Chapter 3:

Cisco NX-OS Software Virtual PortChannel: Fundamental Concepts

Contents

Virtual PortChannel Technology

Virtual PortChannels (vPCs) allow 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 PortChannels.

Cisco NX-OS Software vPCs and Cisco Catalyst® Virtual Switching Systems (VSS) are similar technologies. For Cisco EtherChannel® technology, the term **multichassis EtherChannel** (MCEC) refers to either technology interchangeably.

vPC allows the creation of Layer 2 PortChannels 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 Cisco Nexus 2000 Series Fabric Extenders).

vPC Topologies

Cisco Nexus vPC topologies can be categorized as follows:

- **vPC on the Cisco Nexus 7000 Series (topology A):** This topology comprises access layer switches dual-homed to the Cisco Nexus 7000 Series with a switch PortChannel with Gigabit Ethernet or 10 Gigabit Ethernet links. This topology can also consist of PortChannels with each host connected either with one or more links to each Cisco Nexus 7000 Series.
- **vPC on Cisco Nexus 5000 (topology B):** This topology comprises switches dual-connected to the Cisco Nexus 5000 (switch PortChannels) with 10 Gigabit Ethernet links, either with one or more links to each Cisco Nexus 5000 Series Switch.. Like topology A, topology B can consist of PortChannels where servers connect at 10 Gigabit Ethernet with one or more ports to each Cisco Nexus 5000 Series.
- **vPC on the Cisco Nexus 5000 Series with a Cisco Nexus 2000 Fabric Extender single-homed** (also called **straight-through mode**) **(topology C):** This topology consists of a Cisco Nexus 2000 Series Fabric Extender single-homed with one to four times 10 Gigabit Ethernet links to a single Cisco Nexus 5000 Series Switch, and of Gigabit-Ethernet-connected servers that form PortChannels to the fabric extender devices. It is important to notice that each fabric extender connects to a single Cisco Nexus 5000 Series Switch and not to both, and that the PortChannel can be formed only by connecting each of the two server network interface cards (NICs) to two fabric extenders, where fabric extender 1 depends on Cisco Nexus 5000 Series Switch 1 and fabric extender 2 depends on Cisco Nexus 5000 Series Switch 2. If both fabric extender 1 and fabric extender 2 depend on switch 1 or both of them depend on switch 2, the PortChannel cannot be formed. In the Cisco Nexus 2000 Series family, the Cisco Nexus 2148T has the restriction that only 2-Ports PortChannels are supported.
- **Dual-homing of the Cisco Nexus 2000 Fabric Extender (topology D)** requires special considerations and is described at the end of this chapter. With this topology the server cannot create a PortChannel split between two fabric extenders. The servers can still be dual-homed with active/standby or active/active transmit load balancing (TLB) teaming.

Note: Topologies B, C, and D are not mutually exclusive. You can have an architecture that uses the three of these topologies concurrently.

Figure 1 illustrates topologies A and B. Figure 2 illustrates topologies C and D.

Figure 1. vPC Topologies A and B

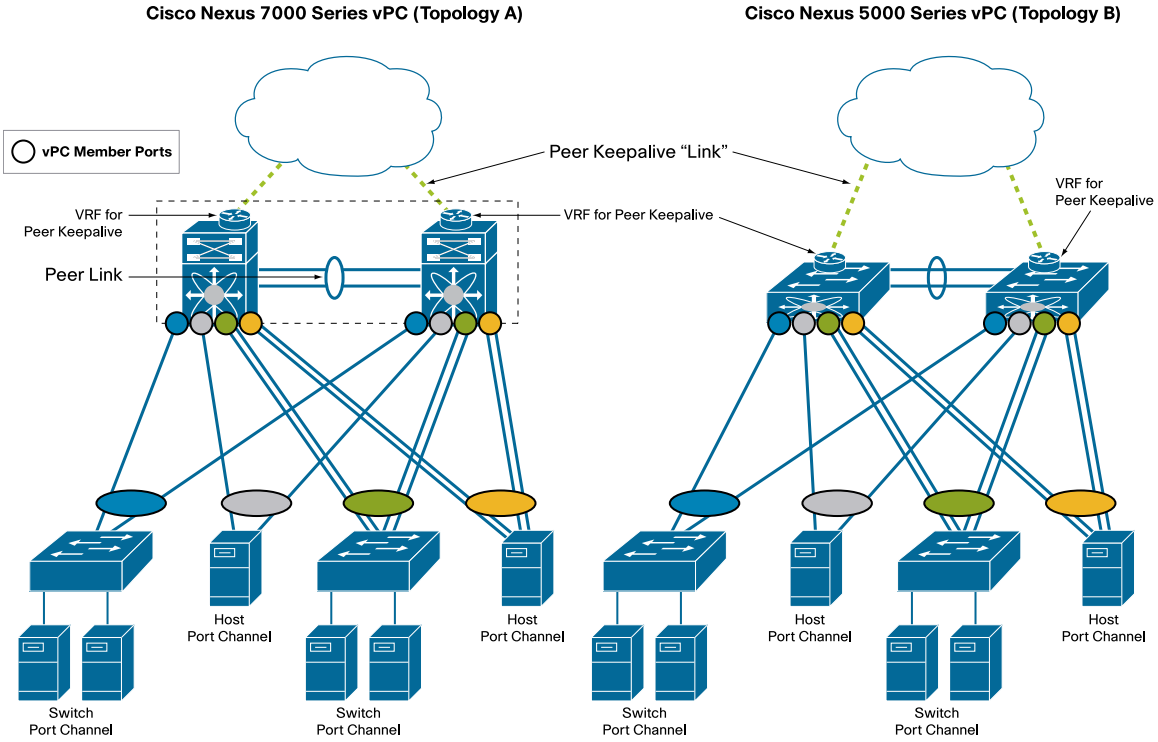
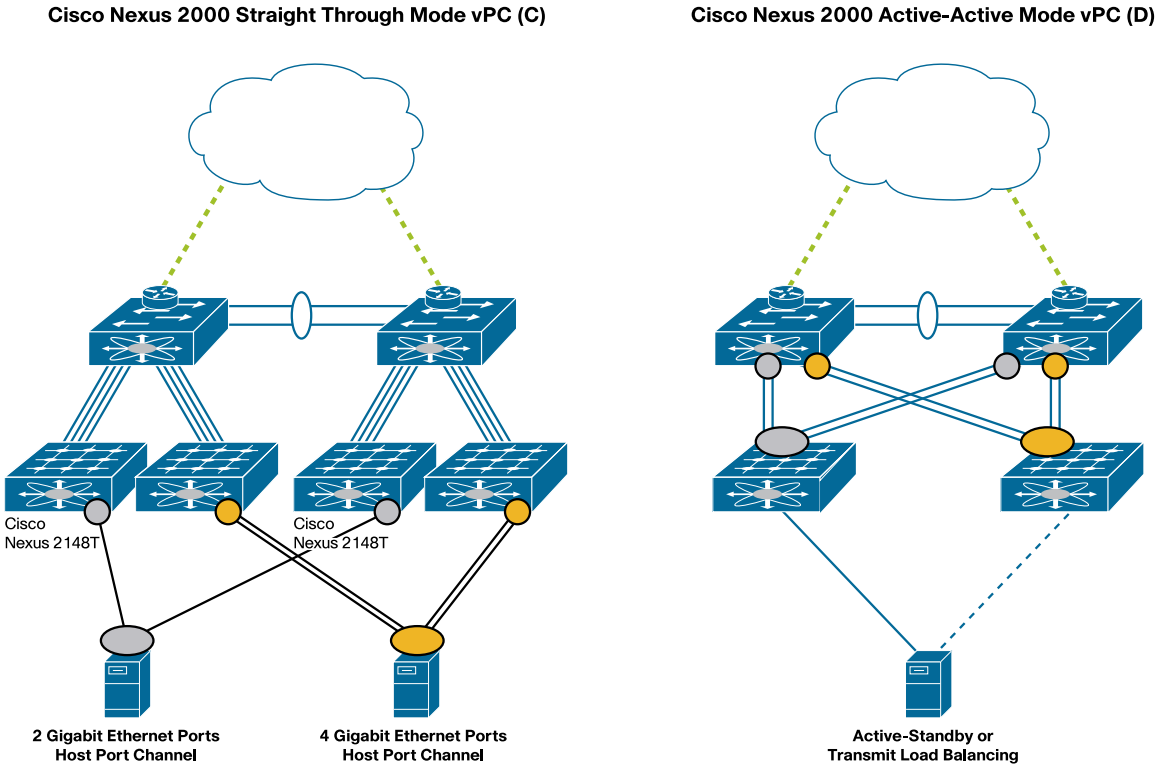


Figure 2. vPC Topologies C and D



Virtual PortChannel Components

The fundamental concepts of vPC are described at: http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-516396.html.

vPCs consist of two vPC peer switches, which are referred to as Switch1 and Switch2 in Figure 3. Of the vPC peers one is primary and one is secondary. The system formed by Switch1 and Switch2 is referred to as a vPC domain.

The vPC peer switches are connected through a link called **peer link**, also known as **multichassis EtherChannel trunk (MCT)**.

The peer link has several uses, including synchronizing MAC addresses between Agg1 and Agg2, providing the necessary transport for multicast traffic, and also for the communication of orphaned ports.

The ports that form the PortChannel are split between the vPC peers and are referred to as **vPC member** ports.

A routed “link” (it is more accurate to say “path”) is used to resolve dual-active scenarios where the peer link connectivity is lost. This link is referred to as **vPC peer-keepalive** or **fault-tolerant link**. The peer-keepalive traffic is often transported via the management network via the port management 0 of the Nexus 5000 or the management 0 ports on each Cisco Nexus 7000 supervisor. The peer-keepalive traffic is typically routed over a dedicated VRF (which could be the management VRF as an example).

Figure 3 shows devices (switch 3, switch 4 and server 2) that are connected to the vPC peers (which could be Cisco Nexus 7000 or 5000 Series Switches). These devices are configured with a normal PortChannel configuration.

Some devices are single-attached to the topology, like server 1 and server 3. The ports connecting devices in a non-vPC mode to a vPC topology are referred to as **orphaned ports**.

Figure 3. vPC Components

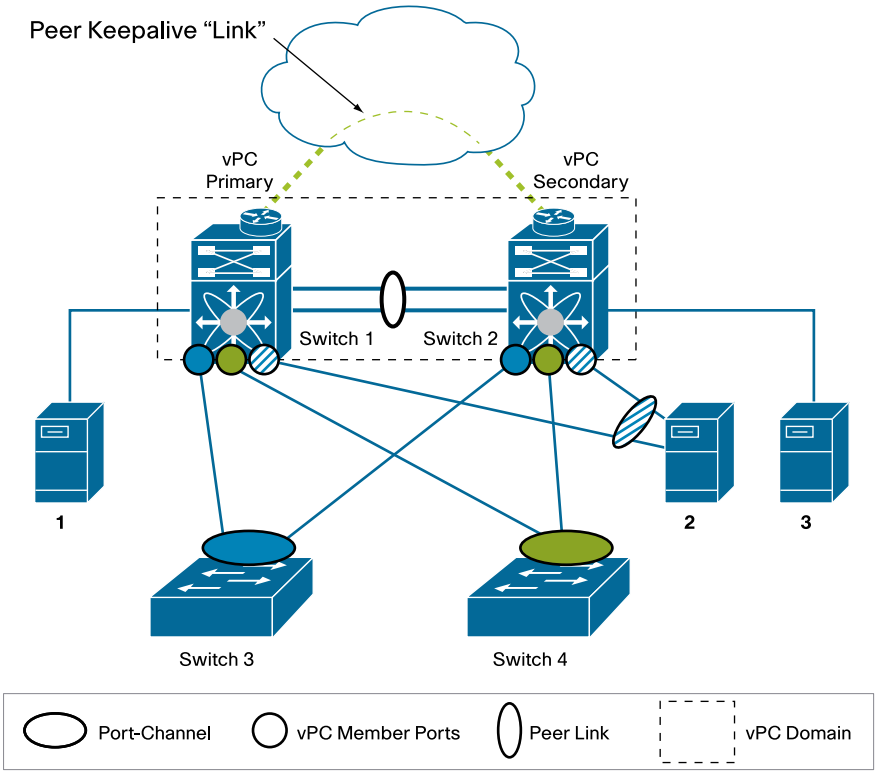
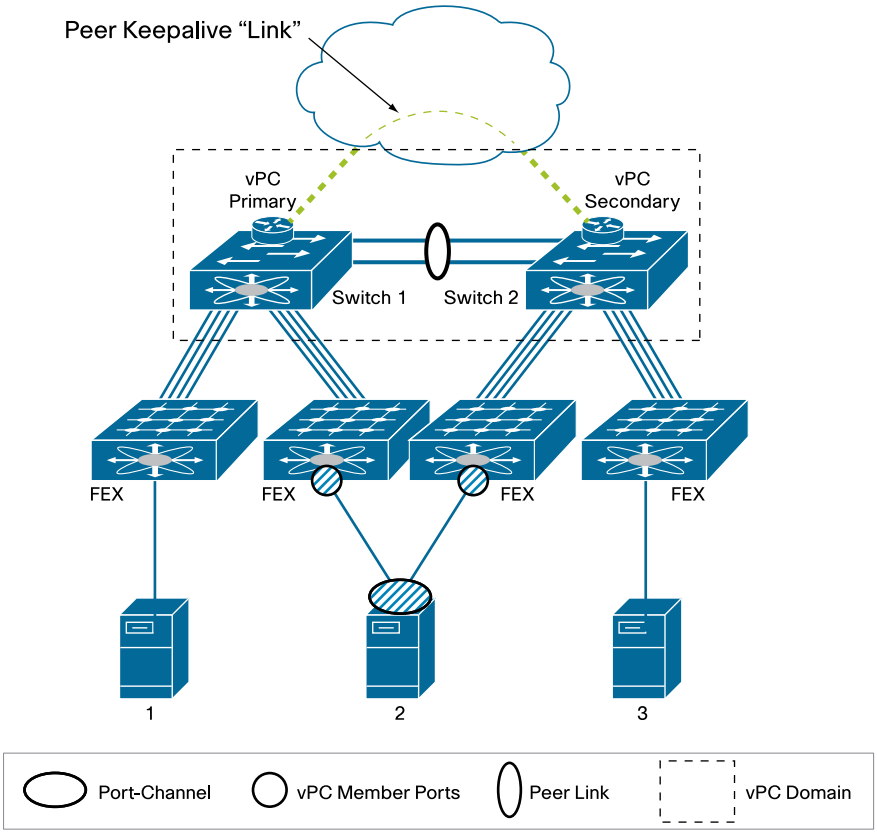


Figure 4 illustrates another vPC topology consisting of Cisco Nexus 5000 Series Switches and Cisco Nexus 2000 Series Fabric Extenders (in straight-through mode—that is, each fabric extender is single-attached to a Cisco Nexus 5000 Series Switch).

Figure 4 shows devices that are connected to the vPC peer (5k01 and 5k02) with a PortChannel (a vPC): like server 2 (configured for NIC teaming with the 802.3ad option).

Server 1 and server 3 connect to orphan ports.

Figure 4. vPC Components with the Fabric Extender



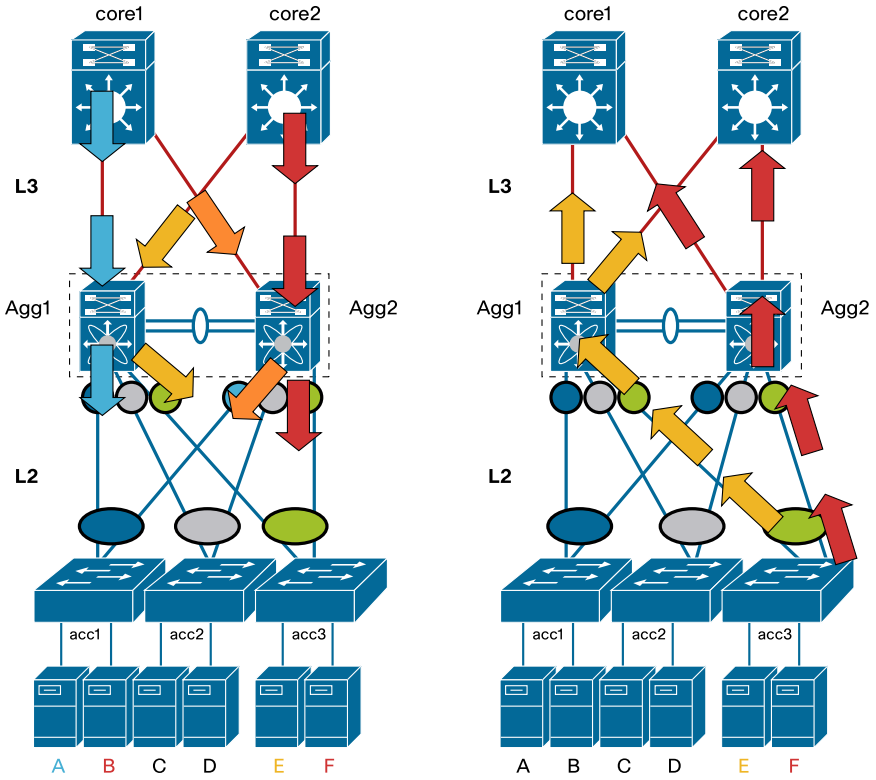
To summarize, a vPC system consists of the following components:

- Two peer devices: the vPC peers of which one is primary and one is secondary and are part of a vPC domain.
- A Layer 3 Gigabit Ethernet link called **peer-keepalive link** to resolve dual-active scenarios
- A redundant 10 Gigabit Ethernet PortChannel called a **peer link** to carry traffic from one system to the other when needed
- vPC member ports forming the PortChannel

Traffic Flows

vPC configurations are optimized to ensure that traffic through a vPC-capable system is symmetric. In Figure 5, for example, the leftmost flow (in blue) reaching a Cisco Nexus switch (Agg1 in the figure) from the core is forwarded toward the access (Acc1 in the figure) without traversing the peer Cisco Nexus switch device (Agg2). Similarly, traffic from the server directed to the core reaches a Cisco Nexus 7000 Series Switch (Agg1) and the receiving Cisco Nexus 7000 Series Switch routes it directly to the core without unnecessarily passing it to the peer Cisco Nexus 7000 Series Device. This happens regardless of which Cisco Nexus 7000 Series device is the primary Hot Standby Router Protocol (HSRP) device for a given VLAN.

Figure 5. Traffic Flows with vPC



Dual-Control Plane with Single Layer 2 Node Behavior

While still operating with two separate control planes, vPC ensures that the neighboring devices connected in vPC-mode perceive the vPC peers as a single spanning-tree and **Link Aggregation Control Protocol** (LACP) entity. For this to happen, the system has to perform IEEE 802.3ad control plane operations in a slightly modified way (which is not noticeable to the neighbor switch).

The Link Aggregation Group Identifier

IEEE 802.3ad specifies the standard implementation of PortChannels. Port channeling specifications provide LACP as a standard protocol, which makes it possible to negotiate port bundling.

LACP makes misconfiguration less likely, in that if ports are mismatched, they will not form a PortChannel.

Consider example A in Figure 7, in which switch 1 connects to switch 2. Port 1 on switch 1 connects to port 4 on switch 2, and port 2 on switch 1 connects to port 6 on switch 2.

Now imagine that the administrator configured PortChannel on switch 1 between ports 1 and 2, while on switch 2 the PortChannel is configured between ports 5 and 3. Without LACP, it would be possible to just put the ports in channel-group mode, and you would not discover that this is a misconfiguration until you notice that traffic has dropped.

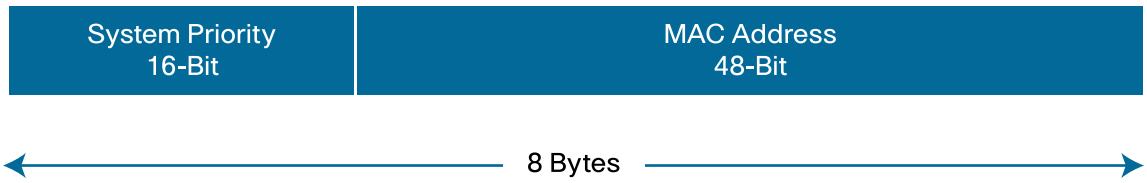
LACP discerns that the only ports that can be bundled are port 1 going to port 4. According to the IEEE specifications, to allow Link Aggregation Control Protocol to determine whether a set of links connect to the same system, and to determine whether those links are compatible from the point of view of aggregation, it is necessary to be able to establish:

- A globally unique identifier for each system that participates in link aggregation (that is, the switch itself needs to be “unique.”) This number is referred to as system ID and is composed of a priority and a MAC address that uniquely identifies the switch. Figure 6 illustrates what the System ID looks like.

- A means of identifying a link aggregation group.

For more information please refer to IEEE 802.3ad standard, Amendment to Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications— Aggregation of Multiple Link Segments

Figure 6. The Components of the System ID



In Figure 7, switch 1 announces ports 1and 2 as part of the same aggregation group, and similarly switch 2 announces ports 4 and 5 as part of the same aggregation group. Because ports 3 and 6 are not part of the group, they cannot bundle with the PortChannel.

Example A in Figure 7 shows an extreme case in which the PortChannel consists of an individual port only. If the negotiation had failed between switches 1 and 2, the links would still operate as normal, individual IEEE 802.3 links.

The way that switch 1 and switch 2 decide which ports can bundle together is based on the identifier LAGID, which stands for link aggregation group identifier. This number includes the system identifier (in other words, an ID for the physical switch) and a key that identifies the aggregation group itself (that is, the equivalent of the channel group number).

As a first approximation, the LAGID is composed of the system ID of both systems and the channel group number used in both systems.

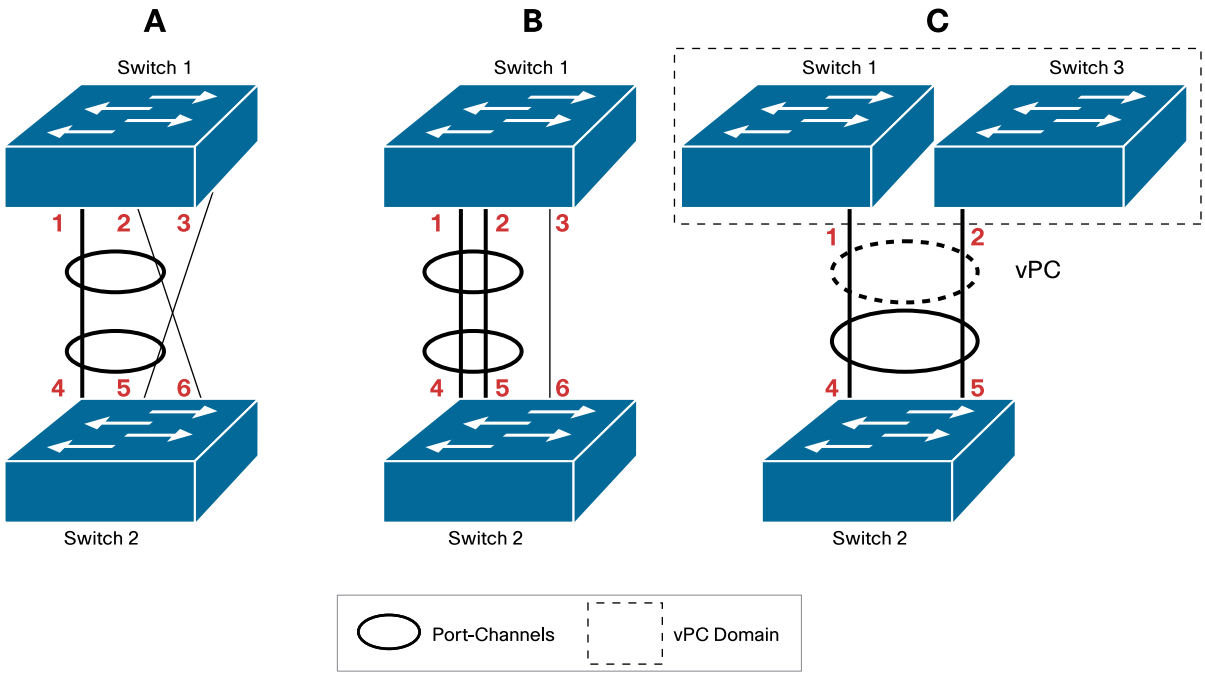
Example B in Figure 7 shows the case in which the ports are correctly wired. Assuming that the ports on switch 1 (system ID 1) are bundled as channel group 100 and the ports on switch 2 (system ID 2) are bundled as channel group 200, the LAGID would appear to be as follows: **[1, 100, 2, 200]**.

Example C in Figure 7 illustrates how the PortChannel is requested between switch 2 and two separate upstream switches, switch 1 and switch 3, where switch 1 and 3 form a vPC system.

The system ID for switch 1 differs from the system-id for switch 3 because the MAC address of the two switches is different.

With vPC, the system ID of switch 1 and switch 3 is identical so that switch 2 believes it is connected to a single upstream device.

Figure 7. LACP Behavior with Various Wiring Configurations



System ID in a vPC System

Spanning tree and LACP use the switch MAC address for bridge ID field in the spanning tree bridge protocol data unit (BPDU) and as part of LACP LAGID, respectively. In a single chassis, they use the systemwide MAC address for this purpose. For systems that use vPCs, using the systemwide MAC address would not work because the vPC peers needs to appear as a single entity as exemplified in Figure 7-C. In order to address this requirement, vPC offers either an automatic configuration or a manual configuration of the system-id for the vPC peers.

The automatic solution implemented by vPC consists in generating a system ID composed of a priority and MAC address, where the MAC is derived from a reserved pool of MAC addresses combined with the domain ID specified in the vPC configuration. The domain ID is encoded in the last octet and the trailing 2 bits of the previous octet of the MAC address.

By configuring domain IDs to be different on adjacent vPC complexes (and to be identical on each vPC peer complex), you will ensure the uniqueness of the system ID for LACP negotiation purposes. You also guarantee that the Spanning Tree Protocol BPDUs use a MAC address that is representative of the vPC complex.

It is also possible to override the automatic generation of the system ID by using the CLI and configuring the system-id on both vPC peers, as follows:

```
(config-vpc-domain)#system-mac <mac>
```

Primary and Secondary vPC Roles

In a vPC system, one vPC switch is defined as primary and one is defined as secondary, based on defined priorities. The **lower number has higher priority**, so it wins.

Also, these roles are **nonpreemptive**, so a device may be operationally primary, but secondary from a configuration perspective.

Spanning Tree

vPC modifies the way in which spanning tree works on the switch in two ways:

- It makes sure that the **peer link is always forwarding**. In fact, even if the switch has a direct path to the root, the secondary vPC peer **always sees the peer link as the root port towards the primary vPC device**.
- It ensures that only the primary switch forwards BPDUs on vPCs, so that the other switches connected to the vPC system perceive the two peers as a single entity from a spanning tree perspective. This modification is strictly limited to vPC-member ports.

As a result, the BPDUs that may be received by the secondary vPC peer on a vPC port are forwarded to the primary vPC peer through the peer link for processing.

Note: Non-vPC ports operate like regular ports. The special behavior of the primary vPC member applies uniquely to ports that are part of a vPC.

CDP

From a Cisco Discovery Protocol perspective, the presence of vPC doesn't hide the fact that the two Cisco Nexus 7000 Series systems are two distinct devices, as illustrated by the following output:

```
tc-nexus5k01# show cdp neigh
Capability Codes: R - Router, T - Trans-Bridge, B - Source-Route-Bridge
                  S - Switch, H - Host, I - IGMP, r - Repeater,
                  V - VoIP-Phone, D - Remotely-Managed-Device,
                  s - Supports-STP-Dispute

Device-ID            Local Intrfce Hldtme Capability  Platform      Port ID

tc-nexus7k01-vdc2(TBM12162254)Eth2/1    158    R S I s   N7K-C7010     Eth2/9
tc-nexus7k02-vdc2(TBM12193229)Eth2/2    158    R S I s   N7K-C7010     Eth2/9
```

Cisco Fabric Services over Ethernet Synchronization Protocol

The vPC peers use the Cisco Fabric Services protocol to synchronize forwarding-plane information and implement necessary configuration checks or both.

vPC peers must be synchronized because the Layer 2 forwarding table — that is, the MAC address information between the vPC peers—must be synchronized. This way, if one vPC peer learns a new MAC address, that MAC address is also programmed on the Layer 2 forwarding table of the other peer device.

The Cisco Fabric Services protocol travels on the peer link and does not require any configuration from the user.

In order to ensure that the peer link communication for the Cisco Fabric Services over Ethernet protocol is always available, spanning tree has been modified to keep the peer link ports always forwarding.

As described in the next section, the Cisco Fabric Services over Ethernet protocol is also used to perform the compatibility checks to validate vPC member ports compatibility to form the channel, to synchronize the IGMP snooping status, and to monitor the status of the vPC member ports.

COMPAT Checks

COMPAT check stands for **compatibility check**. During a compatibility check, one vPC peer conveys configuration information to the other vPC peer in order to verify that vPC member ports can actually form a PortChannel. As an example, if two ports that are going to join the channel carry a different set of VLANs, this is a misconfiguration.

Depending on the severity of the misconfiguration, vPC may either warn the user (Type 2 misconfiguration) or suspend the PortChannel (Type 1 misconfiguration). In the specific case of a VLAN mismatch, only the VLAN that differs between the vPC member ports is going to be suspended on all the vPC PortChannels.

You can verify the status of consistency between the vPC peers by using the command **show vpc consistency-parameter global**, as follows:

```
tc-nexus5k02# show vpc consistency-parameter
```

Inconsistencies can be global or interface specific:

- Global inconsistencies: Type 1 global inconsistencies affect all vPC member ports (but don't affect non-vPC ports).
- Interface-specific inconsistencies: Type 1 interface-specific inconsistencies affect only the interface itself.

Examples of global inconsistencies include **spanning-tree mode mst** on one peer and **spanning-tree mode rapid-pvst** on the other peer. Another example of a global inconsistency is mismatched MST regions.

vPC Peer Link

The vPC peer link is the most important connectivity element of the vPC system. This link is used to create the illusion of a single control plane by forwarding BPDUs or LACP packets to the primary vPC switch from the secondary vPC switch. It is also used for forwarding traffic that originates at or is destined for orphan ports. It carries flooded frames, and in the case of the Cisco Nexus 7000 Series, Hot Standby Router Protocol (HSRP) frames.

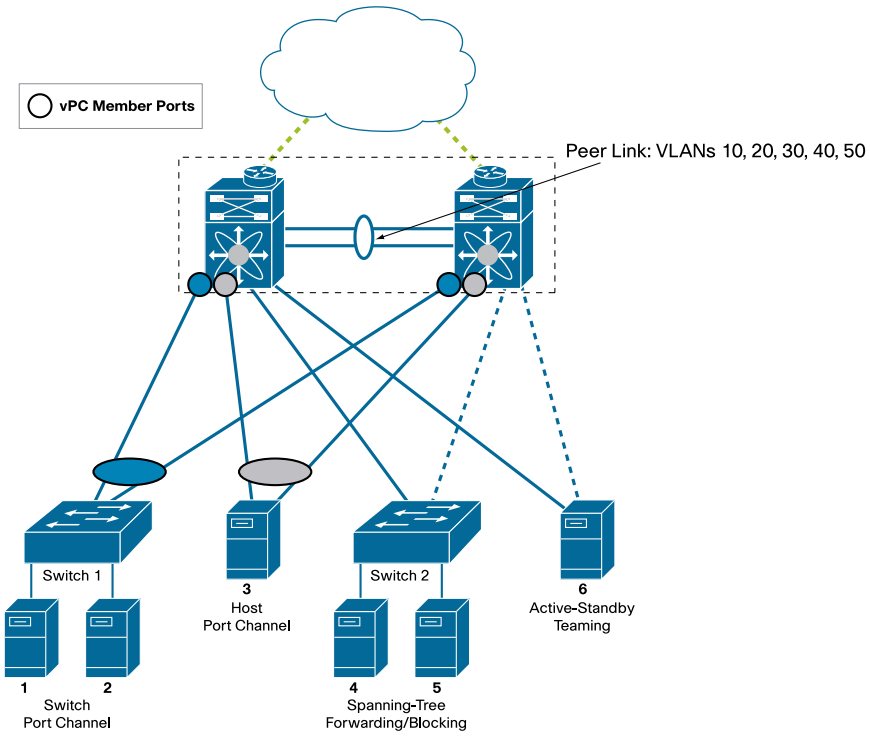
vPC Peer-Keepalive or Fault-Tolerant Link

The peer-keepalive link is used to resolve dual-active failures (that is, failures where the peer link between vPC peers is lost). The keepalive can be carried over a routed infrastructure; it does not need to be a direct point-to-point link, and, in fact, it is desirable to carry the peer keepalive traffic on a different network than on a straight point-to-point link.

vPC VLANs, vPC Ports, and Orphaned Ports

The concept of a vPC VLAN is important when a peer link is lost and is mostly relevant for the Cisco Nexus 7000 Series case. For a VLAN to be categorized as a vPC VLAN, it is sufficient that this VLAN is configured on the peer link. By default, being configured on the peer link makes a VLAN a vPC VLAN. As an example in Figure 8, VLANs 10, 20, 30, 40 and 50 are vPC VLANs.

Figure 8. vPC VLANs and Orphaned Ports



For a VLAN to be on the peer link, it must exist on both vPC peers, and it must appear in the allowed list of the switch port trunk. If either of these conditions is not met, the VLAN isn't displayed when you enter the command **show vpc brief**, nor is it a vPC VLAN.

A vPC port is a port that is assigned to a vPC channel group. A non-vPC port, also known as orphaned port, is a port that is not part of a vPC.

Figure 8 shows different types of Orphaned Ports. Switch 2 connects to the Cisco Nexus 7000 with a regular Spanning-Tree configuration, hence one link is forwarding and one link is blocking. These links connect to the Cisco Nexus 7000 with "orphaned" ports.

Server 6 connects to the Cisco Nexus 7000 with an active/standby teaming configuration. The ports that Server 6 connects to on the Cisco Nexus 7000 are "orphaned" ports.

Duplicate Frames Prevention Technique in vPC

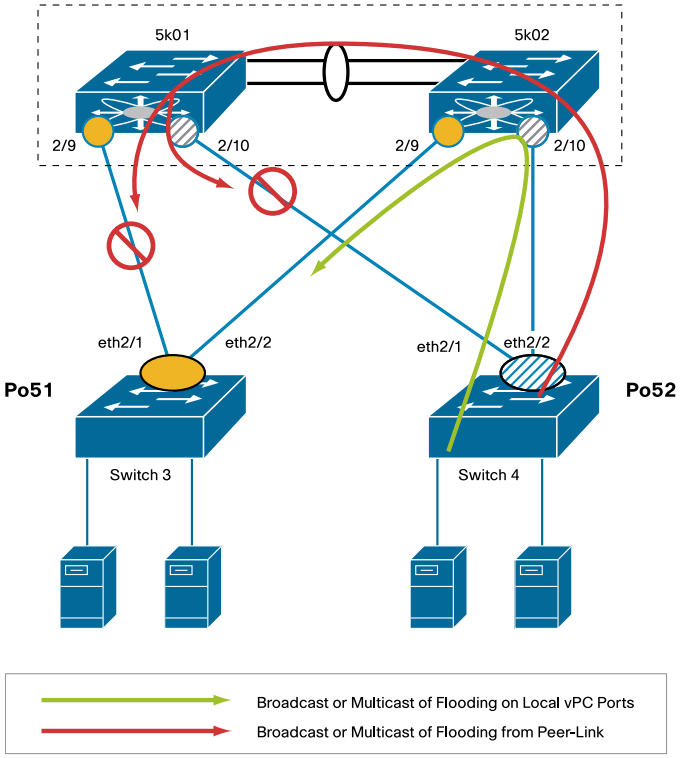
One of the most important forwarding rules of vPC is the fact that a frame that entered the vPC peer switch from the peer link cannot exit the switch out of a vPC member port (except if this is coming from an orphaned port).

Figure 9, shows switch 3 and switch 4 connected to switches 1 and 2 with vPCs Po51 and Po52. A host sending either an unknown unicast or a broadcast that gets hashed to port Ethernet2/2 on switch 3 on PortChannel 52. Switch 2 receives the broadcast and needs to forward it to the peer link for the potential orphan ports on switch 1 to receive it.

Upon receiving the broadcast, switch 1 detects that this frame is coming from a vPC member port. Therefore, it does not forward it to port 2/10, or a duplicate frame on switch 3 would be created.

If a host on switch 4 sends a broadcast, switch 2 will correctly forward it to Po51 on port 2/9 and place it on the peer link. Switch 1 will prevent this broadcast frame from exiting onto ports 2/9 or 2/10 because this frame entered switch 2 from a vPC member port. Should port 2/2 on switch 3 go down, the **masking information** sent by switch 2 gets updated in order for switch 1 to forward the frame out of port 2/9.

Figure 9. vPC Doesn't Introduce Duplicate Frames

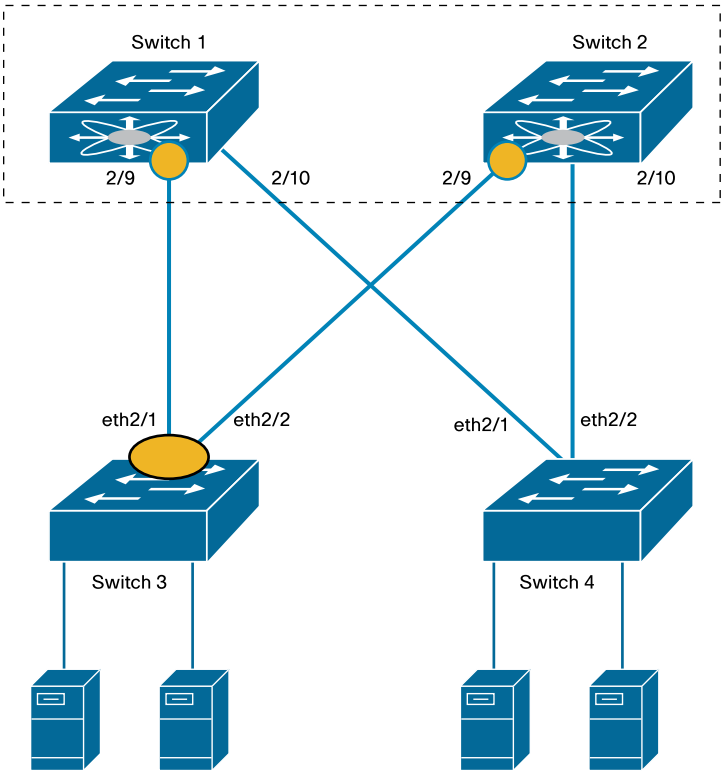


Finally, it is also important to realize that a topology based on PortChannels doesn't introduce loops, even if the peer link is lost and all the ports are forwarding. Figure 10 explains why.

Figure 10 illustrates the worst-case scenario of a vPC dual-active failure when both peer-link and peer-keepalive connectivity is lost. In this particular case one switch is running spanning tree (switch 4) with links that are not in PortChannel mode, while the other switches are configured in PortChannel mode. Even with this scenario, the worst that can happen is duplicate frames.

With all links forwarding, a broadcast frame or an unknown unicast generated on switch 4, for example, is forwarded only on both links directed to switch 1 and switch 2. When these two frames arrive on switch 3, they are not sent back to the PortChannel because that breaks the basic rule of Layer 2 forwarding—namely, that a frame cannot return to the port from which it originated.

Figure 10. Worst Case of Dual Active Failure



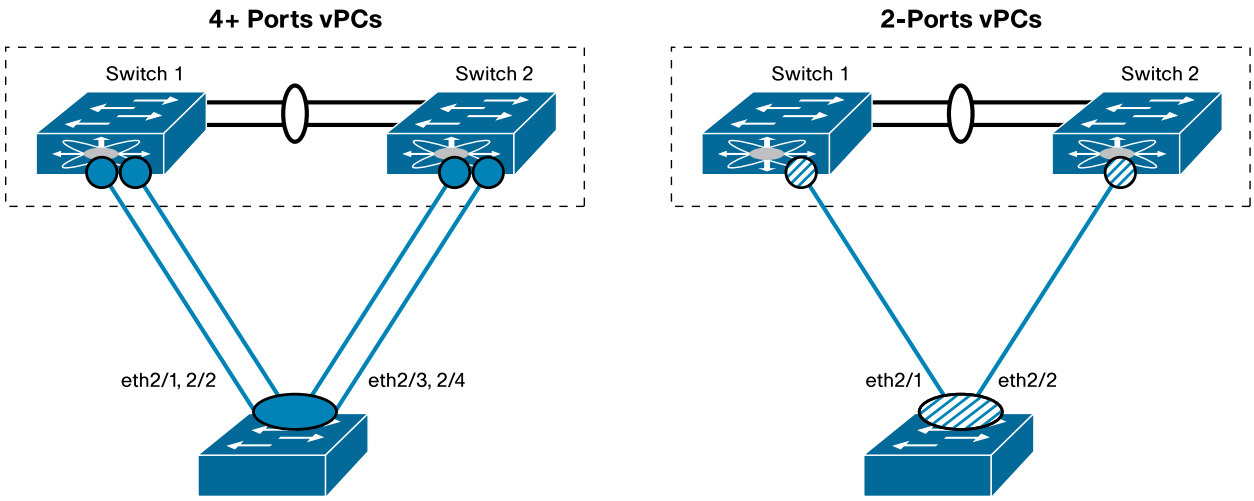
2-Port vPC Versus 4+-Port vPC

When you use virtual PortChannels, it is important and useful to categorize two configuration types:

- **4+-port vPCs with two links to each Cisco Nexus switch:** On the Cisco Nexus 5000 this configuration consumes one full hardware PortChannel out of each individual switch. For example, if the vPC consists of four ports, each switch locally owns a 2-port PortChannel.
- **2-port vPC with one link to each Cisco Nexus switch:** With this configuration, each switch individually owns an individual link. On the Cisco Nexus 5000 Series platform, this type of EtherChannel technology doesn't consume any hardware resources.

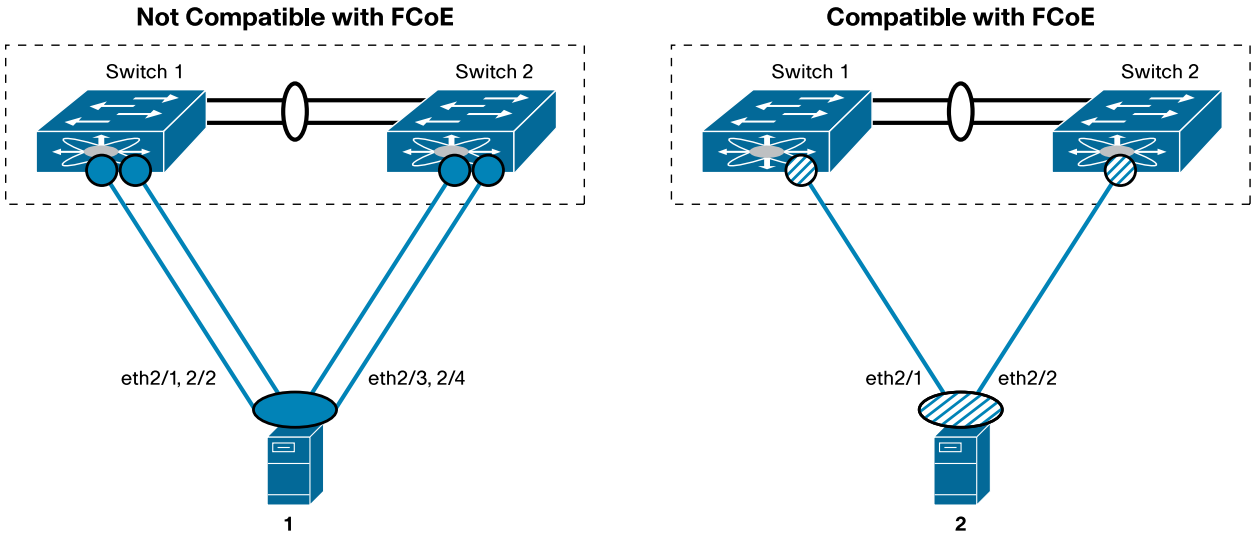
As the example in Figure 11 shows, at the time of this writing on the Cisco Nexus 5000 Series you can configure a maximum of 16 hardware port channels. But you can configure more than 16 virtual PortChannels, where the vPC has one port connected to one Cisco Nexus 5000 Series device and the other port connected to the other Cisco Nexus 5000 Series, either directly or through the fabric extender.

Figure 11. 4+ Ports vPCs and 2 Ports vPCs



In the case of Unified Fabric (Fibre Channel over Ethernet [FCoE]) deployments, it is important to distinguish 4+ ports Host vPCs from 2-Ports Host vPCs as depicted in Figure 12. In order for FCoE to work, the server adapter needs to see two separate Fibre Channel fabrics. This is only possible when the Converged Network Adapter card has only 1 link per Cisco Nexus 5000 as in the case of server 2 in Figure 12.

Figure 12. 4+ Ports vPCs and 2 Ports vPCs and FCoE



In-Service Software Upgrade and vPC

In presence of vPC, it is possible to upgrade a device such as the Cisco Nexus 7000 Series using an In-Service Software Upgrade (ISSU) with no disruption to the traffic. However, if someone modifies the vPC configuration during the upgrade, it will cause an inconsistency between the vPC peer devices (the one being upgraded and the other device).

To avoid this undesirable situation, vPC is capable of locking the configuration on the device that is not undergoing the upgrade and releasing it when the upgrade is complete.

vPC Failure Scenarios

This section describes the expected behavior of a vPC design for various link failures.

vPC Member Port Failure

If one vPC member port goes down, it is removed from the PortChannel without bringing down the virtual PortChannel entirely. Conversely, the switch on which the port went down will properly unmask this vPC number when sending frames over the peer link (recall the vPC duplicate avoidance technique) in order for the vPC peer to forward the traffic to the remaining vPC member port. The Layer 2 forwarding table on the switch that detected the failure is also updated to point the MAC addresses that were associated with the vPC port toward the peer link.

vPC Peer Link Failure

The following happens when the peer link fails:

- The operational secondary vPC peer (which may not match the configured secondary because vPC is nonpreemptive) brings down the vPC member ports, including the vPC member ports located on the fabric extender in the case of a Cisco Nexus 5000 Series design with fabric extender in straight-through mode.
- On the Cisco Nexus 7000 Series, the secondary vPC peer brings down the vPC VLANs SVIs, that is, all SVIs for the VLANs that happen to be configured on the vPC peer link, whether or not they are used on a vPC member port.
- On the Cisco Nexus 7000 Series, the primary vPC peer brings down the SVI for vPC VLANs for which there is no forwarding vPC member interface.

Note: To keep the SVI interface up when a peer link fails, use the command **dual-active exclude interface-vlan**.

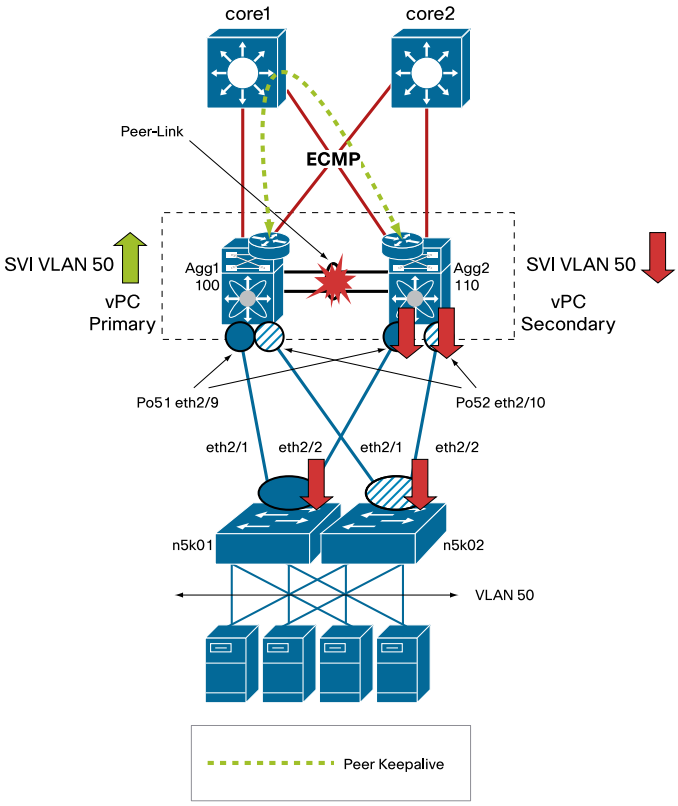
Cisco Nexus 7000 Series Example

Figure 13 illustrates what happens during vPC peer link failure for vPC ports. Agg1 is the vPC primary and Agg2 is the vPC secondary.

The sequence of events is as follows:

1. The vPC peer link fails, but Agg1 and Agg2 can still communicate via the routed path with the vPC peer keepalive protocol
2. Ethernet 2/9 and 2/10 on Agg2 are brought down because they are part of vPC Po51 and Po52 respectively and Agg2 is the operational secondary vPC device.
3. SVI VLAN50 (vPC-VLAN) is brought down on the operational secondary device in order to prevent traffic from the core routers from reaching the vPC secondary device where the vPC ports are shut down.

Figure 13. Peer-Link Failure



As a result of the peer-link failure all traffic in Figure 13 takes the leftmost path via the vPC primary device. This is true both for the client-to-server traffic and the server-to-client traffic.

The following show command issued on the secondary vPC peer illustrates the result of the vPC Peer Link failure.

```
tc-nexus7k02-vdc2# show vpc br

vPC domain id                     : 1
Peer status                       : peer link is down
vPC keep-alive status             : peer is alive
vPC role                          : secondary
Dual Active Detected

vPC Peer Link Status
--------------------------------------------------------------------
id   Port   Status Active vLANs
--   ----   ------ -----------------------------------------------
1    Po10   down   -

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vLANs
--   ----   ------ ----------- -------------------------- ------------
51   Po51   down   success     success                    -

The access switch uses the remaining link:
```

```
tc-nexus5k01# show port channel summary
--------------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports
      Channel
--------------------------------------------------------------------------------
51    Po51(SU)    Eth      LACP      Eth2/1(P)    Eth2/2(D)
```

The peer keepalive communication ensures that the loss of the peer-link path doesn't introduce any unwanted flooding, or split subnet scenarios.

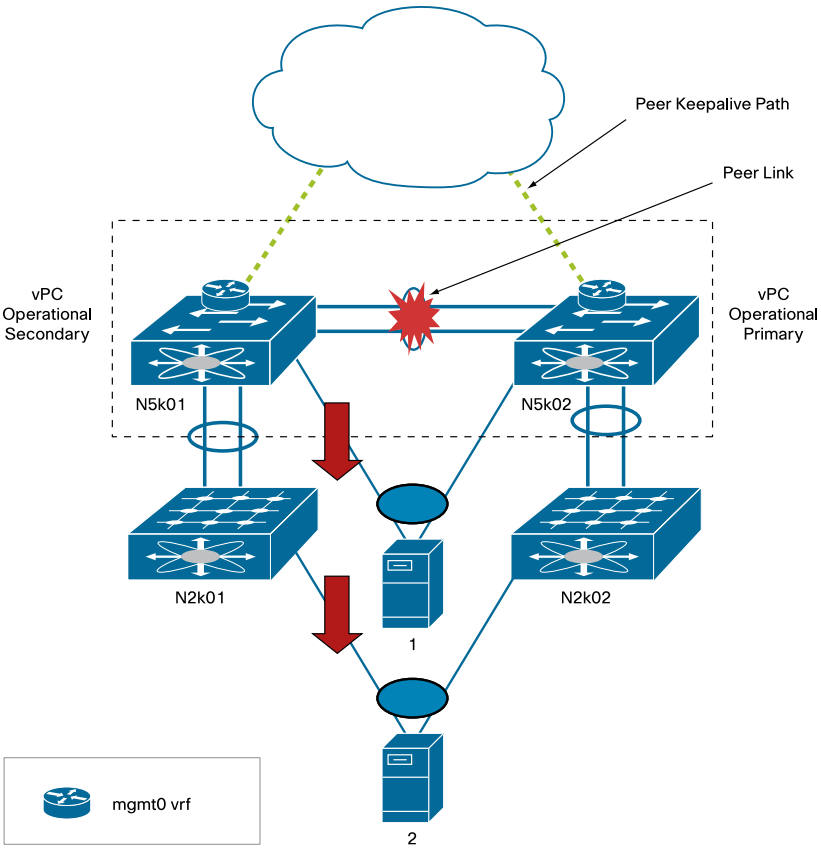
Cisco Nexus 5000 Series Example

In the case of the Cisco Nexus 5000 Series, a peer link failure causes the operational secondary switch to shut down the vPC member ports so that the host and switch PortChannels can continue forwarding along the active path.

Figure 14 shows that the operational secondary shuts down the vPC member port to host 1, which is directly attached to Cisco Nexus 5000 Series N5k01 and the vPC member port of host 2 connected to the Cisco Nexus 2000 Series N2k01.

Because of this, in vPC configurations, hosts connected to a Cisco Nexus switch should be configured for port channeling as a NIC teaming option. If hosts are single-homed their ports are treated as orphaned ports, hence the failure would be similar to the well known split subnet scenario of traditional layer 2 switching. This applies to both the Cisco Nexus 5000 and 7000 Series.

Figure 14. vPC Peer-Link Failure on the Cisco Nexus 5000



vPC Complete Dual-Active Failure (Double Failure)

In case both the peer link and the peer keepalive link get disconnected, the Cisco Nexus switch does not bring down the vPC, because each Cisco Nexus Switch cannot discriminate between a vPC device reload and a peer-link- plus peer-keepalive failure. This means that each vPC member port keeps advertising the same LACP ID as before the dual-active failure.

As described previously, a vPC topology intrinsically protects from loops in case of dual-active scenarios. In the worst case, there will be duplicate frames. Despite this, as a loop prevention mechanism, each switch starts forwarding BPDUs with the same BPDU Bridge ID as prior to the vPC dual active failure.

vPC Dual-Active Failure

The following output illustrates the status of the former primary vPC peer:

```
nexus7k01-vdc2# show vpc br
Legend:
                (*) - local vPC is down, forwarding via vPC peer link

vPC domain id                   : 1
Peer status                     : peer link is down
vPC keep-alive status           : Suspended (Destination IP not reachable)
Configuration consistency status: success
vPC role                        : primary
```

The following output illustrates the status of the former secondary vPC peer, now the operational primary:

```
nexus7k02-vdc2# show vpc br
Legend:
                (*) - local vPC is down, forwarding via vPC peer link

vPC domain id                   : 1
Peer status                     : peer link is down
vPC keep-alive status           : Suspended (Destination IP not reachable)
Configuration consistency status: success
vPC role                        : secondary, operational primary

vPC Peer link status
------------------------------------------------------------------------
id   Port   Status Active vLANs
--   ----   ------ -----------------------------------------------------
1    Po10   down   -

vPC status
----------------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vLANs
--   ----   ------ ----------- -------------------------- ------------
51   Po51   up     success     success                    11-14,21-24
                                                         ,50,60
52   Po52   up     success     success                    11-14,21-24
```

As you can see, the vPCs are kept up on both systems, and they are recognized as valid PortChannels from the access layer device:

```
51     Po51(SU)     Eth      LACP       Eth2/1(P)     Eth2/2(P)
```

While not intuitive, if both the vPC peer link and Peer Keepalive link fail, it is still possible and desirable to continue forwarding traffic from the access layer to the aggregation layer without drops for existing traffic flows, provided that the Address Resolution Protocol (ARP) tables are already populated on both Cisco Nexus 7000 Series peers for all needed hosts.

If new MAC addresses are to be learned by the ARP table, issues may arise because the ARP response from the server may always be hashed to one Cisco Nexus 7000 Series device and not the other, making it impossible for the traffic to flow correctly.

Suppose, however, that before the failure in the situation just described, traffic was equally distributed to both Cisco Nexus 7000 Series by a correct PortChannel and by Equal Cost Multipath (ECMP) configuration. In that case, server-to-server and client-to-server traffic continues with the caveat that single-attached hosts connected directly to the Cisco Nexus 7000 Series will not be able to communicate (for the lack of the peer link). Also, new MAC addresses learned on one Cisco Nexus 7000 Series cannot be learned on the peer, as this would cause flooding for the return traffic that arrives on the peer Cisco Nexus 7000 Series device.

Virtual PortChannel Design Considerations

vPC Role and Priority

First, vPC needs to be enabled:

```
agg(config)# feature vpc
```

A domain needs to be defined (as indicated by the domain-id) as well as priorities **to define primary and secondary roles** in the vPC configuration. The **lower number has higher priority**, so it wins. For two switches (vPC peers) to form a vPC system, the domain-id of these switches need to match. As previously described, the domain-id is used to generate the LAGID in the LACP negotiation.

```
agg1(config)# vpc domain <domain-id>

agg1(config-vpc-domain)# role priority 100

agg2(config)# vpc domain <domain-id – same as agg1>

agg2(config-vpc-domain)# role priority 110
```

It should also be noted that the role is **nonpreemptive**, so a device may be operationally primary but secondary from a configuration perspective. Because spanning tree is preemptive, this may result in a mismatch between the spanning tree root and the vPC operational primary device with no consequences to traffic forwarding.

While mismatched Spanning-Tree and vPC priorities do not any impact on traffic forwarding, it is still desirable to keep the priorities matched so as to have Spanning-Tree root and vPC primary on the same device and Spanning-Tree secondary root and vPC secondary on the same device. The main benefit is easier management.

In case after failovers the vPC operational primary and vPC operational secondary do not match the original configuration, you can restore it following these easy configuration steps:

From the vPC *operational* primary you can change the role priority to the highest value 32768, then do a shut/no shut on the peer-link PortChannel.

You can also script this as follows:

```
7k-1(config)# cli alias name vpcpreempt conf t ; vpc domain <domain-id> ;
role priority 32767 ; int <peer-link> ; shut ; no sh *
```

vPC Peer Link

A PortChannel connects agg1 with agg2 and carries all access VLANs (defined by the user). This link also carries additional traffic that the user does not need to define, more specifically BPDUs and HSRP hellos, and MAC address synchronization between the vPC peers. This link is called **peer link**.

On the Cisco Nexus 7000 Series, this PortChannel should be configured on **dedicated-mode** 10 Gigabit Ethernet interfaces **across two different 10 Gigabit Ethernet line cards**.

This is by far the most important component in the vPC system, in that its failure, while not disruptive to existing vPC flows, may impair the establishment of new flows and isolate orphan ports. Configuring the peer link in a redundant fashion ensures virtually uninterrupted connectivity between the vPC peers. The following configuration illustrates how to configure the peer-link, which in this case is PortChannel 10.

```
agg(config)# interface port-channel10
agg(config-if)#  vpc peer-link
agg(config-if)#  switchport trunk allowed vLAN <all access vLANs>
```

vPC VLANs and non-vPC VLANs

The PortChannel connecting the vPC peers should carry all the VLANs used by the vPC member ports.

It's also possible to carry the VLAN used by orphaned ports on this same link with the same caveat as a regular non-vPC topology, which is, that upon losing the peer-link, communication between orphaned ports is interrupted (split subnet). If you want to avoid this problem you should make sure that servers are dual connected with a PortChannel to vPC ports.

Alternatively, if you want to decouple vPC and non-vPC failure scenarios, you can use different VLANs for vPC-connected devices and single-port attached devices (orphaned ports), and put the non-vPC VLANs and the peer link on different trunks.

vPC Peer Keepalive

Finally, a dual-active detection configuration needs to be put in place. The keepalive that resolves dual-active scenarios should never be carried as a VLAN on the peer link. Instead, it can be carried over a routed infrastructure, and it doesn't need to be a direct point-to-point link.

The following configuration illustrates the use of a dedicated Gigabit Ethernet interface for this purpose:

```
vrf context vpc-keepalive

interface Ethernet8/16
  description tc-nexus7k02-vdc2 - vPC Heartbeat Link
  vrf member vpc-keepalive
  ip address 192.168.1.1/24
  no shutdown

vpc domain 1
peer-keepalive destination 192.168.1.2 source 192.168.1.1 vrf vpc-keepalive
```

You should not use the mgmt0 interface for a direct back-to-back connection between Cisco Nexus 7000 systems because you cannot discern which supervisor is active at any given time. You can use it instead on the Cisco Nexus 5000 Series.

The mgmt0 interface can be used both for management and for routing the peer keepalive through the out-of-band management network. In this case, each Cisco Nexus 7000 Series is connected to the management network through both the mgmt0 of supervisor slot 5 and supervisor slot 6 and the Cisco Nexus 5000 Series through the single mgmt0 interface.

By following this approach, regardless of which supervisor is active, the Cisco Nexus 7000 Series has one of the mgmt0 interfaces connected to the management network, which can then be used for peer keepalive purposes.

vPC Ports

PortChannels are configured by bundling Layer 2 ports (switch ports) on each Cisco Nexus switch through the command **vpc**, as shown in the following code. The system issues an error message if the PortChannel wasn't previously configured as a **switchport**.

```
agg1(config)#interface ethernet2/9
agg1(config-if)# channel-group 51 mode active
agg1(config)#interface Port-channel 51
agg1(config-if)# switchport
agg1(config-if)# vpc 51
!
agg2(config)#interface ethernet2/9
agg2(config-if)# channel-group 51 mode active
agg2(config)#interface Port-channel 51
agg2(config-if)#switchport
agg2(config-if)# vpc 51
```

If the consistency check doesn't show **success**, it is recommended verifying the **consistency-parameters**. Typical reasons for the vPC not to form include:

- The VLAN that is defined in the trunk doesn't exist, or it is not defined on the peer link
- One member port is configured as access and the other as trunk
- **Mismatches in the VLANs that are carried on the trunk**, and so on

The example below illustrates how to verify that the vPC configuration is consistent between two vPC peers for the Global Consistency parameter as well as for a specific PortChannel.

```
tc-nexus7k01-vdc2# show vpc consistency-parameters global

tc-nexus7k01-vdc2# show vpc consistency-parameters int port-channel 51

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
STP Port Type               1     Default                Default
STP Port Guard              1     None                   None
STP MST Simulate PVST       1     Default                Default
Allowed vLANs               -     10-14,21-24,50,60      10-14,21-24,50,60
```

After a port is defined as part of a PortChannel, any further configurations, such as activation or disablement of bridge assurance or trunking mode are performed under the interface PortChannel configuration mode. Trying to configure spanning tree properties for the physical interface instead of the PortChannel will result in an error message.

Link Aggregation Control Protocol

It is always good practice to use the Link Aggregation Control Protocol (LACP) for dynamic bundling of the ports in the vPC group. This is because LACP determines that the ports being bundled are actually part of the same physical or virtual switch, preventing erroneous configurations.

As an example, if the PortChannel is configured as **active** on the Cisco Nexus 7000 Series and the downstream switch is not configured for port channeling, the PortChannel ports will show as **Individual** state (**I**) and run regular spanning tree.

Once the access layer switches are configured for LACP, the negotiation completes the PortChannel forms:

```
tc-nexus5k01(config)# int eth2/1-2
tc-nexus5k01(config-if-range)# channel-group 51 mode passive
```

The PortChannel on the Cisco Nexus 5000 Series access switches goes up. This indicates that the LACP protocol negotiation is functioning between the upstream vPC system and the Cisco Nexus 5000 Series:

```
tc-nexus5k01# show port-channel summary
Flags:  D - Down        P - Up in port-channel (members)
        I - Individual  H - Hot-standby (LACP only)
        s - Suspended   r - Module-removed
        S - Switched    R - Routed
        U - Up (port-channel)
--------------------------------------------------------------------------------
Group Port-       Type     Protocol  Member Ports
      Channel
--------------------------------------------------------------------------------
51    Po51(SU)    Eth      LACP      Eth2/1(P)    Eth2/2(P)
```

The PortChannel on the Cisco Nexus 7000 Series also goes up because of the LACP negotiation:

```
tc-nexus7k01-vdc2# show vpc br
[…]
vPC Peer-link status
---------------------------------------------------------------------
id   Port   Status Active vLANs
--   ----   ------ -----------------------------------------------------
1    Po10   up     10-14,21-24,50,60

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vLANs
--   ----   ------ ----------- -------------------------- ------------
51   Po51   up     success     success                    10-14,21-24
                                                         ,50,60
```

If the PortChannel ports are suspended, there must be a mismatch between the PortChannel ports between the switches that are supposed to bring up the PortChannel. As an example, a vPC on the Cisco Nexus 7000 Series is configured with ports that individually connect to two different PortChannels on the Cisco Nexus 5000 Series.

Alternatively, if the access layer ports are not configured for a channel, the Cisco Nexus 7000 and 5000 Series will operate normally with spanning tree. If the ports on the Cisco Nexus 5000 Series are configured in passive channel-group mode and the Cisco Nexus 7000 Series ports are not configured for port channeling, the Cisco Nexus 7000 and 5000 Series run spanning tree once more on those ports.

vPC Considerations Specific to the Cisco Nexus 7000 Series

10 Gigabit Ethernet Card Considerations and Tracking

It is possible to use a single 10 Gigabit Ethernet card on the Cisco Nexus 7000 Series for both core connectivity and the peer link, but it is not the most desirable option. If you lose the 10 Gigabit Ethernet card on the vPC primary, you lose not only core connectivity, but also the peer link. As a result, ports will be shut down on the peer vPC device, isolating the servers completely.

It is possible to address this specific configuration requirement with a tracking configuration. The objects being tracked are the uplinks to the core and the peer link. If these links are lost, vPCs local to the switch are brought down so that traffic can continue on the vPC peer.

To configure this feature, use the following command syntax:

```
! Track the vpc peer link
track 1 interface port-channel110 line-protocol

! Track the uplinks to the core
track 2 interface Ethernet7/9 line-protocol
track 3 interface Ethernet7/10 line-protocol

! Combine all tracked objects into one.
! “OR” means if ALL object are down, this object will go down
! --> we have lost all connectivity to the core and the peer link

track 10 list boolean OR
  object 1
  object 2
  object 3

! If object 10 goes down on the primary vPC peer,
! system will switch over to other vPC peer and disable all local vPCs
vpc domain 1
  track 10
```

HSRP

The use of HSRP in the context of vPC does not require any special configuration. With vPC, only the active HSRP interface answers ARP requests, but both HSRP interfaces (active and standby) can forward traffic.

If an ARP request coming from a server arrives on the secondary HSRP device, it is forwarded to the active HSRP device through the peer link.

HSRP Configuration and Best Practices for vPC

The configuration on the primary Cisco Nexus 7000 Series device looks like this:

```
interface vLAN50
  no shutdown
  ip address 10.50.0.251/24
  hsrp 50
    preempt delay minimum 180
    priority 150
    timers  1  3
    ip 10.50.0.1
```

The configuration on the secondary Cisco Nexus 7000 Series device looks as follows:

```
interface vLAN50
  no shutdown
  ip address 10.50.0.252/24
  hsrp 50
    preempt delay minimum 180
    priority 130
    timers  1  3
    ip 10.50.0.1
```

The most significant difference between the HSRP implementation of a non-vPC configuration compared with a vPC configuration is that the HSRP MAC addresses of a vPC configuration are programmed with the **G** (gateway) flag on both systems, compared with a non-vPC configuration where only the active HSRP interface can program the MAC address with the **G** flag.

Given this fact, routable traffic can be forwarded by both the vPC primary device (where HSRP is primary) and the vPC secondary device (where HSRP is secondary), with no need to send this traffic to the HSRP primary device.

Without this flag, traffic sent to the MAC address would not be routed.

The following CLI captures show the MAC address table programming on the vPC peer that is HSRP active for a given VLAN and the vPC peer that is HSRP standby for that same VLAN.

vPC HSRP on active:

```
G     -    0000.0c07.ac01    static
```

vPC HSRP on standby:

```
G     -    0000.0c07.ac01    static
```

In a non-vPC environment, the HSRP MAC looks as follows:

```
On Active: G     -    0000.0c07.ac01    static
On Standby: *    -    0000.0c07.ac01    static
```

Layer 3 Link Between vPC Peers

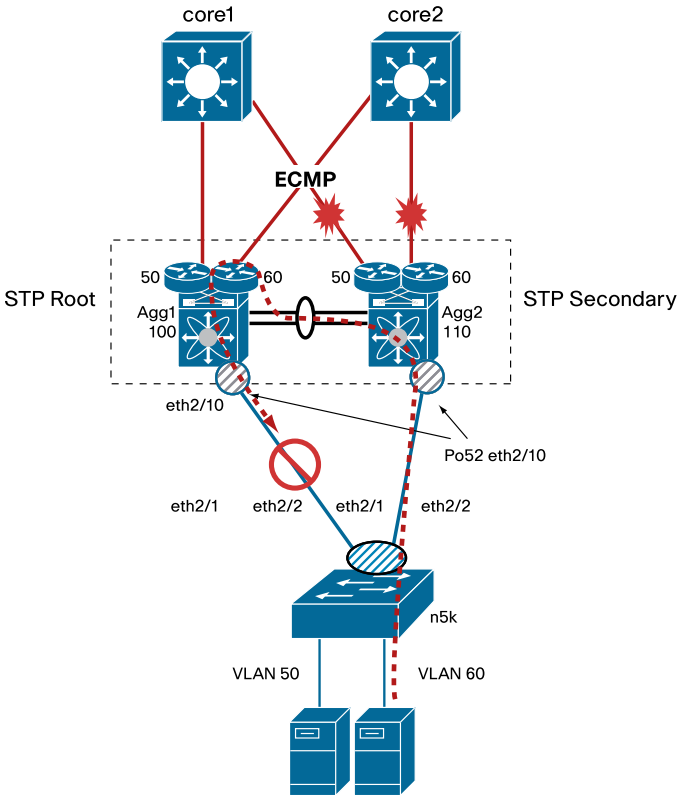
In vPC designs, you should make sure to include a Layer 3 link or VLAN between the Cisco Nexus 7000 Series systems so that the routing areas are adjacent. Also, you may consider HSRP tracking in non-vPC designs, but not in vPC designs.

HSRP tracking is not recommended for the reasons illustrated in Figure 13. Imagine that traffic from n5k on VLAN60 needs to be routed to n5k on VLAN 50. As a result of core links failure, HSRP tracking shuts down the SVI 60 on

Agg2 and forces the VLAN60-to-VLAN50 traffic to Agg1. Agg1 routes from SVI60 to SVI50 and then forwards to Po51 in order to reach n5k. vPC prevents this forwarding behavior as previously explained.

Because of this, it is recommended to create a Layer 3 path on the peer link between the routing engines on Agg2 and Agg1 instead of using HSRP tracking.

Figure 15. HSRP Tracking Is Not Needed nor Suitable for vPC Designs



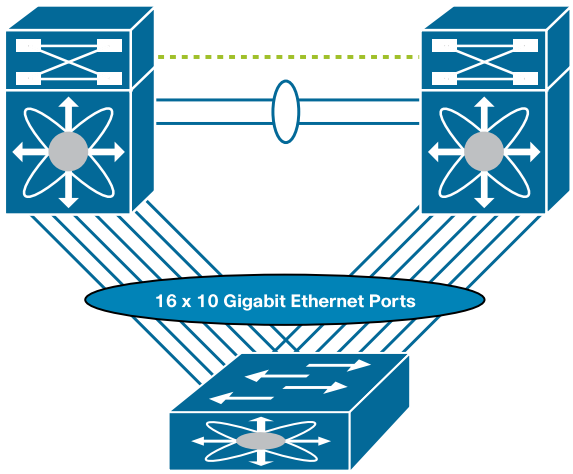
The following illustrates how to create a Layer 3 "link" to connect the Aggregation Layer switches in order to reroute the traffic to Agg1 if the routed uplinks of Agg2 go down.

```
tc-nexus7k01-vdc2(config)# vLAN 3
tc-nexus7k01-vdc2(config-vLAN)# name l3_vLAN
tc-nexus7k01-vdc2(config-vLAN)# exit
tc-nexus7k02-vdc2(config)# int vLAN 3
tc-nexus7k02-vdc2(config-if)# ip address 10.3.0.2 255.255.255.252
tc-nexus7k02-vdc2(config-if)# ip router ospf 1 area 0.0.0.0
tc-nexus7k02-vdc2(config-if)# no shut

tc-nexus7k01-vdc2(config)# int Port channel 10
tc-nexus7k01-vdc2(config-if)# switchport trunk allowed vLAN add 3

tc-nexus7k01-vdc2# show ip ospf neigh
 OSPF Process ID 1 VRF default
 Total number of neighbors: 3
 Neighbor ID     Pri State            Up Time  Address         Interface
 128.0.0.3         1 FULL/DR          01:03:05 10.51.35.126    vLAN10
```

160-Gbps vPC Between the Cisco Nexus 5000 and Cisco Nexus 7000 Series

The Cisco Nexus 5000 Series hardware supports 16-port PortChannels. The Cisco Nexus 7000 Series supports 16-port virtual PortChannels with 8 ports per Cisco Nexus 7000 Series. This allows the implementation of the topology shown in Figure 16.

Figure 16. Illustrates the Configuration of a 16 Times 10 Gigabit Ethernet PortChannel from the Cisco Nexus 5000 to the Cisco Nexus 7000



vPC Considerations for a Cisco Nexus 2000 Series Fabric Extender Dual-Attached to a Cisco Nexus 5000 Series Switch

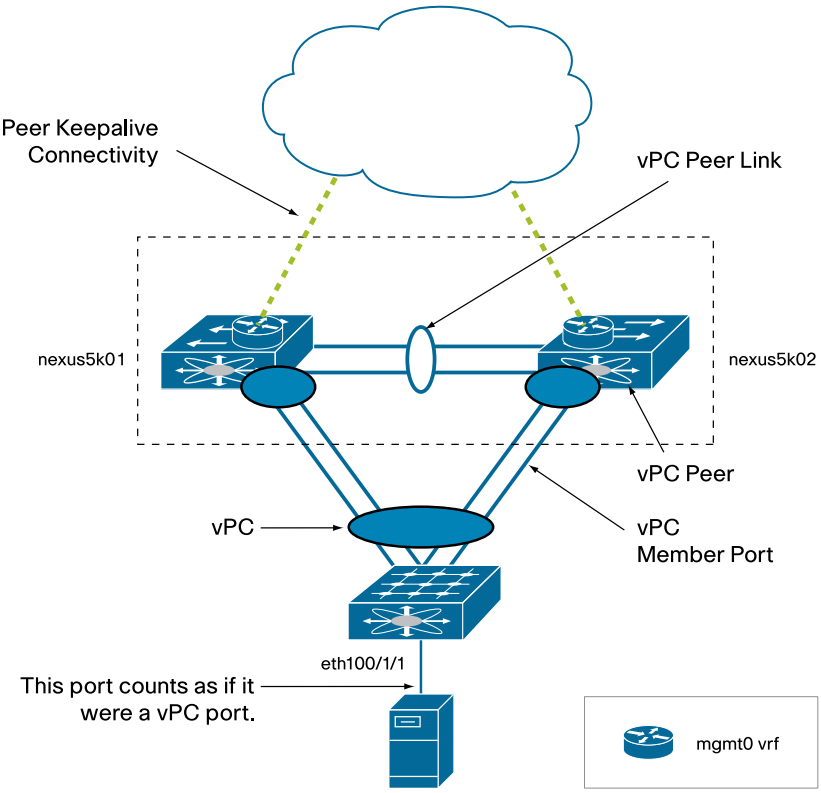
Starting from Cisco NX-OS Software Release 4.1(3), it is possible to connect a fabric extender to two Cisco Nexus 5000 Series devices configured for virtual PortChannels. The fabric extender is a satellite switch that depends on the Cisco Nexus 5000 Series for both configurations and forwarding. As a result, this vPC configuration is fundamentally different from all others, starting with the fact that the fabric extender line card must now control plane devices that have equal rights to configure the fabric extender switch ports.

The other significant difference is that, because the fabric extender is not an independent switching entity, no LACP protocol is configured on the links connecting the fabric extender to the Cisco Nexus 5000 Series Switches. The bundling of fabric extender ports is handled with internal protocols.

Finally, in this kind of configuration, each individual port on the fabric extender is treated internally as if it was a vPC port.

Figure 17 illustrates the configuration elements of such a topology. The Cisco Nexus 5000 Series devices are configured like regular vPC topologies with a peer link, a peer-keepalive link, and so on.

Figure 17. Fabric Extender Active-Active Design



The main difference is that the vPC configuration is associated with the 10 Gigabit Ethernet ports connecting to the fabric extender (switch port mode fabric extender fabric). As a result, port 100/1/1 in Figure 17 shows both in the nexus5k01 and nexus5k02 configurations.

With this topology, PortChannels are not supported across fabric extenders, while regular active-standby NIC teaming or active-active transmit load balancing from servers to fabric extenders is still supported.

The failure scenarios previously described for vPC member ports apply equally to the fabric extender ports. If the peer link is lost, the vPC secondary device shuts down the fabric ports that are connected to the secondary Cisco Nexus 5000 Series device.

Dual-active failures don't require any change in the forwarding topology.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

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

Printed in USA C07-572835-00 01/10