Chapter 5

Data Center Aggregation Layer Design and Configuration with Cisco Nexus Switches and Virtual PortChannels

Contents

Introduction **3**

Virtual Device Contexts **3**

VDC Versus VLANs 3
VDC Context Types 4
VDC Port Allocation 4
VDC Best Practices 5
Choosing the High-Availability Policy: Restart or Switchover 6
VDC Isolation and Management Network 6

Virtual PortChannel Design **7**

vPC Topology Choice 7
More on Double-Sided vPC Topologies 8
More on domain-id with Double-Sided vPC Topologies 10
Data Center-to-Data Center Connectivity 10
MST Configuration 11
HSRP Configuration 11
Layer 2 Best Practices 12
vPC Best Practices 12
Spanning Tree Best Practices 13
vPC VLANs and non-vPC VLANs 15
Cisco Nexus 7000 Series Example 15
Layer 3 Best Practices 17

Configuration Steps **18**

Spanning Tree Configuration 18
vPC Role and Priority 18
vPC Peer Link 19
Configuration for Single 10 Gigabit Ethernet Card 19
CFSoE 21
vPC Peer-Keepalive 21
vPC Ports 21
Orphaned Ports with non-vPC VLANs 23
HSRP 23
HSRP Configuration and Best Practices for vPC 23
Advertising the Subnet 24
L3 Link Between vPC Peers 24

Fine-Tuning the Design **24**

Ensuring Proper Traffic Distribution with PortChannels 24

Sample Configurations **25**

Reference Topology 25
Nexus 7k01 VDC2 Configuration 26
Nexus 7k02 VDC2 Configuration 29

Introduction

This chapter covers the design recommendations for a data center design deployment consisting of a Cisco Nexus® 7000 Series Switch at the aggregation layer and a Cisco Nexus 5000 Series Switch at the access layer. The content of this chapter focuses on the aggregation layer design with the Cisco Nexus 7000 Series.

This chapter assumes that the reader is familiar with Virtual PortChannel (vPC) technology. If some portions of this document aren't clear, we suggest that you refer to Chapter 3, "Cisco NX-OS Software Virtual PortChannel: Fundamental Concepts." For more about Spanning Tree Protocol, see Chapter 4, "Spanning Tree Design Guidelines for Cisco NX-OS Software and Virtual PortChannels."

Virtual Device Contexts

Most Nexus-based data center designs today use the concept of Cisco® virtual device context (VDC), which allows the creation of separate control-plane domains in a single switch.

From a forwarding perspective, vPC is deployed in the context of a VDC. In other words, vPC as a feature and the optimizations that are performed by vPC are fully contained within a VDC.

The Cisco virtual device context (VDC) feature allows the virtualization of a single physical device into one or more logical devices. Each of the provisioned logical devices is configured and managed as if it were a separate physical device. This logical partitioning of the device throughout control, data, and management planes observes similar fault domain isolation.

You can read more about VDCs at: http://www.cisco.com/en/US/partner/docs/switches/datacenter/sw/4_2/nx-os/virtual_device_context/configuration/guide/vdc_nx-os_cfg.html.

VDC Versus VLANs

Cisco has provided the ability to separate data plane traffic broadcast domains by using VLANs. While VLANs provide data plane security, they do not provide any segmentation at the control plane level. As an example, if a router has a routed interface in multiple VLANs (such as VLAN A and VLAN B), and it has not been secured properly, an attacker can log in to the switch virtual interface (SVI) of VLAN A (the gateway IP address on VLAN A) and start changing configurations that may compromise VLAN B as well.

VDCs, on the other hand, provide control plane isolation. If VLAN A is on context VDC 2 and VLAN B is on context VDC 3, an attacker managing to gain access to context VDC 2 through SVI A (gateway IP address for VLAN A) will not be able to perform changes that could affect VLAN B.

This property makes VDCs particularly suitable in multi-tenant environments, and, in particular, on the device that provides the default gateway function, which naturally exposes its control plane access by servers and clients. On a Layer 2 device, having VDC is less important, considering that Layer 2 devices provide VLAN switching and do not have routable interfaces in these VLANs.

VDCs are a finite resource and the Cisco Nexus 7000 Series provides a total of four VDCs.

VDC Context Types

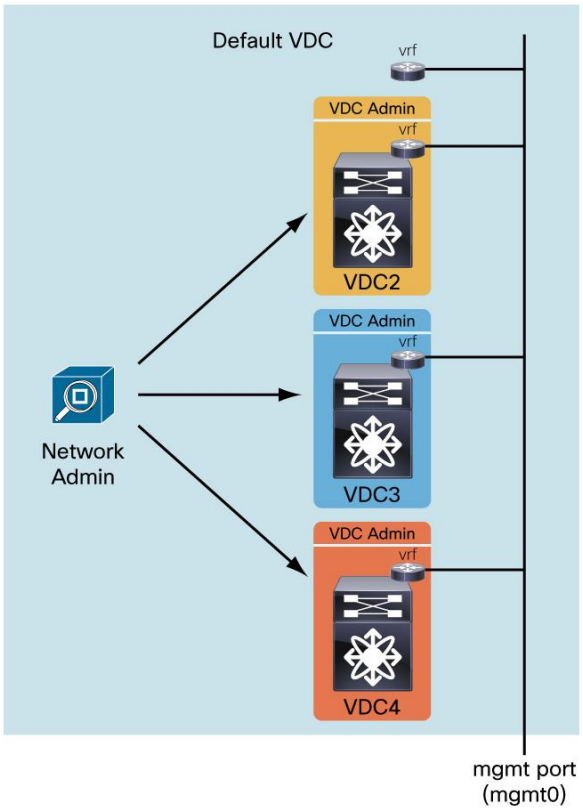
The four VDCs are not equal. The default VDC (VDC 1) is used to manage the other VDCs. The **network-administrator** on the **master or default** VDC can:

- Create and destroy other VDCs.
- Allocate resources (such as **physical interfaces**, for example) to VDCs.
- Enter other VDCs.

Users of VDCs numbered 2 through 4 do not enjoy these privileges. They cannot change resource allocations, nor can they change VDC context. The default user role for VDCs 2-4 is vdc-admin. (For more information on roles, please see Chapter 2, “Cisco NX-OS Software Command-Line Interface Primer.”)

As Figure 1 shows, each VDC is accessible through the **mgmt0** interface provided by each supervisor. The VRF management on each VDC has an IP address on the **virtual Layer 2** segment that connects the mgmt0 interface to all VDCs.

Figure 1. Hierarchy of VDCs in a Cisco Nexus 7000 and Management Access (via VRF)



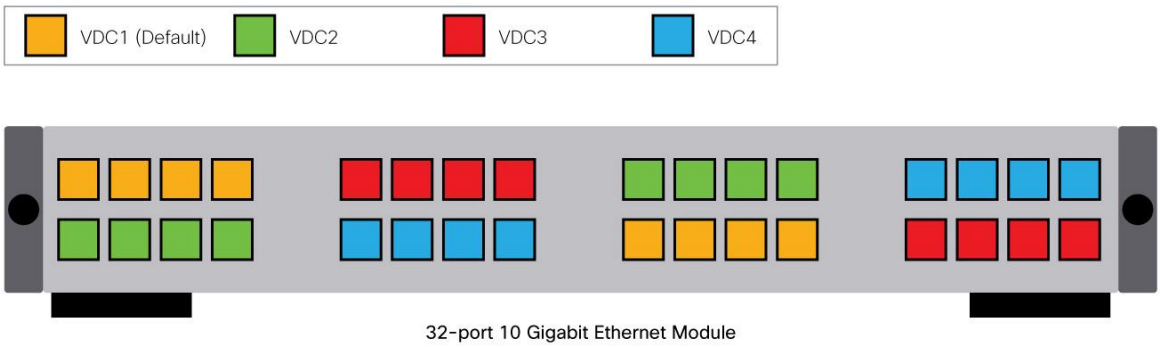
VDC Port Allocation

The physical ports of a Cisco Nexus 7000 Series Switch can be partitioned and assigned to one of the four VDCs. Each port can belong to one and only one VDC at a time.

The Cisco 10/100/1000 Gigabit Switch ports can be individually assigned to a VDC.

The 10 Gigabit Ethernet ports can be assigned to VDCs in groups of four, as shown in Figure 2.

Figure 2. Assigning 10 Gigabit Ethernet Ports to VDC in Groups of Four



VDC Best Practices

The following list summarizes some best practices for using VDCs:

- For high-security, sensitive environments, it is recommended that the default VDC be reserved as a **master VDC** and strictly used for the administration of the other VDCs. You should not run data path traffic through the master VDC unless absolutely necessary.
- In all VDC environments, it is recommended that access to the default VDC be restricted; assign to each user the fewest privileges necessary to accomplish operational tasks required by the job. For example, unless a particular user must configure global VDC parameters or provision other VDCs, that user should be assigned the **vdc-admin** role and **not** the **network-admin** role.
- If the master VDC must be used for traffic, allocate the VDCs such that the domain with the highest availability requirements, or highest priority, is carried in the master VDC. This minimizes the likelihood that operations on a lower priority or less critical VDC (for example, resource reallocation and system reload) will impact the highest-priority or more critical domain.
- If VDCs have separate administrative domains (different **vdc-admin** roles defined), you should be very careful when using AAA for authentication and authorization. Authenticating to the same authentication, authorization, and accounting (AAA) server across VDCs will implicitly apply authentication as if all VDCs are managed by a single administrative domain.
- To correctly segregate administrative domains between VDCs, and prevent an admin account from one VDC accessing another VDC, you should either use a different access control server (ACS) server per VDC, or create different admin user groups on the AAA server and limit the access of those user groups. To accomplish this, use a feature such as Network Access Restrictions (on Cisco Secure ACS), to specify AAA client IP addresses or groups. This works because each VDC sources its AAA traffic from its local management interface IP address, distinguishing VDCs from each other.
- Explicitly configure the high-availability policy of newly created VDCs to restart or bring down the system in a dual-supervisor configuration to minimize the impact of the failure within a single VDC. The default high-availability policy for VDCs in a dual-supervisor system is **switchover**. This will initiate a supervisor switchover of all VDCs if there is a failure in a single VDC.
- Review the **control plane policing (CoPP)** policy and rate limits to ensure that they are appropriate for the deployment environment. The system will apply CoPP collectively for all VDCs since there is only one logical, in-band control plane interface. Ensure that the configured limits will satisfy the requirements of all necessary control plane traffic for all active VDCs.

Choosing the High-Availability Policy: Restart or Switchover

The **ha-policy** is a configuration parameter that you can define from the default or master VDC in order to decide whether VDCs 2-4 should individually restart, or if the full supervisor should perform a stateful switchover when necessary.

The conditions that would trigger such an action are the repeated crash of a process that cannot be recovered by simply restarting the process (that is, the process cannot be restarted statefully or statelessly). In other words, upon process crash, Cisco NX-OS Software first attempts a stateful process restart. If this process crashes again soon thereafter, the process is restarted statelessly. If this measure is still not sufficient for the process to function properly, the ha-policy decides whether Cisco NX-OS Software should restart the VDC (without affecting the other VDCs) or perform a stateful supervisor switchover. Performing a switchover affects all VDCs, but because it's stateful, it causes no traffic disruption or Spanning Tree Protocol or routing reconvergence.

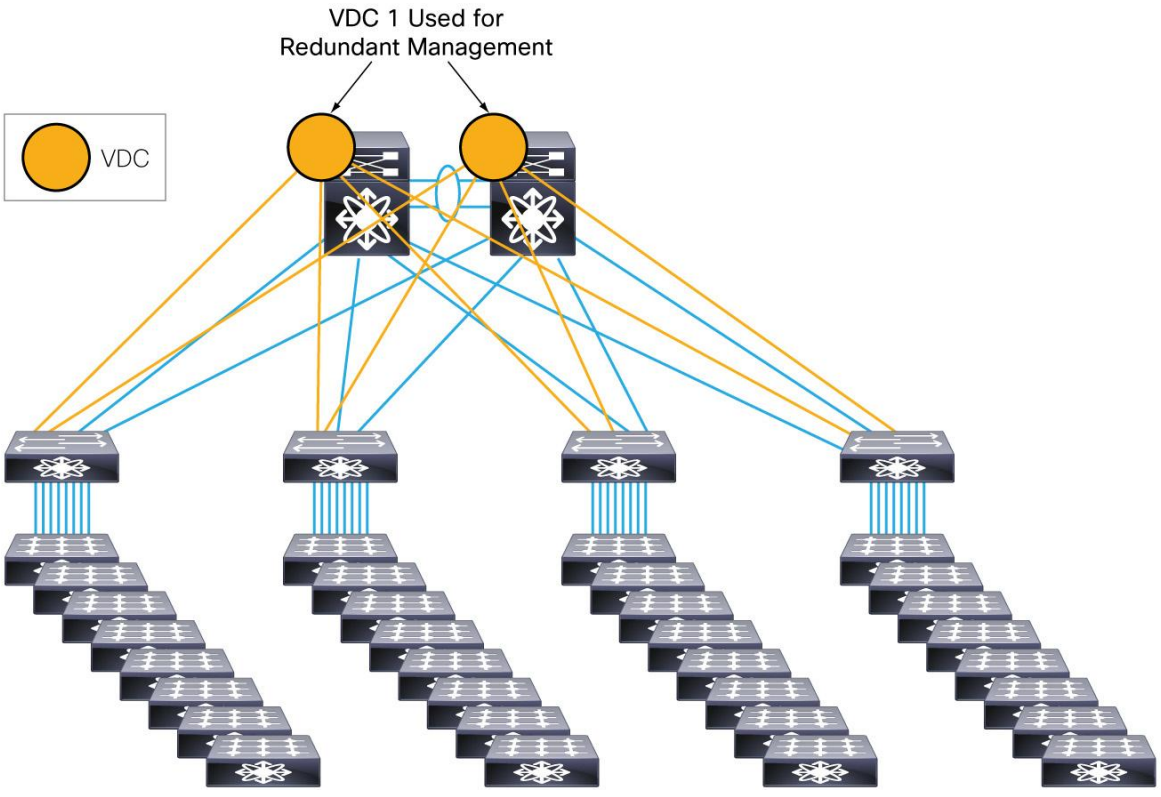
VDC Isolation and Management Network

VDCs have many uses. One of them is that you can use the default or master VDC to create a fully redundant management network.

The default or master VDC has higher privileges than all the others, and it can forward traffic just as the other VDCs do. In a multi-tenant environment, as a network administrator (role **network-admin**), you may want to use the default VDC for management of all network devices in the data center.

To accomplish this, you can implement a fully redundant management topology, which uses the default-admin VDC from each Cisco Nexus 7000 Series in the aggregation layer, as shown in Figure 3.

Figure 3. An Example of Use of VDC for Management



Virtual PortChannel Design

The Cisco Nexus 7000 Series supports 768 hardware PortChannels (usage can be verified with the command **show port-channel capacity**). Release notes indicate how many port channels are supported based on control-plane capacity - that is, depending on how many VLANs are used, how many SVIs, and so on. These numbers increase as more software optimizations are introduced in new releases.

The Cisco Nexus 7000 Series supports 8-port PortChannels per chassis, which allows a vPC design where you can have a 16-port PortChannel from the downstream Cisco Nexus 5000 Series device split into two times 8-port PortChannels on the Cisco Nexus 7000 Series.

vPC Topology Choice

When using the Cisco Nexus 7000 Series in vPC mode, several topologies become possible, including the following:

- Using dual-connected access switches, for example, the Cisco Nexus 5000 Series dual- connected to the Cisco Nexus 7000 Series with PortChannels. (see Design 1 in Figure 4). With this topology, you can create 16-port PortChannels from each Cisco Nexus 5000 Series device toward the Cisco Nexus 7000 Series. If using FEX in this topology it should be used in straight-through mode (with no support for host vPC).
- Using Cisco Nexus 5000 Series pairs in vPC mode (refer to chapter 6 for more information) with a unique PortChannel configured between the Cisco Nexus 5000 Series pair and the Cisco Nexus 7000 Series pair (see Design 2 in Figure 4). This topology is often referred to as “Double-sided vPC”. With this topology you can create a unique 16-port PortChannel from the Cisco Nexus 5000 Series pair connecting to the Cisco Nexus 7000 Series Switches. This topology can support both fabric extender dual-homed (active-active mode) and fabric extender single-homed (straight-through mode) with or without Host vPC support.

Figure 4. Cisco Nexus 5000 and Cisco Nexus 7000 Connectivity Options

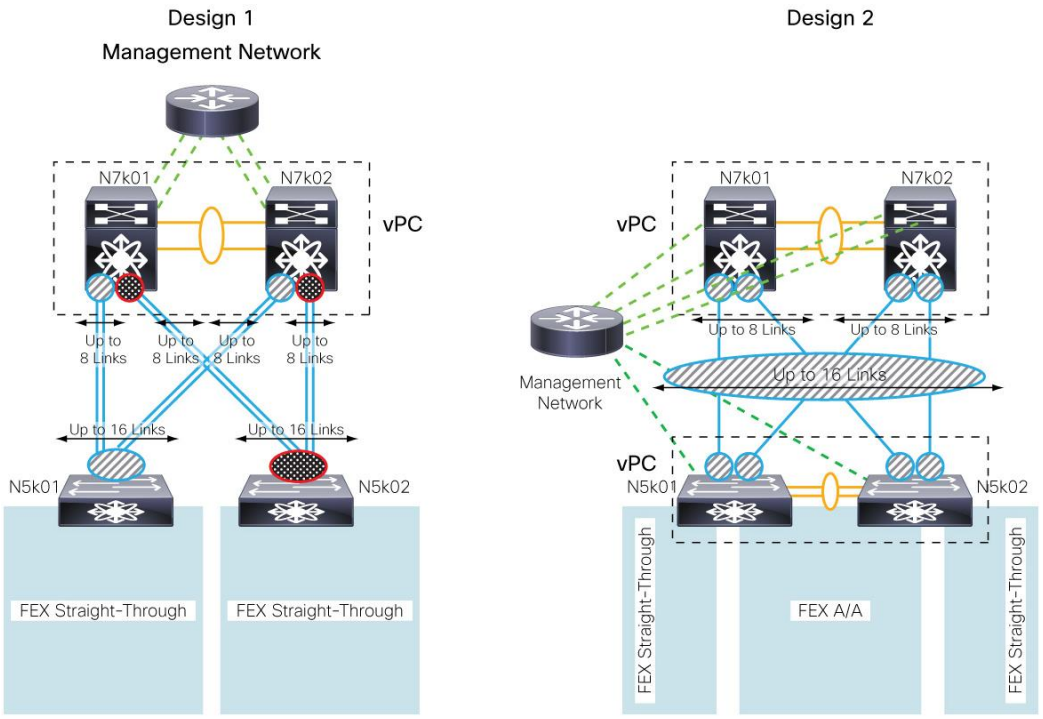


Figure 5 shows a variation of Design 2 in Figure 4.

Figure 5. Variation of the Double-Sided vPC Topology of Figure 4 Design 2

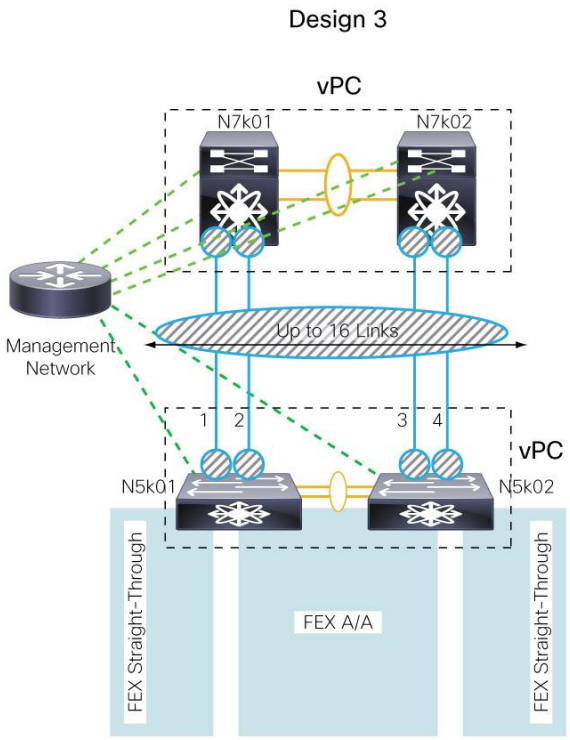


Table 1 summarizes the design issues depicted in topologies in Figures 4 and 5.

Table 1. vPC Topology Choices

	All Links Forwarding	Support for Fabric Extender Dual-Homed	Support for Host Port-Channeling on Fabric Extender
Design 1	Yes	No	No
Designs 2 and 3	Yes	Yes	Yes

More on Double-Sided vPC Topologies

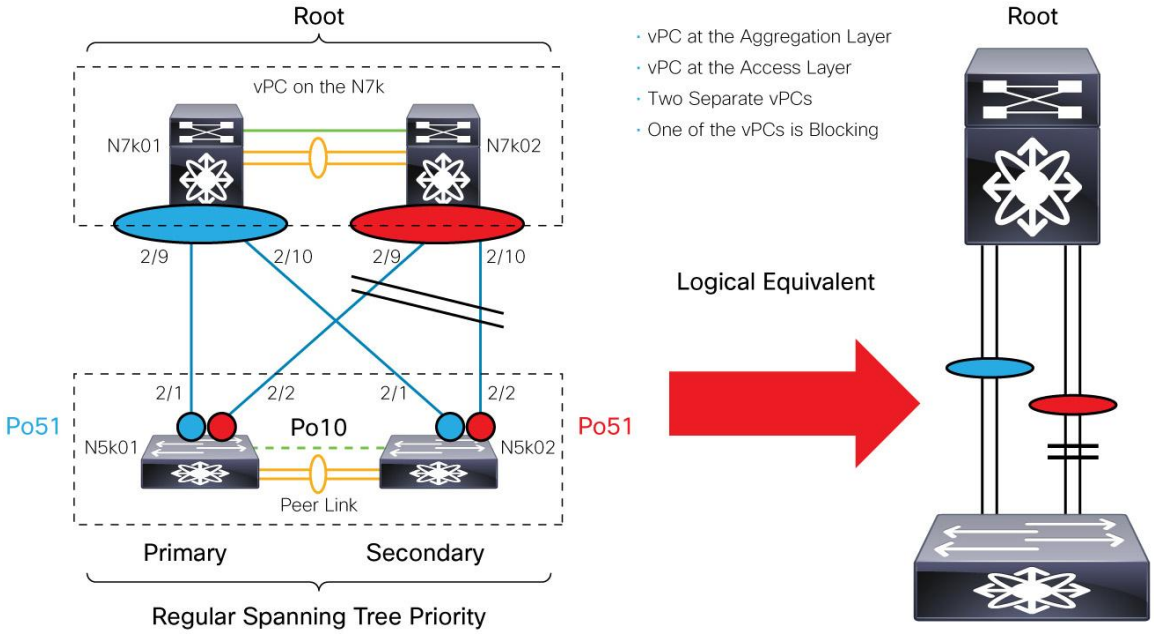
A Cisco Nexus 5000 Series running in vPC mode that connects to a Cisco Nexus 7000 Series operating in vPC mode is called a **double-sided vPC topology** because PortChannels connect two vPC systems.

Consider the diagram in Figure 6, in which the machines are not yet operating in double-sided vPC mode. In this topology, the Cisco Nexus 7000 Series Switches are configured for vPC mode. Cisco Nexus 5000 Series devices are also configured in vPC mode to support host PortChannels.

Connectivity between the Cisco Nexus 5000 Series and Cisco Nexus 7000 Series consists of two separate PortChannels, Po51 and Po52, which you would expect to find in forwarding mode. In reality, either Po51 or Po52 will be blocking.

The reason is that when you configure the Cisco Nexus switches in vPC mode, the peer link is never blocking. So even if Po51 and Po52 are **not** configured for vPC, the VLANs they carry are on the peer link, which, by definition, is always forwarding.

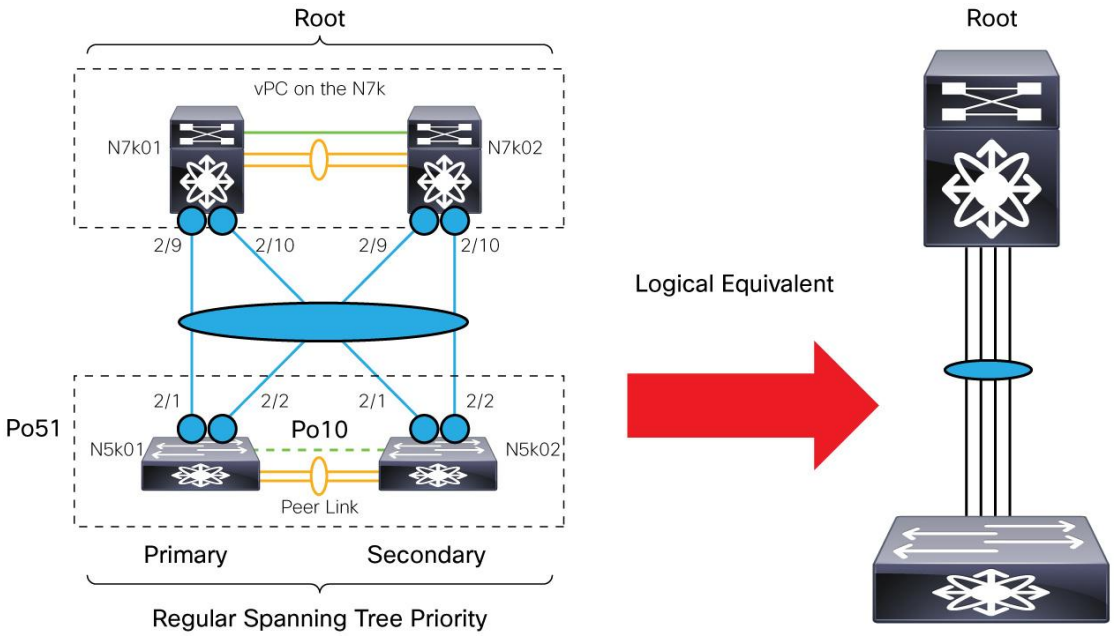
Figure 6. Example of Suboptimal Configuration Where One vPC "Link" is Blocking



To make all the links between the Cisco Nexus 5000 Series and the Cisco Nexus 7000 Series forwarding, you have these options:

- Clear the VLANs used by Po51 and Po52 from the peer link. This option may not be viable if these VLANs are also used to create vPCs to the Cisco Nexus 5000 Series device from the fabric extender or servers.
- Make sure that the ports 2/1-2 on n5k01 and 2/1-2 on n5k2 are part of the same PortChannels as depicted in Figure 7.

Figure 7. Example of Correct vPC Configuration for Double-Sided vPC



More on domain-id with Double-Sided vPC Topologies

As described in Chapter 3 “System ID in a vPC System”, the domain-id defined for the vPC domain is used to generate the system-id of the system comprised of the vPC peers. Because of this it is important to ensure that each vPC “system” in a topology such as the one illustrated in Figure 7 utilizes a different domain-id number. This ensures uniqueness in the Bridge ID used for BPDUs and the LAGID used by LACP. Alternatively, as described in Chapter 3, it is possible to define the same domain-id as long as the user configures manually a different system-id which uniquely identifies the vPC domain.

Data Center-to-Data Center Connectivity

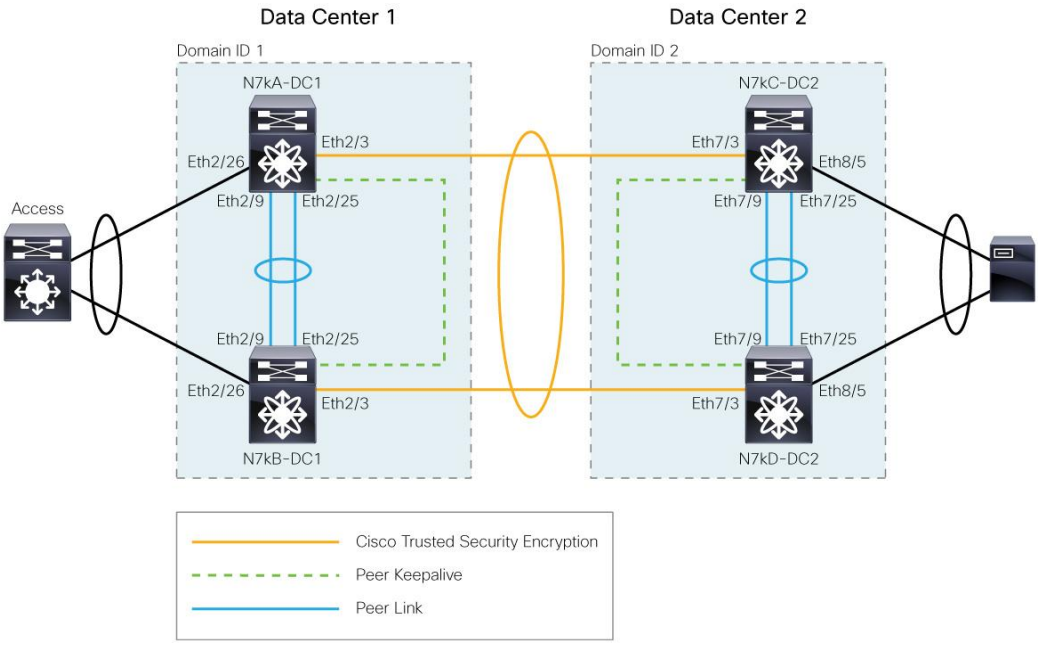
The multichassis EtherChannel configuration that is often required to connect distant data center sites consists of four systems: two switching devices per site, connected with a multichassis EtherChannel as depicted in Figure 8.

In Figure 8, there are two data centers, DC1 and DC2. Each site is comprised of two Cisco Nexus 7000 Series Switches that have a local peer link and a local peer-keepalive link. The two data centers are connected with two fibers. One fiber connects N7kA-DC1 to N7kC-DC2 and one fiber connects N7kB-DC1 to N7kC-DC2. An EtherChannel ensures that both links are forwarding and that DC1 and DC2 have Layer 2 connectivity. Incidentally, the links are secured with Cisco Trusted Security, so what traverses the public space is encrypted at wire speed at Layer 2.

As Figure 8 indicates, not only are back-to-back PortChannels supported, but also dual layers of PortChannels, in that each Cisco Nexus 7000 Series system provides vPC toward the other data center and vPC from the access layer.

When configuring the vPC domain IDs, make sure they differ on data center 1 and data center 2.

Figure 8. Layer 2 Extension Between Data Centers with vPC



MST Configuration

For data-center-to-data-center connectivity, it is advisable that MST be used as the Spanning Tree Protocol and that two MST regions be defined, each one with its own regional root, as follows (the device names refer to Figure 8):

- **Data Center 1:** MST region 1, with N7kA as the regional root, 7kB as the regional secondary root
- **Data Center 2:** MST region 2, with N7kC as the regional root, N7kD as the regional secondary root

HSRP Configuration

The routing configuration for the extended VLANs can be accomplished with two HSRP groups that have four different priorities for each group. Basically, this means that local servers use the local HSRP gateway, and if both gateways of a given site fail, the remote site can take over the processing.

Figure 9 illustrates the concept. HSRP group 1 is active and standby in DC1 and listen mode in DC2. HSRP group 2 instead is active and standby in DC2 and in listen mode in DC1.

The programming of the HSRP MAC is as follows:

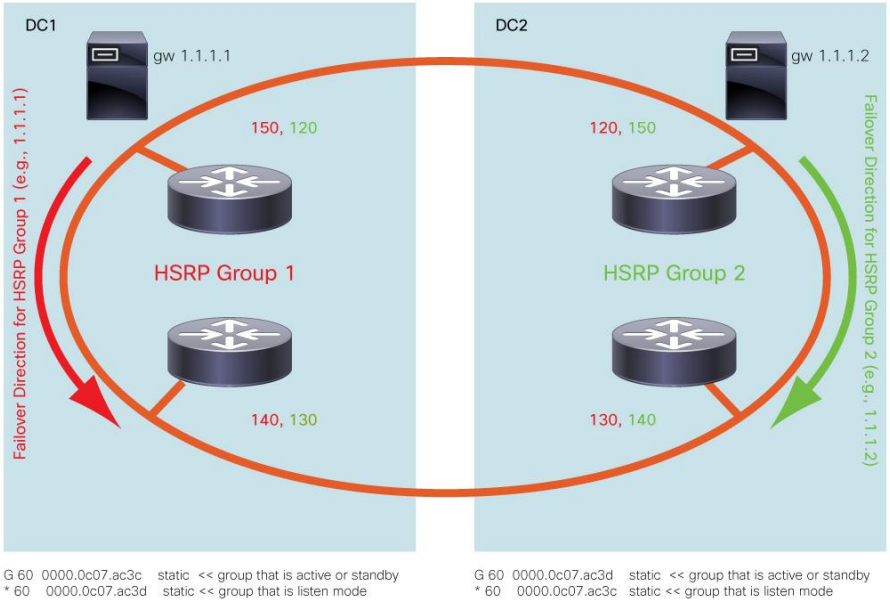
```
G 60       0000.0c07.ac3c   static  -   False  False sup-eth1(R) << group that is
active or standby

* 60       0000.0c07.ac3d    static -   False  False Po30 << group that is listen
mode (i.e. active or stdby in the remote DC)
```

If traffic from a server in DC1 is directed to the gateway 1.1.1.1 (HSRP group 1) it gets routed locally. If a server (a VM as an example) from DC1 moves to DC2 and sends traffic with a destination MAC of HSRP group 1, DC2 bridges it to DC1 where HSRP group 1 is active for this traffic to be routed.

If a server in DC2 uses the gateway 1.1.1.2 (HSRP group 2), the traffic directed to the HSRP group 2 is routed locally in DC2. Similarly if a VM from DC2 moves to DC1 and sends traffic to HSRP group 2 (1.1.1.2), the traffic is bridged from DC1 to DC2 where HSRP group 2 is active for this traffic to be routed.

Figure 9. HSRP Groups and Priorities in an Extended Data Center



Layer 2 Best Practices

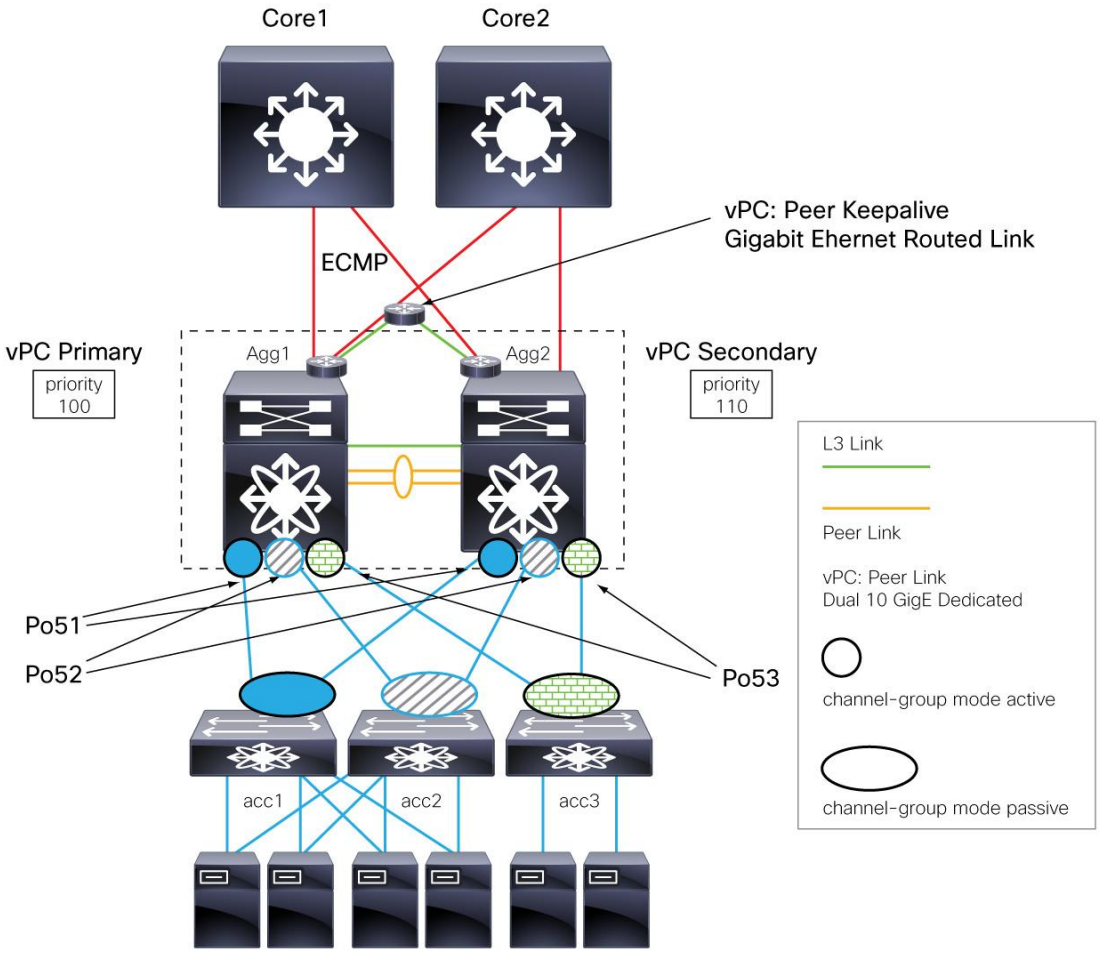
Remember that vPCs are switch ports, not Layer 3 ports, and if they are not configured as such, the vPC doesn't come up.

vPC Best Practices

The following list summarizes the best practice recommendations for vPC configurations. As a reference and summary for vPC configurations, Figure 10 highlights the components of a vPC design.

- Connect the two Cisco Nexus 7000 Series Switches through redundant 10 Gigabit Ethernet links (operated in dedicated mode) from 10 Gigabit Ethernet line cards for the purpose of forming the peer link between vPC peers. Preferably this link carries only vPC VLANs.
- A single 10 Gigabit Ethernet card providing both the Layer 3 core links and vPC peer links is not recommended, as the card failure on a primary vPC Cisco Nexus 7000 Series would disconnect it from core connectivity and also create a failure scenario where the primary device would keep the vPC member ports up, thus blackholing traffic. If you deploy such a topology, you should use the object-tracking capabilities of Cisco NX-OS Software Release 4.2 and track the core links under the vPC role configuration.
- An additional Layer 2 link is allocated to carry non-vPC VLANs between the vPC members. In a Rapid PVST+ deployment, you can trunk the non-vPC VLANs on a separate PortChannel connecting the vPC peers. If vPC and non-vPC VLANs share the same link, you should consider using **dual-active exclude interface-vlan <non-vPC vlans list>** to decouple the SVI status from the peer-link failure.
- Configure the spanning tree root and secondary root priorities as usual. The adjacent switches will see the root switch bridge ID (regardless of whether the root is the primary or secondary vPC). Matching primary root and vPC primary is recommended.
- A Layer 3 VLAN connecting the Cisco Nexus 7000 Series routing engines makes the Open Shortest Path First (OSPF) Protocol area contiguous and does not require HSRP tracking.
- The peer keepalive traffic should never be carried as a VLAN over the peer link.
- Use a routed Layer 3 connection between the Agg1 and Agg2 for the peer keepalive, in order to resolve dual-active scenarios.
- mgmt0 can be used if you route the peer keepalive through the out-of-band management network, in which case each Cisco Nexus 7000 Series is connected to the management network through both the mgmt0 of supervisor slot 5 and supervisor slot 6. If you follow this approach regardless of which supervisor is active, the Cisco Nexus 7000 Series will have one of the mgmt0 interfaces connected to the management network, which can then be used for peer-keepalive purposes.
- Direct connectivity of the peer keepalive through mgmt0 from one vPC peer to the other should never be utilized.
- If direct connectivity for the peer keepalive is required between vPC peers, you should use a dedicated Gigabit Ethernet port from one of the line cards.
- Configure Layer 2 links from acc1, acc2, acc3 as Layer 2 EtherChannels.
- Port channels on the Cisco Nexus 7000 Series side are configured for LACP active mode.
- If the access switch is a Cisco Catalyst platform, you may have to disable the EtherChannel misconfiguration guard (unless you are using a Cisco NX-OS release higher than 4.2(X) on the Cisco Nexus 7000).

Figure 10. Summary vPC Best Practices



Spanning Tree Best Practices

This section summarizes the best practices for spanning-tree configurations in the presence of vPCs. For more details, see “Spanning Tree Design Guidelines for Cisco NX-OS Software and Virtual PortChannels,” Chapter 4 in this guide.

When configuring spanning tree with a Cisco Nexus 7000 Series vPC deployment, remember that for vPC member ports only, the vPC operational primary device generates and processes BPDUs. Also remember that the vPC peer link is always forwarding, no matter which connectivity you put in place.

Spanning-tree best practices for a vPC configuration are as follows:

- Choose the spanning-tree algorithm, keeping in mind that Multiple Spanning Tree (MST) scales better, but that Rapid Per-VLAN Spanning Tree Plus (Rapid PVST+) is easier to deploy.
- Verify the VLAN range that is used in the topology and make sure that all devices are configured for the VLAN range that is common to all platforms. Cisco NX-OS platforms support this VLAN range: **1-3967,4048-4093**.
- Be mindful of non-vPC ports (orphaned ports) and consider the topology behavior for these ports when the peer link is lost.

- Be sure to preprovision MST. Create the region configuration for the deployment as well as all the VLAN mappings. VLAN creation and assignment to trunk links and host ports can be performed after the deployment with no disruption. Region modifications should be limited to the deployment time to reduce the need for topology recalculations and the need to deal with Type-1 misconfigurations.
- At the aggregation layer, create a root or a secondary root device as usual. Matching root and primary vPC switch is preferred.
- Make sure **pathcost method long** is enabled.

Prior to configuring vPCs, the spanning-tree configuration on the access layer appears with Nexus7k01 (agg1) as the root switch on port Ethernet2/1 and with Nexus7k02 (agg2) as the secondary root switch on port Ethernet2/2. The following output illustrates the connectivity from the Cisco Nexus 5000 Series to the Cisco Nexus 7000 Series and the “show spanning-tree” command illustrates the spanning-tree topology.

```
tc-nexus5k01# show cdp neigh
Device-ID             Local Intrfce Hldtme Capability  Platform      Port ID
tc-nexus7k01-vdc2(TBM12162254)Eth2/1     133    R S I s   N7K-C7010     Eth2/9
tc-nexus7k02-vdc2(TBM12193229)Eth2/2     175    R S I s   N7K-C7010     Eth2/9

tc-nexus5k01# show spanning-tree vlan 50

VLAN0050
  Spanning tree enabled protocol rstp
  Root ID    Priority    24626
             Address     001b.54c2.80c2
             Cost        2000
  Bridge ID  Priority    32818  (priority 32768 sys-id-ext 50)
             Address     000d.eca3.477c
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Eth2/1           Root FWD 2000      128.257  Network P2p
Eth2/2           Altn BLK 2000      128.258  Network P2p
```

The direct path to the root has a cost of 2000, while the alternate path (indicated as blocking) offers a cost of 3000 which indicates a path made of two hops: first a 10 Gigabit Ethernet link and then a two times 10 Gigabit Ethernet link PortChannel.

After configuring the vPCs, the two Cisco Nexus 7000 Series Switches appear as a single switching device from a spanning-tree perspective (remember that the vPC primary is the Cisco Nexus 7000 Series Switch that processes BPDUs). The following output illustrates the Spanning-Tree topology after the vPC configuration.

```
tc-nexus5k01# show spanning-tree vlan 50

VLAN0050
  Spanning tree enabled protocol rstp
  Root ID    Priority    24626
```

```
                Address     001b.54c2.80c2
                Cost        1000
                Port        4146 (port-channel51)
                Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

  Bridge ID  Priority    32818  (priority 32768 sys-id-ext 50)
             Address     000d.eca3.477c
             Hello Time  2  sec  Max Age 20 sec  Forward Delay 15 sec

Interface        Role Sts Cost      Prio.Nbr Type
---------------- ---- --- --------- -------- --------------------------------
Po51             Root FWD 1000      128.4146 Network P2p
```

vPC VLANs and non-vPC VLANs

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

In addition, it is possible to carry also the VLAN used by orphaned ports with some special considerations.

As a general best practice, the VLANs you use for vPC-connected devices should be different from those you use for single-port attached devices (orphaned ports), and you should put those VLANs (the non-vPC VLANs) on a trunk that's different from the one on which the peer link resides.

On the Cisco Nexus 7000 Series, when carrying vPC and non-vPC VLANs on the peer link, you may want to exclude the orphaned ports SVIs from the default behavior by using the command **dual-active exclude interface-vlan <non-vPC vlans list>**. Alternatively, 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.

This recommendation applies to the aggregation layer only, not to the access layer and the following example illustrates why you should consider either a separate trunk for non-vPC VLANs or the use of the **dual-active exclude interface-vlan** command.

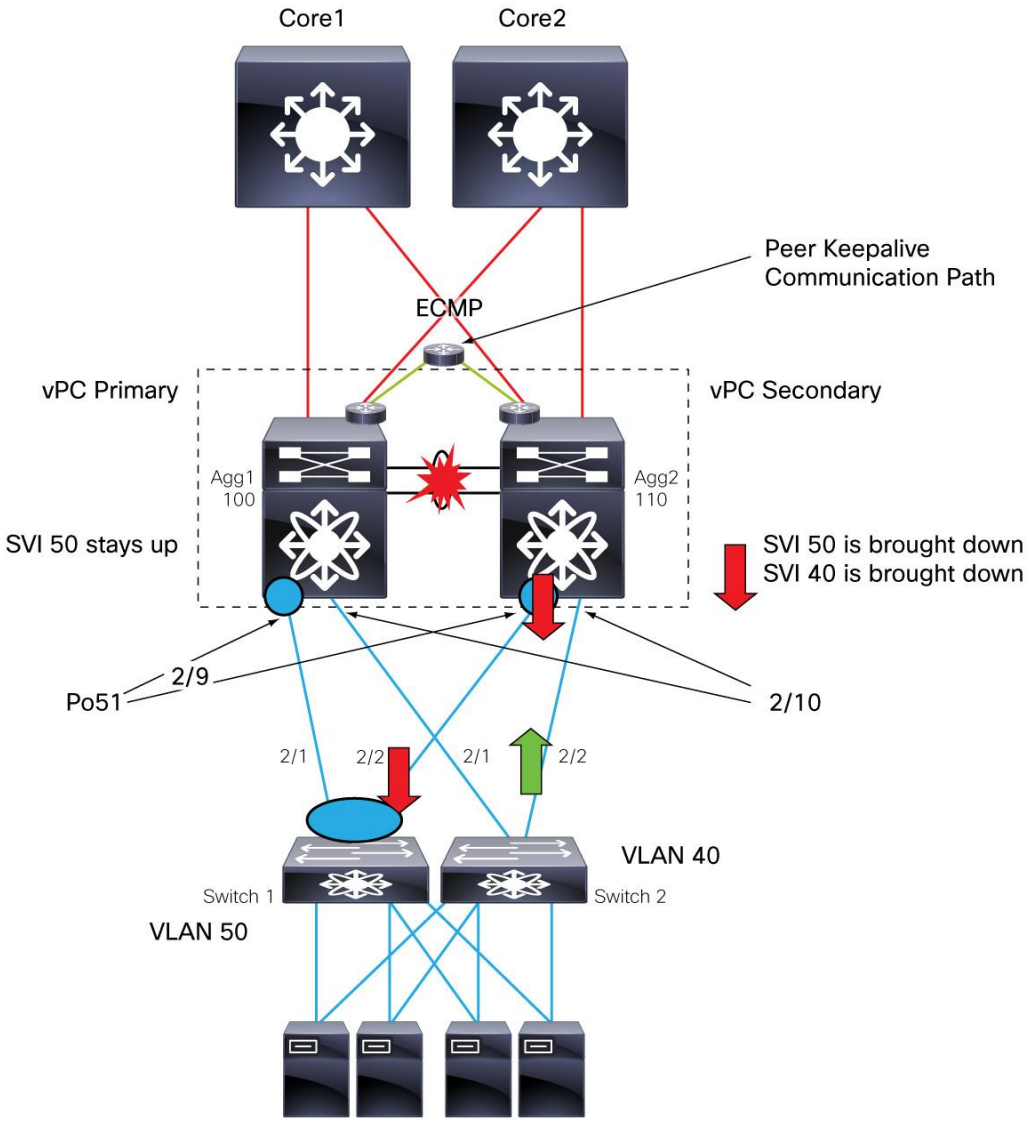
Cisco Nexus 7000 Series Example

Figure 11 illustrates what happens during vPC peer link failure for vPC and non-vPC ports. In the figure, Ethernet 2/10 is not part of a vPC; it operates as a regular spanning-tree port. Port Ethernet 2/10 carries VLAN 40. Port Ethernet2/9 is a vPC ports and it carries VLAN 50

The sequence of events is as follows:

- The vPC peer link fails (Po10).
- Ethernet 2/9 on Agg2 is brought down because it is part of vPC Po51 and belongs to the operational secondary vPC device.
- Ethernet 2/10 is not part of a vPC, so it stays up and unblocks (indicated by the green arrow in the figure).
- SVI VLAN50 (vPC-VLAN) and SVI VLAN 40 (not used for vPC, but trunked on the peer link, and as such, considered a vPC VLAN) are both brought down on the operational secondary device.

Figure 11. Peer Link Failure Example and Behavior of vPC VLANs



The secondary vPC peer brings down all interface VLANs regardless of whether they belong to a virtual PortChannel configuration or not, as long as these VLANs are trunked on the peer link, as well as all **vPC member** ports:

```
tc-nexus7k02-vdc2# show int eth2/9
Ethernet2/9 is down (vpc peerlink is down)
```

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)
```

Bringing down SVI 50 is desirable, because doing this the traffic from the core destined to VLAN 50 takes the path via the primary vPC device.

For VLAN 40, the failure of the peer link has the effect that on the secondary vPC device SVI 40 is brought down because all SVIs for vPC VLANs are brought down, the SVI 40 on primary vpc device will stay up because there is at least one active vpc (or link) carrying vlan 40.

You can modify this behavior in such a way that the SVI 40 stays up on secondary as well in two possible ways:

- Using a separate trunk between the Cisco Nexus 7000 devices for non-vPC VLANs and including vlan 40 on that trunk instead of the vpc peer-link
- Excluding the non-vPC VLANs from the vPC autostate behavior as described in this section

If the desired behavior is to keep VLAN 40 up, because it is not used on any vPC, you should just modify the vPC domain configuration as follows:

```
vpc domain 1
role priority 100
dual-active exclude interface-vlan 40
```

Layer 3 Best Practices

With Cisco NX-OS Software Release 4.2, the HSRP protocol on the Cisco Nexus 7000 Series can support upgrades to the supervisor without flapping HSRP. This is achieved by incrementing the hold-time value (extended hold-time) automatically during the upgrade. Apart from this, the user needs to configure nothing special compared to regular HSRP configurations.

The following list summarizes Layer 3 best practices for vPC configurations:

- HSRP is configured normally, ideally with primary and secondary roles matching the vPC role priorities. Preemption should be configured, again mostly to maintain the active configuration as closely as possible to that specified in the network design.
- Should a peer link failure occur, the SVIs on the operational secondary will be shut down, which enforces only one possible active topology: that is, the SVI on the primary Cisco Nexus 7000 Series Switch is going to be the active HSRP interface. If you want to remove an SVI from this behavior you should use the command **dual-active exclude interface-vlan <non-vPC vlans list>** in the vPC domain configuration.
- Avoid using static routing with HSRP tracking of core links in order to avoid the situation where routed traffic from access switch to access switch is dropped due to the duplicate prevention technique used by vPC. See Chapter 3, “L3 Link Between vPC Peers” for more information.
- Dynamic routing with L3 VLANs or links between aggregation Cisco Nexus 7000 Series devices is recommended.
- Routing from the core to the aggregation layer should use Layer 3 links with no vPCs. Traffic distribution is going to use Equal Cost Multipath (ECMP).

Configuration Steps

Spanning Tree Configuration

Regardless of whether you are using Rapid PVST+ or MST, make sure to configure spanning tree to use **pathcost method long**, as follows:

```
spanning-tree pathcost method long
```

Configuring Rapid PVST+ follows usual well-known guidelines. In addition to these guidelines, if you configure MST, you will have to follow a configuration similar to the following (and with different priorities on primary and secondary root):

```
spanning-tree mode mst
spanning-tree mst 0-1 priority 24576
spanning-tree mst configuration
  name dc1
  revision 1
instance 1 vlan 1-3967,4048-4093
```

If you are using MST, remember to configure the region mapping for all 4000 VLANs, even if you are not using them all. Doing so uses no hardware or CPU resources until the VLAN is created and assigned to interfaces. But because in MST changing the region configuration can cause a temporary glitch, you should plan the MST region at the deployment time. And you should create VLANs and decide where to trunk them whenever a new VLAN needs to be provisioned (see Chapter 4 in this design guide for more information). If you follow this approach, there is no spanning-tree reconvergence when you need to provision a new VLAN.

Remember to ensure that if you are using MST, the **region configuration** for primary and secondary vPC peers matches.

Note: If the topology includes other switches that are not based on Cisco NX-OS Software, make sure to change the MST region mappings on the adjacent switches to match the Cisco NX-OS Software range of supported VLANs: 1-3967, 4048-4093.

In order to verify that the configuration is correct from a vPC perspective, make sure to issue the following command:

```
nexus7000# show vpc consistency-paramters global
```

vPC Role and Priority

Within the VDC, the following configurations are required.

First, the vPC needs to be enabled, as follows:

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

A domain needs to be defined as well as priorities **to define primary and secondary roles** in the vPC configuration. The **lower number has higher priority**, so it wins.

Note: 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:

```
agg(config)# vpc domain 1

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

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

There are no functional issues when the spanning-tree root and vPC primary node do not match. This can only cause some suboptimal convergence time due to spanning-tree resynchronization when the peer link is flapped or a vPC device is reloaded.

Because of this, in case you want to restore the original mapping between the spanning-tree root and vPC primary device, you can follow this procedure on the **secondary, operational primary** device:

1. Enter the vPC domain configuration, **vpc domain <domain_id>** (same vPC domain you are using).
2. Reset the vPC role priority with the command **vpc role priority <priority_number>** (reentering the same priority is all right).
3. Perform a shut-no shut operation over the peer link.

Alternatively, you can create a script (which you should customize):

```
7k-1(config)# cli alias name vpcpreempt conf t ; vpc domain  <number> ; role
priority 32767 ; int po 10 ; shut ; no sh

7k-1(config)# show cli alias

CLI alias commands
==================
alias       :show cli alias
vpcpreempt  :conf t ; vpc domain  10 ; role priority 32767 ; int po 10 ; shut ;
no sh
```

vPC Peer Link

This PortChannel should be configured on dedicated-mode 10 Gigabit Ethernet interfaces across two different 10 Gigabit Ethernet line cards:

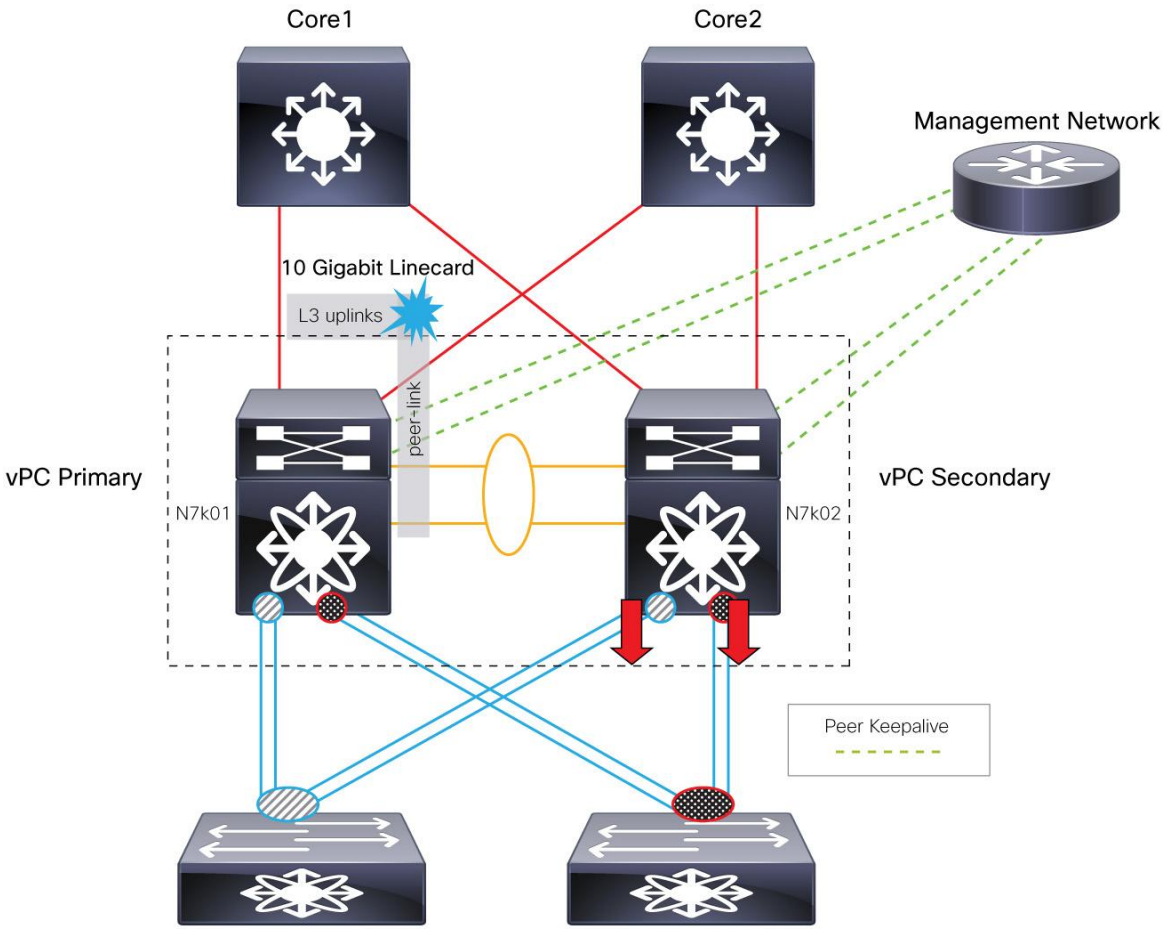
```
agg(config)# interface port-channel10
agg(config-if)#  vpc peer-link
agg(config-if)#  switchport trunk allowed vlan <all access vlans>
```

Configuration for Single 10 Gigabit Ethernet Card

Although 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 is possible, 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.

Figure 12 illustrates the issues.

Figure 12. Failure Scenario with 1 Single 10 Gigabit Ethernet Card on N7k01



In the topology shown in Figure 12, the failure of the 10 Gigabit Ethernet card that provides both peer-link connectivity and core connectivity, causes the vPC secondary to shut down the vPC member ports, so that traffic flows to the vPC primary. The vPC primary doesn't have any core connectivity, however, so traffic gets blackholed with a single failure.

The best solution is naturally to have two 10 Gigabit Ethernet line cards, but alternatively you can use the object tracking functionality.

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
```

```
! 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

! 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
```

CFSoE

Cisco Fabric Services over Ethernet (CFSoE) provides several infrastructure services for vPC, including MAC synchronization, configuration verification for potential mismatch in the configurations, and locking of the configuration while a vPC peer is being upgraded.

The CFSoE configuration does not need to be specifically enabled, but just as a reference, the configuration appears automatically when you enable vPC. It looks like this:

```
agg1(config)#cfs region 10
agg1(config-cfs-region)# vpc
agg1(config)#cfs ethernet distribute
```

vPC Peer-Keepalive

Finally, a dual-active detection configuration needs to be put in place. The keepalive link that is used to resolve dual-active scenarios can be carried over a routed infrastructure; it doesn't need to be a direct point-to-point link. The keepalives are sent every two seconds.

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
```

vPC Ports

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

```
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
```

You can verify the success of the configuration by issuing the command:

```
agg1#show vpc brief

tc-nexus7k02-vdc2# show vpc br
[…]

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason                     Active vlans
--   ----   ------ ----------- -------------------------- ------------
51   Po51   down*  failed      vPC type-1 configuration   -
                               incompatible - STP
                               interface port type
                               inconsistent
```

If the **consistency check** doesn't show **success**, it is recommended that you verify 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.

The following commands illustrate how to verify that the configuration on the vPC member ports is correct.

```
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 vPC, any further configurations, such as enabling or disabling bridge assurance or trunking mode, are performed under the interface port channel configuration mode. Trying to configure spanning-tree properties for the physical interface instead of the PortChannel will result in an error message.

Orphaned Ports with non-vPC VLANs

As previously described, when the peer link is lost, vPC shuts down the SVI on the secondary switch and, as a result, orphaned ports on the operational secondary may become isolated. For this reason you can either trunk the non-vPC VLANs on a different link, or you should remove the non-vPC VLANs from this behavior, as described next.

First, you may want to execute the following command to learn which ports are considered orphan ports from the Cisco Nexus 7000 Series perspective:

```
Nexus7000#show vpc orphan-ports
```

Second, you can remove the non-vPC VLANs in the vPC domain configuration:

```
vpc domain 1
  role priority 100
  dual-active exclude interface-vlan <non-vPC VLANs>
  peer-keepalive destination 192.168.1.2 source 192.168.1.1 vrf vpc-keepalive
```

HSRP

The use of HSRP in the context of vPC doesn't 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 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 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
```

Advertising the Subnet

The configuration is completed by including the subnet in the routing advertisements and making sure that the VLANs used for server connectivity are not used to create neighbor relationship between the aggregation layer devices. Here's how to do this:

```
interface Vlan50
  no shutdown
  ip address 10.50.0.251/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp 50
    preempt delay minimum 180
    priority 150
    timers  1  3
    ip 10.50.0.1
```

L3 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 so that the routing areas can be adjacent. You may also consider HSRP tracking in non-vPC design, but not in vPC designs.

You should, therefore, create a Layer 3 path on the peer link between the routing engine on Agg2 and Agg1 instead of using HSRP tracking:

```
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
```

You can then verify that the Cisco Nexus 7000 Series are OSPF neighbors by issuing the following command:

```
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
```

Fine-Tuning the Design

Ensuring Proper Traffic Distribution with PortChannels

As part of a vPC design, you need to make sure that traffic flows are properly distributed over all the available forwarding paths. This helps to ensure not only that all links are utilized, but also that all hardware tables in every redundant network device are programmed correctly, which, in turn, helps to prevent failures. As an example, if all

traffic takes the path through the vPC primary, the ARP table on the vPC secondary may end up being out of date despite the synchronization feature.

For this reason, proper traffic distribution is recommended both upstream and downstream of a vPC system.

The client-to-server traffic load distribution happens on the core devices, in which case you need to check the Equal Cost Multipath configuration to ensure that this is properly tuned (for example, if you want to include the Layer 4 port information in the Layer 3 hashing from the core to the Cisco Nexus 7000 Series):

```
tc-core01(config)#mls ip cef load-sharing full
```

To verify which path is taken from the core to the Cisco Nexus 7000 Series, you can use the following command:

```
tc-core01#show mls cef exact-route 10.40.0.10 10.50.0.103
```

Because server-to-client traffic load distribution happens on the Cisco Nexus 5000 Series in the access switch, you need to check the PortChannel configuration on the Cisco Nexus 5000 Series. For example, to configure source IP and source layer port-based hashing from server to the Cisco Nexus 7000 Series, use the following command:

```
port-channel load-balance ethernet source-port
```

To verify which path is taken, from the access switch to the Cisco Nexus 7000 Series, use the following command:

```
show port-channel load-balance forwarding-path
```

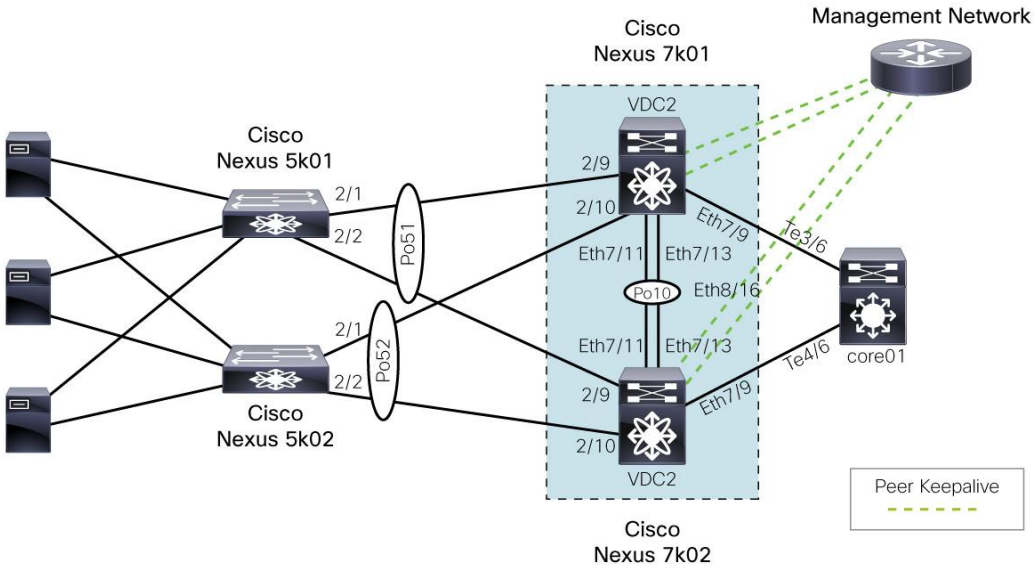
Sample Configurations

Reference Topology

The following configurations refer to the topology shown in Figure 13. The main difference between the previous recommendations and this configuration example is the use of a direct connection for the vPC peer keepalive over port Ethernet 8/16 instead of through mgmt0 routed over the management network.

As previously described, the VDC used for production differs from the default VDC. For this reason, the sample topology in Figure 13 uses VDC 2 and not VDC1, but nothing prevents you from using the same, exact configurations on the default or master VDC.

Figure 13. Sample Reference Topology



Nexus 7k01 VDC2 Configuration

```
feature ospf
feature udld
feature interface-vlan
feature hsrp
feature lacp
feature vpc

username admin role vdc-admin

vrf context management
  ip route 0.0.0.0/0 10.51.35.129

vrf context vpc-keepalive

vlan 3
  name l3_vlan
vlan 50
vlan 60

spanning-tree mode mst
spanning-tree pathcost method long
spanning-tree mst 0-1 priority 24576
spanning-tree mst configuration
  name dc1
  revision 3
  instance 1 vlan 1-3967,4048-4093

vpc domain 1
  role priority 100
  peer-keepalive destination 192.168.1.2 source 192.168.1.1 vrf vpc-keepalive


interface Vlan3
  no shutdown
  ip address 10.3.0.1/30
  ip router ospf 1 area 0.0.0.0

interface Vlan50
  no shutdown
  ip address 10.50.0.251/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp 50
```

```
    preempt delay minimum 180
    priority 150
    timers  1  3
    ip 10.50.0.1

interface Vlan60
  no shutdown
  ip address 10.60.0.254/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp 60
    preempt delay minimum 180
    priority 150
    timers  1  3
    ip 10.60.0.1

interface port-channel10
  description peer-link
  switchport
  switchport mode trunk
  vpc peer-link
  spanning-tree port type network
  spanning-tree guard loop

interface port-channel51
  switchport
  switchport mode trunk
  vpc 51
  switchport trunk allowed vlan 10-14,21-24,50-51,60

interface port-channel52
  switchport
  switchport mode trunk
  vpc 52
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60

interface Ethernet2/9
  description tc-nexus5k01 - Eth2/1
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 10-14,21-24,50-51,60
  channel-group 51 mode active
  no shutdown
```

```
interface Ethernet2/10
  description tc-nexus5k02 - Eth2/2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60
  channel-group 52 mode active
  no shutdown

interface Ethernet7/9
  description tc-core01 - TenGiga3/6
  ip address 10.1.1.1/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown

interface Ethernet7/11
  description tc-nexus7k02-vdc2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3,10-14,21-24,50,60
  channel-group 10 mode active
  no shutdown

interface Ethernet7/13
  description tc-nexus7k02-vdc2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3,10-14,21-24,50,60
  channel-group 10 mode active
  no shutdown

interface Ethernet8/16
  description peer-keepalive
  vrf member vpc-keepalive
  ip address 192.168.1.1/24
no shutdown

interface mgmt0
vrf member management
ip address <ip>

interface loopback0
ip address 128.0.0.2/24

router ospf 1
```

```
auto-cost reference-bandwidth 1000000
```

Nexus 7k02 VDC2 Configuration

```
feature ospf
feature udld
feature interface-vlan
feature hsrp
feature lacp
feature vpc

username admin role vdc-admin

vrf context management
  ip route 0.0.0.0/0 10.51.35.129

vrf context vpc-keepalive

vlan 3
  name l3_vlan
vlan 50
  name production1
vlan 60
name production2

spanning-tree mode mst
spanning-tree pathcost method long
spanning-tree mst 0-1 priority 28672
spanning-tree mst configuration
  name dc1
  revision 3
  instance 1 vlan 1-3967,4048-4093

vpc domain 1
  role priority 110
  peer-keepalive destination 192.168.1.1 source 192.168.1.2 vrf vpc-keepalive


interface Vlan3
  no shutdown
  ip address 10.3.0.2/30
  ip router ospf 1 area 0.0.0.0


interface Vlan50
  no shutdown
```

```
  ip address 10.50.0.252/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp 50
    preempt delay minimum 180
    priority 130
    timers  1  3
    ip 10.50.0.1

interface Vlan60
  no shutdown
  ip address 10.60.0.253/24
  ip ospf passive-interface
  ip router ospf 1 area 0.0.0.0
  hsrp 60
    preempt delay minimum 180
    priority 140
    timers  1  3
    ip 10.60.0.1

interface port-channel10
  description peer-link
  switchport
  switchport mode trunk
  vpc peer-link
  spanning-tree port type network
  spanning-tree guard loop

interface port-channel51
  switchport
  switchport mode trunk
  vpc 51
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60

interface port-channel52
  switchport
  switchport mode trunk
  vpc 52
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60

interface Ethernet2/9
  description tc-nexus5k01 - Eth2/2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60
```

```
  channel-group 51 mode active
  no shutdown

interface Ethernet2/10
  description tc-nexus5k02-eth2/2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 10-14,21-24,30,50-51,60
  channel-group 52 mode active
  no shutdown

interface Ethernet7/9
  description tc-core01 - TenGiga4/6
  ip address 10.1.1.13/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  no shutdown

interface Ethernet7/11
  description tc-nexus7k01-vdc2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3,10-14,21-24,30,50-51,60
  channel-group 10 mode active
  no shutdown

interface Ethernet7/13
  description tc-nexus7k01-vdc2
  switchport
  switchport mode trunk
  switchport trunk allowed vlan 3,10-14,21-24,30,50-51,60
  channel-group 10 mode active
  no shutdown


interface Ethernet8/16
  description tc-nexu7k02-vdc2 - vPC Heartbeat Link
  vrf member vpc-keepalive
  ip address 192.168.1.2/24
  no shutdown
```

```
interface mgmt0
  vrf member management
  ip address <ip>

interface loopback0
ip address 128.0.0.3/24

router ospf 1
  auto-cost reference-bandwidth 1000000
```

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 C07-572830-01 04/12