one host to the other. One of the constraints to using VMotion is that the target host must have the same characteristics as the source. In terms of networking, this means that if a virtual machine needs two FC HBAs and 4-Gigabit Ethernet NICs, these must be available on the host to where this virtual machine is to be moved. The unified fabric concept makes things very simple here. As soon as a host features a CNA, it can provide both virtual HBA and Ethernet NICs to its virtual machines, introducing maximum flexibility in the data center regarding to where virtual machines can be moved.

Using the sample data center presented in Figure 20, VMotion can be successfully performed between any of the four types of hosts. The goal for this particular test was to prove not only the flexibility introduced by FCoE in the data center, but also to validate the seamless interaction between FCoE and HBA/NIC-attached hosts. FCoE can be implemented incrementally at the edge of the network without causing problems to the existing equipment in place.

Equipment List

This section provides a list of equipment used in the solution described in this document.

Aggregation Switches

The following aggregation switches are used:

- Cisco Nexus 7000, NX-OS 4.2(2a)
- Cisco MDS 9500, supervisor 2, NX-OS 4.2(1a)

Access Switches

- Cisco Nexus 5000, NX-OS 4.1(3)N1(1a)

Servers/Filer

The following servers/filers are used:

- HP ProLiant DL380 G5 servers
- VMware ESX 4.0 Enterprise Plus
- Windows Enterprise Server 2003
- Second generation CNA—QLogic 8152 (FCoE driver 831k1.23vmw, Ethernet 1.0.0.36)
- First generation CNA—Emulex LP2100
- Netapp Filer, FAS6070, 7.3.1D9

Conclusion

Fibre Channel over Ethernet (FCoE) is simply Fibre Channel. The rules for designing and administering an FC network are not fundamentally changed by this latest T11 standard. FCoE requires a lossless Ethernet network and this function is easy to achieve at the edge of the network.

Combining the various data center I/O technologies over Ethernet as a unified fabric eliminates the cost associated with the maintenance of several parallel networks. The savings in term of equipment (switches, cabling, HBAs/NICs) also generates an economy of space, power, and cooling. The unified fabric model also maximizes the flexibility in the allocation of resources, a critical goal in the data center.

In addition, introducing FCoE does not require a complete overhaul of the data center. FCoE at the edge not only provides most of the benefits of unified fabric, but it also interacts without causing problems with the existing FC-attached devices.

References

- *I/O Consolidation in the Data Center*, Silvano Gai & Claudio De Santi, Cisco Press, ISBN: 158705888X
- FIP white paper—http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-560403.html
- Priority Flow Control: Build Reliable Layer 2 Infrastructure—http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-542809.html